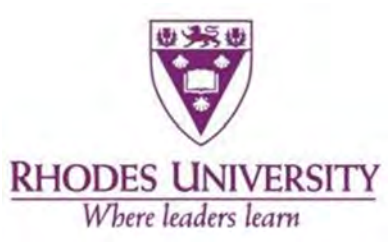


Sequence, structure, dynamics, and substrate specificity analyses of bacterial Glycoside Hydrolase 1 enzymes from several activities

A thesis submitted in fulfilment of the requirement for the degree

of

DOCTOR OF PHILOSOPHY
IN BIOINFORMATICS

RHODES UNIVERSITY, SOUTH AFRICA

Department of Biochemistry and Microbiology
Faculty of Science

by

Wayde Veldman

ORCID iD: 0000-0002-2747-7146

December 2021

Abstract

Glycoside hydrolase 1 (GH1) enzymes are a ubiquitous family of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety. Despite their conserved catalytic domain, these enzymes have many different enzyme activities and/or substrate specificities as a change of only a few residues in the active site can alter their function. Most GH1 active site residues are situated in loop regions, and it is known that enzymes are more likely to develop new functions (broad specificity) if they possess an active site with a high proportion of loops. Furthermore, the GH1 active site consists of several subsites and cooperative binding makes the binding affinity of sites difficult to measure because the properties of one subsite are influenced by the binding of the other subsites. Extensive knowledge of protein-ligand interactions is critical to the comprehension of biology at the molecular level. However, the structural determinants and molecular details of GH1 ligand specificity and affinity are very broad, highly complex, not well understood, and therefore still need to be clarified.

The aim of this study was to computationally characterise the activity of three newly solved GH1 crystallographic structures sent to us by our collaborators, and to provide evidence for their ligand-binding specificities. In addition, the differences in structural and biochemical contributions to enzyme specificity and/or function between different GH1 activities/enzymes was assessed, and the sequence/structure/function relationship of several activities of GH1 enzymes was analysed and compared. To accomplish the research aims, sequence analyses involving sequence identity, phylogenetics, and motif discovery were performed. As protein structure is more conserved than sequence, the discovered motifs were mapped to 3D structures for structural analysis and comparisons. To obtain information on enzyme mechanism or mode of action, as well as structure-

function relationship, computational methods such as docking, molecular dynamics, binding free energy calculations, and essential dynamics were implemented. These computational approaches can provide information on the active site, binding residues, protein-ligand interactions, binding affinity, conformational change, and most structural or dynamic elements that play a role in enzyme function.

The three new structures received from our collaborators are the first GH1 crystallographic structures from *Bacillus licheniformis* ever determined. As phospho-glycoside compounds were unavailable for purchase for use in activity assays, and as the active sites of the structures were absent of ligand, *in silico* docking and MD simulations were performed to provide evidence for their GH1 activities and substrate specificities. First though, the amino acid sequences of all known characterised bacterial GH1 enzymes were retrieved from the CAZy database and compared to the sequences of the three new *B. licheniformis* crystallographic structures which provided evidence of the putative 6Pβ-glucosidase activity of enzyme *Bl*BglH, and dual 6Pβ-glucosidase/6Pβ-galactosidase (dual-phospho) activity of enzymes *Bl*BglB and *Bl*BglC. As all three enzymes were determined to be putative 6Pβ-glycosidase activity enzymes, much of the thesis focused on the overall analysis and comparison of the 6Pβ-glucosidase, 6Pβ-galactosidase, and dual-phospho activities that make up the 6Pβ-glycosidases. The 6Pβ-glycosidase active site residues were identified through consensus of binding interactions using all known 6Pβ-glycosidase PDB structures complexed complete ligand substrates.

With regards to the 6Pβ-glucosidase activity, it was found that the L8b loop is longer and forms extra interactions with the L8a loop likely leading to increased L8 loop rigidity which would prevent the displacement of residue Ala423 ensuring a steric clash with galacto-configured ligands and may engender substrate specificity for gluco-configured ligands only. Also, during molecular dynamics simulations using enzyme *Bl*BglH (6Pβ-glucosidase

activity), it was revealed that the favourable binding of substrate stabilises the loops that surround and make up the enzyme active site.

Using the *Bl*BglC (dual-phospho activity) enzyme structure with either galacto- (PNP6Pgal) or gluco-configured (PNP6Pglc) ligands, MD simulations in triplicate revealed important details of the broad specificity of dual-phospho activity enzymes. The ligand O4 hydroxyl position is the only difference between PNP6Pgal and PNP6Pgal, and it was found that residues Gln23 and Trp433 bind strongly to the ligand O3 hydroxyl group in the PNP6Pgal-enzyme complex, but to the ligand O4 hydroxyl group in the PNP6Pglc-enzyme complex. Also, His124 formed many hydrogen bonds with the PNP6Pgal O3 hydroxyl group but had none with PNP6Pglc. Alternatively, residues Tyr173, Tyr301, Gln302 and Thr321 formed hydrogen bonds with PNP6Pglc but not PNP6Pgal.

Lastly, using multiple 3D structures from various GH1 activities, a large network of conserved interactions between active site residues (and other important residues) was uncovered, which most likely stabilise the loop regions that contain these residues, helping to retain their positions needed for binding molecules. Alternatively, there exists several differing residue-residue interactions when comparing each of the activities which could contribute towards individual activity substrate specificity by causing slightly different overall structure and malleability of the active site.

Altogether, the findings in this thesis shed light on the function, mechanisms, dynamics, and ligand-binding of GH1 enzymes – particularly of the 6Pβ-glycosidase activities.

Declaration

I Wayde Veldman, declare that this thesis is my own, unaided work, unless otherwise stated. It is being submitted for the degree of Doctor of Philosophy at Rhodes University. It has not been submitted before for any degree or examination in any other university.

20/01/2022

Acknowledgements

A huge thank you to my supervisor, Professor Özlem Taştan Bishop, for her guidance, motivation, and support (academic, financial, and moral) throughout my PhD. I am truly grateful for everything Prof Özlem, you gave me an opportunity and helped to make it a reality. Words are not enough to express my appreciation.

I thank my co-supervisor, Dr. Vuyani Moses, for his teachings during my transition into bioinformatics and for his advice at the start of my PhD.

Thank you to Professor Kevin Lobb, Dr. Magambo Phillip Kimuda, Dr. Olivier Sheik Amamuddy, Dr. Thommas Musyoka, Dr. Taremekedzwa Allan Sanyanga, Dr. Arnold Amusengeri, and Afrah Khairallah for their advice and assistance during my PhD.

I appreciate all the friendships I have made with RUBi members during my PhD, thank you to everyone.

A big thank you to our collaborators at the University of São Paulo, especially Dr. Marcelo Liberato and Professor Igor Polikarpov.

To the Centre for High Performance Computing (CHPC), South Africa, thank you for the computing resources.

The financial assistance of the National Research Foundation (Grant number 105267) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the National Research Foundation.

Thank you to the Ernst & Ethel Eriksen Trust for additional funding support.

Dedication

I dedicate this work to my parents, Hans and Ronel Veldman. Thank you for your love, support, and encouragement.

Table of contents

List of figures

List of tables

List of abbreviations

3D	Three-Dimensional
6Pβ-glycosidase	6-phospho-β-glycosidase
AMBER	Assisted Model Building with Energy Refinement
CAZy	Carbohydrate-Active enZyme database
CAZyme	Carbohydrate-Active enZyme
CHPC	Centre for High Performance Computing
CPU	Central Processing Unit
Dual-phospho	Dual 6-phospho-β-glucosidase/6-phospho-β-galactosidase
GH1	Glycoside Hydrolase family 1
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
LINCS	Linear Constraint Solver
MD	Molecular Dynamics
MAST	Motif Alignment and Search Tool
MEGA	Molecular Evolutionary Genetic Analysis
MEME	Multiple Expectation maximizations for Motif Elicitation
MM/PBSA	Molecular Mechanics/Poisson-Boltzmann Surface Area
MSA	Multiple Sequence Alignment
NCBI	National Centre for Biotechnology Information
NJ	Neighbour Joining
PBC	Periodic Boundary Conditions
PCA	Principal Component Analysis
PNP6Pgal	p-Nitrophenyl-beta-D-galactoside-6-phosphate
PNP6Pglc	p-Nitrophenyl-beta-D-glucoside-6-phosphate

PROMALS	PROfile Multiple Alignment with Local Structure
R_g	Radius of Gyration
RMSD	Root Mean Squared Deviation
RMSF	Root Mean Squared Fluctuation
vdW	van der Waals

List of tools and webservers

ACPYPE	https://github.com/llazzaro/acpype
AutoDock Vina	http://vina.scripps.edu/
CAZy	http://www.cazy.org/
DALI	http://ekhidna2.biocenter.helsinki.fi/dali/
DiscoveryStudio	https://discover.3ds.com/discovery-studio-visualizer-download
GROMACS	http://www.gromacs.org/
H++	http://biophysics.cs.vt.edu/
Jalview	http://www.jalview.org/
MATLAB	https://www.mathworks.com/products/matlab.html
Matplotlib	https://matplotlib.org/
MEGA	https://www.megasoftware.net/
MEME	https://meme-suite.org/
MODELLER	https://salilab.org/modeller/
Open Babel	http://openbabel.org
PDB	http://www.rcsb.org/
PROCHECK	https://saves.mbi.ucla.edu/
PyMOL	https://pymol.org
Python	https://www.python.org/
QMEAN	https://swissmodel.expasy.org/qmean/
R	https://www.r-project.org/
VERIFY3D	https://saves.mbi.ucla.edu/
VMD	https://www.ks.uiuc.edu/Research/vmd/
X-Score	http://www.umich.edu/~shaomengwanglab/software/xtool/

Research outputs

Research articles

1. **Wayde Veldman**, Marcelo Liberato, Vitor Almeida, Valquiria Souza, Maira Frutuoso, Sandro Marana, Vuyani Moses, Özlem Tastan Bishop, and Igor Polikarpov. "X-ray Structure, Bioinformatics Analysis, and Substrate Specificity of a 6-Phospho-β-glucosidase Glycoside Hydrolase 1 Enzyme from *Bacillus licheniformis*". *Journal of Chemical Information and Modeling*. 2020. 60 (12) 6392-6407. DOI: 10.1021/acs.jcim.0c00759.

   Contribution: I performed all computational experiments and data analyses. With Marcelo Liberato I wrote the first draft of the article, responded to reviewer's comments, and finalised the article. The direction and ideas in the article were my own.

2. **Wayde Veldman**, Marcelo Liberato, Valquiria Souza, Vitor Almeida, Sandro Marana, Özlem Tastan Bishop, and Igor Polikarpov. "Differences in Gluco and Galacto Substrate-Binding Interactions in a Dual 6Pβ-Glucosidase/6Pβ-Galactosidase Glycoside Hydrolase 1 Enzyme from *Bacillus licheniformis*". *Journal of Chemical Information and Modeling*. 2021. DOI: 10.1021/acs.jcim.1c00413.

   Contribution: I performed all computational experiments and data analyses. I wrote the first draft of the article, responded to reviewer's comments, and finalised the article. The direction and ideas in the article were my own.

Review article

Olivier Sheik Amamuddy, **Wayde Veldman**, Colleen Manyumwa, Afrah Khairallah, Steve Agajanian, Odeyemi Oluyemi, Gennady M. Verkhivker, and Özlem Tastan Bishop.

“Integrated Computational Approaches and Tools for Allosteric Drug Discovery”. *International Journal of Molecular Sciences.* 2020. 21 (3) 847. DOI: 10.3390/ijms21030847.

Poster presentation

Wayde Veldman, Vuyani Moses, Igor Polikarpov, and Özlem Tastan Bishop. “Elucidating the function and mechanism of Glycoside Hydrolase 1 enzymes from *Bacillus licheniformis*”. Centre for High Performance Computing (CHPC) national conference 2018, Cape Town, South Africa.

Oral presentations

Wayde Veldman, Vuyani Moses, Igor Polikarpov, and Özlem Tastan Bishop. “Elucidating the function and mechanism of Glycoside Hydrolase 1 enzymes from *Bacillus licheniformis*”. Rhodes University 2019 Postgraduate Conference.

Wayde Veldman, Marcelo Liberato, Vuyani Moses, Valquiria Souza, Vitor Almeida, Sandro Marana, Özlem Tastan Bishop, and Igor Polikarpov. “Differences in sequence and structure between bacterial Glycoside Hydrolase 1 enzyme-activities cause variations in function and substrate specificity”. SASBi-SC/SAGS 2021 Virtual Student Symposium.

Other:

Conference student volunteer

ISC High Performance conference 2019, Frankfurt, Germany.

Summer school attendance

Joint 9th CHPC introductory scientific programming school and the 30th Chris Engelbrecht (NITheP) Theoretical Physics Summer School, 2019, Drakensberg, South Africa.

Thesis overview and structure

The thesis is based on the role of the relationship between sequence, structure and function on enzyme activity and ligand specificity of bacterial Glycoside Hydrolase 1 (GH1) enzymes. As all three GH1 crystallographic structures received from our collaborators were shown to be putative 6Pβ-glycosidase activity enzymes, much of the thesis focuses on the overall analysis and comparison of the 6Pβ-glucosidase, 6Pβ-galactosidase, and dual-phospho activities that make up the 6Pβ-glycosidases. Chapters 2 and 3 utilise many of the same sequences for their sequence analysis sections, therefore the sequence analysis of Chapter 2 is far more in depth compared to Chapter 3. Although the methodology throughout the thesis is much the same, differences depending on the chapter merit a separate methodology section for each research chapter.

In Chapter 1, literature is reviewed. The relationship between sequence, structure, and function is introduced, as well as the process of characterising enzymes and the determination of enzyme substrate specificity. Information is given on GH1 enzymes, and the 6Pβ-glycosidase active site residues are identified through consensus of binding interactions using 6Pβ-glycosidase PDB structures complexed with ligands. Then, background information is provided on the methods used in the thesis.

In Chapter 2, the crystallographic structure of enzyme *Bl*BglH is analysed. Its sequence is compared to characterised bacterial GH1 enzymes, and *in silico* docking and molecular dynamics simulations reveal the activity and specificity of the enzyme. The contribution of the L8a loop to the substrate specificity of GH1 enzymes is researched. The dynamics of the enzyme is also discussed.

In Chapter 3, crystallographic structures of enzymes *Bl*BglB and *Bl*BglC are analysed. Their sequences are compared to characterised bacterial GH1 enzymes, and *in silico*

docking and molecular dynamics simulations reveal their activities and specificities. Important details of broad specificity are discovered and elaborated upon.

In Chapter 4, the residues and structures of many enzymes of the 6Pβ-glucosidase, 6Pβ-galactosidase, and dual-phospho activities are compared. Conserved differences and similarities between the activities in residue-residue interactions are discovered.

In Chapter 5, the findings in the thesis are reported, and potential future work is discussed.

Chapter 1: Literature review

This chapter is divided into three main parts. Part 1 provides background information regarding the main theme of the thesis. It describes the relationship between enzyme sequence, structure, and function, as well as the process of characterising enzymes and the determination of enzyme substrate specificity. In Part 2, relevant information is given surrounding the Glycoside Hydrolase 1 (GH1) enzymes obtained from the collaborators. It starts off with the bacterium *Bacillus licheniformis*, the phosphoenolpyruvate-dependent phosphotransferase system, carbohydrate-active enzymes, and carbohydrates. Then, GH1 enzymes are described in detail and the 6Pβ-glycosidase active site residues are identified through consensus of binding interactions using 6Pβ-glycosidase PDB structures complexed with ligands. Lastly, Part 3 provides extensive information concerning the computational methods used in the thesis.

- ***PART 1***

1.1 Enzymes: The relationship between sequence, structure, and function

The analyses of a wide variety of proteins from different families have conclusively found that the amino acid sequence of a protein/enzyme determines its three-dimensional (3D) structure, and that the 3D structure determines its function [1,2]. In other words: function is derived from structure, and structure is derived from sequence.

Generally, the higher the sequence similarity between two enzymes, the higher the chances are of the enzymes having similar 3D structures and therefore similar functions [1–3]. More related proteins have higher sequence similarity compared to less related

proteins due to fewer accumulated genetic mutations over evolutionary time. So, proteins that share a relatively recent common ancestor have similar sequences, structures, and functions [1–3]. Principally, a 30% sequence identity between protein sequences is likely to translate into similar 3D structures. However, care must be taken when assigning a function to a protein based only on sequence similarity as some examples exist of functionally unrelated proteins with similar sequences, and vice-versa. An example of this is the yeast Gal1 and Gal3 proteins that share 73% sequence identity but have largely different functions; Gal1 is a galactokinase and Gal3 is a transcriptional inducer [4]. There are even some instances of proteins having unrelated sequences *and* functions but have similar structures. Even so, the comparison of enzyme sequences is very valuable because the order of amino acids is the fundamental starting point of analysis of all proteins (primary structure).

The “native state” or 3D structure of an enzyme is established when its linear chain of amino acids completes the spontaneous folding process, whereby mostly noncovalent interactions between regions in the sequence of amino acids causes the chain to fold into a functional enzyme [1,5,6]. The initial step in protein folding is the formation of secondary structure (α-helices and β-sheets), which is stabilised by hydrogen bonds. Then, tertiary structure arises when amino acid nonpolar side groups form hydrophobic interactions, as well as ionic interactions and hydrogen bonding between polar side groups and the polypeptide backbone. The hydrophobic core of the protein is formed as the protein folds, after the hydrophobic contacts promote the expulsion of water from the immediate vicinity of nonpolar residues. All these interactions contribute to the folding and stabilisation of the protein. Interestingly, the overall structures of proteins have constant but minor fluctuations due to the mostly weak stabilising interactions [1].

With all this in mind, it is obvious that studying the sequence, structure, and function of proteins is extremely important in the field of biological science and for understanding life.

1.2 Enzyme substrate specificity

A ligand, or substrate, is the molecule to which proteins bind. To catalyse a chemical reaction, enzymes are first required to bind to their substrates [7]. The enzyme active site is composed of the substrate-binding site and the catalytic site. The substrate-binding site recognises and binds the substrate, whereas the catalytic site executes the chemical reaction [1].

Just like the forces that fold polypeptides and determine native protein structure, substrates and other molecules use noncovalent forces to bind to enzymes [8]. Enzyme substrate specificity is generally described as the preference of a protein to bind one particular molecule or a very small group of molecules, and it depends on the structure and chemical properties of both the substrate-binding site and the substrate itself (molecular complementarity) [1]. A substrate-binding site is usually found on the surface of an enzyme molecule in the form of a pocket or crevice that has a complementary shape to that of the substrate (geometric complementarity). Often the charged amino acid residues in the binding site, or residues around the binding site, are organized in a particular fashion to attract the substrate (electronic complementarity) [8]. Specificity can also be accomplished when the binding site has complementary hydrophilic/hydrophobic features to the substrate [7]. Therefore, molecules that have a slightly different geometry or functional group distribution from the native substrate do not bind effectively to the enzyme (chemoselective, regioselective and stereospecific).

Enzyme-substrate complementarity or specificity was first proposed by Emil Fischer in 1894 and is the foundation of the “lock-and-key” model of enzyme function [9]. This early model explains enzyme specificity but does not account for the stabilisation of the transition state that enzymes reach [10]. In 1958, Daniel Koshland proposed the “induced fit” model, which suggests that the substrate-binding site is flexible and changes shape in

order to optimize catalysis during substrate binding [1,11]. X-ray studies now confirm substrate-binding sites to be mostly preformed but change conformations upon substrate binding [8]. The structure of the substrate-binding site fluctuates until the substrate is optimally bound. For certain enzymes like glycosidases, the substrate also changes conformation when entering the substrate-binding site [12]. Because molecular recognition occurs in a congested biological environment filled with many molecules and potential substrates, the induced fit mechanism may improve enzyme-substrate specificity [13].

The current third and final model – the conformational selection model – draws from the free energy landscape (FEL) theory of protein structure and dynamics [14–17]. This model proposes that the native state of a protein exists as an extensive ensemble of conformational states/substates that coexist in equilibrium with various population distributions, and that the ligand binds selectively to the most suitable conformational state/substate, eventually shifting the equilibrium towards this state/substate [18].

All three of these distinct conceptual models have been observed experimentally, so it is noteworthy that all three mechanisms can occur in both a simultaneous and sequential way that includes a range of binding events [19].

1.2.1 Broad specificity

As mentioned before, enzymes have a substrate preference of one particular molecule or a very small group of molecules. The ability of enzymes to catalyse and convert more than one substrate is called broad substrate specificity. Most enzymes are capable of catalysing the reactions of a limited number of related compounds, but at varying efficiencies [8,20–22]. Sometimes however, they can possess a secondary enzyme activity that is unrelated to the primary activity that the protein evolved to perform, this is called enzyme promiscuity and it is a form of broad specificity [23]. Generally, enzymes have evolved to catalyse a single reaction or one class of reaction, and the function of the enzyme will determine the

degree of specificity [24]. Enzymes with broad substrate specificity can be very helpful to us, such as cytochrome P-450. The various possible uses of cytochrome P-450 enzymes involve the biocatalytic generation of drug metabolites, the breakdown of recalcitrant toxins, and they catalyse the hydroxylation and/or epoxidation of several classes of compounds, ranging from alkanes to heterocycles [24–26]. On the other hand, the restriction endonuclease enzymes have only one substrate.

Broad specificity, or enzyme promiscuity, can develop by chance or can be derived from evolution [27,28]. The structural and molecular basis of broad substrate specificity has been widely researched in many fields such as drug design and in industrial applications [29–33]. Broad specificity provides functional benefits to the cell using several mechanisms like scavenging of nutrients, proofreading, removal of antimetabolites, balancing of metabolite pools, and establishing system redundancy [34]. Studying the broad substrate specificity of enzymes leads to an understanding of how enzymes evolve and could be a useful starting point in directed evolution research. Novel or transformed enzymatic activities can emerge suddenly which demonstrates the speed at which some enzymes can evolve, even in a natural environment [24]. For instance, enzymes are often the targets of antibiotics, and sometimes resistant forms of these enzymes can manifest within months [35,36]. Certain synthetic chemicals have existed in nature for only a limited period, yet enzymes have been found to utilise and degrade these chemicals. For example, bacterial enzymes used by the cell to break down lactones have evolved to utilise organophosphate pesticides [37].

Interestingly, a computational analysis performed by Dellus-Gur *et al.* [38] found that enzymes are more likely to have multiple functions or substrates if they possess an active site with a high proportion of loops. In other words, enzyme folds with a high percentage of active site residues that are not part of the protein scaffold may be better at acquiring new functions. This is probably why the TIM barrel and Rossmann folds, among others, are

associated with multiple functions [39]. In contrast, with enzymes that have only one known function like dihydrofolate reductase (DHFR), the active site and scaffold co-evolve because they are mostly joined – this leads to greater constraints and fewer chances of novel function acquisition [38]. Furthermore, the mobility of active site loops makes them adaptable, leading to the utilisation of different substrates [34,40,41].

1.3 Characterisation of enzyme activity or function

There are many ways to characterise or predict enzyme activity or function. Although one can determine protein function using experimental methods, technological leaps forward in sequencing has dramatically increased the number of new sequences that need to be characterised [42]. New sequences are therefore mainly annotated by prediction using computational methods.

1.3.1 *In vitro* enzyme assays

The most accurate method of protein characterisation remains the use of laboratory enzyme activity assays. After cloning, expression, and purification of an enzyme, activity assays can be performed to identify the substrates that the enzyme can utilise, as well as the enzyme's rate of reaction under specific conditions. Enzyme assays are laboratory methods that are used to measure enzymatic activity regarding either the consumption of substrate or production of product over time [43,44]. Leonor Michaelis and Maud Menten [45] were the ones to discover that enzyme activity is influenced by certain factors such as temperature, pH, the nature and strength of ions, and the concentrations of the enzymes and substrates [46]. The reaction progression can be observed continuously (continuous assay) using spectroscopic [47–49] or electrochemical methods which show the full progress curve [50].

Recently, Helbert *et al*. [51] measured the degradation of several substrates with 564 carbohydrate-active enzymes using colorimetric reducing assays and size exclusion

chromatography. This expanded the collection of biochemically characterised subfamilies and resulted in the discovery of new enzyme families and unknown substrate specificities. Today, the major impediment to biochemical functional annotation is not protein production, but the accessibility of substrates [51].

1.3.2 Protein sequencing

Introduced in 1967, Edman degradation [52] was the process used to sequence most proteins prior to the early 1990's. During Edman degradation, several steps of chemical degradation are performed using a single peptide in order to determine its sequence. When performed correctly, the Edman process is accurate with >99% efficiency per amino acid. However it is rather slow, as one cycle runs for approximately an hour and is limited to peptides consisting of 30 residues or fewer [53–56]. Edman degradation has mostly been replaced by higher throughput technologies. Mass spectrometry methods are currently principally utilised for protein sequencing, although Edman degradation is still important for characterising a protein's N-terminus.

Today, a bottom-up approach to mass spectrometry (BU-MS) is mostly used [53,57–59]. BU-MS entails protein enzymatic digestion, ionization of the peptide products, ion separation based on their mass/charge ratio, followed by ion detection. The masses of the "tryptic peptides" are analysed by electrospray ionization or matrix-assisted laser desorption/ionization (MALDI). The ions are then fragmented to acquire details about the peptide sequences from MS [59,60]. BU-MS is not very sensitive because it does not actually sequence the protein, it infers the primary structure or classifies the protein [59–63].

1.3.3 Sequence analyses to predict enzyme function

The function prediction of the enormous amount uncharacterised proteins is crucial for understanding the role of enzymes and is a mammoth job for bioinformaticians. Protein

functions are mostly predicted from protein sequences [64–70]. This is mainly done using sequence similarity, searching for sequence domains, or performing multiple sequence alignments (MSA) to infer functions based on homologous proteins with known functions. Programs like BLAST [71] are probably used the most for computational function-prediction, which is based on the assumption that two highly similar sequences most likely evolved from a common ancestor and therefore have similar functions. In other words, if a query protein shares significant sequence similarity to a protein that has a known function, then the function of the latter can be transferred to the former [72].

To obtain potential enzyme functional information, known domains within a query sequence can be searched for using protein domain databases like Pfam (Protein Families Database) [73]. And within protein domains, shorter sequence signatures called motifs are linked to certain functions [72]. Motif databases like PROSITE can be used to search for enzyme sequence motifs [74]. Even when a great difference exists between the whole sequences of two enzymes, they can often share important motifs (such as active site motifs) from which protein function can be inferred [75–77]. If the identified short, conserved sequence motifs are crucial to the function of the protein, this method can predict the function and activity with a large amount of accuracy and determine function-related subfamilies of glycoside hydrolases [77] .

1.3.4 Other ways of characterising protein function

1.3.4.1 Structure-based

Protein structure is known to have higher conservation as compared to protein sequence, therefore proteins with high structural similarity most likely have a similar function [72,78]. An unknown protein structure can be screened against the Protein Data Bank [79] with a multitude of programs like CE-ME [80], DALI [81], and FATCAT [82] to find a characterised

enzyme structure that most closely resembles the unknown structure, thus inferring function.

1.3.4.2 Genomic context-based

Several novel ways of protein function prediction do not use sequence or structure comparisons but rely on links between unknown genes or proteins and those have been annotated – this is sometimes called phylogenomic profiling. The concept is that the dual presence or dual absence of two traits throughout many species can infer a significant biological association, like the participation of two different proteins in the same biological pathway [72,83]. Unlike homology-based methods where molecular functions of a protein are determined, context-based methods predict the cellular function or biological process in which a protein operates [42,83].

Automated predictions of protein function from DNA sequences by computer algorithms have resulted in the establishment of large databases containing protein sequences and their functional information such as UniProt [84].

1.3.4.3 Network-based methods

Computer algorithms can also create a functional association network for a certain cluster of genes or proteins [85,86]. Many nodes representing genes or proteins are linked by edges that represent evidence of shared function [87]. Many networks using various sources of data can be integrated and subsequently utilised by prediction algorithms to annotate unknown genes or proteins [88]. Recently, machine learning has been used to possibly improve these methods [64,89,90].

1.3.4.4 In silico substrate determination

Assigning protein function based on sequence or structure is remarkably difficult [78,91]. Even if two proteins are highly homologous and have similar structures, a change of only a

few residues in the active site can change the functional specificity [78,91]. In addition, sequence similarity and/or genome/operon context can only provide indications of enzyme function, they do not yield information regarding substrate specificity and the catalysed reaction [92]. Many studies have used computational screening/docking to investigate the substrate specificity of enzymes [93–98]. Protein experimental 3D structures or homology models are used to screen putative substrates. Computational screening is much faster and cheaper than physical assays. Also, it is not limited to commercially available compounds or readily synthesized compounds, because any substrate imaginable can be computationally constructed and screened [92]. The *in silico* metabolite docking method has proven feasible and valuable in both retrospective and prospective tests [99]. However, very little testing of docking and scoring methods for enzyme-substrate recognition has taken place compared to the binding of drug-like molecules to drug targets [99]. Even so, while studying the glycolysis pathway Kalyanaraman and Jacobson [99] showed that computational methods are viable and can exclude a big proportion of the metabolome. Caution must be exercised though, as the accuracy of *in silico* docking is not perfect; in fact, new research shows that in datasets classically utilised in virtual screening challenges, there are accidental biases that cause overestimation of virtual screening accuracy [100,101].

Molecular dynamics (MD) simulations can complement computational substrate screening/docking. If the docked substrate remains in the protein active site, in the correct orientation, throughout an MD simulation of length at least equal to the substrate-protein half-life, and both the substrate and protein remain stable, it is a conservative indication of substrate specificity.

1.4 Determination of substrate interactions with active site and the effects of substrate binding

Knowledge of the inner workings of enzymes can be helpful to understand and improve their use in industrial applications [102]. Extensive knowledge of protein-ligand interactions is critical to the comprehension of biology at the molecular level. Several experimental methods exist that can research many features of protein-ligand binding. X-ray crystallography, nuclear magnetic resonance (NMR), Laue X-ray diffraction, small-angle X-ray scattering, and cryo-electron microscopy supply protein structures at atomic-resolution or near-atomic-resolution. These protein structures can be resolved with and without ligands, and could show differences in structure and/or dynamics between the ligand-free and ligand-bound proteins. These methods also provide important binding details like interactions with the enzyme. However, researching binding affinity using these methods are difficult, long, and costly [18]. Also, studying the dynamics of enzyme-substrate interactions is very challenging when using experimental methods alone [103].

In silico methods can help to resolve and predict experimental results and have become a sophisticated research tool to use side by side with experimental methods [104]. It is very important to establish theoretical methods that will help to understand existing experimental data, and also form the basis of new experiments. Theoretical/computational techniques are starting to become very useful in elaborating the function/mechanism of enzymes, and the future possibilities are even more exciting [18].

In order to obtain information on an enzyme's mechanism or mode of action, as well as its structure-function relationship, computational methods such as docking, molecular dynamics and binding free energy calculations can be implemented [92,103,105–116]. These computational approaches can provide information on the active site, protein-ligand interactions, binding residues, binding affinity, carbohydrate processivity (dependent on

simulation length and system characteristics), conformational change, and many structural or dynamic elements that play a role in the enzyme's function. When researching protein-ligand binding, structure-based computational methods are beneficial to all situations. MD simulations show time-dependent changes in atomic coordinates of the protein and ligand, in bound and unbound forms – this can be used to obtain information on the conformational entropy change upon binding. MD simulations are used to study the non-equilibrium effects that give rise to transient protein conformers, which have a role in the binding event but is difficult to observe experimentally [18]. Today, simulation times are matching those applicable to most biological occurrences [117], but these situations are still very limited; MD simulations exploring the conformational space of peptides and proteins can fold small proteins (less than 80 amino acids) to their native structures [118], and the special-purpose supercomputer *Anton* used all-atom MD simulation to obtain a millisecond time-scale [119,120]. One can conservatively identify significant/functional residues by running molecular dynamics simulations on an enzyme structure after having first computationally substituted residues [110]. The optimal enzyme functional conditions with regards to its environment can be determined using simulations at various pH and/or temperature values [121,122]. Multiple MD simulations can be run at specific pH's, where the protonation of titratable residues will change depending on the pH. Recently, researchers have developed a method whereby the pH changes slowly during a consecutive series of MD simulations in a defined direction [120]. To fully encapsulate structural diversity, the simulations must be of sufficient length [106]. Fortunately, if a desired protein structure is unavailable, homology modelling, threading, or *ab initio* prediction approaches can construct a protein model for use in computational studies.

In summary, evidence clearly suggests that MD simulations are useful for investigating enzyme function and mechanisms, including estimating substrate binding affinities of

glycosidases, and predicting the amino acid residues involved in their substrate recognition [105,106,116,123,124].

- ***PART 2***

1.5 Bacillus licheniformis

In this thesis, the 3D structures and substrate specificity details of three GH1 enzymes from *B. licheniformis* were investigated using structural biology, structural bioinformatics, and wet-lab approaches. *B. licheniformis* is a Gram-positive mesophilic bacterium that is mostly found in soil but is also abundantly found on the feathers of birds. The bacterium generates a wide range of extracellular enzymes that have a role in cycling nutrients found in nature [125], including carbohydrate-active enzymes [126–128]. This bacterial species is well suited for use in industry because of properties such as its high yield of target proteins, its ease of genetic manipulation, its favourable fermentation conditions, its current status as *generally recognized as safe* (GRAS), and its probiotic attributes [129–132]. The optimal temperatures of growth and enzyme secretion for *B. licheniformis* are 50 °C and 37 °C respectively, which contributes to a low contamination risk and low consumption of energy for cooling the fermentation vessel [133].

1.6 The phosphoenolpyruvate-dependent phosphotransferase system

Glycoside hydrolases are involved in the first step the phosphoenolpyruvate (PEP):carbohydrate phosphotransferase system (PTS), also known as PEP group translocation. The system was initially discovered in the laboratory of Saul Roseman at the University of Michigan in 1964. The subsequent publication concerned the Horseradish Peroxidase (HPr) enzyme from *E. coli* and its function in hexose phosphorylation [134].

The system is utilised by a multitude of bacteria (both Gram-positive and Gram-negative bacteria), as well as some archaea, for the uptake of organic compounds such as sugars and sugar derivatives including cellobiose, lactose, sucrose, and trehalose [135–138]. PEP acts as a phosphoryl donor and an energy source while the PTS transport system catalyses the concomitant phosphorylation and translocation of these compounds through the cytoplasmic membrane in a single energy-coupled step [136,137,139]. Inside the cell, the majority of the phosphorylated compounds are used up in glycolysis; the initial and determinant step to enter the pathway is the cleavage of the glycosidic bond promoted by glycoside hydrolases such as 6-phospho-beta-glucosidases (6Pβ-glucosidases) and 6-phospho-beta-galactosidases (6Pβ-galactosidases).

Generally, the PTS consists of one membrane-spanning protein and four soluble proteins. Most microorganisms use enzymes I (EI) and HPr for the uptake of all PTS carbohydrates; these two enzymes are the main elements of the cytoplasmic PTS. The EIIA, EIIB, and EIIC enzymes, however, are mostly specific for only one substrate. EIIC is the integral membrane sugar permease [136,137,139,140]. A three letter code is added as superscript to the enzyme names that signifies their substrate specificity [141]. For instance, EIIA Glc would be glucose-specific, EIIB Fru is fructose-specific, and so on. Not long after the PTS was discovered it was shown that the PTS performs regulatory functions related to carbon metabolism and sugar transport, in addition to transporting and phosphorylating carbohydrates.

Interestingly, as yet, there is no known eukaryote that uses the PTS and, therefore, the enzymes in this system could be potential targets of antimicrobial agents [142].

1.7 Carbohydrate-active enzymes

The enzymes responsible for the synthesis, degradation, and modification of carbohydrates are the Carbohydrate-Active enZymes (CAZymes) [143–145]. Very many

CAZymes exist, having many different enzyme families each with several activities (specificities) – this is because there are more carbohydrates on our planet than any other biomolecule, with extreme structural diversity. Living organisms use carbohydrates for a wide range of things, from cell wall structure (cellulose, chitin, starch, or glycogen) to intra- and intercellular recognition within one organism or between organisms. In fact, about 1-2% of any organism's genome codes for CAZymes [146–149]. The synthesis and degradation of glycosidic bonds is mostly performed by glycosyltransferases and glycoside hydrolases, respectively [146], and CAZymes have various industrial applications ranging from biofuel production to drug design [150]. Interestingly, CAZymes as well as their substrates are unique in that they are both particularly flexible [151].

1.7.1 Carbohydrates

Even though carbohydrates share similar chemical composition, they can be arranged in a huge amount of combinations [147]. Individual monosaccharide units can be assembled to form oligo- and polysaccharides, with a glycosidic bond between the anomeric position of one sugar and the hydroxyl group of another [152,153]. In nature there exists a combinatorially-large amount of carbohydrate structures because of the many hydroxyl groups on every sugar, the potential of two possible anomeric configurations of sugars, and the possibility of different ring sizes [154]. Also, many noncarbohydrate substituents can be linked to carbohydrates, resulting in a great number and range of glycoconjugates [147].

1.7.2 The Carbohydrate-Active Enzymes database

The classification of CAZymes and information regarding protein-sugar interactions can be found on databases. The databases and associated tools can be utilised to plan and initiate high-resolution computer simulation and modelling [147,155–157], including the

study of carbohydrate recognition [112]. Based on sequence, CAZymes have been classified into families as early as the year 1991 [158].

Introduced in 1999, the Carbohydrate-Active Enzymes database (CAZy) provides continuously updated information regarding the classification of CAZymes [147], and is the only database that correlates the sequence, structure and molecular mechanism of CAZymes. A big contribution of CAZy is the establishment of a stable nomenclature for CAZymes; the enzymes are grouped into families and subfamilies and this classification has become the standard of the field [149]. In the CAZy classification system, CAZyme families are defined by sequences that cluster around at least one biochemically characterised member [147], and is based on the concept that sequence defines protein structure, and protein structure defines function. Sequence-based classification methods are quite different, but also complementary, to the Enzyme Commission classification scheme (EC numbers) which categorizes proteins according to the reactions that they catalyse. The CAZy classification (a) is based more on structural characteristics than substrate specificity, (b) aids in revealing the evolutionary associations amongst CAZymes and (c) provides a useful foundation to decipher mechanistic features [159].

The solution to substrate specificity prediction is to research how CAZymes realize selective recognition of ligands that have very subtle stereochemical differences. Although this is now a reality for several subfamilies, the CAZy team admit that they are still far from accurate automated substrate (and/or product) prediction for all CAZymes encoded by a genome [147]. As is the case for general protein databases, a similarity search used with the CAZy database mostly produces uncharacterised or unreliably named gene products, and therefore no reliable functional inference [51]. Following sequence-based functional predictions, CAZymes of differing substrate specificity frequently fall into the same enzyme family [158]. The outcome is broad functional categorization, like “putative glycoside hydrolase,” and no reliable prediction of the enzyme substrate is given. However, it has

been shown that dividing large multifunctional glycoside hydrolase families into subfamilies results in far fewer substrate specificities in each subfamily which improves functional prediction [160–163].

1.8 Glycoside Hydrolase enzymes

Glycoside hydrolases (GHs), or glycosidases, catalyse the hydrolysis of terminal sugar residues from the non-reducing ends of a broad-spectrum of glycosylated compounds. GHs are omnipresent in all domains of life and are used for many functions [164]. There exists a large amount of genes that code for GHs in the genomes of most organisms. In fact, GHs account for approximately 44% of the enzymes on the entire CAZy database. In prokaryotes, they act as intracellular and extracellular enzymes that are mostly involved in nutrient acquisition. In higher organisms, glycoside hydrolases have a role in the biosynthesis and degradation of glycogen in the body, and they are found in the endoplasmic reticulum and Golgi apparatus where they process N-linked glycoproteins. GHs are the most widely studied and the best biochemically characterised CAZymes due to their use (current and potential) in biotechnological applications [159], including the degradation of biomass such as cellulose (cellulase), hemicellulose, and starch (amylase).

Glycoside hydrolases are usually named after their substrate, for example glucosidases catalyse the hydrolysis of glucosides and xylanases catalyse the cleavage of xylan. Other examples include lactase, amylase, chitinase, sucrase, maltase, etc. The CAZy database holds information regarding the putative function and reaction mechanisms (activity) of all currently available GH sequences [165], and is a good resource for researching GH evolution and biology [166]. Unfortunately, it is not very easy to predict GH enzymatic activity based on sequence [167], so CAZy splits the GHs into 172 protein families (as of thesis submission) based on sequence and structural information [106,147,168,169]. Because of convergent evolution, most GH families consist of enzymes with various

functions [170]. The GH1 and GH5 families possess great plasticity and can utilise a wide array of substrates, meaning it is not possible to predict GH activity by assigning a protein to a GH family [77]. However, approximately only one third of GH families are somewhat specific and hydrolyse only one type of substrate [103].

GH classification into families does enable numerous predictions with regards to the catalytic machinery and molecular mechanism as these are conserved in nearly all GH families [171], this also includes the geometry around the glycosidic bond [168]. Although there are very many known GH families, they share a common catalytic mechanism: acid/base catalysis with either retention or inversion of the anomeric configuration. Exceptions are family GH97 which contain both retaining and inverting enzymes [172], and families GH4 and GH109 which use an NAD-dependent hydrolysis mechanism [173,174].

1.9 Glycoside Hydrolase Family 1 enzymes

CAZymes utilise β-strands in their secondary structure to form recognition motifs and generally shape the binding site [112]. The 3-dimensional structure of glycoside hydrolase 1 (GH1) enzymes exist as a conserved $(\beta/\alpha)_8$-barrel core made up of consecutive (β/α) motifs that are joined by short loops (Figure 1.1) [175]. Despite their conserved catalytic domain, GH1 enzymes have many different substrate specificities or activities. Currently, there are 22 defined GH1 activities on the CAZy database, which include: β-glucosidase (EC 3.2.1.21), β-galactosidase (EC 3.2.1.23), β-mannosidase (EC 3.2.1.25), β-glucuronidase (EC 3.2.1.31), β-xylosidase (EC 3.2.1.37), β-D-fucosidase (EC 3.2.1.38), phlorizin hydrolase (EC 3.2.1.62), exo-β-1,4-glucanase (EC 3.2.1.74), 6-phospho-β-galactosidase (EC 3.2.1.85), 6-phospho-β-glucosidase (EC 3.2.1.86), strictosidine β-glucosidase (EC 3.2.1.105), lactase (EC 3.2.1.108), amygdalin β-glucosidase (EC 3.2.1.117), prunasin β-glucosidase (EC 3.2.1.118), vicianin hydrolase (EC 3.2.1.119), raucaffricine β-glucosidase (EC 3.2.1.125), thioglucosidase (EC 3.2.1.147), β-

primeverosidase (EC 3.2.1.149), isoflavonoid 7-O-β-apiosyl-β-glucosidase (EC 3.2.1.161), ABA-specific β-glucosidase (EC 3.2.1.175), DIMBOA β-glucosidase (EC 3.2.1.182), and protodioscin 26-O-I^2-D-glucosidase (EC 3.2.1.186). The multiple functions most likely exist because most active site residues of GH1 enzymes are located in loop regions [31], [35], [37], [38].

β-glucosidases are the most intensely studied due to their application in biofuel production [176]. Despite the vital importance of 6-phospho-β-glucosidases (6Pβ-glucosidases) and 6-phospho-β-galactosidases (6Pβ-galactosidases) for bacterial energetic balance, a reduced number of these enzymes have been previously characterised in comparison to β-glucosidases. Our collaborators sent us three X-ray crystallographic structures from the bacterium *B. licheniformis* and based on initial sequence and structural comparisons they were all classified as putative 6Pβ-glycosidases.

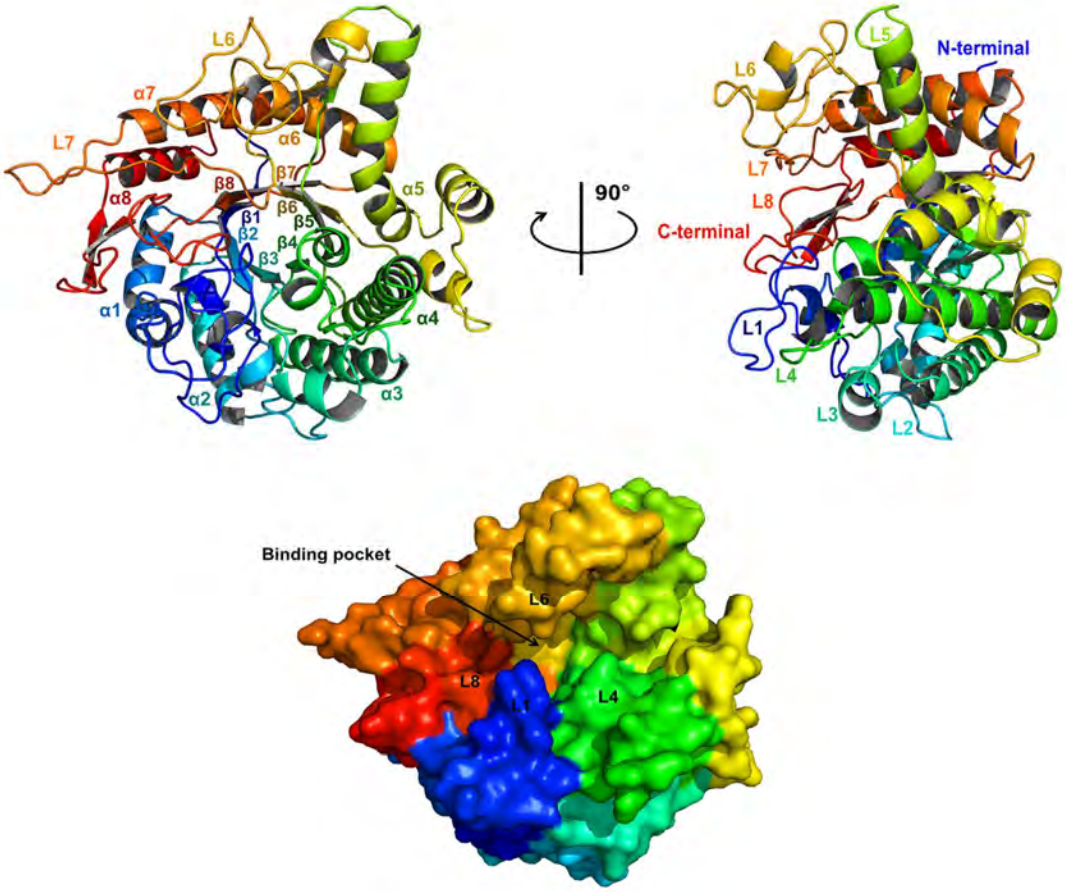


Figure 1.1. *3D structure of GH1 enzymes. Conserved (β/α)8-barrel core formed by consecutive (β/α) motifs joined by short loops. Here, a 6Pβ-glucosidase is shown (PDB ID 6WGD). Adapted from Veldman et al. 2020 [103].*

1.9.1 GH1 substrates

To understand the structure, activities, and specificity of GH1 enzymes, one must have knowledge of the GH1 substrates. GH1 enzymes (i.e., β-glycosidases) cleave β-glycosidic bonds in cello-oligosaccharides and other small substrates resulting in glucose and other monosaccharides. One monosaccharide is removed from the nonreducing end of the substrate in each catalytic cycle [177]. All members of the GH1 family cleave b-glycosidic bonds between a pyranosyl glycon and an aglycon [178]. The glycon is the monosaccharide of the substrate nonreducing end, and the remaining moiety is called the aglycon (Figure 1.2). Additionally, substrates of the 6Pβ-glycosidase activities have a phosphate group attached to the glycon.

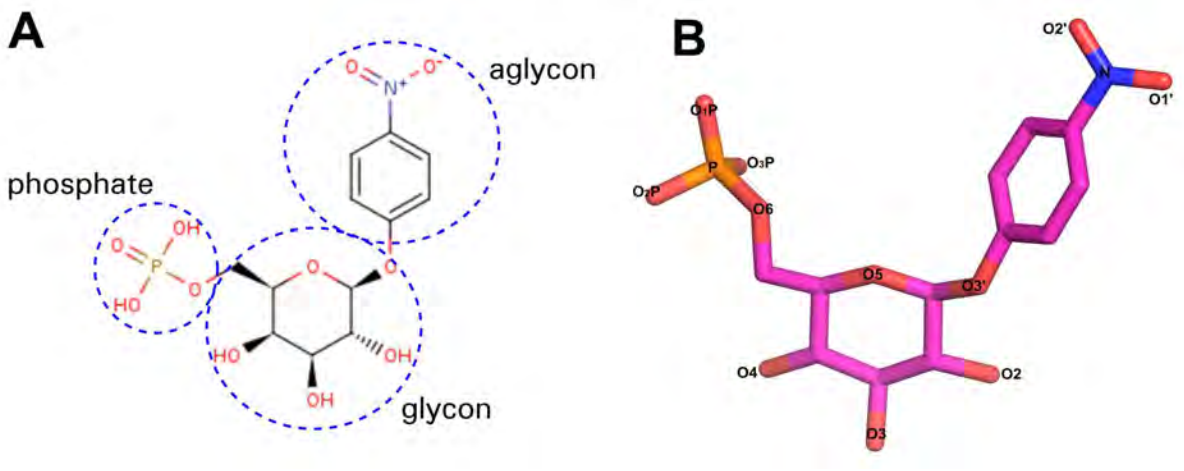


Figure 1.2. *GH1 enzyme ligand information. (A) Phospho-glycoside showing the phosphate, glycon and aglycon ligand groups. (B) Atom identifiers of ligand.*

The 21 different activities of GH1 enzymes highlight the diversity of substrates that these enzymes hydrolyse. Glucose, galactose, fucose, mannose, xylose, 6-phospho-glucose and 6-phospho-galactose are all glycones that are recognized by the GH1 family. Slight modifications in the glycon structure can cause considerable changes in GH1 activity. There is even greater diversity of aglycons, which include monosaccharides, oligosaccharides, and aryl or alkyl groups [177]. Some GH1 enzymes are specific for only one type of aglycon whereas others show broad specificity.

As the work in this thesis mostly concerns GH1 enzymes of the β-glucosidase, 6Pβ-glucosidase, and 6Pβ-galactosidase activities, the main reactions of these activities according to MetaCyc database [179] are shown in Figure 1.3. The only difference between galacto- and gluco-configured ligands is the position of the glycon O4 hydroxyl group. The galacto epimer's O4 hydroxyl group has an axial position, whereas the gluco epimer's O4 hydroxyl group has an equatorial position.

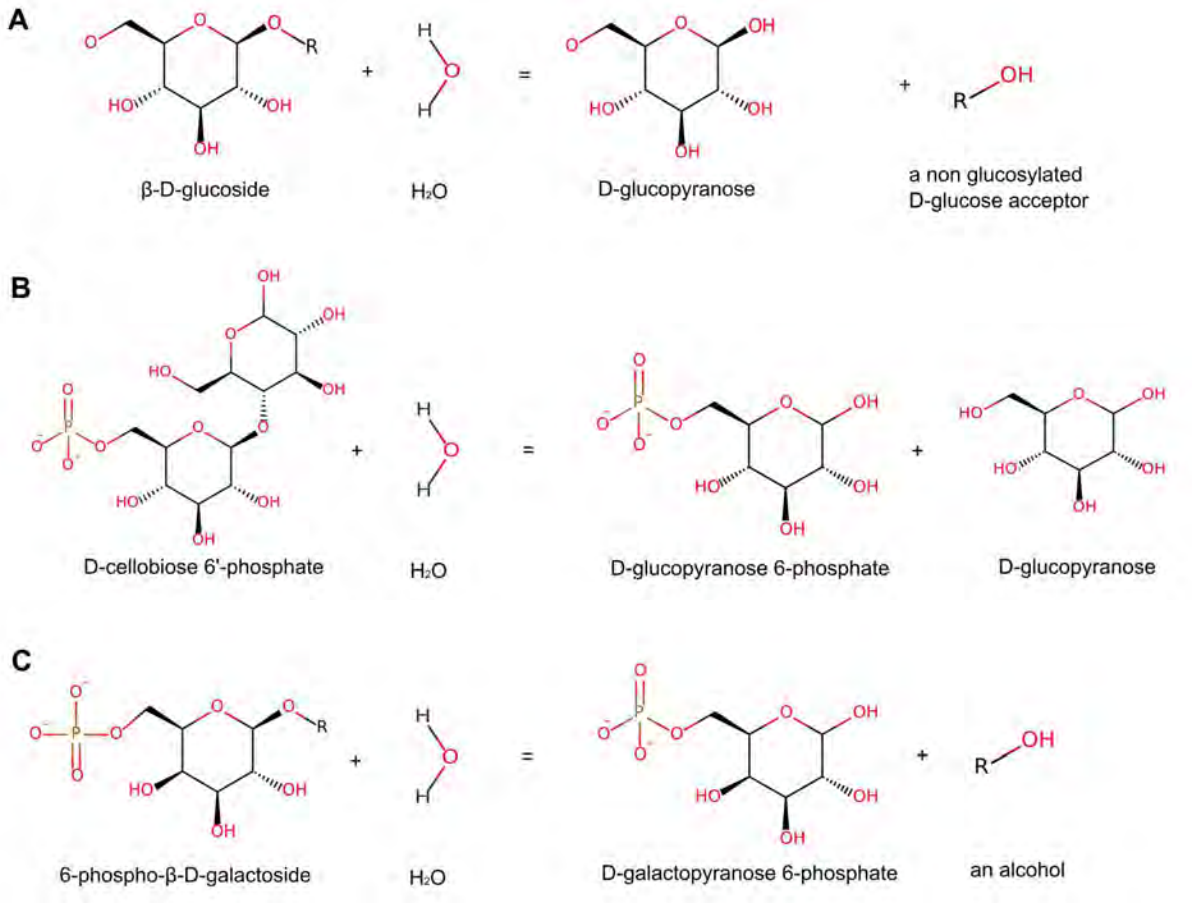


Figure 1.3. *Main reactions of the activities studied in the thesis. (A) β-glucosidase, (B) 6Pβ-glucosidase, and (C) 6Pβ-galactosidase.*

1.9.2 Koshland mechanism for retaining enzymes

GH1 enzymes use a double-displacement mechanism of catalysis to hydrolyse their substrates by retention of configuration at the anomeric carbon atom (Figure 1.4) [180]. Two conserved glutamic acid catalytic residues play a major role in this mechanism – one residue acts as an acid/base and the other a nucleophile. In the first reaction step, the nucleophilic catalytic glutamic acid attacks an electrophilic anomeric carbon atom resulting in the formation of a covalent glycosyl-enzyme intermediate while at the same time the other catalytic residue acts as an acid that protonates the glycosidic oxygen atom, facilitating the exit of the aglycon. During the next reaction step, the now deprotonated acidic carboxylate catalytic residue acts as a base that deprotonates a water molecule to

hydrolyse the glycosyl enzyme intermediate, thereby releasing the glycon moiety and the free enzyme. Although GH1 members share this conserved catalytic mechanism, some differ in the recognition of substrates which leads to a diversity of functions. Therefore, the enzyme structural features causing the differences in substrate recognition need to be elaborated.

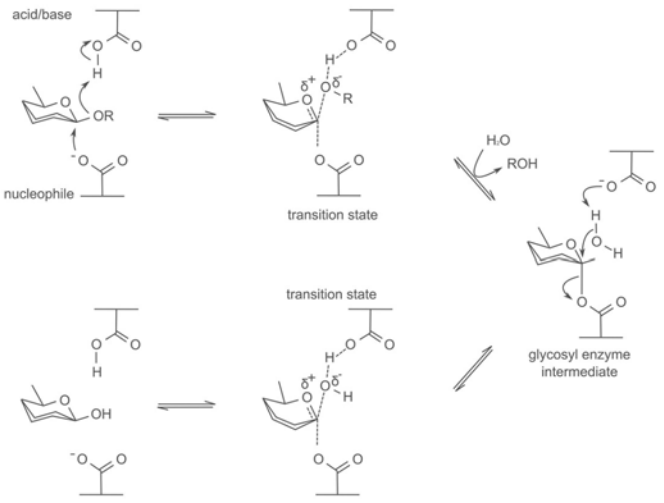


Figure 1.4. *Classical Koshland double-displacement retaining mechanism of catalysis for a GH1 enzyme.*

1.9.3 GH1 active site and key residues

The active site of GH1 enzymes is situated at the C-terminal side of the β-barrel and it is surrounded by loops that connect the α-helices to the β-strands. The active site consists of several subsites big enough to bind one monosaccharide unit. The subsite that binds the monosaccharide of the substrate nonreducing end is called subsite -1 (or glycon subsite), and the rest of the substrate (aglycon) binds to one or more subsites (+1, +2, +3 and so on) depending on the length of the oligosaccharide (Figure 1.5). The point at which the substrate is cleaved is between the glycon and aglycon binding regions [170]. Because of the mostly weak sugar-protein interactions, multiple binding sites frequently combine to enhance the signal. Cooperative binding makes the binding affinity of sites difficult to measure, as the properties of one subsite are influenced by the binding of the other

subsites [112]. It is possible that the orientation of the aglycon in the active cleft is important for glycon binding/positioning [177].

The predicted catalytic domain of a GH1 enzyme is a good indicator of its glycon specificities (e.g., glucosidase, galactosidase, rhamnosidase, etc.), however, glycon specificity is not absolute [178]. On the other hand, most GH1 enzymes have little specificity for the aglycon or for the bond configuration – the molecular details of aglycon specificity and affinity are very broad, highly complex, and are therefore far less understood [164]. Based on sequence or structure it is currently not possible to reliably predict the aglycon-specificity of a specific GH class, which is probably due to the functional diversity of GHs and limited experimental data [106]. Unfortunately, substrates needed for a broad biochemical characterisation are frequently commercially unavailable (e.g., phospho-glycosides) or too expensive for detailed kinetic analyses.

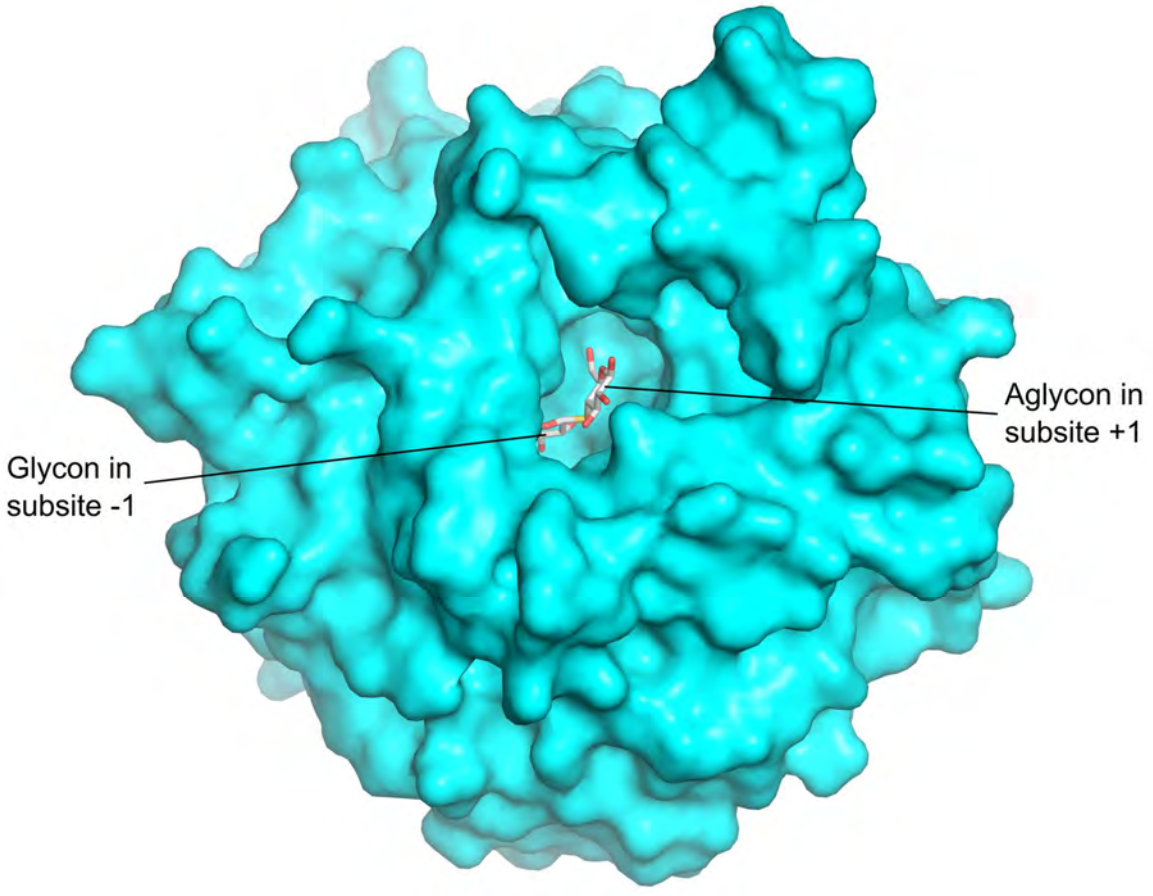


Figure 1.5. *GH1 binding site. PDB structure 2O9R (β-glucosidase) showing substrate with glycon and aglycon in subsites -1 and +1, respectively.*

A diversity of amino acid residues exists in different GH1 members, yet their conserved active site structures guarantee that their analogous residues will have most of the same interactions [178]. The catalytic acid/base (Glu170) and nucleophile (Glu378) residues are both glutamic acids and they are located at the centre of the TIM-barrel, at the C-terminal ends of the β4 and β7 strands, respectively. Data from crystallographic structures have revealed that within each subsite, the side chain of an apolar amino acid residue (usually the indole ring of a tryptophan residue) forms a platform which is a support base for the ligand stabilised by hydrophobic stacking interactions [177,181,182].

1.9.3.1 Substrate glycon interactions

GH1 enzymes bind to the glycon group of substrates using a network of highly conserved residues at the −1 subsite: Gln22, His122, Asn166, Glu356, Trp402, Glu409, Trp410 (PDB ID 2O9T numbering – a β-glucosidase; Figure 1.6) [175,177,183–191]. These residues form hydrogen bonds with ligands, except for Trp402 which acts a basal platform using hydrophobic interactions. In addition to ligand binding, Glu356 acts as the catalytic nucleophile residue. The β-glucosidase Glu409 residue (replaced by serine in 6Pβ-glycosidases) is an invariant phosphomimetic residue [175] – it differentiates between phosphorylated and nonphosphorylated substrates by clashing with the ligand-phosphate in terms of spatial position and like-charge repulsion (both the glutamate and phosphate are negatively-charged) (Figure 1.7).

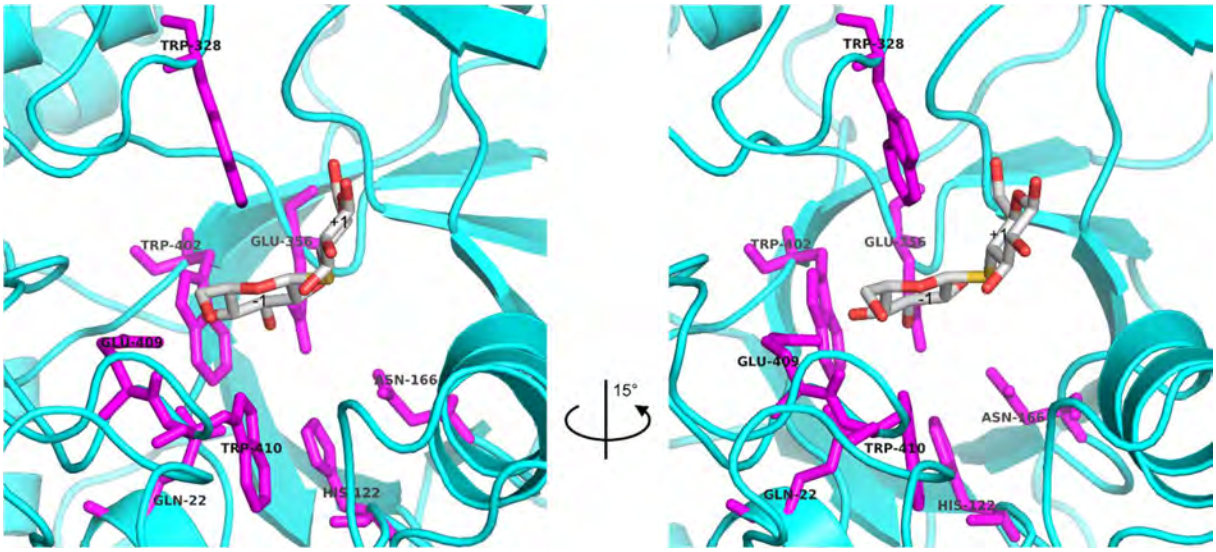


Figure 1.6. *Conserved binding residues in GH1 enzymes. PDB structure 2O9R (β-glucosidase) is used here. The Glu409 residue is replaced by serine in 6Pβ-glycosidases. All residues, except for Trp328, bind to the glycon (subsite -1).*

Despite very high conservation of residues interacting with the glycon, several phenomena are still not understood. For instance, a *Spodoptera frugiperda* β-glycosidase hydrolyses fucosides 40 times more rapidly than galactosides, even though the two ligands differ in only their hydroxyl 6, which is missing in fucosides [192]. In another case, mostly identical ligands which differ in only their hydroxyl 4 position displayed vastly contrasting results when subjected to *in silico* docking and molecular dynamics simulations whereby one ligand exited the enzyme while the other showed strong binding affinity. Additionally, it is still not understood how a GH1 enzyme can primarily be a β-glucosidase or a β-mannosidase [164]. GH1 specificity is a fascinating issue. The differential substrate preference, in combination with structural and mechanistic data, makes GH1 enzymes a suitable model to study enzymatic specificity which is a central property of biological systems [177].

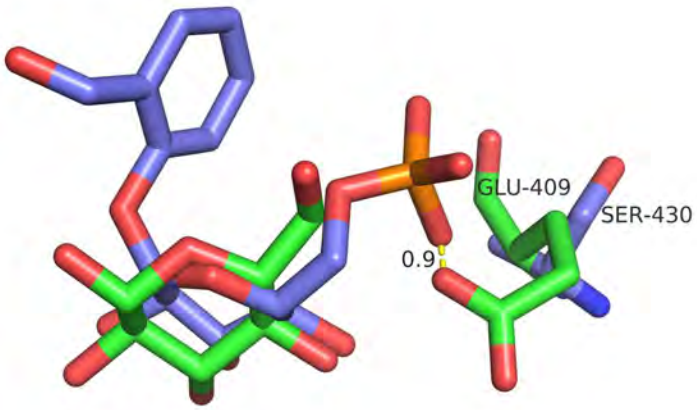


Figure 1.7. *β-glucosidase invariant phosphomimetic residue. Superposition of PDB structures 2O9T (green; β-glucosidase) and 4F77 (purple; 6Pβ-glucosidase). The β-glucosidase Glu409 residue (replaced by serine in 6Pβ-glycosidases) is an invariant phosphomimetic residue which differentiates between phosphorylated and nonphosphorylated substrates by clashing with the phosphate in terms of position and like-charge repulsion (both the glutamate and phosphate are negatively-charged).*

1.9.3.2 Substrate aglycon interactions

In contrast to the glycon, most GH1 enzymes have different sets of active site residues interacting with the aglycon [164,193] which are mostly hydrophobic interactions, and the way that these interactions control specificity is still unknown [177]. Therefore, researching aglycon specificity is crucial to understanding the function and mechanism of individual GH1 enzymes, and it is vital to keep hydrophobic interactions in mind when analysing the substrate specificity [106,194–196].

Most of the non-covalent interactions with the aglycon are found in the aglycon enzyme-subsite (subsite +1), which displays a variable spatial structure and amino acid composition. This structural variability is probably due to the high diversity of aglycons that are recognized by GH1 enzymes [177]. Structural data concerning the GH1 aglycon binding site are limited [175,181,187,190,197–201]. The only constant aglycon interaction is with a conserved tryptophan residue (Trp328 in PDB ID 2O9T numbering), which acts as a main hydrophobic platform that forms stacking interactions with the +1 sugar ring. The only exception is LpPgb1 from PDB ID 3QOM [175] where hydrogen bonds are

formed with the tryptophan residue instead of hydrophobic interactions. The rest of the residues that makeup the aglycon-binding pocket are not conserved.

1.9.3.3 Substrate phosphate interactions

In GH1 enzymes that are active upon 6Pβ-glycoside ligands (6Pβ-glycosidases), the substrate phosphate group binds to three conserved residues that are situated in the L8a loop (Figure 1.8). These residues are Ser430, Lys438 and Tyr440 (PDB ID 4F79 numbering – a 6Pβ-glucosidase). The serine residue replaces a glutamate that would bind to the glycon in non-phospho-glycosides, as mentioned in section 1.9.3.1. Only a handful of 6Pβ-glycoside PDB structures exist [105,116,175,183,199,201–203]. However, the phosphate interactions with the three binding residues are conserved. The serine and tyrosine make hydrogen bonds with the phosphate group, whereas the positively-charged lysine forms an attractive charge interaction with the negatively-charged phosphate. The charged lysine residue is thought to attract the phosphate to the L8a loop, and once there it will bind tightly due to the serine, lysine and tyrosine residues forming a three-point "anchor" [199].

In the crystallographic structures of a dual 6Pβ-glucosidase/6Pβ-galactosidase (dual-phospho) enzyme called Gan1D [199], the active sites of Gan1D complexed with several different ligands were compared. The glycon-binding residues do not move upon different-ligand binding; however, the phosphate-binding residues change positions depending on the specific binding modes seen for the different ligands. This is evidence that the phosphate-binding site allows for some binding flexibility in the Gan1D active site. Gan1D is a dual-phospho GH1 enzyme though, and this observation may not be true for the other GH1 activities.

1.9.3.4 6Pβ-glycosidases

DiscoveryStudio was used to determine the conserved interactions among 6Pβ-glycosidase enzymes. Only PDB structures of 6Pβ-glycosidases with ligands that have all three groups (phosphate, glycon, and aglycon) were compared. An exception is PDB 4PBG which has no aglycon – this is the only existing 6Pβ-galactosidase structure with a ligand. The other PDB structures used were 4F79, 4GPN, 4IPN (6Pβ-glucosidase), and 5OKE (dual-phospho). In Table 1.1 it is seen that the same 13 residue sequence positions interact with the ligand in all of the five PDB structures. These residues are also shown in the active site of the 4F79 structure in Figure 1.8.

Table 1.1. *Protein-ligand interactions of all current GH1 6Pβ-glycosidase enzyme PDB structures binding to ligands having all three groups (phosphate, glycon, and aglycon). Protein-ligand interactions were obtained using DiscoveryStudio. Residues in each column have the same sequence position. The PDB 4PBG ligand has no aglycon, it has been added because it is the only 6Pβ-galactosidase structure with a ligand. Blue residues indicate analogous interactions seen in all five structures. Hydrogen bonds with ligand atoms are indicated in red colour, in the subsequent column to the interacting residue. Green residues indicate analogous interactions seen in three of the five structures and could also have a role in enzyme function.*

4F79 (AAN59243.1)	Q18	O4	H130	O2/O3	F131	N175	O2	E176	O2	N179							A239	C241			Y313				
4GPN (AAN59243.1)	Q18	O3/O4	H130	O3	F131	N175	O2	E176	O2												Y313	O5			
4IPN (AAK74732.1)	Q18	O3	H125	O3	Y126	N170	O2	E171		S174	E177	L178			M227	L229					Y303				
4PBG - no aglycon (AAA25183.1)	Q19	O3	H116	O3	F117	N159	O2	E160					K173						N297	Y299					
5OKE (AHL67640.1)	Q23	O3/O4	H124	O3	W125	N169		E170G*	O2	I173		F177				A226			N299	Y301					

4F79 (AAN59243.1)	M314	W349	E375G*		W423			S430		A431	P	G432	P	T433	K438	Y440	P								
4GPN (AAN59243.1)	M314	W349	E375G*	O2	W423			S430	P	A431	P	G432	P		K438	Y440	O6/P								
4IPN (AAK74732.1)		W338	E364	O2	W415			S422		M423	P	S424	P		K430	Y432									
4PBG - no aglycon (AAA25183.1)		W347			W421			S428	P	W429	O3			N431	K435	Y437	O6/P								
5OKE (AHL67640.1)		W352	E378	O2	W425	O4	L430	S432[a]		W433	O3			N435	K439	Y441	P								

[a]No interaction due to structure missing residue
*Mutation

The 13 active site residues are highly conserved among GH1 6Pβ-glycosidases; the exceptions are Phe131 and Ala431 (PDB 4F79 numbering). In GH1 enzymes, the Phe131

residue is sometimes replaced with tyrosine or tryptophan, meaning the residue in this position is always an aromatic residue with a hydrophobic side chain. This residue, together with Tyr313 and Trp423, surround the ligand forming hydrophobic interactions.

The other non-conserved residue, Ala431, is sometimes replaced with methionine or phenylalanine in 6Pβ-glucosidases but is always a tryptophan in 6Pβ-galactosidases and dual-phosphate enzymes. In the 6Pβ-glucosidase activity the alanine residue forms a hydrogen bond with the ligand-phosphate group, whereas in the dual-phospho and 6Pβ-galactosidase enzymes the analogous tryptophan forms a hydrogen bond with the glycon-O3 atom (Table 1.1). More information about this residue position is discussed in subsequent sections.

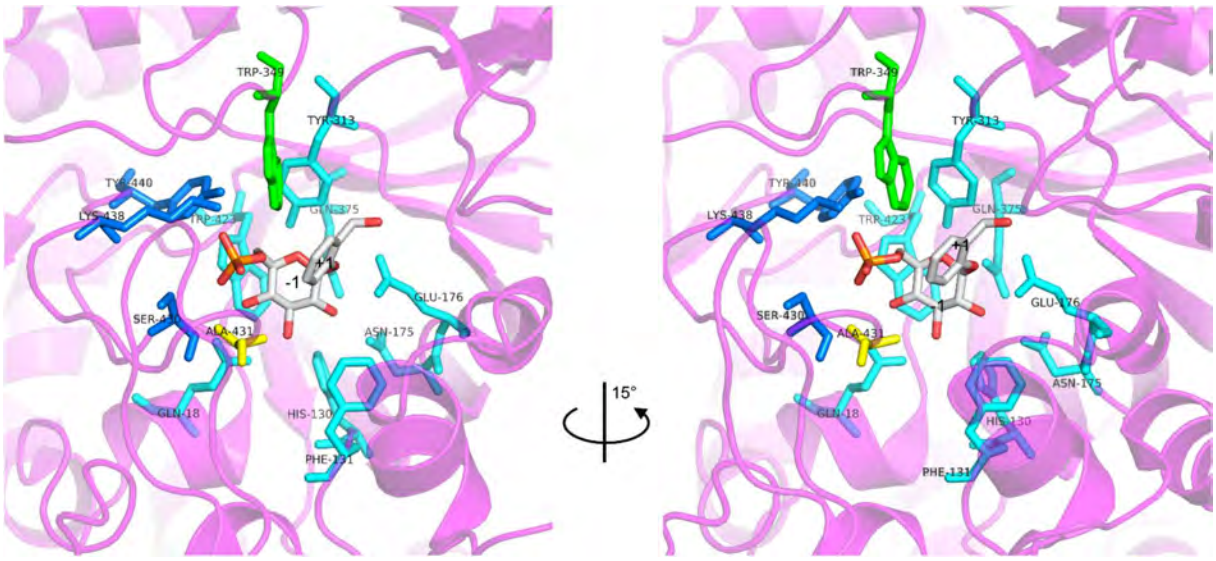


Figure 1.8. *6Pβ-glycosidase active site (PDB 4F79), showing residue sequence positions that interact with the ligand in all five 6Pβ-glycosidase PDB structures with a complete ligand. PDB 4PBG is an exception whose ligand has no aglycon. Cyan residues are glycon-binding. Green residue is aglycon-binding. Blue residues are phosphate-binding. Yellow residue position is glycon-binding in 4PBG (6Pβ-galactosidase) and 5OKE (dual-phospho) but is phosphate-binding in the remaining three structures (6Pβ-glucosidase).*

Three of the five structures show ligand interactions with the Asn179, Ala239, Gly432, and Thr433 residues (PDB 4F79 numbering). The residues in these positions could also have a role in GH1 enzyme function. All three 6Pβ-glucosidase structures form hydrogen bonds between the Gly432 residue and the ligand-phosphate, but this interaction is absent in the 6Pβ-galactosidase and dual-phosphate structures; the Gly432 residue position interaction could be important for the specificity of the 6Pβ-glucosidase activity, helping to position bound substrate slightly differently in the active site compared to the other activities. The 6Pβ-glycosidase active site residues are also shown in a multiple sequence alignment in Figure 1.9.

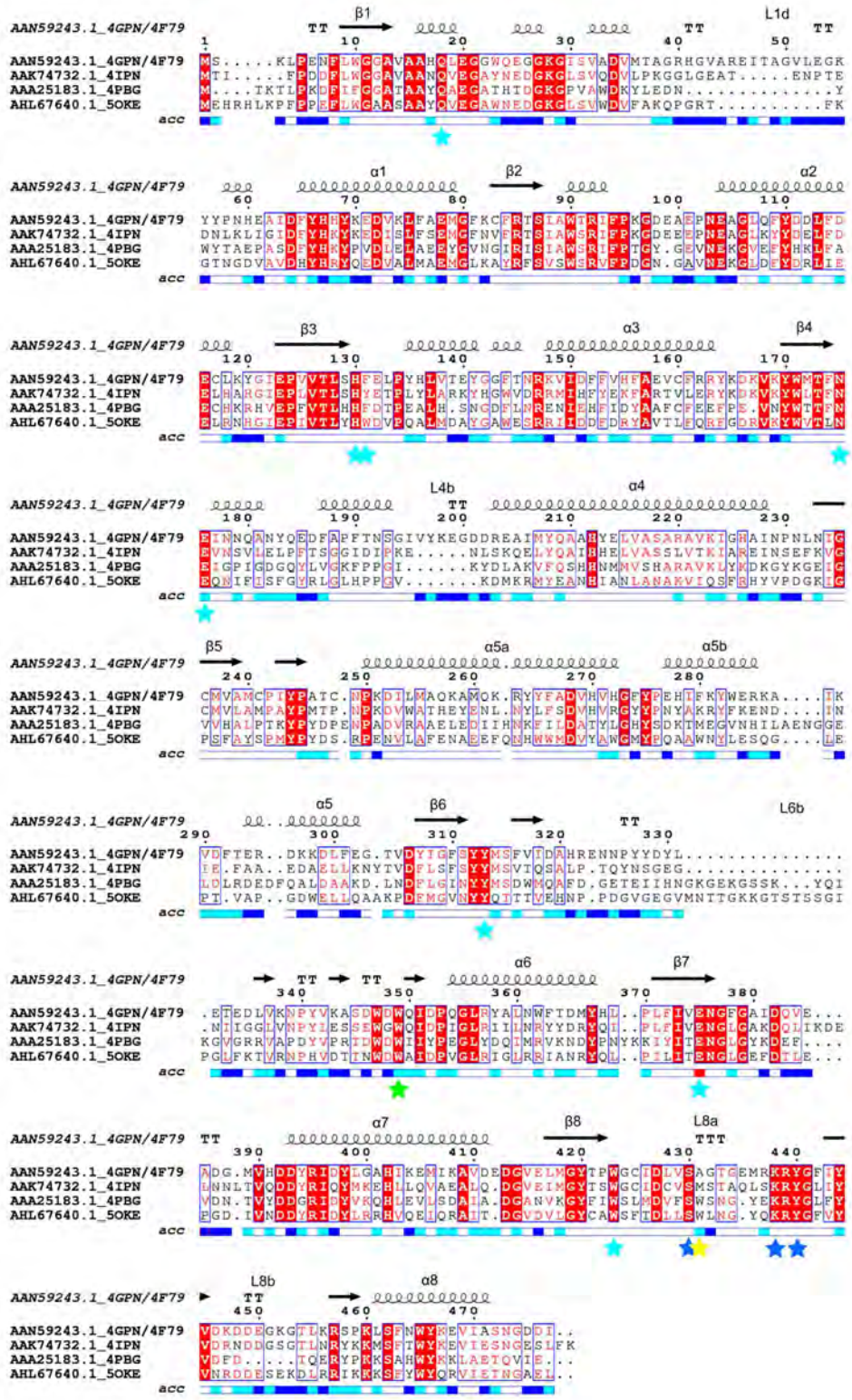


Figure 1.9. *Multiple sequence alignment using 6Pβ-glycosidase enzymes. Identical residues are highlighted in red, similar residues are shown as red letters. Glycon-binding residues are shown with cyan stars. Aglycon-binding residue is shown with green star. Phosphate-binding residues are shown with blue stars. Yellow star shows residue position that is glycon-binding in 4PBG (6Pβ-galactosidase) and 5OKE (dual-phospho), but is phosphate-binding in the remaining three structures (6Pβ-glucosidase). Secondary structure elements derived from 4F79 are depicted. The figure was produced with ESPript 3.0.*

1.9.3.4.1 Glucose- vs galactose-configured ligands: 6Pβ-glucosidases vs 6Pβ-galactosidases

There is evidence that a specific residue position in loop L8a (Ala431 - 4F79 numbering) is responsible for, or at least has a role to play in, substrate specificity between the 6Pβ-galactosidases and 6Pβ-glucosidases [178,183,199,201]. This residue is a conserved tryptophan in 6Pβ-galactosidases and an alanine, phenylalanine, or methionine in 6Pβ-glucosidases. The one and only difference between galacto- and gluco-configured substrates is their O4 hydroxyl group arrangement; Michalska and coauthors [175] have explained that the closer O4 hydroxyl group of the galacto epimer would clash with the 6Pβ-glucosidase residue in loop L8a (Ala431 - 4F79 numbering) and that this would prevent binding and catalysis (Figure 1.10).

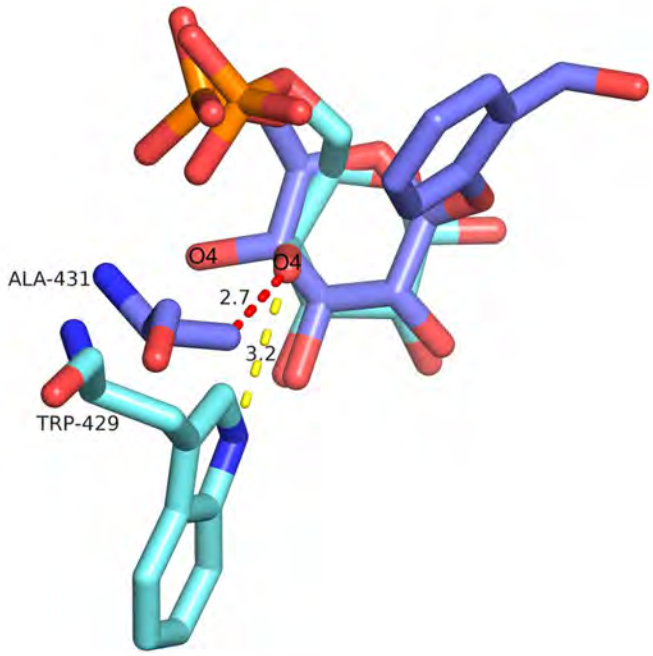


Figure 1.10. *6Pβ-glucosidase specificity-inducing Ala431 residue position (4F79 numbering). Superposition of PDB structures 4F79 (purple; 6Pβ-glucosidase) and 4PBG (cyan; 6Pβ-galactosidase). 6Pβ-glucosidase residue would have a steric clash (red dashes) with the axial OH4 of a galactoside ligand. This would prevent galactosides from binding to 6Pβ-glucosidases.*

Residue mutation followed by activity measurements provides additional evidence for the specificity-determination of this L8a loop residue position. In a 6Pβ-galactosidase enzyme, the tryptophan was mutated to alanine which shifted the substrate preference towards gluco-configured substrates [204]. On the other hand, in a 6Pβ-glucosidase, mutating its homologous Met423 to tryptophan changed its specificity to galactoside substrates [201]. These results provide evidence for the significance of this L8a loop residue position in the specificity distinction between the 6Pβ-galactosidase and 6Pβ-glucosidase activities.

1.9.3.4.2 Glucose- vs galactose-configured ligands: Dual 6Pβ-glucosidase/6Pβ-galactosidase enzymes

The Gan1D enzyme [199] is the only dual-phospho GH1 enzyme with previously published crystallographic structures. Gan1D shows promiscuous activity and has approximately equal specificity for substrates that contain either galactose6P or glucose6P as their glycon moiety. The researchers responsible for the Gan1D structures state that Gln23 and Trp433 in Gan1D are important for glycon binding and recognition and propose that the specific hydrogen bonding of these residues enables the change of preference toward a galactose or glucose sugar. They note that according to their Gan1D-galactose6P complex, the O4 hydroxyl prefers to interact with Trp433 whereas the Gan1D-glucose6P O4 hydroxyl prefers to interact with Gln23.

The Gan1D-Trp433 residue was mutated to either alanine or methionine, followed by kinetic measurements using both galactose and glucose substrates [205]. Although the catalytic activity was reduced for both types of substrates, both mutant enzymes were more active toward glucose substrates compared to galactose substrates. This suggests that galactose binding to Gan1D depends strongly on its specific hydrogen bond with Trp433, and that this tryptophan residue helps discriminate between glucose- vs galactose-configured ligands in dual-phospho activity enzymes.

1.9.3.4.3 Loops around active site

The GH1 structures mostly differ in the loops that surround the active site (Figure 1.11). Difficulty has been shown in determining the coordinates of these loops in several PDB structures, resulting in missing residues [183,199,201,202]. The difficulty modelling the loops could mean that they are not important for the structure and function of the enzymes. However, it could also mean they have a more general and dynamic role to play [199]. The loops form the entrance to the active site of the enzyme which indicates a functional role. The 6Pβ-galactosidase activity and dual-phospho enzymes possess an additional β-hairpin loop (L6 loop) that covers the front of the enzyme and it is thought to control access to the active site [183]. When this β-hairpin loop blocks the opening to the active site, the substrate and the glycon product cannot pass through and only the aglycon product can be released. 6Pβ-glucosidases do not possess the β-hairpin loop so the active site cavity is slightly larger [175]. Nevertheless, the 6Pβ-glucosidase L1d and L6c loops may potentially act as a small “gate”, also controlling access to the active site.

Considering the conservation of the overall structures and active sites of GH1 enzymes in general, the differences at the +1 subsite (aglycon) and the entrance to the active site are likely to be contributing determinants of substrate specificity.

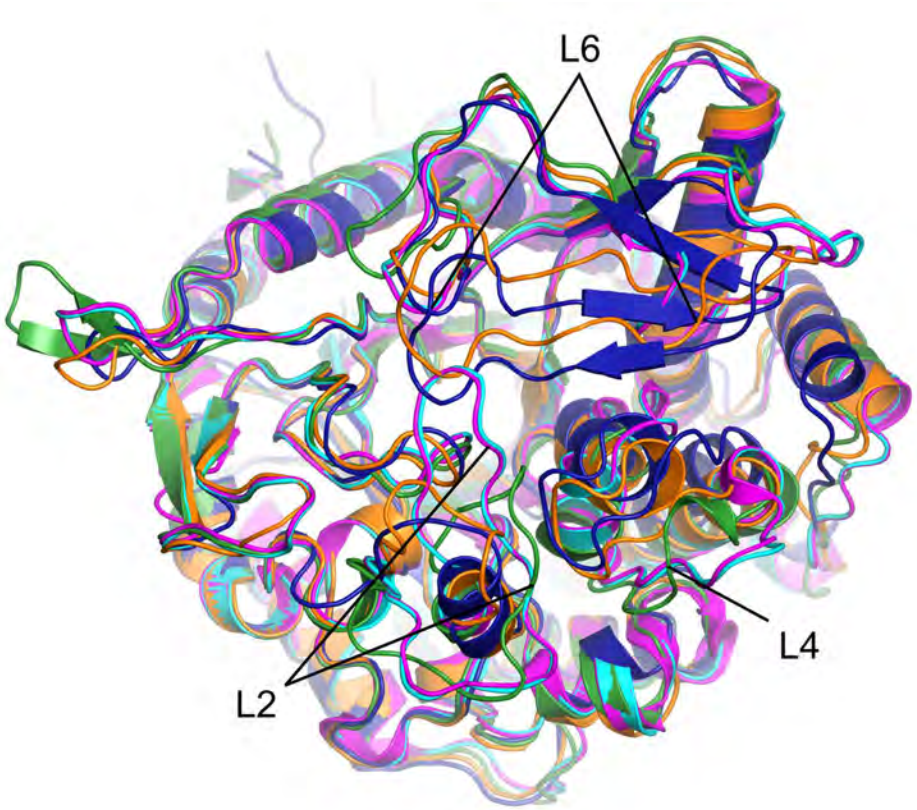


***Figure 1.11**. Structural differences in the loops that surround the active site of GH1 enzymes. Superposition of PDB structures 4F79 (purple), 4GPN (cyan), 4IPN (green), 4PBG (dark blue), and 5OKS (orange). Dual-phospho 5OKS was used instead of 5OKE in this case because it has no missing residues in important regions.*

- ***PART 3***

1.10 Sequence analysis

Use of sequence comparisons to deduce protein structure and function has expanded substantially in recent years as the genomes and messenger RNAs of more and more organisms have been sequenced, permitting a vast array of protein sequences to be deduced [1]. Proteins that have a common ancestor are referred to as homologs. The main evidence for homology among proteins, and hence for their common ancestry, is similarity in their sequences, which is often reflected in similar structures (particularly when confirmed by phylogenetic analyses). We can describe homologous proteins as belonging

to a “family” and can trace their lineage (how closely or distantly they are related to one another in an evolutionary sense) from comparisons of their sequences. The folded 3D structures of homologous proteins may be similar even if some parts of their primary structure show little evidence of sequence homology. It is generally thought that proteins with about 30 percent sequence identity are likely to have similar 3D structures; although, such high sequence identity is not required for proteins to share similar structures. Revised definitions of family and superfamily have been proposed, in which a family comprises proteins with a clear evolutionary relationship (>30 percent identity or additional structural and functional information showing common descent but <30 percent identity).

1.10.1 Multiple sequence alignment

In bioinformatics, sequence alignment is commonly the initial step in revealing the molecular phylogeny of an unknown biological sequence [206]. During the production of multiple sequence alignments (MSAs), a series of algorithms are utilised to align evolutionarily related sequences while considering evolutionary occurrences like mutations, insertions, deletions, and rearrangements. This is applicable to DNA, RNA, or protein sequences. The function of an MSA is to align the sequences so that their evolutionary, functional, or structural relationship can be seen more clearly. This is accomplished by aligning homologous positions with each other by inserting gaps within the sequences. The inserted gaps represent insertions and deletions (indels) within the sequences that are caused by evolution from a common ancestor [207]. MSAs have become integral in areas of molecular biology and bioinformatics such as phylogenetic tree reconstruction, 3D structure prediction, conserved regions identification, and molecular function [208].

Protein sequence alignment holds more significance compared to nucleotide sequence alignment because proteins are vital functional biological molecules and are the source of structural and/or functional information [209]. Therefore, the alignment is intertwined with

structural biology [210]. Usually, protein alignment is utilised to locate conserved sequence regions that have functional importance by comparing proteins that have similar characteristics [211]. This works well when comparing closely related sequences, but the alignment of very distantly related sequences can be unreliable and challenging to interpret. Advanced sequence alignment methods misalign 11-19% of sequences [212], and is worse when comparing many divergent sequences [213]. Unfortunately, current MSA tools do not produce biologically perfectly accurate MSAs because the task is 1) biologically complex (complex relationships often exist between related sequences), 2) computationally intensive, and 3) difficult due to occasional lack of evolutionary history [214]. Consequently, there is a lot of ongoing research in this area [215,216]. Well over 100 different MSA methods have been developed [217]. The use of a particular method is dependent on the type and length of the sequences and on user preferences [218].

Pairwise sequence alignment (PSA) and multiple sequence alignment (MSA) exists. PSA only utilises two sequences at once. MSA uses multiple sequences (more than two) and therefore produces added biological information compared to PSA. MSA is also required to compare genomic analyses for identification and quantification of conserved regions or functional motifs in a whole sequence family, assessment of evolutionary divergence, and for ancestral sequence profiling [219]. In MSA generation, differing scoring methods are used to determine the level of identity or similarity between any two sequences. Nucleotide scoring is a straightforward identification process whereby identical bases in both sequences are assigned positive scores. On the other hand, protein sequence similarity scores are included in addition to identity scores – sequence similarity indicates amino acids that have similar physicochemical properties. The substitution matrices that are used the most for protein sequence alignment are Point Accepted Mutation (PAM) [220] and BLOcked SUbstitution Matrix (BLOSUM) [221], and two methods are utilised: global alignment [222] and local alignment [223]. Global alignment uses the whole length of the

sequences to calculate sequence similarity whereas local alignment uses local shorter strands within the whole sequences. Many global alignment methods exist [224,225] but problems occur between sequences that are only homologous over local regions, when the lengths of the sequences are largely different, or when there are shuffled domains between the sequences [226]. This is when local alignment is used [214,227–229].

Progressive alignment is the most popular heuristic used for MSA generation [215,230]. It first performs pairwise alignments with methods like the Needleman-Wunsch algorithm, Smith-Waterman algorithm, k-tuple algorithm [231], or k-mer algorithm [232]. Next, the relationship between the sequences is revealed by clustering them with methods like mBed and k-means [233]. Distance scores are drawn from similarity scores and used to construct guide trees using Neighbour-Joining (NJ) [206] and Unweighted Pair Group Method with Arithmetic Mean UPGMA guide tree building methods [234]. Based on the guide tree, sequences are added to the alignment one at a time starting with the most similar sequences. A disadvantage of the progressive alignment heuristic is that it only looks at two sequences at a time, meaning that any errors that occur at the start will progressively become worse throughout the alignment process which cannot be fixed later. The progressive alignment heuristic is the basis of many alignment algorithms, including ClustalW [235], Clustal Omega [233], MAFFT [236], MULTAL [237], PCMA [238], MULTALIGN [239], Kalign [240], Probalign [241], MUSCLE [232], T-Coffee [225], and PROMALS [242].

The disadvantage of progressive alignment, mentioned above, can be solved using iterative alignment which continuously updates the guide tree. Dynamic programming repeatedly realigns the initial sequences while adding new sequences which improves their overall alignment quality [243]. Some iterative methods repeatedly divide the aligned sequences into two groups and realign the groups until the alignment process has converged [244,245]. Iterative alignments are limited to a few hundred sequences but are

5–10% more accurate [233]. Popular iterative alignment algorithms include PRRP [246], MUSCLE [232], Dialign [227], SAGA [247], and T-COFFEE [225]. MSA accuracy can be improved even more by taking protein structural information into account – this is because protein structure is more conserved than sequence [248]. Popular tools that incorporate structure include 3D-COFFEE [249], EXPRESSO [250], PRRP [244], MICAlign [251] and PROMALS3D [252].

PROfile Multiple Alignment with predicted Local Structures and 3D constraints (PROMALS3D) is a protein MSA tool strengthened by adding evolutionary and structural information from databases. Progressive alignment is initially used whereby similar sequences are aligned using a scoring function of weighted sum-of-pairs of BLOSUM62 [221] scores, resulting in differing pre-aligned sequence groups. A sequence representative (called 'target sequence') from each pre-aligned group is then chosen and subjected to PSI-BLAST searches to obtain extra homologs from the UNIREF90 [253] database and to PSIPRED [254] for the prediction of secondary structure. A hidden Markov model (HMM) of profile–profile alignment that includes predicted secondary structures is implemented using pairs of the target sequences to obtain posterior probabilities of residue matches. A probabilistic consistency scoring function is drawn from sequence-based constraints that are derived from the probabilities. The target sequences are then progressively aligned using the consistency scoring function, and the pre-aligned groups are merged with the target sequence alignment [252]. In short, PROMALS3D uses input proteins to automatically find homologs from sequence and structure databases. It then derives structure-based constraints from alignments of 3D structures and merges them with sequence-based constraints of profile–profile alignments in a consistency-based framework to generate high-quality multiple sequence alignments.

1.10.2 Phylogenetic analysis

Phylogenetics has been used for more than 50 years and is now involved in almost every biological discipline [255]. It was developed to organize objects around a collection of cladistic rules but now has become the basis of evolutionary biology and a useful tool for research in many different fields [256], which include genomics [257], community ecology [258], epidemiology [259], conservation biology [260], and population dynamics [261]. In phylogenetic trees, the evolutionary history of groups of species are depicted as tree structures. The tips of the trees represent extant species, and the internal nodes represent speciation events. The tree model is rather simplistic, but it encapsulates the intricacy of the underlying phenomena [262].

The bioinformatics tool Molecular Evolutionary Genetics Analysis (MEGA) analyzes molecular sequences to construct phylogenies after determining evolutionary distances. From sequence alignments, MEGA draws useful information using statistical techniques to calculate specific physiognomies of nucleotide or proteins and predicts evolutionary relationships [263,264]. First, distance-based methods like UPGMA and Neighbor-Joining group all the taxa in a single node and separates with every repetition so that pairs of nodes are chosen that are grouped at iterations to minimize the overall branch length. Trees with the greatest likelihood are found using character-based methods like parsimony and probalistic methods (maximum likelihood and bayesian inference) by evaluating the possibility that a specific evolutionary model (e.g., BLOSSUM or PAM matrices) has produced the observed data [265,266]. MEGA performs bootstrapping, a statistical re-sampling procedure, to test tree reliability by calculating the probability of branch recovery if the taxa were re-sampled.

1.10.3 Sequence motif discovery

Random mutagenesis occurs during evolution which alters the biological molecular sequences that encode proteins. Amino acids can be substituted, inserted, or deleted, which could adversely affect protein function if the mutation is located in an important region of the sequence. As a result, functionally important amino acids are conserved over time. However, sometimes mutations can result in a new protein function, bringing about new traits and even species [267].

Sequence motifs are highly conserved regions within the sequences of a cluster of related proteins and they most likely have the same function in a protein [268]. Therefore, motifs can help to understand gene function which could be useful for example in medicine, forensic DNA analysis, and agricultural biotechnology. Motifs may represent important segments in proteins like enzyme active sites, ligand-binding sites, cleavage sites, post-translational modification sites, transcription binding sites, splice junctions, or interaction interfaces. They exist in an exact or approximate form within a family or a subfamily of sequences [267].

Firstly, there are short functional motifs containing specific residues that have mostly evolved independently from the surrounding structural context (e.g., myristilation sites, glycosylation sites, Src homology [SH]2-binding sites). Secondly, there are short structural motifs that correspond to sequence-level topological constraints (e.g., N and C caps of α-helices). Thirdly, there are functional motifs that are not based on invariant residues and are more constrained at the sequence level (e.g., transmembrane regions, signal sequences). Finally, some motifs represent unique, detectable sequence features that can differentiate a cluster of related proteins from all other proteins. These motifs indicate functional and structural constraints within the protein cluster, and therefore common evolutionary descent (homology) [269].

Motif representation models try to provide generalizations about known functional motifs and are used to characterise functional sites which helps to identify them in unknown protein sequences. The motif representation models are split into deterministic models and probabilistic models. Deterministic models use consensus sequences which are generated simply by choosing the amino acid found most frequently at each sequence position. Since consensus sequences do not account for amino acid variability at each position, regular expressions are sometimes used. Probabilistic models go a step further and calculate the residue identity frequency of every sequence position. A popular probabilistic model is the position-specific scoring matrix (PSSM) [270], also known as the probability weight matrix (PWM), which uses a matrix where each entry (i,a) is the probability of finding an amino acid a at the ith position in the sequence motif. Another popular probabilistic model is the hidden Markov model (HMM) [271].

Already-characterised motifs can be found in some databases such as PRINTS [272], PROSITE [273], or ELM [274]. During motif detection, the deterministic and probabilistic models are used to identify known motifs in other sequences, which is called the pattern recognition problem. A known motif that is found in an unknown sequence provides evidence of protein function and family classification. Frequently, deterministic models are either too specific resulting in many false negative predictions, or too degenerate resulting in many false positives. In probability matrices or HMM-based methods, a log-odds score is used to measure the probability that a sequence is generated by a model rather than by a random null model [267].

Motif discovery is a much more complicated task [275]. There are an enormous number of potential combinations of the different amino acids, and motifs are often degenerate which complicates things even more. Generally, motif discovery algorithms use a cluster of related sequences to look for patterns that are unlikely to occur by chance. Motif discovery uses alignment-based methods and unaligned sequence methods. There is no upper limit

to motif length when using alignment-based methods, and no maximum threshold value for motif distance from the sequences. However, this approach is only accurate if the sequences are similar enough and the patterns are present in the identical order in all of the sequences. Consequently, motif discovery mainly uses alignment-free methods that generally find patterns that are overrepresented in a cluster of sequences. This is complicated because of factors such as the precise start and end boundaries of the motif, the size variability (presence of gaps or not), or stronger or weaker motif conservation during evolution [267]. De novo motif discovery programs are usually based on enumeration, probabilistic optimization, or deterministic optimization algorithms.

Enumeration counts all substrings of a specific length (known as words or k-mers) and locates overrepresentations. It determines all instances of motifs, but the exponential complexity makes this computationally exhaustive and only suitable for short motifs [276]. Enumeration requires user input like motif length, the number of mismatches allowed, and the minimum number of sequences in which the motif must be present [277]. Probabilistic optimization is an iterative method where a random subsequence is extracted from each sequence to build an initial model. During each iteration, the *i*th sequence is removed, the model is recalculated, and a new motif is extracted from the *i*th sequence. This is repeated until convergence [267]. Deterministic optimization is based on the expectation-maximization (EM) algorithm which involves two main steps. The first is the “Expectation step” that reconstructs the hidden motif structure using a set of parameters. The second is the “Maximization step” which utilises the motif structure to refine the parameters over numerous iterations. In this way, alternate sequences representing the motif are found and the motif model updated.

MEME (Multiple EM for Motif Elicitation) is a popular program that is based on deterministic optimization, and it optimizes PWMs using the EM algorithm [229]. MEME locates an initial motif and uses the expectation and maximization steps to improve the

motif until the PWM values cease to improve or the iteration maximum is reached. MEME only needs a set of unaligned sequences and a motif-width parameter as input. It usually finds the most statistically significant (low E-value) motifs first and returns a model of each motif and a threshold which together can be used as a Bayes-optimal classifier for searching for occurrences of the motif in other databases. MEME also outputs an alignment of the occurrences of the motifs. To identify and omit any overlapping motifs, the Motif Alignment Search Tool (MAST) can be used which is part of the MEME suite [278].

As protein structure is more conserved than protein sequence [279], the discovered motifs can be mapped to 3D structures for further analysis. Important details regarding protein function can be revealed when performing structural analyses on a protein or when comparing many proteins.

1.11 Protein 3D structural analysis

The analysis of protein structure or the comparison of protein structures is important for understanding structural, functional and evolutionary relationships of proteins, particularly in case of novel proteins [280]. Protein structure visualization is crucial to utilising, studying and understanding macromolecular structure data [281]. It is often informative to visualize two related structures superimposed, whereby the Cα atoms of two or more proteins are aligned as much as possible [282]. The behaviour of a protein can be understood or predicted using comparisons with members of its protein family because they have similar structure. The structural differences or similarities between proteins having an evolutionary relationship can expose many insights [283]. Several molecular graphics tools can be used for visualization, and some can perform automatic superposition such as PyMOL or VMD. And obtaining experimentally determined protein 3D structures for analyses is easy as they are all combined in one data repository, the Worldwide Protein Data Bank (wwPDB).

Structural comparisons can be characterised in certain ways, depending on the objective:

Level of detail – The different positions of several types of structure may be compared such as individual atoms, residues and their side chains, position and orientation of secondary structural elements, and similarity of folds at the tertiary structure level. We may analyse the residues and geometry of binding sites, and the interactions with ligands [283,284]. One could also identify structural motifs and functional sites [285].

Physicochemical properties – The function of an enzyme is influenced by the physicochemical attributes of atoms such as being a hydrogen bond donor, a hydrogen bond acceptor, having a positive charge or a negative charge, or being hydrophobic. Whole residues can also be hydrophobic or hydrophilic. The property could be affiliated with a certain point in space or covering a part of an enzyme surface [283]. Many tools can generate molecular surfaces that can map a range of properties like residue conservation scores, depth-cue information, mean-force potentials, in addition to hydrophobicity and electrostatics. These coloured surfaces, also called texture mappings, display regions with complementary shape and physicochemical attributes to provide details like molecular interactions and conformational changes [284].

Extent of comparison – The structure comparison can be global or more local. Global comparisons use the whole protein, whereas local comparisons are restricted to a protein domain, a contiguous subsequence of amino acids, or the subset of atoms in a binding pocket [283]. Structures that have very good global similarity might not have good local similarity. For instance, frequently inadequately predicted flexible or disordered protein sections like long loops and/or termini can undermine the similarity between structures. Relative domain movements of multi-domain proteins could also affect global similarity scores. Therefore, targeting specific areas using local similarity can prevent these problems [286].

Number of proteins – Structure comparisons could involve one, two, or even more proteins. When a single protein is used, conformational changes are usually compared or changes in residue side chain positions of the binding pocket caused by ligand docking. When two proteins are used, differences of sequence composition in relation to structure are compared. For instance, distant but homologous proteins can be compared to evaluate the level of structural similarity that has been conserved in an evolutionary descent. Or perhaps conformational changes could be assessed that are caused by the substitution of residues. Studies where several proteins are involved often use closely related proteins as established by sequence alignment and are usually used to determine similarities and/or differences in binding sites. On the other hand, structural similarity studies can be performed independent of sequence similarity. Very low sequence similarity but high structural similarity is evidence of evolutionary convergence [283].

Additionally, protein structure comparison algorithms can be used for searching similar structures in structural databases, prediction of unknown protein structures, hierarchical classification of proteins, evaluating sequence alignment methods, and evaluation of predicted 3D structure of a protein [285,287–291].

1.12 Homology modelling

The experimental determination of protein structure has relied on X-ray crystallography and NMR spectroscopy [292]. These methods provide high-resolution atomistic detail, but they are costly, time consuming, and still not applicable to all types of proteins, like intrinsically disordered proteins [293] and some membrane proteins [294,295]. Click or tap here to enter text.On the other hand, there has been an explosion of protein sequencing, meaning that many important proteins have no structures available, and the gap between sequences and solved structures is ever widening [296,297]. Fortunately, we can predict protein structure using either homology modelling (also called comparative modelling),

threading, or *ab initio* methods [117]. Homology modelling (not applicable to intrinsically disordered proteins) predicts protein structure using one or more template 3D structures of proteins that have some sequence and structural similarity to the protein to be modelled while *ab initio* methods predict structure by simulating protein folding based primarily on physical chemistry principles [298,299]. Significant progress has been made with ab initio methods. However, homology modelling is more efficient and currently generates the most reliable and accurate models of protein structure [117,300–302], but a good quality template structure is needed that has sufficient sequence identity and covers most of the target sequence.

Using its amino acid sequence, homology modelling determines the 3D structure of a protein (target protein) based on homologous or similar template structures. We can do this because all members of a protein family share the same fold with a core that is more conserved than sequence [303]. Although, high template sequence identity is very important for model accuracy and quality. Templates with over 50% sequence identity are considered reliable enough for most applications [304]. Templates with 30-50% similarity share a minimum of 80% structural similarity, so modelling errors will mostly occur in loop regions. Sequence identity of below 30% will lead to speculative models [305]. Also contributing to model quality is template structure resolution [306], phylogenetic similarity [307], and environmental factors such as pH, solvent type, and existence of bound ligand [296].

Before homology modelling, a sequence alignment must be generated using the target and template proteins. An accurate alignment is essential as an error of one residue causes a shift of an α-carbon, and a single residue gap in an α-helix causes a rotation of the other helix residues [308]. If the required sequence similarity is unattainable, more than one template structure is used [309].

Most homology modelling algorithms have common steps with minor differences. The steps are usually repeated until a satisfactory model is obtained [309]. The steps include backbone generation, loop modelling, side chain modelling, and model optimization/refinement.

1.12.1 Backbone generation

Backbone, or framework, building approaches are grouped into rigid body-assembly methods, segment matching methods, spatial restraint methods and artificial evolution methods. In rigid-body assembly the protein structure is divided into conserved core regions, loops and side chains [310]. The target backbone is constructed by averaging the backbone atom positions of the template structures, weighted by the target sequence similarity [311]. Rigid-body assembly is used by tools like 3D-JIGSAW [312] and SWISS-MODEL [313].

In segment matching, the positions of a subset of atoms (usually C-alpha atoms) from the template structure are used as a guide to locate and construct short all-atom segments that correspond to the reference atoms [296]. The whole atom model is then constructed using the leading structure as a pillar to lay the segments. This can be done by using SEGMOD/ENCAD [314].

The spatial restraint method builds the model by satisfying constraints/restraints based on the spacing between atoms, bond lengths, bond angles, dihedral angles, van der Waals contact distances etc., seen in the template protein structures [311]. The most popular spatial restraint program is MODELLER [315]. Additional restraints can be used that are not based on homology such as analyses of hydrophobicity [316], data from NMR experiments [317], site-directed mutagenesis and cross-linking experiments [318], and fluorescence spectroscopy [319].

Artificial evolution method uses rigid-body assembly with stepwise template evolutionary mutations until the template structure is as accurate as possible [296]. According to the alignment, iterative template structure alterations are performed together with energy minimization to offset the energy cost caused by mutation, deletion or insertion in the template sequence. The template structure alterations not having serious energy penalties are kept [320]. The NEST homology modelling program was the first to use artificial evolution [321].

1.12.2 Loop Modelling

Loop structure prediction is more difficult compared to strands and helices [322] as loop structure is not evolutionarily conserved, even in proteins without loop deletions or insertions. Loop modelling accuracy is very important as loops often contribute to active and binding sites and thus functional specificity [305]. The database search approach and the conformation search approach are used to model loops. The database search method scans all known protein structures to locate segments giving the endpoint regions of the protein backbone. MODELLER [315] and SWISS-MODEL [313] use database searches. The conformation search approach is based on ab initio fold prediction using scoring function optimization [302] with Monte Carlo or molecular dynamics techniques [323]. The scoring function accounts for conformational energy, steric hindrance, and favourable interactions like hydrogen bonds [311]. Usually, 4-7 residues are modelled at a time because the conformation variation increases as the loop length increases. In addition, methods exist that combine database and conformation searches [324,325].

Although there has been much progress in flexible loop modelling, it is still an open problem and is the most difficult aspect of protein modelling [326,327]. The major limitation for the creation of higher accuracy modelling methods is the shortage of experimental data. Due to low electron density, x-ray crystallography (the most widely used experimental method to obtain high-resolution protein structures) often cannot solve the

structure of flexible regions like loops. The latest methods can predict stable conformations of relatively short loops (less than 12 residues); however, accurately sampling, scoring, and representing the many numbers of loop conformations remains a challenge, especially for long loops. Recently, machine learning methods are looking promising in loop modelling and structural bioinformatics in general [328–331].

1.12.3 Side Chain Modelling

The prediction of residue side chain orientation is usually simple for conserved residues because they mostly have well-defined Cα-Cβ bond torsion angles and are copied straight from template to model [286]. Alternatively, libraries of common backbone-dependent rotamers taken from high-resolution X-ray structures are tested sequentially and scored using energy functions based on hydrogen bonds, disulphide bridges, and steric hindrance [311]. The hydrophobic core residues are usually predicted less accurately than the water-exposed residues [332]. Selecting one rotamer affects the choice of neighbouring ones, making the procedure computationally expensive. However, position-specific rotamer libraries can be used when some backbone conformations favour specific rotamers [333,334].

1.12.4 Model Optimization/Refinement

Distortions such as clashes between atoms, steric hindrance, stretching of bond lengths, unfavourable bond angles and dihedral angles etc., can exist in the model before this stage of homology modelling. Energetically unfavourable conformations occur when atoms clash together. Steric hindrance causes strong repulsive forces when the van der Waals spheres affiliated with the residues overlap. Optimization normally starts with iterative energy minimization using molecular mechanics force fields [335–337] where the model falls into a lower energy conformation and comes closer to the native structure [311]. Every minimization step fixes some big errors, although, simultaneously many other small

errors are introduced. To decrease the accumulation of errors, atom positions can be restrained, the number of minimization steps can be reduced to only a few hundred, and more precise force fields can be used such as quantum force fields [338] and self-parameterizing force fields. Additionally, Monte Carlo simulation and all-atom molecular dynamics simulations that imitate the natural folding process are sometimes used for model refinement [339–343]. A general rule is that more attention should be given to the first homology modelling steps, as model refinement usually has a disappointing return on investment [344].

1.12.5 Model evaluation

Model validation must be performed to determine its accuracy for further application. The protein models generated can have different accuracy depending on sequence similarity, environmental parameters, and template quality. Low-accuracy models are satisfactory for mutagenesis experiments whereas virtual screening in drug discovery, for instance, needs better accuracy [345], and very high accuracy is needed for mechanistic studies [297,346]. Model evaluation methods generally use a combination of stereochemical plausibility checks, knowledge-based statistical potentials, physics-based energy functions, or model consensus approaches [347–350]. Popular tools for stereochemistry evaluation are WHATCHECK [351], PROCHECK [352] and Molprobity [353]. The Ramachandran plot is a very good indicator of model quality: residues with inaccurate stereochemistry will fall out of the acceptable regions of a Ramachandran plot [354]. VERIFY3D [355] is a tool that determines the model spatial features using 3D conformations and mean force statistical potentials, and considers model construction environmental parameters relative to expected environmental conditions. Model evaluation results can be compared to the same evaluation performed on the template/s [309].

1.12.6 The MODELLER program

All modelling tools and servers have pros and cons [356]. MODELLER is known to be one of the best tools for homology modelling [357]. MODELLER predicts protein structures by satisfying spatial restraints. The restraints include 1) homology-derived restraints of distances and dihedral angles, 2) stereochemical restraints taken from the CHARMM-22 molecular mechanics force field [358], 3) statistical preferences for dihedral angles and non-bonded interatomic distances [359], and 4) user-defined restraints such as from NMR spectroscopy, rules of secondary structure packing, cross-linking experiments, fluorescence spectroscopy, image reconstruction from electron microscopy, site-directed mutagenesis and intuition. The input is generally a sequence alignment of the target and template structure/s, the atomic coordinates of the template/s, as well as a small script file. MODELLER then automatically generates a model consisting of all non-hydrogen atoms [309].

1.13 *In silico* docking

Molecular docking is a rapid, popular, and economical computational method used to predict the binding modes and affinities of molecular recognition events like protein-ligand and protein-protein binding [360]. Protein-ligand docking is an intense research topic and many docking software are currently available, such as AutoDock [361], GOLD [362], DOCK [363], FlexX [364], and Glide [365], which use various algorithms to perform docking. In general, protein-ligand docking methods consist of a search algorithm and a scoring function. The search algorithm goes through several ligand positions and poses with a target protein, and the scoring function estimates the binding affinities of the various poses and ranks them [18]. Docking reproduces chemical potentials based on the bound conformation preference and the free energy of binding [366]. The sum of the van der Waals interactions, coulombic interactions and the formation of hydrogen bonds are

approximated by a docking score which is an indication of binding potential [367]. Docking accuracy is influenced by many things, such as protein structure quality, protein environment, binding site environment, etc [368]. Docking programs are usually ranked by ligand pose prediction and binding affinity estimation; these are two very different measurements, the second being dependent on the first. But good geometry does not imply a good estimation of affinity.

1.13.1 AutoDock Vina

Autodock Vina is an easy to use, free open-source package, with parallel computing ability that can quickly calculate ligand-binding affinity [369,370]. In the CASF-2013 ligand-binding affinity benchmark [371], Vina showed better accuracy compared to its parent software AutoDock4 and has therefore become more popular. In the year 2016, an extensive test of ten docking programs on a broad range of protein-ligand complexes showed AutoDock Vina to have the best binding affinity estimation, beating five commercial docking programs as well as AutoDock4 [372]. Recently however, Nguyen et al. 2020 [370] tested AutoDock4 against AutoDock Vina using 800 protein-ligand complexes having PDB structures and experimental binding affinity data. They found that Autodock Vina's binding poses are more accurate, but Autodock4 does better in binding affinity.

The overall functional arrangement of the conformation-dependent section of the scoring function that Vina uses is

$$c = \sum_{i<j} f_{t_i t_j}\left(r_{ij}\right)$$

where the summation includes all of the atom pairs that can move relative to each other, usually omitting 1–4 interactions. Every atom *i* is given a type t_i, and a symmetric set of

interaction functions $f_{t_i t_j}$ of the interatomic distance r_{ij} should be defined. Therefore, the value can also be represented as a sum of intermolecular and intramolecular contributions:

$$c = c_{inter} + c_{intra}$$

The optimization algorithm tries to locate the global minimum of *c,* as well as other low-scoring conformations, and subsequently ranks them. The predicted free energy of binding is determined using the intermolecular section of the lowest-scoring conformation, designated here as 1:

$$s_1 = g(c_1 - c_{intra1}) = g(c_{inter1})$$

where the function *g* can be an arbitrary strictly increasing smooth possibly nonlinear function.

The derivation of the Vina scoring function integrates advantages of knowledge-based potentials and empirical scoring functions by obtaining empirical data from conformational preferences of the receptor-ligand complexes and the experimental affinity measurements.

The interaction functions are defined relative to the surface distance, where R_t is the van der Waals radius of atom type *t*.

$$\text{Interaction functions: } f_{t_i t_j}(r_{ij}) \equiv h_{t_i t_j}(d_{ij})$$

$$\text{Surface distance: } d_{ij} = r_{ij} - R_{t_i} - R_{t_j}$$

In the scoring function, $h_{t_i t_j}$ is a weighted sum combining steric interactions (identical for all atom pairs), hydrophobic interactions between hydrophobic atoms, and, where applicable, hydrogen bonding.

The steric terms are:

$$gauss_1(d) = e^{-(d/0.5\text{Å})^2}$$

$$gauss_2(d) = e^{-((d-3\text{Å})/2\text{Å})^2}$$

$$repulsion(d) = \begin{cases} d^2, if\, d < 0 \\ 0, if\, d \geq 0 \end{cases}$$

The hydrophobic term equals 1, when d < 0.5Å; 0, when d > 1.5Å, and is linearly interpolated between these values. The hydrogen bonding term equals 1, when d < −0.7Å; 0, when d > 0, and is also linearly interpolated between the values. All interaction functions $f_{t_i t_j}$ are cut off at r_{ij} = 8Å.

The Vina optimization algorithm uses an Iterated Local Search global optimizer [373,374] with a series of steps involving a mutation and a local optimization, with step approval based on the Metropolis criterion [375]. For local optimization, Vina utilises an efficient quasi-Newton method called the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [376] method which uses the scoring function gradient in addition to the scoring function value, *i.e.,* the derivatives of the scoring function with respect to its arguments. The coordinates and pose of the ligand are the arguments, including the ligand active rotatable bond torsion values, as well as flexible residues. The amount of steps of a run depends on the seeming intricacy of the problem, and numerous runs originating from random conformations are carried out.

1.13.2 AutoDock Vina-Carb

Most molecular docking programs are created to align rigid, drug-like compounds into the binding sites of macromolecules, but they often perform inadequately when docking flexible carbohydrate molecules. Therefore, in this thesis molecular docking was performed using Vina-Carb [377] which is a variant of AutoDock Vina used specifically for carbohydrate docking. Vina-Carb incorporates particular carbohydrate characteristics, such as glycosidic torsion angle preferences and intramolecular energies of glycosidic

linkages. In this way, it integrates the advantage of Vina's large search space with a powerful intramolecular energy evaluation [103].

Oligosaccharide glycosidic linkages are an important source of flexibility, caused primarily by the two or three rotatable bonds that link monosaccharides [378,379]. However, glycosidic linkage angles only adopt a known subset of conformations that are dictated by carbohydrate-specific stereo-electronic and solvent-solute interactions [380–382]. Carbohydrate Intrinsic (CHI) energy functions are used to quantify the glycosidic torsion angle preferences [383] by assigning relative energies to the torsion angles. The distribution of glycosidic torsion angles in protein-carbohydrate complexes taken from the Protein Data Bank correlate with the CHI-energy profiles. By applying the CHI energies, a ligand orientation with either favourable or unfavourable glycosidic torsion angles would be promoted or penalized, respectively.

During the development and testing of Vina-Carb, taking into account the intramolecular energies of the glycosidic linkages resulted in a pose prediction success rate of 74% for Vina-Carb, compared to 55% for AutoDock Vina [377].

1.14 Molecular dynamics simulations

Proteins and other molecules are flexible, dynamic, and exhibit a complex network of interactions – this must be considered when studying their functionality and for understanding living systems [11,384]. The binding/recognition process is dynamic, not only structurally but also energetically [117]. Physics-based MD simulations use computers to model a physical system that predicts the movement of every atom in the system based on interatomic interactions, and these simulations are improving all the time, helping to elucidate protein function and many other natural phenomena [385,386]. MD produces very fine spatial and temporal resolution that is currently not possible in physical experiments. The simulations are able to provide details regarding a range of significant

biomolecular processes such as conformational change, protein folding, and ligand binding. They are also able to predict how biomolecules will respond to perturbations such as mutation, phosphorylation, protonation, or the addition or removal of a ligand [387]. Generally, an MD user selects a starting molecular configuration, specifies the atomic interactions and model physics (force field), simulates the system, and then analyses the trajectory [388]. Before a simulation, the molecular system must be prepared by adding missing atoms such as hydrogen atoms which are usually not resolved in crystal structures (as well as any missing residues, loops, domains), and adding solvent molecules such as water, salt ions, and (for a membrane protein) lipids [387].

The planning and analysis of MD simulations should take into account certain limitations. For instance, although the accuracy of MD simulations and their force fields are ever improving, the results are still inherently approximate [389]. In addition, accurate simulation is usually influenced by the accuracy of the experimental protein structure or homology model utilised. It must be kept in mind that covalent bonds do not break or form during classical MD simulations and the protonation states of titratable amino acid residues are fixed – these should be set thoughtfully before a simulation [390]. However, quantum MD simulations take explicitly into account the quantum nature of the chemical bond using equations of quantum mechanics and therefore can elucidate additional biological problems such as enzymatic reactions. Quantum MD simulations are much more computationally expensive and currently cannot be used when simulating very large biomolecular systems [384].

During an MD simulation, forces are usually calculated using a potential energy function of the system. At thermal equilibrium, the populated states are the low-energy regions of the potential energy function and the forces acting on individual atoms are related to the gradient of this function, which is why these functions are called “force fields” [391]. A protein force field consists of terms for bonded (bond lengths, bond angles, and dihedral

angles) and non-bonded interactions (van der Waals and electrostatics). Force fields have rather straightforward mathematical formulae; however, they have a number of empirical parameters that affect its accuracy [392]. The uncomplicated force field representation of molecular properties enables rapid energy and force calculations, even for large systems: Springs for bond length and angles, periodic functions for bond rotations, Lennard–Jones potentials for van der Waals interactions, and Coulomb's law for electrostatic interactions. After computing the forces that act on individual atoms, classical Newton's laws of motion are utilised to determine accelerations and velocities and then the atom coordinates are updated. Integration of movement is performed numerically, so to prevent instability a time step must be used that is shorter than the fastest movements in the system; usually between 1 and 2 fs for atomistic simulations. Microsecond-long simulations need to iterate calculations 10^9 times! Fortunately, the performance of MD simulations has been strengthened by fine-tuning energy calculations, parallelization, or the use of graphical processing units (GPUs) [117].

Solvent representation is important when setting up the system, as water molecules are essential for protein structure and dynamics [393]. The explicit-solvent representation of molecules has shown to be the best and is the simplest. However, due to the bigger size of the simulated systems most of the computational resources are utilised calculating forces on water molecules [392]. Explicit-solvent captures most of the solvation effects of real solvent, even entropic effects such as hydrophobicity [117]. The TIP3P and TIP4P water models are widely used [394]. For efficiency purposes, implicit-solvent methods have been developed, mostly based on the Generalized Born (GB) solvation model [395–397]. Also, a solvent-accessible surface areas (SASA) energy term is occasionally used for estimating non-polar contributions to solvation [395].

The protein force fields that are used most frequently, integrate the following somewhat basic potential energy function:

$$V(r) = \sum_{bonds} k_b\,(b - b_0)^2 + \sum_{angles} k_\theta\,(\theta - \theta_0)^2 + \sum_{torsions} k_\varphi\,[cos(n\varphi + \delta) + 1]$$

$$+ \sum_{nonbondpairs} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right]$$

where the first three summations represent the bonds (1-2 interactions), angles (1-3 interactions), and torsions (1-4 interactions). The final sum, regarding pairs of atoms *i* and *j*, represents electrostatic interactions via Coulomb's law that use partial charges q_i on each atom. A Lennard-Jones 6-12 potential controls the dispersion and exchange repulsion forces which is sometimes referred to as the ''van der Waals'' term. This basic potential energy function is able to create the fundamental features of protein energy landscapes at atomic detail. Thus, a force field is comprised of a potential energy function and its parameters [391].

In classical MD simulations, the laws of classical mechanics define the dynamics of the system where atoms are represented by soft balls and bonds are represented by elastic springs (ball and spring representation) [384]. Using the coordinates of all the atoms that constitute a biomolecular system (e.g., a protein in water), the force exerted on each atom by all of the other atoms can be determined. Newton's laws of motion are used to predict the spatial position of each atom as a function of time by repeatedly calculating the forces on each atom and then using those forces to update the position and velocity of each atom [387]. The time evolution in a system of interacting particles is followed using the solution of Newton's equations of motion (EOM):

$$F_i = m_i \frac{d^2 r_i(t)}{dt^2}$$

where $r_i(t) = (x_i(t), y_i(t), z_i(t))$ and is the position vector of the *i*th particle and F_i is the force acting on the *i*th particle at time *t*, and m_i is the mass of the particle. The particles are

usually atoms, but they could represent any distinct entities such as specific chemical groups [384].

In the thesis, the software package Amber was used to setup the MD simulations, and the software package GROMACS was used to run and analyse the simulations. The term “Amber” sometimes also refers to the empirical force fields that are very widely used, one of which (ff14SB) was used in the thesis [398]. The main Amber preparation programs are *antechamber* and *LeaP*. Antechamber provides force fields for residues or organic molecules which are not in standard libraries by inputting a 3D structure and automatically assigning charges, atom types, and force field parameters. LEaP builds biopolymers from the component residues, solvates the system, and prepares lists of force field terms and their associated parameters. The Amber preparation produces a coordinate file (prmcrd) with the atom Cartesian coordinates, and a parameter-topology file (prmtop) with all the other information required to calculate energies and forces including atom names and masses, force field parameters, lists of bonds, angles, and dihedrals [398]. The main Amber molecular dynamics program is called sander, and the analysis program is called ptraj. However, GROMACS was used for running and analysing the simulations in the thesis. GROMACS is a very popular open-source and free software used mostly for dynamical simulations of biomolecules. It has a wide range of calculation types, preparation, and analysis tools [388].

1.14.1 MD analysis

The analysis of protein MD simulations usually requires quantitative analysis and visual analysis [387]. In the thesis, MD trajectories were analysed using protein and ligand Root Mean Square Deviation (RMSD), protein radius of gyration (R_g), and residue Root Mean Square Fluctuation (RMSF) calculations. The overall movement of the proteins and ligands was visually inspected using VMD software [399]. Protein-ligand hydrogen bonding

information was extracted, the binding free energy between the protein-ligand complexes was calculated, and principal-component analysis (PCA) was performed.

1.14.1.1 RMSD

Protein and ligand RMSD were calculated to determine their stability during the simulations. RMSD measures the deviation of a structure from a specific conformation; in MD simulations it is usually the initial conformation. The RMSD is the root mean squared Euclidean distance in *3N* configuration space as function of the time step. With GROMACS, RMSD is computed using the *gmx rms* command with respect to the reference starting structure, using the equation:

$$\rho^{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=l}^{N}(r_i(t) - r_i{}^{ref})^2}$$

where $r_i(t)$ represents the current coordinates at time *t*, and $r_i{}^{ref}$ represents the reference coordinates [400].

1.14.1.2 R_g

The expansion or contraction (compactness) of proteins during simulations can be assessed by calculating R_g. The deviation of the R_g values is normally very low unless there are major conformational changes of the protein. The GROMACS program *gmx gyrate* calculates R_g using the equation:

$$R^2{}_{gyr} = \frac{1}{M}\sum_{i-l}^{N} m_i\,(r_i - R)^2$$

where $M=\sum^{N}_{i=1} m_i$ is the total mass and $R=N^{-1}\sum^{N}_{i=1} r_i$ is the centre of mass of the protein consisting of *N* atoms [401].

1.14.1.3 RMSF

RMSF is a measure of the flexibility of individual residues and indicates the regions of the system that are the most mobile. While RMSD is normally calculated to an initial state, RMSF is calculated to an average structure of the simulation. RMSF is usually restricted to backbone or alpha-carbon atoms because they better represent conformational changes compared to the more flexible side chains. RMSF is the square root of the variance of the fluctuation around the average position and can be determined with the *gmx rmsf* command which uses the equation:

$$\rho_i^{RMSF} = \sqrt{\langle (r_i - \langle r_i \rangle)^2 \rangle}$$

where r_i is the coordinates of particle i, and $\langle \boldsymbol{r}_i \rangle$ is the ensemble average position of i [402].

1.14.1.4 Hydrogen bond analysis

Protein-ligand hydrogen bonding information was extracted using the *hbond* command of AMBER's *cpptraj* program [403]. Hydrogen bonding is crucial for numerous chemical and biological processes, including ligand binding and enzyme catalysis [404]. The abundance and flexibility of hydrogen bonds make them the most essential physical interactions in biomolecule systems in aqueous solution. Biological macromolecules and solvating water consist of about 50% and 66% hydrogen atoms, respectively – therefore, in a biological system, hydrogen atoms (protons) exist between nearly every pair of noncovalently-bonded heavy atoms.

1.14.1.5 Binding free energy

Free energy calculations provide an estimation of protein-ligand binding affinity, as the calculations consider all thermodynamically relevant phenomenon such as protein dynamics/flexibility, explicit inclusion of the solvent, and the difference between protein-ligand interactions in the complex and their interactions with water and counterions when

unbound [18]. Free energy calculations are not completely accurate. Calculating protein-ligand binding free energies is extremely challenging as the large change in configurational enthalpy is difficult to assess in brute-force simulations [405]. Mikulskis *et al.* [406] estimated the relative free energies between 91 pairs of ligands from 10 different proteins and found that only 54% of the estimates have errors of 4 kJ/mol or less compared to experimental results. Although computational and experimental researchers have made good use of the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) binding free energy method, it is not as accurate as more computationally intensive methods. Accuracy depends on the system. Using MM/PBSA, Hou *et al*. [407] compared binding free energies with experimental values in various systems, in terms of correlation coefficients. They found that binding free energies between the protein avidin and its ligands correlate well with experimental values (correlation coefficient r = 0.92). For α-thrombin, neuraminidase, and P450cam, MM/PBSA displayed satisfactory performance (r = 0.68−0.81). MM/PBSA did not perform well on cytochrome C peroxidase (r = 0.27) and penicillopepsin (r = 0.41). However, MM/PBSA is a fast, effective, and reproducible way of investigating protein-ligand binding interactions in terms of binding free energies [408–410]. The majority of MM/PBSA experiments have been performed in high-throughput screening for drug discovery studies, but the MM/PBSA method was implemented in the thesis to obtain an approximation of ligand binding affinity for comparison between the different protein-ligand complexes.

Utilising the g_mmpbsa software tool [411], the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) method was employed to calculate and compare the binding free energies of the protein-ligand complexes during MD simulations. Using g_mmpbsa, the total binding free energy of a protein-ligand complex can be summed up in the following equation:

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand})$$

where $G_{binding}$ is an approximation of the total binding free energy between protein and ligand, $G_{complex}$ is the total free energy of the protein-ligand complex, and $G_{protein}$ and G_{ligand} are the total free energies of the isolated protein and isolated ligand, respectively. This general equation factors in van der Waals, electrostatic, polar solvation and apolar solvation energy constituents. The free energy of each of these variables is given by:

$$G_x = \langle E_{MM} \rangle - TS + \langle G_{solvation} \rangle$$

where $\langle E_{MM} \rangle$ represents the average molecular mechanics potential energy in a vacuum. T and S are the temperature and entropy, respectively, and together represent the entropic contribution to the free energy in a vacuum. Lastly, $\langle G_{solvation} \rangle$ is the free energy of solvation.

The vacuum potential energy (E_{MM}) consists of both the bonded and nonbonded interaction energies, and it is calculated using the following molecular mechanics (MM) force field parameters:

$$E_{MM} = E_{bonded} + E_{nonbonded} = E_{bonded} + (E_{vdW} + E_{elec})$$

where E_{bonded} represents bonded interactions which include bond, angle, dihedral and improper interactions. $E_{nonbonded}$ represents electrostatic and van der Waals interactions which are modelled using a Coulomb and Lennard-Jones potential function, respectively.

The energy needed to transfer a solute from vacuum into the solvent is called the free energy of solvation and is calculated with an implicit solvent model. The solvation free energy is denoted as:

$$G_{solvation} = G_{polar} + G_{nonpolar}$$

where G_{polar} and $G_{nonpolar}$ are the electrostatic and non-electrostatic contributions to the solvation free energy, respectively.

1.14.1.6 Essential dynamics and free energy landscapes

Utilised in the majority of scientific fields, principal component analysis (PCA) is very likely the most popular multivariate statistical technique [412] and was used in the thesis to determine protein essential dynamics. PCA can systematically decrease the number of dimensions required to characterise protein dynamics using a decomposition procedure to sort observed protein motions from the largest to smallest spatial scales [412–414]. PCA is a linear transform that withdraws the most significant data components using a covariance matrix or a correlation matrix (normalized PCA) generated using the atomic coordinates that depict the accessible degrees of freedom (DOF) of the protein, namely the Cartesian coordinates that describe atomic displacements in every conformation having a trajectory [415]. Essential dynamics is the method of applying PCA to a protein trajectory, as the principle/essential motions are determined using the set of sampled conformations [416–418]. Conformational transitions in proteins have functional significance, such as substrate binding. The large-scale protein motions are determined using the lowest PCA modes that have the largest variances [414,416] and a 2D plot can be generated. 2D projections provide insight into the formation of clusters and the accessible conformational space explored, however free energy landscapes (FELs) can also show the transition subspace along with conformer associated abundance [419,420]. The FEL theory of protein structure and dynamics proposes that the native state of a protein exists as an extensive ensemble of conformational states/substates that co-exist in equilibrium with various population distributions, and that the ligand binds selectively to the most suitable conformational state/substate, eventually shifting the equilibrium towards this state/substate [18]. In the thesis, the essential dynamics and FELs of ligand-bound and ligand-free proteins were analysed and compared. The covariance matrices were built and diagonalized by the GROMACS tool *gmx covar*, and the principal components and overlap was determined with *gmx anaeig*. Gibbs free energy landscapes were plotted using *gmx sham* [421].

1.15 Knowledge gap

Extensive knowledge of protein-ligand interactions is critical to the comprehension of biology at the molecular level. Despite their conserved catalytic domain, GH1 enzymes have many different enzyme activities and/or substrate specificities, as a change of only a few residues in the active site can change the functional specificity. The structural determinants and molecular details of GH1 ligand specificity and affinity are very broad, highly complex, not well understood, and therefore still need to be clarified.

Our collaborators sent us three X-ray crystallographic structures from the bacterium *B. licheniformis* and based on initial sequence and structural comparisons they were all classified as putative 6Pβ-glycosidases. Despite the vital importance of 6Pβ-glucosidases and 6Pβ-galactosidases for bacterial energetic balance, a reduced number of these enzymes have been previously characterised in comparison to β-glucosidases. Unfortunately, the experimental substrates needed for a broad biochemical characterisation are frequently commercially unavailable (e.g., phospho-glycosidases) or too expensive for detailed kinetic analyses.

The GH1 active site consists of several subsites big enough to bind one monosaccharide unit. Because of the mostly weak sugar-protein interactions, multiple binding sites frequently combine to enhance the signal. Cooperative binding makes the binding affinity of sites difficult to measure, as the properties of one subsite are influenced by the binding of the other subsites. Most GH1 enzymes have little conserved specificity for the aglycon portion of ligands due to the variable spatial structure and amino acid composition of the aglycon subsite (subsite +1). Considering the conservation of the overall structures and active sites of GH1 enzymes in general, the differences at the +1 subsite and the entrance to the active site (surrounding loops) are likely to be involved in engendering substrate specificity to individual GH1 activities.

1.16 Research aims

The aim of this study was to classify the GH1 activity of three solved crystal structures from the bacterium *B. licheniformis*, and to provide evidence for their ligand-binding specificities. In addition and complimentary to this, the differences in structural and physicochemical contributions to enzyme specificity and/or function between different GH1 activities/enzymes had to be assessed. As the amino acid sequence of an enzyme determines its 3D structure which in turn determines its function, the sequence/structure/function relationship of the GH1 enzymes had to be analysed. The knowledge generated here would contribute to the body of research seeking to improve GH1 enzyme biotechnological applications, such as the production of biofuel from biomass.

To accomplish the research aims, sequence analyses involving sequence identity, phylogenetics, and motif discovery were performed. As protein structure is more conserved than sequence, the discovered motifs were mapped to 3D structures for structural analysis and comparisons. To obtain information on enzyme mechanism or mode of action, as well as structure-function relationship, computational methods such as docking, molecular dynamics, binding free energy calculations, and essential dynamics were implemented. These computational approaches can provide information on the active site, binding residues, protein-ligand interactions, binding affinity, conformational change, and most structural or dynamic elements that play a role in enzyme function.

Chapter 2: Sequence, structure, function relationship of enzyme *Bl*BglH

2.1 Introduction

This chapter delves into the sequence, structure, function relationship of a new GH1 enzyme crystallographic structure from the bacterium *B. licheniformis*. Sequence and structural analyses, together with *in silico* docking and MD simulations were performed to provide evidence for the specific GH1 activity and dynamic function of the enzyme.

Most of the work in this chapter is reproduced from the publication below. All of the work is my own, except for the enzyme 3D crystallographic structure determination and activity assays. All original writing and figure generation in this chapter is my own except for sections 2.2.6, 2.3.2 and 2.3.3.

> **Wayde Veldman**, Marcelo Liberato, Vitor Almeida, Valquiria Souza, Maira Frutuoso, Sandro Marana, Vuyani Moses, Özlem Tastan Bishop, and Igor Polikarpov. "X-ray Structure, Bioinformatics Analysis, and Substrate Specificity of a 6Pβ-glucosidase Glycoside Hydrolase 1 Enzyme from *Bacillus licheniformis*". *Journal of Chemical Information and Modeling*. 2020. 60 (12) 6392-6407. DOI: 10.1021/acs.jcim.0c00759.

2.2 Materials and Methods

2.2.1 Enzyme crystallographic structure from collaborators

From our collaborators at the University of São Paulo we obtained an unpublished crystallographic enzyme structure in PDB format from the bacterium *B. licheniformis*. The GenBank accession of the enzyme is AAU43027.1. The enzyme will be referred to as

*Bl*BglH hereafter. Enzyme *Bl*BglH was checked for rotamers, missing atoms, and missing residues. The crystallographic structure did not have any ligand in the active site; therefore, enzyme function and specificity were unclear. The *Bl*BglH crystallographic structure has now been published on the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) web server with PDB ID 6WGD. X-ray diffraction was used to determine the structure, and the quality metrics are as follows: Resolution – 2.20 Å; R-Value Free – 0.244; R-Value Work – 0.210; R-Value Observed – 0.212; Clashscore – 2; Ramachandran outliers – 0; Sidechain outliers – 0.9%; RSRZ outliers – 0.6%.

2.2.2 Sequence data retrieval

The CAZy database was used to retrieve protein sequences of characterised bacterial GH1 enzymes. The β-glucosidase (EC 3.2.1.21), 6Pβ-galactosidase (EC 3.2.1.85), 6Pβ-glucosidase (EC 3.2.1.86) sequences, as well as the dual-phospho activity sequences, were downloaded via the GenBank [422] links that are found in CAZy. A total of 29 β-glucosidase, 18 6Pβ-glucosidase, eight 6Pβ-galactosidase, and three dual-phospho activity sequences were retrieved (Supplementary Table 2.1). The remaining GH1 activities on CAZy were not incorporated into the sequence dataset because the enzymes from those activities have different overall structure and sequence lengths as compared to *Bl*BglH – these sequences are few and far between though. The *Bl*BglH sequence (AAU43027.1) was then added to the retrieved sequences, making a total of 59 sequences in the final dataset.

2.2.3 Multiple sequence alignment

The PROfile Multiple Alignment with predicted Local Structures and 3D constraints (PROMALS3D) alignment tool, along with the GH1 enzyme sequence dataset, were used to generate a multiple sequence alignment (MSA). Alignment accuracy is promoted when

crystallographic secondary structure information is added to PROMALS3D, therefore crystallographic structures (if any) of the enzymes in the dataset were uploaded with its full GenBank enzyme sequence. The PDB files of the crystallographic structures were first checked for residue rotamers to prevent sequence residue duplicity. Once the MSA was generated, the alignment was inspected and fine-tuned using Jalview alignment viewer [423]. Lastly, the sequences in the alignment that were from the crystallographic structures were removed.

2.2.4 Phylogenetic analysis

The MSA was used with Molecular Evolutionary Genetic Analysis (MEGA) v7.0.26 [424] software to construct phylogenetic trees. The evolutionary models with the three lowest Bayesian information criterion (BIC) scores were chosen, and for each model three different gap deletions (90, 95 and 100%) were used. A strong branch swap filter was utilised with each tree, with 1000 bootstrap replicates. The best models for all three gap deletions were determined to be the Whelan and Goldman model [425] with Gama distribution and Invariant sites (WAG+G+I), the Le and Gascuel model [426] with Gama distribution (LG+G) and the Le and Gascuel model with Gama distribution and Invariant sites (LG+G+I). Three models at three different gap deletions meant that a total of nine maximum likelihood phylogenetic trees were constructed. In order to identify the best tree overall, the branching pattern and branch support of the nine trees were compared to their corresponding bootstrap consensus trees. The phylogenetic tree constructed employing the WAG+G+I model with 100% gap deletion was selected as the most accurate tree.

2.2.5 Motif analysis and structure mapping

The discovery of conserved motifs within the 59 enzyme sequences was achieved using Multiple Expectation Maximization for Motif Elicitation (MEME) v4.9.1 [427] software. A motif width of between five and ten residues was set. To identify and omit any overlapping

motifs, the Motif Alignment Search Tool (MAST) was utilised, resulting in 70 significant motifs (Supplementary Table 2.2). By parsing the MEME log file, a motif heatmap was generated that shows the conservation of motifs between the different GH1 enzyme activities and sequences – this was done using a Python script previously written by Faya *et al*., [428]. Motifs that were present in the majority of sequences in each respective activity were then mapped to representative enzyme crystallographic structures.

2.2.6 Activity assays

The enzyme activity assays were performed by our collaborators at the University of São Paulo using the substrate *p*-nitrophenyl β-glucopyranoside (PNPβglc), from Sigma-Merck, prepared in 100 mM MES (2-(N-morpholino)ethanesulfonic acid) pH 6 buffer containing either 50 mM NaCl or 50 mM sodium phosphate. Reactions were interrupted by adding 0.5 M $NaCO_3$ and the product (*p*-nitrophenolate) detected by absorbance at 415 nm. Initial rates of substrate hydrolysis (v_0) and substrate concentration data were analysed using the Lineweaver-Burk plot. In order to describe the mechanism of enzyme activation by phosphate, a kinetic equation relating the v_0 to the substrate and phosphate concentration was deduced assuming rapid equilibrium conditions following standard procedures previously described [19].

2.2.7 Homology modelling

Several enzymes from the 6Pβ-galactosidase activity had to be modelled since only one crystallised GH1 enzyme from the 6Pβ-galactosidase activity has been published [183] and its 3D structures (PDB IDs 1PBG, 2PBG, 3PBG and 4PBG) deposited in the Protein Data Bank (PDB). More than one 6Pβ-galactosidase activity enzyme structure was required for a part of this study, therefore three target 6Pβ-galactosidase enzymes with GenBank accessions AAA16450.1, AAD15134.1, and BAA07122.1 were structurally modelled. HHpred [429] and PRotein Interactive MOdeling (PRIMO) [430] webservers

aided in identifying the best suited template structures. A PDB structure from a β-glucosidase enzyme (PDB ID 3W53), and two different structures from the same 6Pβ-galactosidase enzyme (PDB IDs 1PBG and 4PBG), all chain A, were used as templates with MODELLER version 9.23 [315] to generate 100 models per target enzyme. The top three models per target enzyme, ranked by normalised z-DOPE (Discrete Optimized Protein Energy) score [315], were further evaluated using PROCHECK [352], Qualitative Model Energy Analysis (QMEAN) [431] and Verify3D [355] tools. Supported by a consensus using all three model quality evaluations, the top model of each target protein was selected. The templates were of exceptional quality, which produced great-quality models (Supplementary Table 2.3).

2.2.8 *In silico* docking

2.2.8.1 Docking validation

The program AutoDock Vina-Carb [377] was employed to validate the docking procedure using the 6Pβ-glucosidase crystallographic structure of *Streptococcus mutans UA150* (PDB ID 4GPN, chain A), containing a gentiobiose-6-phosphate ligand. The gentiobiose-6-phosphate ligand was first removed from the protein and then re-docked into the protein, utilising blind docking. A grid box search space of dimension size: x = 75.0 Å, y = 75.0 Å, z = 75.0 Å was defined, the ligands were centred at x = 0.27, y = 6.16, z = 28.41, and a search exhaustiveness value of 300 was applied. The RMSD between the original ligand and the docked ligand was 0.43 Å, computed with the GROMACS RMSD command *gmx rms*. Both the original crystallised ligand and the docked ligand exhibited the same pose, and they both interacted with a great majority of the same enzyme residues (Supplementary Figure 2.1). Ultimately, the docking procedure was considered reliable.

2.2.8.2 *Ligand docking*

To provide evidence of enzyme activity, blind docking of a positive- and a negative-control ligand to the *Bl*BglH enzyme was executed. p-Nitrophenyl-beta-D-galactoside-6-phosphate (PNP6Pgal) and p-Nitrophenyl-beta-D-glucoside-6-phosphate (PNP6Pglc) were the positive- and negative-control ligands used, respectively. One difference exists between the ligands: the C4 atom configuration. The ligand O4 hydroxyl group is in an equatorial position in the gluco epimer, but an axial position in the galacto epimer (Supplementary Figure 2.2). Prior to the docking calculations, the acid/base catalytic glutamate (Glu175) of the *Bl*BglH structure was protonated (Glh175) which is in agreement with the retention of configuration catalysis mechanism of GH1 enzymes [146,432]. For the docking, AutoDock Vina-Carb and chain B of the *Bl*BglH crystallographic structure were used. A grid box search space of dimension size: x = 75.0 Å, y = 75.0 Å, z = 75.0 Å was defined, the ligands were centred at z = -40.7, y = -29.6, z = 41.6, and a search exhaustiveness value of 300 was applied.

2.2.9 Molecular dynamics

Ligand partial atomic charges were assigned to a fully protonated ligand based on the AM1-BCC method of the AmberTools17 [433] antechamber program. The generated ligand mol2 file was converted to a PDB file using DiscoveryStudio [434]. On the other hand, the protein was protonated using the H++ webserver [435] at a pH of 6. H++ outputs the finished structure in AMBER inpcrd/prmtop format, which was then converted to a PDB file using the AmberTools17 ambpdb program. The ligand and protein PDB files were concatenated to produce a protein-ligand complex PDB file. AMBER topology files were then generated with programs parmchk2 and tleap of the AmberTools17 package. At this step, solvent in the form of water molecules was added to a cubic periodic box with a minimum distance of 10 Å from the protein edge and modelled with the TIP3P water

model. Using 0.15 M NaCl, the systems were then neutralised. The general amber force field 2 (GAFF2) was utilised for the van der Waals and bonded parameters. ACPYPE (AnteChamber PYthon Parser interfacE), also of the AmberTools17 package, converted these topology files to GROMACS format.

Before the MD production runs, all molecular systems were energy minimized using a conjugate-gradient, being energy relaxed with a force tolerance of 1000 kJ/mol/nm and capped at a maximum of 50,000 steps. The temperature of the systems was then equilibrated at 303.15 K over a period of 100 ps, utilising the velocity rescale thermostat (modified Berendsen thermostat) [436]. The pressure was equilibrated using the Parrinello-Rahman barostat [437] to maintain the system pressure at 1 bar. Production runs used the all-atom AMBER ff14SB force field [438] with GROMACS 2016.4 [388] on 240 cores (CPU: Intel® Xeon®) at the Centre for High Performance Computing (Cape Town, South Africa). Each simulation was run at a temperature of 303.15 K (30 °C) and a pH of 6 – these match the conditions that were used during the *in vitro* activity assay. The simulations were run for a period of 400 ns with 2 fs time steps, and the coordinates were written every 2 ps. To correct for rotational bond lengthening, the Linear Constraint Solver (LINCS) algorithm [439] was employed during the simulations. Using a Fourier grid spacing of 0.16 nm, long-distance electrostatic interactions were calculated with the Particle-Mesh Ewald (PME) algorithm. Coulombic and van der Waals short-range cut-offs were set to 1.0 nm. To promote the credibility of the results, duplicate MD simulations of 200 ns were executed.

2.2.10 Analysis of MD trajectories

After the completion of the MD runs, the trajectories were corrected for periodic boundary conditions, any jumps across boundaries were removed, and the protein was centred inside the simulation box. The RMSD, RMSF and R_g were computed with the GROMACS tools *gmx rms*, *gmx rmsf* and *gmx gyrate*, respectively. Using the backbone atoms, protein

RMSD was calculated after least square fitting along the backbone atoms. On the other hand, ligand RMSD was calculated by least square fitting to the ligand itself – this was done to visualize the stability of the ligand graphically. RMSF calculations were computed in order to observe the individual residue motion during the simulation. The R_g was calculated which indicates protein expansion or contraction during the course of the simulation. In addition, protein motion and dynamics were visually examined with the aid of the visual molecular dynamic (VMD) software tool [399]. The protein-ligand hydrogen bonding data was extracted using the hbond command from AMBER's cpptraj [403] program, with standard/default geometric criteria/definitions. Molecular interactions other than hydrogen bonds, and including hydrogen bonds, were monitored via DiscoveryStudio.

2.2.11 Binding free energy calculations

The strength of a biomolecular interaction, such as catalysis or recognition, can be quantified using binding free energy calculations [411]. The molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) method [440] was executed with the g_mmpbsa software tool [411] to determine the binding free energies of the protein-ligand complexes.

The final 15 ns of the 400 ns MD trajectories were used for the binding free energy calculations, sampled at 10 ps time intervals. The g_mmpbsa MmPbSaStat.py Python script was used to provide an overview of the final energies.

2.3 Results and discussion

2.3.1 Sequence analysis

2.3.1.1 Sequence identity

A multiple sequence alignment was generated using the *Bl*BglH sequence (AAU43027.1) and 58 retrieved GH1 sequences (Supplementary Figure 2.3). The MSA was then used to create a sequence identity heatmap (Figure 2.1). A high correlation between sequence

identity and GH1 activity exists, which was expected. However, certain exceptions were observed.

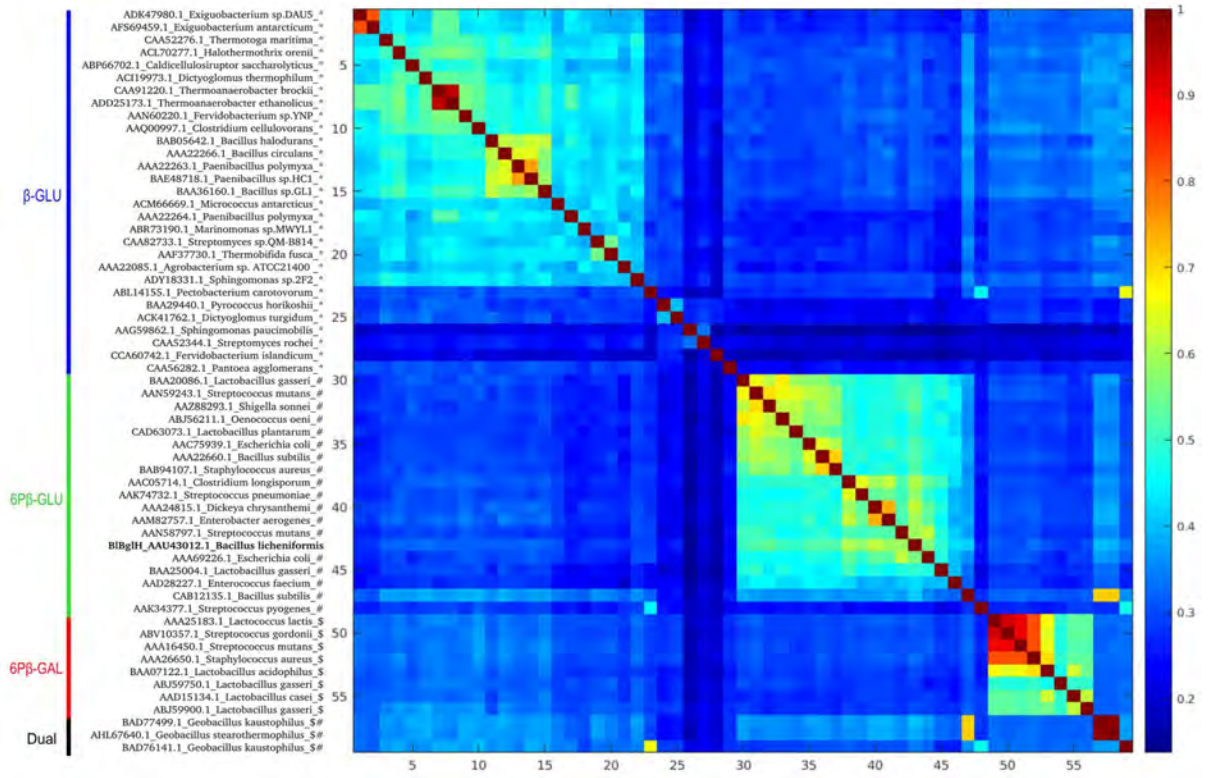


Figure 2.1. *Sequence identity heatmap showing the pairwise percentage identity between all 59 sequences used in Chapter 2. The X and Y axes indicate the 59 different GH1 enzymes. Identity scores are shown as a colour-coded matrix, calculated by comparing every sequence to each other (every sequence vs every sequence). Sequence identity increases from blue to red. 29 β-glucosidases (*), 18 6Pβ-glucosidases (#), 8 6Pβ-galactosidases ($) and 3 dual activity ($#) enzymes are labelled. Adapted from Veldman et al. 2020 [103].*

The top 22 β-glucosidases in the heatmap share a higher sequence identity than the bottom seven β-glucosidases. The lower sequence identity seen in the bottom seven β-glucosidases could point to a different phylogenetic lineage. Sequence ABL14155.1 (*Pectobacterium carotovorum)* is an odd β-glucosidase enzyme, as it shares its highest percentage identity (66%) with an enzyme characterised by CAZy as having dual activity: BAD76141.1 (*Geobacillus kaustophilus*). Further, ABL14155.1 shares its second highest percentage identity (47%) with AAK34377.1 (*Streptococcus pyogenes)*, which has 6Pβ-

glucosidase activity. The three sequences named above are unique among the all the sequences in the dataset; they form a group in the phylogenetic tree and motif heatmap that do not fall into a defined enzyme activity, as will be seen later in the study.

All of the 6Pβ-glucosidase sequences share a high sequence identity, except for CAB12135.1 (*Bacillus subtilis)* and AAK34377.1. Interestingly, CAB12135.1 shares its highest sequence identity by far (70%) with two of the three dual activity enzymes. Because of this high sequence identity, CAB12135.1 could be a dual activity enzyme and not just a 6Pβ-glucosidase enzyme as described on CAZy. The 6Pβ-galactosidase activity seems to be the most conserved: a higher shared sequence identity exists between the sequences of this activity as compared to the other activities – with a minimum of 53% shared identity, and the first four sequences sharing at least 77% identity. Curiously, the dual phospho-activity enzymes share a higher sequence identity with the β-glucosidases rather than the 6Pβ-glucosidases or 6Pβ-galactosidases, which is seen by looking at the bottom-left and top-right corners of the sequence identity heatmap in Figure 2.1. With regards to enzyme *Bl*BglH, it shares its highest average sequence identity with the enzymes of the 6Pβ-glucosidase activity (53%) as compared to the β-glucosidases (31%), 6Pβ-galactosidases (33%) and the enzymes with dual-phospho activity (39%).

2.3.1.2 Phylogenetic analysis

A phylogenetic tree was generated using the entire sequence dataset (Figure 2.2). It was seen that there were distinct phylogenetic clusters which correspond exactly to the sequence identity heatmap. The majority of the enzymes exhibited successful clustering based on their activity, however, unique sequences ABL14155.1 (*Pectobacterium carotovorum)*, CAB12135.1 (*Bacillus subtilis)* and AAK34377.1 (*Streptococcus pyogenes)* were found to deviate from this observation and did not fall into an activity cluster. These three enzymes, along with the dual activity enzymes, were seen forming two distinct phylogenetic clusters. In one of these “unique” clusters, dual-phospho activity enzymes

BAD77499.1 and AHL67640.1 group with 6Pβ-glucosidase enzyme CAB12135.1. In the other unique cluster, dual-phospho activity enzyme BAD76141.1 groups with β-glucosidase enzyme ABL14155.1 and 6Pβ-glucosidase enzyme AAK34377.1.

The β-glucosidase activity is split into five smaller sub-clusters in the phylogenetic tree which means this activity has greater diversity as compared to the other activities. The 6Pβ-glucosidase activity is split into two sub-clusters, and the 6Pβ-galactosidase activity has only one major cluster. Enzyme *Bl*BglH was found to group into the 6Pβ-glucosidase activity with a strong bootstrap confidence.

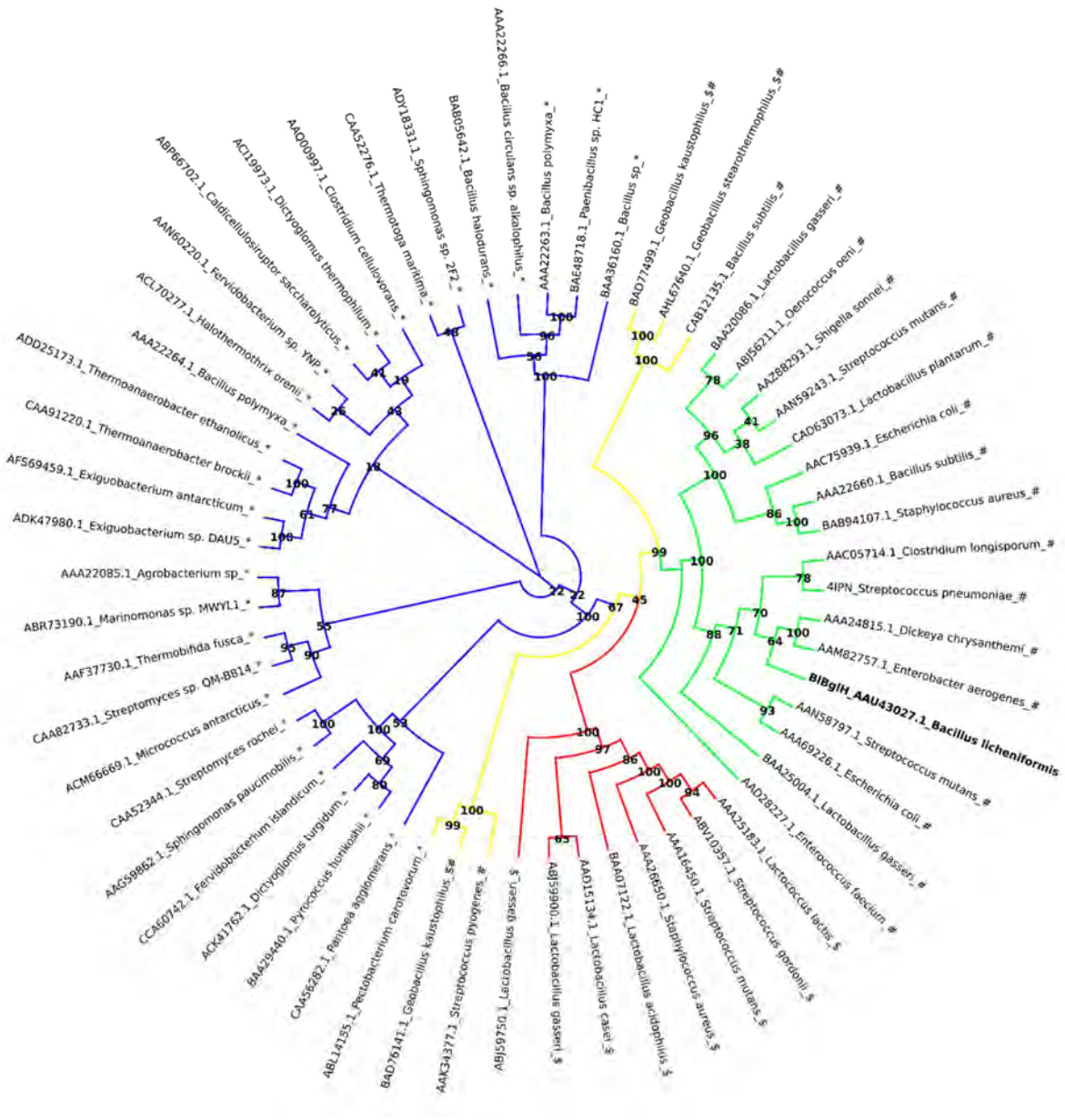


Figure 2.2. *Maximum likelihood phylogenetic tree consisting of all 59 sequences used in Chapter 2, generated using MEGA v7.0.26 program. Branch numbers indicate bootstrap values. Colour code: β-glucosidases – Blue. 6Pβ-glucosidases – Green. 6Pβ-galactosidases – Red. Unique clusters – Yellow. Adapted from Veldman et al. 2020 [103].*

2.3.1.3 Motif discovery

The motif searching software MEME v4.9.1 was used to locate sequence regions that are shared amongst the GH1 sequences. A motif heatmap was then generated showing the conservation of the shared sequence motifs (Figure 2.3). The groupings of the enzyme

activities were once again consistent with the previous analyses: the motif-separated groups in the motif heatmap match the groups seen in the phylogenetic tree. The β-glucosidases clustered into two major motif-separated groups, whereas the 6Pβ-glucosidases and 6Pβ-galactosidases formed one motif group each. The β-glucosidase activity enzymes have far fewer motifs compared to the other activities, especially the smaller β-glucosidase group, which is additional evidence of the greater diversity of this activity. In the motif heatmap, *Bl*BglH fell into the 6Pβ-glucosidase activity and was found to possess motifs that were only exhibited in the sequences of this activity. Motifs 29 and 34 are significant motifs as they are found only in the 6Pβ-glucosidase activity, *and* they are totally conserved in this activity. These and other motifs will be explored in a later section of this chapter, where they will be mapped to enzyme 3D structures. In total, the various sequence analyses point to enzyme *Bl*BglH having 6Pβ-glucosidase activity.

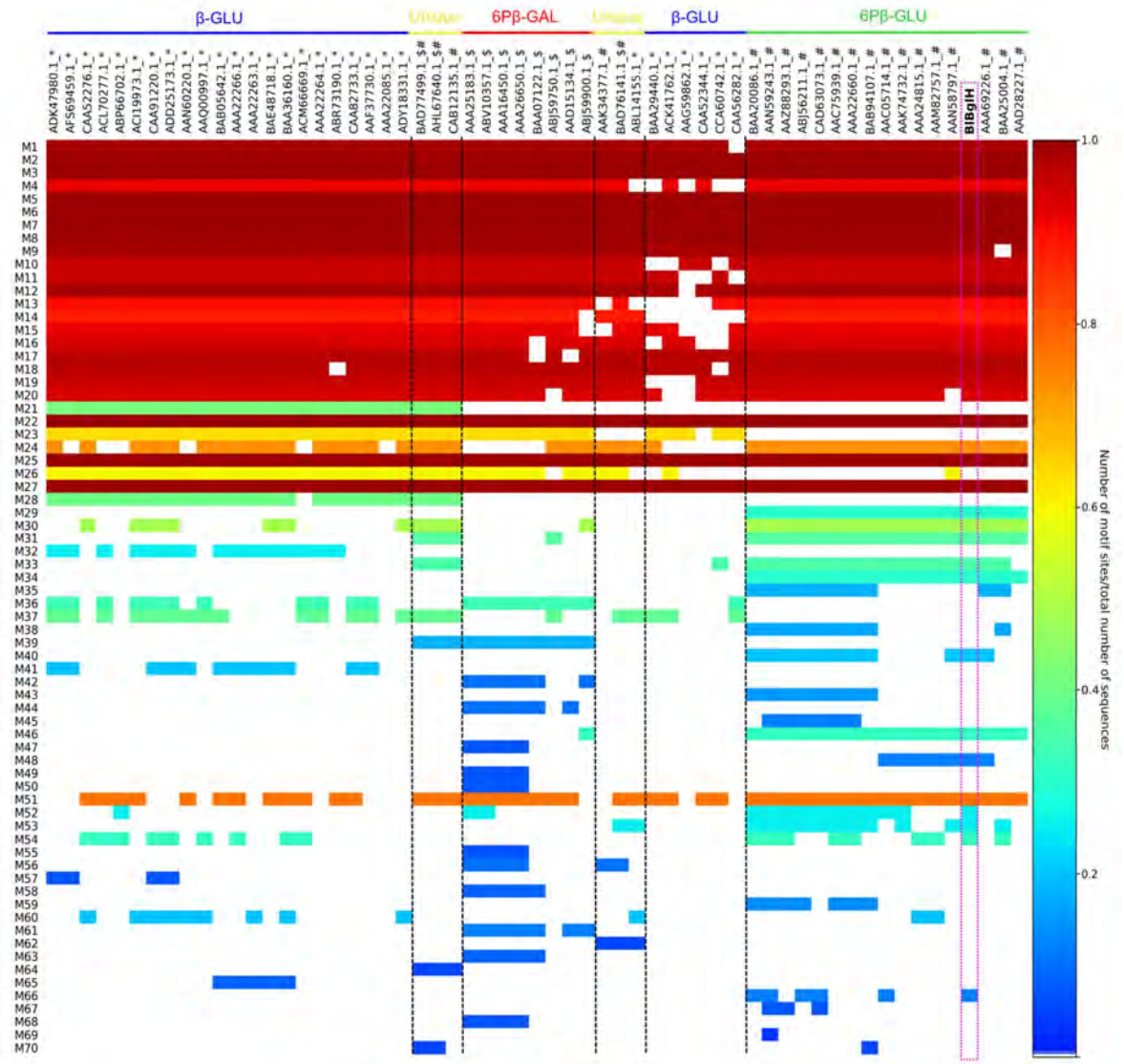


Figure 2.3. *Heatmap representing the 70 MEME discovered motifs in the 59 GH1 enzyme sequences used in Chapter 2. The colours show the motif conservation amongst all sequences, which increases from blue to red. White shows absence of motifs. Divisions are shown as dotted lines to indicate the sub-groups that are formed by similar sequences in terms of shared motifs. Adapted from Veldman et al. 2020 [103].*

2.3.2 Enzyme production and activity assay

To substantiate the predictions of the sequence analyses, our collaborators at the University of São Paulo expressed and purified enzyme *Bl*BglH, and its activity was verified. Seeing that 6-phospho-cellooligosaccharides and PNP-6-phospho-monosaccharides were unavailable for purchase, enzyme activity was deduced utilising non-phosphorylated substrates. On account of this, inclusion of 50 mM phosphate was

required to detect *Bl*BglH catalytic activity against PNPβglc – this is an indicator of the ion acting as an essential activator. Additionally, Lineweaver-Burk plots (1/v_0 *versus* 1/[PNPβglc]) in the presence of 2 to 50 mM phosphate showed that both slope and intercept of those lines decrease as a function of increasing concentrations of phosphate (Figure 2.4).

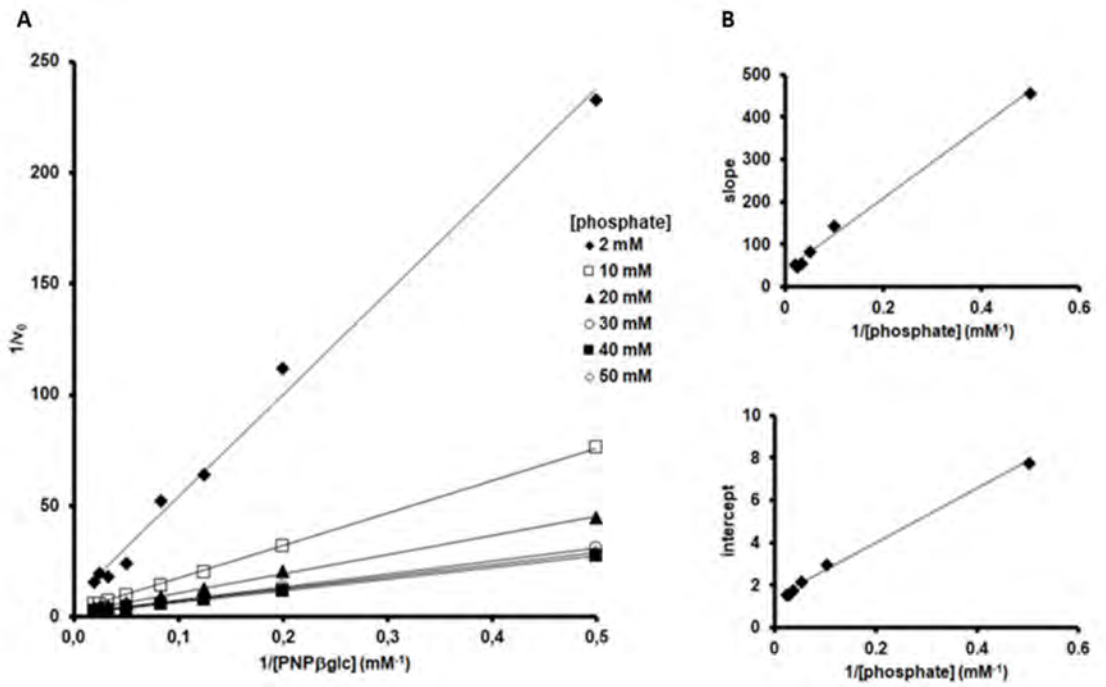


Figure 2.4. *Effect of the substrate PNPβglc and phosphate concentration on the activity of the enzyme BlBglH. (A) Lineweaver-Burk plots for the enzyme activity in the presence of different phosphate concentrations. (B) Plot of the slope and intercept of the lines observed in the Lineweaver-Burk analysis for different phosphate concentrations. Reproduced with permission from Veldman et al. 2020 [103].*

Based on the above, we studied various activation mechanisms in pursuit of one that would act accordingly, converging to a model in which only the enzyme-substrate-phosphate complex (ESA) would be productive. Furthermore, the model shows that phosphate can bind to the enzyme which would increase the affinity for the substrate, an effect reflected in the dissociation constant αK_s assuming $\alpha < 1$. Likewise, the binding of substrate may also enhance the enzyme affinity for phosphate, as seen by the dissociation

constant αK_A. Therefore α is a factor that quantitatively describes the mutual contribution of phosphate and PNPβglc to improve their affinities for the enzyme. PNPβgal was used with the same assay, but no activity was seen.

In agreement, the kinetic equation deduced from this model (Supplementary Figure 2.4) predicted that the presence of increasing concentrations of phosphate would decrease both the slope and intercept of lines of a Lineweaver-Burk plot ($1/v_0$ *versus* 1/[PNPβglc]), which was observed. In addition, the kinetic equation revealed that the slope and intercept of those lines are both a linear function of the phosphate concentration. As a result, based on secondary plots of slope *versus* 1/[phosphate] and intercept *versus* 1/[phosphate], the complete set of parameters describing the activation mechanism were determined, i.e., K_S = 66 mM, K_A = 21 mM, k_{cat} = 3.2 s^{-1} and α = 0.4.

Intriguingly, the parameter α indicates that the *Bl*BglH affinity for the substrate PNPβglc increased 2.5 times when phosphate was added. The substrate produced the same increment in the affinity for the phosphate. Also, phosphate is needed to form an active enzyme-substrate complex. This is an indication that the phosphate interactions within the active site have a significant role in stabilising the substrate binding in the ES and $ES^{\ddagger}$ complexes, which improves both the substrate affinity and the catalysis. This would be expected if *Bl*BglH's natural substrate is a "phospho-glucoside", as seen in 6Pβ-glucosidases. The activity assays performed here support the evidence of 6Pβ-glucosidase activity of enzyme *Bl*BglH.

2.3.3 *Bl*BglH crystallographic structure

The general protein folding of *Bl*BglH is a $(\beta/\alpha)_8$-barrel, which is a conserved structure that is seen in all the families affiliated with Clan-A of glycoside hydrolase enzymes, and this includes GH1 [159]. In a barrel shape, eight β-strands are encompassed by eight or more α-helices (Figure 2.5 A). The N-terminal portion of the loops connecting the secondary

structures at the C-terminal ends of the β-strands form a pocket where the substrate fits for catalysis (Figure 2.5 B), and the entrance of the binding site is shaped by loops L1, L4, L6 and L8. The DALI server [441] was used to search for proteins with similar structure to *Bl*BglH which resulted in a large number of GH1 enzymes with high similarity. The RMSD values between *Bl*BglH and the numerous GH1 enzymes were between 0.9 Å to 1.35 Å, substantiating the conserved folding in the GH1 family. Further, the *Bl*BglH structure is most similar to the 6Pβ-glucosidase activity. The lowest RMSD of 0.9 Å was seen with the 6Pβ-glucosidase enzyme BglA-2 from *Streptococcus pneumoniae*, which contains the ligand thiocellobiose-6P (PDB ID 4IPN). Like chain A and C of *Bl*BglH's structure, the loop L6 from BglA-2 was partially absent due to high flexibility (Figure 2.5 C).

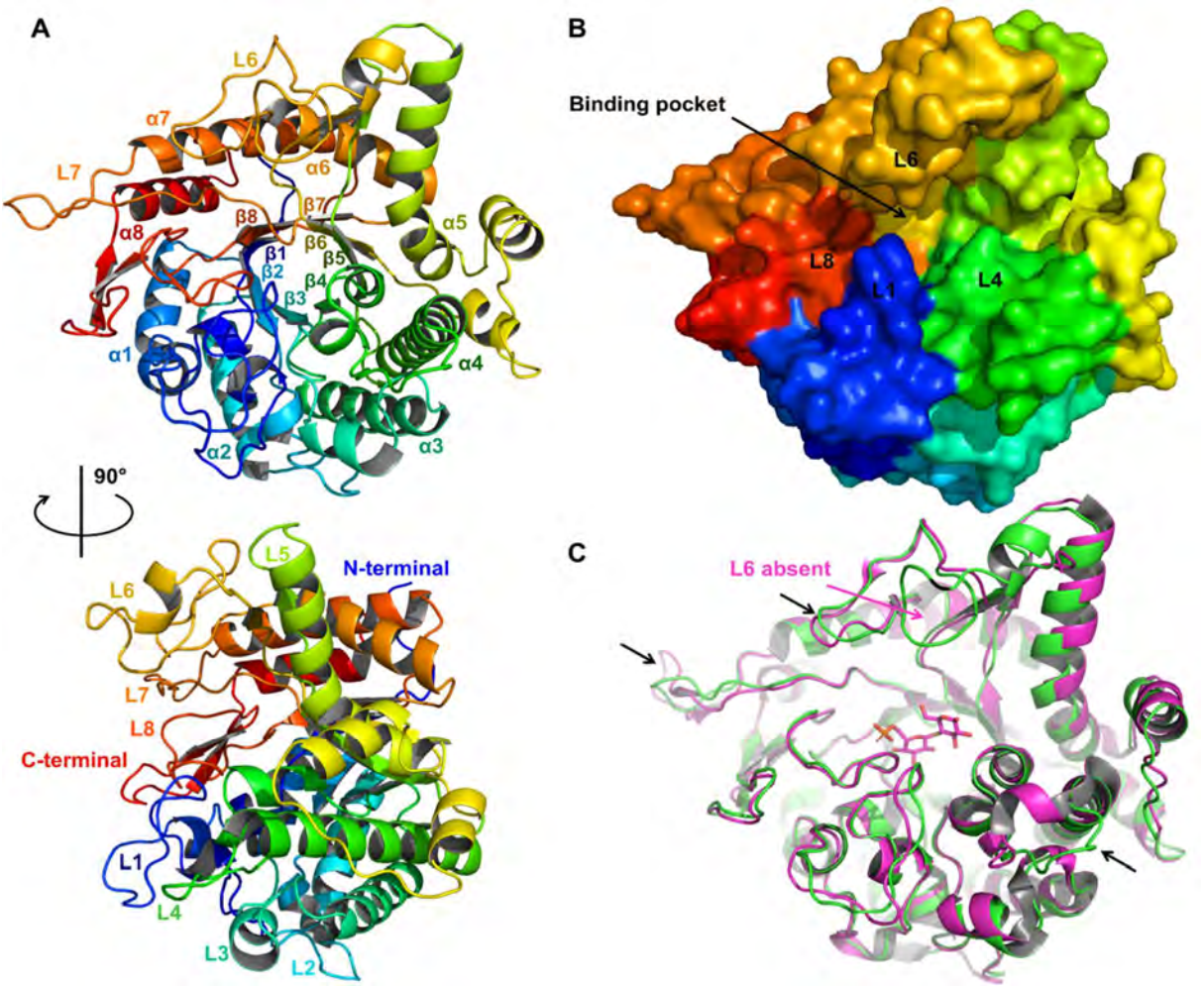


Figure 2.5. *Crystallographic structure of BlBglH (PDB ID 6WGD). (A) Cartoon representation of BlBglH coloured in rainbow gradient from dark blue (N-terminal end) to dark red (C-terminal end). BlBglH has a (ß/α)8-barrel fold that is conserved in all families from Clan-A, including GH1. The main α-helixes, β-strands and loops are labelled following the colours. (B) Surface representation of BlBglH with the same colour pattern used in A. Here, the substrate binding-pocket is shown with the entrance composed of loops L1, L4, L6 and L8. (C) Superposition of BlBglH and BglA-2 complexed with thiocellobiose-6P. The structures are highly similar with small differences in loops indicated by black arrows. Loop L6 from BglA-2 is partially absent due to high flexibility. Reproduced with permission from Veldman et al. 2020 [103].*

The *Bl*BglH and BglA-2 structures were superimposed, and it is seen that both enzymes are generally largely similar (Figure 2.6 A). Despite the amino acid sequence identity of 59.6%, the superposition shows that the residues interacting with thiocellobiose-6P are completely conserved between the structures. The phosphate group of the ligand is coordinated by hydrogen bonds with main chain nitrogens from Ala423 and Ser424, and

with side chain oxygens from Ser424 and Tyr432. Whereas the thiocellobiose is coordinated by hydrogen bonds with Gln22, His129, Tyr130, Asn174, Glu175, Glu368 and Trp415. Only one difference in the binding site is noted: BglA-2's M423 is substituted by Ala423 in *Bl*BglH.

Another binding site comparison was performed by superposing the *Bl*BglH structure with that of PGALase^{E375C} (Figure 2.6 B), a 6Pβ-galactosidase from *Lactococcus lactis* (PDB ID 4PBG) complexed with the ligand β-galactose-6-phosphate (6Pgal). Although PGALase has 6Pβ-galactosidase activity, the RMSD between the two structures is 1.25 Å and the majority of the residues are conserved. However, a few discrepancies at the side chains of conserved amino acids are noticed, most likely because of the absence of a ligand in the *Bl*BglH crystal (Figure 2.6 B – blue arrows). The binding sites do differ with regards to four amino acids though (Figure 2.6 B – red arrows): 1) PGALase^{E375C} has a tryptophan at position 429 as opposed to Ala423 and Met423 from *Bl*BglH and BglA-2, respectively. A hydrogen bond is formed between Trp429 and 6Pgal, while Ala423 and Met423 may not be able to accommodate 6Pgal; 2) PGALase^{E375C} has its native catalytic Glu375 mutated to cysteine; 3) PGALase^{E375C} has an asparagine (Asp297) interacting with 6Pgal at O1. This interaction seems to exist only in the mutated form of PGALase, as the Asn297 residue is more easily accommodated due to the mutation opening up additional space for the ligand; 4) PGALase^{E375C} has a phenylalanine at position 117 in contrast to tyrosines (Tyr130) from *Bl*BglH and BglA-2, which forms hydrogen bond with the ligand at +1 subsite.

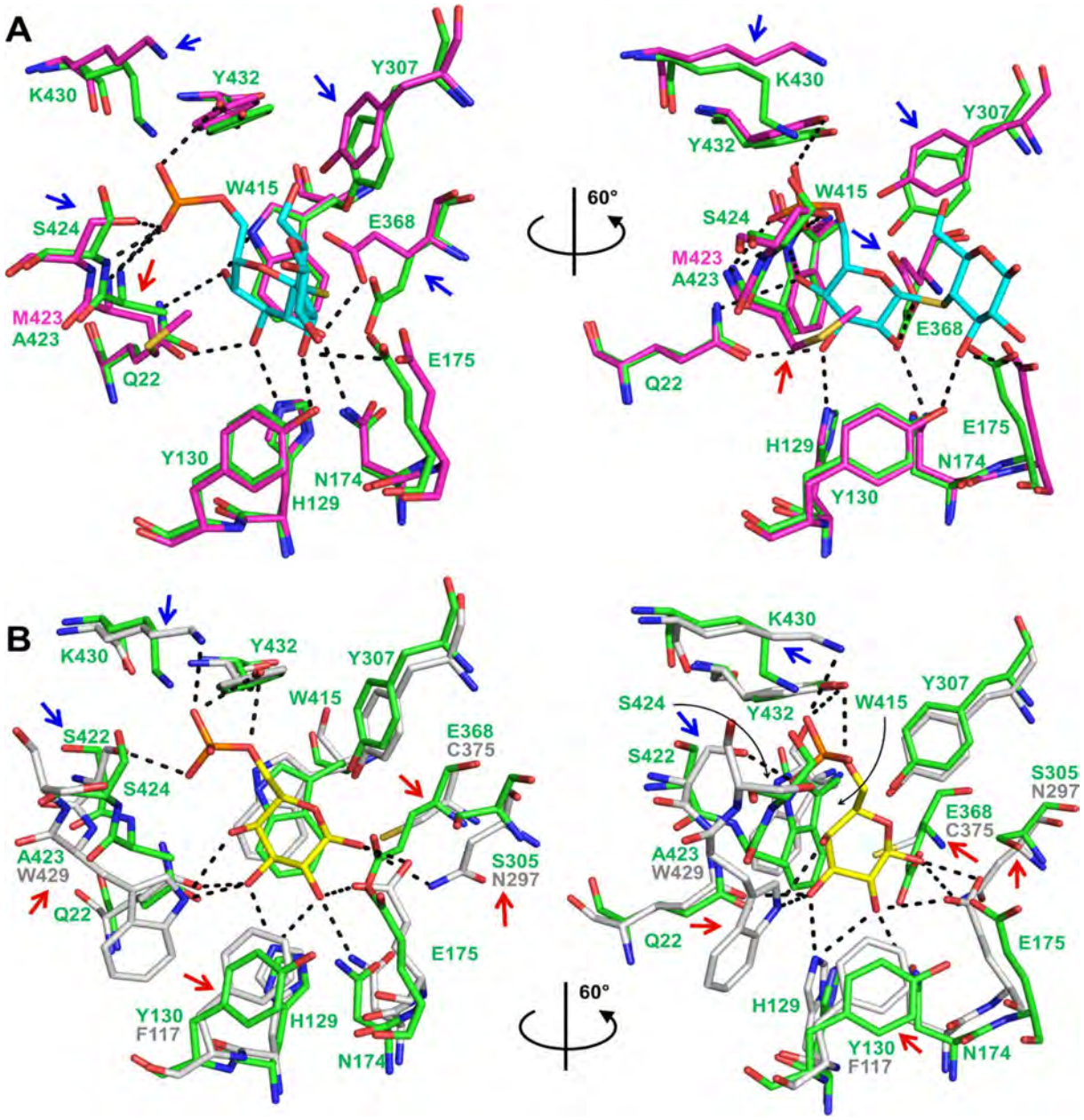


Figure 2.6. *Details of BlBglH ligand-binding site (PDB ID 6WGD – green residues) in comparison with: (A) BglA-2 (PDB ID 4IPN – magenta residues) complexed with thiocellobiose-6P (cyan). The structures have the same amino acids in each position, with exception of A423 from BlBglH that superposes with M423 from BglA-2 (red arrows). Discrepancies in side chain position are indicated with blue arrows; (B) PGALaseE375C (PDB ID 4PBG – grey residues) complexed with 6Pgal (yellow). Most of the amino acids from the binding site is conserved and have the same positioning with some differences likely caused by the presence of the ligand (blue arrows). The significant differences are indicated with red arrows. Reproduced with permission from Veldman et al. 2020 [103].*

2.3.4 Conserved motifs mapped to enzyme structures

To link the GH1 sequence motifs to structural function, the motifs were mapped to a respective crystallographic enzyme structure from each GH1 activity (Figure 2.7). Only the motifs that were present in 1 or 2 (but not all 3) of the GH1 activities were mapped. The next requirement was that the motifs had to be present in most of the sequences in their activity.

2.3.4.1 6Pβ-galactosidase activity

Motif 39 is totally conserved in the 6Pβ-galactosidase activity, and absent in the other activities. The motif consists of a beta-hairpin loop that covers the front of the enzyme and is known to control access to the active site in the 6Pβ-galactosidase activity. When this beta-hairpin loop blocks the opening to the active site, the substrate and the glycon product cannot pass through – only the aglycon product can be released. Motif 44, unique to the 6Pβ-galactosidase activity, forms part of this beta-hairpin loop also.

Motif 42 is located nearer to the N-terminal of the sequence and forms an α-helix; the residues in this motif were checked for possible significance. Residue Trp34 is totally conserved in the 6Pβ-galactosidase and β-glucosidase activities but is absent in the 6Pβ-glucosidase activity (Supplementary Figure 2.3). With DiscoveryStudio, the Trp34 interactions were analysed using crystallographic structure PDB IDs 1PBG (ligand-free) and 4PBG (ligand-bound) from *Lactococcus lactis*; this is the only 6Pβ-galactosidase enzyme with known PDB structures. It was seen that in the ligand-free form, Trp34 makes four different pi-pi stacked interactions with residue Trp429 simultaneously; however, in the ligand-bound form no interactions between these two residues were seen. Residue Trp429, and its sequence position in the L8a loop, is known to be important for substrate specificity within GH1 enzymes [175,183,199,201,204]. Trp34 in motif 42 may therefore be responsible for keeping residue Trp429 in place until it binds to a substrate. Furthermore,

ligand or not, Trp34 interacts with different protein regions (Tyr18, Phe117, Pro175), keeping the area surrounding the active site in place.

Motif 63 contains residues Asp231 and Ala234 that are totally conserved only within the 6Pβ-galactosidase activity; these residues bind to motif 44 (Met304 and Ala306), presumably providing the very long loop (L6 loop) with some support and could even be involved in the dynamics of the beta-hairpin loop that controls access to the active site. Motif 58 most likely is also involved in supporting loop L6, as Ile266 (totally conserved in 6Pβ-galactosidase) forms interactions with the loop (Val331) and with motif 63 (Leu237).

2.3.4.2 6Pβ-glucosidase activity

There were two motifs in the 6Pβ-glucosidase activity that were specific to this activity *and* were totally conserved in this activity – they were deemed highly significant for these reasons. On the enzyme structure, these motifs form the L8a and L8b loops (motifs 29 and 34, respectively). From Figure 2.7 A it is seen that the secondary structure of the 6Pβ-glucosidase L8 loop differs from the other activities. The L8 loop has known importance as it is home to residue Ala423 (*Bl*BglH numbering) – this particular residue position is considered to have a role in substrate specificity in 6Pβ-glucosidase enzymes [175], where Ala423 sterically clashes with galacto-configured ligands thereby ensuring the binding of only gluco-configured ligands. We go into detail about the L8 loop and its importance in section 2.3.5 .

Motif 31 consists of the C-terminal portion of loop L6. The conserved residues in motif 31 were analysed for possible significance. It seems the conserved residues are responsible for binding to other parts of the same loop, perhaps for stability and 3D positional reasons. The other activities do not have a motif in the same structural region as motif 31. However, the residues in the loop also bind to each other just like the 6Pβ-glucosidase activity. The residues within the motif are just slightly more conserved within this particular activity. The

same can be said of motifs 33 and 40. Regardless of activity, it seems these regions are also responsible for maintaining general tertiary structure by linking secondary structural elements. The important GH1 conserved tryptophan residue, Trp342 (*Bl*BglH numbering), that acts as a main hydrophobic platform that forms stacking interactions with the +1 sugar ring, is only three residues down the sequence from motif 31.

2.3.4.3 β-glucosidase activity

Although not conserved in all β-glucosidase enzymes, motif 21 is conserved within the first group of 22 β-glucosidases and is not present in the other activities. This motif forms an α-helix and is in the same protein region as motif 42 from the 6Pβ-galactosidase activity. In addition, the motif 21 region makes many connections with the surrounding regions, linking secondary structural elements. The remaining motifs 28, 32, and 37 do not seem to have any large significance. The conserved residues in these motifs bind to surrounding areas and probably maintain the tertiary structure of the enzyme.

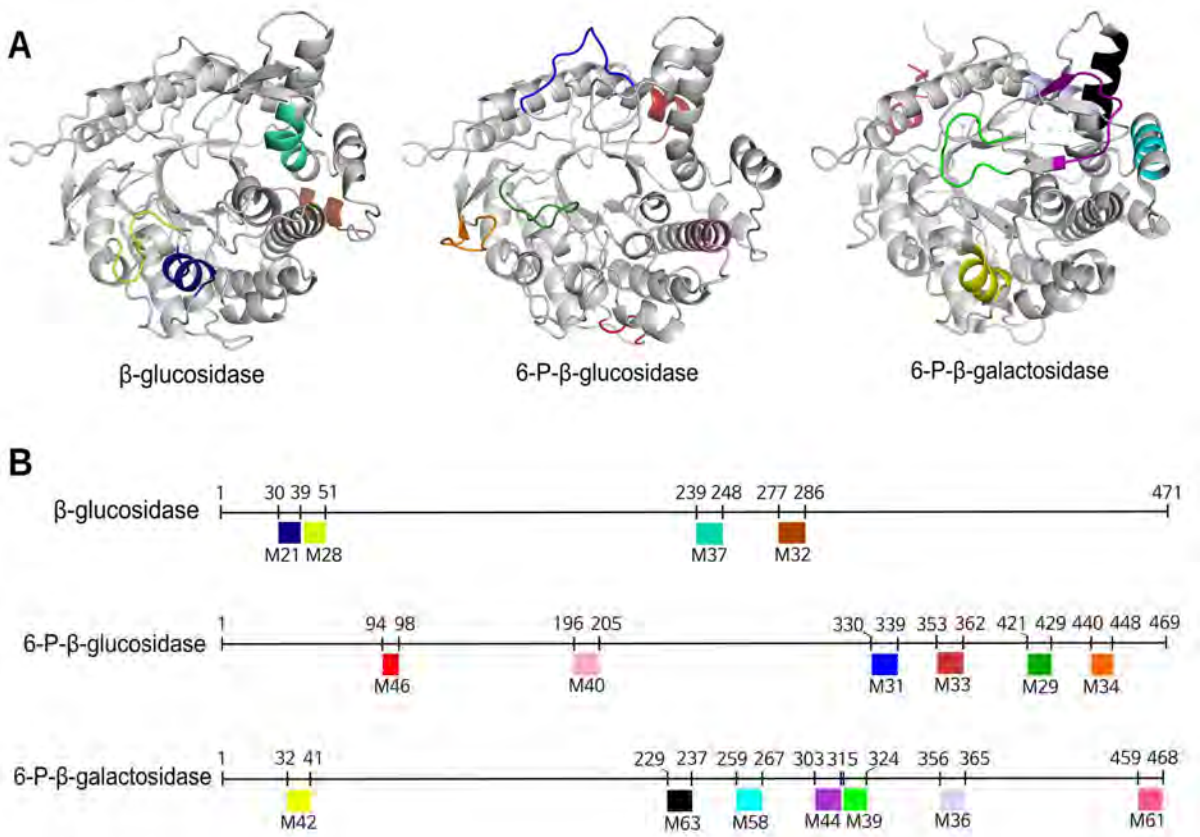


Figure 2.7. *(A) Motifs mapped to respective crystal structures: β-glucosidase – PDB ID 5DT7, 6Pβ-glucosidase – BlBglH (PDB ID 6WGD) and 6Pβ-galactosidase – PDB ID 4PBG. (B) Linear representation of the enzyme sequences that show the positions of the motifs in the sequence – the residue numbering is based on each individual crystallographic structure. Reproduced with permission from Veldman et al. 2020 [103].*

2.3.5 Analysis of loop L8

Motifs 29 and 34 are thought to be important because they are unique to the 6Pβ-glucosidase activity, and they are totally conserved within the activity. Motifs 29 and 34 were discovered in the L8 loop sequence region and make up loop L8a and loop L8b, respectively. There is evidence that a specific residue position in loop L8a (Ala423; *Bl*BglH numbering) has a role to play in substrate specificity 6Pβ-glycosidases [175]. The one and only difference between galacto- and gluco-configured substrates is their O4 hydroxyl group arrangement; Michalska and coauthors [175] have explained that the closer O4 hydroxyl group of the galacto epimer would clash with the 6Pβ-glucosidase residue in loop

L8a (Ala423 - *Bl*BglH numbering) and that this would prevent binding and catalysis (Figure 2.8).

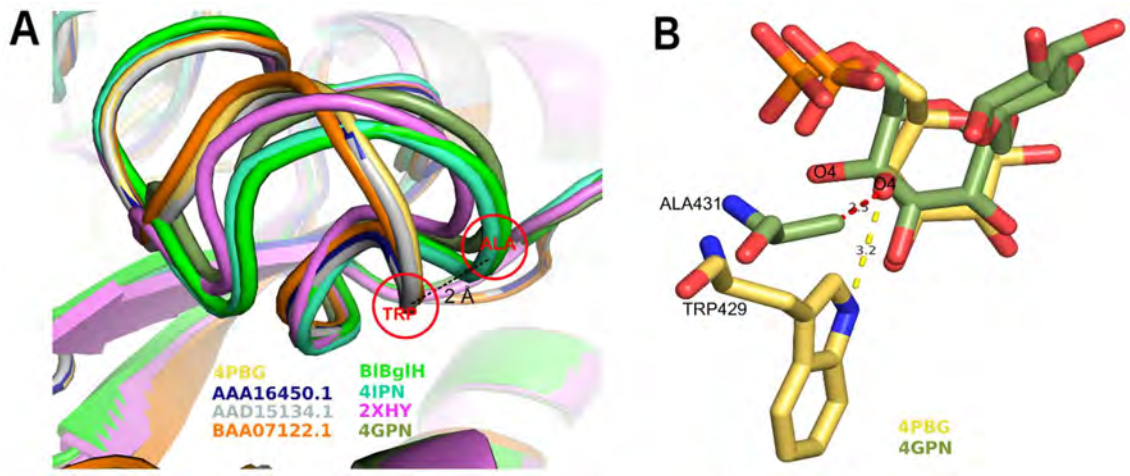


Figure 2.8. *Difference in loop L8a structure between the 6Pβ-galactosidase and 6Pβ-glucosidase activities, contributing to steric clash between 6Pβ-glucosidase enzymes and galacto-configured substrates. (A) Superposition of PDB ID 4PBG and three homology models (6PB-galactosidases) with BlBglH, 4IPN, 2XHY, and 4GPN structures (6Pβ-glucosidases). Red circles show the different spatial positions (~2 Å apart) of the specificity-inducing Ala423 residue-position (BlBglH numbering). (B) 4GPN and 4PBG are co-crystallised with 6PB-cellobiose and 6PB-galactose, respectively. The 6Pβ-glucosidase residue would have a steric clash (red dashes) with the OH4 of 6PB-galactose, which has axial position. While tryptophans from 6PB-galactosidases have hydrogen bond between OH4 of 6PB-galactose (black dashes).*

Three different 6Pβ-galactosidase enzymes were modelled, as only one known 6Pβ-galactosidase enzyme has crystallographic structures (PDB ID 4PBG). We note that the L8a loop backbone position at Ala423 differs by a distance of 2 Å between 6Pβ-galactosidases and 6Pβ-glucosidases (Figure 2.8 A) – this difference in secondary structure of the L8a loop (motif 29) very likely contributes to the clash. To find the reason behind the structural difference in the L8a loop between the 6Pβ-galactosidase and 6Pβ-glucosidase activities, crystallographic structures were used to compare the sequence positions of residue-residue interactions in the L8a loop. We checked for L8a loop residue interactions where *all* enzymes within each respective activity: (1) had the same

interactions with the same sequence positions, and (2) did not share any of these interactions with any one enzyme from the opposite activity (Table 2.1 & Figure 2.9).

There are many more residue-residue interactions that are unique to the 6Pβ-glucosidase activity. The 6Pβ-glucosidase activity has five unique interactions whereas the 6Pβ-galactosidase activity has only one. Contributing to the additional interactions in the 6Pβ-glucosidase activity is the longer L8b loop (motif 34; blue block in Figure 2.9 A; blue residues in Figure 2.9 C) and a single residue insert Glu427. All the differing loop-residue interactions between the two activities most likely cause the differing loop structure and spatial positioning of residue Ala423 (red block in Figure 2.9 A; red residue in Figure 2.9 C). Furthermore, the extra interactions between the L8a and L8b loops in the 6Pβ-glucosidase activity likely give rise to additional rigidity of the L8a loop, which may inhibit the movement of Ala423, and guarantee the clash with galacto-configured ligands.

Table 2.1. *Differing L8a loop residue interactions (based on sequence position), between the 6Pβ-galactosidase and 6Pβ-glucosidase activities.*

Same residue position (6Pβ-galactosidases/6Pβ-glucosidases)	**6Pβ-galactosidases**	**6Pβ-glucosidases**
Phe427 / Val421	Tyr433 (pi-pi stacked interaction)	Leu57 (Alkyl interaction)
Ser428 / Ser422		Glu427 (H-bond)
Glu427 (Insert In 6PGLU)		Ser422 (H-bond)
Tyr433 / Met428	Phe427 (pi-pi stacked interaction)	Asn441 (H-bond)
Phe444 / Arg439		Arg431 (H-bond) & Gly443 (H-bond)

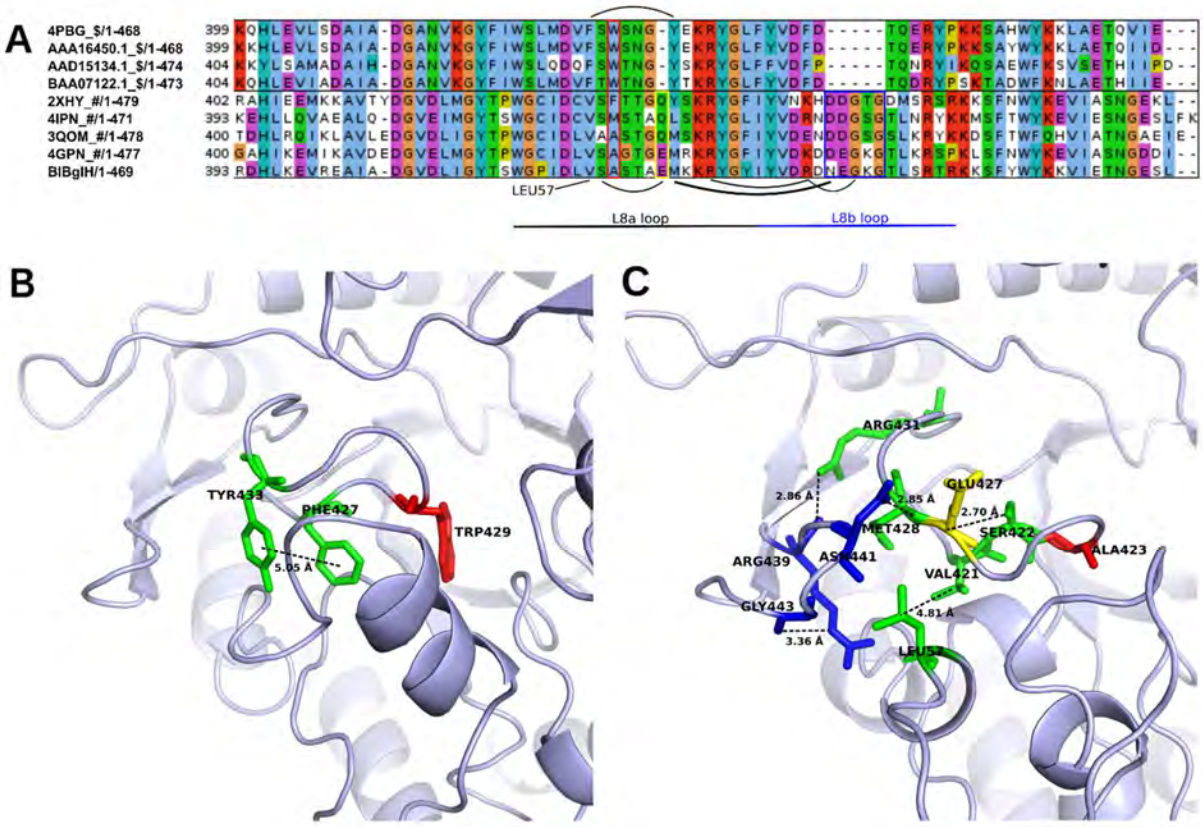


Figure 2.9. *Differing residue-residue interactions (based on sequence position) in the L8a loop between the 6Pβ-galactosidase and 6Pβ-glucosidase activities, contributing to differing loop 3D structure. (A) MSA of sequences from the 6Pβ-galactosidase (top 4 sequences) and 6Pβ-glucosidase (bottom 5 sequences) activities. Curved lines show interactions between residues. (B) 3D structure of L8a loop interactions of 6Pβ-galactosidase activity (4PBG numbering; PDB ID 4PBG). (C) 3D structure of L8a loop interactions of 6Pβ-glucosidase activity (BlBglH numbering; PDB ID 6WGD). The 6Pβ-glucosidase L8a loop has far more unique interactions compared to the 6Pβ-galactosidase L8a loop. Red coloured residue = Trp429/Ala423 (steric clash residue). Yellow coloured residue = 6Pβ-glucosidase activity Glu427 insert. Blue coloured residues = L8b loop residues (L8b is longer in 6Pβ-glucosidase activity). Reproduced with permission from Veldman et al. 2020 [103].*

2.3.6 *In silico* docking

AutoDock Vina-Carb was utilised for blind docking. The test ligands PNP6Pgal and PNP6Pglc were docked into the active site of the *Bl*BglH crystallographic structure (chain B). First though, prior to the docking and MD runs, the acid/base catalytic glutamate (Glu175) was protonated (Glh175) in agreement with the GH1 enzyme retention of configuration mechanism of catalysis [146,432]. After the docking, it was found that the

orientation of the PNP6Pgal ligand was incorrect and was the wrong way around (Figure 2.10 A). Despite this, there was a difference of only 0.3 kcal/mol binding energy between the ligands: -7.5 and -7.8 kcal/mol for PNP6Pgal and PNP6Pglc, respectively. The protein-ligand residue interactions are shown in Figure 2.10 B and Figure 2.10 C. Taken from Table 1.1 in Chapter 1, the important active site residues for *Bl*BglH (putative 6Pβ-glucosidase activity) should be Gln22, His129, Tyr130, Asn174, Glu175, Tyr307, Trp342, Glu368, Trp415, Ser422, Ala423, Ser424, Lys430 and Tyr432. The PNP6Pgal and PNP6Pglc ligands interact with 12 and 13 of the 14 active site residues, respectively. His129 and Tyr130 is missing from the PNP6Pgal interactions, and Lys430 is the only residue missing from PNP6Pglc. Although PNP6Pgal displayed 12 active site residue interactions, the binding of the three different ligand groups (phosphate, glycon and aglycon) to the protein subsites were mismatched. Conventionally, the two catalytic glutamates (Glu175 and Glu368) should interact with the glycon group [175,201] but these catalytic glutamates interact with the PNP6Pgal phosphate group. In fact, the majority of the other active site residues have mismatched PNP6Pgal binding site interactions suggesting the unsuitability of the ligand for the protein. On the other hand, PNP6Pglc shows favourable active site interactions as well as a correct orientation (Figure 2.10 C). These findings provide evidence that enzyme *Bl*BglH prefers gluco-configured substrates.

Due to the docked PNP6Pgal ligand having an incorrect orientation, we manually built PNP6Pgal into the active site to have the correct pose. This was done by transforming the docked PNP6Pglc ligand (having the correct pose) by altering the configuration of the O4 hydroxyl group. This new pose of PNP6Pgal will be named PNP6Pgal-pose2. Figure 2.10 D shows the interactions of PNP6Pgal-pose2 with the enzyme active site. The His129 residue

interaction is lost when PNP6Pglc is transformed to PNP6Pgal-pose2. To determine the

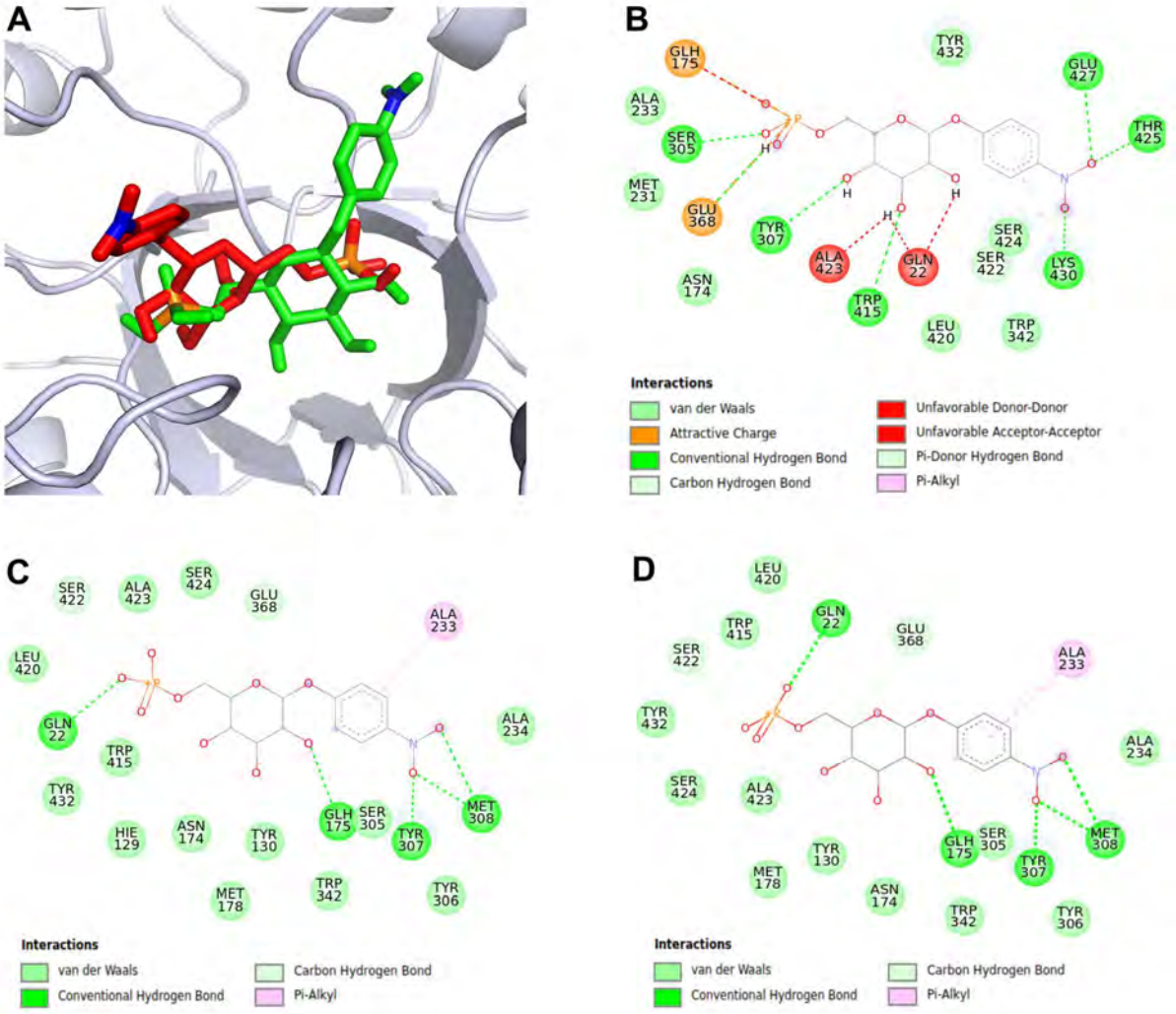


Figure 2.10. *BlBglH protein-ligand docking. (A) 3D representation of the PNP6Pgal (red) and PNP6Pglc (green) ligand orientations within the enzyme active site. (B) PNP6Pgal interactions, with binding energy of -7.5 kcal/mol. PNP6Pgal is in incorrect orientation within the active site. (C) PNP6Pglc interactions, with binding energy of -7.8 kcal/mol. (D) PNP6Pgal-pose2 interactions, with binding energy of -7.6 kcal/mol. The PNP6Pgal ligand was transformed from the docked PNP6Pglc ligand by changing the configuration of the O4 hydroxyl group from equatorial to axial. Adapted from Veldman et al. 2020 [103].*

PNP6Pgal-pose2 binding score, the “vina-carb –score_only” command was utilised. This score was -7.6 kcal/mol which is a marginal improvement compared to the original docked PNP6Pgal (-7.5 kcal/mol), but not as good as PNP6Pglc (-7.8 kcal/mol). Considering the superior binding score and orientation of the PNP6Pgal-pose2 ligand compared to the

docked PNP6Pgal ligand, the PNP6Pgal-pose2 complex was simulated using molecular dynamics in addition to the PNP6Pgal and PNP6Pglc complexes.

2.3.7 Molecular dynamics

Enzyme *Bl*BglH alone, and in complex with the docked and transformed ligands, was simulated for 400 ns using MD. This was done to analyse the dynamics of *Bl*BglH, and to provide evidence of the activity and substrate specificity of *Bl*BglH. PNP6Pglc stayed in the active site of the enzyme throughout the entire 400 ns of MD simulation. In contrast, PNP6Pgal was unstable and at 77 ns it departed the active site and even exited the enzyme at 287 ns. The PNP6Pgal-pose2 ligand, however, was found to be in the enzyme at 400 ns, but only the phosphate group was still bound to the active site. Hereafter, the results from the PNP6Pgal-pose2 ligand and its complex were compared to the PNP6Pglc-complex because of the PNP6Pgal exit from enzyme *Bl*BglH.

2.3.7.1 Trajectory analysis

RMSD, R_g and RMSF were used to analyse the MD simulations from the PNP6Pgal-pose2 and PNP6Pglc complexes (Figure 2.11). Protein RMSD gives an indication of conformational variation and overall stability throughout the MD simulations. Figure 2.11 A shows the protein RMSD of enzyme *Bl*BglH alone (apo) and complexed with the two ligands. It is seen that *Bl*BglH complexed with PNP6Pglc had higher stability compared to both the apo protein and the protein complexed with PNP6Pgal-pose2 – this is demonstrated by the smaller and more consistent RMSD values. Ligand binding did not cause any major protein conformational change – we know this because there was no significant change in protein RMSD throughout the simulation, and no change in conformation was observed with visual inspection using the program VMD. This observation is in line with a previous study [178] where it was established that the binding of ligands within GH1 enzymes have a minimal effect on conformational change. In the

study, 16 GH1 crystallographic structures were very tightly superimposed, despite some structures having occupied active sites and some not. The *ligand* RMSD (Figure 2.11 B) of PNP6Pglc was consistent and did not have values higher than 0.3 nm, meaning that PNP6Pglc was stable throughout the MD simulation. On the other hand, the PNP6Pgal-pose2 RMSD values were not consistent – at 25 ns and again at 275 ns, the ligand changed its position within the enzyme, gradually increasing its distance from the initial position. The R_g was then used to measure the compactness of the enzyme during MD. All of the MD runs had an extremely stable R_g during the whole 400 ns, never deviating more than 0.05 nm. However, the *Bl*BglH enzyme complexed with PNP6Pglc was moderately more compact in general when compared to the apo protein and PNP6Pgal-pose2 complex (Figure 2.11 C). RMSF measures the fluctuation of each individual residue during the simulations. The residues of enzyme *Bl*BglH complexed with PNP6Pglc showed less fluctuation overall compared to the apo protein and PNP6Pgal-pose2 complex (Figure 2.11 D). The difference in fluctuation is more prominent in the protein loop regions, however. Overall, the smaller RMSD, R_g and RMSF values of the PNP6Pglc complex, as well as the greater stability, are most likely due to the sustained favourable binding of PNP6Pglc within the enzyme and is evidence of the suitability of the ligand for the enzyme. Figure 2.11 E clearly shows the *Bl*BglH structural regions with significant differences in RMSF values between the ligand-free and ligand-bound proteins, or between the two different complexes. Most of these regions are the loops that surround the active site. The right flank of the enzyme (L5 & α5 – cyan colour) shows slightly less fluctuation in the PNP6Pglc complex, this could mean that this side of the enzyme opens up very slightly when not binding stably to a ligand.

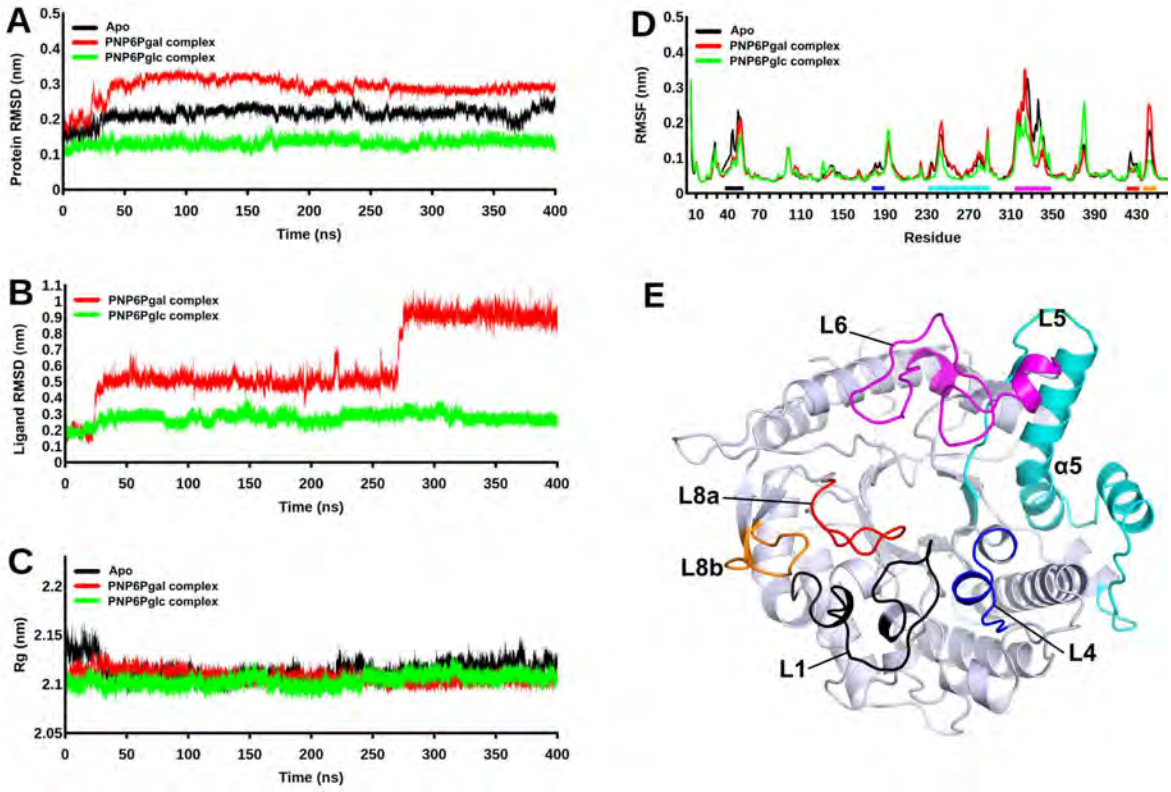


Figure 2.11. *BlBglH MD trajectory analysis. (A) Protein backbone RMSD after least square fitting to protein backbone, (B) ligand RMSD after least square fitting to protein backbone, (C) protein radius of gyration, (D) protein residue RMSF, and (E) colour coded sequence regions that are mapped to the BlBglH structure where significant differences in RMSF values are seen between either the ligand-free and ligand-bound proteins, or the two different complexes. Adapted from Veldman et al. 2020 [103].*

In Figure 2.12, snapshots of the MD simulations at 0 ns (grey), 15 ns (cyan), 250 ns (magenta) and 400 ns (dark blue) are superimposed for the PNP6Pgal-pose2 complex (Figure 2.12 A) and PNP6Pglc complex (Figure 2.12 B). The times of the snapshots were chosen over periods of PNP6Pgal-pose2 stability and show the various positions that the PNP6Pgal-pose2 ligand occupied during the simulation. As time progresses, the PNP6Pgal-pose2 ligand shifts, and by the end of the simulation only the phosphate group remained in the active site (Figure 2.12 A). The loops around the active site are stabilised when PNP6Pglc is bound, shown with orange arrows in Figure 2.12. The PNP6Pglc-complex had lower RMSF values in these loop regions also (Figure 2.11 D). Regarding the PNP6Pgal-pose2 complex, loops L1, L6a, L8a and L8b move further from their initial

positions over time. This is especially noticeable in loop L8b – the loop discussed earlier to be important for the structure of loop L8a which contains Ala423 (*Bl*BglH numbering), the residue believed to clash with a galacto-configured O4 hydroxyl group. The L6 loop is thought to obstruct the entrance to the active site where the substrate and the glycon product cannot pass through – only the aglycon product can be released [183]. In the PNP6Pglc complex, this loop covered the active site entrance for the duration of the MD simulation, whereas the loop was much more scattered in the PNP6Pgal-pose2 complex. The results show that the favourable binding of substrate stabilises the loops that surround the *Bl*BglH active site. The enzyme without a bound substrate would need to open itself up slightly to accommodate a potential substrate and would be stabilised when bound to a suitable substrate.

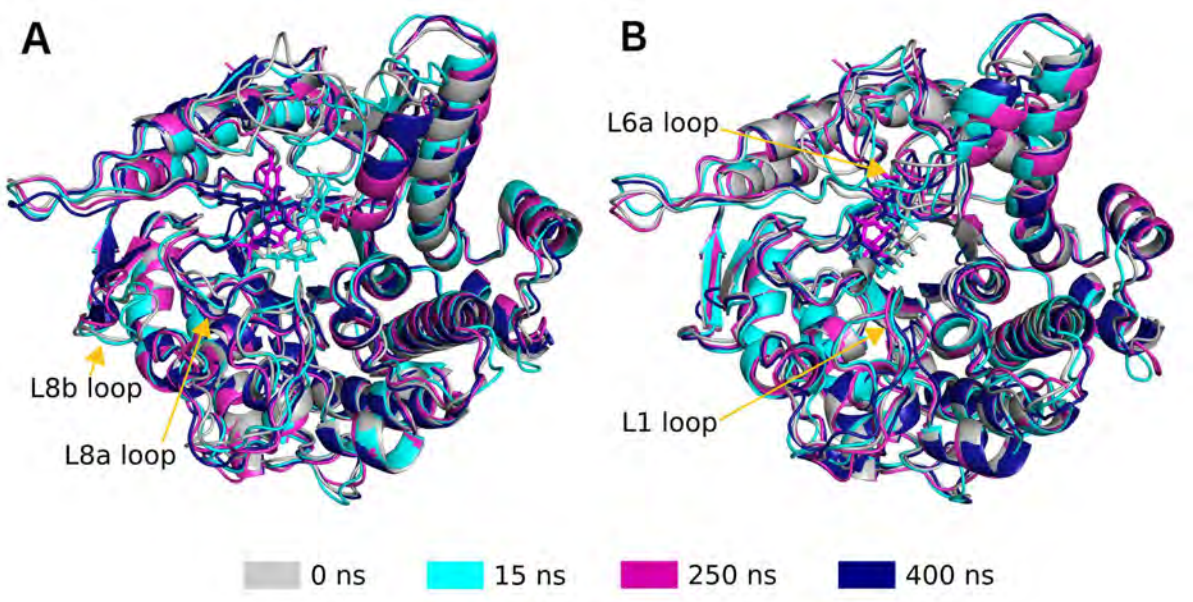


Figure 2.12. *Static snapshots of the (A) PNP6Pgal-pose2 and (B) PNP6Pglc complexes during the MD simulations at 0 ns (grey), 15 ns (cyan), 250 ns (magenta) and 400 ns (dark blue) are superimposed. Orange arrows show the location of loops that are all stabilised by PNP6Pglc* binding and *destabilised by the mobile PNP6Pgal-pose2. Reproduced with permission from Veldman et al. 2020 [103].*

2.3.7.2 PNP6Pglc ligand interactions

The interactions between *Bl*BglH and the PNP6Pglc ligand at 400 ns of MD simulation are shown in Figure 2.13. The ligand interacts with 13 *Bl*BglH active site residues and is in a good orientation within the active site. From the start until the end of the simulation, PNP6Pglc had very minimal change in its position/orientation and interactions (Figure 2.13 C). The only missing active site residue was Lys430, which would normally form a charge-charge attraction with the negatively-charged ligand-phosphate. Since charge-charge interactions can be strong even at 5-10 Å [442], the distance was measured between the positively-charged Lys430 nitrogen atom and the two negatively-charged oxygen atoms on the ligand-phosphate (Figure 2.14). During the majority of the MD simulation this distance was less than 10 Å, meaning that Lys430 was most likely still attracting the ligand. Hydrogen bonds, at 400 ns, are formed between the ligand and four different enzyme residues, namely Trp342, Tyr307, Glu368 and Ser424. In addition to its hydrogen bond contributions, Tyr307 forms a pi-pi stacked interaction with the aglycon ring. The Ser424 residue position interaction seen here is also seen in all three 6Pβ-glucosidase activity enzymes from Table 1.1, but is absent in the other activities, therefore the Ser424 residue position could be important for the substrate specificity of 6Pβ-glucosidase activity enzymes.

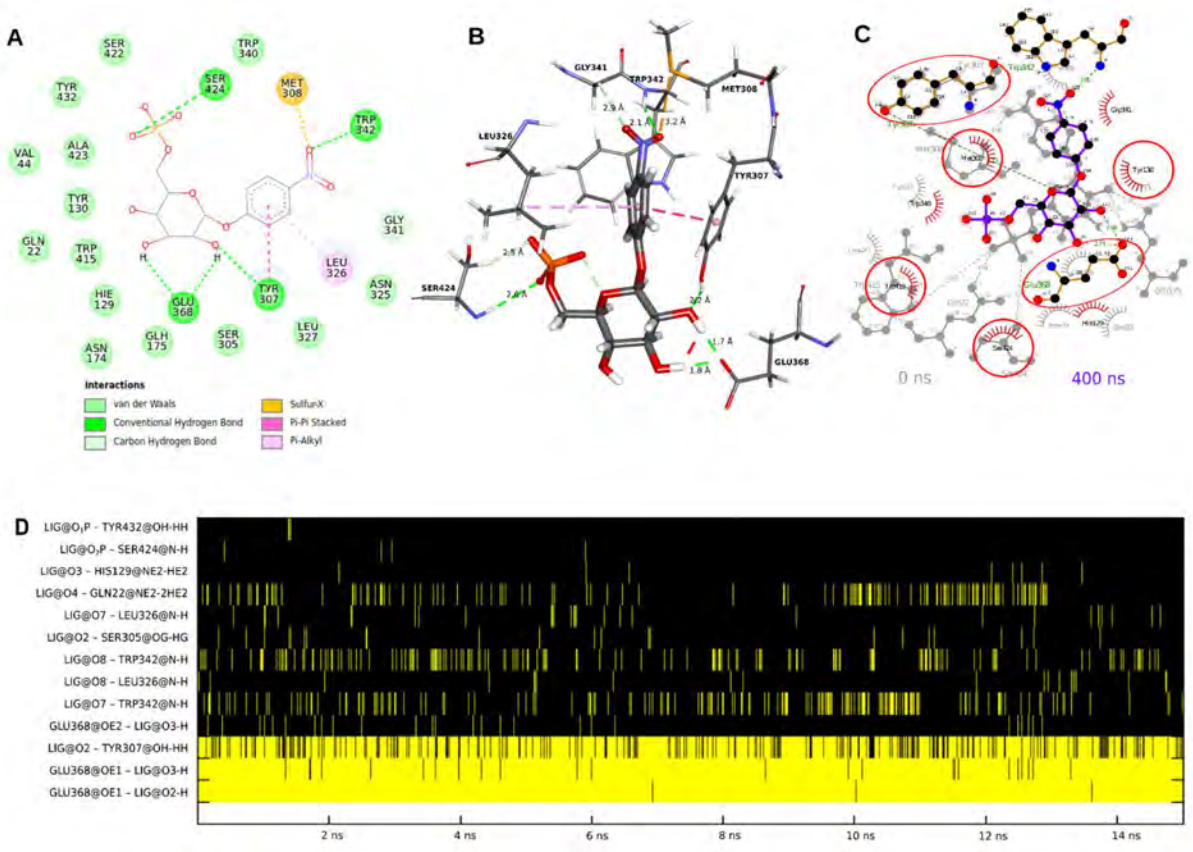


Figure 2.13. *BlBglH-PNP6Pglc interactions (A) 2D representation of PNP6Pglc protein-ligand interactions at 400 ns of MD simulation. (B) 3D representation of PNP6Pglc protein-ligand interactions at 400 ns of MD simulation. (C) Overlay of PNP6Pglc protein-ligand interactions at 0 ns (grey) and 400 ns (colour). (D) Hydrogen bonding of the PNP6Pglc ligand-protein complex during the last 15 ns of the MD simulation, using standard/default geometric criteria/definitions. Hydrogen bonds are shown in yellow. Adapted from Veldman et al. 2020 [103].*

For more detail, we analysed the hydrogen bonding in the *Bl*BglH-PNP6Pglc complex during the final 15 ns of MD simulation (Figure 2.13 D). For the vast majority of the time, the catalytic Glu368 residue forms hydrogen bonds with two different positions on the ligand glycon ring; namely the hydrogen atoms of the ligand O2 and O3 hydroxyls. There is also consistent hydrogen bonding between Tyr307 and the ligand glycon O2 atom. Residues Trp342, Leu326 and Gln22 also form many hydrogen bonds in more than one position on the ligand. Residues Trp342 and Leu326 are located in the large L6 loop that may control access to the entrance of the substrate-binding pocket. The L6 loop could

then be responsible for ligand binding, in addition to having a role in controlling which ligands enter and exit the enzyme. The side chain of Trp342 is a large contributor of GH1 interactions with the aglycon group of bound substrates. By experimentally mutating the Trp342 residue and then performing catalysis studies, this residue has been shown to be significant in cellobiose-6P hydrolysis [201]. Residue Trp342 is conserved in 58 of the 59 sequences used in this study (Supplementary Figure 2.3), an indication of its significance in bacterial GH1 enzymes.

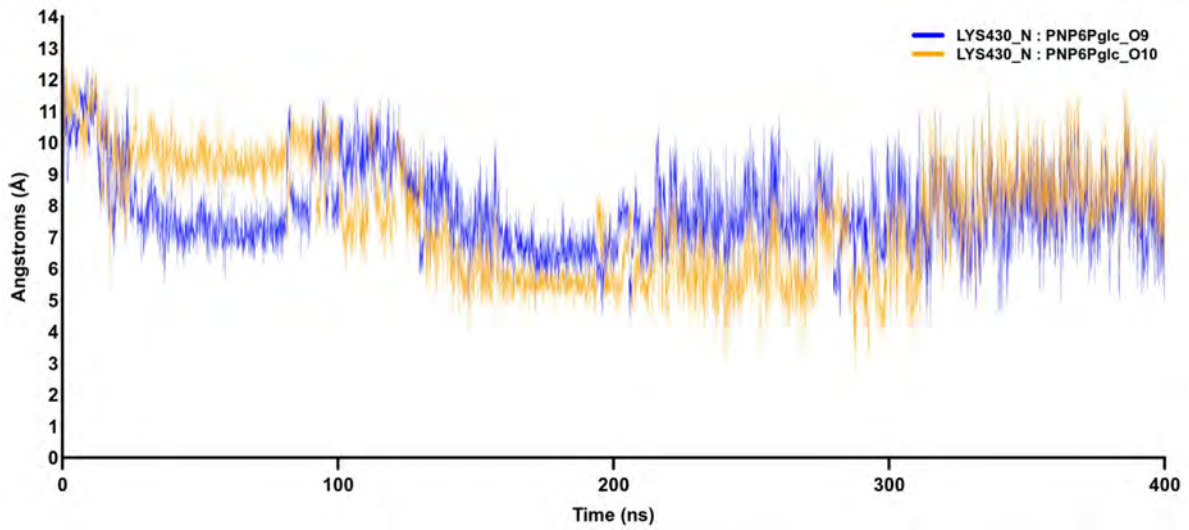


Figure 2.14. *Distance measured between the positively-charged Lys430 nitrogen atom and the negatively-charged ligand-phosphate.*

2.3.7.3 Binding free energy

The measurement of binding strength of protein-ligand or protein-protein interactions can be performed with binding free energy calculations [409,410]. Although not as accurate as more computationally intensive methods, the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) method has been useful to computational and experimental researchers. The MM/PBSA method is a quick but reliable way to explore protein-ligand binding interactions [408]. Despite being used mostly for drug discovery research, the

MM/PBSA method was used here to provide an estimation of ligand binding affinity of the two different ligands (Table 2.2).

Table 2.2: *Contribution of individual energies towards the total binding free energy, resulting from the MM/PBSA analyses of the two complexes. All values are in kJ/mol. The final 15 ns of the MD simulations were used (385-400 ns).*

Complex	Van der Waal's	Electrostatic	Polar solvation	Solvent Accessible Surface Area (apolar)	Total binding energy
PNP6Pgal-pose2 – *Bl*BglH	-79.92 ± 0.44	339.44 ± 2.96	199.94 ± 3.05	-12.08 ± 0.04	**447.12 ± 1.12**
PNP6Pglc – *Bl*BglH	-159.39 ± 0.36	-108.77 ± 0.26	191.22 ± 0.26	-18.03 ± 0.02	**-94.98 ± 0.30**

The final 15 ns (385-400 ns) of the MD simulations were used during the energy calculations. The difference in the total binding energy between the complexes was large: the PNP6Pgal-pose2 complex had a poor binding energy total of 447.12 kJ/mol compared to the PNP6Pglc complex which had a binding energy total of -94.98 kJ/mol. The disparity in the total energies is mostly caused by a big difference in electrostatic energy. The electrostatic component of binding free energy can be greatly affected by a modest fluctuation in charge distribution [442,443], therefore the large difference between the complexes in electrostatic binding energy could be explained by the like-charge interaction between the negatively-charged ligand phosphate group and the negatively-charged catalytic residue Glu368 in the PNP6Pgal-pose2 complex (Figure 2.15). Some types of hydrogen bonds can contribute to electrostatic binding energy [443]. At any one time, the PNP6Pglc complex had around two more hydrogen bonds on average when compared to the PNP6Pgal-pose2 complex (Figure 2.15 B & Figure 2.15 C). Thus, the reduced number of hydrogen bonds in the PNP6Pgal-pose2 complex could also account for the disparity in the electrostatic binding energy. The binding free energy results further support the putative 6Pβ-glucosidase activity of enzyme *Bl*BglH.

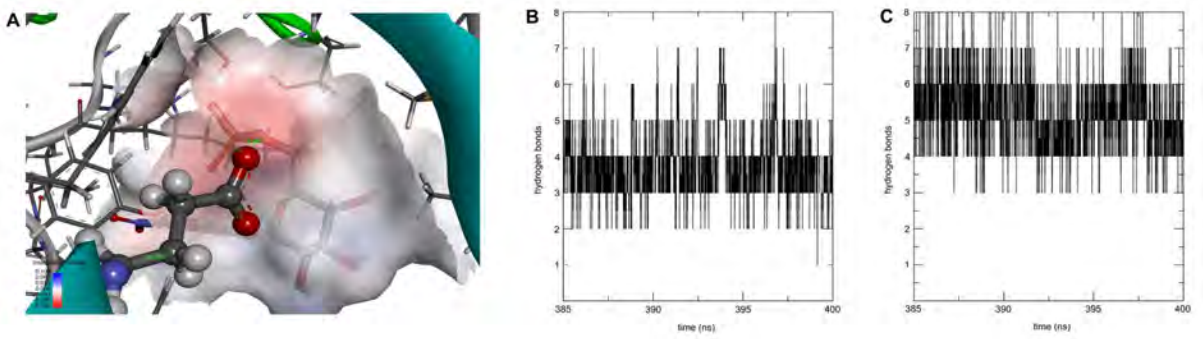


Figure 2.15. *Contributors to the large difference in electrostatic binding energy between the two BlBglH complexes during the final 15 ns of the MD simulation. (A) Like-charge interaction between the negatively-charged ligand phosphate group and the negatively-charged catalytic protein residue Glu368. (B) Number of hydrogen bonds formed in the PNP6Pgal complex. (C) Number of hydrogen bonds formed in the PNP6Pglc complex.*

2.3.7.4 MD duplication

The structural difference between the two test ligands is small, namely just one hydroxyl group arrangement. However, the difference in the MD simulations and ensuing binding free energy results were striking. To enhance the plausibility of the results, the MD simulations and MM/PBSA calculations were duplicated (Supplementary Figure 2.5 and Supplementary Table 2.4, respectively). In the duplicates, just like before, the MD simulation of the PNP6Pglc complex had lower and more consistent RMSD, R_g and RMSF values compared to the PNP6Pgal-pose2 complex, as well as better ligand stability. Loops L1, L6a and L8a fluctuate less in the PNP6Pglc complex. The catalytic Glu368 residue once again forms hydrogen bonds with the PNP6Pglc ligand glycon group in two different locations. At the end of the simulation, PNP6Pglc was found to be in a good pose within the active site although the ligand phosphate group interacted with fewer active site residues compared to the original MD simulation. The MM/PBSA results from the duplicated MD's showed the PNP6Pgal-pose2 and PNP6Pglc complexes to have total binding energies of 430.89 kJ/mol and -74.38 kJ/mol, respectively: quite similar to the first calculations. The uniform results from two different sets of MD simulations proves the validity of the results overall.

2.4 Conclusion

The sequence analyses used a dataset of 59 bacterial GH1 enzymes and predicted enzyme *Bl*BglH to have 6Pβ-glucosidase activity. In fact, *Bl*BglH fell into the 6Pβ-glucosidase group in each separate sequence-based evaluation. It shared an average of 53% sequence identity with the 6Pβ-glucosidase group; 20% more than any other activity. During the phylogenetic analysis, *Bl*BglH clustered into the 6Pβ-glucosidase activity with strong bootstrap confidence. *Bl*BglH was also found to possess motifs that were seen only in the sequences of the 6Pβ-glucosidase activity. Despite the high structural similarity between GH1 members, it was found that the L8 loop secondary structure contributes to the substrate specificity between 6Pβ-glucosidases and 6Pβ-galactosidases. In contrast to PNP6Pgal, PNP6Pglc showed sustained favourable binding during MD, providing evidence of the 6Pβ-glucosidase activity of *Bl*BglH. Additionally, the favourable binding of PNP6Pglc stabilised the loops that surround the active site. Seeing as correspondent transport and utilisation systems are conserved, this study could influence other research and industrial applications using *B. licheniformis* as well as other bacteria. As far as we are aware, this is the earliest known investigation to simulate a 6Pβ-glycosidase GH1 enzyme using MD.

Chapter 3: Bioinformatics analysis and substrate specificity of enzymes *Bl*BglC and *Bl*BglB

3.1 Introduction

At the time of writing, 21 various GH1 enzymatic activities are classified in the CAZy database. Most often, bacteria utilise only one preferred carbon source [444], therefore most bacteria possess one specific activity. Bacteria usually act together in order to metabolize a cascade of substrates. The 6Pβ-galactosidase activity hydrolyses 6Pβ-galactosides (e.g., phosphorylated lactose) forming galactose-6-phosphate [199], whereas the 6Pβ-glucosidases activity hydrolyses 6Pβ-glucosides forming glucose-6-phosphate [175]. On the other hand enzymes can also be promiscuous, having broad specificity, which means they are able to utilise more than one substrate and catalyse more than one reaction [28]. In the CAZy database, only three GH1 enzymes have been characterised as having dual-phospho activity, and only one previous dual-phospho crystallographic structure exists: Gan1D from *G. stearothermophilus* [199].

Galacto- and gluco-configured substrates differ only in the position of their O4 hydroxyl group and are therefore extremely similar. As seen in the previous chapter, this slight difference has enough importance to cause very different outcomes in MD simulations where the gluco-configured substrate remained stable in the active site, but the galacto-configured substrate exited the enzyme completely [105]. However, dual-phospho activity enzymes are able to hydrolyse both galacto- and gluco-configured substrates. The Gln23 and Trp433 residues of the Gan1D enzyme structure (referred to above) are thought to be significant for ligand recognition and binding, and that the particular Gln23 and Trp433 hydrogen bonding enables the modulation of preference toward a galactose or glucose sugar.

This chapter seeks to provide insight into the function of two new crystallographic structures of GH1 enzymes from the bacterium *B. licheniformis* (*Bl*BglB & *Bl*BglC). They are analysed to provide evidence for their enzyme activity and to study their substrate specificity. Sequence analyses, MD simulation analysis and binding free energy calculations provide evidence for the dual-phospho activity of *Bl*BglC. In triplicate, test ligands PNP6Pgal and PNP6Pglc are used to establish the details of *Bl*BglC substrate specificity. Important details of the broad specificity of dual-phospho activity GH1 enzymes are revealed. In contrast, the *Bl*BglB enzyme is very unique and activity determination of the enzyme is elusive.

Most of the work in this chapter is reproduced from the publication below. All of the work is my own, except for enzyme 3D crystallographic structure determination and activity assays. All original writing and figure generation from this chapter is my own except for section 3.3.2 and Figure 3.3.

> **Wayde Veldman**, Marcelo Liberato, Valquiria Souza, Vitor Almeida, Sandro Marana, Özlem Tastan Bishop, and Igor Polikarpov. “Differences in Gluco and Galacto Substrate-Binding Interactions in a Dual 6Pβ-Glucosidase/6Pβ-Galactosidase Glycoside Hydrolase 1 Enzyme from *Bacillus licheniformis*”. *Journal of Chemical Information and Modeling*. 2021, 61, 9, 4554–4570. DOI: 10.1021/acs.jcim.1c00413.

3.2 Methodology

3.2.1 Enzyme crystallographic structure from collaborators

From our collaborators at the University of São Paulo we obtained two unpublished crystallographic enzyme structures in PDB format from the bacterium *B. licheniformis*. The GenBank accessions of the enzymes are AAU39345.1 (*Bl*BglC) and AAU43012.1 (*Bl*BglB). The structures were checked for rotamers, missing atoms, and missing residues. The structures did not have any ligand in their active sites and therefore enzyme function

and specificity were unclear. The *Bl*BglC enzyme crystallographic structure is published on the RCSB PDB webserver with PDB ID 7M1R. X-ray diffraction was used to determine the structures. The quality metrics for 7M1R are as follows: Resolution – 1.98 Å; R-Value Free – 0.188; R-Value Work – 0.155; R-Value Observed – 0.157; Clashscore – 2; Ramachandran outliers – 0; Sidechain outliers – 0.7%; RSRZ outliers – 0.8%. The quality metrics for the *Bl*BglB crystallographic structure are as follows: Resolution – 2.43 Å; R-Value Free – 0.251; R-Value Work – 0.225; Clashscore – 6; Ramachandran outliers – 0.

3.2.2 GH1 enzyme sequence data retrieval

The amino acid sequences of characterised bacterial GH1 enzymes were retrieved from the CAZy database. 35 β-glucosidase (EC 3.2.1.21), eight 6Pβ-galactosidase (EC 3.2.1.85), 20 6Pβ-glucosidase (EC 3.2.1.86), and three dual-phospho activity sequences were downloaded via the GenBank links in CAZy. The *Bl*BglC and *Bl*BglB sequences were then added to the retrieved characterised homolog GH1 enzyme sequences to make a total of 69 sequences in the final dataset (Supplementary Table 3.1).

3.2.3 Multiple sequence alignment

The PROMALS3D alignment webserver was used with the GH1 sequence dataset to produce an MSA (Supplementary Figure 3.1). It is possible to upload crystallographic structures to PROMALS3D so that the webserver includes the secondary structure information in order to improve the alignment. For this reason, if any of the sequences had structures, they were added to the webserver in addition to the full GenBank sequences, but only after they were checked for residue rotamers to prevent the duplication of residues in the sequences. Following alignment, the MSA was viewed, analysed and appropriately adjusted with Jalview (version 2) alignment viewer. The last step was the removal of the crystallographic structure sequences from the MSA.

3.2.4 Phylogenetic analysis

With the MSA and the MEGA v7.0.26 tool, phylogenetic trees were constructed. The evolutionary models with the three lowest Bayesian information criterion (BIC) scores were selected for phylogenetic tree construction and three different gap deletions (90, 95 and 100%) were used for each model. For each construction of a tree, 1000 bootstrap replicates and a strong branch swap filter were used. The top models for all three gap deletions were determined to be the Whelan and Goldman model with Gama distribution and Invariant sites (WAG+G+I), the Le and Gascuel model with Gama distribution (LG+G), and the Le and Gascuel model with Gama distribution and Invariant sites (LG+G+I). Nine maximum likelihood phylogenetic trees were constructed using the three models at the three different gap deletions. To select the optimal tree, each of the nine constructed phylogenetic trees were compared to their respective bootstrap consensus trees – overall branching pattern and branch support were taken into consideration. The most accurate phylogenetic tree was chosen to be the one using the WAG+G+I model with 90% gap deletion.

3.2.5 Motif analysis and structure mapping

The MEME v4.9.1 tool was utilised to search for sequence motifs within the enzyme sequences (Supplementary Table 3.2). A search for motifs with a size range of between five and ten residues was set, as this range was established to be optimal for both the identification of activity-specific motifs as well as shared motifs between different activities. To locate and exclude any intersecting motifs, the Motif Alignment Search Tool (MAST; part of the MEME suite) was used; this resulted in 80 significant motifs. By parsing the MEME log file with a Python script that was formerly written by Faya *et al*., a motif heatmap was generated showing the conservation of motifs that are shared between the GH1 enzyme sequences. Motifs that were found to be totally conserved within a particular

enzyme activity were mapped to an enzyme crystallographic structure from the same activity.

3.2.6 Activity assays

The enzyme activity assays were performed by our collaborators at the University of São Paulo. P-nitrophenyl β-fucopyranoside (PNPβfuc), p-nitrophenyl β-galactopyranoside (PNPβgal), p-nitrophenyl β-glucopyranoside (PNPβglc), p-nitrophenyl β-xyloside, p-nitrophenyl β-mannopyranoside, and the oligosaccharides cellobiose, cellotriose and cellotetraose were used. Regrettably, no phosphorylated PNP-monosaccharides nor phosphorylated cellooligosaccharides were commercially available.

3.2.7 Homology modelling

Two regions of the enzyme *Bl*BglB crystallographic structure were missing due to high flexibility – these regions were loops L4 and L6, and they had to be modelled in order to obtain a complete enzyme structure for use in *in silico* docking and MD. Three template structures were utilised, including the original *Bl*BglB structure, the *Bl*BglC structure (PDB ID 7M1R), and finally PDB ID 6WGD. Chain A of enzyme *Bl*BglB was modelled.

HHpred and PRotein Interactive MOdeling (PRIMO) webservers were used to identify suitable template structures for the models. 100 models per target enzyme were generated using MODELLER version 9.23. The top three models per target enzyme, ranked by normalised z-DOPE (Discrete Optimized Protein Energy) score, were evaluated further using PROCHECK, Qualitative Model Energy Analysis (QMEAN) and Verify3D webservers. According to the consensus from all three model quality evaluation tools, the best model from each target protein was selected. The templates were all of high quality, and a very high-quality model was produced (Supplementary Table 3.3).

3.2.8 *In silico* docking

To validate the docking procedure, the Autodock Vina-Carb version 1.0 docking program was used with the crystallographic structure of Gan1D from *G. stearothermophilus T-1* (PDB ID 5OKG, chain D) and its ligand cellobiose-6-phosphate. After removing cellobiose-6-phosphate from the protein, it was docked back into the protein using a blind docking approach. A grid box search space of dimension size of x = 75.0 Å, y = 75.0 Å, z = 75.0 Å was defined, the ligands were centred at x = -15.4, y = -34.8, z = -88.4, and a search exhaustiveness value of 300 was applied. Using the GROMACS RMSD command *gmx rms*, the RMSD between the crystallographic structure ligand and the same ligand docked into the crystallographic structure was 0.56 Å. Docking validation poses are conventionally regarded as successful if the RMSD is below 2 Å from the known ligand conformation [445–447].

Due to the absence of missing residues in important regions, chain A of the crystallographic structure 7M1R (*Bl*BglC) was used to perform docking. To evaluate enzyme activity, two ligands that represent either the 6Pβ-galactosidase or 6Pβ-glucosidase activity were docked into the active site of *Bl*BglC using blind docking. These ligands were p-nitrophenyl-beta-D-galactoside-6-phosphate (PNP6Pgal) and p-nitrophenyl-beta-D-glucoside-6-phosphate (PNP6Pglc), respectively. The two ligands have identical structure, with only one exception: the galacto epimer's O4 hydroxyl group has an axial position, whereas the gluco epimer's O4 hydroxyl group has an equatorial position (Supplementary Figure 2.2). Initial attempts at docking into the active site of the original conformation of the *Bl*BglC crystallographic structure failed – the suspected cause being the L6 loop that is thought to block the entrance to the active site [175,183]. To overcome this obstacle, a 200 ns apo MD simulation was run at an increased temperature of 315 K using 7M1R chain A in order to open up the L6 loop and expose the active site. The distance between the loop and the active site during the course of the simulation was

monitored using VMD software. The *Bl*BglC structure at the point where this distance was the greatest (115 ns; 6.5 Å) was used for the second docking attempt. On the other hand, chain A of enzyme *Bl*BglB was modelled, and the model directly used for docking. For both *Bl*BglC and *Bl*BglB, a grid box search space of dimension size: x = 75.0 Å, y = 75.0 Å, z = 75.0 Å was defined, with a search exhaustiveness value of 300. The ligands were centred at x = 36.3, y = 39.9, z = 41.9 and x = -1.4, y = -2.9, z = 27.2 for *Bl*BglC and *Bl*BglB, respectively. In agreement with the retention of configuration catalysis mechanism of GH1 enzymes [146,432], the acid/base catalytic glutamate residues of the enzyme structures were protonated before docking.

3.2.9 Molecular dynamics

Using the AM1-BCC method of the AmberTools17 antechamber program, ligand partial atomic charges were assigned to a fully protonated ligand and a mol2 file created. Making use of the H++ webserver, the proteins were protonated at a pH of 6. The generated H++ inpcrd/prmtop files were converted to a PDB file with the AmberTools17 ambpdb program. The ligand and protein PDB files were then concatenated to create a protein-ligand complex PDB file that was used to generate topology files with programs parmchk2 and tleap of the AmberTools17 package. Modelled with the TIP3P water model during the previous step, water molecules were added as solvent to a cubic periodic box with a minimum distance of 10 Å from the protein edge. The systems were then neutralised using 0.15 M NaCl. The van der Waals and bonded parameters from the general amber force field 2 (GAFF2) were implemented. Acpype, of the AmberTools17 package, converted the topology files to GROMACS format.

Prior to production runs, all molecular systems were energy minimised using a conjugate-gradient being energy relaxed with a force tolerance of 1000 kJ/mol/nm and capped at a maximum of 50,000 steps. The temperature of the systems was then equilibrated at 303.15 K over 100 ps utilising the velocity rescale thermostat (modified Berendsen

thermostat). The pressure was equilibrated using the Parrinello-Rahman barostat in order to maintain 1 bar of system pressure. Production runs were executed using GROMACS 2018.2 with the all-atom AMBER ff14SB force field on one Nvidia Tesla v100 GPU in conjunction with 10 CPU cores at the Centre for High Performance Computing (Cape Town, South Africa). All simulations were run at a pH of 6 and a temperature of 303.15 K (30 °C). Each of the MD simulations was run for 500 ns with 2 fs time steps, and coordinates written every 10 ps. The Linear Constraint Solver (LINCS) algorithm was employed for the duration of the simulations in order to correct for rotational bond lengthening. Coulombic and van der Waals short-range cut-offs were set to 1.4 nm. The long-distance electrostatic interactions were calculated using the Particle-Mesh Ewald (PME) algorithm, with a Fourier grid spacing of 0.16 nm.

3.2.10 Analysis of MD trajectories

Once the MD simulations concluded, the periodic boundary conditions of the trajectories were corrected, jumps across boundaries were removed, and the protein was centred inside the simulation box. The GROMACS tools *gmx rms*, *gmx rmsf* and *gmx gyrate* were used to determine RMSD, RMSF, and R_g, respectively. The protein-ligand hydrogen bonding information during the final 20 ns of the MD's was extracted with the *hbond* command of AMBER's *cpptraj* program, with standard/default geometric criteria/definitions. Molecular interactions other than hydrogen bonds, and including hydrogen bonds, were monitored via DiscoveryStudio. In certain instances, the distances between catalytic residues and bound ligand were monitored over the course of the MD simulations using VMD.

3.2.11 Binding free energy calculations

The molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) method was executed with the g_mmpbsa software tool, to determine the binding free energies of the

protein-ligand complexes. The final 20 ns (480-500 ns) of the MD trajectories were run during the binding free energy calculations, sampled at 10 ps time intervals, and once completed the g_mmpbsa MmPbSaStat.py Python script was run to obtain a summary of the final energies.

3.2.12 Free energy landscape analyses

The global protein motions throughout the MD simulations were investigated using PCA [416]. First, the *gmx covar* tool was used to construct a covariance matrix from the coordinates of the protein Cα atoms. Then, using the *gmx anaeg* tool, eigenvectors and eigenvalues were obtained after diagonalization of the covariance matrix. Finally, Gibbs free energy profiles were constructed using the two eigenvectors with the largest eigenvalues (PC1 and PC2) as these embody the slowest modes (large-scale movements) – this was done using the *gmx sham* tool, and xpm2txt.py and sham.pl scripts. Figures were generated using Python and R.

3.3 Results

3.3.1 Sequence analysis

The 67 characterised GH1 sequences from various enzyme activities, together with the sequences of enzymes *Bl*BglC and *Bl*BglB, were used to create an MSA from which a sequence identity heatmap (Figure 3.1 A) and phylogenetic tree (Figure 3.1 B) were generated. Also, discovered sequence motifs were placed in a motif heatmap (Figure 3.1 C). Quite fittingly, the enzymes showed identical groupings within the sequence identity heatmap, the phylogenetic tree, and the motif heatmap. Primarily, the enzyme sequences formed groups that correlate to enzyme activity, with the exception of nine “unique” sequences. The unique sequences are composed of *Bl*BglC, *Bl*BglB, the 3 dual-phospho activity enzymes, as well as CAB12135.1, AAX76617.1, ABL14155.1, and ACK41762.1.

The unique sequences form their own groups, which we have named Groups 1, 2, and 3. Group 1 contains *Bl*BglC, two of the three dual-phospho activity enzymes, and CAB12135.1. The sequence identity is extremely high (87%) between the *Bl*BglC and CAB12135.1 sequences (Supplementary Figure 3.2). The CAZy database has characterised CAB12135.1 as a 6Pβ-glucosidase, yet its high shared sequence identity with the two dual-phospho activity enzymes in Group 1 (~70%) suggests CAB12135.1 may actually have dual-phospho activity. The phylogenetic tree (Supplementary Figure 3.3) and the many shared motifs between the four enzyme sequences in Group1 (Figure 3.1 C) also suggest that CAB12135.1 is a dual-phospho activity enzyme. In addition, CAB12135.1 is in possession of the longer L6 loop that is usually found in the dual-phospho activity enzymes, therefore CAB12135.1 is not a 6Pβ-glucosidase enzyme as these enzymes do not have this longer loop. Thus, Group 1 can be classified as the dual-phospho activity group. Group 2 contains enzyme *Bl*BglB and AAX76617.1 which share a very high 79% sequence identity. Although characterised as a β-glucosidase, AAX76617.1 has a higher sequence identity with the dual-phospho activity enzymes in Group 1 than with the β-glucosidases. Compared to the other activities/groups, Group 2 shares its highest sequence identity with Group 1 (50%), indicating that *Bl*BglB could be a dual-phospho enzyme. Sequences BAD76141.1 and ABL14155.1 form Group 3. BAD76141.1 is characterised by CAZy as a dual-phospho activity enzyme, even though only 41% sequence identity is shared between this sequence and the other 2 dual-phospho activity enzymes. Furthermore, BAD76141.1 lacks the longer L6 loop, meaning it is not a dual-phospho enzyme. ABL14155.1 is characterised as a β-glucosidase but does not fall into the β-glucosidase group. The enzymes in Group 3 share their highest sequence identity with the β-glucosidase enzymes.

The enzymes in Group 1, which includes *Bl*BglC, possess a combination of motifs that are present in all of the other three GH1 activities. These motifs are 24 and 26 from β-

glucosidases, 26 and 43 from 6Pβ-galactosidases, and 34 from 6Pβ-glucosidases. This shows that dual-phospho activity GH1 enzymes such as *Bl*BglC could depend on sequence regions from all of these activities to perform their broad functions. On the other hand, Group 1 is in possession of motifs 57 and 70 which are found only in Group 1 and could have independent functional importance for dual-phospho activity enzymes. Motif 78 is interesting because it exists only in the enzyme sequences of Groups 1 and 2, providing further evidence of the putative dual-phospho activity of enzyme *Bl*BglB.

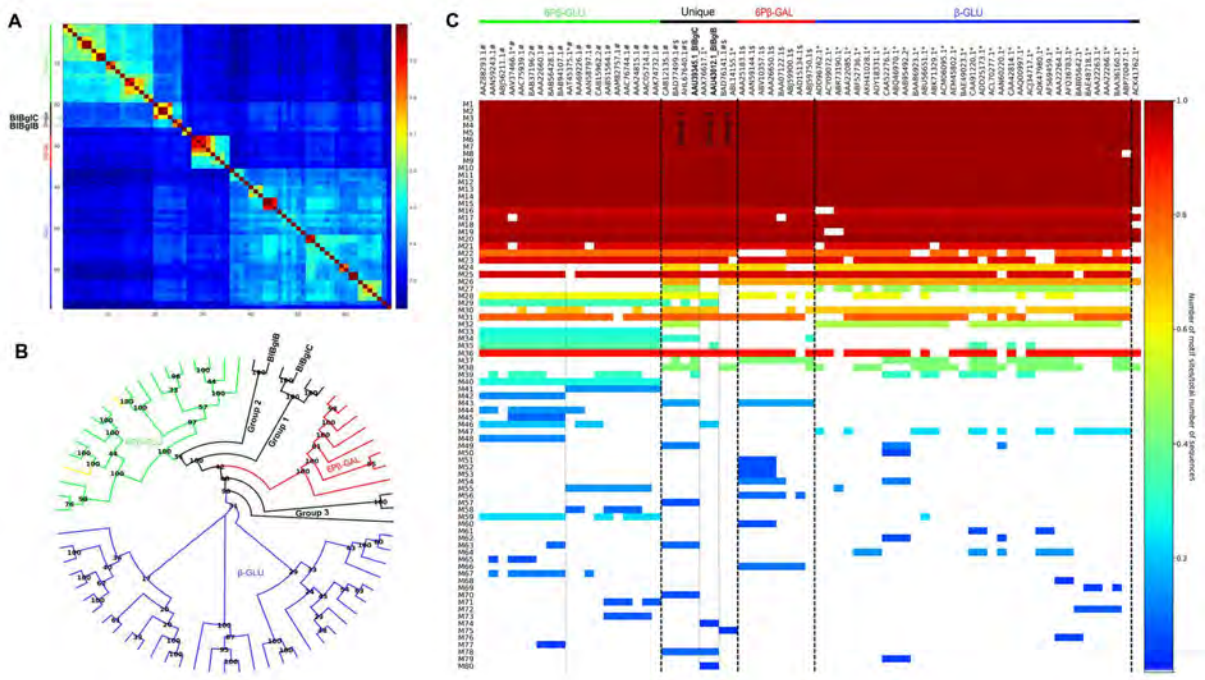


Figure 3.1. *Sequence analyses using BlBglB and BlBglC. Sequence labels for Figure 3.1 A and B can be found in Supplementary Figures 3.2 and 3.3, respectively. (A) Pairwise percentage-identity heatmap between all 69 sequences. The X and Y axes comprise each of the 69 different GH1 enzymes. Identity scores are shown as a colour-coded matrix (increases from blue to red), calculated by comparing every sequence to one another. Major groupings of the different enzyme activities are labelled on the Y axis with colour code: β-glucosidases – Blue; 6Pβ-glucosidases – Green. 6Pβ-galactosidases – Red. Unique groupings – Black. (B) Maximum likelihood phylogenetic tree. Branch numbers indicate bootstrap values. (C) Heatmap representing the 80 discovered motifs in the GH1 enzyme sequences dataset. The colour-coded conservation of motifs increases from blue to red. White colour shows the absence of the motif. Dotted lines indicate sub-groups formed by similar sequences in terms of shared motifs. Reproduced with permission from Veldman et al. 2021 [114].*

Further along in the study, the GH1 motifs discovered here will be analysed in greater detail by mapping the motifs to enzyme structures to uncover any potential function. Overall, the sequence analysis has shown *Bl*BglC to be a putative dual-phospho activity enzyme, whereas *Bl*BglB has a more unique sequence but points towards dual-phospho activity also.

3.3.2 Activity assays

*Bl*BglC and *Bl*BglB enzyme activity was tested against several PNP-monosaccharides and oligosaccharides, as no phosphorylated substrates were available. Activity was tested using PNPβgal, PNPβglc, p-nitrophenyl β-xyloside, p-nitrophenyl β-mannopyranoside and the oligosaccharides cellobiose, cellotriose and cellotetraose. *Bl*BglB was unable to cleave any of the substrates. However, *Bl*BglC was able to cleave PNPβfuc and PNPβgal, with PNPβfuc being used for further characterisation considering it had the highest initial activity. The enzyme showed a typical bell-shaped curve for the pH effect on the activity, which is evidence that two essential glutamate residues are important for its activity. The maximum activity was observed at pH 6. Additionally, it was found that *Bl*BglC has a transition temperature for thermal denaturation of 46 °C (Supplementary Figure 3.4). Using these conditions, the enzyme kinetic parameters were established for the substrates PNPβfuc and PNPβgal. *Bl*BglC followed a Michaelis-Menten kinetics for both substrates (Supplementary Figure 3.5). The K_m for PNPβfuc was 0.19 ± 0.03 mM and the k_{cat} is 5.6 ± 0.1 min^{-1}, whereas the K_m for PNPβgal was 4.6 ± 0.4 mM and the k_{cat} is 2.3 ± 0.1 min^{-1}. Therefore, the k_{cat}/K_m for PNPβfuc was 58 times higher than for PNPβgal. Overall, the activity assays exhibited either low or no activity against the tested unphosphorylated substrates which supports the predicted activity of *Bl*BglC and *Bl*BglB as using phosphorylated substrates.

3.3.3 Analysis of conserved sequence motifs

The sequence motifs that were *totally conserved* in an activity or group were mapped to enzyme crystallographic structures from each activity/group (Figure 3.2). Group 1 (dual-phospho activity), which contains *Bl*BglC, possesses an assortment of motifs that are seen separately in all three of the β-glycosidase activities. These motifs are 24 and 26 from β-glucosidases, 26 and 43 from 6Pβ-galactosidases, and 34 from 6Pβ-glucosidases. It has been suggested that motif 43 (L6 loop) could control access to the opening of the active site, whereby the substrate and the glycon product cannot move through – only the aglycon product can be released [175,183]. Akin to motif 43, motif 27 (L1 loop) is also thought to influence access to the enzyme [175]. Motif 27 is totally conserved only in the dual-phospho activity, although, ~90% of the sequences in the β-glucosidase activity also possess motif 27. The L8a loop contains several important binding residues, one of which (Trp433; *Bl*BglC numbering) is thought to play a major role in substrate specificity in GH1 enzymes. Motifs 26, 33, and 74 all form part of the L8a loop but in different activities/groups: motif 26 is present in β-glucosidases, 6Pβ-galactosidases, and Group 1; motif 33 is unique to 6Pβ-glucosidases; and motif 74 is unique to Group 2. It was revealed in the previous chapter that the 6Pβ-glucosidase L8a loop forms far more bonding interactions with the L8b loop (motif 40), compared to the other activities. The additional inter L8 loop binding most likely increases L8 loop rigidity, limiting the displacement of the specificity-promoting residue and therefore ensuring a steric clash with ligands having an axial O4 hydroxyl group. The residues and function of the L8a loop, with regards to *Bl*BglC and *Bl*BglB, will be further analysed later in the study.

Motifs 24, 34 and 43 are all found in Group 1, and found individually in the β-glucosidase, 6Pβ-glucosidase and 6Pβ-galactosidase activities, respectively. The cause of the broad specificity of dual-phospho activity enzymes could be because they possess motifs that are found individually in each of the other activities. The potential function of these motifs,

and their containing residues, will be analysed later. Moving on from motifs that are shared between the various activities, we now look at motifs that are unique to one individual activity (activity-specific motifs). *Bl*BglC exhibits motifs 57 and 70 which exist only in Group 1, and motifs 74 and 80 are seen only in Group 2 which contains *Bl*BglB (Figure 3.1 C). Not much is known about the function of these motifs, except for motif 74 which forms part of the L8a loop. Motif 80 is near the same region as motif 24 but does not form any part of the L4 loop like motif 24 does. Again, the potential function of these motifs and their containing residues will be analysed later.

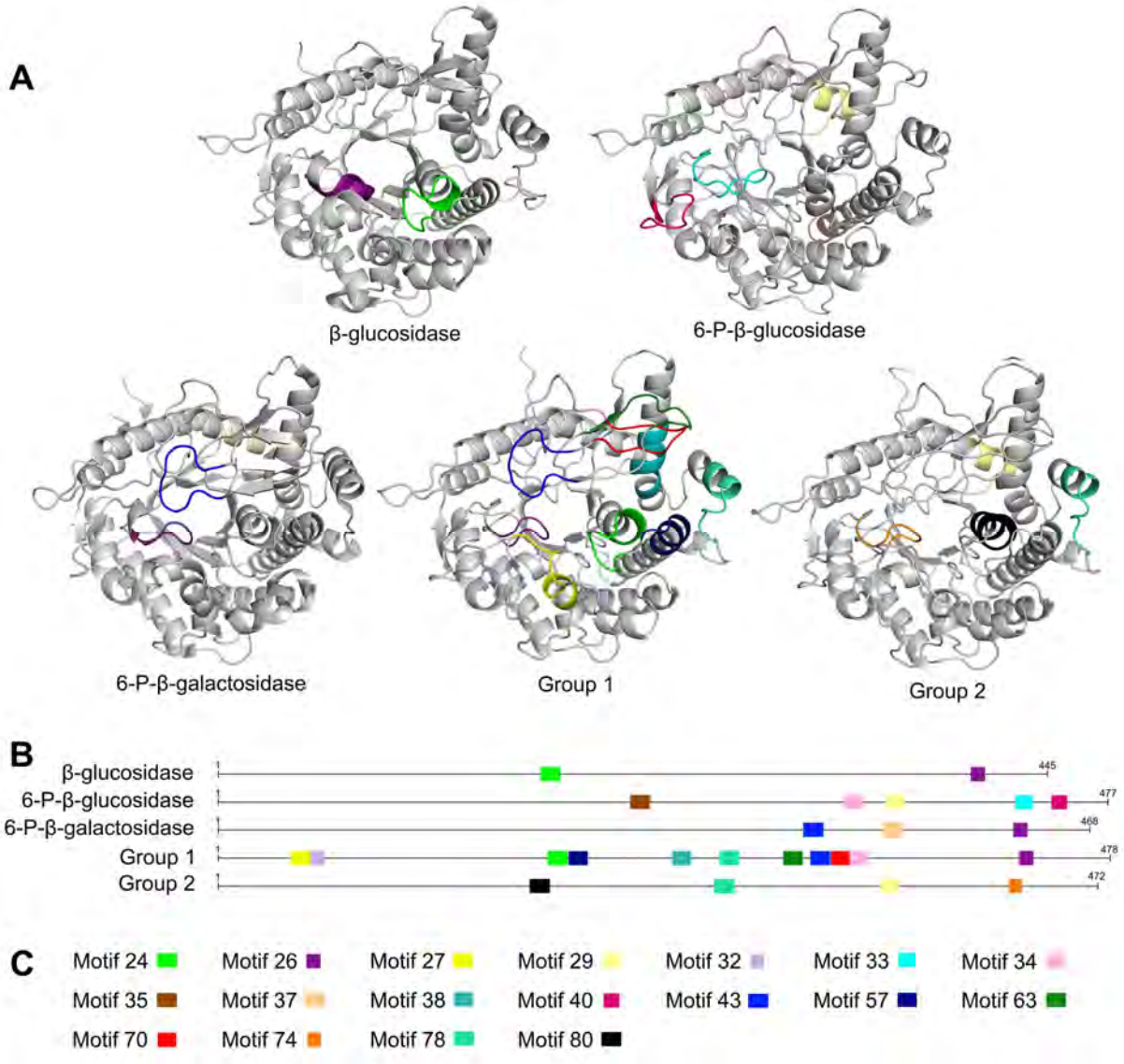


Figure 3.2. *(A) Motifs mapped to crystallographic structures respective of activity/group. Only completely conserved motifs in each activity/group were mapped. PDB IDs used: β-glucosidase – 3AHX, 6Pβ-glucosidase – 4GPN, 6Pβ-galactosidase – 4PBG, Group 1 – 7M1R (BlBglC), and Group 2 – BlBglB. (B) Linear representation of the protein sequences showing the motif locations within the sequences. (C) Motif colour codes. Adapted from Veldman et al. 2021 [114].*

3.3.4 Further analysis of *Bl*BglC

3.3.4.1 Crystallographic structure

The asymmetric unit of the *Bl*BglC crystallographic structure consists of four near-identical chains with RMSD's varying from 0.134 to 0.173 Å. Most of the amino acids could be built, apart from the N-terminal residues 1-4 in chain A, 1 in chain B, and 1-5 in chains C and D, as well as the C-terminal residue 478 in chains C and D. *Bl*BglC has the typical TIM-barrel fold that is found in the GH1 family and is comprised of eight inner parallel beta-strands, surrounded by eight outer alpha helices. The binding site is formed by the loops and smaller alpha helices that connect the barrel (Figure 3.3). Three molecules of ethylene glycol were modelled into conserved sites in chains A, B and C, with just one molecule found in chain D occupying the binding site. The unspecific binding was caused by the high concentration of ethylene glycol (20 %) used as a cryogenic agent.

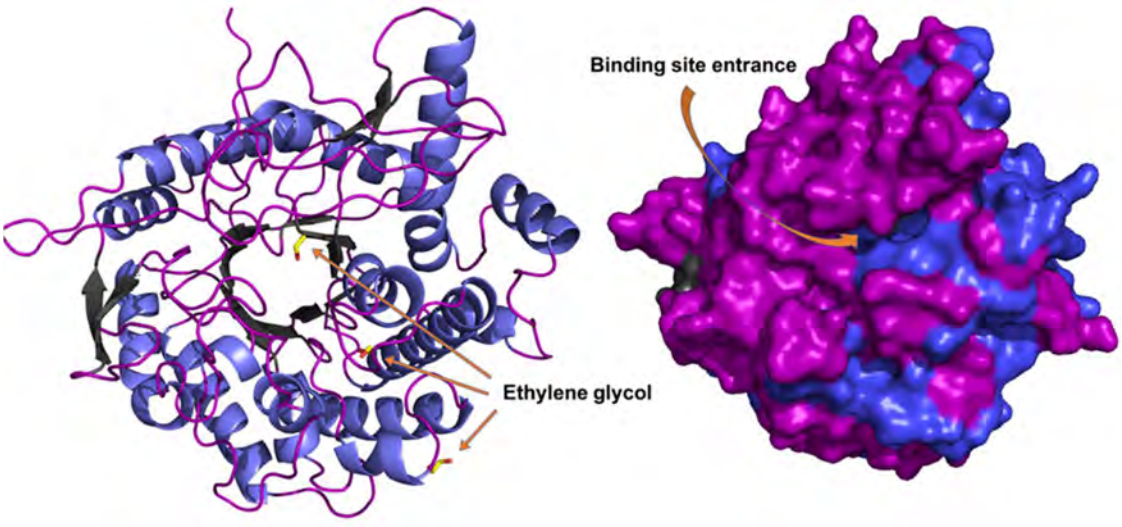


Figure 3.3. *Crystallographic structure of BlBglC. On the left, a cartoon representation of the TIM-barrel fold created by secondary structures: the core beta strands (black); the outer alpha helices (blue); and the loops connecting them (magenta). On the right, a surface model with an arrow indicating the location of the binding site entrance. Reproduced with permission from Veldman et al. 2021 [114].*

A structural similarity search was carried out with the DALI server using *Bl*BglC chain B (most complete structure) as a reference: Gan1D, a dual-phospho GH1 from *G. stearothermophilus* [199] (PDB ID 5OKB), was found to be the most similar enzyme. *Bl*BglC and Gan1D share 70 % amino acid sequence identity and a structural RMSD of 0.48 Å. In Figure 3.4 the binding residues of Gan1D (PDB ID 5OKE) are compared to the residues of the *Bl*BglC crystallographic structure (PDB ID 7M1R). In PDB ID 5OKE, Gan1D is in complex with the substrate cellobiose-6-phosphate (Cell6P) and is the only dual-phospho conservatively refined PDB structure having a ligand with an aglycon group. We are able to compare the apo *Bl*BglC and ligand-bound 5OKE structures because ligand binding does not cause significant changes in the side chain positions of the Gan1D residues (Supplementary Figure 3.6). The first thing we notice is that the residue numbering between *Bl*BglC and Gan1D matches; in fact, the residue numbering matches for the entire length of their sequences. Excluding the mutation (red arrow), all but two binding site residues (black arrows) are conserved between *Bl*BglC and Gan1D: residues Ile173 and Phe177 in Gan1D are replaced with Tyr173 and His177 in *Bl*BglC. In PDB 5OKE, Ile173 and Phe177 form hydrophobic interactions with the ligand aglycon. In GH1 enzymes, the aglycon-binding residues are less conserved compared to the glycon- and phosphate-binding residues – the reason for this is most likely due to the variability of the aglycon in GH1 ligands. Interestingly, Tyr173 is only found in two of the 69 sequences in the bacterial GH1 dataset, namely in *Bl*BglC and CAB12135.1. In the remaining 67 sequences, the residues in this position are much smaller in size compared to tyrosine: Asparagine, isoleucine, methionine, valine, serine, proline, cysteine and alanine. The larger Tyr173 residue could have influence on the function of *Bl*BglC.

The analogous *Bl*BglC and Gan1D residues that have low discrepancies in side chain position are shown with cyan arrows (Figure 3.4), namely Gln23, Trp125, Trp425, and Trp433. Because these residues are far more tightly superimposed compared to the

remaining residues, it could mean that their positions are important in dual-phospho enzymes, and that the remaining residues change their positions according to either galacto- or gluco-substrates. However, the researchers responsible for the Gan1D structures put forward the idea that Gln23 and Trp433 in Gan1D allow for the modulation of preference toward a galactose or glucose sugar because they found that both residues mostly bind to different ligand locations depending on galacto- or gluco-configured ligands. This could be the case, whereby Gln23 and Trp433 do not move but their interactions change due to a change of ligand spatial position (galacto vs gluco) and therefore also a change of interactions with the more mobile residues. The other two tightly superimposed residues are Trp125 and Trp425. In PDB 5OKE, both of these residues form hydrophobic interactions with Cell6P, with Trp425 also forming a hydrogen bond. Trp125 and Trp425 are positioned on opposite sides of bound ligand and most likely help to initially place substrates into a reasonably accurate position within the active site before being “locked in”.

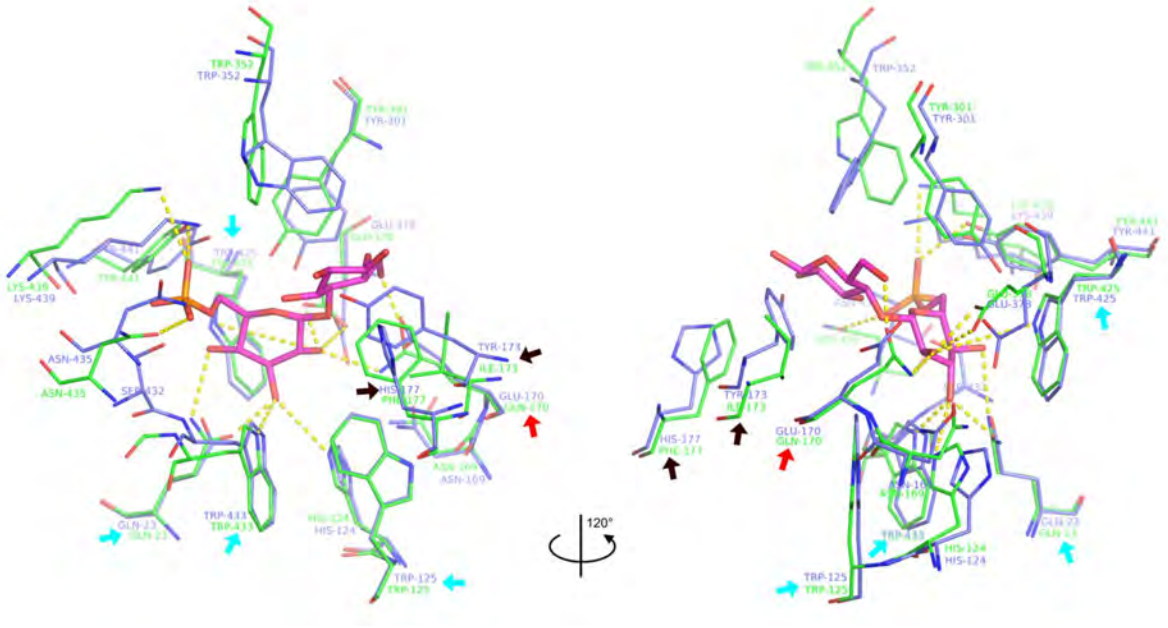


Figure 3.4. *Details of BlBglC ligand-binding site (purple residues), in comparison with Gan1D (PDB ID 5OKE - green residues) complexed with cellobiose-6P (magenta). Black arrows indicate sequence positions with different amino acids. Red arrow indicates mutant Gan1D-E170Q. Cyan arrows indicate sequence positions with the same amino acid that do not have discrepancies in side chain position. Here, the apo BlBglC residues can be compared to the ligand-bound Gan1D residues because both the apo and ligand-bound Gan1D residue side chain positions are very similar (Supplementary Figure 3.6). Reproduced with permission from Veldman et al. 2021 [114].*

3.3.4.2 In silico docking

Complying with the retention of configuration catalysis mechanism of GH1 enzymes, the acid/base catalytic glutamate (Glu170) was protonated (Glh170) before docking and MD. Using blind docking with the program Vina-Carb, initial attempts at docking into the active site of *Bl*BglC failed with the suspected cause being the L6 loop that is thought to block the entrance to the active site [175,183]. To overcome this obstacle, an apo MD simulation was run at an increased temperature of 315 K in order to open up the L6 loop and expose the active site (Figure 3.5). During this process, the L8a loop also changed positions which opened up the enzyme even more.

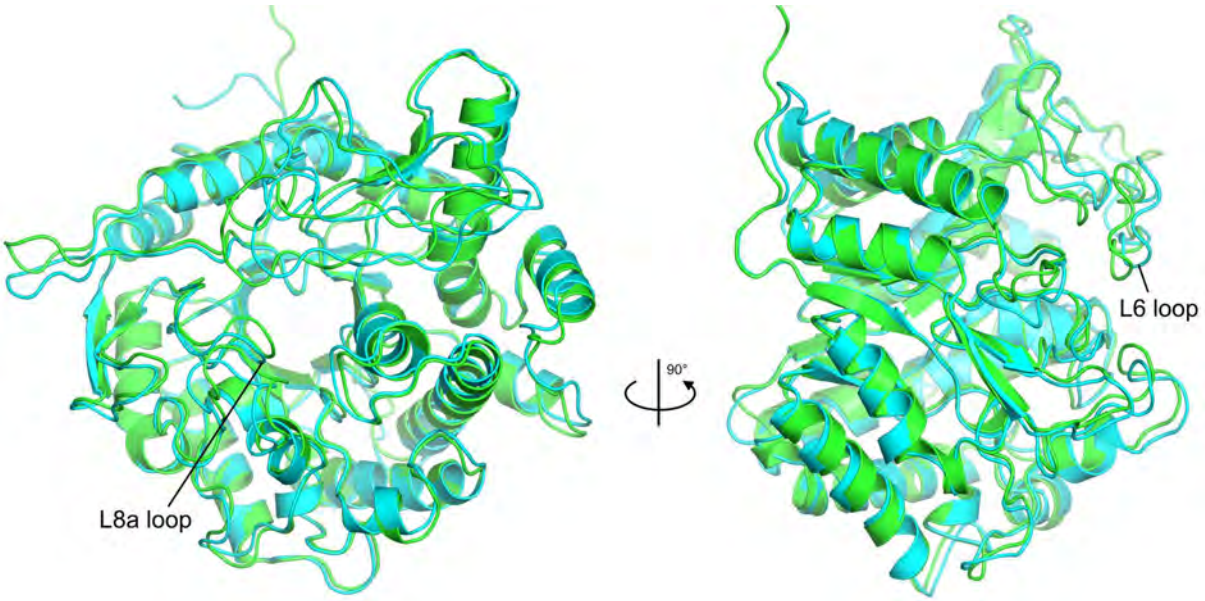


Figure 3.5. *Expanding the BlBglC active site cavity. An apo MD simulation was run at an increased temperature of 315 K in order to open up the enzyme and expose the active site to facilitate in silico docking. Original BlBglC structure – Green; BlBglC structure at 115 ns of simulation – Cyan.*

The second attempt at docking was successful: the PNP6Pgal and PNP6Pglc ligands were docked into the active site with similar positions and similar binding energies of -8.2 and -8.3 kcal/mol, respectively (Figure 3.6). The ligand-residue interactions were compared with the known binding residues of 6Pβ-glycosidase enzymes established in Table 1.1 from Chapter 1. The two different ligands share many residue interactions with the enzyme, however PNP6Pgal interacted with eleven active site residues compared to only seven for the PNP6Pglc ligand. Residue-binding with Asn169 and Lys439 was absent for both ligands, with PNP6Pglc also missing both catalytic residues Glu170 and Glu378 as well as His124 and Trp125. Although the binding energies and orientations of the ligands are similar, the docking of PNP6Pgal is considered superior since it had four additional active site residue interactions, which include the two catalytic glutamate residues. The adequate docking of the two ligands into the active site permitted the subsequent execution of MD simulations.

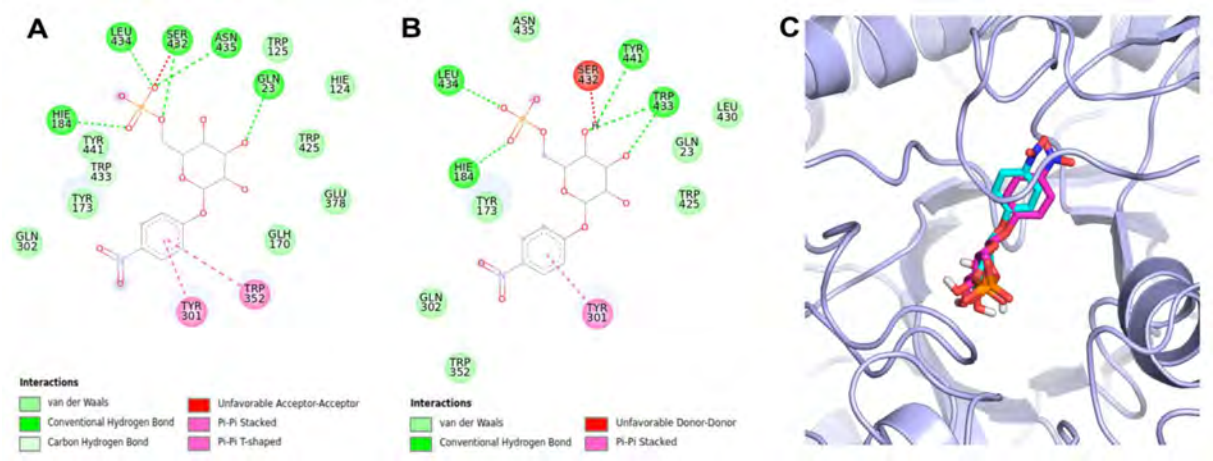


Figure 3.6. *BlBglC in silico blind-docking results. (A) PNP6Pgal-BlBglC interactions, -8.2 kcal/mol, (B) PNP6Pglc-BlBglC interactions, -8.3 kcal/mol, and (C) superimposition of both complexed post-docking. PNP6Pgal – cyan; PNP6Pglc – magenta. Reproduced with permission from Veldman et al. 2021 [114].*

3.3.4.3 Molecular dynamics

In both a mechanistic and energetic manner, it is feasible to perform a full analysis of protein-ligand recognition and binding using MD [448]. In triplicate, 500 ns MD simulations were executed using the docked ligands to further study the dual-phospho activity and specificity of *Bl*BglC. At 500 ns of simulation all triplicates of both ligands were in the enzyme active site, having deviated only slightly from their initial positions. Interestingly, in the previous chapter where 6Pβ-glucosidase *Bl*BglH was used, PNP6Pgal exited the enzyme while PNP6Pglc exhibited high affinity [105] – a big contrast to the result of *Bl*BglC.

3.3.4.3.1 Trajectory analysis

Calculations of RMSD, R_g and RMSF were used to analyse the MD trajectories (Figure 3.7). From RMSD, conformational change and overall stability of the protein can be evaluated. Throughout all the simulations the protein RMSD was mostly stable and deviated very little. There was an insignificant difference in protein RMSD between the apo protein and the ligand-bound proteins (Figure 3.7 A), even though the PNP6Pglc triplicates

displayed a lower RMSD on average. Additionally, VMD was used to visually examine the simulations and no significant conformational change caused by ligand binding was observed. The ligand RMSD was stable for all the ligands during the simulations (Figure 3.7 B). The RMSD of the PNP6Pgal triplicates were a little higher on average compared to the PNP6Pglc triplicates, an indication that the PNP6Pgal triplicates moved further from their initial docked position compared to the PNP6Pglc triplicates. R_g is a quantification of protein compactness, and no significant difference in the R_g was observed between the apo protein and the proteins complexed with the ligands (Figure 3.7 C). RMSF is a measure of individual residue displacement during the MD run and provides insight on local protein movement. Overall, the residues of the ligand-free protein fluctuated marginally more than the residues of the ligand-bound proteins (Figure 3.7 D). Three protein locations experienced a notable difference in RMSF due to ligand binding – residues 250-275, residues 321-350, and residues 388-391. The 250-275 region flanks the enzyme (Figure 3.7 E) but could have a role to play in opening the enzyme slightly when no ligand is bound. Region 321-350 contains the L6 beta-hairpin loop, which is believed to act as a gatekeeper to the entrance of the active site [175,183] – in a ligand-bound state this loop may stabilise because the enzyme does not require a ligand from its environment to pass through the gate. Alternatively, ligand binding increased the RMSF of residues 388-391 that form part of peripheral loop L7 which has unknown significance for enzyme function. The reason why ligand binding provides this loop with a degree of freedom remains unclear, but it could have to do with the catalytic Glu378 residue that is just eight residues away in the sequence. On the other hand, when comparing the RMSF of the PNP6Pgal and PNP6Pglc complexes, the residues of the PNP6Pgal complexes showed slightly larger fluctuations on average. In general, the stability of both of the ligands during MD in triplicate is evidence of their suitability as substrates for *Bl*BglC.

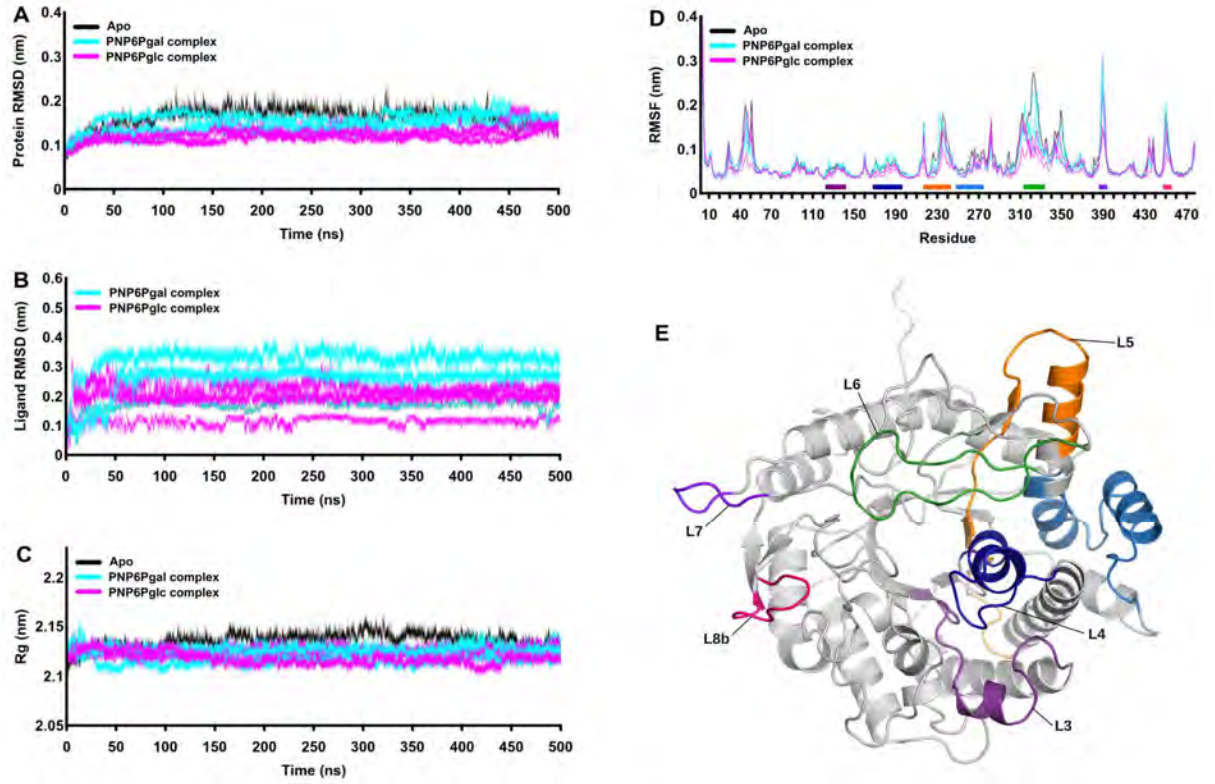


Figure 3.7. *Trajectories of the MD simulations. (A) Protein backbone RMSD after least square fitting to protein backbone, (B) ligand RMSD after least square fitting to protein backbone, (C) protein radius of gyration, (D) protein residue RMSF, and (E) regions mapped to BlBglC structure where disparities exist in the RMSF values between either the ligand-free and ligand-bound simulations, or between the PNP6Pgal and PNP6Pglc simulations. Reproduced with permission from Veldman et al. 2021 [114].*

3.3.4.3.2 Ligand positions and interactions at 500 ns of simulation

Relative to each other, the spatial positions of all the ligands at 500 ns of MD simulation are shown in Figure 3**.8** A. The PNP6Pglc triplicates clustered closer together with an average RMSD between the triplicates of 0.31 Å, whereas the PNP6Pgal triplicates exhibited more sporadic positions and had an average RMSD of 0.7 Å. With Discovery Studio, 2D depictions of the protein-ligand interactions were produced (Figure 3**.8** B-D). The PNP6Pgal triplicates all interacted with 12 of 13 active site residues, with the first two triplicates missing residue Trp352 and the third triplicate missing residue Asn169. On the other hand, the first PNP6Pglc triplicate interacted with all 13 active site residues, while the other two PNP6Pglc triplicates were only missing the Asn169 interaction. The 2D

representations of interactions are merely snapshots in time. Nevertheless, several patterns were observed when comparing the interactions of the PNP6Pgal and PNP6Pglc complexes. The positively-charged Lys439 appears to be very important, as it forms two attractive charge interactions with the negatively-charged ligand-phosphate oxygen atoms from all three triplicates from both ligands. Concerning PNP6Pgal, all triplicates form O2-hydroxyl hydrogen bonds and O4-hydroxyl hydrogen bonds with the His124 and Trp433 residues, respectively. Concerning PNP6Pglc, all triplicates form O2-hydroxyl hydrogen bonds and O4-hydroxyl hydrogen bonds with the Glu378 and Gln23 residues, respectively. Residue Tyr173 appears to be more important for PNP6Pglc binding, as it had two hydrogen bond interactions and one pi-pi interaction with all three PNP6Pglc triplicates, whereas it had only van der Waals interactions with the PNP6Pgal triplicates. Only *Bl*BglC and CAB12135.1 have a tyrosine residue in this position (Tyr173), which is bulkier than the residues in this position from the sequences and could play a role in the specificity distinction between the two kinds of ligands. Another difference is that residues Tyr301 and Tyr302 have conventional hydrogen bonds with all three PNP6Pglc triplicates, whereas the PNP6Pgal triplicates have none. Ala226 may hold significance in *Bl*BglC binding as it forms a pi-alkyl interaction with five of the six ligands. The residue in the sequence position of Ala226 forms a ligand interaction in three of the five crystallographic structures from Table 1.1, meaning that it could be a relatively common interaction found in GH1 enzymes. Curiously, in both Figure 3.8 and Table 1.1, the Tyr173 and Ala226 interactions were either both present or both absent. The aglycon portion of the ligand is situated in between the Tyr173 and Ala226 residues and, therefore, these two residues may work together to bind and position the ligand. Finally, His184 forms van der Vaals interactions with all of the PNP6Pglc triplicates but not any of the PNP6Pgal triplicates.

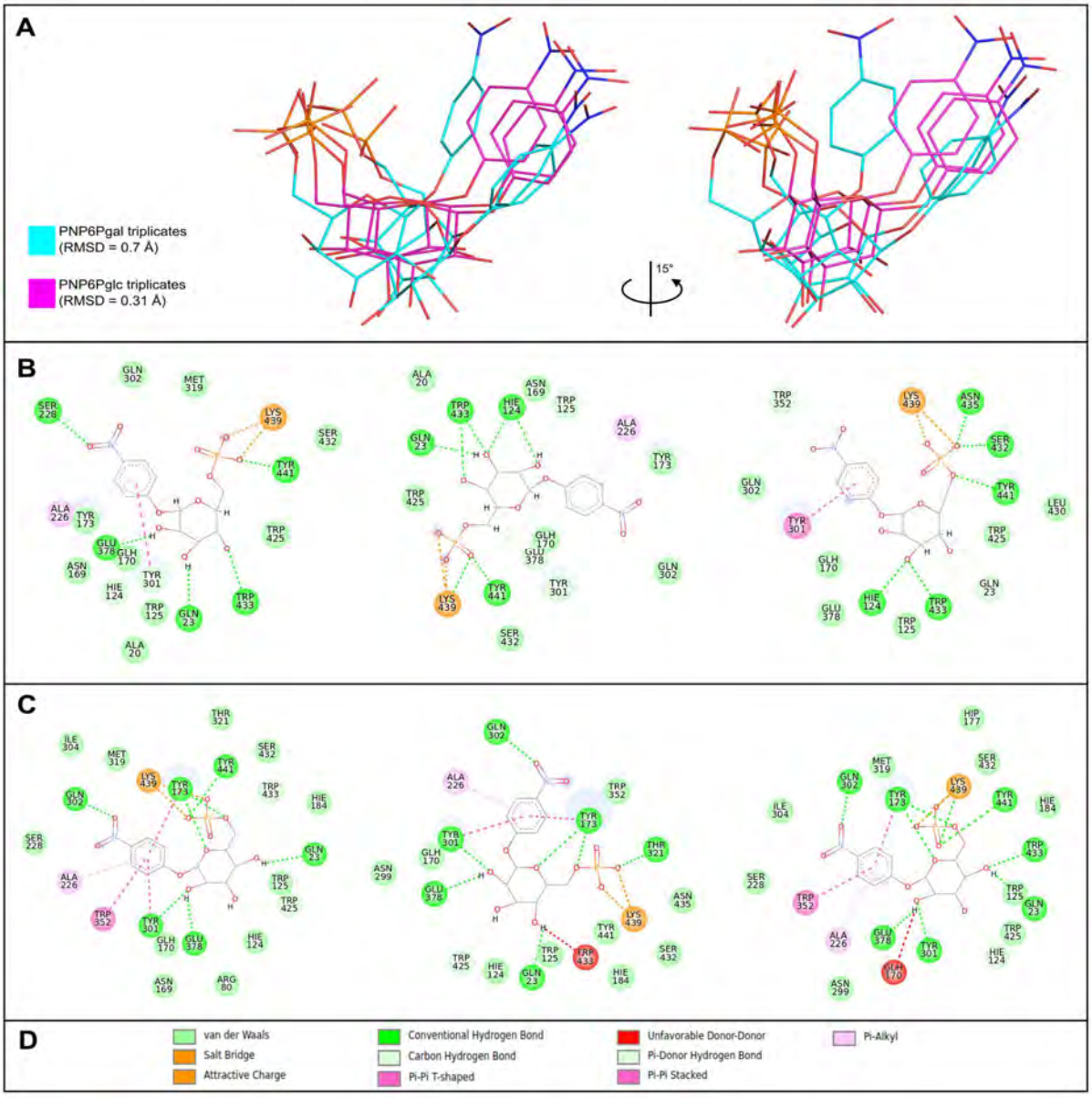


Figure 3.8. *BlBglC interactions with PNP6Pgal and PNP6Pglc triplicates at 500 ns of MD simulation. (A) Relative 3D positions of all the ligands, (B) PNP6Pgal triplicate interactions, (C) PNP6Pglc triplicate interactions, and (D) interaction colour-code. Reproduced with permission from Veldman et al. 2021 [114].*

3.3.4.3.3 Comprehensive hydrogen bond comparison between the PNP6Pgal and PNP6Pglc complexes

The molecular recognition of ligands by protein binding sites relies heavily on hydrogen bonds [449–451] and they are also crucial for catalysis [449,452,453]. To further study the *Bl*BglC-ligand interactions, the hydrogen bonding during the final 20 ns of the MD

simulations was recorded. The average frequency of each of the two sets of triplicate ligands forming hydrogen bonds with *Bl*BglC are shown in Figure 3**.9** and Table 3.1, but depicted in two different ways. The hydrogen bonding for each triplicate is shown in Supplementary Figure 3.7.

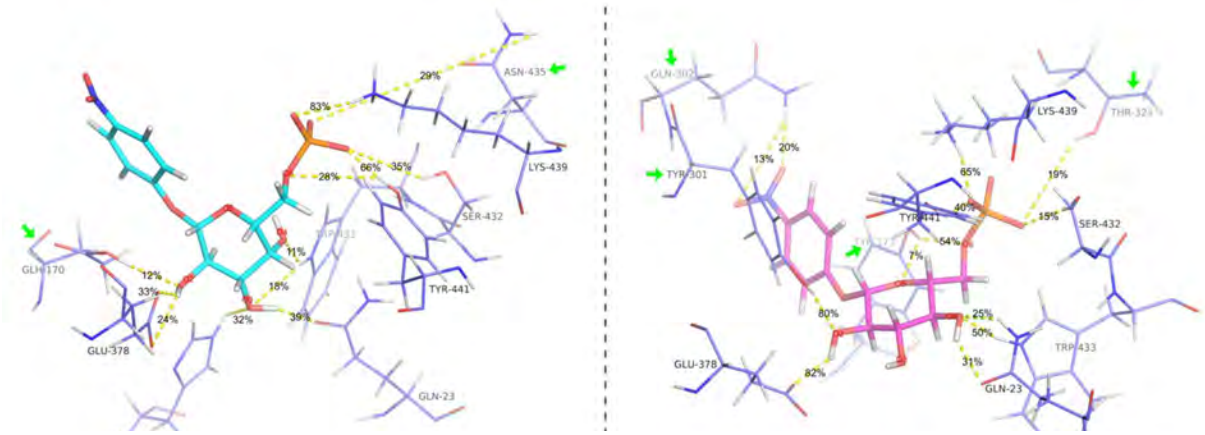


Figure 3.9. *Triplicate average frequency of hydrogen bonding interactions (%) between BlBglC and the PNP6Pgal (left) and PNP6Pglc (right) ligands throughout the last 20 ns of MD simulations. Green arrows show hydrogen bonding residues that are present in one complex but absent in the other. Reproduced with permission from Veldman et al. 2021 [114].*

The largest disparity between the PNP6Pgal and PNP6Pglc triplicates is the hydrogen bonding interactions with their O3 and O4 hydroxyl groups. When combining the hydrogen bonding from all three interacting residues the PNP6Pgal O3 hydroxyl group has a residue-combined frequency of 89%, while the PNP6Pglc O3 hydroxyl group shows no hydrogen bonding. On the other hand, a residue-combined hydrogen bonding frequency of 106% is formed with the PNP6Pglc O4 hydroxyl group, in contrast to just 11% with the PNP6Pgal O4 hydroxyl group. The residues that form hydrogen bonds with one ligand but do not form any with the other ligand are shown with green arrows in Figure 3.9. His124 is one of the three residues that hydrogen bonds to the PNP6Pgal O3 hydroxyl group, but this residue does not hydrogen bond with PNP6Pglc at all. Another large disparity between the PNP6Pgal and PNP6Pglc triplicates is their residue-combined O2 hydroxyl group

hydrogen bonding frequencies of 69% and 162%, respectively. This is more than double the hydrogen bonding for the PNP6Pglc O2 hydroxyl group compared to the PNP6Pgal O2 hydroxyl group. Although the O2-H atom of both ligands bound to the catalytic Glu378 residue, the O2 atom of PNP6Pgal and PNP6Pglc bound to either Glu170 or Tyr301, respectively. However, there was far more Tyr301-PNP6Pglc bonding compared to the Glh170-PNP6Pgal bonding. Tyr301 also forms pi-pi interactions with the aglycon group in both ligands. Another significant difference between the two ligands is the very active involvement of Tyr173 with PNP6Pglc. Tyr173 has hydrogen bonding with the PNP6Pglc O5 and O6 atoms, whereas hydrogen bonding with Tyr173 is absent with PNP6Pgal. In addition to hydrogen bonds, Tyr173 forms pi-pi interactions with PNP6Pglc but van der Waals interactions with PNP6Pgal. Now on to the ligand phosphate group: this is the only part of the ligand where PNP6Pgal had more hydrogen bonding compared to PNP6Pglc, with a residue-combined frequency of 213% and 139% respectively. Lys439 appears to have a large role in binding to the ligand phosphate group, as it forms hydrogen bonds *and* charge attraction forces with all triplicates from both ligands (Figure 3**.8** & Figure 3.**9**). The PNP6Pglc phosphate group displays an extra interaction with Thr321 that is absent in PNP6Pgal. However, the PNP6Pgal-Ser432 hydrogen bonding frequency is more than double the PNP6Pglc-Ser432 frequency. The O6 ligand atom links the ligand phosphate group to the glycon group – it is interesting that the PNP6Pgal O6 atom forms hydrogen bonds with Tyr441, whereas the PNP6Pglc O6 atom forms hydrogen bonds with Tyr173. The final significant discrepancy is that only PNP6Pglc forms hydrogen bonds with the ligand aglycon group, with residue Gln302.

All the differences discussed above most likely originate from the only difference between the two ligands: the position of the O4 hydroxyl group. Hence, it is highly probable that the decisive residues enabling the broad specificity of *Bl*BglC are Gln23 and Trp433. When the O4 hydroxyl group is in the axial position (PNP6Pgal), Gln23 and Trp433 bind strongly

to the O3 hydroxyl group. However, when the O4 hydroxyl group is in the equatorial position (PNP6Pglc), Gln23 and Trp433 bind strongly to this equatorial O4 hydroxyl group. Interestingly, the researchers responsible for the Gan1D structures propose that the specific hydrogen bonding of Gln23 and Trp433 in Gan1D enables the change of preference toward a galactose or glucose sugar. In their Gan1D-galactose6P complex they note that the O4 hydroxyl prefers to interact with Trp433, while the Gan1D-glucose6P O4 hydroxyl prefers to interact with Gln23. This is also the case in the *Bl*BglC hydrogen bonding frequency with the two different ligands (Figure 3.**9**). Additionally, in Gan1D-Gal6P the ligand forms one hydrogen bond with Gln23 and two hydrogen bonds with Trp433, whereas the opposite is true in their Gan1D-Glc6P structure. This is also the case for the *Bl*BglC hydrogen bonds. The *Bl*BglC results provide additional evidence of the influence that residues Gln23 and Trp433 have in the broad specificity of dual-phospho activity GH1 enzymes. Overall, the plethora of protein-ligand hydrogen bonds in both complexes during the last 20 ns is an indication that both PNP6Pgal and PNP6Pglc are natural binders to *Bl*BglC.

Table 3.1. *Average frequency of hydrogen bonding (%) between BlBglC and the PNP6Pgal and PNP6Pglc triplicates throughout the last 20 ns of the MD runs.*

Substrate binding site	**Protein residue**	**Residue atom**	**Substrate atom**	**Average hydrogen bond frequency (%) followed by the number of triplicates which had interaction**	
				PNP6Pgal	**PNP6Pglc**
Phosphate	THR321	OG1-HG1	O_1P/O_2P		19% 3
	SER432	OG-HG	O_1P/O_3P	35% 2	15% 2
	ASN435	ND2-2HD2	O_3P	29% 1	
	LYS439	NZ-HZ1/HZ2/HZ3	$O_1P/O_2P/O_3P$	83% 3	65% 3
	TYR441	OH-HH	$O_1P/O_2P/O_3P$	66% 3	40% 3
Glycon	GLN23	NE2-2HE2	O4		25% 3
		OE1	O4-H		31% 3
		OE1	O3-H	39% 2	
	HIE124	NE2-HE2	O3	32% 3	
	GLH170	OE2-HE2	O2	12% 1	

	TYR173	OH-HH	O5		7% 2
		OH-HH	O6		54% 3
	TYR301	OH-HH	O2		80% 3
	GLU378	OE1	O2-H	24% 2	
		OE2	O2-H	33% 1	82% 3
	TRP433	NE1-HE1	O3	18% 3	
		NE1-HE1	O4	11% 2	50% 3
	TYR441	OH-HH	O6	28% 2	
Aglycon	GLN302	NE2-2HE2	O7		13% 3
		NE2-2HE2	O8		20% 3
Total				**410%**	**501%**

3.3.4.3.4 Binding free energy calculations

In the process of ligand recognition or enzyme catalysis, the strength of the biomolecular interaction can be estimated using calculations of binding free energy [411]. With the last 20 ns (480-500 ns) of the *Bl*BglC-ligand MD simulations, the molecular mechanics Poisson-Boltzmann surface area (MM/PBSA) method was used to determine the binding free energies. Although not as accurate as the more computationally intensive methods, MM/PBSA is a quick, effective and reproducible means of studying protein-ligand binding interactions [408–410]. Most studies implementing MM/PBSA involve high-throughput screening for drug discovery, but here MM/PBSA was utilised here to ascertain an estimation of ligand binding affinity. Every one of the PNP6Pgal and PNP6Pglc triplicates complexed with *Bl*BglC exhibited a significantly low total binding free energy (Table 3.2), which is an indication that *Bl*BglC has very high affinity for both of the ligand types. The average binding free energy values for the PNP6Pgal and PNP6Pglc triplicates were -31.2 kJ/mol and -146.3 kJ/mol, respectively. This is additional evidence of the dual-phospho activity of *Bl*BglC.

Table 3.2. *MM/PBSA binding free energy values between BlBglC and the PNP6Pgal and PNP6Pglc triplicates during the last 20 ns of the MD simulations. All values are in kJ/mol.*

PNP6Pgal triplicate no.	Van der Waal's	Electrostatic	Polar solvation	Solvent Accessible Surface Area (apolar)	Total binding energy
1	-141.9 ± 0.3	-578.7 ± 0.9	613.9 ± 0.9	-17.6 ± 0.02	**-124.3 ± 0.7**
2	-144.7 ± 0.3	-545.6 ± 1.5	558.2 ± 1.6	-16.5 ± 0.01	**-148.6 ± 0.5**
3	-135 ± 0.3	-610.9 ± 0.7	642.1 ± 0.6	-17.1 ± 0.02	**-120.9 ± 0.5**
Average					**-131.2**

PNP6Pglc triplicate no.	Van der Waal's	Electrostatic	Polar solvation	Solvent Accessible Surface Area (apolar)	Total binding energy
1	-163.7 ± 0.4	-562.1 ± 0.7	583.6 ± 0.6	-18.4 ± 0.02	**-160.6 ± 0.5**
2	-163.2 ± 0.3	-516.02 ± 0.9	565.6 ± 0.9	-18.6 ± 0.02	**-132.2 ± 0.5**
3	-175.1 ± 0.3	-548.7 ± 0.8	596.4 ± 0.6	-18.6 ± 0.02	**-146 ± 0.6**
Average					**-146.3**

3.3.4.3.5 Distance between *Bl*BglC catalytic residues and ligands throughout MD simulations

To confirm ligand stability and substantiate catalytic capacity, the distances between the catalytic residues and the ligands were checked during the course of the MD simulations (Figure 3.10). These distances were more consistent and stable for the PNP6Pglc ligands as compared to the PNP6Pgal ligands, and they were also smaller on average (Figure 3.10 B). After the first 20 ns, the measurements between the PNP6Pglc ligands and the catalytic residues regularly stay within 3.5 Å. The measurements between the PNP6Pgal ligands and the catalytic residues also remain within 3.5 Å for most of the simulation time, although they were not as steady as the PNP6Pglc measurements.

The many MD simulation analyses, including trajectories, binding interactions, binding free energy, and catalytic residue distance, indicate that *Bl*BglC has the potential to bind and hydrolyse both PNP6Pgal and PNP6Pglc ligands. The results have also revealed that *Bl*BglC is marginally more suited to PNP6Pglc, as the ligand showed higher RMSD

stability and triplicate-consistency compared to PNP6Pgal, as well as having more hydrogen bonding and lower binding free energy.

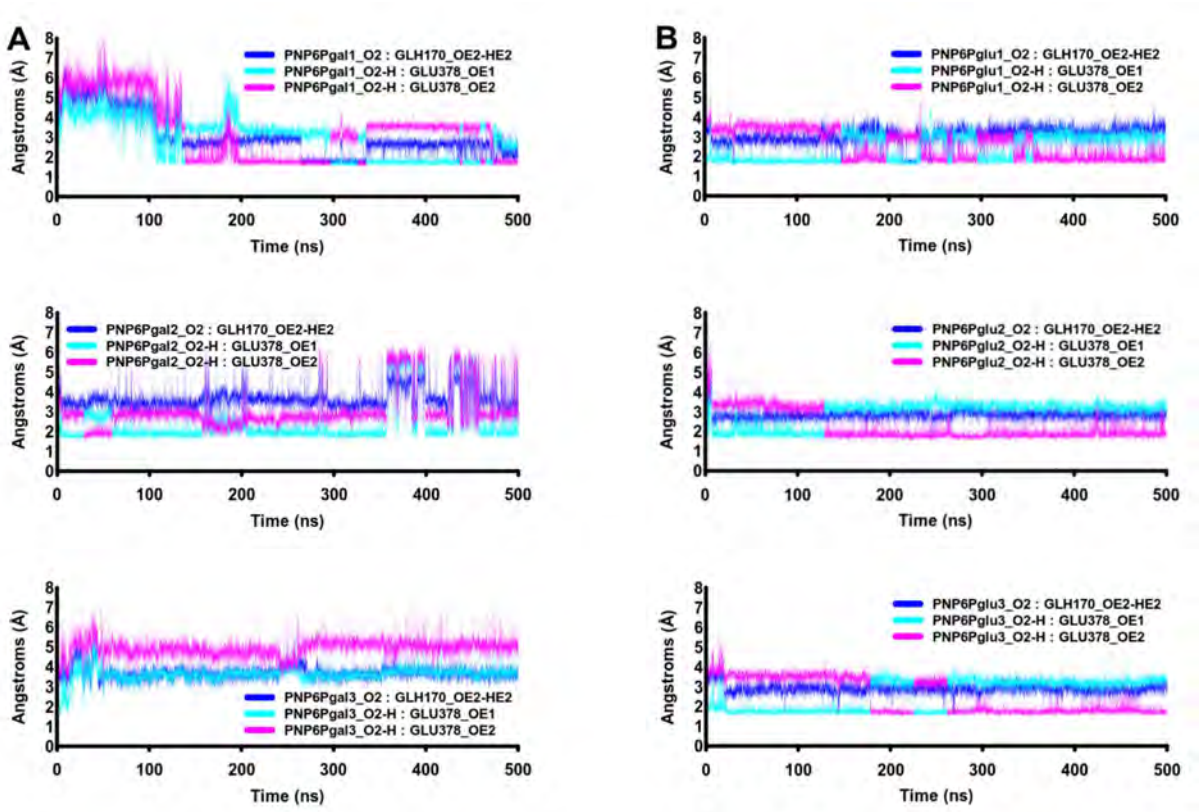


Figure 3.10. *Distance between the ligands and the BlBglC catalytic residues throughout the triplicate MD simulations using (A) PNP6Pgal and (B) PNP6Pglc. Reproduced with permission from Veldman et al. 2021 [114].*

3.3.4.3.6 Essential dynamics investigations using PCA and FEL

Using PCA and FEL, it is possible to investigate if any structural conformation shifts occur that are caused by ligand binding [454,455]. The first triplicate MD run from both PNP6Pgal and PNP6Pglc ligands were extended to 1000 ns (1 μs), as conformational shifts can occur on longer timescales. From the MD trajectories, the primary movement of the protein was determined by extracting the correlated motions during conformational sampling. Figure 3.11 shows PCA scatter plots and FEL's of the apo protein and both ligand-bound proteins. The apo protein was more structurally dispersed throughout the trajectory as seen by the larger area compared to the ligand-bound proteins. The smaller

more distinct dark-blue FEL regions of the ligand-bound proteins show that these proteins progressed through more prominent conformations. The FEL's of the PNP6Pgal and PNP6Pglc complexes have four and three separate energy wells respectively, seen as the dark-blue regions. This means that the structural conformation of the PNP6Pglc complex was slightly more stable as compared to the PNP6Pgal complex, which correlates with the study thus far. The more structurally dispersed apo protein also correlates with the study, in that ligand binding slightly promoted conformational stability of *Bl*BglC.

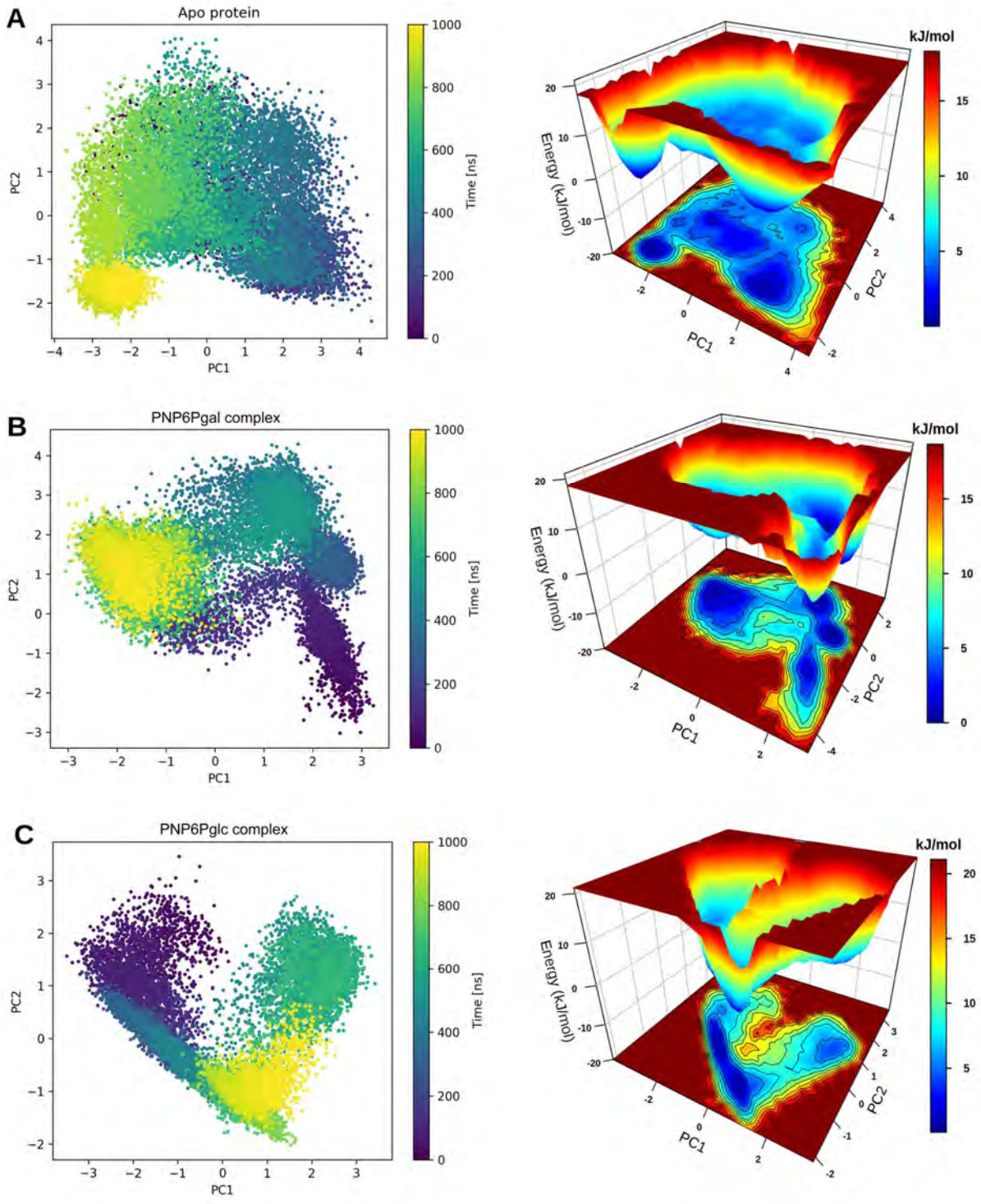


Figure 3.11. *Essential dynamics of BlBglC throughout 1000 ns MD simulations: PCA scatter plot (left) and FEL (right), of (A) apo protein, (B) PNP6Pgal complex and (C) PNP6Pglc complex.*

3.3.5 Further analysis of enzyme *Bl*BglB

Two regions of the *Bl*BglB crystallographic structure were missing due to high flexibility – these regions were loops L4 and L6, and they had to be modelled in order to obtain a complete enzyme structure for use in *in silico* docking and MD simulations (Figure 3.12). Both the original *Bl*BglB crystal structure and the model were used with the DALI server to search for structural homologs. The top hit for the crystal structure and the model was PDB ID 5OKB from the Gan1D enzyme, a dual-phospho GH1 enzyme. Using the *Bl*BglB model structure, *in silico* docking and MD simulations were performed in order to obtain evidence of enzyme activity and/or substrate specificity.

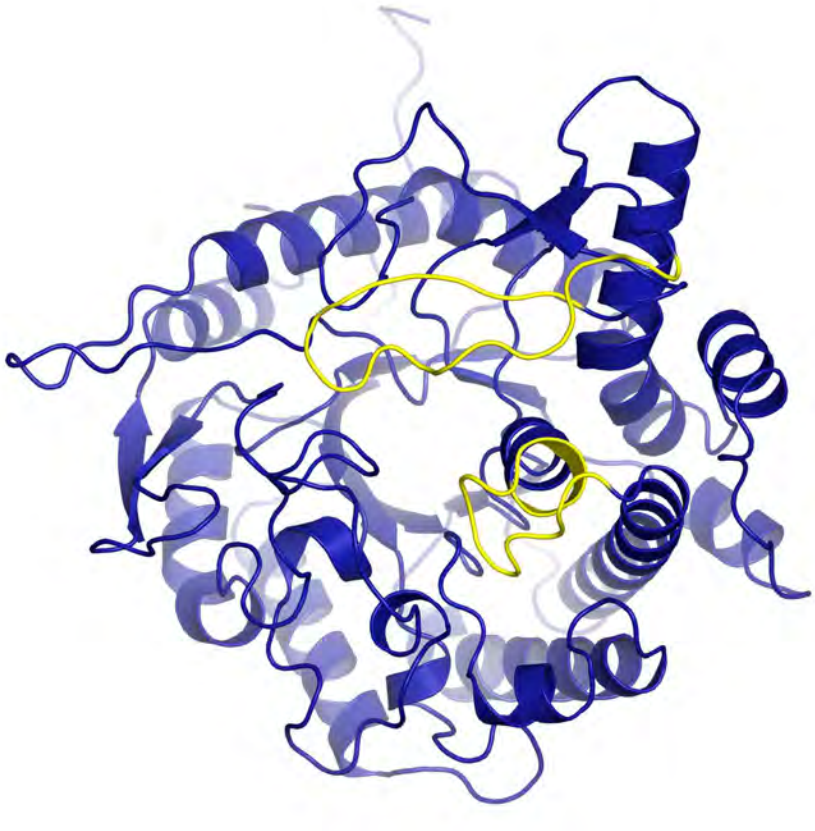


Figure 3.12. *Homology modelling of BlBglB. Loops L4 and L6 of the BlBglB crystallographic structure were missing due to high flexibility and were modelled to obtain a complete enzyme structure.*

3.3.5.1 *In silico docking and MD using enzyme BlBglB*

As the sequence analysis showed *Bl*BglB to be unique by not grouping into any of the GH1 activities, all of the different types of ligands that exist in all crystallographic structures of 6Pβ-glycosidase enzymes were used for docking into *Bl*BglB, in addition to ligands PNP6Pgal and PNP6Pglc. This resulted in a total of 15 ligands all of which docked into the active site (Figure 3.13). However most displayed incorrect orientations (Supplementary Table 3.4).

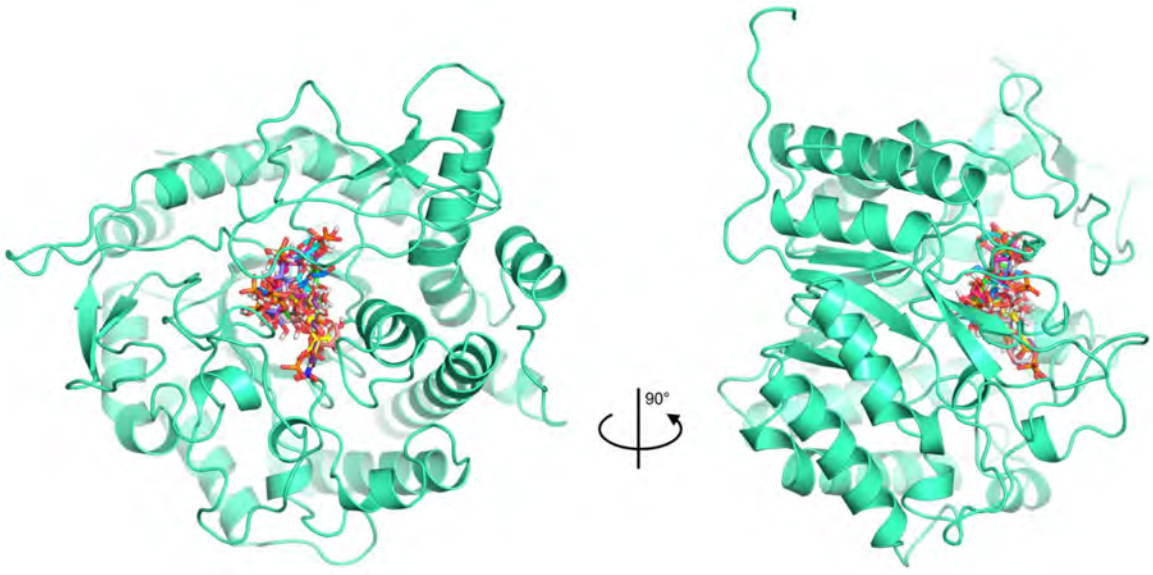


Figure 3.13. *BlBglB with positions of all 15 docked ligands in the active site.*

The ligands PNP6Pgal and PNP6Pglc showed very similar docking positions but they bound to the active site in an upside-down orientation (Supplementary Table 3.4). Despite relatively stable protein and ligand RMSD values over 500 ns of MD simulations, PNP6Pgal and PNP6Pglc were found to be in incorrect orientations within the active site and, therefore, the simulations were extended another 500 ns. However, the ligands did not correct their orientations after 1000 ns. As a result of this, the docking and MD results using these two ligands provided no certainty as to the preference of *Bl*BglB for either galacto- or gluco-configured substrates.

Despite docking in an incorrect orientation, salicin-6P was the only ligand that was stable during MD simulations *and* was found to be in a correct orientation within the active site at 1000 ns (Supplementary Table 3.4). Salicin-6P superimposes relatively well onto the reference ligand (pink) at 1000 ns and interacts with all but one binding residue, namely Asn165. However, binding residue Thr426 forms an unfavourable interaction at 1000 ns. The Thr426 residue-position is the important specificity-inducing residue-position found in loop L8a. Curiously, of all the 69 enzymes in the dataset, only *Bl*BglB and AAX76617.1 have a threonine residue in the *Bl*BglB-Thr426 sequence-position (Trp433 in *Bl*BglC). Despite having an unfavourable interaction at 1000 ns, the Thr426 residue had the most persistent hydrogen bonding of all the residues with salicin-6P (Figure 3.14). Two different atoms of Thr426 form near continuous hydrogen bonds with the O4 ligand-atom throughout the final 20 ns of the MD simulations. A close second is the Arg316 residue, which is located in loop L6. This residue also had two atoms forming many hydrogen bonding interactions with the O5 and O6 ligand-atoms. Also located in loop L6, Gln320 forms hydrogen bonds with the salicin-6P phosphate group for the vast majority of time. In addition, many hydrogen bonds are formed between the salicin-6P phosphate group and the phosphate binding ligands Ser425, Lys432, and Tyr434. Alternatively, Tyr298 forms many hydrogen bonds with the aglycon portion of the ligand. Although there are many hydrogen bonds between salicin-6P and *Bl*BglB, the two catalytic glutamates are missing from the hydrogen bonding, suggesting that the function and activity of enzyme *Bl*BglB is still unclear. Nonetheless, MD simulations indicate a potential affinity for the compound salicin-6P.

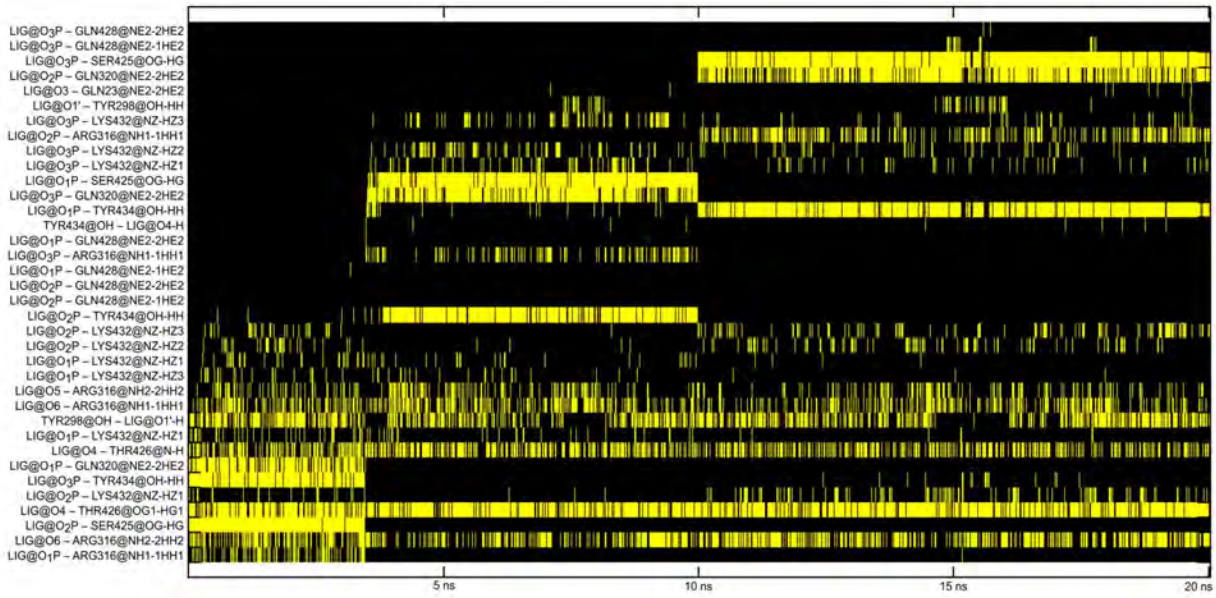


Figure 3.14. *Hydrogen bonding between salicin-6P and BlBglB during the last 20 ns of the MD simulation (980-1000 ns). Hydrogen bonds are shown in yellow.*

3.4 Conclusion

Enzyme *Bl*BglC is very likely a member of the dual-phospho activity, as determined by sequence analysis, MD simulation analysis, and binding free energy calculations. During *Bl*BglC MD simulations in triplicate, the orientations and interactions of the PNP6Pglc ligand were moderately more consistent compared to the PNP6Pgal ligand, although both ligands were stable and showed strong affinity for *Bl*BglC. Residues Gln23 and Trp433 likely have an important role in the broad specificity of dual-phospho activity GH1 enzymes as the two residues bind strongly to the ligand O3 hydroxyl group in the PNP6Pgal-*Bl*BglC complex, but to the ligand O4 hydroxyl group in the PNP6Pglc-*Bl*BglC complex. Also, the *Bl*BglC-His124 residue forms many hydrogen bonds with the PNP6Pgal O3 hydroxyl group but does not form any with the PNP6Pglc triplicates. On the other hand, *Bl*BglC residues Tyr173, Tyr301, Gln302 and Thr321 form hydrogen bonds with PNP6Pglc but not PNP6Pgal. These results contribute important aspects of the broad specificity of dual-phospho activity GH1 enzymes.

In contrast, the activity determination of *Bl*BglB has been elusive, although sequence and structure comparisons hint at dual-phospho activity. *Bl*BglB MD simulations showed significant affinity for salicin-6P, however the two catalytic glutamates were missing from hydrogen bonding.

Chapter 4: Extensive sequence and structure analyses of GH1 activities

4.1 Introduction

With many 3D structures from various activities of bacterial GH1 enzymes, active site residues as well as conserved residues were analysed in terms of differences and similarities between activities in sequence identity and residue-residue interactions.

A conserved and complex network of active site residue-residue interactions was found in all of the 6Pβ-glycosidase activities. However, many differences in interactions between residues were seen when comparing the different activities. The differences likely lead to variation in the active sites of the different activities, thereby contributing to substrate specificity.

Some of the work in this chapter is reproduced from the publication below. All of the work is my own.

> **Wayde Veldman**, Marcelo Liberato, Valquiria Souza, Vitor Almeida, Sandro Marana, Özlem Tastan Bishop, and Igor Polikarpov. “Differences in Gluco and Galacto Substrate-Binding Interactions in a Dual 6Pβ-Glucosidase/6Pβ-Galactosidase Glycoside Hydrolase 1 Enzyme from *Bacillus licheniformis*”. *Journal of Chemical Information and Modeling*. 2021, 61, 9, 4554–4570. DOI: 10.1021/acs.jcim.1c00413.

4.2 Materials and methods

4.2.1 Homology modelling

Dual-phospho activity GH1 crystallographic structures are currently limited to *Bl*BglC (PDB ID 7M1R) and Gan1D (PDB IDs: 5OKB, 5OKE, etc.). Additional and different dual-phospho enzyme structures were needed for use in this section of the study. Two target dual-phospho enzymes with GenBank accessions BAD77499.1 and CAB12135.1 were structurally modelled. PDB structures 7M1R and 5OKB, both chain A, were used as templates. HHpred and PRIMO webservers were used to identify suitable template structures for the models. 100 models per target enzyme were generated using MODELLER version 9.23. The top three models per target enzyme, ranked by normalised z-DOPE score, were evaluated further using PROCHECK, QMEAN and Verify3D webservers. According to the consensus from all three model quality evaluation tools, the best model from each target protein was selected. (Supplementary Table 4.1).

The three 6Pβ-galactosidase enzyme models from Chapter 2 (AAA16450.1, AAD15134.1 and BAA07122.1) were also used for this section of the study.

4.2.2 Sequence and structure comparisons

The MSA containing 69 GH1 sequences from the previous chapter was used to look for conserved residues across all activities as well as conserved residues in individual activities. With DiscoveryStudio, five crystallographic structures and/or model structures per 6Pβ-glycosidase activity were then used to check the interactions of active site residues with other residues, as well as the interactions of conserved residues with other residues. This was done to compare any differences or similarities between each of the activities in terms of their inter-residue interactions which may have an influence on the rigidity/malleability of their structures, spatial positions of their residues, and the enzyme's overall function. The structures included: 7M1R (*Bl*BglC), 5OKB, 5OKE, and models

BAD77499.1 and CAB12135.1 (dual-phospho activity); 3PBG, 4PBG, and models AAA16450.1, AAD15134.1 and BAA07122.1 (6Pβ-galactosidase); 2XHY, 4IPN, 3QOM, 4GPN and 6WGD (6Pβ-glucosidase).

4.3 Results and discussion

4.3.1 Interactions between active site residues in GH1 6Pβ-glycosidase bacterial enzymes

Comparing the dual-phospho, 6Pβ-galactosidase, and 6Pβ-glucosidase activities, enzyme crystallographic structures and models were used to record the active site residue-residue interactions (Figure 4.1; Table 4.1; Supplementary Table 4.2). Across the three activities, patterns of similarities and differences were recorded that were in terms of residue sequence positions, not residue identity. Five 3D structures from the dual-phospho activity, five from the 6Pβ-galactosidase activity, and five from the 6Pβ-glucosidase activity were compared. Interactions were conservatively and strictly recorded (Table 4.1): 1) Similar interactions between all three activities were recorded when all activities showed the interaction in at least four of five of their structures. 2) Similar interactions between two activities were recorded when both activities showed the interaction in at least four of five of their structures and the remaining activity did not show the interaction in any structure. 3) Differences were recorded when all the structures in at least one activity did not show the interaction and at least four of five structures in at least one other activity showed the interaction.

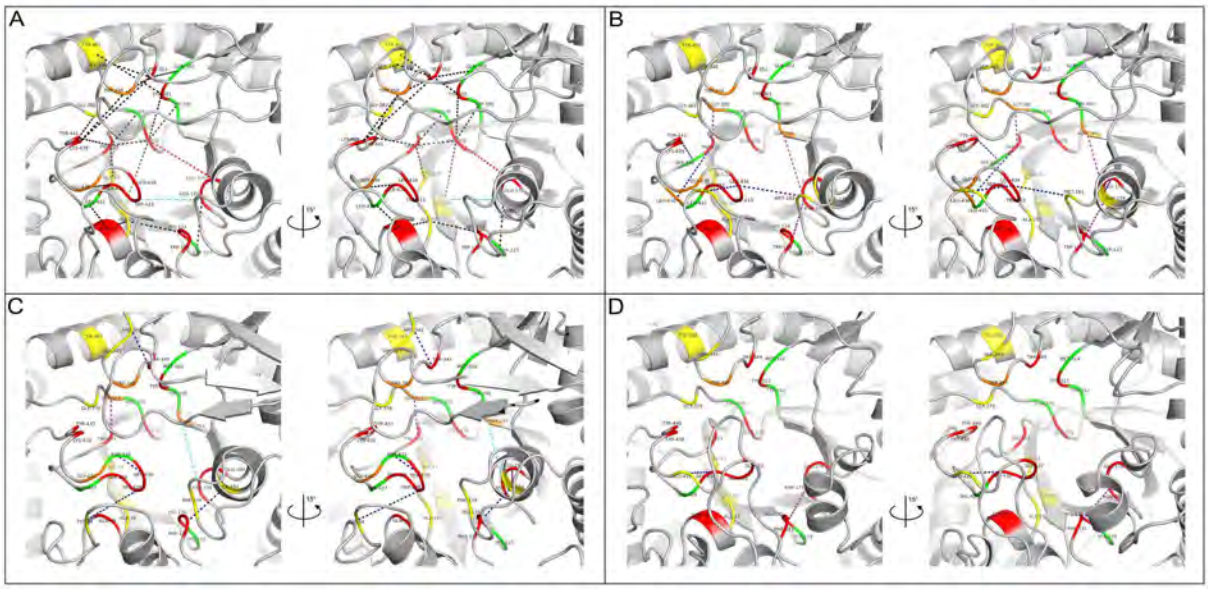


Figure 4.1. *Similarities and differences between the dual-phospho, 6Pβ-galactosidase and 6Pβ-glucosidase activities in terms of residue-residue interactions of active site residues. Red residues are active site residues, green residues are next to active site residues, orange residues are two residues away from active site residues, and yellow residues are further than two residues away. Black dotted lines show interactions that are conserved throughout all three enzyme activities, purple dotted lines show interactions that are shared between two activities, darkblue dotted lines show an interaction unique to only one activity, cyan dotted lines indicate rare case that three of five structures in one activity shows the interaction, red dotted lines indicates a clash. (A) Interactions that are conserved throughout all three enzyme activities. PDB structure 7M1R was used. (B) Dual-phospho activity interactions that are unique or shared with only one other activity. PDB structure 7M1R was used. (C) 6Pβ-galactosidase activity interactions that are unique or shared with one other activity. PDB structure 4PBG was used. (D) 6Pβ-glucosidase activity interactions that are unique or shared with one other activity. PDB structure 4GPN was used. Adapted from Veldman et al. 2021 [114].*

A vast network of active site residue-residue interactions is conserved throughout all the 6Pβ-glycosidase enzyme activities (Figure 4.1 A), especially in one corner of the active site concerning residues Glu378 and Trp425 (*Bl*BglC numbering). Five different residue-residue interactions are formed with catalytic Glu378, regardless of enzyme activity. The same can be said of Trp425 and Trp352, except four instead of five interactions are formed. There exists a link between six different active site residues, namely Tyr301, Trp352, Glu378, Trp425, Lys439 and Tyr441. This link and the interactions around it most likely stabilise the loop regions that contain these residues, helping to retain the positions

of these loops. In addition to ligand-phosphate binding, Lys439 seems to play a role in binding to residues Trp350 and Trp352 – residue Trp352 is known to form stacking interactions with bound ligand in many GH1 bacterial enzymes (PDB IDs: 3QOM, 5OKE, 4IPN). Trp352 forms conserved interactions with four surrounding residues and this residue is conserved in all the enzyme sequences in the study dataset (Supplementary Figure 3.1). Trp350 is conserved in the three activities used in this section (6Pβ-glycosidases; Table 4.1).

The opposite side of the active site displays far fewer conserved interactions. The two catalytic residues, Glu170 and Glu378, exhibit an unfavourable negative-negative interaction (red dotted lines) – the only conserved unfavourable interaction observed. The cyan dotted line between Arg80 and Asn169 represents a mostly-conserved interaction (rare exception), where three of five dual-phospho structures have the interaction and all structures in the remaining activities have the interaction. Interestingly, there exists a conserved link between Leu431, Gln23, Ala20 and His124, followed by a link between Tyr123 and Asn169 – this link could be important for maintaining the position of this part of the active site.

Alternatively, there exists a number of differing interactions between each of the activities mentioned here (Figure 4.1 B-D). These interactions could contribute towards individual activity substrate specificity by causing slightly different overall structure and malleability of the active site. The differing interactions mainly involve two regions: the L8a loop, and the Glu170 (catalytic) and Asn169 residues (*Bl*BglC numbering). The interactions unique to an activity (darkblue dotted lines) mostly involve loop L8a, yet another indication of the importance of this loop in engendering substrate specificity between GH1 enzymes. In the dual-phospho activity, Leu434 (loop L8a) forms unique interactions with Phe49 and Met183, and Leu430 (loop L8a) with active site residue Trp425. In the 6Pβ-galactosidase activity, the special Trp429 residue (loop L8a) interacts with Tyr44 (4PBG numbering), the

only conserved interaction in any one activity related to this residue position. In the 6Pβ-glucosidase activity, Ser430 forms a unique interaction with the loop L8a insert Glu435 (4GPN numbering). Loop L8a seems to be the largest contributing factor to the differences of the active site of GH1 bacterial enzymes.

The dual-phospho activity shares certain interactions with either of the other two activities (purple dotted lines) which may be a reason why the dual-phospho enzymes are capable of hydrolyzing galacto- and gluco-configured substrates. The dual-phospho and 6Pβ-galactosidase activities share the Glu170-Asn174 and Asn169-Asn299 interactions (*Bl*BglC numbering), whereas the dual-phospho and 6Pβ-glucosidase activities share the Trp125-Asn169 interaction. In Table 4.1, more information on the differences can be found. With regards to the conserved Tyr299-Cys375 interaction (4PBG numbering), the 6Pβ-galactosidase activity forms an amide-pi stacked interaction whereas the other two activities form hydrogen bonds. Also, with regards to the conserved Ala13-Trp423 interaction (4GPN numbering), the 6Pβ-glucosidase activity forms an amide-pi stacked interaction whereas the other two activities form hydrogen bonds. Lastly, concerning the shared Trp125-Asn169 interaction (*Bl*BglC numbering) between the dual-phospho and 6Pβ-galactosidase activities, the dual-phospho enzymes form a conventional hydrogen bond whereas the 6Pβ-galactosidase enzymes form a pi-donor hydrogen bond.

Table 4.1. *Interaction similarities and differences between active site residues of the dual-phospho (BlBglC numbering), 6Pβ-galactosidase (4PBG numbering) and 6Pβ-glucosidase (4GPN numbering) activities. A large network of interactions between active site residues is conserved throughout all the enzyme activities, particularly involving residues Glu378 and Trp425 (BlBglC numbering). Each row in the table shows interactions with the same residue positions in the enzyme sequences. Similar interactions between all three activities were recorded when all activities showed the interaction in at least four of five of their structures. Similar interactions between two activities were recorded when both activities showed the interaction in at least four of five of their structures and the remaining activity did not show the interaction in any structure. Differences were recorded when all the structures in at least one activity did not show the interaction and at least four of five structures in at least one other activity showed the interaction. Pink coloured residues are conserved in all 69 enzymes in the study dataset, green coloured residues are conserved*

in the three activities in the table, and orange-coloured residues are conserved in the particular activity in which it belongs. Differences among the activities are highlighted in grey colour.

Active site residues (DUAL / 6PGAL / 6PGLU)	DUAL activity	6PGAL activity	6PGLU activity	Significance / notes
GLN23 / GLN19 / GLN18	ALA20 (Hbond)	ALA16[a] (Hbond)	ALA15 (Hbond)	
	LEU431 (Hbond)	PHE427 (Hbond)	VAL429 (Hbond)	One residue away from active site residue. Two residues away from important TRP433
HIS124 / HIS116 / HIS130	ALA20 (pi-sigma)	ALA16 (pi-sigma)	ALA15 (pi-sigma)	
TRP125 / PHE117 / PHE131	ASN169 (Hbond)		ASN175 (pi-donor Hbond)	Active site residue; pi-donor Hbond (6PGLU)
		ILE164[a]		
N169 / N159 / N175	TRP125 (Hbond)		PHE131 (Pi-donor Hbond)	Active site residue
	TYR123 (Hbond)	HIS115 (Hbond)	SER129 (Hbond)	One residue away from active site residue
	ASN299 (Hbond)	#		Two residues away from active site residue
	#	ARG72 (Hbond)	ARG85 (Hbond)	
GLU170 / GLU160 / GLU176	GLU378	CYS375[a]	GLN375[a]	Catalytic residue; Unfavourable negative-negative interaction
	ASN174 (Hbond)	ILE164[a] (Hbond)		
TYR301 / TYR299 / TYR313	GLU378 (Hbond)	CYS375:ASN376[a]	GLN375 (Hbond)	Catalytic residue; Amide-pi stacked interaction (6PGAL)
	TYR401[a] (Hbond)	TYR397 (Hbond)	TYR398[a] (Hbond)	
W352 / W347 / W347	SER348 (Hbond)	THR343 (Hbond)	SER345 (Hbond)	
	TRP350[a] (C-H bond)	TRP345 (C-H bond)	TRP347 (C-H bond)	Two residues away from active site residue
	LYS439	LYS435	LYS438	Active site residue
	GLN302	MET300[a]	MET314[a]	One residue away from active site residue
		ARG342[a] (C-H bond)		
GLU378 / CYS375 (Mut) / GLN375 (Mut)	ARG80 (Hbond)	ARG72 (Hbond)	ARG85 (Hbond)	
	TYR300 (Hbond)	TYR298 (Hbond)	TYR312 (Hbond)	One residue away from active site residue
	TRP425	TRP421	TRP423	Active site residue
	TYR301 (Hbond)	TYR299:ASN376[a]	TYR313 (Hbond)	Active site residue; Amide-pi stacked interaction (6PGAL)
	GLU170	GLU160[a]	GLU176[a]	Catalytic residue; Unfavourable negative-negative interaction
TRP425 / TRP421 / TRP423	GLU378	CYS375	GLN375	Catalytic residue
	ASN379 (Hbond)	ASN376 (Hbond)	ASN376 (Hbond)	One residue away from catalytic residue
	ALA18	ALA14	ALA13:VAL14[a]	Amide-pi stacked interaction (6PGLU)
	TYR441	TYR437	TYR440	Active site residue
	GLY380 (C-H bond)	GLY377 (C-H bond)		Two residues away from catalytic residue
	LEU430[a]		*	Two residues away from active site residue
SER432 / SER428 / SER430	GLY436[a] (Hbond)	GLY432 (Hbond)	GLY434[a] (Hbond)	
			GLU435 (Hbond)	Insert (6PGLU)
W433 / W429 / A431		TYR44[a]	*	
L434 / S430 / G432	PHE49[a] (Hbond)			

	MET183 (Pi-alkyl)	*		
		ASN431[a] (Hbond)		One residue away from active site residue
LYS439 / LYS435 / LYS430	TRP350 (x3 bonds)	TRP345 (x3 bonds)	TRP347 (x2 bonds)	At least 2 bonds
	TRP352	TRP347 (x2 bonds)	TRP349[a] (x2 bonds)	
TYR441 / TYR437 / TYR440	LEU381:GLY382	LEU378:GLY379	PHE378:GLY379	Amide-pi stacked interaction
	TRP425	TRP421	TRP423	Active site residue
	SER426 (Hbond)			One residue away from active site residue

* One structure within activity has residue interaction (rare exception)
# Three structures within activity have residue interaction (rare exception)
[a] Four out of five structures within activity show residue interaction

4.3.2 Conserved residue sequence comparisons between GH1 bacterial enzymes

The multiple sequence alignment containing the 69 sequences from the characterised bacterial GH1 enzymes (Supplementary Figure 3.1) was analysed in order to record the highly conserved residues in each activity. The major and meaningful differences as well as similarities of residue identity between the activities at the same residue positions were compared (Table 4.2; Supplementary Table 4.3).

As expected, the conserved residues across all activities are active site residues, with the exception of Asn379 (*Bl*BglC numbering) which is the residue that follows the catalytic Glu378 residue in the sequence. Some active site residues are not conserved across all activities. For instance, the three phosphate-binding residues have a different identity in the β-glucosidase activity as these enzymes do not catalyse substrates with a phosphate group. With regards to the 6Pβ-glycosidase activities, all active site residues are conserved except for Trp125 and Trp433 (*Bl*BglC numbering). The Trp433 residue-position is the specificity-inducing residue-position that differentiates between gluco- and galacto-configured ligands in the 6Pβ-glucosidase activity – this is why this residue is never a tryptophan in the 6Pβ-glucosidase activity but always a tryptophan in the other activities. However, it is not known why the residue in the Trp125 residue-position (*Bl*BglC numbering) is different among the activities. The reason could be the unique inter active

site interactions from the previous section (Table 4.1). The dual-phospho tryptophan forms a hydrogen bond with the active site residue Asn169, whereas the analogous 6Pβ-glucosidase residue forms a pi-donor with the same active site residue, and the analogous 6Pβ-galactosidase residue forms a hydrophobic interaction with a different residue in a different sequence position (Ile164; 4PBG numbering). The split between either phenylalanine or tyrosine in this position in the 6Pβ-glucosidase activity (Phe131; 4GPN numbering) suggests this active site residue is not as important for enzyme function as compared to the other activities where the residue in this position is conserved. However, phenylalanine is nonpolar whereas tyrosine is polar – the type of residue in this position in a 6Pβ-glucosidase enzyme may then depend on the particular substrate it hydrolyses. All three residue identities in this position across the activities (tryptophan, tyrosine, and phenylalanine) have hydrophobic side chains and are aromatic. The residue position forms hydrophobic interactions with substrate, and therefore the residue identity here may not have too much influence. Crystallographic structures from each of the activities were superimposed to compare the spatial positions of the residues in this position (Figure 4.2). Each of the respective residue types cluster well together, except one of the phenylalanines from the 6Pβ-glucosidase activity (PDB ID 2XHY; brown colour) that superimposes well onto the tyrosines. The mix of phenylalanine and tyrosine in this position in the 6Pβ-glucosidase activity has no effect on the inter active site interactions of this activity, as the interaction is conserved (Figure 4.1; Phe131-Asn175), and even shares this conserved interaction with the dual-phospho activity (Trp125-Asn169).

Table 4.2. *Major and meaningful differences (blue residues) and similarities of residue identity between the activities at certain sequence positions. Rows of active site residue positions are highlighted light red.*

***Bl*BglC res no.**	**Dual-phospho**	**6Pβ-Galactosidase**	**6Pβ-Glucosidase**	**β-Glucosidase**
23	100% Q	100% Q	100% Q	100% Q

123	100% Y	100% H	84% S / 16% C	94% Y
124	100% H	100% H	100% H	100% H
125	100% W	100% F	63% F / 37% Y	97% W
169	100% N	100% N	100% N	100% N
170	100% E	100% E	100% E	97% E
205	100% N	100% H	100% S	97% H
299	100% N	88% N / 12% D	100% S	100% N
301	100% Y	100% Y	100% Y	100% Y
350	100% W	100% W	100% W	mix M I T F
352	100% W	100% W	100% W	100% W
377	100% T	100% T	100% V	94% T
378	100% E	100% E	100% E	100% E
379	100% N	100% N	100% N	100% N
425	100% W	100% W	100% W	100% W
426	100% S	100% S	100% G	100% S
431	100% L	100% F	100% V	91% F / 9% L
432	100% S	100% S	95% S	100% E
433	100% W	100% W	mix A F M	100% W

439	100% K	100% K	100% K	79% K / 15% M / 6% Q
441	100% Y	100% Y	100% Y	100% F

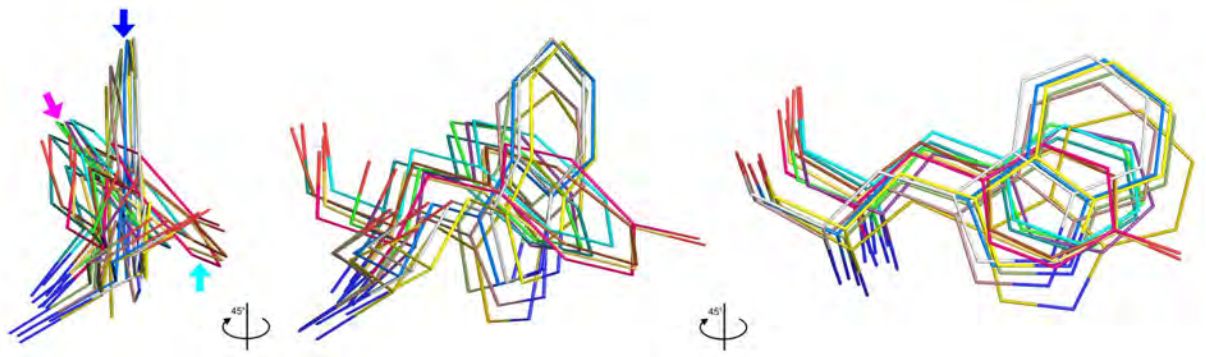


Figure 4.2. *Superposition of residues in the substrate-binding Trp125 residue (BlBglC numbering) position from different activities. The residue in this position is a conserved tryptophan in the dual-phospho and β-glucosidase activities, a conserved phenylalanine in the 6Pβ-galactosidase activity, and a mix between phenylalanine (63%) and tyrosine (37%) in the 6Pβ-glucosidase activity. Each of the different residue types cluster well together, except one of the phenylalanines from the 6Pβ-glucosidase activity (PDB ID 2XHY; brown colour) that superimposes well onto the tyrosines. The blue arrow shows tryptophan residues from the dual-phospho (right of blue arrow) and β-glucosidase (left of blue arrow) activities. The purple arrow shows all phenylalanine residues except one (PDB ID 2XHY; brown colour) that grouped with the tyrosine residues (cyan arrow). Crystallographic structures used are 7M1R, 5OKB, 5OKE (dual-phospho), 3PBG, 4PBG (6Pβ-galactosidase), 2XHY, 3QOM, 4GPN, 4IPN, 6WGD (6Pβ-glucosidase), and 2O9T, 3AHX, 3W53 (β-glucosidase).*

There are also non-active site residue positions where at least one activity has a conserved residue that is different to the other activities (Table 4.2) and may cause slight differences in function between the activities. Most of these residues are next to, or close to, active site residue positions in the sequence. To obtain information about why the residues in these positions differ between the activities, crystallographic structures and models were used to record the residue interactions with other residues of the enzyme,

using DiscoveryStudio (Figure 4.3 & Table 4.3). As this study is focussed mainly on the 6Pβ-glycosidase activities, only they will be analysed from here on.

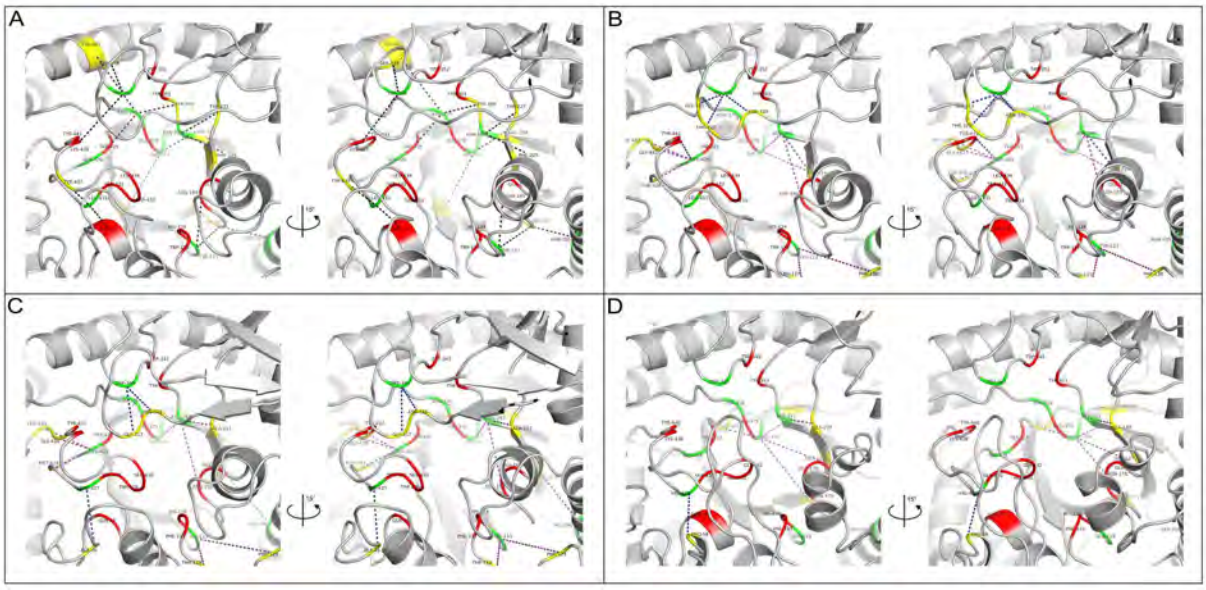


Figure 4.3. *Similarities and differences between the dual-phospho, 6Pβ-galactosidase and 6Pβ-glucosidase activities in terms of residue-residue interactions of non-active site conserved residues; interactions are related to residue sequence positions. Red residues are active site residues, green residues are conserved non-active site residue positions, and yellow residues show interacting residues. Black dotted lines show interactions that are conserved throughout all three enzyme activities; purple dotted lines show interactions that are shared between two activities; darkblue dotted lines show an interaction unique to only one activity. (A) Interactions that are conserved throughout all three enzyme activities. PDB structure 7M1R was used. (B) Dual-phospho activity interactions that are unique or shared with only one other activity. PDB structure 7M1R was used. (C) 6Pβ-galactosidase activity interactions that are unique or shared with one other activity. PDB structure 4PBG was used. (D) 6Pβ-glucosidase activity interactions that are unique or shared with one other activity. PDB structure 4GPN was used.*

In the Tyr123 residue position (*Bl*BglC numbering), although the residue identities differ between the activities, they all still share interactions with Ser167 and active site residue Asn169. However, the dual-phospho and 6Pβ-galactosidase activities share the Tyr123-Leu127 and Tyr123-Phe138 interactions which are missing in the 6Pβ-glucosidase activity.

The Asn205 residue position (*Bl*BglC numbering) has a different residue identity that is conserved in each of the respective activities; however, all three of the activities show a conserved interaction whereby the Asn205 residue position indirectly links with the Tyr123

residue position through the Ser176 residue position. On the other hand, only the 6Pβ-galactosidase activity exhibits a His196-Val214 interaction (4PBG numbering).

The Asn299 residue position (*Bl*BglC numbering) has a conserved interaction across all of the activities with residue positions Phe225 and Tyr227. In contrast, Asn299 forms an interaction with Ser224 that is only seen in the dual-phospho activity. Additionally, the dual-phospho and 6Pβ-galactosidase activities share the Asn299-Asn169 and Asn299-Thr377 interactions, while the 6Pβ-galactosidase and 6Pβ-glucosidase activities share the Asn297-Ala217 interaction (4PBG numbering).

The residue in the Thr377 sequence position (*Bl*BglC numbering) is next to the catalytic Glu378 residue and has conserved interactions with Arg80 and Val298 in all three of the activities. An interaction of Thr377-Ser224 (*Bl*BglC numbering) is shared between the dual-phospho and 6Pβ-glucosidase activities, whereas a Thr374-Tyr372 interaction (4PBG numbering) is shared between the 6Pβ-galactosidase and 6Pβ-glucosidase activities. The only unique interactions involving this residue position is with regards to the 6Pβ-glucosidase activity, namely Val374-Met172 and Val374-Thr421 (4GPN numbering), meaning this residue has two extra conserved interactions in the 6Pβ-glucosidase activity as compared to the other activities, which may be important.

A serine residue is conserved in the Ser426 (*Bl*BglC numbering) sequence position except in the 6Pβ-glucosidase activity where this residue is a glycine. Ser426 forms a unique interaction among the activities with active site residue Tyr441 in the dual-phospho activity. Three other Ser426 interactions are shared between the dual-phospho and 6Pβ-galactosidase activities, namely with residues Thr428, Gly442, and Phe443 (*Bl*BglC numbering). No interactions are shared with the 6Pβ-glucosidase activity glycine.

In the Leu431 residue position (*Bl*BglC numbering), although the residue identities are all different between the activities, they all still share interactions with Tyr437 and active site

residue Gln23. Unique interactions include Phe427-Ala46 in the 6Pβ-galactosidase activity and Val429- Pro58 in the 6Pβ-glucosidase activity.

Table 4.3. *Interaction similarities and differences of conserved residues that are not active site residues, between the dual-phospho (BlBgIC numbering), 6Pβ-galactosidase (4PBG numbering) and 6Pβ-glucosidase (4GPN numbering) activities. Each row in the table shows interactions with the same residue positions in the enzyme sequences. Similar interactions between all three activities were recorded when all activities showed the interaction in at least four of five of their structures. Similar interactions between two activities were recorded when both activities showed the interaction in at least four of five of their structures and the remaining activity did not show the interaction in any structure. Differences were recorded when all of the structures in at least one activity did not show the interaction and at least four of five structures in at least one other activity showed the interaction. Pink coloured residues are conserved in all 69 enzymes in the study dataset, green coloured residues are conserved in the three activities in the table, and orange-coloured residues are conserved in the particular activity in which it belongs. Differences among the activities are highlighted in grey colour.*

Active site residues (DUAL / 6PGAL / 6PGLU)	DUAL activity	6PGAL activity	6PGLU activity	Significance / notes
TYR123 / HIS115 / SER129	SER167	THR157	THR173	
	LEU127[a]	THR119		
	ASN169	ASN159	ASN175	Active site residue
	PHE138[a]	PHE129		
ASN205 / HIS196 / SER218	SER167	THR157[#]	THR173	
		VAL214		
ASN299 / ASN297 / SER311	TYR227	LEU218	MET240	
	PHE225	HIS216	VAL238	
	ASN169	ASN159[#]		Active site residue
	SER224[a]			
		ALA217[a]	ALA239[a]	
	THR377[#]	THR374		
TRP350 / TRP345 / TRP347	SER348	THR343	SER345[a]	
	ASN320			
		GLY317		
	LYS439 x3	LYS435 x3	LYS438 x2	Active site residue
		ASN316[a]		
	THR322:GLY323			Amide-pi stacked interaction
THR377 / THR374 / VAL374	ARG80[a]	ARG72	ARG85	
	SER224[a]		MET237	
	VAL298	ILE296[a]	PHE310	
	ASN299[#]	ASN297		

			MET172[a]	
		TYR372[#]	PHE372[a]	
			THR421[a]	
ASN379 / ASN376 / ASN376	TYR300	TYR298	TYR312[a]	Residue next to active site residue
	TYR401[a]	TYR397[a]	TYR420	
	TYR422	TYR418		
	TRP425	TRP421	TRP423	Active site residue
SER426 / SER422 / GLY424	TYR441			Active site residue
	THR428[a]	MET424		
	PHE443	LEU439		
	GLY442[a]	GLY438		
LEU431 / PHE427 / VAL429	GLN23	GLN19	GLN18	Active site residue
	TYR437	TYR433	MET436[#]	
		ALA46	ASN59*	
			PRO58	

* One structure within activity has residue interaction (rare exception)
[#] Three structures within activity have residue interaction (rare exception)
[a] Four out of five structures within activity show residue interaction

The only non-active site residues that are conserved across all 6Pβ-glycosidase activities are Trp350 and Asn379 (*Bl*BglC numbering).

Trp350 is two residues away from active site residue Trp352 in the sequence. Across all the activities, Trp350 forms interactions with Ser348 and active site residue Lys439 (*Bl*BglC numbering). On the other hand, this residue forms unique interactions with the long L6 loop in the dual-phospho (Asn320 & Thr322:Gly323) and 6Pβ-galactosidase activities (Asn316 & Gly317) which is not seen in the 6Pβ-glucosidase activity as this activity does not have the longer L6 loop that is thought to act as a gate to the active site. The L6 loop is slightly longer in the dual-phospho activity (Figure 4.4) and the asparagine residues are not in the same sequence position in the dual-phospho and 6Pβ-galactosidase activities. Nonetheless, the residue is in a similar location in terms of 3D structure.

Asn379 follows the catalytic Glu378 residue in the sequence. In all of the activities this residue forms conserved interactions with three different residues, namely Tyr300, Tyr401, and active site residue Trp425 (*Bl*BglC numbering). Alternatively, the dual-phospho and 6Pβ-galactosidase activities share the Asn379-Tyr422 interaction which is missing in the 6Pβ-glucosidase activity.

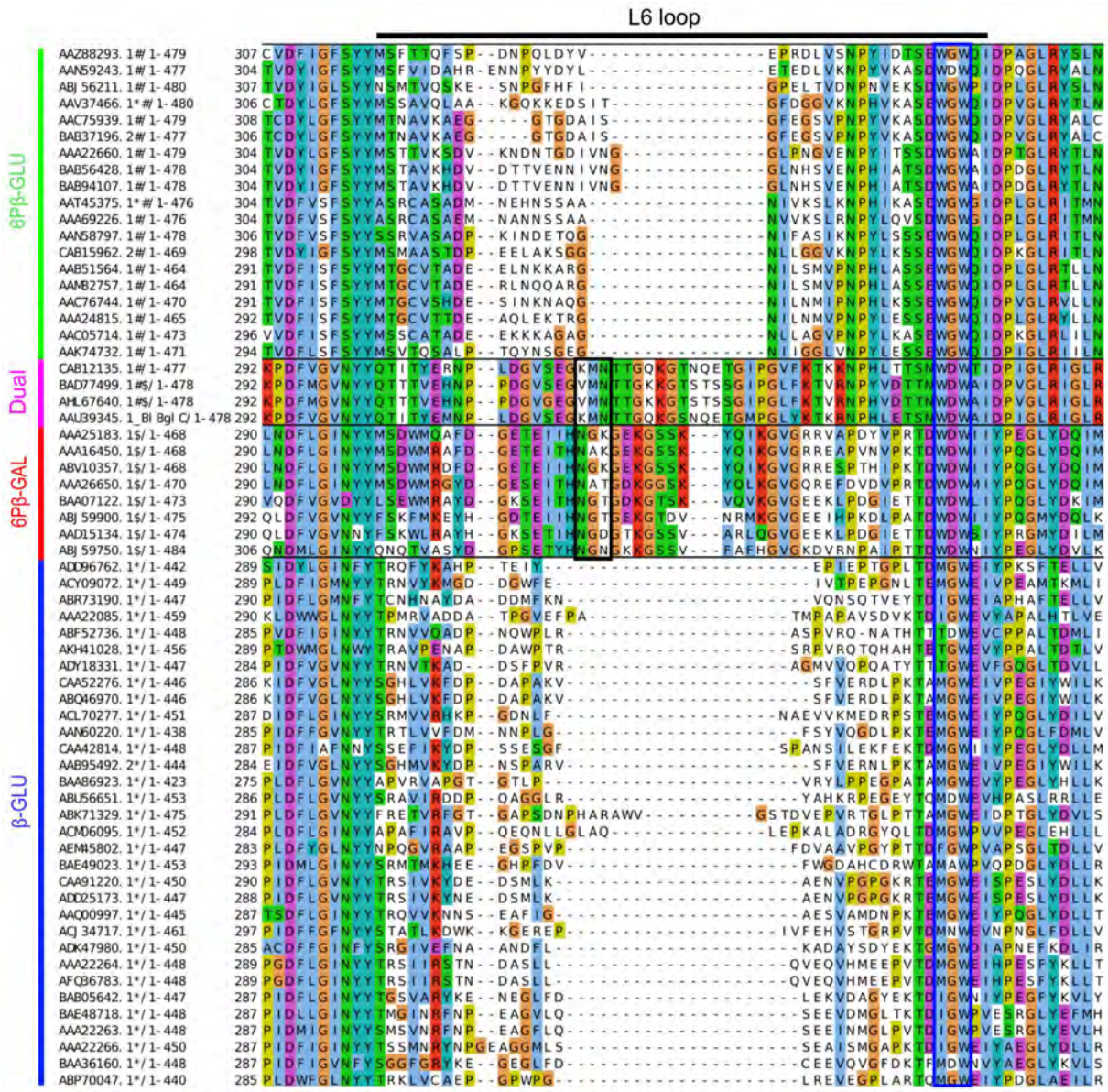


Figure 4.4. *MSA section containing the L6 loop region of the GH1 enzyme sequences. The L6 loop is slightly longer in the dual-phospho activity. The blue block shows residues Trp350 and Trp352 (BlBglC numbering) and the black block shows the different positions of the asparagine residues of the dual-phospho and 6Pβ-galactosidase activities that bind to residue Trp350 (BlBglC numbering). The L6 loop is thought to control access to the active site in the dual-phospho and 6Pβ-galactosidase activities.*

4.4 Conclusion

Using multiple 3D structures from various bacterial GH1 enzymes, active site residues as well as conserved residues (across all activities and individual activities) were analysed in terms of differences and similarities in sequence identity and residue-residue interactions.

A conserved and complex network of active site residue-residue interactions was found in all of the 6Pβ-glycosidase activities, particularly in one corner of the active site relating to residues Glu378 and Trp425 (*Bl*BglC numbering). There is a conserved link of interactions between the Tyr301, Trp352, Glu378, Trp425, Lys439 and Tyr441 active site residues. This link and the interactions around it most likely stabilise the loop regions that contain these residues, helping to retain the positions of these loops. In addition to ligand-phosphate binding, Lys439 seems to play a role in binding to residues Trp350 and Trp352, with at least two Trp350-Lys439 bonds being conserved in all activities. In the dual-phospho and 6Pβ-galactosidase activities, Trp350 forms conserved interactions with the long L6 loop that is thought to act as a gate to the active site. The Trp350 residue may bond to the L6 loop when a closed L6 loop gate is required.

There are several different interactions when comparing the activities that could cause slight variations of overall structure and malleability of the active site resulting in a separate substrate specificity. The differing active site interactions mainly involve two regions: the L8a loop, and the Glu170 (catalytic) and Asn169 residues (*Bl*BglC numbering). The L8a loop contains the Trp433 residue-position that differentiates between gluco- and galacto-configured ligands in the 6Pβ-glucosidase activity. The Trp433 residue-position is one of two 6Pβ-glycosidase active site residue positions that are not conserved, the other being the Trp125 residue-position (*Bl*BglC numbering). The reason for the differing residue identity of the Trp125 residue-position could be the unique inter active site interactions that the residue makes in each of the different activities.

Many differences and similarities in conserved interactions between residues were discovered among the different GH1 activities. These interactions likely have a role in forming slightly nuanced active sites depending on the GH1 activity, thereby contributing to substrate specificity.

Chapter 5: Conclusions and Future Work

Three new GH1 enzyme crystallographic structures from the bacterium *B. licheniformis* were obtained from collaborators. These are the first GH1 crystallographic structures from *B. licheniformis* ever determined. As the active sites of these structures were absent of ligand, *in silico* docking and MD simulations were performed to provide evidence for their GH1 activities and substrate specificities. First though, the amino acid sequences of all known characterised bacterial GH1 enzymes were retrieved from the CAZy database and compared to the sequences of the three new *B. licheniformis* crystallographic structures to obtain putative enzyme activity. Sequence identity, phylogeny, and sequence motif analyses provided evidence of the putative dual-phospho activity of *Bl*BglC and 6Pβ-glucosidase activity of *Bl*BglH. The activity of *Bl*BglB at this stage was more difficult to obtain as the sequence was found to be unique among the characterised bacterial GH1 enzyme sequences. However, the sequence is most similar to the dual-phospho activity and a structural comparison using the DALI server outputted dual-phospho Gan1D (PDB ID 5OKB) as the top hit for both the *Bl*BglB crystal structure and model.

As all three enzymes were shown to be putative 6Pβ-glycosidase activity enzymes, much of the thesis focuses on the overall analysis and comparison of the 6Pβ-glucosidase, 6Pβ-galactosidase, and dual-phospho activities that make up the 6Pβ-glycosidases. In the thesis literature review, the 6Pβ-glycosidase active site residues are identified through consensus of binding interactions using all known 6Pβ-glycosidase PDB structures containing ligands possessing all three ligand groups (phosphate, glycon, and aglycon). An exception is PDB 4PBG whose ligand is absent of the aglycon group; however, this is the only existing 6Pβ-galactosidase structure with a ligand. Thirteen residue sequence positions, some with differing identities, interact with the ligand in all of the 6Pβ-glycosidase PDB structures. This information was used to analyse and compare the

activities, binding interactions, and ligand specificities of the new GH1 enzyme crystallographic structures received from our collaborators.

Although GH1 members share high structural similarity, it was found that the secondary structure of the 6Pβ-glucosidase L8a loop is different to the other activities which most likely contributes to the ability of this activity to differentiate between gluco- and galacto-configured ligands. Comparing the 6Pβ-glucosidase and 6Pβ-galactosidase activities, it was seen that the 6Pβ-glucosidase L8b loop is longer and forms additional interactions with the L8a loop likely leading to increased L8 loop rigidity which prevents the displacement of residue Ala423 ensuring a steric clash with galacto-configured ligands. During MD simulations with *Bl*BglH the PNP6Pglc ligand showed sustained favourable binding while the PNP6Pgal ligand was unstable, providing evidence of the 6Pβ-glucosidase activity of *Bl*BglH. Also, the favourable binding of PNP6Pglc stabilised the loops that surround the active site. This was the earliest known study to simulate a 6Pβ-glycosidase GH1 enzyme using molecular dynamics, as far as we are aware.

During *Bl*BglC MD simulations in triplicate, both PNP6Pgal and PNP6Pglc ligands were stable and showed strong affinity for *Bl*BglC. However, the orientations and interactions of PNP6Pglc were moderately more consistent. It is plausible that the Gln23 and Trp433 residue positions (*Bl*BglC numbering) have an important role in engendering the broad specificity of dual-phospho activity GH1 enzymes, as the two residues bind strongly to the ligand O3 hydroxyl group in the PNP6Pgal-*Bl*BglC complex but to the ligand O4 hydroxyl group in the PNP6Pglc-*Bl*BglC complex. This too was seen in the dual-phospho crystallographic structures of the Gan1D enzyme. Also, the *Bl*BglC-His124 residue forms many hydrogen bonds with the PNP6Pgal O3 hydroxyl group but forms none with PNP6Pglc. Alternatively, *Bl*BglC residues Tyr173, Tyr301, Gln302 and Thr321 form hydrogen bonds with PNP6Pglc but not PNP6Pgal. The findings present important information of the broad specificity of dual-phospho activity GH1 enzymes.

The activity determination of the unique *Bl*BglB enzyme was difficult. The docked PNP6Pgal and PNP6Pglc ligands bound to the *Bl*BglB active site in an incorrect orientation and remained in an incorrect orientation at an extended MD simulation of 1000 ns. Therefore, it was unclear as to the preference of *Bl*BglB for either galacto- or gluco-configured substrates. Fourteen different 6Pβ-glycosidase ligands were tested against *Bl*BglB. Despite docking in an incorrect orientation, salicin-6P was the only ligand that was stable during MD simulations *and* was found to be in a correct orientation within the active site at 1000 ns. At 1000 ns salicin-6P interacts with all binding residues except Asn165; however, during the final 20 ns of the simulation the two catalytic glutamates were missing from hydrogen bonding. The function and activity of enzyme *Bl*BglB remains uncertain.

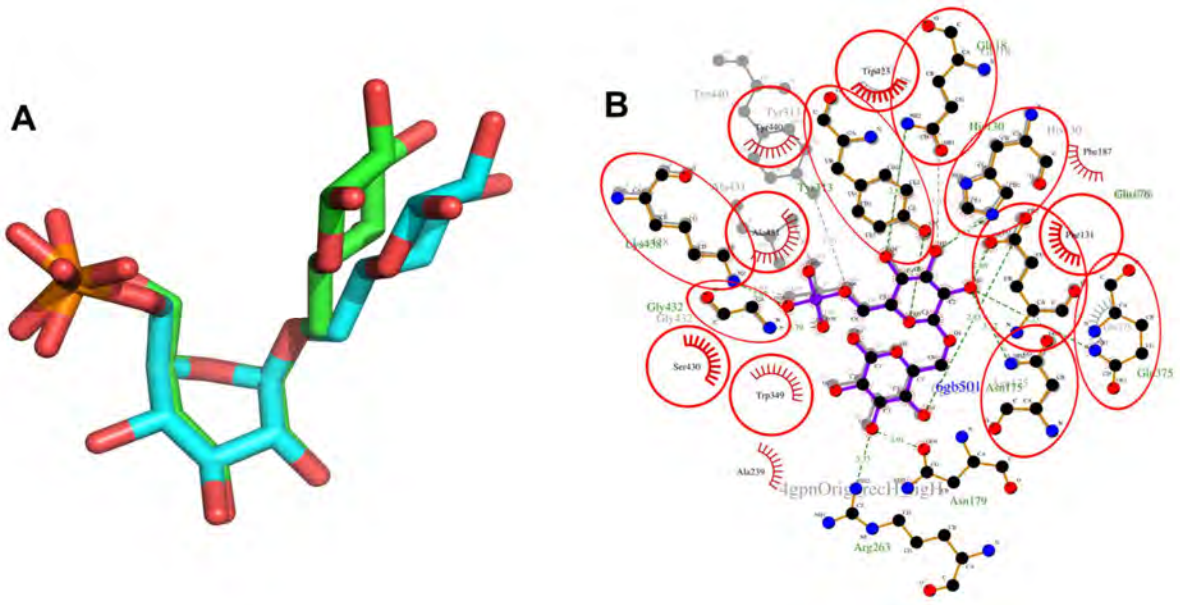
Bacterial GH1 enzyme sequences and structures from various activities were meticulously compared. Active site residues, as well as conserved residues (across all activities or individual activities), were analysed in terms of differences and similarities in sequence identity and residue-residue interactions. A large network of conserved interactions among active site and conserved residues was discovered. Also, there exists a number of differing interactions when comparing each of the activities which could contribute towards individual activity substrate specificity by causing slightly different overall structure and malleability of the active site.

As a disclaimer, we must mention that the evidence obtained from molecular modelling is not definitive, but indicative, and must be validated by additional methods.

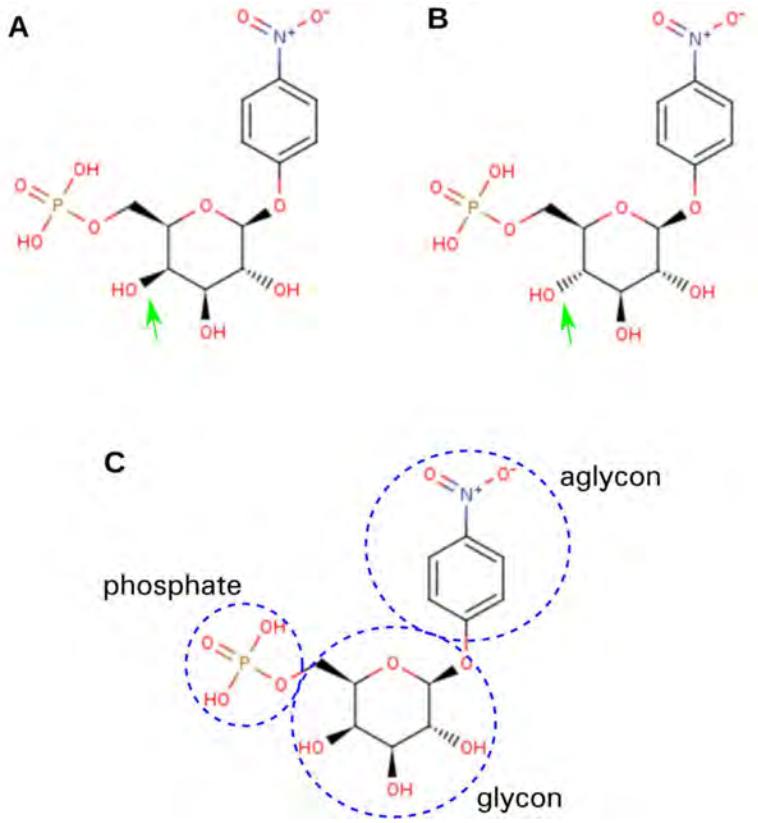
Possible future work could include computationally mutating specific residues in structures of a particular activity and running MD simulations to observe the effects. The mutated residues could be binding residues, or residues important for structure and function such as the Trp350 residue that forms conserved interactions with the long L6 loop that is thought to act as a gate to the active site. In the future, additional enzyme structures could be modelled from GH1 activities that are lacking crystallographic structures like the 6Pβ-

galactosidase and dual-phospho activities. These models could help strengthen GH1 analyses related to structure and dynamics. A possible idea for the future could be the running of very long simulations using enzymes of the 6Pβ-galactosidase and dual-phospho activities with a ligand that is a product to observe the potential release of the ligand and the movement and binding of the L6 loop as well as the ligand. As GH1 activities differ mostly in loop regions, future work could include the use of machine-learning models, trained only on the number of residues in the active site loops as features, to discriminate between GH1 activities. Also in the future, the Gromacs rerun feature could be used to obtain energy fluctuation for the interactions during the trajectories; this could allow the ranking of residues by importance, or the analysis of the steric clash between the 6Pβ-glucosidase loop L8a specificity-inducing residue and the axial OH4 of a galactoside ligand. Finally, relative binding free energy (RBFE) calculations could be performed to obtain more accurate binding free energy comparisons of the protein-ligand complexes.

Supplementary data

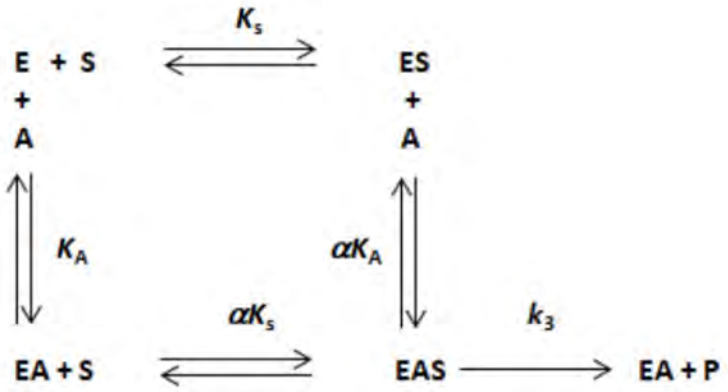


Supplementary Figure 2.1. Docking validation (A) Crystalised ligand from PDB ID 4GPN superimposed with the same ligand docked into the 4GPN protein using Vina-Carb. An RMSD of 0.43 Å was achieved. Crystalised ligand – green. Docked ligand – cyan. (B) Comparison of the protein residue interactions using the same ligands. The program Ligplot was used.



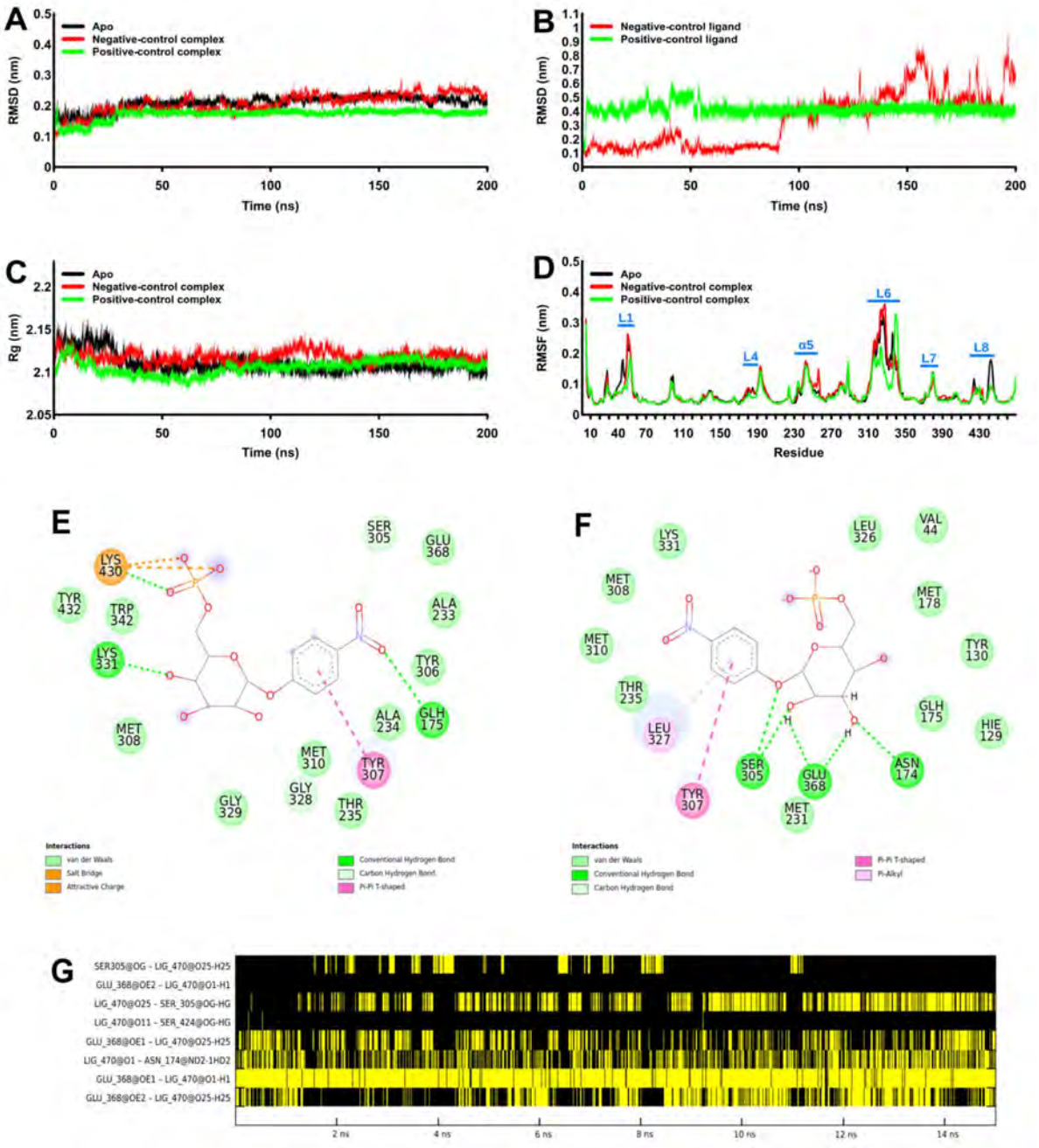
Supplementary Figure 2.2. 2D representation of the control ligands. The only difference between the ligands is the configuration of the O4 hydroxyl group (axial vs equatorial), shown using green arrows. (A) Negative-control ligand p-Nitrophenyl-beta-D-galactoside-6-phosphate, (B) positive-control ligand p-Nitrophenyl-beta-D-glucoside-6-phosphate and (C) phospho-glycoside showing the phosphate, glycon and aglycon ligand groups. These ligand groups bind to corresponding binding subsites within the active site of 6Pβ-glycosidase enzymes.

Supplementary Figure 2.3. Multiple sequence alignment containing all 59 GH1 sequences from Chapter 2 aligned with PROMALS3D and viewed with Jalview. Order of the sequences are based on the PROMALS3D alignment. The sequences fall into groups that are consistent throughout the sequence analysis in this study. β-glucosidases (*) – Blue; 6Pβ-glucosidases (#) – Green; 6Pβ-galactosidases ($) – Red; Unique groupings – Yellow.

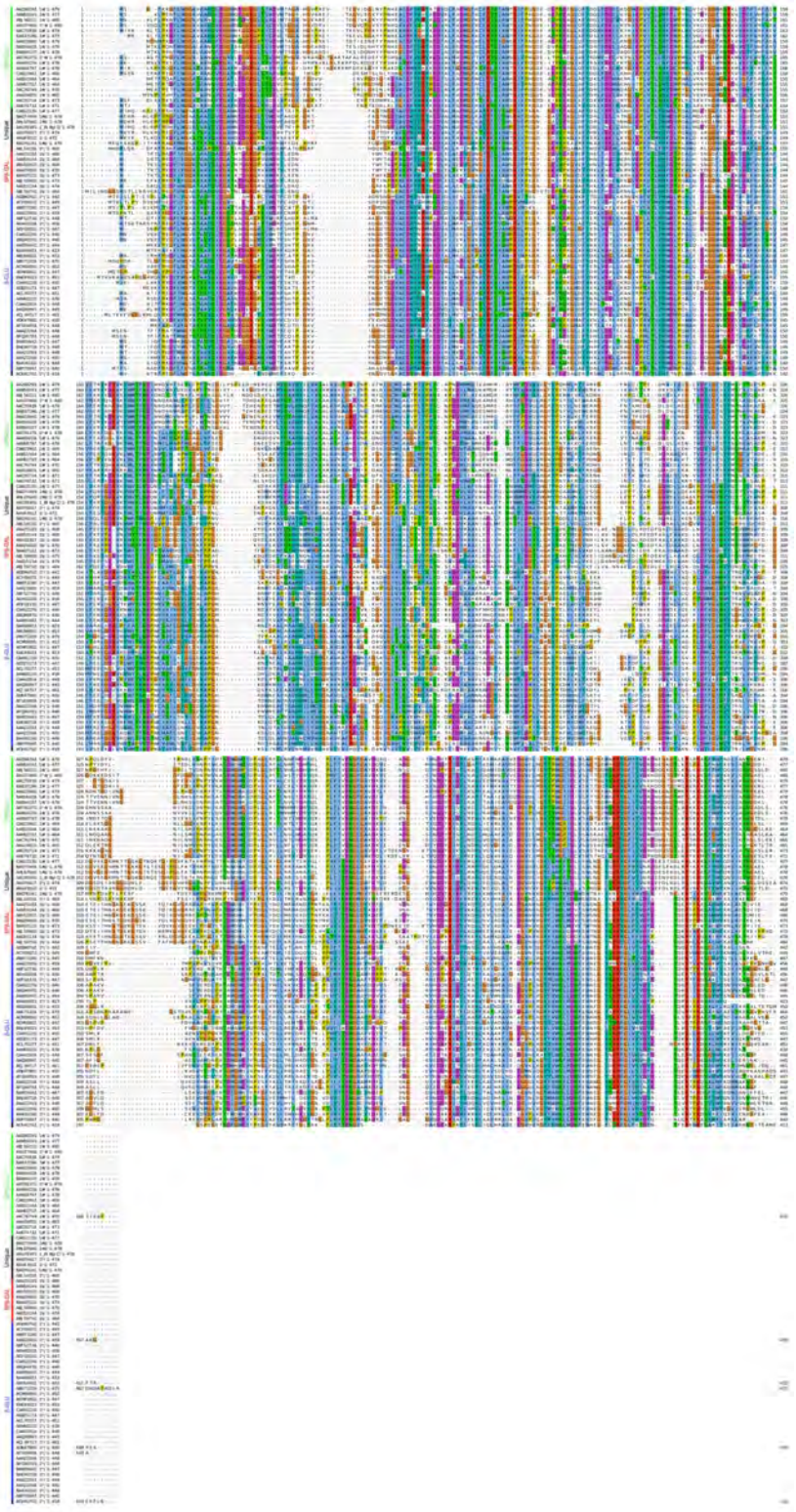


$1/v_0 = \alpha K_s/V_{max}\ (1 + K_A/[A])\ 1/[S] + 1/V_{max}\ (1+\alpha K_A/[A])$

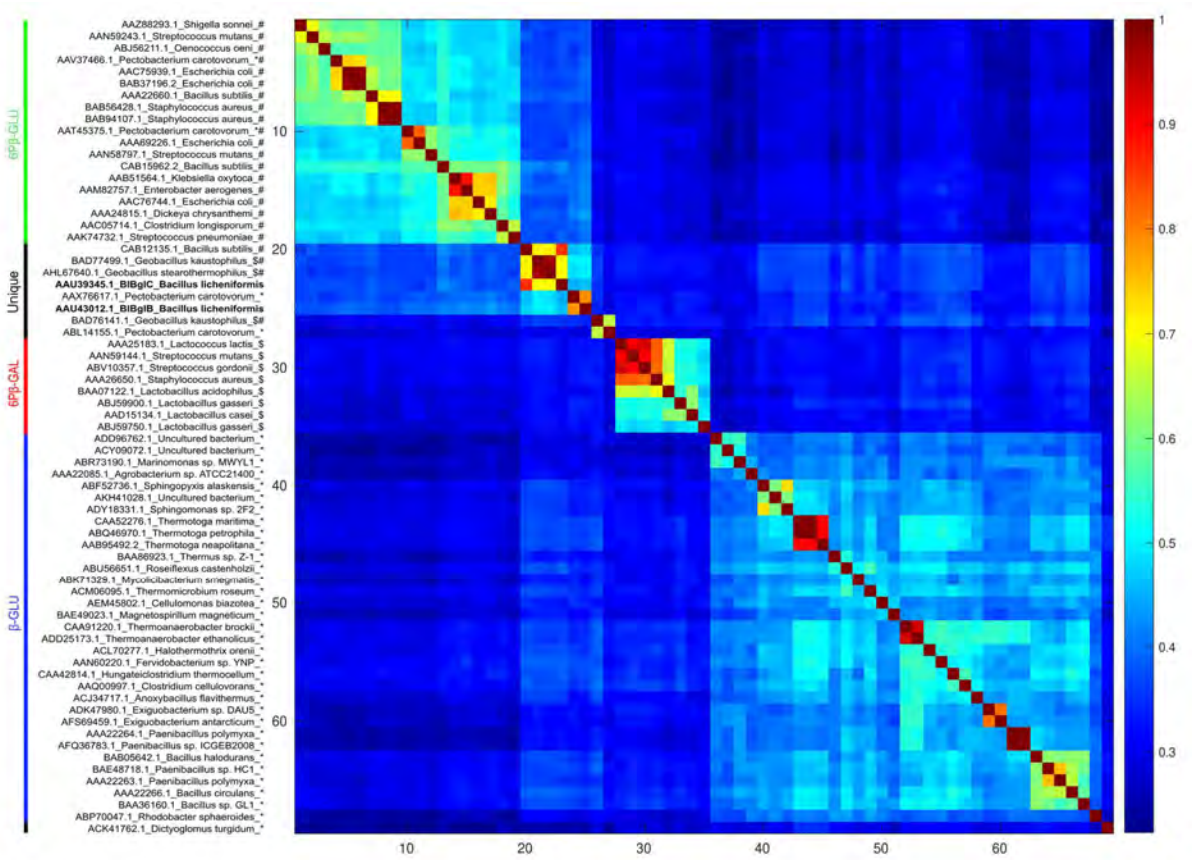
Supplementary Figure 2.4. Kinetic mechanism of the *Bl*BglH activation by phosphate and the development of the corresponding kinetic equation.



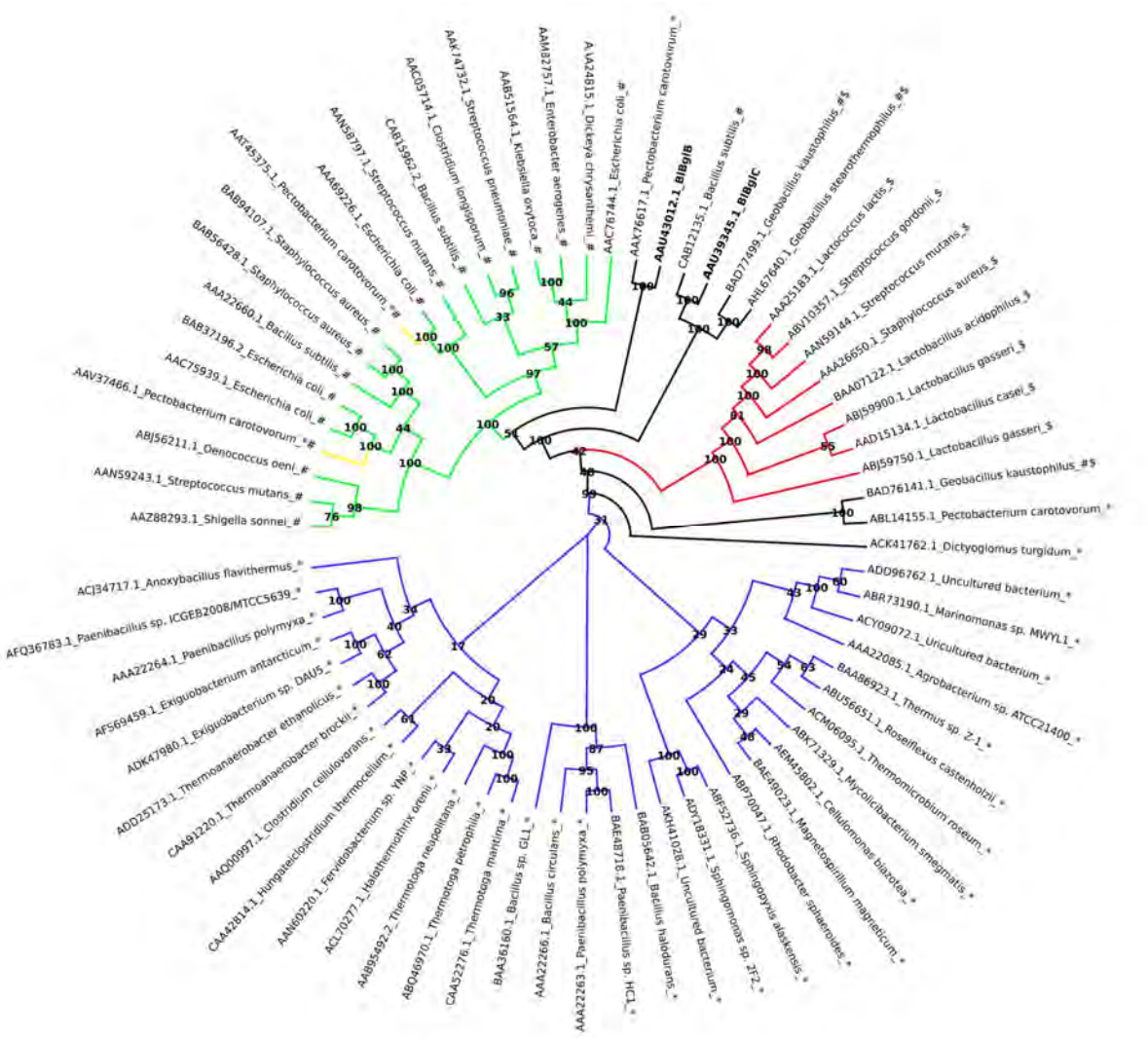
Supplementary Figure 2.5. Duplicated 200 ns MD simulation results of PNP6Pgal-pose2 and PNP6Pglc complexes. (A) Protein backbone RMSD after least square fitting to protein backbone, (B) ligand RMSD after least square fitting to protein backbone, (C) protein radius of gyration, and (D) protein residue RMSF, (E) PNP6Pgal-pose2 complex interactions, (F) PNP6Pglc complex interactions, and (G) hydrogen bonding of the PNP6Pglc ligand-protein complex during the last 15 ns of the MD simulation. Hydrogen bonds are shown in yellow.



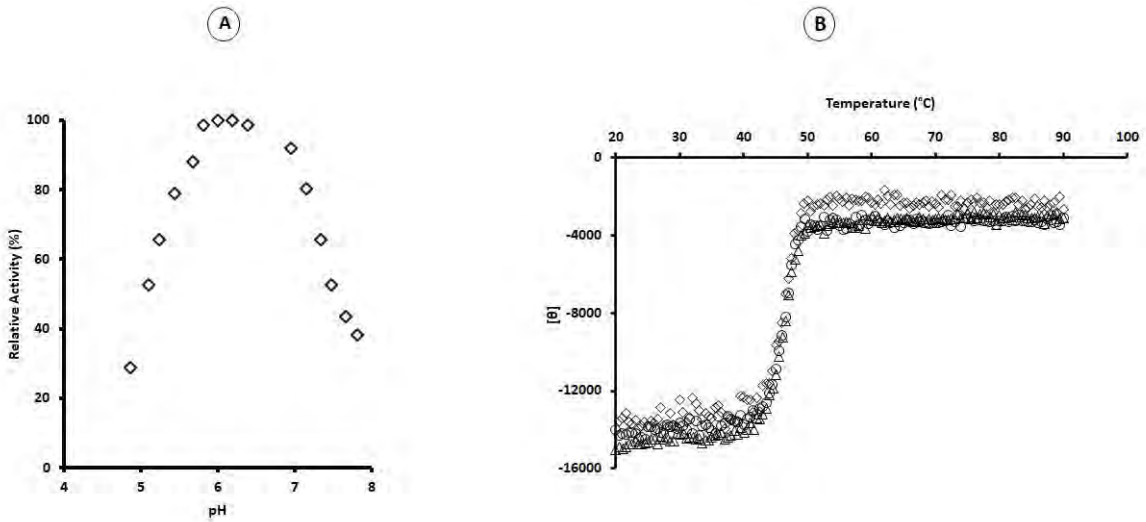
Supplementary Figure 3.1. Multiple sequence alignment containing all 69 GH1 sequences from Chapter 3 aligned with PROMALS3D and viewed with Jalview. Order of the sequences are based on the PROMALS3D alignment. The sequences fall into groups that are consistent throughout the sequence analysis in this study. β-glucosidases (*) – Blue; 6Pβ-glucosidases (#) – Green; 6Pβ-galactosidases ($) – Red; Unique groupings – Black.



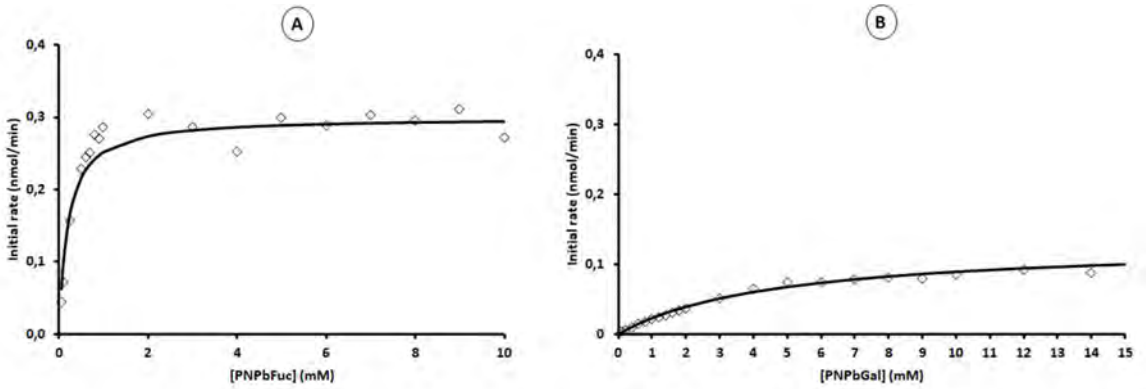
Supplementary Figure 3.2. Sequence identity heatmap showing the pairwise percentage identity between all 69 sequences from Chapter 3. The X and Y axes indicate the 69 different GH1 enzymes. Identity scores are shown as a colour-coded matrix, calculated by comparing every sequence to each other (every sequence vs every sequence). Sequence identity increases from blue to red. 35 β-glucosidases (*), 8 6Pβ-galactosidases ($), 20 6Pβ-glucosidases (#), and 3 dual activity ($#) enzymes are labelled.



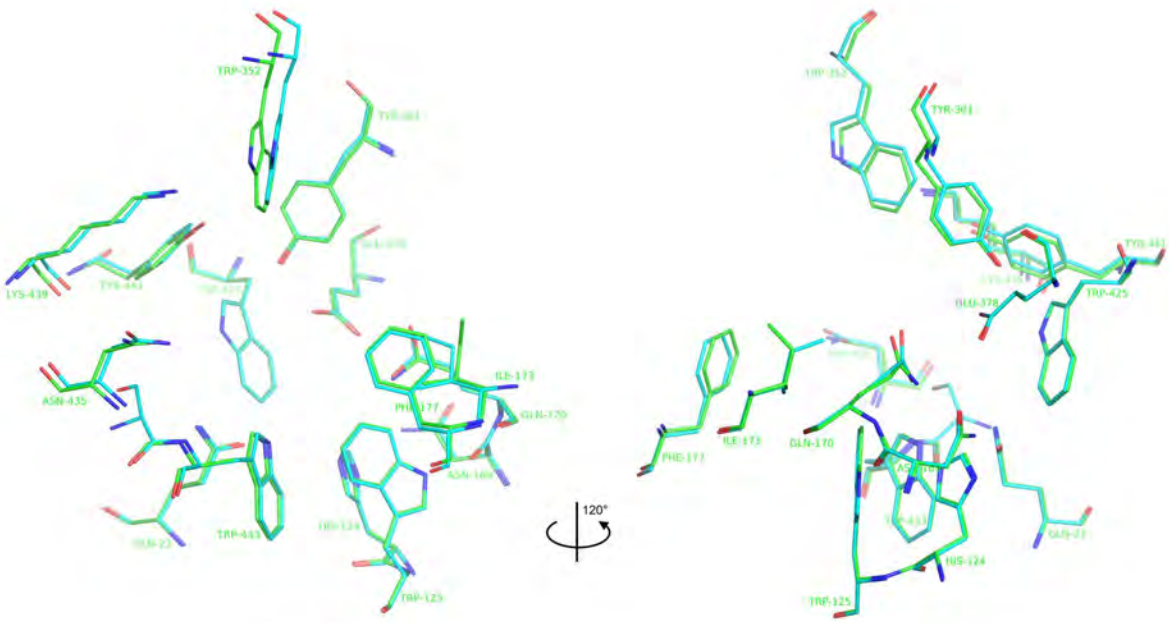
Supplementary Figure 3.3. Maximum likelihood phylogenetic tree consisting of all 69 sequences from Chapter 3, generated using MEGA v7.0.26 program. Branch numbers indicate bootstrap values. Colour code: β-glucosidases (*) – Blue. 6Pβ-galactosidases ($) – Red. 6Pβ-glucosidases (#) – Green. Unique groupings – Black.



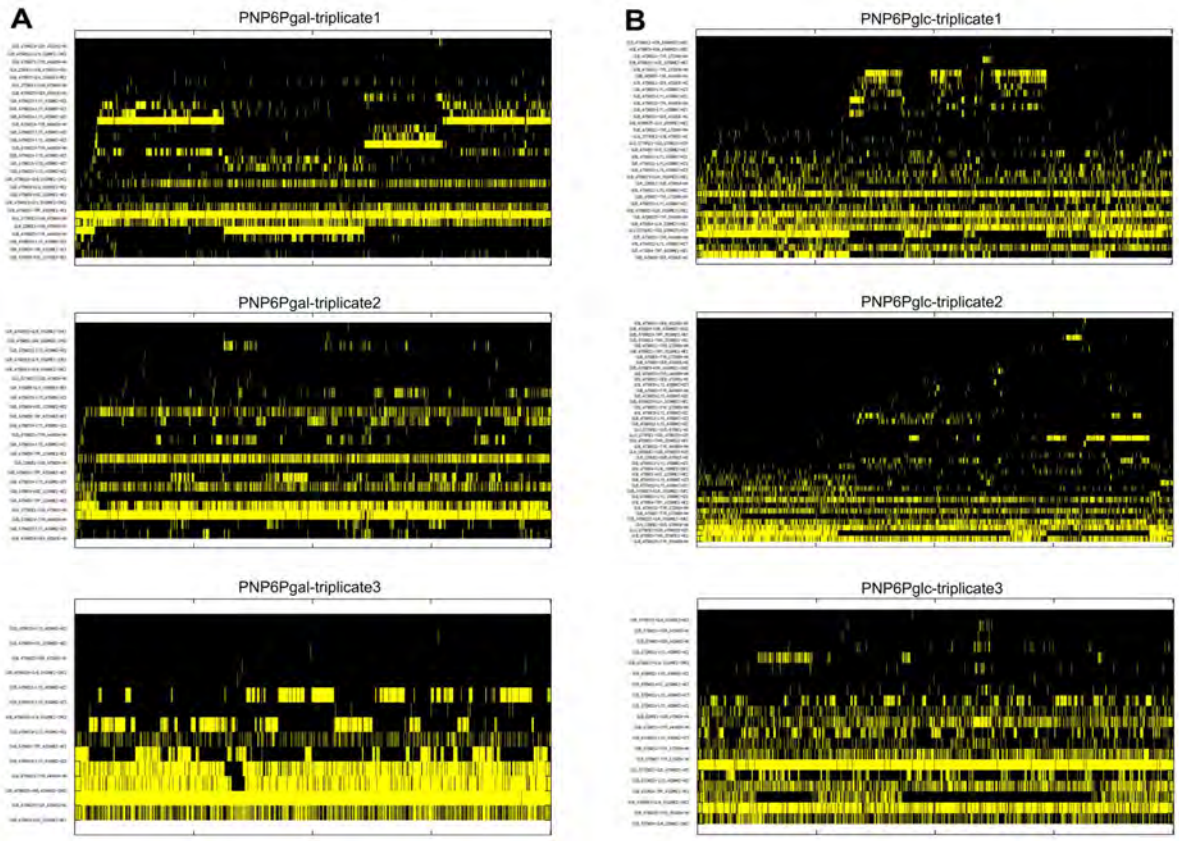
Supplementary Figure 3.4. Determination of the conditions for the enzymatic assays. (A) The pH effect on the *Bl*BglC activity. Assays were performed at 30 °C using 1.5 mM *p*-nitrophenyl β-fucopyranoside prepared in 50 mM citrate-phosphate buffers presenting pH from 4.5 to 8.0. (B) The temperature stability of the *Bl*BglC was probed using circular dichroism. The protein sample was prepared in 10 mM potassium phosphate pH 7. Sample temperature was increased from 20 to 90 °C (0.5 °C/min) and readings (θ) collected at 208 (◊), 215 (○) and 222 (Δ) nm.



Supplementary Figure 3.5. Determination of the substrate concentration effect on the activity of the *Bl*BglC. Assays were performed at 30 °C using *p*-nitrophenyl β-fucopyranoside (PNPβfuc; A) and *p*-nitrophenyl β-glucopyranoside (PNPβgal; B) prepared in 100 mM citrate-phosphate buffer pH 6.



Supplementary Figure 3.6. Comparison of Gan1D apo protein (PDB ID 5OKB) to the ligand-bound Gan1D protein (PDB ID 5OKE). It is seen that ligand binding does not cause significant changes in the side chain positions of the Gan1D residues.



Supplementary Figure 3.7. Hydrogen bonding of each of the triplicates of the PNP6Pgal and PNP6Pglc ligand-protein complexes during the final 20 ns of the MD simulations. Hydrogen bonds are shown in yellow.

Supplementary Table 2.1. GenBank accession number, species and PDB ID (if available) of the 59 GH1 sequences used in Chapter 2.

GenBank accession no.	Species	PDB ID, if available
Enzyme *Bl*BglH		
AAU43012.1	*Bacillus licheniformis*	6WGD
β-glucosidases		
ADK47980.1	*Exiguobacterium sp. DAU5*	
AFS69459.1	*Exiguobacterium antarcticum*	5DT7
CAA52276.1	*Thermotoga maritima*	1OIF
ACL70277.1	*Halothermothrix orenii*	4PTV
ABP66702.1	*Caldicellulosiruptor saccharolyticus*	
ACI19973.1	*Dictyoglomus thermophilum*	
CAA91220.1	*Thermoanaerobacter brockii*	
ADD25173.1	*Thermoanaerobacter ethanolicus*	
AAN60220.1	*Fervidobacterium sp. YNP*	
AAQ00997.1	*Clostridium cellulovorans*	3AHX
BAB05642.1	*Bacillus halodurans*	
AAA22266.1	*Bacillus circulans*	1QOX
AAA22263.1	*Paenibacillus polymyxa*	1BGG
BAE48718.1	*Paenibacillus sp. HC1*	
BAA36160.1	*Bacillus sp.*	
ACM66669.1	*Micrococcus antarcticus*	3W53
AAA22264.1	*Paenibacillus polymyxa*	2O9T
ABR73190.1	*Marinomonas sp. MWYL1*	
CAA82733.1	*Streptomyces sp. QM-B814*	1GNX
AAF37730.1	*Thermobifida fusca*	
AAA22085.1	*Agrobacterium sp.*	
ADY18331.1	*Sphingomonas sp. 2F2*	
ABL14155.1	*Pectobacterium carotovorum*	
BAA29440.1	*Pyrococcus horikoshii*	1VFF
ACK41762.1	*Dictyoglomus turgidum*	
AAG59862.1	*Sphingomonas paucimobilis*	

CAA52344.1	*Streptomyces rochei*	
CCA60742.1	*Fervidobacterium islandicum*	
CAA56282.1	*Pantoea agglomerans*	
6-phospho-β-glucosidases		
CAB12135.1	*Bacillus subtilis*	
AAK34377.1	*Streptococcus pyogenes*	5FOO
BAA20086.1	*Lactobacillus gasseri*	
AAN59243.1	*Streptococcus mutans*	4GPN
AAZ88293.1	*Shigella sonnei*	
ABJ56211.1	*Oenococcus oeni*	
CAD63073.1	*Lactobacillus plantarum*	3QOM
AAC75939.1	*Escherichia coli*	2XHY
AAA22660.1	*Bacillus subtilis*	
BAB94107.1	*Staphylococcus aureus*	
AAC05714.1	*Clostridium longisporum*	
AAK74732.1	*Streptococcus pneumoniae*	4IPN
AAA24815.1	*Dickeya chrysanthemi*	
AAM82757.1	*Enterobacter aerogenes*	
AAN58797.1	*Streptococcus mutans*	
AAA69226.1	*Escherichia coli*	
BAA25004.1	*Lactobacillus gasseri*	
AAD28227.1	*Enterococcus faecium*	
6-phospho-β-galactosidases		
AAA25183.1	*Lactococcus lactis*	4PBG
ABV10357.1	*Streptococcus gordonii*	
AAA16450.1	*Streptococcus mutans*	
AAA26650.1	*Staphylococcus aureus*	
BAA07122.1	*Lactobacillus acidophilus*	
ABJ59750.1	*Lactobacillus gasseri*	
AAD15134.1	*Lactobacillus casei*	
ABJ59900.1	*Lactobacillus gasseri*	
dual activity 6-phospho-β-galactosidase/6-phospho-β-glucosidases		
BAD77499.1	*Geobacillus kaustophilus*	

AHL67640.1	*Geobacillus stearothermophilus*	5OKG
BAD76141.1	*Geobacillus kaustophilus*	

Supplementary Table 2.2. Motifs discovered in the 59 sequences from Chapter 2 using MEME v4.9.1. The regular expression shows the motif sequence, where the residues in square brackets represent the different residues that can exist in the various sequences in that particular position. The "width" column shows the length of the motif and the "sites" column represents the number sequences that possess the motif.

Motif no.	Regular expression	Width	Sites	E-value	*Bl*BglH sequence location
1	R[FT]SI[AS]W[PST]RIF	10	58	4.1e-363	84-93
2	VTL[YSH]H[FW][DE][LM]P[QL]	10	59	1.2e-315	125-134
3	L[WF]G[GT]ATA[AS][YH]Q	10	59	1.2e-313	13-22
4	[CI]D[FH]YHRY[KE]ED	10	54	1.8e-298	62-71
5	VKYW[IM]TFNE[PI]	10	59	6.2e-295	167-176
6	P[LI][YF]I[TV]ENG[AL][GA]	10	59	3.4e-285	363-372
7	KR[YF]G[LFI][IV]YVD[YFR]	10	59	2.2e-282	430-439
8	GY[FT]VW[SG][LC][MI]D[NL]	10	59	1.6e-284	411-420
9	D[FY][LI]G[IVF][NS]YY[TM][SR]	10	58	8.8e-249	300-309
10	R[TYI][PK]K[DK]S[FA]YWY	10	56	8.7e-240	449-458
11	[WM][GD]WEI[DY]P[EQ]GL	10	56	5.2e-229	340-349
12	[VI][HN]DDYRIDYL	10	58	8.2e-205	383-392
13	EGA[TWY]NE[DG]G[KR]G	10	53	2.0e-198	24-33
14	E[PV]N[EPQ][KAE]GL[DA]FY	10	52	4.8e-187	100-109
15	H[HN][LE][LM][VL]A[HS][AG][LKR]A	10	54	1.7e-193	206-215
16	GGWLNR[EKD][TV][IV]D	10	55	1.5e-173	141-150
17	L[MF][KA]E[LM]G[FVL][KN]A[YF]	10	57	2.3e-152	74-83
18	DE[LC]L[KA][NY]GIEP	10	57	8.0e-159	114-123
19	[FY]VR[YF]AE[TV][CV]F[EK]	10	56	3.1e-128	154-161
20	D[PV]xxKGKYP[EQ]	10	55	2.4e-112	262-271
21	SIWD[TRV]F[ASC][HK]TP	10	25	1.6e-109	
22	H[LI][EK][AQ][VA]H[KR]AIE	10	59	5.7e-101	395-404
23	F[ES]W[AS][EN]GY	7	38	2.3e-089	
24	VYP[AYL][ST]C[SK][PE]ED	10	44	1.5e-084	45-54
25	P[DK]G[KQ][IV]G[IC][MTV]L[NA]	10	59	1.2e-103	224-233

26	**[GS]YLLGV[HF][AP]PG**	10	36	4.9e-080	
27	**KK[FL]PK[DG]F**	7	59	7.0e-061	6-12
28	**[VT]F[NK]G[DH][NT]GD[VI]A**	10	24	1.8e-055	
29	**VSA[GST]T[GA][EQ]M[SK]**	9	17	3.1e-054	421-429
30	**[KQ][EK]VI[AE][ST]NGE[EDS]**	10	28	4.8e-052	459-468
31	**VKNPY[LV]K[TSA]S[DE]**	10	21	1.6e-046	330-339
32	**GD[ML][ED][IT]IS[QT]PI**	10	15	6.6e-039	
33	**LN[RW][LFI][YT]DRY[QH][KL]**	10	20	1.1e-043	353-362
34	**D[DN][ED]G[KN]G[TS]L[KE]**	9	19	3.2e-024	440-448
35	**[GN][KE]YYPNHEA**	9	10	9.5e-023	
36	**D[LQ][LI][MK]R[LVI]K[KNR][DE]Y**	10	20	3.1e-020	
37	**AD[GA][FL]xNRWFL**	10	23	9.2e-021	
38	**S[VI]AD[VI]MTAG[ARS]**	10	10	3.4e-020	35-44
39	**[HM]N[GT][TK]G[EK]KG[ST]S**	10	11	2.4e-025	
40	**[ND]RE[EA][VI]MYQAA**	10	12	2.1e-012	196-205
41	**WC[SA][AS]F**	5	12	3.7e-010	
42	**VAWDKYLE[DE]N**	10	6	4.6e-010	
43	**[HN]G[VK][PA]REIT[DAK]G**	10	8	5.6e-010	
44	**WMRA[FY]DG[EK][ST]E**	10	6	3.7e-009	
45	**[MF]A[QE][KE]AM[QR][KR]RY**	10	6	4.3e-008	
46	**P[KNQ]GDE**	5	18	6.4e-008	94-98
47	**EDIIHNKFIL**	10	4	1.3e-007	
48	**N[MS][SV]LHAPF[TM][GS]**	10	7	1.4e-007	177-186
49	**YD[LF][AE]KVFQSH**	10	4	3.0e-007	
50	**PTKYP[YF]DP[ENS]N**	10	4	8.1e-007	
51	**DGV[DN][LV]**	5	45	1.3e-006	405-409
52	**RYKDK**	5	15	1.9e-006	162-166
53	**D[RKL][KE][ILD]LKE[GN]TV**	10	15	2.7e-006	290-299
54	**[AFY][KV]DK[VL][ET][EA]DG[KRS]**	10	19	8.3e-010	373-382
55	**KYQ[IL]KGVG[RQ]R**	10	4	1.2e-005	
56	**YWY[TE]AEP**	7	6	2.8e-004	
57	**DMM[EN]L[FY]SK[IY][IV]**	10	4	6.1e-004	
58	**T[ML][EA][GA]V[NKQ][HD]IL**	9	5	8.4e-004	
59	**[FAW]TNSG[IV][VL][YL][KT][EN]**	10	7	1.6e-003	
60	**VKA[FY]RE[LM]**	7	12	2.8e-002	
61	**KKLAET[QH][IE]I[EP]**	10	7	3.2e-002	

62	**MPGR[KR]MN[PV]**	8	3	4.3e-002	
63	**P[AE]D[VI][RG]AAEL**	9	5	4.5e-002	
64	**KD[MR]KR[MF]YEAN**	10	3	6.3e-002	
65	**[NL][TK][DN]LQ[LT][AS][ITV][ND]V**	10	5	3.1e-001	
66	**VK[IL]GHAI**	7	6	1.7e+001	216-222
67	**NNQ[ATV]N[FY][QES][ES][DS]**	9	3	1.8e+001	
68	**[AQ][AI]L[DE]AAKDLN**	10	4	2.9e+001	
69	**HIFKYWERKA**	10	2	3.8e+001	
70	**TTVE[HN]N[PI][PV][DN]G**	10	3	2.4e+001	

Supplementary Table 2.3. Homology model quality evaluation scores of the 6Pβ-galactosidase models

Structure	**z-DOPE**	**PROCHECK**	**QMEAN**	**Verify3D**
Templates				
1PBG	-2.26	86.80%	1.02	81.06%
4PBG	-2.19	84.80%	0.94	99.79%
3W53	-2.44	90.30%	0.96	97.87%
Models				
AAA16450.1	-2.04	92.20%	-0.12	100.00%
AAD15134.1	-1.73	92.30%	-1.13	94.94%
BAA07122.1	-1.91	91.10%	-1.19	96.19%

Supplementary Table 2.4. Duplicated binding free energy results, using duplicated MD simulations. All values are in kJ/mol. The final 15 ns of the MD simulations were used (185-200 ns).

Complex	**Van der Waal's**	**Electrostatic**	**Polar solvation**	**Solvent Accessible Surface Area (apolar)**	**Total binding energy**
PNP6Pgal-pose2 – *B*/BglH	-112.11 ± 0.33	246.03 ± 1.70	312.73 ± 1.99	-15.76 ± 0.03	**430.89 ± 0.86**
PNP6Pglc – *B*/BglH	-132.42 ± 0.33	-108.05 ± 0.48	183.02 ± 0.37	-16.93 ± 0.02	**-74.38 ± 0.29**

Supplementary Table 3.1. GenBank accession number, species and PDB code (if available) of the 69 GH1 sequences used in Chapter 3.

GenBank accession no.	Species	PDB ID, if available
Enzyme *Bl*BglC/*Bl*BglB		
AAU39345.1 (*Bl*BglC)	*Bacillus licheniformis*	7M1R
AAU43012.1 (*Bl*BglB)	*Bacillus licheniformis*	
β-glucosidases		
ABL14155.1	*Pectobacterium carotovorum*	
ADD96762.1	Uncultured bacterium	
ACY09072.1	Uncultured bacterium	
ABR73190.1	*Marinomonas sp. MWYL1*	
AAA22085.1	*Agrobacterium sp.ATCC21400*	
ABF52736.1	*Sphingopyxis alaskensis*	
AKH41028.1	Uncultured bacterium	5GNX, 5GNY
ADY18331.1	*Sphingomonas sp. 2F2*	
CAA52276.1	*Thermotoga maritima*	1OIF
ABQ46970.1	*Thermotoga petrophila*	
AAB95492.2	*Thermotoga neapolitana*	5IDI
BAA86923.1	*Thermus sp. Z-1*	
ABU56651.1	*Roseiflexus castenholzii*	
ABK71329.1	*Mycolicibacterium smegmatis*	
ACM06095.1	*Thermomicrobium roseum*	
AEM45802.1	*Cellulomonas biazotea*	
BAE49023.1	*Magnetospirillum magneticum*	
CAA91220.1	*Thermoanaerobacter brockii*	
ADD25173.1	*Thermoanaerobacter ethanolicus*	
ACL70277.1	*Halothermothrix orenii*	4PTV, 4PTX
AAN60220.1	*Fervidobacterium sp. YNP*	
CAA42814.1	*Hungateiclostridium thermocellum*	5OGZ
AAQ00997.1	*Clostridium cellulovorans*	3AHX
ACJ34717.1	*Anoxybacillus flavithermus*	
ADK47980.1	*Exiguobacterium sp. DAU5*	
AFS69459.1	*Exiguobacterium antarcticum*	5DT7
AAA22264.1	*Paenibacillus polymyxa*	2O9R, 2O9T
AFQ36783.1	*Paenibacillus sp. ICGEB2008/MTCC5639*	

BAB05642.1	*Bacillus halodurans*	
BAE48718.1	*Paenibacillus sp. HC1*	
AAA22263.1	*Paenibacillus polymyxa*	1BGG
AAA22266.1	*Bacillus circulans*	1QOX
BAA36160.1	*Bacillus sp. GL1*	
ABP70047.1	*Rhodobacter sphaeroides*	
ACK41762.1	*Dictyoglomus turgidum*	
AAX76617.1	Pectobacterium carotovorum	
6-phospho-β-glucosidases		
AAZ88293.1	*Shigella sonnei*	
AAN59243.1	*Streptococcus mutans*	4GPN, 4F66, 4F79
ABJ56211.1	*Oenococcus oeni*	
AAC75939.1	*Escherichia coli*	2XHY
BAB37196.2	*Escherichia coli*	
AAA22660.1	*Bacillus subtilis*	
BAB56428.1	*Staphylococcus aureus*	
BAB94107.1	*Staphylococcus aureus*	
AAA69226.1	*Escherichia coli*	
AAN58797.1	*Streptococcus mutans*	
CAB15962.2	*Bacillus subtilis*	
AAB51564.1	*Klebsiella oxytoca*	
AAM82757.1	*Enterobacter aerogenes*	
AAC76744.1	*Escherichia coli*	
AAA24815.1	*Dickeya chrysanthemi*	
AAC05714.1	*Clostridium longisporum*	
AAK74732.1	*Streptococcus pneumoniae*	4IPN
CAB12135.1	*Bacillus subtilis*	
6-phospho-β-galactosidases		
AAA25183.1	*Lactococcus lactis*	4PBG
AAN59144.1	*Streptococcus mutans*	
ABV10357.1	*Streptococcus gordonii*	
AAA26650.1	*Staphylococcus aureus*	
BAA07122.1	*Lactobacillus acidophilus*	
ABJ59900.1	*Lactobacillus gasseri*	
AAD15134.1	*Lactobacillus casei*	
ABJ59750.1	*Lactobacillus gasseri*	

Dual activity 6-phospho-β-galactosidase/6-phospho-β-glucosidases		
BAD77499.1	*Geobacillus kaustophilus*	
AHL67640.1	*Geobacillus stearothermophilus*	5OKE, 5OKB, 5OKK, 5OKR
BAD76141.1	*Geobacillus kaustophilus*	
Dual activity β-galactosidase/6-phospho-β-glucosidases		
AAV37466.1	Pectobacterium carotovorum	
AAT45375.1	Pectobacterium carotovorum	

Supplementary Table 3.2. Motifs discovered in the 69 sequences from Chapter 3 using MEME v4.9.1. The regular expression shows the motif sequence, where the residues in square brackets represent the different residues that can exist in the various sequences in that particular position. The "width" column shows the length of the motif and the "sites" column represents the number sequences that possess the motif.

Motif no.	Regular expression	Width	Sites	E-value	*Bl*BglB sequence location	*Bl*BglC sequence location
1	YR[FT]SI[AS]W[PST]RI	10	69	4.4e-439	76-85	79-88
2	[CI]D[HF]YHRY[KP]ED	10	69	3.0e-393	55-64	58-67
3	VTL[YS]H[WF][DE][LM]PQ	10	69	1.7e-392	117-126	120-129
4	AT[AS][AS]YQ[IV]EGA	10	69	1.6e-384	18-27	18-27
5	KR[YF]G[LIF][VI][YH]V[DN][YRF]	10	69	4.2e-366	432-441	439-448
6	GY[FT]VW[SG]L[MIL]D[NL]	10	69	1.3e-346	414-423	421-430
7	P[LI][YF]I[TV]ENG[AL][GA]	10	69	2.3e-340	366-375	373-382
8	VKYWIT[FL]NE[PI]	10	68	1.7e-338	158-167	162-171
9	DF[LI]G[FIV][NS]YY[TM][SR]	10	69	2.6e-320	291-300	294-303
10	R[TI][PK]K[KD]S[FA]YWY	10	69	1.3e-309	451-460	458-467
11	[WM][GD]WEI[DY]PEGL	10	69	1.0e-285	343-352	350-359
12	H[HN]LL[VL]A[HS][AG][LK]A	10	69	1.0e-272	196-205	199-208
13	x[FL]P[KE][DG]FL[WF]G[GT]	10	69	9.7e-263	8-17	8-17
14	[PV]N[EP][KAE]GL[DA]FYD	10	69	1.4e-260	93-102	96-105
15	[VI][HN]DDYRIDYL	10	69	4.4e-257	386-395	393-402
16	E[DG]G[KR]G[PL]SIWD	10	66	5.0e-256	30-39	30-39

17	**GGWLNR[DE][TV][IAV]D**	10	67	4.1e-219	132-141	136-145
18	**[IV]AL[FM][KA]E[LM]G[FVL]K**	10	68	9.7e-175	65-74	68-77
19	**DELL[KE]xGIEP**	10	67	7.1e-189	106-115	109-118
20	**LD[PV]Q[FL][RK]GxYP**	10	69	3.2e-167	252-261	255-264
21	**F[VA][RE]YA[RE][VT][VCL]F[EK]**	10	64	1.0e-156	143-152	147-156
22	**YP[AY][ST][CE]K[PE]ED[VI]**	10	54	6.1e-147	228-237	231-240
23	**RAIEDGV[DN][LV]K**	10	64	1.9e-134	404-413	411-420
24	**[GS][YH]LLGVH[AP]PG**	10	44	5.2e-127		178-187
25	**[KEQ]IGI[VTM]L[NA]LxP**	10	65	9.6e-129	217-226	220-229
26	**F[ES]W[AS][EN]GY**	7	49	2.3e-125		431-437
27	**TF[SCA][HK][TI]PG[KN][VT][FK]**	10	32	6.2e-093		40-49
28	**EVIA[ST]NG[EA][ES]L**	10	41	3.8e-071	462-471	469-478
29	**LNEL[YW]DRYQ[KL]**	10	23	6.7e-068	356-365	
30	**GD[ML]E[ILT][IL][ASQ]QPI**	10	46	2.9e-069	280-289	283-292
31	**FPDGD[GE][EA]**	7	55	1.5e-056	86-92	89-95
32	**G[DH][NT]GD[VI]A**	7	34	1.4e-061		51-57
33	**VS[AF][STG]T[GA][EQ]M[SK]**	9	19	5.0e-055		
34	**V[KR]NP[YH][LVI]K[TAS]S[DE]**	10	21	2.9e-046		340-349
35	**VK[ALI][GAC][HR]E[IM]NP[EK]**	10	24	1.6e-041	206-215	
36	**[KR]EH[LI]E[AQ]V**	7	62	9.1e-040	396-402	403-409
37	**[LI][LI][KM]R[VIL][KH][RK][DE]Y[PGT]**	10	29	1.2e-025		362-371
38	**[RN]AD[GA][FY]xNRWF**	10	30	8.9e-032		245-254
39	**[AF][KD]DKVE[EA]DG[SK]**	10	22	4.3e-024		383-392
40	**[DH][DN][DE]G[NT]G[TS][LM][EK]**	9	19	8.7e-021		
41	**N[MIS][SV]LH[AS]PF[TMS][GS]**	10	10	5.3e-020		
42	**[ML]T[AGS]GAHG[VK][AP]R**	10	9	3.1e-020		
43	**[HM]N[GT][TK]G[EK]KG[ST]S**	10	11	3.1e-026		319-328
44	**[EPD][GND][EKH][YF]YP[NS]H[EQ][AG]**	10	10	3.8e-016		
45	**P[LI]F[GL][WY][CT][CN]SGV**	10	6	8.0e-015		
46	**[ED]NP[EK][EQ]V[ML]YQ[AV]**	10	14	1.4e-014	185-194	

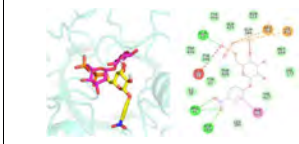
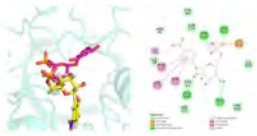
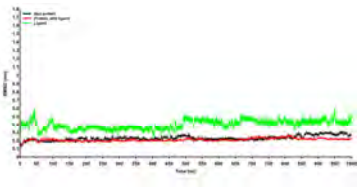
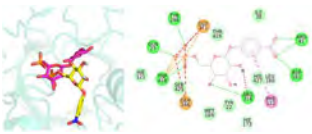
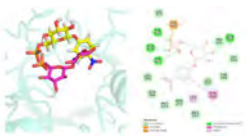
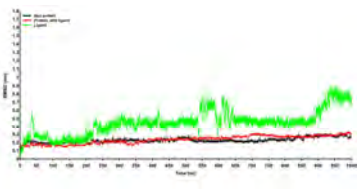
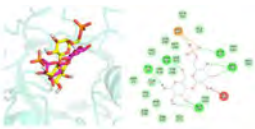
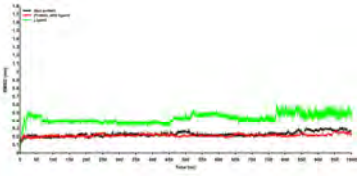
47	**WC[SA][AS]F**	5	16	6.2e-013		
48	**[QE][EKI][AS][MN][RQ][ELR]R[YF][FVW]F**	10	9	4.8e-010		
49	**[KQ][SV]FR[EH][YT]V[KP]DG**	10	8	2.5e-011		210-219
50	**RFMHQFNNYP**	10	3	3.5e-009		
51	**LEDIIHNKFI**	10	4	6.4e-008		
52	**E[DE]NYWYTAEP**	10	4	9.8e-008		
53	**YD[LF][AE]KVFQSH**	10	4	3.0e-007		
54	**[PK][FY]DP[DENS][NA]PA[DK][VI]**	10	8	4.9e-006		
55	**[LM][EQR][EKQ][NDE]R[ES][NW][LQ]FF**	10	10	3.0e-006		
56	**WMRA[FY]DG[EK][ST]E**	10	6	1.5e-008		
57	**KD[MR]KR[FM]YEAN**	10	4	5.1e-006		189-198
58	**[KQ][IV]G[NC]ML[LA]GG**	9	6	6.0e-006		
59	**RY[KQ][DH]K**	5	16	1.1e-005		
60	**KYQ[IL]KGVG[RQ]R**	10	4	1.4e-005		
61	**DMM[EN]L[FY]SK[IY][IV]**	10	4	3.9e-004		
62	**S[FY]V[EQ][RG][DN]LPKT**	10	4	8.6e-004		
63	**[TI]T[VY]E[HN]N[PI][LPV][DN]G**	10	6	2.5e-004		304-313
64	**H[REK][ND][MWY][RFN]EA[VFLP][IRT]A**	10	10	1.7e-009		
65	**NNQ[RA]N[WY][QR]**	7	4	1.1e-002		
66	**KKLAET[QH][IE]I**	9	7	7.0e-002		
67	**SY[AIV][LK][KN][EM][WFL][EA]R[KR]**	10	8	3.8e-001		
68	**QVEQVHMEEP**	10	2	4.2e-001		
69	**QF[ML]VDWF**	7	3	6.4e-001		
70	**[ES][ST]G[IM]PG[LV][FY]KT**	10	4	1.1e+000		330-339
71	**[SL][SG][KE][AQG][ED][VL]YQA[IM]**	10	5	9.6e-002		
72	**[NL][TK][DN]LQ[LT][AS][ITV][ND]V**	10	5	3.0e-001		
73	**[GS]C[VA][TS][AHT]DE**	7	5	1.2e+000		
74	**STHQGCS**	7	2	1.2e+000	425-431	
75	**MPGRKMNPY**	9	2	2.5e+000		

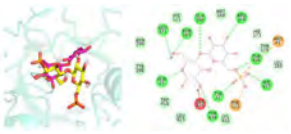
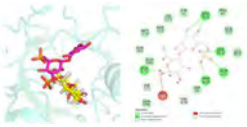
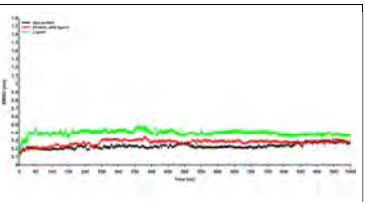
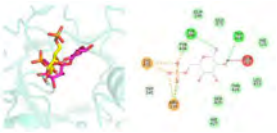
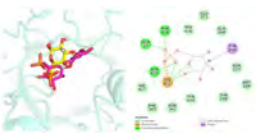
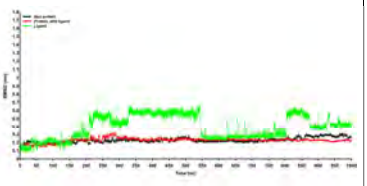
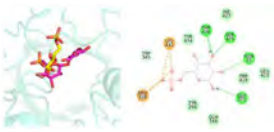
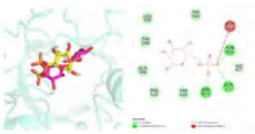
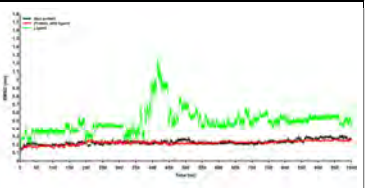
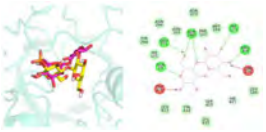
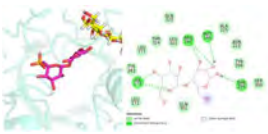
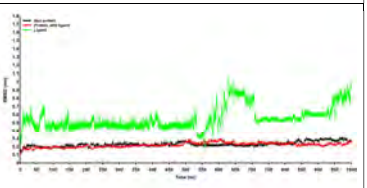
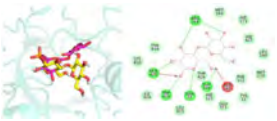
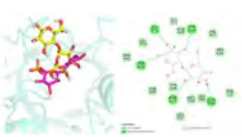
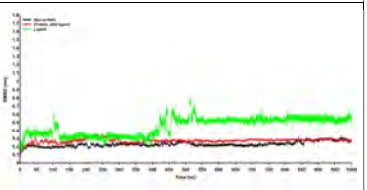
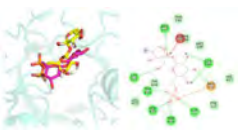
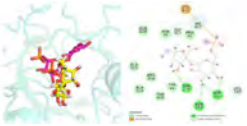
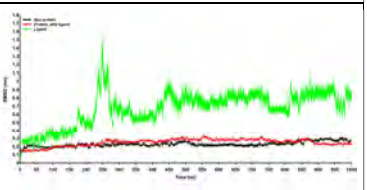
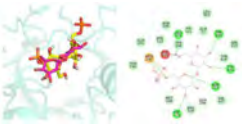
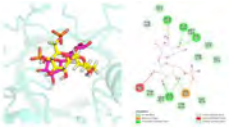
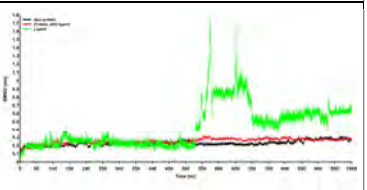
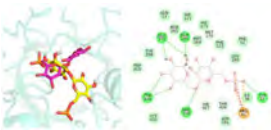
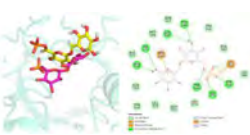
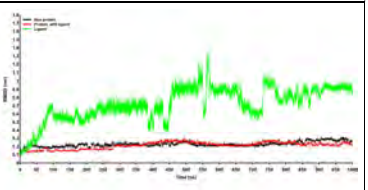
76	**KQM[MI]A[KN][NQ][GF]F**	9	3	2.3e+000		
77	**NNQMD[TV][SN]**	7	3	3.6e+000		
78	**Y[LM][EK][ES][KQ]G[LW][AET]PT**	10	6	4.3e+000	267-276	270-279
79	**VLEFAREYLP**	10	3	5.5e-001		
80	**NMMILHGSAL**	10	2	7.8e+000	168-177	

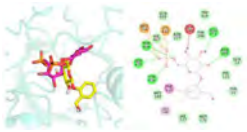
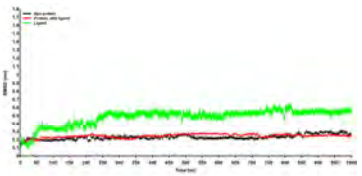
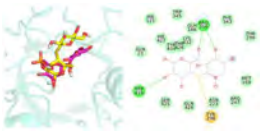
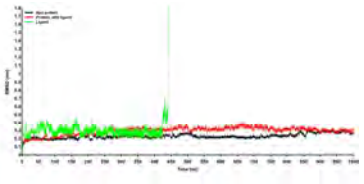
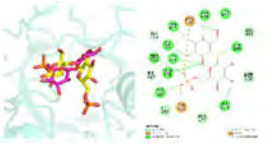
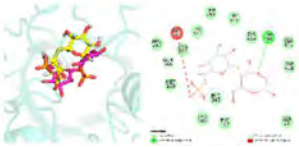
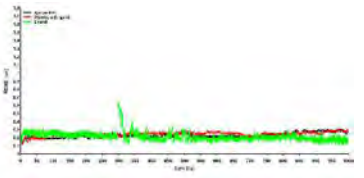
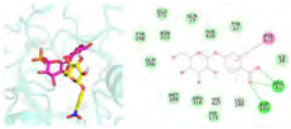
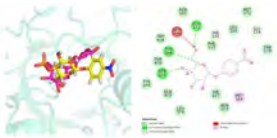
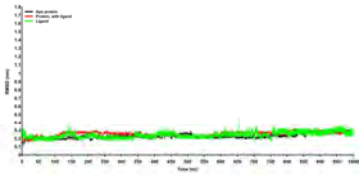
Supplementary Table 3.3. Homology model quality evaluation scores of *Bl*BglB.

Structure	**z-DOPE**	**PROCHECK**	**QMEAN**	**Verify3D**
Templates				
*Bl*BglB chain A	-1.95	89.90%	0.79	89.80%
7M1R chain A (*Bl*BglC)	-2.38	90.50%	0.44	87.21%
6WGD chain B (*Bl*BglH)	-1.97	91.80%	0.90	89.89%
Models				
*Bl*BglB	-1.70	92.30%	0.78	94.69%

Supplementary Table 3.4. Orientation in the *Bl*BglB active site and residue interactions of ligands that were docked and then simulated for 1 ms of MD. Protein and ligand RMSD graphs are also shown, compared to the apo simulation.

	Docking	**1 ms of MD**	**Protein and ligand RMSD**
PNP6Pgal Docking: 10 active site residues 1 ms of MD: 8 active site residues			
PNP6Pglc Docking: 10 active site residues 1 ms of MD: 6 active site residues			
Lactose-6P Docking: 13 active site residues 1 ms of MD: 4 active site residues			

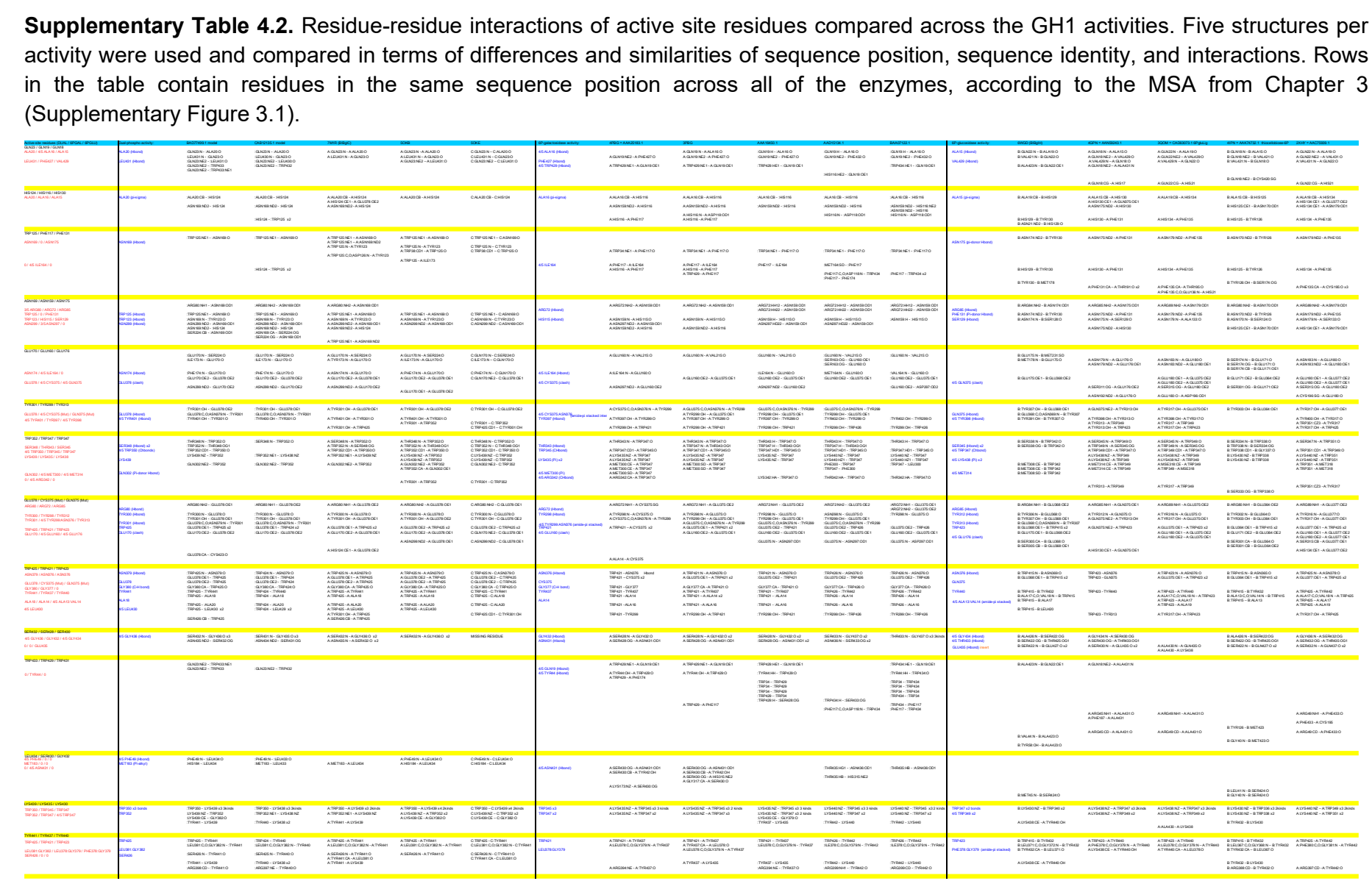
Cellobiose-6P Docking: 12 active site residues 1 ms of MD: 9 active site residues			
Galactose-6P Docking: 11 active site residues 1 ms of MD: 10 active site residues			
Glucose-6P Docking: 11 active site residues 1 ms of MD: 9 active site residues			
Lactose Docking: 11 active site residues 1 ms of MD: 1 active site residue			
Cellobiose Docking: 11 active site residues 1 ms of MD: 7 active site residues			
Sucrose-6P Docking: 10 active site residues 1 ms of MD: 3 active site residues			
Trehalose-6P Docking: 11 active site residues 1 ms of MD: 7 active site residues			
Gentiobiose-6P Docking: 12 active site residues 1 ms of MD: 6 active site residues			

Salicin-6P Docking: 12 active site residues 1 ms of MD: 12 active site residues			
Thiocellobiose Docking: 11 active site residues 1 ms of MD: Ligand exited enzyme		EXITED ENZYME	
Thiocellobiose-6P Docking: 10 active site residues 1 ms of MD: 8 active site residues			
PNPglc Docking: 7 active site residues 1 ms of MD: 10 active site residues			

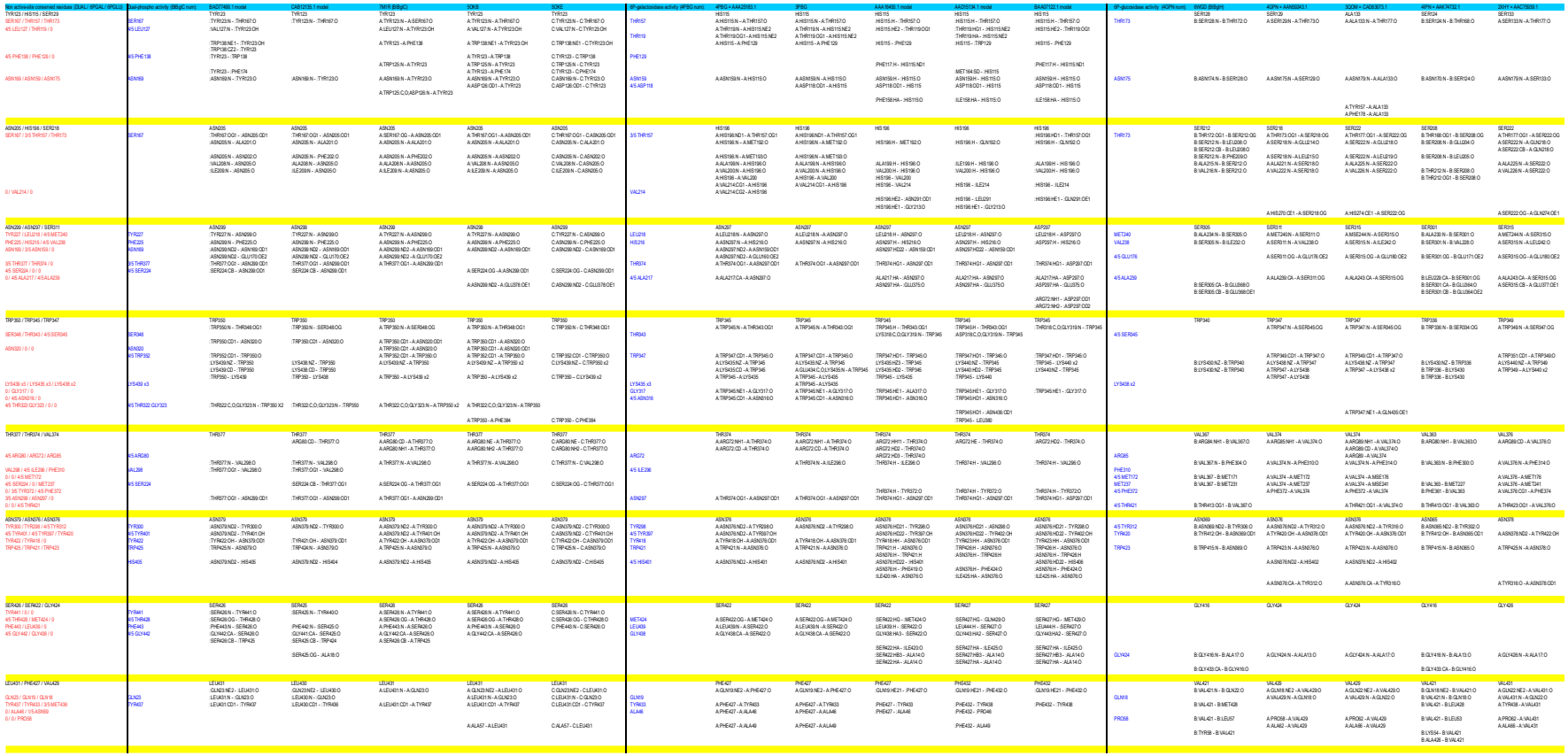
Supplementary Table 4.1. Homology model quality evaluation scores of the dual-phospho activity models.

Structure	z-DOPE	PROCHECK	QMEAN	Verify3D
Templates				
7M1R chain A (*Bl*BglC)	-2.38	90.50%	0.44	87.21%
5OKB	-2.45	89.20%	0.83	88.84%
Models				
BAD77499.1	-2.17	92.50%	-0.74	90.17%
CAB12135.1	-2.13	93.00%	-1.00	88.89%

Supplementary Table 4.2. Residue-residue interactions of active site residues compared across the GH1 activities. Five structures per activity were used and compared in terms of differences and similarities of sequence position, sequence identity, and interactions. Rows in the table contain residues in the same sequence position across all of the enzymes, according to the MSA from Chapter 3 (Supplementary Figure 3.1).



Supplementary Table 4.3. Residue-residue interactions of non-active site residues that are highly conserved across the GH1 activities or conserved in one or two activities. Five structures per activity were used and compared in terms of differences and similarities of sequence position, sequence identity, and interactions. Rows in the table contain residues in the same sequence position across all of the enzymes, according to the MSA from Chapter 3 (Supplementary Figure 3.1).



References

[1] H. Lodish, Molecular Cell Biology 8th ed., 2016.

[2] P.K. Agarwal, Enzymes: An integrated view of structure, dynamics and function, Microbial Cell Factories. (2006). https://doi.org/10.1186/1475-2859-5-2.

[3] S.D. Brown, P.C. Babbitt, New insights about enzyme evolution from large scale studies of sequence and structure relationships, The Journal of Biological Chemistry. 289 (2014) 30221–30228. https://doi.org/10.1074/JBC.R114.569350.

[4] A. Platt, H.C. Ross, S. Hankin, R.J. Reece, The insertion of two amino acids into a transcriptional inducer converts it into a galactokinase, Proceedings of the National Academy of Sciences of the United States of America. (2000). https://doi.org/10.1073/pnas.97.7.3154.

[5] K.A. Dill, J.L. MacCallum, The protein-folding problem, 50 years on, Science (New York, N.Y.). 338 (2012) 1042–1046. https://doi.org/10.1126/SCIENCE.1219021.

[6] R. Nassar, G.L. Dignon, R.M. Razban, K.A. Dill, The Protein Folding Problem: The Role of Theory, Journal of Molecular Biology. 433 (2021) 167126. https://doi.org/10.1016/J.JMB.2021.167126.

[7] K.E. Jaeger, T. Eggert, Enantioselective biocatalysis optimized by directed evolution, Current Opinion in Biotechnology. (2004). https://doi.org/10.1016/j.copbio.2004.06.007.

[8] D. Voet, J. Voet, C. Pratt, Fundamentals of Biochemistry - Life at the molecular level, 5th ed., 2016.

[9] E. Fischer, Influence of configuration on the action of enzymes, Ber. (1894).

[10] G.M. Cooper, The Central Role of Enzymes as Biological Catalysts, The Cell: A Molecular Approach. (2000).

[11] D.E. Koshland, Jr., Application of a Theory of Enzyme Specificity to Protein Synthesis, Proceedings of the National Academy of Sciences of the United States of America. 44 (1958) 98. https://doi.org/10.1073/PNAS.44.2.98.

[12] A. Vasella, G.J. Davies, M. Böhm, Glycosidase mechanisms, Current Opinion in Chemical Biology. (2002). https://doi.org/10.1016/S1367-5931(02)00380-0.

[13] Y. Savir, T. Tiusty, Conformational proofreading: The impact of conformational changes on the specificity of molecular recognition, PLoS ONE. (2007). https://doi.org/10.1371/journal.pone.0000468.

[14] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, The energy landscapes and motions of proteins, Science. (1991). https://doi.org/10.1126/science.1749933.

[15] K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins, Nature. (2007). https://doi.org/10.1038/nature06522.

[16] J.D. Bryngelson, J.N. Onuchic, N.D. Socci, P.G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: A synthesis, Proteins: Structure, Function, and Bioinformatics. (1995). https://doi.org/10.1002/prot.340210302.

[17] D.W. Miller, K.A. Dill, Ligand binding to proteins: The binding landscape model, Protein Science. (1997). https://doi.org/10.1002/pro.5560061011.

[18] X. Du, Y. Li, Y.-L.L. Xia, S.-M.M. Ai, J. Liang, P. Sang, X.-L.L. Ji, S.-Q.Q. Liu, Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods, International Journal of Molecular Sciences. 17 (2016). /pmc/articles/PMC4783878/ (accessed July 26, 2021).
[19] P.L. Kastritis, A.M.J.J. Bonvin, On the binding affinity of macromolecular interactions: Daring to ask why proteins interact, Journal of the Royal Society Interface. (2013). https://doi.org/10.1098/rsif.2012.0835.
[20] R. Bone, J.L. Silen, D.A. Agard, Structural plasticity broadens the specificity of an engineered protease, Nature. (1989). https://doi.org/10.1038/339191a0.
[21] M. Huse, J. Kuriyan, The conformational plasticity of protein kinases, Cell. (2002). https://doi.org/10.1016/S0092-8674(02)00741-9.
[22] C. Das, Q.Q. Hoang, C.A. Kreinbring, S.J. Luchansky, R.K. Meray, S.S. Ray, P.T. Lansbury, D. Ringe, G.A. Petsko, Structural basis for conformational plasticity of the Parkinson's disease-associated ubiquitin hydrolase UCH-L1, Proceedings of the National Academy of Sciences of the United States of America. (2006). https://doi.org/10.1073/pnas.0510403103.
[23] S. Chakraborty, B. Ásgeirsson, B.J. Rao, A Measure of the Broad Substrate Specificity of Enzymes Based on "Duplicate" Catalytic Residues, PLoS ONE. (2012). https://doi.org/10.1371/journal.pone.0049313.
[24] H.-W. Liu, T.P. Begley, Comprehensive Natural Products III, 2020. https://doi.org/10.1016/c2017-1-02016-0.
[25] E.E.V. Bezirtzoglou, Intestinal cytochromes P450 regulating the intestinal microbiota and its probiotic profile, Microbial Ecology in Health & Disease. (2012). https://doi.org/10.3402/mehd.v23i0.18370.
[26] E.T. Morgan, Impact of infectious and inflammatory disease on cytochrome P450-mediated drug metabolism and pharmacokinetics, Clinical Pharmacology and Therapeutics. (2009). https://doi.org/10.1038/clpt.2008.302.
[27] P.J. O'Brien, D. Herschlag, Catalytic promiscuity and the evolution of new enzymatic activities, Chemistry and Biology. (1999). https://doi.org/10.1016/S1074-5521(99)80033-7.
[28] O.K. and D.S. Tawfik, Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective, Annual Review of Biochemistry. (2010). https://doi.org/10.1146/annurev-biochem-030409-143718.
[29] S.G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P.M. Harrell, Y.T. Trinh, Q. Zhang, I.L. Urbatsch, G. Chang, Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding, Science. (2009). https://doi.org/10.1126/science.1168750.
[30] M. Nukaga, S. Haruta, K. Tanimoto, K. Kogure, K. Taniguchi, M. Tamaki, T. Sawai, Molecular evolution of a class C β-lactamase extending its substrate specificity, Journal of Biological Chemistry. (1995). https://doi.org/10.1074/jbc.270.11.5729.
[31] J.C. Samuelson, S.Y. Xu, Directed evolution of restriction endonuclease BstYI to achieve increased substrate specificity, Journal of Molecular Biology. (2002). https://doi.org/10.1016/S0022-2836(02)00343-1.

[32] I. Berger, C. Guttman, D. Amar, R. Zarivach, A. Aharoni, The molecular basis for the broad substrate specificity of human sulfotransferase 1A1, PLoS ONE. (2011). https://doi.org/10.1371/journal.pone.0026794.
[33] B.A. Kaup, U. Piantini, M. Wüst, J. Schrader, Monoterpenes as novel substrates for oxidation and halo-hydroxylation with chloroperoxidase from Caldariomyces fumago, Applied Microbiology and Biotechnology. (2007). https://doi.org/10.1007/s00253-006-0559-3.
[34] C. Pandya, J.D. Farelli, D. Dunaway-Mariano, K.N. Allen, Enzyme promiscuity: Engine of evolutionary innovation, Journal of Biological Chemistry. (2014). https://doi.org/10.1074/jbc.R114.572990.
[35] A.S. Ibuka, Y. Ishii, M. Galleni, M. Ishiguro, K. Yamaguchi, J.M. Frère, H. Matsuzawa, H. Sakai, Crystal structure of extended-spectrum β-lactamase Toho-1: Insights into the molecular mechanism for catalytic reaction and substrate specificity expansion, Biochemistry. (2003). https://doi.org/10.1021/bi0342822.
[36] E. Sauvage, E. Fonzé, B. Quinting, M. Galleni, J.M. Frère, P. Charlier, Crystal structure of the mycobacterium fortuitum class a β-lactamase: Structural basis for broad substrate specificity, Antimicrobial Agents and Chemotherapy. (2006). https://doi.org/10.1128/AAC.01226-05.
[37] M. Thakur, I.L. Medintz, S.A. Walper, Enzymatic Bioremediation of Organophosphate Compounds—Progress and Remaining Challenges, Frontiers in Bioengineering and Biotechnology. (2019). https://doi.org/10.3389/fbioe.2019.00289.
[38] E. Dellus-Gur, A. Toth-Petroczy, M. Elias, D.S. Tawfik, What makes a protein fold amenable to functional innovation? fold polarity and stability trade-offs, Journal of Molecular Biology. (2013). https://doi.org/10.1016/j.jmb.2013.03.033.
[39] C. Zhang, C. DeLisi, Protein folds: Molecular systematics in three dimensions, Cellular and Molecular Life Sciences. (2001). https://doi.org/10.1007/PL00000779.
[40] M. Ben-David, M. Elias, J.J. Filippi, E. Duñach, I. Silman, J.L. Sussman, D.S. Tawfik, Catalytic versatility and backups in enzyme active sites: The case of serum paraoxonase 1, Journal of Molecular Biology. (2012). https://doi.org/10.1016/j.jmb.2012.02.042.
[41] J. Hiblot, G. Gotthard, M. Elias, E. Chabriere, Differential Active Site Loop Conformations Mediate Promiscuous Activities in the Lactonase SsoPox, PLoS ONE. (2013). https://doi.org/10.1371/journal.pone.0075272.
[42] T. Gabaldón, M.A. Huynen, Prediction of protein function and pathways in the genome era, Cellular and Molecular Life Sciences. (2004). https://doi.org/10.1007/s00018-003-3387-y.
[43] B. Srinivasan, Words of advice: teaching enzyme kinetics, FEBS Journal. (2020). https://doi.org/10.1111/febs.15537.
[44] S. Schnell, M.J. Chappell, N.D. Evans, M.R. Roussel, The mechanism distinguishability problem in biochemical kinetics: The single-enzyme, single-substrate reaction as a case study, Comptes Rendus - Biologies. (2006). https://doi.org/10.1016/j.crvi.2005.09.005.
[45] L. Michaelis, M.L. Menten, R.S. Goody, K.A. Johnson, Die Kinetik der Invertinwirkung/ The kinetics of invertase action, Biochemistry. (1913).

[46] H. Bisswanger, Enzyme assays, Perspectives in Science. (2014). https://doi.org/10.1016/j.pisc.2014.02.005.
[47] R.K. Scopes, Enzyme Activity and Assays, in: Encyclopedia of Life Sciences, 2002. https://doi.org/10.1038/npg.els.0000712.
[48] J. Shi, J. Dertouzos, A. Gafni, D. Steel, Chapter 7 Application of Single-Molecule Spectroscopy in Studying Enzyme Kinetics and Mechanism, Methods in Enzymology. (2008). https://doi.org/10.1016/S0076-6879(08)03407-1.
[49] U. Kettling, A. Koltermann, P. Schwille, M. Eigen, Real-time enzyme kinetics monitored by dual-color fluorescence cross-correlation spectroscopy, Proceedings of the National Academy of Sciences of the United States of America. (1998). https://doi.org/10.1073/pnas.95.4.1416.
[50] J. Boeckx, M. Hertog, A. Geeraerd, B. Nicolai, Kinetic modelling: An integrated approach to analyze enzyme activity assays, Plant Methods. (2017). https://doi.org/10.1186/s13007-017-0218-y.
[51] W. Helbert, L. Poulet, S. Drouillard, S. Mathieu, M. Loiodice, M. Couturier, V. Lombard, N. Terrapon, J. Turchetto, R. Vincentelli, B. Henrissat, Discovery of novel carbohydrate-active enzymes through the rational exploration of the protein sequences space, Proceedings of the National Academy of Sciences of the United States of America. 116 (2019) 6063–6068. https://doi.org/10.1073/pnas.1815791116.
[52] P. Edman, E. Högfeldt, L.G. Sillén, P.-O. Kinell, Method for Determination of the Amino Acid Sequence in Peptides., Acta Chemica Scandinavica. (1950). https://doi.org/10.3891/acta.chem.scand.04-0283.
[53] W. Timp, G. Timp, Beyond mass spectrometry, the next step in proteomics, Science Advances. (2020). https://doi.org/10.1126/sciadv.aax8978.
[54] R. Aebersold, M. Mann, Mass-spectrometric exploration of proteome structure and function, Nature. (2016). https://doi.org/10.1038/nature19949.
[55] J.D. Shannon, J.W. Fox, Identification of phosphorylation sites by Edman degradation, Techniques in Protein Chemistry. (1995). https://doi.org/10.1016/S1080-8914(06)80017-7.
[56] F.E. Ross, T. Zamborelli, A.C. Herman, C.H. Yeh, N.I. Tedeschi, E.S. Luedke, Detection of acetylated lysine residues using sequencing by edman degradation and mass spectrometry, Techniques in Protein Chemistry. (1996). https://doi.org/10.1016/S1080-8914(96)80024-X.
[57] A.I. Nesvizhskii, R. Aebersold, Interpretation of shotgun proteomic data: The protein inference problem, Molecular and Cellular Proteomics. (2005). https://doi.org/10.1074/mcp.R500012-MCP200.
[58] C. Hughes, B. Ma, G.A. Lajoie, De novo sequencing methods in proteomics., Methods in Molecular Biology (Clifton, N.J.). (2010). https://doi.org/10.1007/978-1-60761-444-9_8.
[59] B.T. Chait, Mass spectrometry: Bottom-up or top-down?, Science. (2006). https://doi.org/10.1126/science.1133987.

[60] G.A. Valaskovic, N.L. Kelleher, F.W. McLafferty, Attomole protein characterization by capillary electrophoresis-mass spectrometry, Science. (1996). https://doi.org/10.1126/science.273.5279.1199.
[61] K.A. Resing, N.G. Ahn, Proteomics strategies for protein identification, in: FEBS Letters, 2005. https://doi.org/10.1016/j.febslet.2004.12.001.
[62] M. The, F. Edfors, Y. Perez-Riverol, S.H. Payne, M.R. Hoopmann, M. Palmblad, B. Forsström, L. Käll, A Protein Standard That Emulates Homology for the Characterization of Protein Inference Algorithms, Journal of Proteome Research. (2018). https://doi.org/10.1021/acs.jproteome.7b00899.
[63] J. Griss, Y. Perez-Riverol, S. Lewis, D.L. Tabb, J.A. Dianes, N. Del-Toro, M. Rurik, M. Walzer, O. Kohlbacher, H. Hermjakob, R. Wang, J.A. Vizcano, Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets, Nature Methods. (2016). https://doi.org/10.1038/nmeth.3902.
[64] M. Kulmanov, R. Hoehndorf, DeepGOPlus: Improved protein function prediction from sequence, Bioinformatics. (2020). https://doi.org/10.1093/bioinformatics/btz595.
[65] Y. Liu, Y. Lei, X. Zhang, Y. Gao, Y. Xiao, H. Peng, Identification and Phylogenetic Characterization of a New Subfamily of α-Amylase Enzymes from Marine Microorganisms, Marine Biotechnology. (2012). https://doi.org/10.1007/s10126-011-9414-3.
[66] G. Huang, C. Chu, T. Huang, X. Kong, Y. Zhang, N. Zhang, Y.D. Cai, Exploring Mouse Protein Function via Multiple Approaches, PLoS ONE. (2016). https://doi.org/10.1371/journal.pone.0166580.
[67] M.F.M. Sobri, S. Abd-Aziz, F.D.A. Bakar, N. Ramli, In-silico characterization of glycosyl hydrolase family 1 β-glucosidase from Trichoderma asperellum UPM1, International Journal of Molecular Sciences. (2020). https://doi.org/10.3390/ijms21114035.
[68] Y. Jiang, T.R. Oron, W.T. Clark, A.R. Bankapur, D. D'Andrea, R. Lepore, C.S. Funk, I. Kahanda, K.M. Verspoor, A. Ben-Hur, D.C.E. Koo, D. Penfold-Brown, D. Shasha, N. Youngs, R. Bonneau, A. Lin, S.M.E. Sahraeian, P.L. Martelli, G. Profiti, R. Casadio, R. Cao, Z. Zhong, J. Cheng, A. Altenhoff, N. Skunca, C. Dessimoz, T. Dogan, K. Hakala, S. Kaewphan, F. Mehryary, T. Salakoski, F. Ginter, H. Fang, B. Smithers, M. Oates, J. Gough, P. Törönen, P. Koskinen, L. Holm, C.T. Chen, W.L. Hsu, K. Bryson, D. Cozzetto, F. Minneci, D.T. Jones, S. Chapman, D. Bkc, I.K. Khan, D. Kihara, D. Ofer, N. Rappoport, A. Stern, E. Cibrian-Uhalte, P. Denny, R.E. Foulger, R. Hieta, D. Legge, R.C. Lovering, M. Magrane, A.N. Melidoni, P. Mutowo-Meullenet, K. Pichler, A. Shypitsyna, B. Li, P. Zakeri, S. ElShal, L.C. Tranchevent, S. Das, N.L. Dawson, D. Lee, J.G. Lees, I. Sillitoe, P. Bhat, T. Nepusz, A.E. Romero, R. Sasidharan, H. Yang, A. Paccanaro, J. Gillis, A.E. Sedeño-Cortés, P. Pavlidis, S. Feng, J.M. Cejuela, T. Goldberg, T. Hamp, L. Richter, A. Salamov, T. Gabaldon, M. Marcet-Houben, F. Supek, Q. Gong, W. Ning, Y. Zhou, W. Tian, M. Falda, P. Fontana, E. Lavezzo, S. Toppo, C. Ferrari, M. Giollo, D. Piovesan, S.C.E. Tosatto, A. del Pozo, J.M. Fernández, P. Maietta, A. Valencia, M.L. Tress, A. Benso, S. di Carlo, G. Politano, A. Savino, H.U. Rehman, M. Re, M. Mesiti, G. Valentini, J.W. Bargsten, A.D.J. van Dijk, B. Gemovic, S. Glisic, V. Perovic, V. Veljkovic, N.

Veljkovic, D.C. Almeida-e-Silva, R.Z.N. Vencio, M. Sharan, J. Vogel, L. Kansakar, S. Zhang, S. Vucetic, Z. Wang, M.J.E. Sternberg, M.N. Wass, R.P. Huntley, M.J. Martin, C. O'Donovan, P.N. Robinson, Y. Moreau, A. Tramontano, P.C. Babbitt, S.E. Brenner, M. Linial, C.A. Orengo, B. Rost, C.S. Greene, S.D. Mooney, I. Friedberg, P. Radivojac, An expanded evaluation of protein function prediction methods shows an improvement in accuracy, Genome Biology. (2016). https://doi.org/10.1186/s13059-016-1037-6.

[69] P. Radivojac, W.T. Clark, T.R. Oron, A.M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk, K. Verspoor, A. Ben-Hur, G. Pandey, J.M. Yunes, A.S. Talwalkar, S. Repo, M.L. Souza, D. Piovesan, R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Törönen, J. Nokso-Koivisto, L. Holm, D. Cozzetto, D.W.A. Buchan, K. Bryson, D.T. Jones, B. Limaye, H. Inamdar, A. Datta, S.K. Manjari, R. Joshi, M. Chitale, D. Kihara, A.M. Lisewski, S. Erdin, E. Venner, O. Lichtarge, R. Rentzsch, H. Yang, A.E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kaßner, S. Seemayer, E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Hönigschmid, T.A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Björne, T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M.N. Wass, M.J.E. Sternberg, N. Škunca, F. Supek, M. Bošnjak, P. Panov, S. Džeroski, T. Šmuc, Y.A.I. Kourmpetis, A.D.J. van Dijk, C.J.F. ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. di Camillo, S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P.C. Babbitt, S.E. Brenner, C. Orengo, B. Rost, S.D. Mooney, I. Friedberg, A large-scale evaluation of computational protein function prediction, Nature Methods. (2013). https://doi.org/10.1038/nmeth.2340.

[70] N. Zhou, Y. Jiang, T.R. Bergquist, A.J. Lee, B.Z. Kacsoh, A.W. Crocker, K.A. Lewis, G. Georghiou, H.N. Nguyen, M.N. Hamid, L. Davis, T. Dogan, V. Atalay, A.S. Rifaioglu, A. Dalklran, R. Cetin Atalay, C. Zhang, R.L. Hurto, P.L. Freddolino, Y. Zhang, P. Bhat, F. Supek, J.M. Fernández, B. Gemovic, V.R. Perovic, R.S. Davidović, N. Sumonja, N. Veljkovic, E. Asgari, M.R.K. Mofrad, G. Profiti, C. Savojardo, P.L. Martelli, R. Casadio, F. Boecker, H. Schoof, I. Kahanda, N. Thurlby, A.C. McHardy, A. Renaux, R. Saidi, J. Gough, A.A. Freitas, M. Antczak, F. Fabris, M.N. Wass, J. Hou, J. Cheng, Z. Wang, A.E. Romero, A. Paccanaro, H. Yang, T. Goldberg, C. Zhao, L. Holm, P. Törönen, A.J. Medlar, E. Zosa, I. Borukhov, I. Novikov, A. Wilkins, O. Lichtarge, P.H. Chi, W.C. Tseng, M. Linial, P.W. Rose, C. Dessimoz, V. Vidulin, S. Dzeroski, I. Sillitoe, S. Das, J.G. Lees, D.T. Jones, C. Wan, D. Cozzetto, R. Fa, M. Torres, A. Warwick Vesztrocy, J.M. Rodriguez, M.L. Tress, M. Frasca, M. Notaro, G. Grossi, A. Petrini, M. Re, G. Valentini, M. Mesiti, D.B. Roche, J. Reeb, D.W. Ritchie, S. Aridhi, S.Z. Alborzi, M.D. Devignes, D.C.E. Koo, R. Bonneau, V. Gligorijević, M. Barot, H. Fang, S. Toppo, E. Lavezzo, M. Falda, M. Berselli, S.C.E. Tosatto, M. Carraro, D. Piovesan, H. Ur Rehman, Q. Mao, S. Zhang, S. Vucetic, G.S. Black, D. Jo, E. Suh, J.B. Dayton, D.J. Larsen, A.R. Omdahl, L.J. McGuffin, D.A. Brackenridge, P.C. Babbitt, J.M. Yunes, P. Fontana, F. Zhang, S. Zhu, R. You, Z. Zhang, S. Dai, S. Yao, W. Tian, R. Cao, C. Chandler, M. Amezola, D. Johnson, J.M. Chang, W.H. Liao, Y.W. Liu, S. Pascarelli, Y. Frank, R. Hoehndorf,

M. Kulmanov, I. Boudellioua, G. Politano, S. di Carlo, A. Benso, K. Hakala, F. Ginter, F. Mehryary, S. Kaewphan, J. Björne, H. Moen, M.E.E. Tolvanen, T. Salakoski, D. Kihara, A. Jain, T. Šmuc, A. Altenhoff, A. Ben-Hur, B. Rost, S.E. Brenner, C.A. Orengo, C.J. Jeffery, G. Bosco, D.A. Hogan, M.J. Martin, C. O'Donovan, S.D. Mooney, C.S. Greene, P. Radivojac, I. Friedberg, The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens, Genome Biology. (2019). https://doi.org/10.1186/s13059-019-1835-8.

[71] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, Nucleic Acids Research. (1997). https://doi.org/10.1093/nar/25.17.3389.

[72] R.D. Sleator, P. Walsh, An overview of in silico protein function prediction, Archives of Microbiology. (2010). https://doi.org/10.1007/s00203-010-0549-9.

[73] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, M. Qureshi, L.J. Richardson, G.A. Salazar, A. Smart, E.L.L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S.C.E. Tosatto, R.D. Finn, The Pfam protein families database in 2019, Nucleic Acids Research. (2019). https://doi.org/10.1093/nar/gky995.

[74] C.J.A. Sigrist, L. Cerutti, E. de Castro, P.S. Langendijk-Genevaux, V. Bulliard, A. Bairoch, N. Hulo, PROSITE, a protein domain database for functional characterization and annotation, Nucleic Acids Research. (2009). https://doi.org/10.1093/nar/gkp885.

[75] A. Marchler-Bauer, S. Lu, J.B. Anderson, F. Chitsaz, M.K. Derbyshire, C. DeWeese-Scott, J.H. Fong, L.Y. Geer, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, J.D. Jackson, Z. Ke, C.J. Lanczycki, F. Lu, G.H. Marchler, M. Mullokandov, M. v. Omelchenko, C.L. Robertson, J.S. Song, N. Thanki, R.A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S.H. Bryant, CDD: A Conserved Domain Database for the functional annotation of proteins, Nucleic Acids Research. (2011). https://doi.org/10.1093/nar/gkq1189.

[76] P.K. Busk, B. Pilgaard, M.J. Lezyk, A.S. Meyer, L. Lange, Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function, BMC Bioinformatics. (2017). https://doi.org/10.1186/s12859-017-1625-9.

[77] P.K. Busk, L. Lange, Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs, Applied and Environmental Microbiology. 79 (2013) 3380–3391. https://doi.org/10.1128/AEM.03803-12.

[78] J.C. Whisstock, A.M. Lesk, Prediction of protein function from protein sequence and structure, Quarterly Reviews of Biophysics. (2003). https://doi.org/10.1017/S0033583503003901.

[79] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Research. (2000). https://doi.org/10.1093/nar/28.1.235.

[80] C. Guda, S. Lu, E.D. Scheeff, P.E. Bourne, I.N. Shindyalov, CE-MC: A multiple protein structure alignment server, Nucleic Acids Research. (2004). https://doi.org/10.1093/nar/gkh464.

[81] L. Holm, DALI and the persistence of protein shape, Protein Science. (2020). https://doi.org/10.1002/pro.3749.
[82] Y. Ye, A. Godzik, Flexible structure alignment by chaining aligned fragment pairs allowing twists, in: Bioinformatics, 2003. https://doi.org/10.1093/bioinformatics/btg1086.
[83] D. Eisenberg, E.M. Marcotte, I. Xenarios, T.O. Yeates, Protein function the post-genomic era, Nature. (2000). https://doi.org/10.1038/35015694.
[84] J.M. Hancock, M.J. Zvelebil, M.J. Zvelebil, UniProt, in: Dictionary of Bioinformatics and Computational Biology, 2004. https://doi.org/10.1002/9780471650126.dob0721.pub2.
[85] S. Zhao, A. Sakai, X. Zhang, M.W. Vetting, R. Kumar, B. Hillerich, B. San Francisco, J. Solbiati, A. Steves, S. Brown, E. Akiva, A. Barber, R.D. Seidel, P.C. Babbitt, S.C. Almo, J.A. Gerlt, M.P. Jacobson, Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks, ELife. (2014). https://doi.org/10.7554/elife.03275.
[86] S. Saha, A. Prasad, P. Chatterjee, S. Basu, M. Nasipuri, Protein function prediction from protein-protein interaction network using gene ontology based neighborhood analysis and physico-chemical features, in: Journal of Bioinformatics and Computational Biology, 2018. https://doi.org/10.1142/S0219720018500257.
[87] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Molecular Systems Biology. (2007). https://doi.org/10.1038/msb4100129.
[88] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, Q. Morris, GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function, Genome Biology. (2008). https://doi.org/10.1186/gb-2008-9-s1-s4.
[89] R. Fa, D. Cozzetto, C. Wan, D.T. Jones, Predicting human protein function with multitask deep neural networks, PLoS ONE. (2018). https://doi.org/10.1371/journal.pone.0198216.
[90] C. Nagao, N. Nagano, K. Mizuguchi, Prediction of detailed enzyme functions and identification of specificity determining residues by random forests, PLoS ONE. (2014). https://doi.org/10.1371/journal.pone.0084623.
[91] B. Rost, Enzyme function less conserved than anticipated, Journal of Molecular Biology. (2002). https://doi.org/10.1016/S0022-2836(02)00016-5.
[92] E.C. Meng, Determining enzyme function by predicting substrate specificity, (n.d.). https://pharmchem.ucsf.edu/research/compchem/specificity (accessed February 18, 2021).
[93] C. Kalyanaraman, K. Bernacki, M.P. Jacobson, Virtual screening against highly charged active sites: Identifying substrates of alpha-beta barrel enzymes, Biochemistry. (2005). https://doi.org/10.1021/bi0481186.
[94] J.C. Hermann, E. Ghanem, Y. Li, F.M. Raushel, J.J. Irwin, B.K. Shoichet, Predicting substrates by docking high-energy intermediates to enzyme structures, Journal of the American Chemical Society. (2006). https://doi.org/10.1021/ja065860f.
[95] D.F. Xiang, P. Kolb, A.A. Fedorov, M.M. Meier, L. v. Fedorov, T.T. Nguyen, R. Sterner, S.C. Almo, B.K. Shoichet, F.M. Raushel, Functional annotation and three-dimensional structure of Dr0930 from deinococcus radiodurans, a close relative of

phosphotriesterase in the amidohydrolase superfamily, Biochemistry. (2009). https://doi.org/10.1021/bi802274f.
[96] A.D. Favia, I. Nobeli, F. Glaser, J.M. Thornton, Molecular Docking for Substrate Identification: The Short-Chain Dehydrogenases/Reductases, Journal of Molecular Biology. (2008). https://doi.org/10.1016/j.jmb.2007.10.065.
[97] L. Song, C. Kalyanaraman, A.A. Fedorov, E. v. Fedorov, M.E. Glasner, S. Brown, H.J. Imker, P.C. Babbitt, S.C. Almo, M.P. Jacobson, J.A. Gerlt, Prediction and assignment of function for a divergent N-succinyl amino acid racemase, Nature Chemical Biology. (2007). https://doi.org/10.1038/nchembio.2007.11.
[98] J.F. Rakus, C. Kalyanaraman, A.A. Fedorov, E. v. Fedorov, F.P. Mills-Groninger, R. Toro, J. Bonanno, K. Bain, J.M. Sauder, S.K. Burley, S.C. Almo, M.P. Jacobson, J.A. Gerlt, Computation-facilitated assignment of the function in the enolase superfamily: A regiochemically distinct galactarate dehydratase from Oceanobacillus iheyensis, Biochemistry. (2009). https://doi.org/10.1021/bi901731c.
[99] C. Kalyanaraman, M.P. Jacobson, Studying enzyme-substrate specificity in silico: A case study of the escherichia coli glycolysis pathway, Biochemistry. (2010). https://doi.org/10.1021/bi100445g.
[100] I. Wallach, A. Heifets, Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization, Journal of Chemical Information and Modeling. 58 (2018) 916–932. https://doi.org/10.1021/ACS.JCIM.7B00403.
[101] J. Sieg, F. Flachsenberg, M. Rarey, In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening, Journal of Chemical Information and Modeling. 59 (2019) 947–961. https://doi.org/10.1021/ACS.JCIM.8B00712/SUPPL_FILE/CI8B00712_SI_001.PDF.
[102] J. Polaina, A.P. MacCabe, Industrial enzymes: Structure, function and applications, 2007. https://doi.org/10.1007/1-4020-5377-0.
[103] R. Kumar, B. Henrissat, P.M. Coutinho, Intrinsic dynamic behavior of enzyme:substrate complexes govern the catalytic action of β-galactosidases across clan GH-A, Scientific Reports. 9 (2019) 1–14. https://doi.org/10.1038/s41598-019-46589-8.
[104] W.F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D.P. Geerke, A. Glättli, P.H. Hünenberger, M.A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N.F.A. van der Vegt, H.B. Yu, Biomolecular modeling: Goals, problems, perspectives, Angewandte Chemie - International Edition. (2006). https://doi.org/10.1002/anie.200502655.
[105] W. Veldman, M.V. Liberato, V.M. Almeida, V.P. Souza, M.A. Frutuoso, S.R. Marana, V. Moses, Ö. Tastan Bishop, I. Polikarpov, X-ray Structure, Bioinformatics Analysis, and Substrate Specificity of a 6-Phospho-β-glucosidase Glycoside Hydrolase 1 Enzyme from Bacillus licheniformis, Journal of Chemical Information and Modeling. (2020). https://doi.org/10.1021/acs.jcim.0c00759.
[106] M. Grandits, H. Michlmayr, C. Sygmund, C. Oostenbrink, Calculation of substrate binding affinities for a bacterial GH78 rhamnosidase through molecular dynamics simulations, Journal of Molecular Catalysis B: Enzymatic. (2013). https://doi.org/10.1016/j.molcatb.2013.03.012.

[107] S. Sarkar, S. Gupta, W. Chakraborty, S. Senapati, R. Gachhui, Homology modeling, molecular docking and molecular dynamics studies of the catalytic domain of chitin deacetylase from Cryptococcus laurentii strain RY1, International Journal of Biological Macromolecules. (2017). https://doi.org/10.1016/j.ijbiomac.2017.03.057.
[108] M.H. Momeni, C.M. Payne, H. Hansson, N.E. Mikkelsen, J. Svedberg, A. Engström, M. Sandgren, G.T. Beckham, J. Stahlberg, Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus heterobasidion irregulare, Journal of Biological Chemistry. (2013). https://doi.org/10.1074/jbc.M112.440891.
[109] M. Wu, G.T. Beckham, A.M. Larsson, T. Ishida, S. Kim, C.M. Payne, M.E. Himmel, M.F. Crowley, S.J. Horn, B. Westereng, K. Igarashi, M. Samejima, J. Ståhlberg, V.G.H. Eijsink, M. Sandgren, Crystal structure and computational characterization of the lytic polysaccharide monooxygenase GH61D from the basidiomycota fungus Phanerochaete chrysosporium, Journal of Biological Chemistry. (2013). https://doi.org/10.1074/jbc.M113.459396.
[110] L. Bu, M.F. Crowley, M.E. Himmel, G.T. Beckham, Computational investigation of the pH dependence of loop flexibility and catalytic function in glycoside hydrolases, Journal of Biological Chemistry. (2013). https://doi.org/10.1074/jbc.M113.462465.
[111] C.M. Bianchetti, P. Brumm, R.W. Smith, K. Dyer, G.L. Hura, T.J. Rutkoski, G.N. Phillips, Structure, dynamics, and specificity of endoglucanase D from clostridium cellulovorans, Journal of Molecular Biology. (2013). https://doi.org/10.1016/j.jmb.2013.05.030.
[112] Q.R. Johnson, R.J. Lindsay, L. Petridis, T. Shen, Investigation of carbohydrate recognition via computer simulation, Molecules. (2015). https://doi.org/10.3390/molecules20057700.
[113] F. Calzado, E.T. Prates, T.A. Gonçalves, M. v. Rubio, M.P. Zubieta, F.M. Squina, M.S. Skaf, A.R.L. Damásio, Molecular basis of substrate recognition and specificity revealed in family 12 glycoside hydrolases, Biotechnology and Bioengineering. (2016). https://doi.org/10.1002/bit.26036.
[114] M. Jain, J. Muthukumaran, A.K. Singh, Structural and functional characterization of chitin binding lectin from Datura stramonium: insights from phylogenetic analysis, protein structure prediction, molecular docking and molecular dynamics simulation, Journal of Biomolecular Structure and Dynamics. (2020). https://doi.org/10.1080/07391102.2020.1737234.
[115] L. Briganti, C. Capetti, V.O.A. Pellegrini, S. Ghio, E. Campos, A.S. Nascimento, I. Polikarpov, Structural and molecular dynamics investigations of ligand stabilization via secondary binding site interactions in Paenibacillus xylanivorans GH11 xylanase, Computational and Structural Biotechnology Journal. 19 (2021) 1557–1566. https://doi.org/10.1016/j.csbj.2021.03.002.
[116] W. Veldman, M.V. Liberato, V.P. Souza, V.M. Almeida, S.R. Marana, Ö. Tastan Bishop, I. Polikarpov, Differences in Gluco and Galacto Substrate-Binding Interactions in a Dual 6Pβ-Glucosidase/6Pβ-Galactosidase Glycoside Hydrolase 1 Enzyme from Bacillus licheniformis, Journal of Chemical Information and Modeling. 61 (2021) 4554–4570. https://doi.org/10.1021/ACS.JCIM.1C00413.

[117] A. Hospital, J.R. Goñi, M. Orozco, J.L. Gelpí, Molecular dynamics simulations: advances and applications, 8 (2015) 37. https://doi.org/10.2147/AABC.S70333.
[118] K. Lindorff-Larsen, S. Piana, R.O. Dror, D.E. Shaw, How fast-folding proteins fold, Science. 334 (2011) 517–520. https://doi.org/10.1126/SCIENCE.1208351/SUPPL_FILE/LINDORFF-LARSEN_SOM-REVISION1.PDF.
[119] D.E. Shaw, M.M. Deneroff, R.O. Dror, J.S. Kuskin, R.H. Larson, J.K. Salmon, C. Young, B. Batson, K.J. Bowers, J.C. Chao, M.P. Eastwood, J. Gagliardo, J.P. Grossman, C.R. Ho, D.J. Lerardi, I. Kolossváry, J.L. Klepeis, T. Layman, C. McLeavey, M.A. Moraes, R. Mueller, E.C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S.C. Wang, Anton, a special-purpose machine for molecular dynamics simulation, Communications of the ACM. 51 (2008) 91–97. https://doi.org/10.1145/1364782.1364802.
[120] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y. Shan, W. Wriggers, Atomic-level characterization of the structural dynamics of proteins, Science (New York, N.Y.). 330 (2010) 341–346. https://doi.org/10.1126/SCIENCE.1187409.
[121] F. Zheng, J. v. Vermaas, J. Zheng, Y. Wang, T. Tu, X. Wang, X. Xie, B. Yao, G.T. Beckham, H. Luo, Activity and thermostability of GH5 endoglucanase chimeras from mesophilic and thermophilic parents, Applied and Environmental Microbiology. (2018). https://doi.org/10.1128/AEM.02079-18.
[122] M.C. Childers, V. Daggett, Insights from molecular dynamics simulations for computational protein design, Molecular Systems Design and Engineering. (2017). https://doi.org/10.1039/c6me00083e.
[123] M.V. Liberato, E.T. Prates, T.A. Gonçalves, A. Bernardes, N. Vilela, J. Fattori, G.C. Ematsu, M. Chinaglia, E.R. Machi Gomes, A.C. Migliorini Figueira, A. Damasio, I. Polikarpov, M.S. Skaf, F.M. Squina, Insights into the dual cleavage activity of the GH16 laminarinase enzyme class on β-1,3 and β-1,4 glycosidic bonds, Journal of Biological Chemistry. (2021). https://doi.org/10.1016/j.jbc.2021.100385.
[124] K. Brodsky, M. Kutý, H. Pelantová, J. Cvačka, M. Rebroš, M. Kotik, I.K. Smatanová, V. Křen, P. Bojarová, Dual substrate specificity of the rutinosidase from aspergillus niger and the role of its substrate tunnel, International Journal of Molecular Sciences. (2020). https://doi.org/10.3390/ijms21165671.
[125] B. Veith, C. Herzberg, S. Steckel, J. Feesche, K.H. Maurer, P. Ehrenreich, S. Bäumer, A. Henne, H. Liesegang, R. Merkl, A. Ehrenreich, G. Gottschalk, The complete genome sequence of Bacillus licheniformis DSM13, an organism with great industrial potential, Journal of Molecular Microbiology and Biotechnology. (2004). https://doi.org/10.1159/000079829.
[126] M. Sakka, S. Tachino, H. Katsuzaki, J.S. van Dyk, B.I. Pletschke, T. Kimura, K. Sakka, Characterization of Xyn30A and Axh43A of Bacillus licheniformis SVD1 identified by its genomic analysis, Enzyme and Microbial Technology. (2012). https://doi.org/10.1016/j.enzmictec.2012.06.003.
[127] E.G.S. Farro, A.E.T. Leite, I.A. Silva, J.G. Filgueiras, E.R. de Azevedo, I. Polikarpov, A.S. Nascimento, GH43 endo-arabinanase from Bacillus licheniformis: Structure,

activity and unexpected synergistic effect on cellulose enzymatic hydrolysis, International Journal of Biological Macromolecules. (2018). https://doi.org/10.1016/j.ijbiomac.2018.05.157.
[128] C.K. Bandi, A. Agrawal, S.P. Chundawat, Carbohydrate-Active enZyme (CAZyme) enabled glycoengineering for a sweeter future, Current Opinion in Biotechnology. (2020). https://doi.org/10.1016/j.copbio.2020.09.006.
[129] D. Cai, Y. Rao, Y. Zhan, Q. Wang, S. Chen, Engineering Bacillus for efficient production of heterologous protein: current progress, challenge and prospect, Journal of Applied Microbiology. 126 (2019) 1632–1642. https://doi.org/10.1111/jam.14192.
[130] C.W. Song, R. Chelladurai, J.M. Park, H. Song, Engineering a newly isolated Bacillus licheniformis strain for the production of (2R,3R)-butanediol, Journal of Industrial Microbiology and Biotechnology. (2020). https://doi.org/10.1007/s10295-019-02249-4.
[131] J.-Y. Peng, Y.-B. Horng, C.-H. Wu, C.-Y. Chang, Y.-C. Chang, P.-S. Tsai, C.-R. Jeng, Y.-H. Cheng, H.-W. Chang, Evaluation of antiviral activity of Bacillus licheniformis-fermented products against porcine epidemic diarrhea virus, AMB Express. 9 (2019) 191. https://doi.org/10.1186/s13568-019-0916-0.
[132] Y. Xu, Y. Li, L. Zhang, Z. Ding, Z. Gu, G. Shi, Unraveling the specific regulation of the shikimate pathway for tyrosine accumulation in Bacillus licheniformis, Journal of Industrial Microbiology and Biotechnology. 46 (2019) 1047–1059. https://doi.org/10.1007/s10295-019-02213-2.
[133] L. Li, K. Li, K. Wang, C. Chen, C. Gao, C. Ma, P. Xu, Efficient production of 2,3-butanediol from corn stover hydrolysate by using a thermophilic Bacillus licheniformis strain, Bioresource Technology. (2014). https://doi.org/10.1016/j.biortech.2014.07.101.
[134] W. Kundig, S. Ghosh, S. Roseman, PHOSPHATE BOUND TO HISTIDINE IN A PROTEIN AS AN INTERMEDIATE IN A NOVEL PHOSPHO-TRANSFERASE SYSTEM, Biochemistry. (1964).
[135] A. Pickl, U. Johnsen, P. Schönheit, Fructose degradation in the haloarchaeon Haloferax volcanii involves a bacterial type phosphoenolpyruvate-dependent phosphotransferase system, fructose-1-phosphate kinase, and class II Fructose-1,6-bisphosphate aldolase, Journal of Bacteriology. (2012). https://doi.org/10.1128/JB.00200-12.
[136] J. Deutscher, C. Francke, P.W. Postma, How Phosphotransferase System-Related Protein Phosphorylation Regulates Carbohydrate Metabolism in Bacteria, Microbiology and Molecular Biology Reviews. (2006). https://doi.org/10.1128/mmbr.00024-06.
[137] B. Erni, The bacterial phosphoenolpyruvate: Sugar phosphotransferase system (PTS): An interface between energy and signal transduction, Journal of the Iranian Chemical Society. (2013). https://doi.org/10.1007/s13738-012-0185-1.
[138] P.W. Postma, J.W. Lengeler, G.R. Jacobson, Phosphoenolpyruvate: Carbohydrate phosphotransferase systems of bacteria, Microbiological Reviews. (1993).

[139] J. Deutscher, F.M.D. Aké, M. Derkaoui, A.C. Zébré, T.N. Cao, H. Bouraoui, T. Kentache, A. Mokhtari, E. Milohanic, P. Joyet, The Bacterial Phosphoenolpyruvate:Carbohydrate Phosphotransferase System: Regulation by Protein Phosphorylation and Phosphorylation-Dependent Protein-Protein Interactions, Microbiology and Molecular Biology Reviews. (2014). https://doi.org/10.1128/mmbr.00001-14.
[140] J. Plumbridge, Regulation of gene expression in the PTS in Escherichia coli: The role and interactions of MIc, Current Opinion in Microbiology. (2002). https://doi.org/10.1016/S1369-5274(02)00296-5.
[141] M.H. Saier, J. Reizer, Domain shuffling during evolution of the proteins of the bacterial phosphotransferase system, Research in Microbiology. (1990). https://doi.org/10.1016/0923-2508(90)90077-4.
[142] M.H. Saier, The Bacterial Phosphotransferase System: New Frontiers 50 Years after Its Discovery, Journal of Molecular Microbiology and Biotechnology. (2015). https://doi.org/10.1159/000381215.
[143] J. Larsbrink, T.E. Rogers, G.R. Hemsworth, L.S. McKee, A.S. Tauzin, O. Spadiut, S. Klinter, N.A. Pudlo, K. Urs, N.M. Koropatkin, A.L. Creagh, C.A. Haynes, A.G. Kelly, S.N. Cederholm, G.J. Davies, E.C. Martens, H. Brumer, A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes, Nature. (2014). https://doi.org/10.1038/nature12907.
[144] A.B. Boraston, D.N. Bolam, H.J. Gilbert, G.J. Davies, Carbohydrate-binding modules: Fine-tuning polysaccharide recognition, Biochemical Journal. (2004). https://doi.org/10.1042/BJ20040892.
[145] O. Shoseyov, Z. Shani, I. Levy, Carbohydrate Binding Modules: Biochemical Properties and Novel Applications, Microbiology and Molecular Biology Reviews. (2006). https://doi.org/10.1128/mmbr.00028-05.
[146] A. Ardèvol, C. Rovira, Reaction Mechanisms in Carbohydrate-Active Enzymes: Glycoside Hydrolases and Glycosyltransferases. Insights from ab Initio Quantum Mechanics/Molecular Mechanics Dynamic Simulations, Journal of the American Chemical Society. (2015). https://doi.org/10.1021/jacs.5b01156.
[147] V. Lombard, H. Golaconda Ramulu, E. Drula, P.M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013, Nucleic Acids Research. (2014). https://doi.org/10.1093/nar/gkt1178.
[148] G.J. Davies, T.M. Gloster, B. Henrissat, Recent structural insights into the expanding world of carbohydrate-active enzymes, Current Opinion in Structural Biology. (2005). https://doi.org/10.1016/j.sbi.2005.10.008.
[149] G.W. Hart, R.J. Copeland, Glycomics hits the big time, Cell. (2010). https://doi.org/10.1016/j.cell.2010.11.008.
[150] T.M. Gloster, D.J. Vocadlo, Developing inhibitors of glycan processing enzymes as tools for enabling glycobiology, Nature Chemical Biology. (2012). https://doi.org/10.1038/nchembio.1029.
[151] R.J. Woods, M.B. Tessier, Computational glycoscience: characterizing the spatial and temporal properties of glycans and glycan-protein complexes, Current Opinion in Structural Biology. (2010). https://doi.org/10.1016/j.sbi.2010.07.005.

[152] R. Stick, S.J. Williams, Carbohydrates: The Essential Molecules of Life, 2009. https://doi.org/10.1016/B978-0-240-52118-3.X0001-4.
[153] M. Sinnott, Carbohydrate chemistry and biochemistry: structure and mechanism, Choice Reviews Online. (2008). https://doi.org/10.5860/choice.45-5590.
[154] R.A. Laine, Invited commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 10 structures for a reducing hexasaccharide: The Isomer Barrier to development of single-method saccharide sequencing or synthesis systems, Glycobiology. (1994). https://doi.org/10.1093/glycob/4.6.759.
[155] T. Lütteke, A. Bohne-Lang, A. Loss, T. Goetz, M. Frank, C.W. von der Lieth, GLYCOSCIENCES.de: An internet portal to support glycomics and glycobiology research, Glycobiology. (2006). https://doi.org/10.1093/glycob/cwj049.
[156] S. Pérez, A. Sarkar, A. Rivet, C. Breton, A. Imberty, Glyco3d: A portal for structural glycosciences, Methods in Molecular Biology. (2015). https://doi.org/10.1007/978-1-4939-2343-4_18.
[157] M.M. Kuttel, J. Ståhle, G. Widmalm, CarbBuilder: Software for building molecular models of complex oligo- and polysaccharide structures, Journal of Computational Chemistry. (2016). https://doi.org/10.1002/jcc.24428.
[158] B. Henrissat, A classification of glycosyl hydrolases based on amino acid sequence similarities, Biochemical Journal. (1991). https://doi.org/10.1042/bj2800309.
[159] B.I. Cantarel, P.M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, B. Henrissat, The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics, Nucleic Acids Research. (2009). https://doi.org/10.1093/nar/gkn663.
[160] M.R. Stam, E.G.J. Danchin, C. Rancurel, P.M. Coutinho, B. Henrissat, Dividing the large glycoside hydrolase family 13 into subfamilies: Towards improved functional annotations of α-amylase-related proteins, Protein Engineering, Design and Selection. (2006). https://doi.org/10.1093/protein/gzl044.
[161] F.J. St John, J.M. González, E. Pozharski, Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups, FEBS Letters. (2010). https://doi.org/10.1016/j.febslet.2010.09.051.
[162] H. Aspeborg, P.M. Coutinho, Y. Wang, H. Brumer, B. Henrissat, Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5), BMC Evolutionary Biology. (2012). https://doi.org/10.1186/1471-2148-12-186.
[163] K. Mewis, N. Lenfant, V. Lombard, B. Henrissat, Dividing the large glycoside hydrolase family 43 into subfamilies: A motivation for detailed enzyme characterization, Applied and Environmental Microbiology. (2016). https://doi.org/10.1128/AEM.03453-15.
[164] J.R.K. Cairns, A. Esen, β-Glucosidases, Cellular and Molecular Life Sciences. (2010). https://doi.org/10.1007/s00018-010-0399-2.
[165] G.J. Davies, M.L. Sinnott, Sorting the diverse: the sequence-based classifications of carbohydrate-active enzymes, Biochemical Journal. (2008). https://doi.org/10.1042/bj20080382.
[166] S.C. Chen, K.J. Duan, Production of galactooligosaccharides using β-galactosidase immobilized on chitosan-coated magnetic nanoparticles with

tris(hydroxymethyl)phosphine as an optional coupling agent, International Journal of Molecular Sciences. (2015). https://doi.org/10.3390/ijms160612499.
[167] Q. Husain, β Galactosidases and their potential applications: A review, Critical Reviews in Biotechnology. (2010). https://doi.org/10.3109/07388550903330497.
[168] B. Henrissat, I. Callebaut, S. Fabrega, P. Lehn, J.P. Mornon, G. Davies, Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases, Proceedings of the National Academy of Sciences of the United States of America. (1995). https://doi.org/10.1073/pnas.92.15.7090.
[169] F.A. Quiocho, Carbohydrate-binding proteins: Tertiary structures and protein-sugar interactions, Annual Review of Biochemistry. (1986). https://doi.org/10.1146/annurev.bi.55.070186.001443.
[170] G.J. Davies, K.S. Wilson, B. Henrissat, Nomenclature for sugar-binding subsites in glycosyl hydrolases, Biochemical Journal. 321 (1997) 557–559. https://doi.org/10.1042/bj3210557.
[171] J. Gebler, N.R. Gilkes, M. Claeyssens, D.B. Wilson, P. Beguin, W.W. Wakarchuk, D.G. Kilburn, R.C. Miller, R.A.J. Warren, S.G. Withers, Stereoselective hydrolysis catalyzed by related β-1,4-glucanases and β- 1,4-xylanases, Journal of Biological Chemistry. (1992). https://doi.org/10.1016/s0021-9258(18)42313-7.
[172] T.M. Gloster, J.P. Turkenburg, J.R. Potts, B. Henrissat, G.J. Davies, Divergence of Catalytic Mechanism within a Glycosidase Family Provides Insight into Evolution of Carbohydrate Metabolism by Human Gut Flora, Chemistry and Biology. (2008). https://doi.org/10.1016/j.chembiol.2008.09.005.
[173] V.L.Y. Yip, J. Thompson, S.G. Withers, Mechanism of GlvA from Bacillus subtilis: A detailed kinetic analysis of a 6-phospho-α-glucosidase from glycoside hydrolase family 4, Biochemistry. (2007). https://doi.org/10.1021/bi700536p.
[174] S.A.K. Jongkees, S.G. Withers, Unusual enzymatic glycoside cleavage mechanisms, Accounts of Chemical Research. (2014). https://doi.org/10.1021/ar4001313.
[175] K. Michalska, K. Tan, H. Li, C. Hatzos-Skintges, J. Bearden, G. Babnigg, A. Joachimiak, GH1-family 6-P-β-glucosidases from human microbiome lactic acid bacteria, Acta Crystallographica Section D: Biological Crystallography. D (2013) 451–463. https://doi.org/10.1107/S0907444912049608.
[176] N. Srivastava, R. Rathour, S. Jha, K. Pandey, M. Srivastava, V.K. Thakur, R.S. Sengar, V.K. Gupta, P.B. Mazumder, A.F. Khan, P.K. Mishra, Microbial Beta Glucosidase Enzymes: Recent Advances in Biomass Conversation for Biofuels Application, Biomolecules. 9 (2019) 220. https://doi.org/10.3390/biom9060220.
[177] S.R. Marana, Molecular basis of substrate specificity in family 1 glycoside hydrolases, IUBMB Life. 58 (2006) 63–73. https://doi.org/10.1080/15216540600617156.
[178] A.D. Hill, P.J. Reilly, Computational analysis of glycoside hydrolase family 1 specificities, Biopolymers. (2008). https://doi.org/10.1002/bip.21052.
[179] R. Caspi, R. Billington, C.A. Fulcher, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, P.E. Midford, Q. Ong, W.K. Ong, S. Paley, P. Subhraveti, P.D. Karp,

The MetaCyc database of metabolic pathways and enzymes, Nucleic Acids Research. 46 (2018) D633–D639. https://doi.org/10.1093/nar/gkx935.
[180] D.E. Koshland, Stereochemistry and the mechanism of enzymatic reactions, Biological Reviews. (2008). https://doi.org/10.1111/j.1469-185x.1953.tb01386.x.
[181] M. Czjzek, M. Cicek, V. Zamboni, W.P. Burmeister, D.R. Bevan, B. Henrissat, A. Esen, Crystal structure of a monocotyledon (maize ZMGlu1) β-glucosidase and a model of its complex with p-nitrophenyl β-D-thioglucoside, Biochemical Journal. 354 (2001) 37–46. https://doi.org/10.1042/0264-6021:3540037.
[182] W. Chuenchor, S. Pengthaisong, R.C. Robinson, J. Yuvaniyama, W. Oonanant, D.R. Bevan, A. Esen, C.J. Chen, R. Opassiri, J. Svasti, J.R.K. Cairns, Structural Insights into Rice BGlu1 β-Glucosidase Oligosaccharide Hydrolysis and Transglycosylation, Journal of Molecular Biology. 377 (2008) 1200–1215. https://doi.org/10.1016/j.jmb.2008.01.076.
[183] C. Wiesmann, W. Hengstenberg, G.E. Schulz, Crystal structures and mechanism of 6-phospho-β-galactosidase from Lactococcus lactis, Journal of Molecular Biology. (1997). https://doi.org/10.1006/jmbi.1997.1084.
[184] W.P. Burmeister, S. Cottaz, H. Driguez, R. Iori, S. Palmieri, B. Henrissat, The crystal structures of Sinapis alba myrosinase and a covalent glycosyl-enzyme intermediate provide insights into the substrate recognition and active-site machinery of an S-glycosidase, Structure. 5 (1997) 663–676. https://doi.org/10.1016/s0969-2126(97)00221-9.
[185] M. Cicek, D. Blanchard, D.R. Bevan, A. Esen, The aglycone specificity-determining sites are different in 2,4- dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA)-glucosidase (maize β- glucosidase) and dhurrinase (sorghum β-glucosidase), Journal of Biological Chemistry. 275 (2000) 20002–20011. https://doi.org/10.1074/jbc.M001609200.
[186] D.L. Zechel, A.B. Boraston, T. Gloster, C.M. Boraston, J.M. Macdonald, D.M.G. Tilbrook, R. v. Stick, G.J. Davies, Iminosugar Glycosidase Inhibitors: Structural and Thermodynamic Dissection of the Binding of Isofagomine and 1-Deoxynojirimycin to β-Glucosidases, Journal of the American Chemical Society. 125 (2003) 14313–14323. https://doi.org/10.1021/ja036833h.
[187] L. Verdoucq, J. Morinière, D.R. Bevan, A. Esen, A. Vasella, B. Henrissat, M. Czjzek, Structural determinants of substrate specificity in family 1 β-glucosidases. Novel insights from the crystal structure of Sorghum dhurrinase-1, a plant β-glucosidase with strict specificity, in complex with its natural substrate, Journal of Biological Chemistry. 279 (2004) 31796–31803. https://doi.org/10.1074/jbc.M402918200.
[188] W.Y. Jeng, N.C. Wang, M.H. Lin, C.T. Lin, Y.C. Liaw, W.J. Chang, C.I. Liu, P.H. Liang, A.H.J. Wang, Structural and functional analysis of three β-glucosidases from bacterium Clostridium cellulovorans, fungus Trichoderma reesei and termite Neotermes koshunensis, Journal of Structural Biology. 173 (2011) 46–56. https://doi.org/10.1016/j.jsb.2010.07.008.
[189] K.H. Nam, M.W. Sung, K.Y. Hwang, Structural insights into the substrate recognition properties of β-glucosidase, Biochemical and Biophysical Research Communications. 391 (2010) 1131–1135. https://doi.org/10.1016/j.bbrc.2009.12.038.

[190] P. Isorna, J. Polaina, L. Latorre-García, F.J. Cañada, B. González, J. Sanz-Aparicio, Crystal Structures of Paenibacillus polymyxa β-Glucosidase B Complexes Reveal the Molecular Basis of Substrate Specificity and Give New Insights into the Catalytic Machinery of Family I Glycosidases, Journal of Molecular Biology. 371 (2007) 1204–1218. https://doi.org/10.1016/j.jmb.2007.05.082.
[191] T.M. Gloster, S. Roberts, V.M.A. Ducros, G. Perugino, M. Rossi, R. Hoos, M. Moracci, A. Vasella, G.J. Davies, Structural studies of the β-glycosidase from Sulfolobus solfataricus in complex with covalently and noncovalently bound inhibitors, Biochemistry. 43 (2004) 6101–6109. https://doi.org/10.1021/bi049666m.
[192] S.R. Marana, M. Jacobs-Lorena, W.R. Terra, C. Ferreira, Amino acid residues involved in substrate binding and catalysis in an insect digestive β-glycosidase, Biochimica et Biophysica Acta - Protein Structure and Molecular Enzymology. 1545 (2001) 41–52. https://doi.org/10.1016/S0167-4838(00)00260-0.
[193] S. Tribolo, J.G. Berrin, P.A. Kroon, M. Czjzek, N. Juge, The Crystal Structure of Human Cytosolic β-Glucosidase Unravels the Substrate Aglycone Specificity of a Family 1 Glycoside Hydrolase, Journal of Molecular Biology. 370 (2007) 964–975. https://doi.org/10.1016/j.jmb.2007.05.034.
[194] S. Fiorucci, J. Golebiowski, D. Cabrol-Bass, S. Antonczak, Molecular simulations bring new insights into flavonoid/quercetinase interaction modes, Proteins: Structure, Function and Genetics. 67 (2007) 961–970. https://doi.org/10.1002/prot.21380.
[195] S. Fiorucci, J. Golebiowski, D. Cabrol-Bass, S. Antonczak, Molecular simulations enlighten the binding mode of quercetin to lipoxygenase-3, Proteins: Structure, Function and Genetics. 73 (2008) 290–298. https://doi.org/10.1002/prot.22179.
[196] B. Christelle, B.D.O. Eduardo, C. Latifa, M. Elaine-Rose, M. Bernard, R.H. Evelyne, G. Mohamed, E. Jean-Marc, H. Catherine, Combined docking and molecular dynamics simulations to enlighten the capacity of Pseudomonas cepacia and Candida antarctica lipases to catalyze quercetin acetylation, Journal of Biotechnology. 156 (2011) 203–210. https://doi.org/10.1016/j.jbiotec.2011.09.007.
[197] W. Chuenchor, S. Pengthaisong, R.C. Robinson, J. Yuvaniyama, J. Svasti, J.R.K. Cairns, The structural basis of oligosaccharide binding by rice BGlu1 beta-glucosidase, Journal of Structural Biology. 173 (2011) 169–179. https://doi.org/10.1016/j.jsb.2010.09.021.
[198] S. Sansenya, R. Opassiri, B. Kuaprasert, C.J. Chen, J.R. Ketudat Cairns, The crystal structure of rice (Oryza sativa L.) Os4BGlu12, an oligosaccharide and tuberonic acid glucoside-hydrolyzing β-glucosidase with significant thioglucohydrolase activity, Archives of Biochemistry and Biophysics. 510 (2011) 62–72. https://doi.org/10.1016/j.abb.2011.04.005.
[199] S. Lansky, A. Zehavi, H. Belrhali, Y. Shoham, G. Shoham, Structural basis for enzyme bifunctionality – the case of Gan1D from Geobacillus stearothermophilus, FEBS Journal. 284 (2017) 3931–3953. https://doi.org/10.1111/febs.14283.
[200] M. Czjzek, M. Cicek, V. Zamboni, D.R. Bevan, B. Henrissat, A. Esen, The mechanism of substrate (aglycone) specificity in β-glucosidases is revealed by crystal structures of mutant maize β-glucosidase-DIMBOA, -DIMBOAGlc, and -dhurrin complexes, Proceedings of the National Academy of Sciences of the United

States of America. 97 (2000) 13555–13560. https://doi.org/10.1073/pnas.97.25.13555.
[201] W.-L. Yu, Y.-L. Jiang, A. Pikis, W. Cheng, X.-H. Bai, Y.-M. Ren, J. Thompson, C.-Z. Zhou, Y. Chen, Structural insights into the substrate specificity of a 6-phospho-β-glucosidase BglA-2 from Streptococcus pneumoniae TIGR4., The Journal of Biological Chemistry. 288 (2013) 14949–58. https://doi.org/10.1074/jbc.M113.454751.
[202] M. Totir, N. Echols, M. Nanao, C.L. Gee, A. Moskaleva, S. Gradia, A.T. Iavarone, J.M. Berger, A.P. May, C. Zubieta, T. Alber, Macro-to-micro structural proteomics: Native source proteins for high-throughput crystallization, PLoS ONE. 7 (2012) 32498. https://doi.org/10.1371/journal.pone.0032498.
[203] D.H. Kwan, Y. Jin, J. Jiang, H.M. Chen, M.P. Kötzler, H.S. Overkleeft, G.J. Davies, S.G. Withers, Chemoenzymatic synthesis of 6-phospho-cyclophellitol as a novel probe of 6-phospho-β-glucosidases, FEBS Letters. 590 (2016) 461–468. https://doi.org/10.1002/1873-3468.12059.
[204] D. Schulte, W. Hengstenberg, Engineering the active center of the 6-phospho-β-galactosidase from Lactococcus lactis, Protein Engineering. 13 (2000) 515–518. https://doi.org/10.1093/protein/13.7.515.
[205] Z. A, Biochemical characterization and structure-function analysis of 6-phospho-b-glycosidases from Geobacillus stearothermophilus, 2015.
[206] B. Chowdhury, G. Garai, A review on multiple sequence alignment from the perspective of genetic algorithm, Genomics. 109 (2017) 419–431. https://doi.org/10.1016/j.ygeno.2017.06.007.
[207] M. Chatzou, C. Magis, J.M. Chang, C. Kemena, G. Bussotti, I. Erb, C. Notredame, Multiple sequence alignment modeling: Methods and applications, Briefings in Bioinformatics. 17 (2016) 1009–1023. https://doi.org/10.1093/BIB/BBV099.
[208] M.T. Pervez, M.E. Babar, A. Nadeem, M. Aslam, A. Razaawan, N. Aslam, T. Hussain, N. Naveed, S. Qadri, U. Waheed, M. Shoaib, Evaluating the accuracy and effciency of multiple sequence alignment methods, Evolutionary Bioinformatics. 10 (2014) 205–217. https://doi.org/10.4137/EBo.s19199.
[209] J. Xiong, Essential Bioinformatics, 2006. https://doi.org/https://doi.org/10.1017/CBO9780511806087.
[210] B. Morgenstern, S.J. Prohaska, D. Pöhler, P.F. Stadler, Multiple sequence alignment with user-defined anchor points, Algorithms for Molecular Biology. 1 (2006) 1–12. https://doi.org/10.1186/1748-7188-1-6.
[211] P.K. Busk, L. Lange, Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs, Applied and Environmental Microbiology. (2013). https://doi.org/10.1128/AEM.03803-12.
[212] C. Kim, B. Lee, Accuracy of structure-based sequence alignment of automatic methods, BMC Bioinformatics. 8 (2007) 1–17. https://doi.org/10.1186/1471-2105-8-355.
[213] W. Huang, D.M. Umbach, L. Li, Accurate anchoring alignment of divergent sequences, Bioinformatics. 22 (2006) 29–34. https://doi.org/10.1093/bioinformatics/bti772.

[214] P. Lakshmi, N.J., Gavarraju, P., Jeevana, J.K., Karteeka, A Literature Survey on Multiple Sequence Alignment Algorithms, Int. J. Adv. Res. Comput. Sci. Softw. Eng. 6 (2016) 280–288.
[215] I.M. Wallace, G. Blackshields, D.G. Higgins, Multiple sequence alignments, Current Opinion in Structural Biology. 15 (2005) 261–266. https://doi.org/10.1016/j.sbi.2005.04.002.
[216] C. Notredame, Recent evolutions of multiple sequence alignment algorithms, PLoS Computational Biology. 3 (2007) 1405–1408. https://doi.org/10.1371/journal.pcbi.0030123.
[217] C. Kemena, C. Notredame, Upcoming challenges for multiple sequence alignment methods in the high-throughput era, Bioinformatics. 25 (2009) 2455–2465. https://doi.org/10.1093/bioinformatics/btp452.
[218] J. Daugelaite, A. O' Driscoll, R.D. Sleator, An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics, ISRN Biomathematics. 2013 (2013) 1–14. https://doi.org/10.1155/2013/615630.
[219] S. Kumar, A. Filipski, Multiple sequence alignment: In pursuit of homologous DNA positions, Genome Research. 17 (2007) 127–135. https://doi.org/10.1101/gr.5232407.
[220] M.O. Dayhoff, M.O. Dayhoff, R.M. Schwartz, Chapter 22: A model of evolutionary change in proteins, IN ATLAS OF PROTEIN SEQUENCE AND STRUCTURE. (1978). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315 (accessed May 16, 2021).
[221] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proceedings of the National Academy of Sciences of the United States of America. 89 (1992) 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.
[222] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, Journal of Molecular Biology. 48 (1970) 443–453. https://doi.org/10.1016/0022-2836(70)90057-4.
[223] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, Journal of Molecular Biology. 147 (1981) 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
[224] J.D. Thompson, D.G. Higgins, T.J. Gibson, Improved sensitivity of profile searches through the use of sequence weights and gap excision, Bioinformatics. 10 (1994) 19–29. https://doi.org/10.1093/bioinformatics/10.1.19.
[225] C. Notredame, D.G. Higgins, J. Heringa, T-coffee: A novel method for fast and accurate multiple sequence alignment, Journal of Molecular Biology. 302 (2000) 205–217. https://doi.org/10.1006/jmbi.2000.4042.
[226] J. Heringa, W.R. Taylor, Three-dimensional domain duplication, swapping and stealing, Current Opinion in Structural Biology. 7 (1997) 416–421. https://doi.org/10.1016/S0959-440X(97)80060-7.
[227] B. Morgenstern, DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment, in: Bioinformatics, Bioinformatics, 1999: pp. 211–218. https://doi.org/10.1093/bioinformatics/15.3.211.

[228] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton, Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment, Science. 262 (1993) 208–214. https://doi.org/10.1126/science.8211139.
[229] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers., Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology. 2 (1994) 28–36. http://europepmc.org/article/MED/7584402 (accessed May 17, 2021).
[230] D.F. Feng, R.F. Doolittle, Progressive sequence alignment as a prerequisitetto correct phylogenetic trees, Journal of Molecular Evolution. 25 (1987) 351–360. https://doi.org/10.1007/BF02603120.
[231] W.J. Wilbur, D.J. Lipman, Rapid similarity searches of nucleic acid and protein data banks., Proceedings of the National Academy of Sciences of the United States of America. 80 (1983) 726–730. https://doi.org/10.1073/pnas.80.3.726.
[232] R.C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics. 5 (2004) 1–19. https://doi.org/10.1186/1471-2105-5-113.
[233] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology. 7 (2011) 539. https://doi.org/10.1038/msb.2011.75.
[234] I. Gronau, S. Moran, Optimal implementations of UPGMA and other common clustering algorithms, Information Processing Letters. 104 (2007) 205–210. https://doi.org/10.1016/j.ipl.2007.07.002.
[235] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research. 22 (1994) 4673–4680. https://doi.org/10.1093/nar/22.22.4673.
[236] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability, Molecular Biology and Evolution. 30 (2013) 772–780. https://doi.org/10.1093/molbev/mst010.
[237] W.R. Taylor, Multiple sequence alignment by a pairwise algorithm, Bioinformatics. 3 (1987) 81–87. https://doi.org/10.1093/bioinformatics/3.2.81.
[238] J. Pei, R. Sadreyev, N. v. Grishin, PCMA: Fast and accurate multiple sequence alignment based on profile consistency, Bioinformatics. 19 (2003) 427–428. https://doi.org/10.1093/bioinformatics/btg008.
[239] F. Corpet, Multiple sequence alignment with hierarchical clustering, Nucleic Acids Research. 16 (1988) 10881–10890. https://doi.org/10.1093/nar/16.22.10881.
[240] T. Lassmann, E.L.L. Sonnhammer, Kalign - An accurate and fast multiple sequence alignment algorithm, BMC Bioinformatics. 6 (2005) 1–9. https://doi.org/10.1186/1471-2105-6-298.

[241] U. Roshan, D.R. Livesay, Probalign: Multiple sequence alignment using partition function posterior probabilities, Bioinformatics. 22 (2006) 2715–2721. https://doi.org/10.1093/bioinformatics/btl472.
[242] J. Pei, N. v. Grishin, PROMALS: Towards accurate multiple sequence alignments of distantly related proteins, Bioinformatics. 23 (2007) 802–808. https://doi.org/10.1093/bioinformatics/btm017.
[243] Y. Liu, B. Schmidt, D.L. Maskell, MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities, Bioinformatics. 26 (2010) 1958–1964. https://doi.org/10.1093/bioinformatics/btq338.
[244] O. Gotoh, Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments, Journal of Molecular Biology. 264 (1996) 823–838. https://doi.org/10.1006/jmbi.1996.0679.
[245] Y. Wang, K. bin Li, An adaptive and iterative algorithm for refining multiple sequence alignment, Computational Biology and Chemistry. 28 (2004) 141–148. https://doi.org/10.1016/j.compbiolchem.2004.02.001.
[246] O. Gotoh, Optimal alignment between groups of sequences and its application to multiple sequence alignment, Bioinformatics. 9 (1993) 361–370. https://doi.org/10.1093/bioinformatics/9.3.361.
[247] C. Notredame, D.G. Higgins, SAGA: Sequence alignment by genetic algorithm, Nucleic Acids Research. 24 (1996) 1515–1524. https://doi.org/10.1093/nar/24.8.1515.
[248] A.M. Lesk, C. Chothia, How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins, Journal of Molecular Biology. 136 (1980) 225–270. https://doi.org/10.1016/0022-2836(80)90373-3.
[249] O. O'Sullivan, K. Suhre, C. Abergel, D.G. Higgins, C. Notredame, 3DCoffee: Combining protein sequences and structures within multiple sequence alignments, Journal of Molecular Biology. 340 (2004) 385–395. https://doi.org/10.1016/j.jmb.2004.04.058.
[250] F. Armougom, S. Moretti, O. Poirot, S. Audic, P. Dumas, B. Schaeli, V. Keduas, C. Notredame, Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee, Nucleic Acids Research. 34 (2006) W604–W608. https://doi.org/10.1093/nar/gkl092.
[251] X. Xia, S. Zhang, Y. Su, Z. Sun, MICAlign: A sequence-to-structure alignment tool integrating multiple sources of information in conditional random fields, Bioinformatics. 25 (2009) 1433–1434. https://doi.org/10.1093/bioinformatics/btp251.
[252] J. Pei, N. v. Grishin, PROMALS3D: Multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information, Methods in Molecular Biology. (2014). https://doi.org/10.1007/978-1-62703-646-7_17.
[253] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley Longman Publishing Co., Inc.75 Arlington Street, Suite 300 Boston, MA United States, 1989.

[254] K.B. Gondro C, A simple genetic algorithm for multiple sequence alignment, Genet Mol Res. 6 (2007) 964–982.
[255] V. Lyubetsky, W.H. Piel, D. Quandt, Current advances in molecular phylogenetics, BioMed Research International. 2014 (2014). https://doi.org/10.1155/2014/596746.
[256] W.M. Fitch, Uses for evolutionary trees., Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 349 (1995) 93–102. https://doi.org/10.1098/rstb.1995.0095.
[257] H. Ellegren, Comparative genomics and the study of evolution by natural selection, Molecular Ecology. 17 (2008) 4586–4596. https://doi.org/10.1111/j.1365-294X.2008.03954.x.
[258] C.O. Webb, D.D. Ackerly, M.A. McPeek, M.J. Donoghue, Phylogenies and community ecology, Annual Review of Ecology and Systematics. 33 (2002) 475–505. https://doi.org/10.1146/annurev.ecolsys.33.010802.150448.
[259] S.C. Stearns, Evolutionary medicine: Its scope, interest and potential, Proceedings of the Royal Society B: Biological Sciences. 279 (2012) 4305–4321. https://doi.org/10.1098/rspb.2012.1326.
[260] K.A. Crandall, O.R.R. Bininda-Emonds, G.M. Mace, R.K. Wayne, Considering evolutionary processes in conservation biology, Trends in Ecology and Evolution. 15 (2000) 290–295. https://doi.org/10.1016/S0169-5347(00)01876-0.
[261] P.H. Harvey, A.J. Leigh Brown, J. Maynard Smith, S. Nee, New uses for old phylogenies, in: Oxford University Press, Oxford, UK, 1996.
[262] K.S. John, Review paper: The shape of phylogenetic treespace, in: Systematic Biology, Oxford University Press, 2017: pp. e83–e94. https://doi.org/10.1093/sysbio/syw025.
[263] S. Kumar, K. Tamura, M. Nei, MEGA: Molecular evolutionary genetics analysis software for microcomputers, Bioinformatics. 10 (1994) 189–191. https://doi.org/10.1093/bioinformatics/10.2.189.
[264] S. Kumar, M. Nei, J. Dudley, K. Tamura, MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences, Briefings in Bioinformatics. 9 (2008) 299–306. https://doi.org/10.1093/bib/bbn017.
[265] B.G. Hall, Building phylogenetic trees from molecular data with MEGA, Molecular Biology and Evolution. 30 (2013) 1229–1235. https://doi.org/10.1093/molbev/mst012.
[266] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods, Molecular Biology and Evolution. 28 (2011) 2731–2739. https://doi.org/10.1093/molbev/msr121.
[267] S.A.E.H. Mohamed, M. Elloumi, J.D. Thompson, Motif Discovery in Protein Sequences, in: Pattern Recognition - Analysis and Applications, InTech, 2016. https://doi.org/10.5772/65441.
[268] E. Elayaraja, K. Thangavel, B. Ramya, M. Chitralegha, Extraction of motif patterns from protein sequence using Rough-K-Means algorithm, in: Procedia Engineering, Elsevier, 2012: pp. 814–820. https://doi.org/10.1016/j.proeng.2012.01.932.

[269] P. Bork, E. v. Koonin, Protein sequence motifs, Current Opinion in Structural Biology. 6 (1996) 366–376. https://doi.org/10.1016/S0959-440X(96)80057-1.
[270] S. Henikoff, J.G. Henikoff, Position-based sequence weights, Journal of Molecular Biology. 243 (1994) 574–578. https://doi.org/10.1016/0022-2836(94)90032-9.
[271] G.A. Churchill, Stochastic models for heterogeneous DNA sequences, Bulletin of Mathematical Biology. 51 (1989) 79–94. https://doi.org/10.1007/BF02458837.
[272] T.K. Attwood, A. Coletta, G. Muirhead, A. Pavlopoulou, P.B. Philippou, I. Popov, C. Romá-Mateo, A. Theodosiou, A.L. Mitchell, The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012, Database. 2012 (2012). https://doi.org/10.1093/database/bas019.
[273] C.J.A. Sigrist, E. de Castro, L. Cerutti, B.A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Research. 41 (2013). https://doi.org/10.1093/nar/gks1067.
[274] H. Dinkel, K. Roey, S. Michael, M. Kumar, B. Uyar, B. Altenberg, V. Milchevskaya, M. Schneider, H. Kühn, A. Behrendt, S.L. Dahl, V. Damerell, S. Diebel, S. Kalman, S. Klein, A.C. Knudsen, C. Mäder, S. Merrill, A. Staudt, V. Thiel, L. Welti, N.E. Davey, F. Diella, T.J. Gibson, ELM 2016 - Data update and new functionality of the eukaryotic linear motif resource, Nucleic Acids Research. 44 (2016) D294–D300. https://doi.org/10.1093/nar/gkv1291.
[275] P. Tompa, N.E. Davey, T.J. Gibson, M.M. Babu, A Million peptide motifs for the molecular biologist, Molecular Cell. 55 (2014) 161–169. https://doi.org/10.1016/j.molcel.2014.05.032.
[276] J. Keith, ed., Bioinformatics, in: Volume I, Humana Press, New York, USA, 2008: p. 562.
[277] Y. Zhang, P. Wang, M. Yan, An Entropy-Based Position Projection Algorithm for Motif Discovery, BioMed Research International. 2016 (2016). https://doi.org/10.1155/2016/9127474.
[278] T.L. Bailey, M. Gribskov, Combining evidence using p-values: Application to sequence homology searches, Bioinformatics. (1998). https://doi.org/10.1093/bioinformatics/14.1.48.
[279] K. Illergård, D.H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence--a study of structural response in protein cores, Proteins. 77 (2009) 499–508. https://doi.org/10.1002/PROT.22458.
[280] R. Kolodny, P. Koehl, M. Levitt, Comprehensive evaluation of protein structure alignment methods: Scoring by geometric measures, Journal of Molecular Biology. 346 (2005) 1173–1188. https://doi.org/10.1016/j.jmb.2004.12.032.
[281] D. Sehnal, R. Svobodová, K. Berka, L. Pravda, A. Midlik, J. Koča, Visualization and Analysis of Protein Structures with LiteMol Suite, in: Methods in Molecular Biology, Humana Press Inc., 2020: pp. 1–13. https://doi.org/10.1007/978-1-0716-0270-6_1.
[282] L. Holm, Using Dali for Protein Structure Comparison, in: Methods in Molecular Biology, Humana Press Inc., 2020: pp. 29–42. https://doi.org/10.1007/978-1-0716-0270-6_3.

[283] F.J. Burkowski, Structural Bioinformatics: An Algorithmic Approach, CRC Press, Taylor & Francis Group 4th, Floor, Albert House, 1-4 Singer Street, London, EC2A 4BQ UK, n.d.
[284] S.I. O'Donoghue, D.S. Goodsell, A.S. Frangakis, F. Jossinet, R.A. Laskowski, M. Nilges, H.R. Saibil, A. Schafferhans, R.C. Wade, E. Westhof, A.J. Olson, Visualization of macromolecular structures, Nature Methods. 7 (2010) 1427. https://doi.org/10.1038/nmeth.1427.
[285] S. Srivastava, S.B. Lal, D.C. Mishra, U.B. Angadi, K.K. Chaturvedi, S.N. Rai, A. Rai, An efficient algorithm for protein structure comparison using elastic shape analysis, Algorithms for Molecular Biology. 11 (2016) 27. https://doi.org/10.1186/s13015-016-0089-1.
[286] I. Kufareva, R. Abagyan, Methods of protein structure comparison, Methods in Molecular Biology. 857 (2012) 231–257. https://doi.org/10.1007/978-1-61779-588-6_10.
[287] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, G. Vriend, A database of protein structure families with common folding motifs, Protein Science. 1 (1992) 1691–1698. https://doi.org/10.1002/pro.5560011217.
[288] C.A. Orengo, F.M.G. Pearl, J.M. Thornton, The Cath Domain Structure Database, in: Structural Bioinformatics, John Wiley and Sons Inc., 2005: pp. 249–271. https://doi.org/10.1002/0471721204.ch13.
[289] I.N. Shindyalov, P.E. Bourne, An alternative view of protein fold space, Proteins: Structure, Function and Bioinformatics. (2000). https://doi.org/https://doi.org/10.1002/(SICI)1097-0134(20000215)38:3<247::AID-PROT2>3.0.CO;2-T.
[290] F.S. Domingues, P. Lackner, A. Andreeva, M.J. Sippl, Structure-based evaluation of sequence comparison and fold recognition alignment accuracy, Journal of Molecular Biology. 297 (2000) 1003–1013. https://doi.org/10.1006/jmbi.2000.3615.
[291] I. Friedberg, T. Kaplan, H. Margalit, Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments, Protein Science. 9 (2000) 2278–2284. https://doi.org/10.1110/ps.9.11.2278.
[292] A. Yonath, X-ray crystallography at the heart of life science, Current Opinion in Structural Biology. 21 (2011) 622–626. https://doi.org/10.1016/j.sbi.2011.07.005.
[293] M.A. Martí-Renom, A.C. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Šali, Comparative protein structure modeling of genes and genomes, Annual Review of Biophysics and Biomolecular Structure. 29 (2000) 291–325. https://doi.org/10.1146/annurev.biophys.29.1.291.
[294] E.P. Carpenter, K. Beis, A.D. Cameron, S. Iwata, Overcoming the challenges of membrane protein crystallography, Current Opinion in Structural Biology. 18 (2008) 581–586. https://doi.org/10.1016/j.sbi.2008.07.001.
[295] E. Ghosh, P. Kumari, D. Jaiman, A.K. Shukla, Methodological advances: The unsung heroes of the GPCR structural revolution, Nature Reviews Molecular Cell Biology. 16 (2015) 69–81. https://doi.org/10.1038/nrm3933.

[296] M.T. Muhammed, E. Aki-Yalcin, Homology modeling in drug discovery: Overview, current applications, and future perspectives, Chemical Biology and Drug Design. 93 (2019) 12–20. https://doi.org/10.1111/cbdd.13388.
[297] T. Schmidt, A. Bergner, T. Schwede, Modelling three-dimensional protein structures for applications in drug design, Drug Discovery Today. 19 (2014) 890–897. https://doi.org/10.1016/j.drudis.2013.10.027.
[298] N. Eswar, B. John, N. Mirkovic, A. Fiser, V.A. Ilyin, U. Pieper, A.C. Stuart, M.A. Marti-Renom, M.S. Madhusudhan, B. Yerkovich, A. Sali, Tools for comparative protein structure modeling and analysis, Nucleic Acids Research. 31 (2003) 3375–3380. https://doi.org/10.1093/nar/gkg543.
[299] S. Hongmao, Homology Modeling and Ligand-Based Molecule Design, in: A Practical Guide to Rational Drug Design, Elsevier, 2016: pp. 109–160. https://doi.org/10.1016/b978-0-08-100098-4.00004-1.
[300] A. Lesk, CASP2: report on ab initio predictions, Proteins. 29 (1997) 151–166. https://doi.org/10.1002/(sici)1097-0134(1997)1+<151::aid-prot20>3.3.co;2-j.
[301] C.N. Cavasotto, S.S. Phatak, Homology modeling in drug discovery: current trends and applications, Drug Discovery Today. 14 (2009) 676–683. https://doi.org/10.1016/j.drudis.2009.04.006.
[302] T. Werner, M.B. Morris, S. Dastmalchi, W.B. Church, Structural modelling and dynamics of proteins for insights into drug interactions, Advanced Drug Delivery Reviews. 64 (2012) 323–343. https://doi.org/10.1016/j.addr.2011.11.011.
[303] C. Chothia, A. Lesk, The relation between the divergence of sequence and structure in proteins, EMBO. 5 (1986) 823–6.
[304] A. Hillisch, L.F. Pineda, R. Hilgenfeld, Utility of homology models in the drug discovery process, Drug Discovery Today. 9 (2004) 659–669. https://doi.org/10.1016/S1359-6446(04)03196-4.
[305] A. Fiser, R.K.G. Do, A. Šali, Modeling of loops in protein structures, Protein Science. 9 (2000) 1753–1773. https://doi.org/10.1110/ps.9.9.1753.
[306] A. Saxena, R.S. Sangwan, S. Mishra, Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatics Tools: An Overview, Science International. 1 (2013) 237–252. https://doi.org/10.17311/sciintl.2013.237.252.
[307] J. Peng, Statistical inference for template-based protein structure prediction, 2013. http://arxiv.org/abs/1306.4420 (accessed May 31, 2021).
[308] J.A.R. Dalton, R.M. Jackson, An evaluation of automated homology modelling methods at low target-template sequence similarity, Bioinformatics. 23 (2007) 1901–1908. https://doi.org/10.1093/bioinformatics/btm262.
[309] H. Hasani, K. Barakat, Homology Modeling: an Overview of Fundamentals and Tools, International Review on Modelling and Simulations. 10 (2017).
[310] A.R. Katebi, A. Kloczkowski, R.L. Jernigan, Structural interpretation of protein-protein interaction network, BMC Structural Biology. 10 (2010). https://doi.org/10.1186/1472-6807-10-S1-S4.
[311] M. Wiltgen, G.P. Tilz, Homology modelling: Eine übersicht über die methode am beispiel der strukturbestimmung vom diabetes antigen GAD 65, Wiener

Medizinische Wochenschrift. 159 (2009) 112–125. https://doi.org/10.1007/s10354-009-0662-z.
[312] P.A. Bates, L.A. Kelley, R.M. MacCallum, M.J.E. Sternberg, Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM, Proteins: Structure, Function and Genetics. 45 (2001) 39–46. https://doi.org/10.1002/prot.1168.
[313] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling, Bioinformatics. 22 (2006) 195–201. https://doi.org/10.1093/bioinformatics/bti770.
[314] M. Levitt, Accurate modeling of protein conformation by automatic segment matching, Journal of Molecular Biology. 226 (1992) 507–533. https://doi.org/10.1016/0022-2836(92)90964-L.
[315] B. Webb, A. Sali, Protein structure modeling with MODELLER, in: Methods in Molecular Biology, 2017. https://doi.org/10.1007/978-1-4939-7231-9_4.
[316] A. Aszódi, W.R. Taylor, Secondary structure formation in model polypeptide chains, Protein Engineering, Design and Selection. 7 (1994) 633–644. https://doi.org/10.1093/protein/7.5.633.
[317] J.R. Allison, S. Hertig, J.H. Missimer, L.J. Smith, M.O. Steinmetz, J. Dolenc, Probing the structure and dynamics of proteins by combining molecular dynamics simulations and experimental NMR data, Journal of Chemical Theory and Computation. 8 (2012) 3430–3444. https://doi.org/10.1021/ct300393b.
[318] A. Kahraman, F. Herzog, A. Leitner, G. Rosenberger, R. Aebersold, L. Malmström, Cross-Link Guided Molecular Modeling with ROSETTA, PLoS ONE. 8 (2013) 73411. https://doi.org/10.1371/journal.pone.0073411.
[319] H. Venselaar, R.P. Joosten, B. Vroling, C.A.B. Baakman, M.L. Hekkelman, E. Krieger, G. Vriend, Homology modelling and spectroscopy, a never-ending love story, European Biophysics Journal. 39 (2010) 551–563. https://doi.org/10.1007/s00249-009-0531-0.
[320] Z. Xiang, Advances in Homology Protein Structure Modeling, Current Protein & Peptide Science. 7 (2006) 217–227. https://doi.org/10.2174/138920306777452312.
[321] D. Petrey, Z. Xiang, C.L. Tang, L. Xie, M. Gimpelev, T. Mitros, C.S. Soto, S. Goldsmith-Fischman, A. Kernytsky, A. Schlessinger, I.Y.Y. Koh, E. Alexov, B. Honig, Using Multiple Structure Alignments, Fast Model Building, and Energetic Analysis in Fold Recognition and Homology Modeling, in: Proteins: Structure, Function and Genetics, Proteins, 2003: pp. 430–435. https://doi.org/10.1002/prot.10550.
[322] S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A.E. Dawid, A. Kolinski, Coarse-Grained Protein Models and Their Applications, Chemical Reviews. 116 (2016) 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163.
[323] K. Tappura, Influence of rotational energy barriers to the conformational search of protein loops in molecular dynamics and ranking the conformations, Proteins: Structure, Function and Genetics. 44 (2001) 167–179. https://doi.org/10.1002/prot.1082.

[324] C.M. Deane, T.L. Blundell, Protein Comparative Modelling and Drug Discovery, in: The Practice of Medicinal Chemistry: Second Edition, Elsevier Inc., 2003: pp. 445–458. https://doi.org/10.1016/B978-012744481-9/50031-3.
[325] H.W.T. van Vlijmen, M. Karplus, PDB-based protein loop prediction: Parameters for selection and methods for optimization, Journal of Molecular Biology. 267 (1997) 975–1001. https://doi.org/10.1006/jmbi.1996.0857.
[326] C. Marks, J. Shi, C.M. Deane, Predicting loop conformational ensembles, Bioinformatics. 34 (2018) 949–956. https://doi.org/10.1093/BIOINFORMATICS/BTX718.
[327] A. Barozet, M. Bianciotto, M. Vaisset, T. Siméon, H. Minoux, J. Cortés, Protein loops with multiple meta-stable conformations: A challenge for sampling and scoring methods, Proteins. 89 (2021) 218–231. https://doi.org/10.1002/PROT.26008.
[328] W. Gao, S.P. Mahajan, J. Sulam, J.J. Gray, Deep Learning in Protein Structural Modeling and Design, Patterns. 1 (2020). https://doi.org/10.1016/J.PATTER.2020.100142.
[329] P. Gkeka, G. Stoltz, A. Barati Farimani, Z. Belkacemi, M. Ceriotti, J.D. Chodera, A.R. Dinner, A.L. Ferguson, J.B. Maillet, H. Minoux, C. Peter, F. Pietrucci, A. Silveira, A. Tkatchenko, Z. Trstanova, R. Wiewiora, T. Lelièvre, Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems, Journal of Chemical Theory and Computation. 16 (2020) 4757–4775. https://doi.org/10.1021/ACS.JCTC.0C00355/ASSET/IMAGES/ACS.JCTC.0C00355.SOCIAL.JPEG_V03.
[330] S.C. Pakhrin, B. Shrestha, B. Adhikari, D.B. Kc, Deep Learning-Based Advances in Protein Structure Prediction, International Journal of Molecular Sciences. 22 (2021). https://doi.org/10.3390/IJMS22115553.
[331] J.A. Ruffolo, C. Guerra, S.P. Mahajan, J. Sulam, J.J. Gray, Geometric potentials from deep learning improve prediction of CDR H3 loop structures, Bioinformatics. 36 (6372) i268–i275. https://doi.org/10.1093/BIOINFORMATICS/BTAA457.
[332] R. Samudrala, J. Moult, Determinants of side chain conformational preferences in protein structures, Protein Engineering. 11 (1998) 991–997. https://doi.org/10.1093/protein/11.11.991.
[333] W.E. Stites, A.K. Meeker, D. Shortle, Evidence for strained interactions between side-chains and the polypeptide backbone, Journal of Molecular Biology. 235 (1994) 27–32. https://doi.org/10.1016/S0022-2836(05)80008-7.
[334] R.L. Dunbrack, M. Karplus, Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains, Nature Structural Biology. 1 (1994) 334–340. https://doi.org/10.1038/nsb0594-334.
[335] J. Zhu, H. Fan, X. Periole, B. Honig, A.E. Mark, Refining homology models by combining replica-exchange molecular dynamics and statistical potentials, Proteins: Structure, Function and Genetics. 72 (2008) 1171–1188. https://doi.org/10.1002/prot.22005.

[336] O. Guvench, A.D. MacKerell, Comparison of protein force fields for molecular dynamics simulations, Methods in Molecular Biology. 443 (2008) 63–88. https://doi.org/10.1007/978-1-59745-177-2_4.
[337] Z. Li, H. Yu, W. Zhuang, S. Mukamel, Geometry and excitation energy fluctuations of NMA in aqueous solution with CHARMM, AMBER, OPLS, and GROMOS force fields: Implications for protein ultraviolet spectra simulation, Chemical Physics Letters. 452 (2008) 78–83. https://doi.org/10.1016/j.cplett.2007.12.022.
[338] H. Liu, M. Elstner, E. Kaxiras, T. Frauenheim, J. Hermans, W. Yang, Quantum mechanics simulation of protein dynamics on long timescale, Proteins: Structure, Function and Genetics. 44 (2001) 484–489. https://doi.org/10.1002/prot.1114.
[339] H. Lu, J. Skolnick, Application of Statistical Potentials to Protein Structure Refinement from Low Resolution Ab Initio Models, Biopolymers. 70 (2003) 575–584. https://doi.org/10.1002/bip.10537.
[340] R. Han, A. Leo-Macias, D. Zerbino, U. Bastolla, B. Contreras-Moreira, A.R. Ortiz, An efficient conformational sampling method for homology modeling, Proteins: Structure, Function and Genetics. 71 (2008) 175–188. https://doi.org/10.1002/prot.21672.
[341] R. Ishitani, T. Terada, K. Shimizu, Refinement of comparative models of protein structure by using multicanonical molecular dynamics simulations, Molecular Simulation. 34 (2008) 327–336. https://doi.org/10.1080/08927020801930539.
[342] H. Fan, Refinement of homology-based protein structures by molecular dynamics simulation techniques, Protein Science. 13 (2004) 211–220. https://doi.org/10.1110/ps.03381404.
[343] S. Kannan, M. Zacharias, Application of biasing-potential replicaexchange simulations for loop modeling and refinement of proteins in explicit solvent, Proteins: Structure, Function and Bioinformatics. 78 (2010) 2809–2819. https://doi.org/10.1002/prot.22796.
[344] J.L. Maccallum, A. Pérez, M.J. Schnieders, L. Hua, M.P. Jacobson, K.A. Dill, Assessment of protein structure refinement in CASP9, Proteins: Structure, Function and Bioinformatics. 79 (2011) 74–90. https://doi.org/10.1002/prot.23131.
[345] D. Baker, A. Sali, Protein structure prediction and structural genomics, Science. 294 (2001) 93–96. https://doi.org/10.1126/science.1065659.
[346] T. Schwede, A. Sali, B. Honig, M. Levitt, H.M. Berman, D. Jones, S.E. Brenner, S.K. Burley, R. Das, N. v. Dokholyan, R.L. Dunbrack, K. Fidelis, A. Fiser, A. Godzik, Y.J. Huang, C. Humblet, M.P. Jacobson, A. Joachimiak, S.R. Krystek, T. Kortemme, A. Kryshtafovych, G.T. Montelione, J. Moult, D. Murray, R. Sanchez, T.R. Sosnick, D.M. Standley, T. Stouch, S. Vajda, M. Vasquez, J.D. Westbrook, I.A. Wilson, Outcome of a Workshop on Applications of Protein Models in Biomedical Research, in: Structure, Cell Press, 2009: pp. 151–159. https://doi.org/10.1016/j.str.2008.12.014.
[347] A. Kryshtafovych, A. Barbato, K. Fidelis, B. Monastyrskyy, T. Schwede, A. Tramontano, Assessment of the assessment: Evaluation of the model quality estimates in CASP10, Proteins: Structure, Function and Bioinformatics. 82 (2014) 112–126. https://doi.org/10.1002/prot.24347.

[348] P. Benkert, M. Biasini, T. Schwede, Toward the estimation of the absolute quality of individual protein structure models, Bioinformatics. 27 (2011) 343–350. https://doi.org/10.1093/bioinformatics/btq662.
[349] P. Larsson, M.J. Skwark, B. Wallner, A. Elofsson, Assessment of global and local model quality in CASP8 using Pcons and ProQ, Proteins: Structure, Function and Bioinformatics. 77 (2009) 167–172. https://doi.org/10.1002/prot.22476.
[350] L.J. McGuffin, M.T. Buenavista, D.B. Roche, The ModFOLD4 server for the quality assessment of 3D protein models., Nucleic Acids Research. 41 (2013) W368–W372. https://doi.org/10.1093/nar/gkt294.
[351] R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Errors in protein structures, Nature. 381 (1996) 272. https://doi.org/10.1038/381272a0.
[352] R.A. Laskowski, M.W. MacArthur, D.S. Moss, J.M. Thornton, PROCHECK: a program to check the stereochemical quality of protein structures, Journal of Applied Crystallography. 26 (1993) 283–291. https://doi.org/10.1107/s0021889892009944.
[353] V.B. Chen, W.B. Arendall, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, D.C. Richardson, MolProbity: All-atom structure validation for macromolecular crystallography, Acta Crystallographica Section D: Biological Crystallography. 66 (2010) 12–21. https://doi.org/10.1107/S0907444909042073.
[354] O. Carugo, K. Djinovic Carugo, Half a century of Ramachandran plots, Acta Crystallographica Section D: Biological Crystallography. 69 (2013) 1333–1341. https://doi.org/10.1107/S090744491301158X.
[355] D. Eisenberg, R. Lüthy, J.U. Bowie, VERIFY3D: Assessment of protein models with three-dimensional profiles, Methods in Enzymology. (1997). https://doi.org/10.1016/S0076-6879(97)77022-8.
[356] A. Kryshtafovych, B. Monastyrskyy, K. Fidelis, J. Moult, T. Schwede, A. Tramontano, Evaluation of the template-based modeling in CASP12, Proteins: Structure, Function and Bioinformatics. 86 (2018) 321–334. https://doi.org/10.1002/prot.25425.
[357] B. Wallner, A. Elofsson, All are not equal: A benchmark of different homology modeling programs, Protein Science. 14 (2005) 1315–1327. https://doi.org/10.1110/ps.041253405.
[358] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, Journal of Physical Chemistry B. 102 (1998) 3586–3616. https://doi.org/10.1021/jp973084f.
[359] A. ŠAli, J.P. Overington, Derivation of rules for comparative protein modeling from a database of protein structure alignments, Protein Science. 3 (1994) 1582–1596. https://doi.org/10.1002/pro.5560030923.
[360] S.F. Sousa, A.J.M. Ribeiro, J.T.S. Coimbra, R.P.P. Neves, S.A. Martins, N.S.H.N. Moorthy, P.A. Fernandes, M.J. Ramos, Protein-Ligand Docking in the New

Millennium – A Retrospective of 10 Years in the Field, Current Medicinal Chemistry. 20 (2013) 2296–2314. https://doi.org/10.2174/0929867311320180002.
[361] G.M. Morris, H. Ruth, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, A.J. Olson, Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, Journal of Computational Chemistry. 30 (2009) 2785–2791. https://doi.org/10.1002/jcc.21256.
[362] G. Jones, P. Willett, R.C. Glen, A.R. Leach, R. Taylor, Development and validation of a genetic algorithm for flexible docking, Journal of Molecular Biology. 267 (1997) 727–748. https://doi.org/10.1006/jmbi.1996.0897.
[363] S. Mukherjee, T.E. Balius, R.C. Rizzo, Docking validation resources: Protein family and ligand flexibility experiments, Journal of Chemical Information and Modeling. 50 (2010) 1986–2000. https://doi.org/10.1021/ci1001982.
[364] I. Schellhammer, M. Rarey, FlexX-Scan: fast, structure-based virtual screening, Proteins. 57 (2004) 504–517. https://doi.org/10.1002/PROT.20217.
[365] M. Repasky, M. Shelley, R. Friesner, Flexible ligand docking with Glide, Current Protocols in Bioinformatics. Chapter 8 (2007). https://doi.org/10.1002/0471250953.BI0812S18.
[366] O. Trott, A.J. Olson, AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, Journal of Computational Chemistry. 31 (2009) NA-NA. https://doi.org/10.1002/jcc.21334.
[367] N.S. Pagadala, K. Syed, J. Tuszynski, Software for molecular docking: a review, Biophysical Reviews. 9 (2017) 91–102. https://doi.org/10.1007/s12551-016-0247-1.
[368] Y.C. Chen, Beware of docking!, Trends in Pharmacological Sciences. 36 (2015) 78–95. https://doi.org/10.1016/J.TIPS.2014.12.001.
[369] C. Pozzi, F. di Pisa, M. Benvenuti, S. Mangani, The structure of the human glutaminyl cyclase-SEN177 complex indicates routes for developing new potent inhibitors as possible agents for the treatment of neurological disorders, Journal of Biological Inorganic Chemistry : JBIC : A Publication of the Society of Biological Inorganic Chemistry. 23 (2018) 1219–1226. https://doi.org/10.1007/S00775-018-1605-1.
[370] N.T. Nguyen, T.H. Nguyen, T.N.H. Pham, N.T. Huy, M. van Bay, M.Q. Pham, P.C. Nam, V. v. Vu, S.T. Ngo, Autodock Vina Adopts More Accurate Binding Poses but Autodock4 Forms Better Binding Affinity, Journal of Chemical Information and Modeling. 60 (2019) 204–211. https://doi.org/10.1021/ACS.JCIM.9B00778.
[371] T. Gaillard, Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark, Journal of Chemical Information and Modeling. 58 (2018) 1697–1706. https://doi.org/10.1021/ACS.JCIM.8B00312.
[372] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, T. Hou, Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power, Physical Chemistry Chemical Physics. 18 (2016) 12964–12975. https://doi.org/10.1039/C6CP01555G.
[373] J. Baxter, Local Optima Avoidance in Depot Location, The Journal of the Operational Research Society. 32 (1981) 815. https://doi.org/10.2307/2581397.

[374] C. Blum, M.J. Belsa Aguilera, A. Roli, M. Sampels, Hybrid Metaheuristics: An Emerging Approach to Optimization. Studies in Computational Intelligence, Computational Intelligence. 114 (2008) 290.
[375] Y. Chen, B. Roux, Generalized Metropolis acceptance criterion for hybrid non-equilibrium molecular dynamics-Monte Carlo simulations, The Journal of Chemical Physics. 142 (2015). https://doi.org/10.1063/1.4904889.
[376] Jorge. Nocedal, S.J. Wright, Numerical optimization, (2006) 664.
[377] A.K. Nivedha, D.F. Thieker, S. Makeneni, H. Hu, R.J. Woods, Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking, Journal of Chemical Theory and Computation. 12 (2016) 892–901. https://doi.org/10.1021/acs.jctc.5b00834.
[378] M.L. DeMarco, R.J. Woods, Structural glycobiology: A game of snakes and ladders, Glycobiology. 18 (2008) 426–440. https://doi.org/10.1093/glycob/cwn026.
[379] A. Imberty, Oligosaccharide structures: Theory versus experiment, Current Opinion in Structural Biology. 7 (1997) 617–623. https://doi.org/10.1016/S0959-440X(97)80069-3.
[380] R.U. Lemieux, S. Koto, D. Voisin, The Exo-Anomeric Effect, ACS Symposium Series. 87 (1979) 17–29. https://doi.org/10.1021/BK-1979-0087.CH002.
[381] S. Wolfe, Gauche effect. Stereochemical consequences of adjacent electron pairs and polar bonds, Accounts of Chemical Research. 5 (2002) 102–111. https://doi.org/10.1021/AR50051A003.
[382] K. Kirschner, R. Woods, Solvent interactions determine carbohydrate conformation, Proceedings of the National Academy of Sciences of the United States of America. 98 (2001) 10541–10545. https://doi.org/10.1073/PNAS.191362798.
[383] A. Nivedha, S. Makeneni, B. Foley, M. Tessier, R. Woods, Importance of ligand conformational energies in carbohydrate docking: Sorting the wheat from the chaff, Journal of Computational Chemistry. 35 (2014) 526–539. https://doi.org/10.1002/JCC.23517.
[384] R. Petrenko, J. Meller, Molecular Dynamics, Encyclopedia of Life Sciences. (2010). https://doi.org/10.1002/9780470015902.A0003048.PUB2.
[385] M. Karplus, J. McCammon, Molecular dynamics simulations of biomolecules, Nature Structural Biology. 9 (2002) 646–652. https://doi.org/10.1038/NSB0902-646.
[386] A. Perez, J.A. Morrone, C. Simmerling, K.A. Dill, Advances in free-energy-based simulations of protein folding and ligand binding, Current Opinion in Structural Biology. 36 (2016) 25–31. https://doi.org/10.1016/J.SBI.2015.12.002.
[387] S.A. Hollingsworth, R.O. Dror, Molecular Dynamics Simulation for All, Neuron. 99 (2018) 1129–1143. https://doi.org/10.1016/J.NEURON.2018.08.011.
[388] M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess, E. Lindah, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, SoftwareX. 1–2 (2015) 19–25.
[389] K. Lindorff-Larsen, P. Maragakis, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, Systematic Validation of Protein Force Fields against Experimental Data, PLOS ONE. 7 (2012) e32131. https://doi.org/10.1371/JOURNAL.PONE.0032131.

[390] G.B. Goh, B.S. Hulbert, H. Zhou, C.L. Brooks, Constant pH molecular dynamics of proteins in explicit solvent with proton tautomerism, Proteins: Structure, Function, and Bioinformatics. 82 (2014) 1319–1331. https://doi.org/10.1002/PROT.24499.
[391] J. Ponder, D. Case, Force fields for protein simulations, Advances in Protein Chemistry. 66 (2003) 27–85. https://doi.org/10.1016/S0065-3233(03)66002-X.
[392] H. Geng, F. Chen, J. Ye, F. Jiang, Applications of Molecular Dynamics Simulation in Structure Prediction of Peptides and Proteins, Computational and Structural Biotechnology Journal. 17 (2019) 1162–1170. https://doi.org/10.1016/J.CSBJ.2019.07.010.
[393] M.-C. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, A.E. Garcia, Water Determines the Structure and Dynamics of Proteins, Chemical Reviews. 116 (2016) 7673–7697. https://doi.org/10.1021/ACS.CHEMREV.5B00664.
[394] W.L. Jorgensen, J. Chandrasekhar, J.D. Madura, R.W. Impey, M.L. Klein, Comparison of simple potential functions for simulating liquid water, The Journal of Chemical Physics. 79 (1998) 926. https://doi.org/10.1063/1.445869.
[395] W.C. Still, A. Tempczyk, R.C. Hawley, T. Hendrickson, Semianalytical treatment of solvation for molecular mechanics and dynamics, Journal of the American Chemical Society. 112 (2002) 6127–6129. https://doi.org/10.1021/JA00172A038.
[396] H. Nguyen, D.R. Roe, C. Simmerling, Improved Generalized Born Solvent Model Parameters for Protein Simulations, Journal of Chemical Theory and Computation. 9 (2013) 2020–2034. https://doi.org/10.1021/CT3010485.
[397] A. Onufriev, D. Bashford, D. Case, Exploring protein native states and large-scale conformational changes with a modified generalized born model, Proteins. 55 (2004) 383–394. https://doi.org/10.1002/PROT.20033.
[398] D. Case, T. Cheatham, T. Darden, H. Gohlke, R. Luo, K. Merz, A. Onufriev, C. Simmerling, B. Wang, R. Woods, The Amber biomolecular simulation programs, Journal of Computational Chemistry. 26 (2005) 1668–1688. https://doi.org/10.1002/JCC.20290.
[399] W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics, Journal of Molecular Graphics. 14 (1996) 33–38. https://doi.org/10.1016/0263-7855(96)00018-5.
[400] 7.1.1. RMSD — AdKGromacsTutorial 2.0.2 documentation, (n.d.). https://adkgromacstutorial.readthedocs.io/en/latest/analysis/rmsd.html (accessed July 26, 2021).
[401] 7.1.4. Radius of gyration — AdKGromacsTutorial 2.0.2 documentation, (n.d.). https://adkgromacstutorial.readthedocs.io/en/latest/analysis/rgyr.html (accessed July 26, 2021).
[402] 7.1.2. RMSF — AdKGromacsTutorial 2.0.2 documentation, (n.d.). https://adkgromacstutorial.readthedocs.io/en/latest/analysis/rmsf.html (accessed July 26, 2021).
[403] D.R. Roe, I. Thomas E. Cheatham, T.E. Cheatham, PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data, Journal

of Chemical Theory and Computation. 9 (2013) 3084–3095. https://doi.org/10.1021/ct400341p.
[404] M.A. Williams, J.E. Ladbury, Hydrogen Bonds in Protein-Ligand Complexes, Protein-Ligand Interactions: From Molecular Recognition to Drug Design. (2005) 137–161. https://doi.org/10.1002/3527601813.CH6.
[405] H. Fu, H. Chen, M. Blazhynska, E. Goulard Coderc de Lacam, F. Szczepaniak, A. Pavlova, X. Shao, J.C. Gumbart, F. Dehez, B. Roux, W. Cai, C. Chipot, Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations, Nature Protocols 2022. (2022) 1–28. https://doi.org/10.1038/s41596-021-00676-1.
[406] P. Mikulskis, S. Genheden, U. Ryde, A large-scale test of free-energy simulation estimates of protein-Ligand binding affinities, Journal of Chemical Information and Modeling. 54 (2014) 2794–2806. https://doi.org/10.1021/CI5004027/SUPPL_FILE/CI5004027_SI_001.PDF.
[407] T. Hou, J. Wang, Y. Li, W. Wang, Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations, Journal of Chemical Information and Modeling. 51 (2011) 69–82. https://doi.org/10.1021/CI100275A/SUPPL_FILE/CI100275A_SI_001.PDF.
[408] C. Wang, D. Greene, L. Xiao, R. Qi, R. Luo, Recent Developments and Applications of the MMPBSA Method, Frontiers in Molecular Biosciences. (2018). https://doi.org/10.3389/fmolb.2017.00087.
[409] T. Hou, N. Li, Y. Li, W. Wang, Characterization of domain-peptide interaction interface: Prediction of SH3 domain-mediated protein-protein interaction network in yeast by generic structure-based models, Journal of Proteome Research. 11 (2012) 2982–2995. https://doi.org/10.1021/pr3000688.
[410] H. Gohlke, C. Kiel, D.A. Case, Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes., Journal of Molecular Biology. (2003). https://doi.org/10.1016/S0022-2836(03)00610-7.
[411] R. Kumari, R. Kumar, A. Lynn, G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations, Journal of Chemical Information and Modeling. (2014). https://doi.org/10.1021/ci500020m.
[412] H. Abdi, L.J. Williams, Principal component analysis, Wiley Interdisciplinary Reviews: Computational Statistics. 2 (2010) 433–459. https://doi.org/10.1002/WICS.101.
[413] I.T. Jolliffe, Principal component analysis, in: Springer Series in Statistics, Vol XXIX, 2nd ed., Springer, New York, 2002: p. 487.
[414] C.C. David, D.J. Jacobs, Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins, Methods in Molecular Biology (Clifton, N.J.). 1084 (2014) 193. https://doi.org/10.1007/978-1-62703-658-0_11.
[415] M.A. Balsera, M.A. Balsera, W. Wriggers, Y. Oono, K. Schulten, Principal Component Analysis and long time protein dynamics, J. PHYS. CHEM. (1996).

http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.382.8268 (accessed July 26, 2021).
[416] A. Amadei, A.B.M. Linssen, H.J.C. Berendsen, Essential dynamics of proteins, Proteins: Structure, Function, and Bioinformatics. 17 (1993) 412–425. https://doi.org/10.1002/prot.340170408.
[417] H. Berendsen, S. Hayward, Collective protein dynamics in relation to function, Current Opinion in Structural Biology. 10 (2000) 165–169. https://doi.org/10.1016/S0959-440X(00)00061-0.
[418] A. Amadei, A. Linssen, B. de Groot, D. van Aalten, H. Berendsen, An efficient method for sampling the essential subspace of proteins, Journal of Biomolecular Structure & Dynamics. 13 (1996) 615–625. https://doi.org/10.1080/07391102.1996.10508874.
[419] A. Amusengeri, R.B. Tata, Ö.T. Bishop, Understanding the Pyrimethamine Drug Resistance Mechanism via Combined Molecular Dynamics and Dynamic Residue Network Analysis, Molecules 2020, Vol. 25, Page 904. 25 (2020) 904. https://doi.org/10.3390/MOLECULES25040904.
[420] I. Daidone, A. Amadei, Essential dynamics: foundation and applications, Wiley Interdisciplinary Reviews: Computational Molecular Science. 2 (2012) 762–770. https://doi.org/10.1002/WCMS.1099.
[421] gmx sham — GROMACS 2021.2 documentation, (n.d.). https://manual.gromacs.org/current/onlinehelp/gmx-sham.html (accessed July 26, 2021).
[422] E.W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K.D. Pruitt, I. Karsch-Mizrachi, GenBank, Nucleic Acids Research. (2019). https://doi.org/10.1093/nar/gky989.
[423] A.M. Waterhouse, J.B. Procter, D.M.A. Martin, M. Clamp, G.J. Barton, Jalview Version 2-A multiple sequence alignment editor and analysis workbench, Bioinformatics. (2009). https://doi.org/10.1093/bioinformatics/btp033.
[424] S. Kumar, G. Stecher, K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, Molecular Biology and Evolution. 33 (2016) 1870–1874. https://doi.org/10.1093/molbev/msw054.
[425] S. Whelan, N. Goldman, A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach, Molecular Biology and Evolution. (2001). https://doi.org/10.1093/oxfordjournals.molbev.a003851.
[426] S.Q. Le, O. Gascuel, An improved general amino acid replacement matrix, Molecular Biology and Evolution. (2008). https://doi.org/10.1093/molbev/msn067.
[427] T.L. Bailey, J. Johnson, C.E. Grant, W.S. Noble, The MEME Suite, Nucleic Acids Research. (2015). https://doi.org/10.1093/nar/gkv416.
[428] N. Faya, D.L. Penkler, Ö. Tastan Bishop, Human, vector and parasite Hsp90 proteins: A comparative bioinformatics analysis, FEBS Open Bio. (2015). https://doi.org/10.1016/j.fob.2015.11.003.
[429] L. Zimmermann, A. Stephens, S.Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A.N. Lupas, V. Alva, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core, Journal of Molecular Biology. (2018). https://doi.org/10.1016/j.jmb.2017.12.007.

[430] R. Hatherley, D.K. Brown, M. Glenister, Ö.T. Bishop, PRIMO: An interactive homology modeling pipeline, PLoS ONE. (2016). https://doi.org/10.1371/journal.pone.0166698.
[431] P. Benkert, M. Künzli, T. Schwede, QMEAN server for protein model quality estimation, Nucleic Acids Research. (2009). https://doi.org/10.1093/nar/gkp322.
[432] Q. Wan, J.M. Parks, B.L. Hanson, S.Z. Fisher, A. Ostermann, T.E. Schrader, D.E. Graham, L. Coates, P. Langan, A. Kovalevsky, Direct determination of protonation states and visualization of hydrogen bonding in a glycoside hydrolase with neutron crystallography, Proceedings of the National Academy of Sciences. (2015). https://doi.org/10.1073/pnas.1504986112.
[433] D.A. Case, D.S. Cerutti, T.E.I. Cheatham, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D. Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao, D.M. York, P.A. Kollman, AmberTools2017, University of California, San Francisco. (2017). https://doi.org/citeulike-article-id:2734527.
[434] T. Miyata, Discovery studio modeling environment, Ensemble. (2015). https://doi.org/https://doi.org/10.11436/mssj.17.98.
[435] R. Anandakrishnan, B. Aguilar, A. v. Onufriev, H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, Nucleic Acids Research. (2012). https://doi.org/10.1093/nar/gks375.
[436] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. Dinola, J.R. Haak, Molecular dynamics with coupling to an external bath, The Journal of Chemical Physics. (1984). https://doi.org/10.1063/1.448118.
[437] G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling, Journal of Chemical Physics. 126 (2007). https://doi.org/10.1063/1.2408420.
[438] J.A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K.E. Hauser, C. Simmerling, ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB, Journal of Chemical Theory and Computation. (2015). https://doi.org/10.1021/acs.jctc.5b00255.
[439] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, LINCS: A Linear Constraint Solver for molecular simulations, Journal of Computational Chemistry. (1997). https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.
[440] P.A. Kollman, I. Massova, C. Reyes, B. Kuhn, S. Huo, L. Chong, M. Lee, T. Lee, Y. Duan, W. Wang, O. Donini, P. Cieplak, J. Srinivasan, D.A. Case, T.E. Cheatham, Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models, Accounts of Chemical Research. (2000). https://doi.org/10.1021/ar000033j.
[441] L. Holm, Benchmarking Fold Detection by DaliLite v.5., Bioinformatics (Oxford, England). (2019). https://doi.org/10.1093/bioinformatics/btz536.

[442] H.X. Zhou, X. Pang, Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation, Chemical Reviews. 118 (2018) 1691–1741. https://doi.org/10.1021/acs.chemrev.7b00305.
[443] Z. Zhang, S. Witham, E. Alexov, On the role of electrostatics in protein-protein interactions, Physical Biology. 8 (2011). https://doi.org/10.1088/1478-3975/8/3/035001.
[444] E. Gabor, A.K. Göhler, A. Kosfeld, A. Staab, A. Kremling, K. Jahreis, The phosphoenolpyruvate-dependent glucose-phosphotransferase system from Escherichia coli K-12 as the center of a network regulating carbohydrate flux in the cell, European Journal of Cell Biology. 90 (2011) 711–720. https://doi.org/10.1016/j.ejcb.2011.04.002.
[445] B. Kramer, M. Rarey, T. Lengauer, Evaluation of the FLEXX Incremental Construction Algorithm for Protein-Ligand Docking, n.d. https://doi.org/10.1002/(SICI)1097-0134(19991101)37:2.
[446] M. Kontoyianni, L.M. McClellan, G.S. Sokol, Evaluation of Docking Performance: Comparative Data on Docking Algorithms, Journal of Medicinal Chemistry. 47 (2004) 558–565. https://doi.org/10.1021/jm0302997.
[447] H. Gohlke, M. Hendlich, G. Klebe, Knowledge-based scoring function to predict protein-ligand interactions, Journal of Molecular Biology. 295 (2000) 337–356. https://doi.org/10.1006/jmbi.1999.3371.
[448] D. Gioia, M. Bertazzo, M. Recanatini, M. Masetti, A. Cavalli, Dynamic docking: A paradigm shift in computational drug discovery, Molecules. 22 (2017). https://doi.org/10.3390/molecules22112029.
[449] E. Nittinger, T. Inhester, S. Bietz, A. Meyder, K.T. Schomburg, G. Lange, R. Klein, M. Rarey, Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein-Ligand Interfaces, Journal of Medicinal Chemistry. (2017). https://doi.org/10.1021/acs.jmedchem.7b00101.
[450] T. Sawada, D.G. Fedorov, K. Kitaura, Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations, Journal of the American Chemical Society. (2010). https://doi.org/10.1021/ja105051e.
[451] S. Salentin, V.J. Haupt, S. Daminelli, M. Schroeder, Polypharmacology rescored: Protein-ligand interaction profiles for remote binding site similarity assessment, Progress in Biophysics and Molecular Biology. (2014). https://doi.org/10.1016/j.pbiomolbio.2014.05.006.
[452] M.S. Taylor, E.N. Jacobsen, Asymmetric catalysis by chiral hydrogen-bond donors, Angewandte Chemie - International Edition. (2006). https://doi.org/10.1002/anie.200503132.
[453] A. Natarajan, J.P. Schwans, D. Herschlag, Using unnatural amino acids to probe the energetics of oxyanion hole hydrogen bonds in the ketosteroid isomerase active site, Journal of the American Chemical Society. (2014). https://doi.org/10.1021/ja413174b.
[454] B. Ma, S. Kumar, C.J. Tsai, R. Nussinov, Folding funnels and binding mechanisms, Protein Engineering. (1999). https://doi.org/10.1093/protein/12.9.713.

[455] C.J. Tsai, R. Nussinov, The free energy landscape in translational science: How can somatic mutations result in constitutive oncogenic activation?, Physical Chemistry Chemical Physics. (2014). https://doi.org/10.1039/c3cp54253j.