



DOCTORAL THESIS No. 2023:9
FACULTY OF NATURAL RESOURCES AND AGRICULTURAL SCIENCES

**Structure function studies of GH7
cellulases, key enzymes in the global
carbon cycle**

TOPI HAATAJA

Structure function studies of GH7 cellulases, key enzymes in the global carbon cycle

Topi Haataja

Faculty of Natural Resources and Agricultural Sciences

Department of Molecular Sciences

Uppsala



SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS

Uppsala 2023

Acta Universitatis Agriculturae Sueciae
2023:9

ISSN 1652-6880

ISBN (print version) 978-91-8046-070-5

ISBN (electronic version) 978-91-8046-071-2

<https://doi.org/10.54612/a.672jon1kdu>

© 2023 Topi Haataja, <https://orcid.org/0000-0002-6436-9953>

Swedish University of Agricultural Sciences, Department of Molecular Sciences, Uppsala, Sweden

The summary chapter of this thesis is licensed under CC BY NC 4.0, other licenses or copyright may apply to illustrations and attached articles.

Print: SLU Grafisk service, Uppsala 2023

Structure function studies of GH7 cellulases, key enzymes in the global carbon cycle

Abstract

Enzyme mixtures used for lignocellulosic ethanol production are most commonly derived from filamentous fungi, and enzymes from the glycoside hydrolase family 7 (GH7) constitute the most abundant components in these cocktails. In this thesis I have aimed to increase our understanding of this enzyme family, with focus on the interrelation between their structure and function.

In a study of the two model enzymes *Trichoderma reesei* Cel7A (TreCel7A) and *Phanerochaete chrysosporium* Cel7D (Pch7D), we determined factors governing the idiosyncratic behavior of these enzymes on commonly used model compounds, and by using fluorescence titration, enzyme kinetics, structure determination and molecular dynamics simulations found specific structural features connected to non-productive binding, playing a major role in enzyme activity on these compounds.

We also determined the molecular structure of a GH7 enzyme RsSymEG1, belonging to a group of smaller GH7 endoglucanases with previously unknown structure architecture, and originating from symbiotic protozoa of wood eating lower termites. The X-ray crystal structure revealed a configuration with several key differences to previously known GH7 structures, and will aid in modelling and engineering of enzymes in this so far little-known group of enzymes. A further look into this group of sequences, as well as other GH7 enzymes found in the termite symbiont protists, also revealed previously unknown details about the evolution of this ancient enzyme family.

Furthermore, we explored single molecule imaging of the model enzyme TreCel7A with novel imaging methods, providing a first proof-of-concept of using fluorescence resonance energy transfer (FRET) for the study of inter-domain dynamics of this enzyme, as well as total internal reflection dark-field microscopy (TIRDFM) for imaging enzyme movement on cellulose surface at ultra-high temporal resolutions.

Keywords: Cellulase, lignocellulose, GH7, glycoside hydrolase, biofuel, fungi

Struktur-funktion studier av GH7 cellulaser, nyckelenzymer i kolets globala kretslopp

Sammanfattning

Enzymblandningar som används för produktion av lignocellulosa etanol är oftast härledda från trådsvampar, och enzymer från glykosidhydrolasfamiljen 7 (GH7) utgör de vanligaste komponenterna i dessa cocktails. I detta examensarbete har jag syftat till att öka vår förståelse för denna enzymfamilj, med fokus på sambandet mellan deras struktur och funktion.

I en studie av de två modellenzymerna *Trichoderma reesei* Cel7A (TreCel7A) och *Phanerochaete chrysosporium* Cel7D (Pch7D), bestämde vi faktorer som styr det idiosynkratiska beteendet hos dessa enzymer på vanliga modellföreningar, och genom att använda fluorescenstitrering, enzymkinetik, strukturbestämning och molekylära dynamiksimuleringar fann specifika strukturella egenskaper kopplade till icke-produktiv bindning, som spelar en viktig roll i enzymaktiviteten på dessa föreningar.

Vi bestämde också den molekylära strukturen för ett GH7-enzym RsSymEG1, som tillhör en grupp av mindre GH7-endoglukanaser med tidigare okänd strukturell arkitektur, och som kommer från symbiotiska protozoer av träätande lägre termiter. Röntgenkristallstrukturen avslöjade en konfiguration med flera viktiga skillnader mot tidigare kända GH7-strukturer, och kommer att hjälpa till med modellering och konstruktion av enzymer i denna hittills föga kända grupp av enzymer. En ytterligare titt på denna grupp av sekvenser, såväl som andra GH7-enzymen som finns i termit-symbiontprotister, avslöjade också tidigare okända detaljer om utvecklingen av denna uråldriga enzymfamilj.

Dessutom undersökte vi en molekylärbildning av modellenzymet TreCel7A med nya avbildningsmetoder vilket gav ett första bevis för att använda fluorescensresonansenergiöverföring (FRET) för studien av interdomändynamik för detta enzym, såväl som total intern reflektion mörkfältsmikroskopi (TIRDFM) för avbildning av enzymrörelser på cellulosaytan vid ultrahöga tidsupplösningar.

Nyckelord: Cellulas, cellulosa, GH7, biobränsle

If it weren't for constraints, everything would be easy, and life would be boring.

Contents

List of publications.....	9
Abbreviations	11
1. Introduction.....	13
2. Background	15
2.1 Lignocellulosic biomass.....	15
2.2 Lignocellulose degradation.....	17
2.3 Fungal enzymes for lignocellulose degradation.....	19
2.4 GH7 cellulases.....	28
2.5 Recombinant expression in <i>Trichoderma reesei</i>	37
3. Current investigation.....	41
3.1 Enzyme kinetics by GH7 cellobiohydrolases on chromogenic substrates is dictated by non-productive binding: insights from crystal structures and MD simulation (Paper I).....	41
3.1.1 The use of chromogenic and fluorescent model compounds in GH7 research.....	41
3.1.2 Enzyme kinetics, crystal structures and molecular dynamics simulations.....	43
3.1.3 Conclusions.....	45
3.2 The crystal structure of RsSymEG1 reveals unique form of smaller GH7 endoglucanases alongside GH7 cellobiohydrolases in protist symbionts of termites (Paper II).....	47
3.2.1 GH7 enzymes in protist symbionts of lower termites.....	47
3.2.2 Structure characteristics of RsSymEG1, a short GH7 endoglucanase	47
3.2.3 Repertoire of GH7 enzymes in lower termite symbiont protozoa.....	49
3.2.4 sEGs in the context of GH7 evolution.....	50
3.2.5 Activity measurements	52
3.2.6 Conclusions and future perspectives.....	53

3.3	Production of TreCel7A mutants in <i>Trichoderma reesei</i> for single-molecule imaging of processive enzyme action and FRET analysis of interdomain protein dynamics.....	54
3.3.1	Single molecule imaging of processive GH7 CBHs.....	54
3.3.2	Design and production of TreCel7A mutants for imaging studies	55
3.3.3	Insights from ultra-high speed single molecule imaging and FRET microscopy	56
3.3.4	Conclusions and future perspectives.....	58
4.	Future perspectives & conclusions.....	61
	References.....	65
	Popular science summary	77
	Populärvetenskaplig sammanfattning	81
	Acknowledgements	85

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

- I. Haataja T, Gado JE, Nutt A, Anderson NT, Nilsson M, Momeni MH, Isaksson R, Väljamäe P, Johansson G, Payne CM & Ståhlberg J (2022). Enzyme kinetics by GH7 cellobiohydrolases on chromogenic substrates is dictated by non-productive binding: insights from crystal structures and MD simulation. *FEBS Journal*, 2022 Aug 23. doi: 10.1111/febs.16602. Online ahead of print.
- II. Haataja T, Hansson H, Moriya S, Sandgren M & Ståhlberg J (2023). The crystal structure of RsSymEG1 reveals unique form of smaller GH7 endoglucanases alongside GH7 cellobiohydrolases in protist symbionts of termites. (manuscript)
- III. Haataja T*, Nakamura A*, Subramanian V, Farmer S, Decker S & Ståhlberg J (2023). Production of TreCel7A mutants in *Trichoderma reesei* for single-molecule imaging of processive enzyme action and FRET analysis of interdomain protein dynamics. (manuscript)

*These authors contributed equally.

Paper I is reproduced with the permission of the publishers.

The contribution of Topi Haataja to the papers included in this thesis was as follows:

- I. Took part in the laboratory work: Determined competitive inhibition constants for lactose. Solved, refined and deposited crystal structures of TreCel7A E212Q mutant with pNPL and lactose bound (7OC8, 7NYT, respectively). Took part in data analysis and interpretation. Took part in writing the manuscript with the co-authors. Made protein crystal structure figures.
- II. Planned the work with the co-authors. Performed all the laboratory work as well as data analysis, and interpretation with help from the co-authors. Made figures and tables. Wrote the manuscript with help from the co-authors.
- III. Took part in planning. Performed laboratory work: Made constructs, expressed in *T. reesei* and purified TreCel7A variants with free cysteines for fluorescence and gold nanoparticle labelling and shipped to Japan for single molecule imaging experiments. Took part in data analysis and interpretation. Wrote the manuscript with help from the co-authors.

Abbreviations

AA	auxiliary activity
AuNP	gold nanoparticle
BGL	beta-glucosidase
CAZy	carbohydrate active enzyme database
CBH	cellobiohydrolase
CBM	carbohydrate binding module
CD	catalytic domain
CDH	cellobiose dehydrogenase
CE	carbohydrate esterase
CMC	carboxymethylcellulose
EG	endoglucanase
eGFP	enhanced green fluorescent protein
GH	glycoside hydrolase
GHG	greenhouse gas
GMC	glucose-methanol-choline oxidoreductase
GT	glycosyltransferase
ICE	internal combustion engine
LFER	linear free energy relationship
LnP	lignin peroxidase

LPMO	lytic polysaccharide monooxygenase
MD	molecular dynamics
MnP	manganese peroxidase
MUC	methylumbelliferyl cellobioside
NREL	National Renewable Energy Laboratory
oNPC	o-nitrophenyl cellobioside
PCA	pyrrolidone carboxylic acid
PchCel7D	<i>Phanerochaete chrysosporium</i> Cel7D
PDB	protein data bank
PHBAH	p-hydroxybenzoic acid hydrazide
PL	polysaccharide lyase
pNP	p-nitrophenyl
pNPC	p-nitrophenyl cellobioside
pNPL	p-nitrophenyl lactoside
PsGH7a	<i>Phytophthora sojae</i> GH7a
RsSymEG1	<i>Reticulitermes speratus</i> symbiont EG1
sEG	short endoglucanase
TIRDFM	total internal reflection dark-field microscopy
TloCel7B	<i>Trichoderma longibrachiatum</i>
TreCel7A	<i>Trichoderma reesei</i> Cel7A
TreCel7B	<i>Trichoderma reesei</i> Cel7B
VP	versatile peroxidase

1. Introduction

One of the greatest challenges of the current and coming decades is reducing the dependence on fossil resources for energy. This relates to both limiting the impact of greenhouse gas (GHG) emissions on climate change, as well as the inherently limited nature of these supplies.

Perhaps the greatest challenge is replacing petrochemicals in transportation, where the current global fleet of vehicles is mostly based on internal combustion engines (ICE) and fueled with fossil fuels (International Energy Agency, 2022a). Many governments have set ambitious goals for phasing out ICE vehicles, yet, there are severe limitations on the velocity of adoption and scalability of most currently available alternative technologies (Michaux, 2021). Additionally, given the need for sustainable and feasible solutions during the transition, continued research into a variety of alternatives for fossil fuel replacement is necessary.

One such alternative is ethanol derived from biomass, i.e., bioethanol. This fuel has been touted as an alternative or complement for fossil fuel-based transportation fuels, and has already been widely adopted, often thanks to mandates and incentives for blending with gasoline (Aui et al., 2021; Liu et al., 2020; Mohanty and Swain, 2019; Rastogi and Shrivastava, 2017; Su et al., 2015). The major advantages of bioethanol lie in its lower GHG emissions compared to gasoline, its high degree of compatibility with existing vehicle fleet and fuel infrastructure, as well as the wide availability of raw materials and ability to produce it using conventional, well established technologies (Borrion et al., 2012; Kim and Dale, 2004; Rosales-Calderon and Arantes, 2019). Currently majority of bioethanol used for blending with gasoline globally is based on starch rich biomass such as corn and wheat, and can thus be seen as competing with food production (Aui et al., 2021). Given the lack of access or affordability of food for many, the use of this type of

bioethanol has been problematic, an aspect which is unlikely to abate with the recent rate of food price increases. Currently, this so-called first-generation bioethanol accounts for roughly 4 % of global transport fuel consumption, making its replacement highly challenging while complying with the need to reduce GHG intensive fossil fuels (International Energy Agency, 2022a). However, so-called second-generation bioethanol processes rely on cellulosic raw materials that are not food grade, and are in fact often waste side-streams of food production or forestry (Zabed et al., 2016). These processes often have highly superior GHG emission impact over the whole life cycle of the fuel production, not only compared to crude oil-based gasoline, but also first-generation bioethanol (Borrion et al., 2012; Karp et al., 2021; Kim et al., 2009; Wang et al., 2018). At best they offer an opportunity to produce a highly value-added product from low value side-streams, without additional land use (Zabed et al., 2016). However, the main obstacle faced by second generation bioethanol is its relatively high cost, making it vulnerable to fluctuations in gasoline prices (Aui et al., 2021; Rosales-Calderon and Arantes, 2019). While the raw material costs for cellulosic ethanol are often low, the cost of production is raised by chemicals and energy required for pre-treatment of the recalcitrant biomass, as well as the cost of enzymes used for hydrolysis, both of which are required to achieve adequate ethanol yields (Aui et al., 2021; Rosales-Calderon and Arantes, 2019; Zabed et al., 2016). While cellulose degrading enzymes have found uses in various industrial applications, the main driver behind cellulase research for industrial uses has been the push to yield more efficient enzymes and enzyme mixtures for the production of cellulosic ethanol (Bhat and Bhat, 1997; Ejaz et al., 2021). This thesis focuses on research conducted on some of the most important cellulose degrading enzymes, those classed as glycoside hydrolase family 7 (GH7) in the carbohydrate active enzyme database classification (CAZy; Drula *et al.*, 2022). Background is provided on cellulosic biomass and the overall process of its enzymatic hydrolysis, with more focus laid on GH7 cellulases, perhaps the most crucial enzymes used for cellulose degradation, both in industry and by organisms in nature. Subsequently, studies included in this thesis are described in detail, discussing some of the methods used, as well as the results and their implications.

2. Background

2.1 Lignocellulosic biomass

Cellulose is often referred to as the most abundant organic polymer on earth, and justifiably so, with some estimates suggesting the annual production on earth reaching up to several billion tons (McNamara et al., 2015). Cellulose is a linear polymer of glucose, where the monomeric glucopyranose units are linked together by Beta-1-4 linkages, forming chains of varying lengths. Cellulose chains have a tendency to assume a flat, ribbon-like structure, where each pyranose unit is rotated 180° in relation to the adjacent ones, consequently making cellobiose the repeating unit within polymeric cellulose (Gardner and Blackwell, 1974a). Within each chain, hydrogen bonds are formed between adjacent glucose units between the hydroxyl group of the C6 carbon and the ring oxygen, as well as between the hydroxyl units of C3 and C2 carbons (Gardner and Blackwell, 1974b). Cellulose is largely insoluble in water in chain lengths exceeding six monomer units, and has both hydrophilic and hydrophobic characteristics due to the axially positioned hydrogen atoms and the equatorially oriented hydroxyl groups (Himmel et al., 2007; McNamara et al., 2015).

A large portion of cellulose in nature is embedded in the cell walls of plant cells, where it is synthesized by cellulose synthase complexes (Carpita and McCann, 2020). These hexagonal complexes simultaneously synthesize several cellulose chains, which assemble into cellulose microfibrils with each containing most likely 18 parallel cellulose chains, although many details of the process are still unknown (Kubicki et al., 2018). Cellulose is synthesized to be a part of both the primary, and the secondary cell walls, the first being more elastic and forming during plant cell growth, while the latter

is constructed after cell growth has ended, and adds rigidity to the cells and tissues (Cosgrove, 2005; Zeng et al., 2017; Zhong and Ye, 2015). While the compositions of carbohydrates within the primary and secondary cell walls usually differ somewhat, both contain a mixture of cellulose, and a diverse set of heteropolysaccharides collectively termed hemicelluloses and pectins. Hemicelluloses are a group of polysaccharides containing mainly glucose, xylose, arabinose, mannose and galactose, while pectins are acidic polysaccharides consisting mainly of galacturonic acid (Gírio et al., 2010; Mohnen, 2008). These carbohydrates form interspersed matrices within the cell walls, with cellulose microfibrils clustering into macrofibrils, which in turn are connected by complex hemicellulose and pectin networks (Figure 1). Within the secondary cell walls, this matrix is further strengthened by lignification, the process of forming a covalently linked network of mainly three aromatic components, p-hydroxyphenyl, guaiacyl, and syringyl -units, with this component collectively referred to as lignin (Vanholme et al., 2010). In terms of dry mass, the major portion of woody biomass consists of secondary cell walls, which is also the main contributor to its recalcitrant nature, and thus it is the main focus of study in the context of conversion of cellulosic biomass through chemical and enzymatic processes (Zeng et al., 2017).

The composition of lignocellulosic plant biomass varies significantly based on not only the type and species of plant, but also depending on the parts of the plants used (Carpita and McCann, 2020). Typically, cellulose constitutes roughly 40-45 %, while hemicellulose and lignin respectively comprise approximately 20-30 % and 20-25 % of typical lignocellulosic biomasses used for biofuels (Zabed et al., 2016).

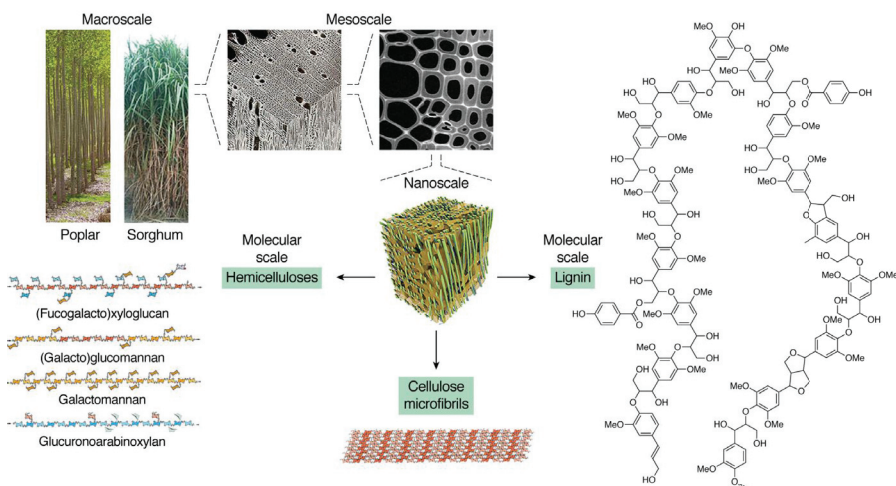


Figure 1. Schematic illustration of the construction of plant cell walls and their main components, cellulose, hemicellulose and lignin. Image from Carpita and McCann, (2020).

2.2 Lignocellulose degradation

Given the vast amounts of plant biomass generated on earth each year, it is natural that many organisms have evolved to directly use these materials as energy sources. The co-evolution of plants and their pathogens has often been described as a continuous battle, where each side iteratively develops means to defend or attack against the other (Anderson et al., 2010; Lee and Hwang, 2020; McCann, 2020). Shifts in this balance can be seen most obviously in outbreaks of plant pathogens, which are often brought about by environmental stress factors such as drought, but sometimes also through introduction of pathogens into new areas, or acquisition of new qualities through gene transfer or rearrangement (Acuña et al., 2012; Ghelardini et al., 2016). Looking back millions of years via fossil records, more dramatic shifts are seen, as decrease in accumulation of organic carbon in sediments has been suggested to have been caused by emergence of less lignin rich plants, as well as lignin degrading peroxidase enzymes in fungi (Floudas et al., 2012; Robinson, 1990), although others have argued tectonics and climate were the main factors behind this development (Nelsen et al., 2016). While carbon sedimentation has likely been mostly dictated by accumulation of lignin, suberin, cutin, and sporopollenin (Hibbett et al., 2016; Nelsen et

al., 2016), cellulose has also been present, and thus a fundamental part of the carbon cycle on earth. Considering the extreme stability of cellulose, with the half-life of its *O*-glycosidic bonds estimated to be 5 million years in the absence of catalysts, cellulose degrading enzymes have consequently played a pivotal role in recycling organic carbon (Wolfenden and Snider, 2001).

Given that cellulose is found as a structural feature not just in plants, but in a wide variety of organisms, the wide-spread presence of cellulose degrading enzymes is not surprising. After all, organisms are often required to modify their own cell walls through different stages of their lifecycles. However, this kind of utilization of cellulases must presumably be very targeted both spatially and temporally, and most likely conducted through expression of only a few necessary enzymes at specific timepoints. An example of an occurrence of this is expression of cellulases in the slime mold social amoeba, *Dictyostelium discoideum*, which uses cellulose and cellulases during the multicellular stages of its life cycle for the formation of a mobile slug and a fruiting body (Freeze and Loomis, 1978, 1977; Hobdey et al., 2016; Zhang et al., 2001). When it comes to organisms utilizing lignocellulose as a major carbon source however, the myriad of different enzymes expressed simultaneously is often much wider, enabling a higher degree of breakdown and utilization of the material (Champreda *et al.*, 2019; Figure 4).

While there is a wide range of organisms capable of degrading cellulose to some extent, filamentous fungi are generally considered to be the main degraders of lignocellulosic biomass in terrestrial ecosystems on earth (Alfaro et al., 2014; Mäkelä et al., 2014). There are various attributes in the lifestyles of cellulose degrading fungi that can be seen to give them distinct advantages in degradation of plant biomass, which is often dense and rigid, with highly crystalline cellulose embedded in a strong lignin network (Himmel et al., 2007). The ability to progressively penetrate into recalcitrant lignocellulose matrix through sequential enzyme secretion and hyphae growth, coupled with the genomic and physiological complexity to produce a wide array of enzymes in a regulated manner, together with the emergence or adoption of some specific key enzymes discussed below, have likely been key factors in the success of these organisms when it comes to utilization of this abundant resource in nature (Demoor et al., 2019; Mäkelä et al., 2014; Martinez et al., 2004; Mendgen et al., 1996). Consequently, the main focus in this work will be lignocellulose degrading enzymes utilized by fungi, with

an objective to present an overview of the range of enzymes these organisms use for these purposes, and thus provide context to the role GH7 cellulases play both in fungi, and the global carbon cycle.

2.3 Fungal enzymes for lignocellulose degradation

Vast majority of cellulase research has been conducted on fungi belonging to two vast divisions, Ascomycota and Basidiomycota (Mäkelä *et al.*, 2014; Rytioja *et al.*, 2014). Several different fungal lifestyles are found within these two clades. Saprotrophic fungi conform to a so-called free-living lifestyle, where they decompose dead or dying plant matter. Other fungi in turn are dependent on interaction with living plants, either through a symbiotic, or parasitic relationships (Anderson *et al.*, 2010; Bödeker *et al.*, 2016). The separation between these different life-styles is not clear-cut, and fungi with very similar enzyme repertoires in their genomes can display highly different behaviors, depending on the environment, or characteristics of specific strains, even within single species (Olson *et al.*, 2012). Major degraders of woody biomass within basidiomycetes are often referred to as white rot, or brown rot fungi, depending on the pattern of lignocellulose degradation (Hatakka and Hammel, 2010). Brown rot species contain some of the most abundant degraders of wood within boreal forests, and utilize mainly the carbohydrate components of lignocellulosic biomass, mostly through non-specific degradation through redox chemistry, often leaving behind a brownish residue rich in lignin (Figure 2; Xu and Goodell, 2001; Penttilä, Siitonen and Kuusinen, 2004; Suzuki *et al.*, 2006; Arantes and Goodell, 2014; Rytioja *et al.*, 2014; Vogel *et al.*, 2017). So-called white rots on the other hand employ a wide repertoire of secreted enzymes, with powerful redox enzymes used for lignin degradation, and hydrolases (as well as redox enzymes) against the carbohydrate components. These fungi sometimes appear to selectively degrade most of the lignin before hydrolysis of the carbohydrates, displaying a light-colored residue, hence the term white rot (Figure 2; Hatakka and Hammel, 2011; Mäkelä, Donofrio and De Vries, 2014). Saprotrophic white rots are perhaps the most potent degraders of lignocellulose in nature, by virtue of their powerful redox enzymes conveying the ability to degrade lignin effectively, in addition to their carbohydrate active enzymes (Mäkelä *et al.*, 2014). Lignocellulose degrading ascomycetes, sometimes referred to as soft rots, while unable to degrade

lignin effectively, are able to produce highly effective carbohydrate degrading enzymes (Hatakka and Hammel, 2010). Many of the early industrial strains used for cellulase production were ascomycete species, and in fact even today, many of the commercial cellulosic enzyme cocktails are produced by these fungi (Fasim et al., 2021; Mäkelä et al., 2014; Payne et al., 2015). Probably the most widely studied and applied is *Trichoderma reesei* (Bischof et al., 2016). This fungus has been fundamental in the study of fungal cellulases, and has functioned both as a model organism, as well as a heterologous expression system for enzymes from other sources (Bischof et al., 2016; Hatakka and Hammel, 2010). One of its main advantages, in addition to the genes it contains encoding highly effective cellulases, has been its ability to secrete high titers of enzymes in liquid cultures (Martinez et al., 2008; Rytioja et al., 2014). In fact, each of the investigations included in this thesis involved cellulases originating from, and produced in *Trichoderma* species. While understanding of the fungal cellulolytic enzyme machinery initially largely stemmed from observations on *T. reesei* (Table 1), further studies have shown that the compositions of cellulose degrading secretomes are remarkably similar in many of the most effective cellulose decomposers (Alfaro et al., 2014; Gritzali and Brown, 1979; Haddad Momeni et al., 2013; Marinović et al., 2018; Olson et al., 2012; Ravalason et al., 2008; Wymelenberg et al., 2010), as well as in optimized synthetic enzyme mixtures (Banerjee et al., 2010a; Kallioinen et al., 2014).



Figure 2. Wood showing a typical degradation pattern by white rot (left) and brown rot fungi (right).

Table 1. The most abundant proteins in the secretome when *Trichoderma reesei* is grown on cellulose as a carbon source.

Enzyme	CAZy family ^{a)}	Enzyme type ^{b)}	Hydrolysis mechanism ^{c)}	CBM1 location ^{d)}	Amount (w/w %) ^{e)}
Cel7A	GH7	CBH	Retaining	C-terminal	40-50
Cel6A	GH6	CBH	Inverting	N-terminal	12-20
Cel7B	GH7	EG	Retaining	C-terminal	5-10
Cel5A	GH5	EG	Retaining	N-terminal	~5
Cel12A	GH12	EG	Retaining	-	1-5
Cel3A	GH3	BGL	Retaining	-	1-2
Lpmo9A	AA9	LPMO	Oxidative	C-terminal	?

^{a)} Enzyme family in the Carbohydrate Active Enzymes database. ^{b)} CBH, Cellobiohydrolase; EG, Endoglucanase; BGL, Beta-glucosidase; LPMO, Lytic polysaccharide monooxygenase. ^{c)} The anomeric configuration at the new reducing end of the product is either retained or inverted. ^{d)} Endoglucanase Cel12A and beta-glucosidase Cel3A consist of a single catalytic module. The other enzymes are bimodular with a small cellulose-binding CBM1 module at either the N- or the C-terminus, connected to the catalytic domain by a flexible linker peptide that is usually highly *O*-glycosylated. ^{e)} Fraction in percentage of the total amount of extracellular protein in the culture filtrate (Gritzali and Brown, 1979).

The CAZy database for carbohydrate active enzymes has greatly facilitated the grouping and classification of enzymes based on enzyme structures and sequence similarities (Drula et al., 2022). As of this writing there are 173 distinct families of glycoside hydrolases, with new ones appointed regularly with the discovery of novel structures. The database also includes numerous families for glycosyltransferases (GT), polysaccharide lyases (PL), and carbohydrate esterases (CE), as well as so called auxiliary activities (AA), families of which include many oxidative enzymes, including *e.g.*, lignin degrading peroxidases and LPMOs. CBMs associated with these enzymes are also covered within separate families. Of all of these, as the name suggests, the GH7 family was identified and characterized early on.

Vast majority of carbohydrate hydrolases utilize so called Koshland mechanisms for cleaving *O*-glycosidic bonds (Koshland, 1953; Payne et al., 2015). The primary mechanisms described by Koshland are denoted inverting, and retaining mechanisms, drawing their names from the effect the hydrolysis event has on the orientation of the oxygen on the anomeric carbon, *i.e.*, the glycosidic oxygen prior to bond cleavage and the corresponding hydroxyl group on the resulting product (Figure 3). The inverting mechanism is considered to be a one-step process, where a water molecule performs a nucleophilic attack on the anomeric carbon of a carbohydrate, while a catalytic base abstracts a proton from the water molecule, and a catalytic acid residue protonates the glycosidic oxygen, leading to cleavage of the *O*-glycosidic bond. As the name implies, this mechanism inverts the orientation of oxygen atom on the anomeric carbon in the process, thus converting the sugar at the newly formed reducing end from a β conformation to an α configuration, or vice versa. A retaining mechanism on the other hand, entails a two-step mechanism where a glycosyl-enzyme intermediate is formed. In the first step, a catalytic nucleophile residue attacks the anomeric carbon, leading to a formation of the sugar-enzyme intermediate, while a catalytic acid protonates the glycosidic oxygen, cleaving it from the anomeric carbon. In a following step, the residue functioning as a catalytic acid in the previous phase, acts as a catalytic base and orients a water molecule to perform a nucleophilic attack on the anomeric carbon while abstracting a proton, thus leading to regeneration of the catalytic acid and the enzyme. Consistent with the name, this mechanism retains the α or β conformation of the carbohydrate.

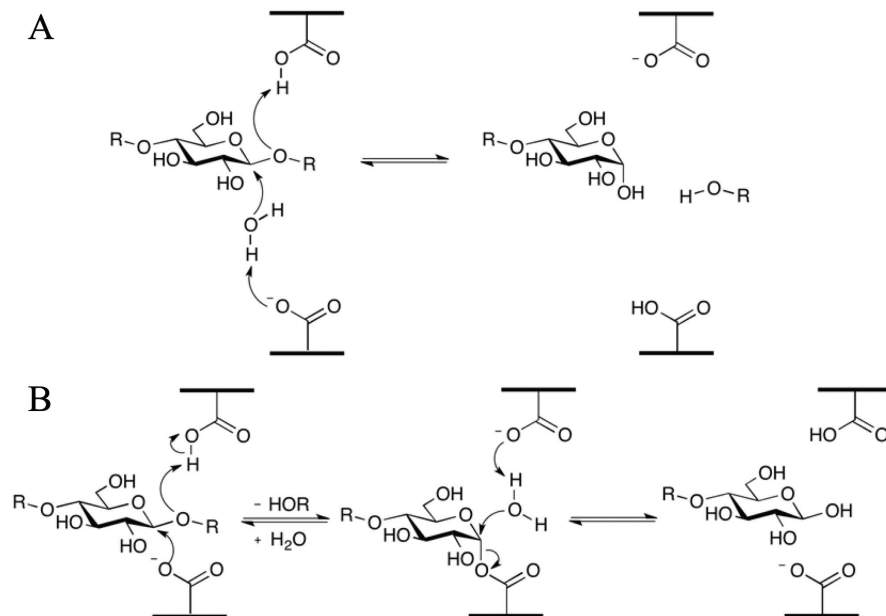


Figure 3. Illustration of the (A) inverting and (B) retaining Koshland mechanisms. Modified from Payne *et al.*, (2015).

Cellulolytic enzymes can be broadly divided into two categories based on the cleaving mechanism, hydrolyzing enzymes performing water mediated hydrolysis of the β -1-4 *O*-glycosidic bonds, and redox enzymes. Glycoside hydrolases cleaving polymeric cellulose can be further divided into two categories based on the main mode of action, cellobiohydrolases (CBH) and endoglucanases (EG). CBHs degrade cellulose in a processive manner, moving along a cellulose chain and cleaving every second *o*-glycosidic bond, with the main product consequently being cellobiose (Rytioja *et al.*, 2014). While CBHs appear to usually have the ability to initiate a processive run from mid-chain positions of cellulose, the predominant mode of starting hydrolysis is threading a cellulose chain end into the active site (Kurasin and Valjamae, 2011; Ståhlberg *et al.*, 1993; Vermaas *et al.*, 2019). These enzymes cleave cellobiose specifically either from the reducing end or non-reducing end, with GH7 CBHs belonging to the former, and glycoside hydrolase family 6 (GH6) enzymes to the latter group (Payne *et al.*, 2015). These two families appear to be among the most abundantly expressed cellulases in many cellulolytic fungi (Alfaro *et al.*, 2014; Gritzali and Brown,

1979; Haddad Momeni et al., 2013; Marinović et al., 2018; Olson et al., 2012; Wymelenberg et al., 2010).

Endoglucanases on the other hand do not possess a processive mechanism, but have a higher tendency to cleave cellulose chains also in medial positions, thus often also providing new entry points for CBHs (Ståhlberg et al., 1993). The structural and functional characteristics of CBHs and EGs will be described in more detail in the following section regarding GH7 enzymes.

Another crucial type of enzyme for cellulose utilization is β -glucosidases, which have most significance in cleaving short soluble cellobiosaccharides into glucose, most importantly cellobiose which is abundant as a result of CBH activity (Lynd et al., 2002). This function is crucial not only to facilitate sugar uptake by the organism, but also to decrease the concentration of cellobiose, which is a strong inhibitor of many cellulases, especially CBHs (Gruno et al., 2004; Kari et al., 2017; Zhang et al., 2013).

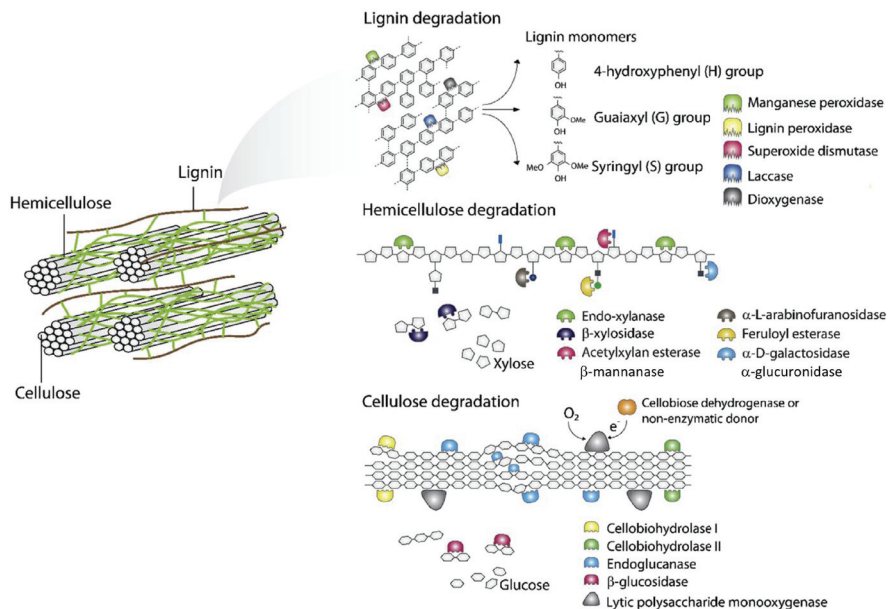


Figure 4. Many enzymes need to co-operate to degrade lignocellulosic plant biomass. Redox-active auxiliary activity (AA) enzymes and non-enzymatic components help to disrupt the material to create more accessible sites for core cellulases, hemicellulose degrading glycoside hydrolases and debranching enzymes, which cooperate synergistically to degrade the lignocellulose. Figure modified from Champreda *et al.*, (2019).

Knowledge of oxidative cellulose degrading enzymes has taken significant steps in recent years, with first reports describing the activity of cellulose specific fungal and bacterial lytic polysaccharide monoxygenases (LPMOs) emerging in 2011 (Forsberg et al., 2011; Langston et al., 2011; Phillips et al., 2011; Westereng et al., 2011). This group of enzymes initially designated GHs (GH61), and CBMs (CBM33), with unknown function, were during this time discovered to act on cellulose through oxidation of the *O*-glycosidic bonds instead of hydrolysis. Subsequently, characterization of various enzymes within this group has revealed that some are specific in oxidizing either the C1 or the C4 carbon of glycosyl units, whereas some show significant activity against both (Beeson et al., 2015). While the first discovered LPMOs were active mostly on insoluble cellulose and chitin, to date LPMOs active on soluble substrates have been characterized as well, with the range of substrates also widening to include hemicellulose and pectin components, as well as starch. In fact, as of the time of this writing there are considered to be eight structural families containing LPMO-like enzymes, comprising the CAZy families AA9-AA11 and AA13-AA17 (Vandhana et al., 2022). Focusing here on cellulose oxidizing LPMOs, the product of C1 oxidation of a cellulose chain is a gluconolactone unit at the reducing end. The gluconolactone rapidly hydrolyses into gluconic acid through a spontaneous reaction. The C4 oxidation product of cellulose is 4-ketoglucose at the non-reducing chain end (Beeson et al., 2015). These relatively small enzymes (commonly <250 amino acids in the catalytic domain) typically possess a flat active site, thus making them suitable for cleaving cellulose chains even in the highly ordered crystalline regions of cellulose microfibrils, something most EGs are unlikely to perform effectively. Indeed, LPMOs have been shown to display significant synergism with processive cellulases (Harris et al., 2010; Hu et al., 2015), likely by opening new chain ends for CBHs to initiate processive runs, and are thus important components of both the fungal cellulolytic machinery as well industrial cellulase mixtures. Initially LPMOs were thought to depend on a constant supply of small compounds as electron donors for their activity via an oxygen dependent reaction (Beeson et al., 2015), but more recently have been shown to also utilize hydrogen peroxide as a co-substrate (Bissaro et al., 2017; Kont et al., 2020). Which is the primary mechanism for these enzymes is a topic that has sparked some fierce debates over recent years, but notably, prerequisites for both are commonly available in environments

where lignocellulose is being degraded by fungi. While a number of studies have shown that the monooxygenase reaction is orders of magnitude slower than the peroxygenase mechanism for these enzymes (Hedison et al., 2021; Jones et al., 2020; Kont et al., 2020; Kuusk et al., 2018), they are most often still referred to as LPMOs by convention.

Another redox enzyme group involved in fungal cellulose degradation is cellobiose dehydrogenases (CDH). These enzymes were initially thought to be oxidases (Ayers et al., 1978), but were later renamed dehydrogenases as oxygen was demonstrated to be a poor electron acceptor in their reactions (Bao et al., 1993). As the name suggests, the main reaction catalyzed by CDHs is the abstraction of electrons from cellobiose, leading to formation of cellobionolactone and subsequent spontaneous reaction into cellobionic acid (Henriksson et al., 2000). It is unlikely this reaction is the primary biological role of these enzymes, as cellobionic acid in itself is not known to play any fundamental role in fungal growth, or degradation of lignocellulose, and disposing of cellobiose via other means is likely to be metabolically more favorable as it can readily be converted to glucose (Henriksson et al., 2000). Instead it is more likely these enzymes exist to function as an electron source for other redox enzymes, and have indeed been shown to transfer electrons to LPMOs (Courtade et al., 2016; Tan et al., 2015). All of the thus far biochemically characterized CDH enzymes in their native form display a two or three domain architecture, where an N-terminal cytochrome domain (AA8 family in the CAZy database) is connected through a flexible linker to a FAD-cofactor containing glucose-methanol-choline oxidoreductase (GMC)-type dehydrogenase domain (CAZy family AA3), with some specimen also containing a C-terminal CBM (Sützl et al., 2019). However, within sequence databases many sequences only containing the AA3-domain are also found. Four subfamilies (I-IV) of CDHs have been denoted based on sequence analysis, with protein structures or detailed characterizations only available for enzymes from subfamilies I and II (Sützl et al., 2019). Within this project I have solved, to my knowledge, the first known structure of a subfamily III CDH, but this work was left out of the scope of this thesis (unpublished data).

In addition to enzymes considered cellulases, primarily active on β -1-4-*O*-glycosidic bonds, there is a wide array of enzymes indirectly facilitating cellulose degradation. These include various enzymes degrading the hemicellulose components, which are an abundant part of the lignocellulosic

matrix, as well as pectinases, and lignin depolymerizing or modifying enzymes (Figure 4; Champreda *et al.*, 2019).

It is important to note the different nature of lignocellulosic substrates in industrial settings and in nature. Biomass used in industrial cellulosic ethanol production is most often subjected to pre-treatment, where usually some combination of physical fragmentation, heat, pressure and acid or base is applied (Agbor *et al.*, 2011). Generally, the main purpose for these processes is to make the material more susceptible to enzymatic degradation, by way of making the carbohydrate components more solvent and enzyme accessible by disturbing the dense lignocellulosic matrix. Given that the method of pretreatment can affect the composition of the biomass, it can also have implications on the optimal composition of enzymes used for hydrolysis (Banerjee *et al.*, 2010b). *E.g.*, pretreatment methods with acidic conditions tend to hydrolyze and solubilize the hemicellulose component, leaving most of the cellulose and lignin in the solid fraction (Agbor *et al.*, 2011). Alkaline treatment on the other hand solubilizes lignin to a greater degree while majority of the cellulose and hemicellulose remain insoluble (Galbe and Wallberg, 2019). Pretreatment methods remain an intense subject of study, with the ultimate goal of improving the economics of lignocellulosic biorefineries through higher levels of biomass utilization, both through higher yields, as well as the production of higher value products (Galbe and Wallberg, 2019). While a concept of biorefineries where the lignin component is prioritized has received increasing attention in recent years, the vast majority of research and the cellulosic bioethanol plants in function today prioritize the hydrolysis and fermentation of the carbohydrate components, while lignin is treated as a leftover residue from the process (Korányi *et al.*, 2020). Therefore, the lignin component is not commonly taken into consideration in the development of enzyme cocktails for these processes (Mäkelä *et al.*, 2014).

While lignin degrading enzymes arguably do not at present constitute crucial components of commercial cellulase mixtures, they are worth mentioning here as they play an important role in the process of lignocellulose recycling in nature. The main lignin degrading fungal enzymes constitute of laccases, manganese peroxidases (MnP), lignin peroxidases (LnP), and versatile peroxidases (VP) (Mäkelä *et al.*, 2014). The latter three belong to so called class II peroxidases, and appear to have evolved within basidiomycetes (Ayuso-Fernández *et al.*, 2019). Of these,

LnPs and VPs are seemingly the ones giving basidiomycete fungi their superior ability to depolymerize lignin, as they can depolymerize even non-phenolic parts of the polymer, and do not require metal ions or other small molecules as redox mediators for their depolymerizing action on lignin (Ayuso-Fernández et al., 2018; Ruiz-Dueñas et al., 2009).

2.4 GH7 cellulases

The GH7 family contains both CBHs (EC 3.2.1.176) and EGs (EC 3.2.1.4), hydrolyzing β -1-4-*O*-glycosidic bonds through a retaining mechanism (Payne et al., 2015). While they do display activity on xylose polymers and some mixed linkage glucose polymers, their primary relevance in nature is most likely to the degradation of cellulose, which they hydrolyze with great efficiency.

The first crystal structure of a GH7 enzyme was published in 1994, when the catalytic domain (CD) of *T. reesei* Cel7A (TreCel7A) was solved (Divne et al., 1994). The structure showed a jelly-roll type β -sandwich forming the core of the single domain catalytic unit, and several loops forming an enclosed tunnel, the eight key loops now often referred to as A1-A4 and B1-B4 (Haddad Momeni *et al.*, 2013; Figure 5). The first published GH7 EG structure, *T. reesei* Cel7B (TreCel7B), showed a similar fold at the core, but fewer loops encompassing the active site, forming rather a groove like shape instead of a substrate binding tunnel (Kleywegt et al., 1997). Subsequently it has become clear that there is a distinct division into CBH and EG structures (Gado et al., 2021), the main differentiator being the aforementioned differences in the shape of the substrate binding site, with EGs having significantly less pronounced, or entirely absent A4, B2, B3 and B4 loops (Figure 5, Figure 6). In further parts of this thesis I will describe what I consider to be a third type of architecture for GH7s and have solved the first structure for, that is, short endoglucanases (sEG) also lacking the B1 loop (Figure 5).

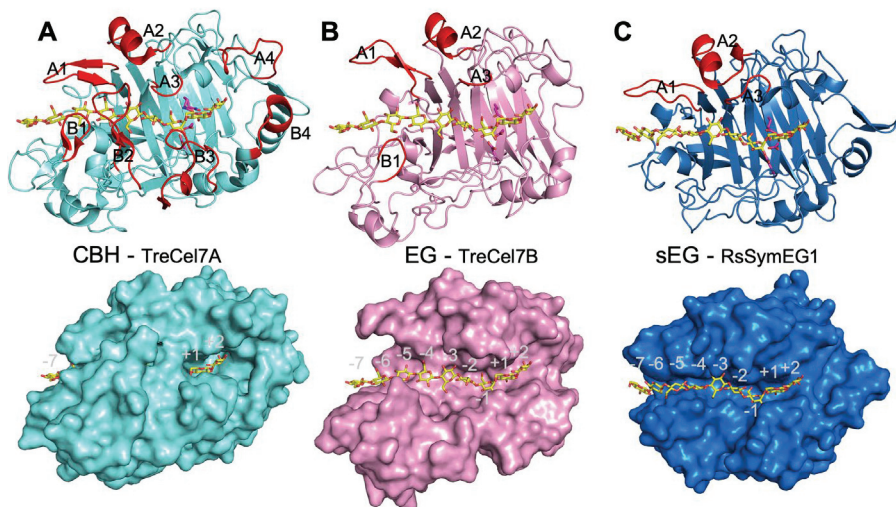


Figure 5. The three types of structure architectures seen in the GH7 family, with the loops surrounding the substrate binding site highlighted in red.

Loops in TreCel7A residue numbering (excl. Signal peptide)		~95-104 ~402-412 ~369-372 ~380-391 ~51-56 ~192-205 ~234-254 ~335-341								
Known structures		PDB accession:	A1-loop	A2-loop	A3-loop	A4-loop	B1-loop	B2-loop	B3-loop	B4-loop
CBH	Trichoderma reesei Cel7A (TreCel7A)	4C4C								
	Geotrichum candidum Cel7A	4Z2U								
	Aspergillus fumigatus	4V20								
	Trichoderma harzianum	2Y0K								
	Limnoria quadripunctata Cel7B	4IPM								
	Heterobasidion irregulare Cel7A	2XSP								
	Melanocarpus albomyces Cel7B	2RFZ								
	Humicola grisea var thermoidea	4CSI								
	Phanerochaete chrysosporium Cel7D (PchCel7D)	123V								
	Rasamsonia emersonii Cel7A	3PL3								
	Dictyostelium discoideum Cel7A	4Z2Q								
	Dictyostelium purpureum Cel7A	4Z2P								
	Daphnia pulex GH7	4XNN								
	Penicillium funiculosum	4XEB								
EG	Trichoderma reesei Cel7B (TreCel7B)	1EG1								
	Humicola insolens Cel7B	1OJJ								
	Fusarium oxysporum EG1	10VW								
	Trichoderma harzianum	5W0A								
	Rasamsonia emersonii Cel7B	6S08								
	Reticulitermes speratus symb. EG1 (RsSymEG1)									

Figure 6. Loop regions surrounding the substrate binding site in known structures of GH7 enzymes. Green: loop is present, yellow: loop is >2 amino acid residues shorter than on the model enzyme TreCel7A, red: loop is not present.

Later structures with bound cellulose substrates have enabled denoting binding sites within the substrate binding tunnel, and have given detailed information about enzyme-substrate interactions (Divne et al., 1998). Based on these and subsequent structures, GH7 enzymes usually contain binding positions for at least 9 glucose units within a cellulose chain, with the naming convention designating the positions from the direction of the non-reducing

end towards the reducing end with ascending numbering, with negative numbers for binding sites on the non-reducing end side of the catalytic site, and positive numbers for positions at the reducing end (Figure 5; (BIELY et al., 1981; Davies et al., 1997)).

For all known GH7 enzymes the catalytic residues consist of three conserved amino acids, two glutamic acid residues acting as a catalytic nucleophile and a catalytic acid/base, as well as an assisting aspartic acid residue critical for catalytic activity (Payne et al., 2015). The active site and the adjacent substrate binding positions contain several conserved residues (Figure 7). Using the residue numbering of the model enzyme TreCel7A as an example, the three catalytic residues at the active site constitute Glu212, Asp 214 and Glu217, with the two residues between the latter two constituting a so called β -bulge in the β -strand forming the base for the catalytic site (Figure 7). A catalytic motif with such a close proximity of the catalytic residues in the primary sequence is not common among GH enzymes. The same configuration can be seen in the related GH16 family, which also shares some of the core structure, with a substrate binding groove formed by an array of β -sheets (Eklöf and Brumer, 2010; Michel et al., 2001; Payne et al., 2015).

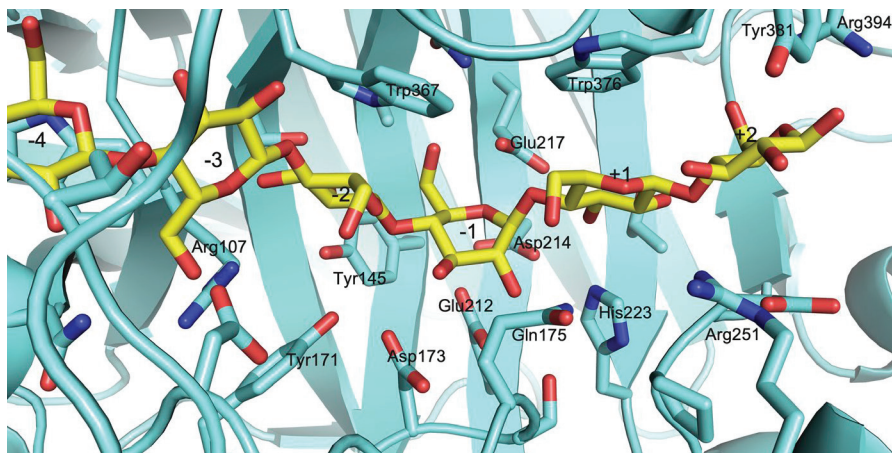


Figure 7. The catalytic site of TreCel7A showing residues involved in interactions with substrates/products, including the catalytic nucleophile Glu212, assisting residue Asp214 and catalytic acid/base Glu217. Protein structure from PDB structure 1CEL, superposed with a cellulose chain from PDB structure 4C4C.

In the immediate vicinity of the catalytic residues, Asp173, Gln175 and His228 contribute to substrate binding, and on the other side of the tunnel, further two residues conserved within GH7s, Trp367 and Trp376, form a “platform/roof” for substrate binding right at the active site, contributing to interactions with the substrate at binding sites -2 and +1, respectively. Most enzymes are considered to contain between 7 and 9 binding sites before the active site, and 1 to 3 product binding sites, with several key aromatic and arginine residues contributing to substrate binding (Payne et al., 2015).

In many cases GH7 enzymes are found in a two-domain architecture, with an N-terminal GH7 domain, combined with a C-terminal type 1 CBM, the two domains connected by a flexible linker region (Gado et al., 2021). The linker peptide is typically heavily glycosylated, contributing to both stability through protease resistance, as well as increased affinity towards substrates through carbohydrate interactions (Amore et al., 2017; Payne et al., 2013). Other than the CBM/ no CMB architectures, there is little to no variation in the domain combinations seen in GH7 enzymes, unlike for some other GH families (Gado et al., 2021). One feature limiting possible domain combinations is a post-translational modification seen at the N-terminus of most GH7 enzymes. This is cyclization of a glutamine residue into pyrroglutamate, also known as pyrrolidone carboxylic acid (PCA). In many GH7 structures this residue is an integral part of the structure, where PCA caps the N-terminus by binding in a hydrophobic pocket, thus hindering initiation of protein unfolding from this position (Payne et al., 2015). Absence of this modification leads to reduced thermal stability, as shown by analysis of enzymes from expression hosts where this modification is not complete (Dana et al., 2014). Naturally however, this capping of the N-terminus is not compatible with domains connected to a GH7 CD through this end of the amino acid chain, perhaps partly explaining the lack of diversity in GH7 domain combinations.

Given the importance of GH7 enzymes in industry and in natural cellulose degradation, these enzymes have been studied extensively. A myriad of biochemical and computational methods have been used to investigate the mechanism of these enzymes, and what makes them such efficient cellulose degraders (Payne et al., 2015).

Various structures of GH7 enzymes have been published since the first CBH and EG structures. While GH7 enzymes have been studied most extensively in ascomycete and basidiomycete fungi, and this is where GH7

sequences seem to be most abundant, representatives of this family have been characterized from various other organisms as well, both structurally and biochemically. Surprisingly, perhaps the most striking feature of these enzymes is in fact their similarity. Enzymes from incredibly diverse eukaryotes show very similar structures, the major distinction being the division into CBH and EG architectures (Gado *et al.*, 2021; Figure 5, Figure 6). While GH7s have only been detected in eukaryotes, looking at a phylogenetic tree of eukaryotic organisms, examples of these enzymes are found in remarkably wide range of organisms (Figure 8). This suggests an ancient origin of this enzyme family, and makes the conserved, uniform nature of GH7 enzymes even more conspicuous. This uniformity is not caused by arbitrary classification into highly similar enzyme families either, as members of the closest related family, GH16, differ significantly from GH7 enzymes, both in sequence and structure, as well as function. While I will not make a comprehensive review of all non-fungal GH7s here, it is worth discussing a few examples of GH7s found in other organism groups in order to illustrate the diversity of host organisms, and the functions of these enzymes in them.

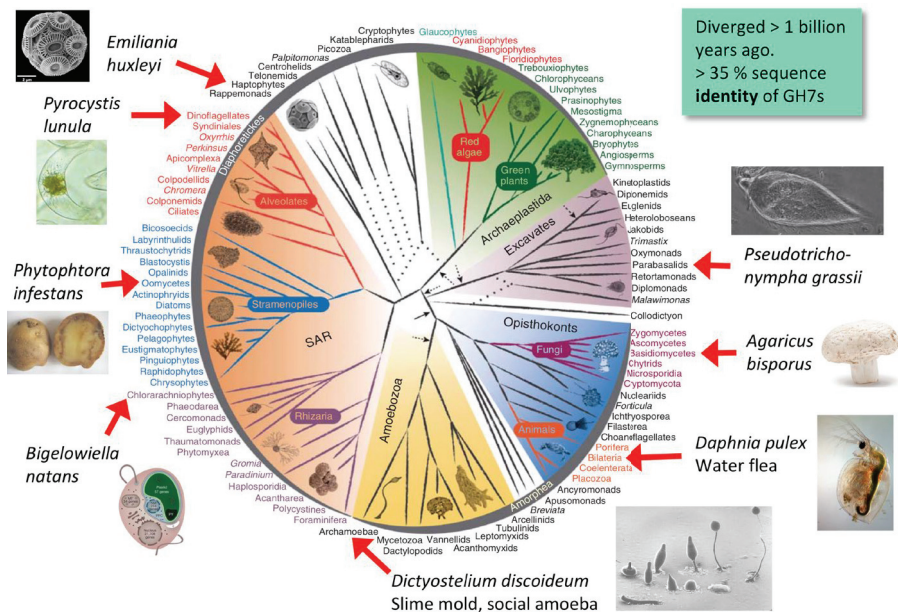


Figure 8. Examples of distant species with GH7s in the evolutionary tree of Eukaryotes. GH7 genes have been reported in all clades, except Archaeplastida (plants and algae), indicating very ancient origin or multiple horizontal gene transfer events. Phylogenetic tree from (Burki, 2014), modified with images from (Hobdey *et al.*, 2016; Nakashima *et al.*, 2002).

Examples of GH7 enzymes utilized by an organism not for subsistence, but most likely for morphological rearrangement, were described by Hobdey *et al.*, (2016) in their article detailing two homologous CBHs from two species of social amoeba of the genus *Dictyostelium*, also known as slime molds. The authors noted the remarkable similarity of these enzymes (in both sequence and structure) to TreCel7A and other GH7 CBHs from ascomycetes, and hypothesized that this could possibly be explained by horizontal gene transfer.

The first GH7 enzyme from an animal was described by Kern *et al.*, (2013) with the characterization of a CBH from the marine wood borer *Limnoria quadripunctata*. Interestingly this enzyme showed a highly acidic surface charge, something the authors hypothesized could be due to the high-salt marine environment the enzyme is utilized in by the organism, and showed it retains its activity at NaCl concentrations as high as 4 M. Notably the enzyme still displayed a pH optimum typical of GH7 enzymes of 4.0-4.5,

indicating that the surface charge profile likely was not an adaptation to alkalinity of the environment.

GH7 enzymes from wood eating termites are a somewhat recent discovery (Nakashima and Azuma, 2000; Watanabe et al., 2002). To be exact, while termites are capable of producing intrinsic cellulases, among others from GH9 and GH16 families, for GH7 enzymes they depend on symbiotic organisms (König et al., 2013). So-called lower termites are a class of wood eating termites which harbor symbiotic protozoa in their hindguts, and depend on them for effective cellulose degradation (Cleveland, 1924, 1923).

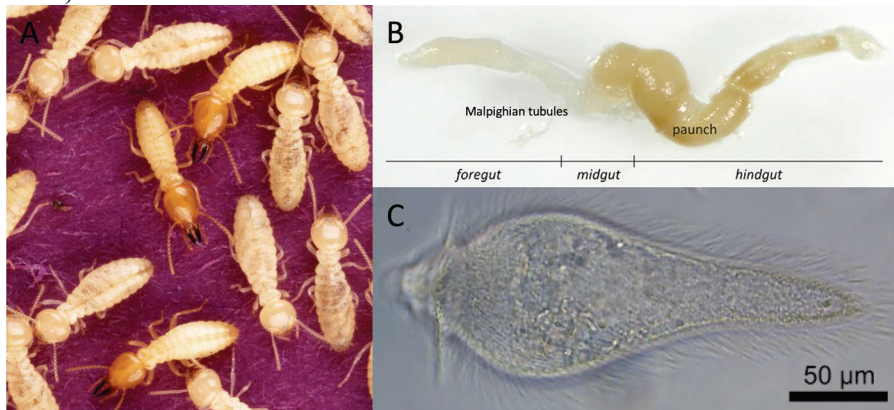


Figure 9. (A) Wood-eating termites of the species *Coptotermes formosanus*. (B) Extracted gut of the termite *Reticulitermes grassei*, showing the foregut, midgut, as well as the hindgut where GH7 producing symbiotic protozoa reside. (C) *Pseudotrichonympha grassii*, a symbiotic parabasalid protist often found in the hindguts of lower termites. Image credits: (A) Scott Bauer, U.S. Department of Agriculture; (B) Modified from Duarte *et al.*, (2018); (C) Nishimura *et al.*, (2020).

These unicellular eukaryotic flagellates belong to the organism groups of parabasalids and oxymonads, and already in the 1920s were shown to be crucial for survival of the host termites when growing on cellulosic substrates (Cleveland, 1924, 1923). Much more recently, metagenomic, genomic, transcriptomic, and protein analysis has shown GH7 enzymes to be major, if not the most important secreted cellulose degrading enzymes in these protists (Geng et al., 2018; Todaka et al., 2010a, 2007; Watanabe et al., 2002). Emergence of sequencing data has shown that these organisms contain both CBHs and EGs from the GH7 family. Interestingly, the majority of the EG sequences are quite different from the archetypal GH7 EGs, being significantly smaller (typically 60-80 residues shorter). Below in the

'Current investigation' -chapter, as well as in **Paper II**, I describe the GH7 enzymes from these protists in more detail, and describe the first known structure for one these short GH7 endoglucanases (sEGs), which I have solved. During this project I also conducted an expression campaign to heterologously express several termite symbiont sEG GH7s in *T. reesei*, in order to further examine the structural and functional diversity within this under-explored enzyme group, but the efforts were not successful in generating adequate quantities of active enzyme.

While conducting phylogenetic analysis on GH7 sequences as part of **Paper II**, I found similar short sEG sequences in another group of organisms in addition to the termite symbiont protozoa, namely, water fleas. These Arthropoda in the genus *Daphnia* have previously been shown to contain GH7s with the deposition of *Daphnia pulex* CBH structures into the Protein Data Bank (PDB; *E.g.*, entry 4XNN). However, the sEGs from this group have not been characterized to my knowledge. It is not clear if these enzymes have emerged independently of the sEGs in the termite symbiont protists, but their existence does provide further evidence supporting the viability and utility of the sEG GH7 architecture, alongside the archetypal EGs and CBHs.

One more group of organisms where GH7s are found, but have received very little attention, is oomycetes. These filamentous eukaryotic organisms have diverse lifestyles, some species being animal pathogens, but the group also containing some highly significant plant pathogens, such as the potato blight causing *Phytophthora infestans* (Judelson, 2017; Sabbadin et al., 2021b, 2021a). Sequencing studies have shown that many species carry a high number of GH7 genes, for example some *Aphanomyces* species showing over 20 unique sequences. However, very little is known of the role of these enzymes within this organism group. Notably the oomycete cell walls contain cellulose, something which could suggest the use of cellulases not only for plant matter degradation, but also for self-regulation of cell wall composition, *e.g.*, for germination or growth (Mélida et al., 2013). One recent study indicated that a CBH, PsGH7a, contributed to the virulence of *Phytophthora sojae* on soybean plants (Tan et al., 2020). A transcriptomics study of the *Pisum sativum* (pea plant) pathogen *Aphanomyces euteiches* confirmed that several GH7 enzymes were expressed, but it is not clear from this study what their exact role is (Gaulin et al., 2018). However, some of the GH7 sequences found in this species are highly interesting, enough so to warrant a closer look. For example, there are EG sequences (EGs based on

the absence of B2 loop) completely missing the A1, A2 and A3 loops, which is not seen in any thus far characterized GH7 enzymes (Figure 6), and implies a substrate binding site with a highly open architecture, even compared to the sEG sequences discussed above. Meanwhile, some of these EGs actually contain CBH-like, long B3 and B4 loops, again, something that is not seen in other GH7 EGs (Figure 6), and could perhaps even suggest a relatively late emergence of EGs within oomycetes through CBH loop deletions. Additionally, at least one EG sequence is seen with three back-to-back type 1 CBMs (GenBank accession: KAF0737786.1), there is a CBH sequence which has five such CBMs in a row (KAF0726178.1), and there is one gene which appears to be a multi-domain GH7 with an N-terminal GH7 CD, a CBM connected by flexible linkers, and an N-terminal GH5 domain (KAF0726201.1), a configuration which interestingly resembles effective synthetic GH7 fusion-protein constructs characterized by Brunecky *et al.*, (2020). While caution should be applied when analyzing sequences purely based on genomic data, and the remarkable multiplication and shuffling of these genes has likely also brought about some inactive pseudogenes, it is noteworthy that several of these sequences are backed by RNA sequencing data as well. To me, all this suggests there could be significant unexplored diversity of GH7s within oomycetes. Interestingly, the hypothesis of somewhat late emergence of EGs in oomycetes is supported by the clustering of oomycete CBHs and EGs in phylogenetic analysis of GH7s, described in more detail in **Paper II**. Within this project, several *Aphanomyces* GH7s were included in an expression campaign in *T. reesei*, but no GH7 activity was detected from any of the transformants obtained. Additionally, a short crude study was conducted to see if GH7 activity could be detected directly from samples of *A. euteiches* infected pea plants. Sampling of filtrates of homogenized plant tissue showed activity on a fluorogenic cellulase substrate methylumbelliferyl lactoside in a zymogram study, but the likely source of this activity turned out to be β -glucosidases belonging to family GH1. It should be mentioned however, that the sensitivity of this crude method is likely quite low, and no further attempts were made to detect GH7s from these samples.

2.5 Recombinant expression in *Trichoderma reesei*

As mentioned earlier, enzymes expressed in species of *Trichoderma* (*T. reesei* and *T. longibrachiatum* to be exact) were utilized in all of the current investigations included in this thesis. This is no accident, as strains of these fungi have been instrumental in cellulase research throughout the several past decades, and still widely in use (Bischof et al., 2016). This has been on one hand due to many industrial cellulase producing strains having been developed from these fungi, making them a logical choice also for expression of new enzymes which are targeted for similar uses, as compatibility with the ultimate expression host is crucial. More crucially however, this has been to a large part due to *Trichodermas* relative simplicity and stability in terms of life-cycle and genome, suitable morphology for expression both in small and large scales, the cumulative effect of iteratively increasing species specific genetic tools available, ability to express fully functional cellulases which are poorly expressed in many other common hosts, and perhaps most importantly, its ability to produce high titers of secreted enzymes (Hatakka and Hammel, 2010; Mäkelä et al., 2014; Martinez et al., 2008; Peterson and Nevalainen, 2012; Seidl et al., 2009). In this section I will give a very brief overview of the use of *T. reesei* for heterologous protein expression, and a short description of the methods I have used in my work with this fungus. It is worth noting this is purely from a point of view of an academic user, as there is very little information available on the systems used in industry.

Ever since the discovery of *T. reesei* during the second world war, and its subsequent characterization as an efficient cellulase producer, efforts have been made to further improve its ability for cellulase secretion (Bischof et al., 2016). Significant advances were made, and by 1970s cellulase hyperproducer strains such as the RUT-C30 were available (Peterson and Nevalainen, 2012). Since then, most of these efforts have resumed in the private domain, with companies developing their own commercial enzyme cocktail producing strains behind closed doors. Academic research efforts have continued however, most importantly to yield molecular tools for using *T. reesei* as a heterologous expression host for expressing high titers of various types of enzymes. Special mention should be given to the continuous efforts made at VTT in Finland, where important work on transformation methods and selection markers was conducted during 1980s and 90s (Penttilä et al., 1987; Singh et al., 2015). In fact, expression strain used in producing

the TreCel7A enzymes used in **Paper I** were constructed at VTT. More recent work there has demonstrated successful use of CRISPR-Cas9 for genome editing, and synthetic promoters and knock-out strains providing high inducible expression of proteins with low background, *i.e.*, reduced quantities of proteins other than the expression target (Rantasalo et al., 2019, 2018). However, to my knowledge these strains with many of the secreted genes silenced or removed are unfortunately not publicly available.

For expression of enzymes used in the work described in **Paper III**, I have utilized a somewhat simpler system, based on a strain denoted Ast1116, a derivative of the original isolate QM6a, where gene for the protein expressed in highest quantities under cellulase induction, TreCel7A, has been silenced (Linger et al., 2015). While the expression levels of this strain are not as impressive as many of the more developed strains, it is publicly available, and due to its TreCel7A deletion is suited for expression and study of other GH7 CBHs. The transformation was conducted using a plasmid (pTrEno) and a spore electroporation method developed at National Renewable Energy Laboratory (NREL) in the United States, utilizing a constitutive enolase promoter native to *T. reesei* (Linger et al., 2015; Figure 10). This plasmid does not contain so called homology arms, *i.e.*, sequences for targeting integration into specific locations in the genome, but instead relies on random integration. Thus, the use of this system requires screening for finding the most suitable transformants. In the case of GH7 enzymes this is a somewhat straightforward task, as the presence of these enzymes in culture filtrates can easily be gauged using colorimetric assays with chromogenic substrates.

On another part of my project I also used a protoplast-based transformation method developed at VTT (Penttilä et al., 1987), during my abovementioned attempts to express several GH7 sequences from oomycetes and termite symbiont protists, as well as unrelated enzymes called loosenins. Unfortunately, none of the transformants yielded from these efforts produced these GH7s in detectable quantities. However, the loosenin constructs were prepared with a 2A-peptide dependent bicistronic version of the pTrEno expression plasmid for co-expression with enhanced green fluorescent protein (eGFP) in order to enable fluorescence based screening (Subramanian et al., 2017). Several transformants were obtained, all showing eGFP expression and thus confirming successful transformation using this method (Figure 10).

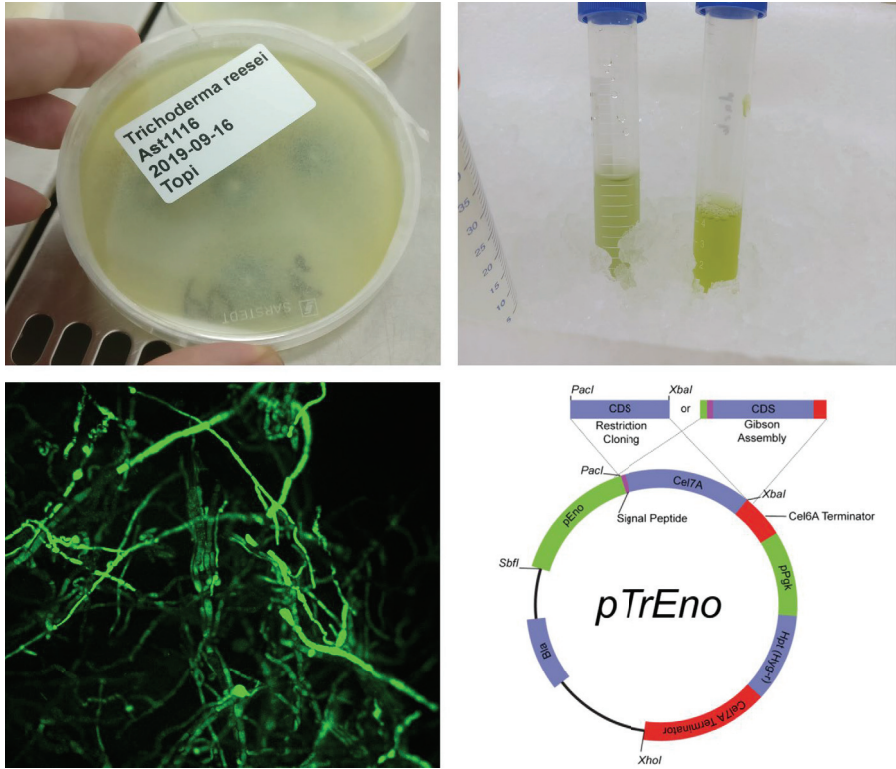


Figure 10. Top left, *T. reesei* grown under light for harvesting spores for transformation using an electroporation method. Top right, harvested spores. Bottom left, confocal microscope image of *T. reesei* mycelium showing fluorescence from eGFP, demonstrating successful transformation with pTrEno-2A-eGFP plasmid by a method utilizing chemical transformation of protoplasts, photo credit Laura Okmane. Bottom right, schematic of the pTrEno plasmid, image from Linger *et al.*, (2015).

3. Current investigation

3.1 Enzyme kinetics by GH7 cellobiohydrolases on chromogenic substrates is dictated by non-productive binding: insights from crystal structures and MD simulation (Paper I)

3.1.1 The use of chromogenic and fluorescent model compounds in GH7 research

The use of chromogenic and fluorescent substrates for assessing cellulase activity became commonplace early on for assessing cellulase activity. These compounds have been widely used most of all due to them being conducive to fast and simple workflows for enzyme activity measurements, with high repeatability even in small volumes. Their use bypasses the problems involved with insoluble cellulosic substrates, where difficulty of handling small sample sizes due to substrate and sample heterogeneity, and multi-step procedures for measuring released sugars lead to low precision and slow sample turnover. Over time it has become clear however, that enzyme behavior on small soluble substrates is poorly correlated to their properties on polymeric substrates, reflected by vastly different values obtained from enzyme kinetics measurements. While this is perhaps not surprising given the inherently different nature of soluble and insoluble substrates, the exact reasons for the somewhat idiosyncratic hydrolysis characteristics of these compounds have remained unclear. Especially curious are the substantial differences in catalytic constants between seemingly very similar model compounds. For example, two widely used p-nitrophenyl (pNP) model compounds pNP-cellobioside (pNPC) and pNP-lactoside (pNPL) differ only

in the orientation of the hydroxyl group (4OH) on the C4 carbon at the non-reducing end of the disaccharide, being in equatorial position on pNPC, and axial orientation in pNPL. How this would have a drastic influence on binding or hydrolysis of these compounds is not immediately clear, and it is uncertain if differences are governed by differences in productive binding, or perhaps inhibition through unproductive interaction between the substrate and the enzyme. In **Paper I**, we set out to study the interactions of two model CBH enzymes of the GH7 family, TreCel7A and PchCel7D, with four model compounds commonly used in cellulase research, pNPC, pNPL, o-nitrophenyl cellobioside (oNPC), and methylumbelliferyl cellobioside (MUC; Figure 11). We utilized enzyme kinetics, x-ray crystal structures of substrate bound enzymes where possible, and molecular dynamics (MD) simulations for attempts to gauge the relative binding strengths of these model compounds in productive and non-productive poses. Inspired by the observation that the hydrolysis rate of oNPC was very low on these enzymes, we also explored the possibility to use this compound for fluorescence titration experiments to determine cellobiose binding constants for these enzymes.

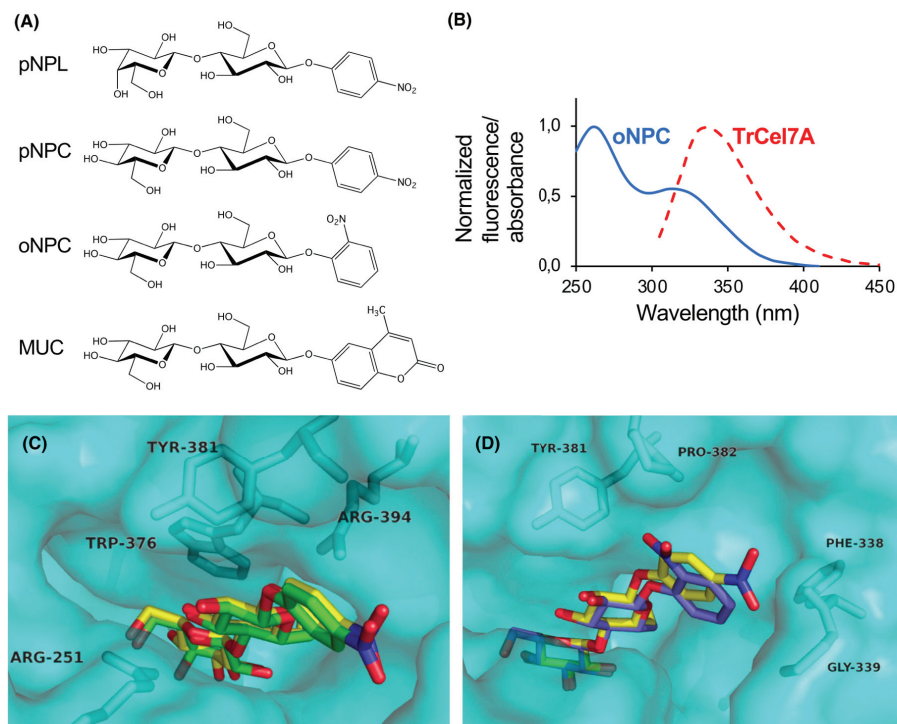


Figure 11. (A) The four model compounds used in the study. (B) The fluorescence spectrum of TrCel7A when excited at 295 nm and the absorbance spectrum of oNPC. (C) pNPC (yellow) and pNPL (green), as well as (D) pNPC (yellow) and oNPC (blue), shown binding at the product binding site of TrCel7A.

3.1.2 Enzyme kinetics, crystal structures and molecular dynamics simulations

Enzyme kinetics values varied significantly between the different substrates and the two enzymes (Table 2). Interestingly, oNPC showed slow hydrolysis turnover rates (k_{cat}) on both enzymes, but especially so on TrCel7A where the k_{cat} was only $66 \times 10^{-6} \text{ s}^{-1}$, compared to that of PchCel7D at 0.015 s^{-1} . Moreover, on TrCel7A pNPL showed an order of magnitude faster hydrolysis than the very similar pNPC, 33 times faster to be exact. There was a similar effect on PchCel7D, but to a much lesser extent, with the difference being 3.7x between k_{cat} values for the two substrates. Overall, k_{cat} values on all substrates were significantly higher on the PchCel7D, but so were the binding constants (K_{M}), making the $k_{\text{cat}}/K_{\text{M}}$ values quite similar for the two enzymes, indicating comparable catalytic efficiencies.

Table 2. Kinetic constants for TreCel7A and PchCel7D on the four tested model compounds.

Enzyme	Substrate	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($\text{s}^{-1}*\text{M}^{-1}$)
<i>TreCel7A</i>	oNPC	$66*10^{-6}\pm 15*10^{-6}$	7.0 ± 4.5	9.5
	pNPC	0.0026 ± 0.0001	26 ± 3	100
	pNPL	0.087 ± 0.002	590 ± 20	147
	MUC	0.013 ± 0.001	12 ± 1	1083
Ratio	pNPC/oNPC	39	3.7	11
	pNPL/pNPC	33	23	1.5
	MUC/pNPC	5.0	0.46	11
<i>PchCel7D</i>	oNPC	0.015 ± 0.002	3200 ± 100	4.6
	pNPC	0.046 ± 0.0021	1300 ± 160	35
	pNPL	0.17 ± 0.01	5500 ± 400	31
	MUC	0.22 ± 0.01	210 ± 20	1048
Ratio	pNPC/oNPC	3.1	0.41	7.5
	pNPL/pNPC	3.7	4.2	0.87
	MUC/pNPC	4.8	0.16	30
Ratio	oNPC	227	457	0.50
PchCel7D/	pNPC	18	50	0.35
TreCel7A	pNPL	2.0	9.3	0.21
	MUC	17	18	0.97

Due to the slow hydrolysis rates of oNPC, we hypothesized that it could be used as a binding probe for these enzymes, utilizing intrinsic protein fluorescence. Fluorescence measurements showed that oNPC indeed quenched the fluorescence emission of the enzymes at 340 nm when excited at 280 nm (Figure 11). This allowed titration experiments to determine the dissociation constants K_d for oNPC binding. Furthermore, in the case of TreCel7A, titration with cellobiose recovered the fluorescence quenched by oNPC binding, allowing assessment of K_d for cellobiose. However, this was not the case for PchCel7D, where fluorescence was not restored even at high concentrations of cellobiose (1 mM). Since the enzyme kinetics or subsequent MD simulations did not suggest this to be due to strong binding

at the substrate or product binding sites, this likely indicates high affinity (of oNPC and not cellobiose) towards a position on the enzyme which is outside the active site. This is also supported by the apparent K_M -values which on both enzymes are higher than the determined K_d constants for oNPC, something that is not consistent with strictly overlapping productive and non-productive binding positions.

We were able to obtain x-ray crystal structures of TreCel7A with three of the model compounds, pNPC, pNPL and oNPC, hoping to see clues as to the causes of the peculiar kinetics on this enzyme. In all cases the substrate was seen binding at the +1, +2, +3 product binding sites, confirming that unproductive binding is likely the dominant mode of binding for these substrates on this enzyme (Figure 11). Significant differences in the k_{cat} and K_M values for the two similar compounds pNPL and pNPC can likely be attributed to the strength of non-productive binding of the substrates in the product binding sites. Weaker interactions between the enzyme and the non-reducing end 4OH group on pNPL likely lead to weaker non-productive binding compared to pNPC, consistent with the higher k_{cat} and K_M values observed with the former. The slower turnover of oNPC on both enzymes is not explained solely by strong non-productive binding at the product binding site, as this should lead to reduction in k_{cat} and K_M by the same factor, k_{cat}/K_M thus remaining more or less unchanged. This is not the case as k_{cat}/K_M for oNPC is roughly an order of magnitude lower than the values for pNPL or pNPC, on both enzymes. It is not obvious from the crystal structures or the MD simulations what other factors contribute to the inefficient hydrolysis of oNPC, but binding of this substrate somewhere outside the active site, as suggested by the K_M and K_d -values could have an influence. It is also not clear what makes MUC vastly superior substrate for these enzymes compared to the pNP based compounds.

3.1.3 Conclusions

The peculiar enzyme kinetics on widely used cellulase model compounds have been observed repeatedly, but the exact reasons behind their characteristics have remained unclear. Through a combination of enzyme kinetics, MD simulations and analysis of protein crystal structures we determined that the strength of non-productive binding is likely to be a major, but not only, factor causing the differences between the activity profiles of TreCel7A and PchCel7A on the nitrophenyl substrates, as well as the

differences between each substrate. Furthermore, a recent study demonstrated a strong correlation between the free energy of binding and free energy of catalysis across several cellulase families, suggesting that tools facilitating assessment of true disassociation constant are highly valuable (Kari et al 2021). We explored the use of oNPC as a fluorescent molecular probe for determining disassociation constants for GH7 enzymes, and while we successfully used this compound to assess K_d of cellobiose on TreCel7A, the same approach did not work in the case of PchCel7A, and is thus not likely to be universally applicable.

3.2 The crystal structure of RsSymEG1 reveals unique form of smaller GH7 endoglucanases alongside GH7 cellobiohydrolases in protist symbionts of termites (Paper II)

3.2.1 GH7 enzymes in protist symbionts of lower termites

As mentioned previously, there is a remarkable conservation among GH7 enzymes. However, the recent discovery of a new type of GH7 EGs from termite symbiont protozoa demonstrates that there is likely still undiscovered diversity even within this long-studied family (Todaka et al., 2007). These protist sequences are shorter than the typical GH7 EGs, and consequently we have referred to them as short endoglucanases, sEGs. The first reports of GH7 enzymes in the termite hindgut protozoa emerged in the early 2000's, and the first GH7 sEGs from these organisms were characterized in detail in 2010, with the characterization of RsSymEG1, an sEG from an unknown symbiont of *Reticulitermes speratus* (Nakashima et al., 2002; Nakashima and Azuma, 2000; Todaka et al., 2010b; Watanabe et al., 2002). Subsequently a few studies have investigated these enzymes, but until now no structure has been published, and the interest they have garnered in academic literature has been surprisingly subdued (Sethi et al., 2013; Woon et al., 2017). In this work we have solved the molecular structure of RsSymEG1 through x-ray crystallography, revealing a structure architecture distinct from hitherto known GH7 structures. We also explored the diversity of GH7 enzymes in the lower termite protozoa, and how they might relate to GH7 enzymes from other organisms.

3.2.2 Structure characteristics of RsSymEG1, a short GH7 endoglucanase

The structure of RsSymEG1 was solved to 1.85 Å resolution, providing good separation of details. The structure showed a core fold typical to GH7s, consisting of a beta-sandwich with 6 beta strands forming the base of the substrate binding site (Figure 5). However, immediately it is clear that there are significant differences to previous GH7 structures, as is to be expected

based on the sequence. The substrate binding face of the enzyme is characterized by shortness of the loops which typically form a substrate binding tunnel in GH7 CBHs and a binding groove in EGs. This leads to a seemingly open structure of the substrate binding site, and a notably flat surface of this side of the enzyme. Looking at the structure in more detail, it is seen that in comparison to other GH7 enzymes this flatness is largely caused by the absence of the so called B1 (typically found in GH7 EGs and CBHs), and B2 loops (found in CBHs). Comparison to the EG TreCel7B illustrates the effect the missing regions have on the shape of the enzyme (Figure 12). Notably, the large volume missing in the region where the B1 loop would be situated, is also where the N-terminal PCA residue and its binding position is situated in GH7 enzymes. Given that the whole region is missing in the enzyme, it is perhaps not surprising that similar N-terminal cap is not seen in the RsSymEG1 structure. While this could make the enzyme less stable, and potentially explains low thermal stability of the enzyme compared to most GH7 cellulases, it also opens up possibilities for new kinds of multi-domain constructs. Meanwhile, another notable feature of the enzyme, and possibly another factor playing into the low heat resistance, is the presence of only four cysteine-bridges in the enzyme, low for a GH7 enzyme, which typically contain between eight and ten disulphide bonds.

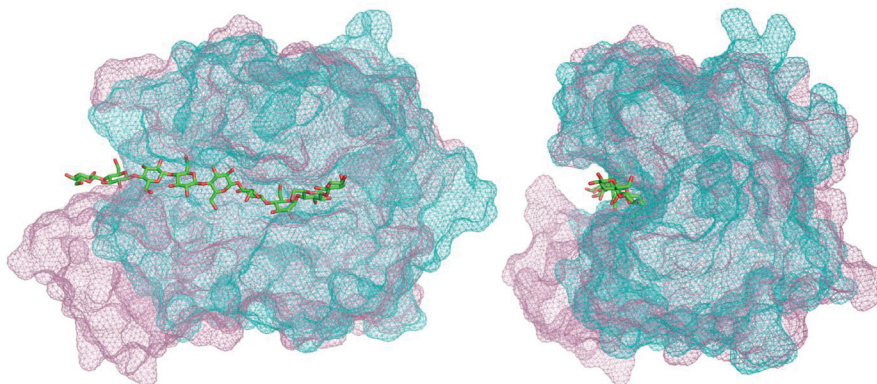


Figure 12. An illustration of the differences in overall structure between RsSymEG1 (cyan mesh) and TreCel7B (pink mesh, PDB structure 1EG1), a typical fungal GH7 EG. A cellononaose molecule (red and green) from PDB structure 4C4C superposed on TreCel7B to visualize substrate binding position on this enzyme.

For the most part the presumed substrate binding residues in RsSymEG1 conform to conserved residues seen in other GH7 EGs (Figure 13). There is an unusual feature at the early part of the binding groove however, where two tryptophan residues are seen at the A1-loop, with the side chains lining the substrate binding positions -6, -5, and -4. While overall the substrate binding site is highly exposed in this structure, these two residues make the early part of the substrate binding groove quite narrow. How this feature affects the substrate binding and activity of this enzyme was outside the scope of our study, but presents an interesting question for further structure-function investigations.

3.2.3 Repertoire of GH7 enzymes in lower termite symbiont protozoa

In this work we also had a look at the array of GH7s produced by the termite gut protists and found that in addition to sEG and CBH sequences which have been characterized previously, there are a few EG sequences conforming to typical GH7 EG features as well. These sequences originate from metagenomic studies of mixed protist samples, so their specific hosts are not known. Notably, they bear low sequence identity to other known sequences (<40%), and only three sequences from these termite-associated sources are found in the NCBI non-redundant database. Interestingly, none of the GH7 sequences found in these organisms contain CBMs. Structure models were constructed for examples of both CBH and EG sequences, and suggested these enzymes contain assortments of residues highly typical of the respective enzyme types.

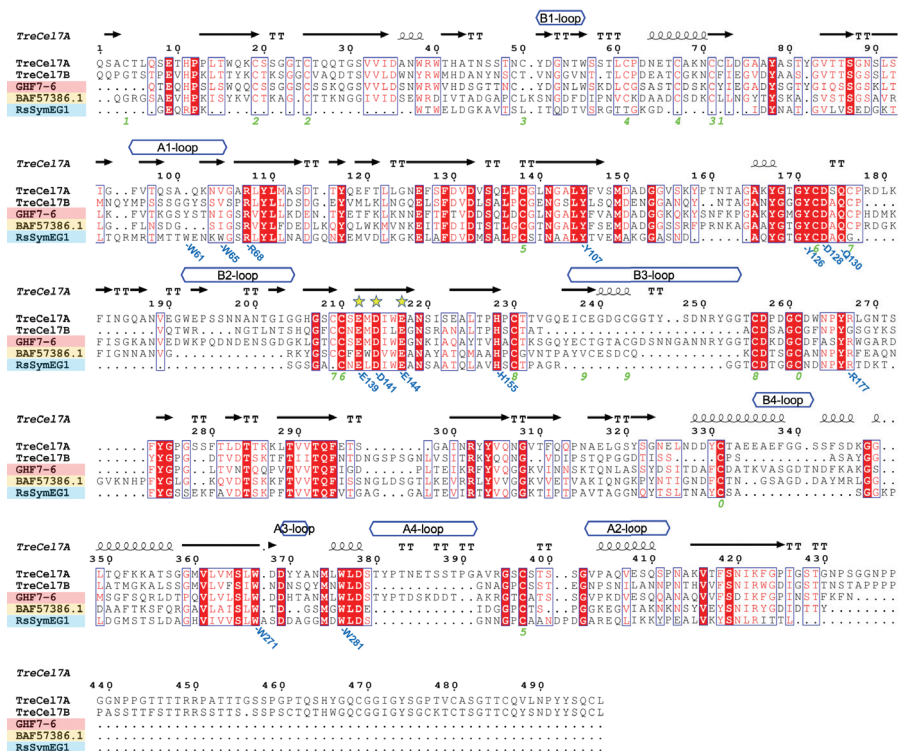


Figure 13. Sequence alignment with a fungal CBH TreCel7A and EG TreCel7B, and termite symbiont GH7 sequences. GHF7-6 and BAF57386.1 are CBH and EG sequences, respectively, from termite symbiont protozoa. RsSymEG1 represents the smaller protist sEG GH7s. The top three rows mark active site loops, secondary structure elements (helices, beta-strands as arrows, and TT are turns) and residue numbers in TreCel7A as reference. White characters on red show identical residues, red letters similar residues, and blue boxes outline conserved regions. Yellow stars mark the conserved catalytic residues, and green numbers the disulphide pairing of cysteines in TrCel7A. Residues of interest in RsSymEG1 are noted in blue at the bottom. The figure was created using ESPript.

3.2.4 sEGs in the context of GH7 evolution

Phylogenetic analysis of GH7 sequences across eukaryotes was conducted in order to shed light on the evolutionary relationships of the studied protist enzymes to their counterparts from other organisms. The results suggest that the three different groups of GH7s found in termite symbiont protists are not closely related to each other, but instead group closer to clades of enzymes of the same type from other organisms (Figure 14).

Since it is possible that the divergence between CBHs and EGs has occurred several times throughout eukaryote evolution, in our study gap regions were removed to prevent clustering of sequences solely based on the insertions/deletions and thus masking the true evolutionary relationships. Preceding the first biochemical characterization of sEG enzymes, these sequences would have easily been dismissed as incomplete sequences. In order to avoid artificially restricting diversity of sequences this way in our study, yet at the same time minimizing inclusion of sequencing or gene annotation artifacts or unfunctional genes, parameters were set for screening sequences which are likely not coding for active enzymes, setting minimum limits for sequence length (300 amino acids), and regions where deletions are not allowed as they would most likely lead to misfolded protein, e.g., the beta-sheets of the enzyme core. A maximum likelihood phylogenetic tree was constructed based on the screened sequence set and alignment. If divergence of the three enzyme types had occurred within the protozoa, it would be expected that these sequences cluster close to each other in a phylogenetic tree containing GH7 sequences all across eukaryotes. However, the results showed clustering of sequences of the three different types of GH7s from termite symbiont protist into distant clades, indicating that the divergence of these enzyme groups has not been a recent event during GH7 evolution. Instead the CBH, EG and sEG sequences from these termite symbionts display a closer relation to enzymes of the same type from other eukaryotes. In fact, to my knowledge GH7 sEG sequences from other organisms have not been described in literature previously, but our analysis highlighted a group of sequences with very similar features to the termite symbiont sEGs in species of *Daphnia* (water fleas). Another highly interesting observation from the analysis was the co-clustering of CBH and EG sequences from oomycetes with other CBHs, possibly indicating a loop deletion event or events within oomycetes, leading to emergence of a line of EGs separate of those seen in other organisms.

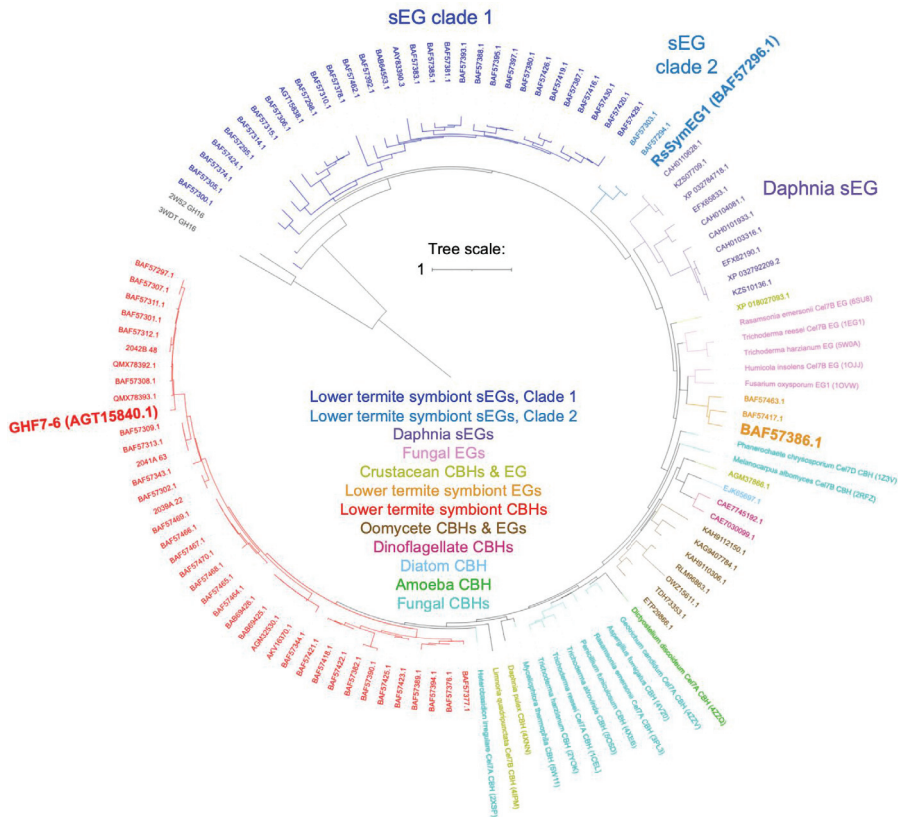


Figure 14. Maximum likelihood phylogenetic tree of GH7 sequences. For visualization, 116 sequences were selected from the original tree of 371 GH7 sequences. Clades containing termite symbiont sequences are included in their entirety. sEG sequences from termite symbionts divide into two clades. RsSymEG1 clusters into the more sparsely populated clade of the two, with only two other closely related sequences. Two sequences from the distantly related family GH16 (black) were included to root the tree.

3.2.5 Activity measurements

The activity of RsSymEG1 was previously analyzed on various substrates, including Avicel and carboxymethylcellulose (CMC), with the enzyme showing high activity on the soluble CMC, yet low activity on the insoluble Avicel (Todaka et al., 2010b). We set out to make a direct comparison of RsSymEG1 activity to another GH7 endoglucanase, TloCel7B. Since the enzyme had shown a clear preference to soluble non-crystalline substrates, we tested the RsSymEG1 on glucomannan, barley β -glucan, and CMC, assessing the formation of reducing ends by a p-

hydroxybenzoic acid hydrazide (PHBAH) assay. On all substrates RsSymEG1 displayed rapid hydrolysis in the early stages (up to 8 hours) of the conversion compared to TloCel7B. However, at later time points activity of RsSymEG1 plateaued faster than TloCel7B, and on glucomannan and CMC TloCel7B reached higher conversion rates by 48 hours, while on barley β -glucan both enzymes showed roughly equal release of reducing ends at the final 74-hour time point.

3.2.6 Conclusions and future perspectives

A recent study demonstrated prediction of cellulase activity in several enzyme families with high precision using computational protein models and MD simulations for assessing binding energy, from which catalytic turnover could be accurately estimated, thanks to so called linear free energy relationship (LFER; Kari et al., 2021). This kind of approach to enzyme discovery and development is likely to gain more ground in the future, as it has the potential to dramatically increase the speed of enzyme improvement by bypassing time consuming and sometimes problematic biochemical screening procedures in the initial discovery phase. Accurate modeling of enzyme structures however, is arguably still dependent on the availability of structural information from relevant template structures. As the first published structure of its type, the RsSymEG1 structure we have presented in **Paper II** is an important step in understanding the properties of this enzyme, but also in assessing other similar enzymes in this group. Our activity measurements confirmed previous reports of high turnover rates on soluble substrates (Todaka et al., 2010b), but also demonstrated lower degrees of conversion over longer time frames when compared to a typical GH7 EG. Screening against various carbohydrate substrates did not demonstrate RsSymEG1 substrate promiscuity, which could have perhaps been expected given the open structure of the substrate binding site. In fact, the substrate range for this enzyme was narrower than for the reference enzyme TloCel7B. While the structure does not offer an immediate rationale for the activity profile, it is an important part of understanding the structure-function relationships in this sparsely characterized group of enzymes. It also demonstrates the potential of these enzymes for novel constructs, as the missing N-terminal PCA cap opens up possibilities for novel multi-domain modular enzyme designs.

3.3 Production of TreCel7A mutants in *Trichoderma reesei* for single-molecule imaging of processive enzyme action and FRET analysis of interdomain protein dynamics

3.3.1 Single molecule imaging of processive GH7 CBHs

The processive nature of GH7 CBHs is a key factor in their ability to hydrolyze cellulose with high efficiency. This mode of action has also made these enzymes a highly interesting target for direct imaging through various microscopy techniques, where visualizing the enzyme movement on substrate has provided important insights into the reaction mechanism, velocity, and bottlenecks in cellulose hydrolysis. Atomic force microscopy (AFM) and different fluorescence microscopy techniques have been successfully used for imaging moving CBH enzymes on cellulose surface (Payne et al., 2015). A seminal study by Igarashi *et al.*, showed the occurrence of “traffic jams”, immobilization of TreCel7A in clusters when several enzymes were working on the same cellulose microfibril (Igarashi et al., 2011). This and subsequent studies have studied the effect of specific residues on enzyme processivity (Igarashi et al., 2011; Nakamura et al., 2021, 2013). Despite improvements in spatial and temporal resolution, studies published to date have left room for improvement in both. In **Paper III** we set out to study the model enzyme TreCel7A with two methods which have not been previously utilized for studying this enzyme class, total internal reflection dark-field microscopy (TIRDFM), and fluorescence resonance energy transfer (FRET). TIRDFM has previously been used for studying a processive chitinase from *Serratia marcescens* (Nakamura et al., 2018). This method offers good resolution for imaging gold nanoparticle (AuNP) labelled enzymes, and can already with current technology be used to achieve frame rates of 1000 frames per second. Relating to reported turnover rates for TreCel7A on cellulose in the range of 3-10 s⁻¹, this provides temporal resolution down to the level of individual processive steps. FRET has been previously successfully utilized for detecting inter- and intra-protein distances, exploiting the overlap between the emission and absorbance ranges of so-called donor and acceptor fluorophores, for gauging their proximity. In **Paper III** we provide a first look into using these methods for investigating GH7 CBH activity, movement and inter-domain dynamics, generating variants of TreCel7A for a proof-of-concept study.

3.3.2 Design and production of TreCel7A mutants for imaging studies

For labeling of TreCel7A for imaging using FRET and TIRDFM, mutants of this enzyme were designed to introduce free (*i.e.*, not involved in forming disulfide bridges) cysteine residues on the surface of the enzyme, with the goal of utilizing them for maleimide mediated labeling. Three mutants, S21C-V233C, S21C-T350C and T332C-I426C, with two introduced free cysteines on the CD were designed for the AuNP-labeling for TIRDFM imaging, in order to achieve binding of two biotin molecules on each enzyme molecule, and thus strong binding of streptavidin (which contains 4 binding sites for biotin) which in turn is bound to biotinylated AuNP via the other two biotin binding sites, thus linking the enzyme and AuNP. While the primary purpose of the mutations was not to study the effects of these mutations on enzyme activity, but to enable labeling the TreCel7A CD with one AuNP per enzyme molecule, three mutants were prepared to allow screening of several constructs in case the mutations or the attached assembly hindered enzyme activity.

Further three mutants, S128C-V478C, S156C-V478C and T231C-V478C, were designed for FRET studies, with each mutant containing one free cysteine located on the CBM and one on the CD. This was to enable labeling the CD and CBM with two different fluorophores, and thus utilization of FRET for gauging distances between the two domains.

The native host of the native enzyme, *T. reesei*, was selected for use as an expression host. The pTrEno plasmid which contains the TreCel7A gene was used as a template for producing all of the mutants. A combination of mutagenesis PCR and synthesized DNA fragment ligation were used to create the mutated TreCel7A constructs, the linearized plasmids transformed into the Ast1116 strain lacking the TreCel7A gene, with the transformants screened for highest expressing strains by pNPL activity assays (for TIRDFM mutants) or TreCel7A specific antibodies (FRET mutants). Larger scale cultivations of the three TIRDFM mutants were done in shake flasks, and purification from the 1-liter culture filtrates yielded between 3-8 mg of purified enzyme. For each of the three FRET mutants cultivations were conducted in 8 liter scale, and two of the mutants were further subjected to purification and FRET imaging trials, yielding 9 mg and 20 mg of purified enzyme for the S128C-V478C and T231C-V478C mutants respectively.

3.3.3 Insights from ultra-high speed single molecule imaging and FRET microscopy

Finding and analyzing moving enzymes from TIRDFM proved to be challenging and extremely time consuming. In our initial experiments for the proof-of-concept study in **Paper III**, we were able to find molecules with a movement pattern corresponding to CBH action for examples of only one of the three mutants. This highlights the need to screen for suitable mutants, due to potential problems caused by the AuNP attachment. We demonstrated the analysis of a movement pattern of one AuNP-labelled TreCel7A molecule, analyzing the stepwise progression on cellulose surface (Figure 15). The analysis illustrates the ability of this method to distinguish individual steps of the enzyme, but also the high noise levels in the imaging, thus highlighting the need for a large number of observations to reliably quantify movement characteristics.

The two mutants selected for FRET studies were labeled with Alexa555 and Alexa647 fluorophores, and imaged on a fluorescence microscope equipped with filters suitable for observing each of the fluorophores. For one of the mutants, T231C-V478C, the results showed consistently fluorescence patterns where fluorescence was initially observed only from the acceptor fluorophore Alexa647, followed by a period of fluorescence only from the donor, Alexa555 (Figure 15). This behavior is in line with 100 % FRET efficiency (*i.e.*, close contact between the fluorophores) followed by photobleaching. The second mutant, S128C-V478C showed a more varied response, with various instances of simultaneous fluorescence from both Alexa555 and Alexa647, indicating intermediate distances between the two labels, *i.e.*, close enough for FRET to occur, but not so close to lead to 100 % FRET efficiency. The distributions of FRET efficiencies from analyzed molecules for the two mutants are seen in Figure 15. Taken together these results imply that the CBM tends to locate closer to the amino acid position 231 than residue 128 in the CD of TreCel7A (Thr and Ser residues respectively in the native enzyme). More data points are necessary to reliably validate our initial findings, but these initial results already demonstrate success in producing and labeling TreCel7A variants for FRET imaging.

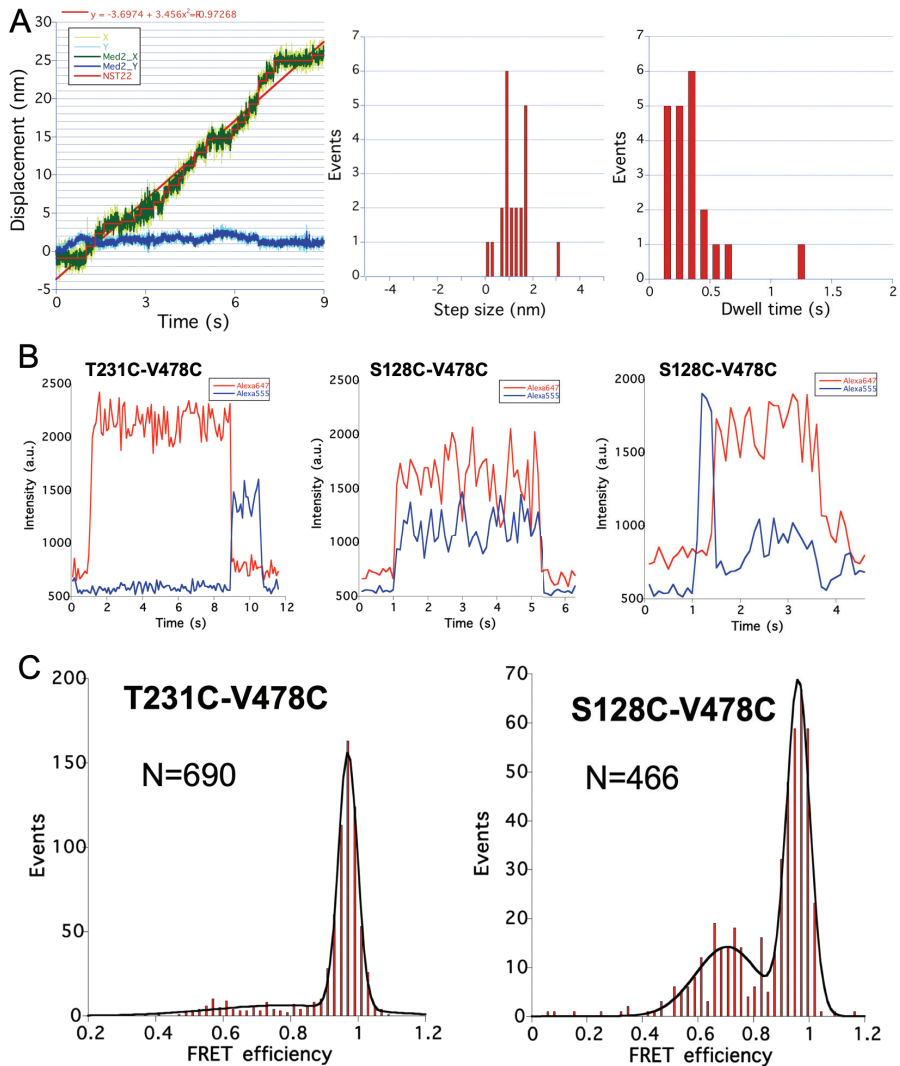


Figure 15. (A) A trajectory of AuNP labelled TreCel7A on cellulose surface with the step sizes and dwell times between each step determined by a step-finding algorithm. (B) Fluorescence intensities corresponding to the fluorophores Alexa555 (donor fluorophore; blue) and Alexa647 (acceptor fluorophore; red) in observations of individual molecules of TreCel7A mutants T231C-V478C and S128C-V478C. (C) Distributions of FRET efficiencies for the two mutants in analyzed images where Alexa647 fluorescence was detected.

3.3.4 Conclusions and future perspectives

We successfully constructed three mutants of TreCel7A, each containing two free cysteine residues on the CD for AuNP labeling for single molecule imaging. Our results from TIRDFM imaging demonstrated ultra-high speed imaging of the movement of one of these mutants on cellulose surface, and provides a starting point for further experiments with this and other variants of this enzyme. This should allow further improving our understanding of the factors behind the processive action of GH7 CBHs.

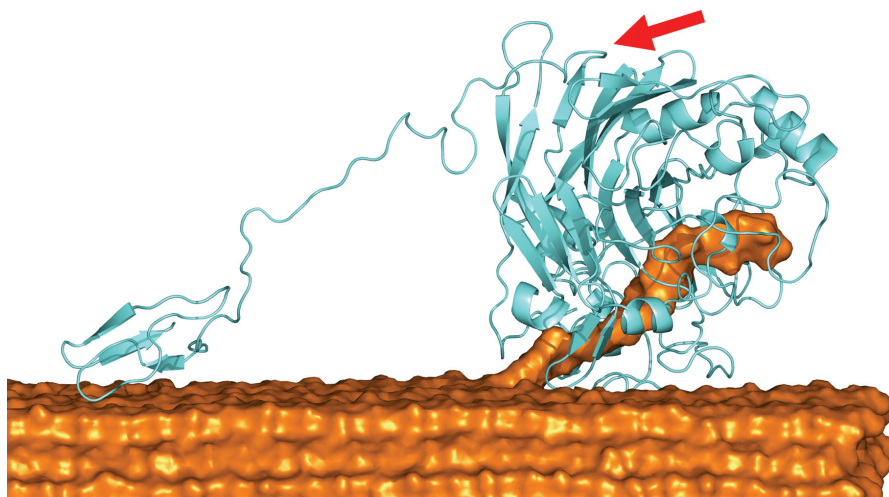


Figure 16. Illustration of TreCel7A on cellulose surface, with the attachment point of the flexible linker to the CD marked with an arrow.

While illustrations of CBM containing GH7 enzymes are often seen in publications, it should be noted that these are merely models of the form the complete enzyme is presumed to take on cellulose surface. Structures of the CD and CBM have been solved separately, but no structural information exists for a complete enzyme including both, and it is not known exactly how the two domains are positioned in relation to each other during enzyme action. Generally, the CBM is depicted to locate “in front of” the CD, *i.e.*, in the direction which a CBH is moving towards during processive action. Given that the flexible linker region is attached to the CD at the opposite side of the enzyme from the substrate binding site, it has roughly equal distance to the substrate surface from all sides of the CD when bound on cellulose surface, and thus could in theory be positioned on any side (Figure 16). Whether during processive CBH action the CBM is “pushed ahead”,

“dragged behind”, or “led beside”, or perhaps all of the above, is a question the FRET imaging approach we have taken should be suitable to elucidate. Our results show successful production, labeling and fluorescence imaging of two mutants containing two free cysteines on the CD and CBM of TreCel7A, and the utilization of FRET for assessing the inter-domain distances in this enzyme. Already these results provide clues to the possible positioning of the CBM during processive action. After this successful initial demonstration, obtaining more imaging data on these and further constructs will provide important insights into the CD-CBM dynamics of multi-domain carbohydrate hydrolases.

4. Future perspectives & conclusions

Despite the need to find less carbon intensive transportation fuels, volumes of global ethanol fuel production have been somewhat stagnant for the last decade (Renewable Fuels Association, 2022). My personal perception is also that in the public discourse and media, bioethanol-based fuels have not gotten much attention in recent years, and funding for academic research projects related to cellulosic ethanol and cellulase enzymes has decreased as well. I think there are several reasons for these developments, some of them related. First of all, a vast majority of the current supply is first generation ethanol (Hoang and Nghiem, 2021), and therefore most of the ethanol production today does not have as good climate credentials as cellulosic ethanol, and could potentially have a negative effect on food availability. Even if these things are much less of a concern with cellulosic ethanol, a fuel that it is not zero emission, releases carbon dioxide during end usage, is used in conventional ICE vehicles, and will most likely not be enough to cover all transport fuel needs without the risk of deforestation or endangering biodiversity, is unfortunately not likely to garner much excitement in the general public. Perhaps most critically, many of the existing cellulosic ethanol plants have struggled to stay economic due to low gasoline (and thus ethanol) prices (Rosales-Calderon and Arantes, 2019), explaining why lignocellulose based ethanol has not gained more market share. Furthermore, the image of ethanol as a transition phase fuel, with high uncertainty concerning the length of this transition period, and the related highly unpredictable regulatory framework, has likely hampered investment in a capital intensive, low-margin industry. In a related anecdote, personal communication with people within or familiar with the enzyme industry for biofuels suggests at least parts of the industry consider the enzyme efficiency problem “solved”. I wonder however, whether this is

more a reflection of low investment activity into cellulosic ethanol projects than lack of potential for improvement of enzyme cocktails, which still are a major cost hurdle in the process (Aui et al., 2021).

As such this is not a new development. Commercial and academic interest in cellulosic ethanol, and consequently to some extent the GH7 enzymes, has historically been somewhat correlated to the price of crude oil, as the economics of cellulosic ethanol production have been highly dependent on the prices of gasoline which it can be used to substitute (Aui et al., 2021; Brethauer and Wyman, 2010; Carpita and McCann, 2020). While oil prices are in no way the focus of this thesis, it is worthwhile to consider the backdrop and broader context of cellulase research, not only for the past but also the future decades. Since 2014 oil prices have been low compared to historical prices (especially in inflation adjusted terms), in large part due to the emergence of large-scale extraction of shale entrapped oil, particularly in The United States. This period of low prices culminated during the Covid-19 pandemic, following wide encompassing travel restrictions and lower economic activity, with West Texas Intermediate crude oil contracts momentarily reaching negative prices in April 2020. Since then oil prices have risen significantly, and while global investments to fossil fuel extraction have decreased since 2019 (International Energy Agency, 2022b), at the same time, with the current pace of policy changes, oil demand and consumption is still projected to keep rising beyond year 2030 (International Energy Agency, 2021). In short, this discrepancy would imply most likely rising prices. So, what does this have to do with GH7 cellulases, or the future of cellulosic ethanol? In order to respond to the demands set by the need to curtail GHG emissions, and the energy demands of a growing global population, massive investments in renewable and low carbon energy forms are required (International Energy Agency, 2022b). Electric vehicles have garnered a lot of attention, often seen as the future mode of transport. However, there are still many open questions regarding this transition, including the sustainability, availability, sufficiency and cost of supply of raw materials for batteries, the energy density for long distance transports, investments needed for energy grids, and availability of consistent electricity supply, which still today is highly dependent on fossil resources in most of the world. Meanwhile, the transition period is likely to take decades, due to the aforementioned issues, as well as the slow turnover of vehicle fleets (Carpita and McCann, 2020; International Energy Agency, 2022b), while

meeting current climate goals necessitates a rapid decline in GHG emissions (Monschauer et al., 2019). Therefore, the need for low carbon fuels which are compatible with the existing infrastructure remains critical, and there are not many, if any feasible alternatives to ethanol when it comes to replacing gasoline in the short term.

While public policy and therefore the incentives for energy investments are extremely difficult to predict, the economic feasibility of second-generation ethanol production is likely to improve significantly if fossil fuel prices continue to rise due to the dynamics described above. Still, significant room remains for efficiency improvements throughout the process, including biomass composition, pre-treatment methodologies, and not least in the enzyme mixtures used. In these enzyme cocktails, GH7 enzymes usually present the most abundant components, making their improvement a key factor in improving the efficiency of these mixtures. While these enzymes have already been studied for decades, I would argue that there is still a lot to explore, evidenced by recent discoveries of highly effective natural and synthetic variants, and perhaps even by the manuscripts included in this thesis.

Arguably one of the major hinders for research on these enzymes is the difficulty to successfully express them in readily available or industrially relevant expression hosts. In the spirit of the old adage about a tree falling in the forest, one might ask: If an enzyme is not readily expressed in *Trichoderma* or *Aspergillus*, does it exist? There is a plethora of highly interesting sequences to be screened, some of them discussed in the above sections, but until they can be obtained in sufficient quantities for characterization, their existence remains somewhat theoretical in the eyes of the academic community, as well as the enzyme industry. While the emergence of computational tools such as AlphaFold and molecular dynamics simulations spur magnificent advances in protein research, they are most often still highly limited in their nature, as I got to discover while working on **Papers I and II**, and caution should be applied when interpreting or applying the results from such methods. Biochemical data feeding these tools and validating their results remains a limiting factor (Drula et al., 2022), and as I also got to experience during my work, the expression of enzymes is perhaps the most critical bottleneck. Thus, I believe developing alternative expression hosts with the necessary genetic toolkits, is a high priority for future enzyme discovery. This problem is of course in no way unique to

cellulase research, and therefore also presents major opportunities to anyone successfully developing widely applicable tools. While I'm wary of uncritical use of the continuously improving computational tools, I do think they hold great potential for generating breakthroughs in cellulase research. Combining these methods with the improving affordability and thus wider availability of robotic systems for high-throughput methods also outside industry, should provide a powerful platform for not only producing effective enzymes for industry, but also improving the fundamental understanding of these molecular machines.

From a more personal perspective, many of the studies and experiments I'm most looking forward to see I have described in the included manuscripts. Building on **Paper III**, I'm very eager to see more FRET imaging conducted on several constructs of TreCel7A, as I believe this can radically improve our understanding of the functioning of these types of enzymes in combination with the associated CBMs. Research on sEGs described in **Paper II** has only scratched the surface into the attributes of this little-known type of GH7s. I would much like to see more detailed studies on the activities of these enzymes, how they function in combination with other cellulases, if they can be incorporated into effective multi-domain enzymes of various forms, and if they can be engineered for higher thermal stability while maintaining their high activities. While GH7 cellulases have been extensively studied, I would posit that there is possibility for many further discoveries, especially looking for enzymes outside the fungal sphere. **Paper II** acts as a demonstration of this point. I have also seen many of the most intriguing candidates come from oomycetes, where multiplications and mutations have produced many sequences with highly interesting characteristics. Where the future of cellulase research takes us remains to be seen, but I remain curious and hopeful.

References

- Acuña, R., Padilla, B.E., Flórez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., Lee, S.J., Yeats, T.H., Egan, A.N., Doyle, J.J., Rose, J.K.C., 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proc. Natl. Acad. Sci. U. S. A.* 109, 4197–4202.
- Agbor, V.B., Cicek, N., Sparling, R., Berlin, A., Levin, D.B., 2011. Biomass pretreatment: Fundamentals toward application. *Biotechnol. Adv.* 29, 675–685.
- Alfaro, M., Oguiza, J.A., Ramírez, L., Pisabarro, A.G., 2014. Comparative analysis of secretomes in basidiomycete fungi. *J. Proteomics* 102, 28–43.
- Amore, A., Knott, B.C., Supekar, N.T., Shajahan, A., Azadi, P., Zhao, P., Wells, L., Linger, J.G., Hobdey, S.E., Vander Wall, T.A., Shollenberger, T., Yarbrough, J.M., Tan, Z., Crowley, M.F., Himmel, M.E., Decker, S.R., Beckham, G.T., Taylor, L.E., 2017. Distinct roles of N- and O-glycans in cellulase activity and stability. *Proc. Natl. Acad. Sci. U. S. A.* 114, 13667–13672.
- Anderson, J.P., Gleason, C.A., Foley, R.C., Thrall, P.H., Burdon, J.B., Singh, K.B., 2010. Plants versus pathogens: An evolutionary arms race. *Funct. Plant Biol.* 37, 499–512.
- Arantes, V., Goodell, B., 2014. Current understanding of brown-rot fungal biodegradation mechanisms: A review. *ACS Symp. Ser.* 1158, 3–21.
- Aui, A., Wang, Y., Mba-Wright, M., 2021. Evaluating the economic feasibility of cellulosic ethanol: A meta-analysis of techno-economic analysis studies. *Renew. Sustain. Energy Rev.* 145.
- Ayers, A.R., Ayers, S.B., Eriksson, K., 1978. Cellobiose Oxidase, Purification and Partial Characterization of a Hemoprotein from *Sporotrichum pulverulentum*. *Eur. J. Biochem.* 90, 171–181.
- Ayuso-Fernández, I., De Lacey, A.L., Cañada, F.J., Ruiz-Dueñas, F.J., Martínez, A.T., 2019. Increase of Redox Potential during the Evolution of Enzymes Degrading Recalcitrant Lignin. *Chem. - A Eur. J.* 25, 2708–2712.
- Ayuso-Fernández, I., Ruiz-Dueñas, F.J., Martínez, A.T., 2018. Evolutionary convergence in lignin-degrading enzymes. *Proc. Natl. Acad. Sci. U. S. A.* 115, 6428–6433.
- Banerjee, G., Car, S., Scott-Craig, J.S., Borrusch, M.S., Aslam, N., Walton, J.D., 2010a. Synthetic enzyme mixtures for biomass deconstruction: Production and optimization of a core set. *Biotechnol. Bioeng.* 106, 707–720.
- Banerjee, G., Car, S., Scott-Craig, J.S., Borrusch, M.S., Walton, J.D., 2010b. Rapid optimization of enzyme mixtures for deconstruction of diverse pretreatment/biomass feedstock combinations. *Biotechnol. Biofuels* 3, 22.
- Bao, W., Usha, S.N., Renganathan, V., 1993. Purification and Characterization of

- Cellobiose Dehydrogenase, a Novel Extracellular Hemoflavoenzyme from the White-Rot Fungus *Phanerochaete chrysosporium*. *Arch. Biochem. Biophys.*
- Beeson, W.T., Vu, V. V., Span, E.A., Phillips, C.M., Marletta, M.A., 2015. Cellulose Degradation by Polysaccharide Monooxygenases. *Annu. Rev. Biochem.* 84, 923–46.
- Bhat, M.K., Bhat, S., 1997. Cellulose degrading enzymes and their potential industrial applications. *Biotechnol. Adv.* 15, 583–620.
- BIELY, P., KRÁTKÝ, Z., VRŠANSKÁ, M., 1981. Substrate-Binding Site of Endo-1,4- β -Xylanase of the Yeast *Cryptococcus albidus*. *Eur. J. Biochem.* 119, 559–564.
- Bischof, R.H., Ramoni, J., Seiboth, B., 2016. Cellulases and beyond: The first 70 years of the enzyme producer *Trichoderma reesei*. *Microb. Cell Fact.* 15, 1–13.
- Bissaro, B., Røhr, Å.K., Müller, G., Chylenski, P., Skaugen, M., Horn, S.J., Vaaje-kolstad, G., Eijsink, V.G.H., 2017. Oxidative cleavage of polysaccharides by monocopper enzymes depends on H₂O₂.
- Bödeker, I.T.M., Lindahl, B.D., Olson, Å., Clemmensen, K.E., 2016. Mycorrhizal and saprotrophic fungal guilds compete for the same organic substrates but affect decomposition differently. *Funct. Ecol.* 30, 1967–1978.
- Borrion, A.L., McManus, M.C., Hammond, G.P., 2012. Environmental life cycle assessment of lignocellulosic conversion to ethanol: A review. *Renew. Sustain. Energy Rev.* 16, 4638–4650.
- Brethauer, S., Wyman, C.E., 2010. Review: Continuous hydrolysis and fermentation for cellulosic ethanol production. *Bioresour. Technol.* 101, 4862–4874.
- Brunecky, R., Subramanian, V., Yarbrough, J.M., Donohoe, B.S., Vinzant, T.B., Vanderwall, T.A., Knott, B.C., Chaudhari, Y.B., Bomble, Y.J., Himmel, M.E., Decker, S.R., 2020. Synthetic fungal multifunctional cellulases for enhanced biomass conversion. *Green Chem.* 22, 478–489.
- Burki, F., 2014. The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* 6.
- Carpita, N.C., McCann, M.C., 2020. Redesigning plant cell walls for the biomass-based bioeconomy. *J. Biol. Chem.* 295, 15144–15157.
- Champreda, V., Mhuantong, W., Lekakarn, H., Bunterngsook, B., Kanokratana, P., Zhao, X.Q., Zhang, F., Inoue, H., Fujii, T., Eurwilaichitr, L., 2019. Designing cellulolytic enzyme systems for biorefinery: From nature to application. *J. Biosci. Bioeng.* 128, 637–654.
- Cleveland, L.R., 1923. Symbiosis between Termites and Their Intestinal Protozoa. *Proc. Natl. Acad. Sci.* 9, 424–428.
- Cleveland, L.R., 1924. THE PHYSIOLOGICAL AND SYMBIOTIC RELATIONSHIPS BETWEEN THE INTESTINAL PROTOZOA OF TERMITES AND THEIR HOST, WITH SPECIAL REFERENCE TO *RETICULITERMES FLAVIPES KOLLAR*. *Biol. Bull.* 46, 178–201.
- Cosgrove, D.J., 2005. Growth of the plant cell wall. *Nat. Rev. Mol. Cell Biol.* 6, 850–861.

- Courtade, G., Wimmer, R., Røhr, Å.K., Preims, M., Felice, A.K.G., Dimarogona, M., Vaaje-Kolstad, G., Sørlie, M., Sandgren, M., Ludwig, R., Eijsink, V.G.H., Aachmann, F.L., 2016. Interactions of a fungal lytic polysaccharide monooxygenase with β -glucan substrates and cellobiose dehydrogenase. *Proc. Natl. Acad. Sci. U. S. A.* 113, 5922–5927.
- Dana, C.M., Dotson-Fagerstrom, A., Roche, C.M., Kal, S.M., Chokhawala, H.A., Blanch, H.W., Clark, D.S., 2014. The importance of pyroglutamate in cellulase Cel7A. *Biotechnol. Bioeng.* 111, 842–847.
- Davies, G.J., Wilson, K.S., Henrissat, B., 1997. Nomenclature for sugar-binding subsites in glycosyl hydrolases [1]. *Biochem. J.*
- Demoor, A., Silar, P., Brun, S., 2019. Appressorium: The breakthrough in Dikarya. *J. Fungi* 5.
- Divne, C., Staahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J.K.C., Teeri, T.T., Jones, T.A., 1994. The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from *Trichoderma reesei*. *Sci. (Washington, D. C.)* 265, 524–528.
- Divne, C., Ståhlberg, J., Teeri, T.T., Jones, T.A., 1998. High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J. Mol. Biol.* 275, 309–325.
- Drula, E., Garron, M.L., Dogan, S., Lombard, V., Henrissat, B., Terrapon, N., 2022. The carbohydrate-active enzyme database: Functions and literature. *Nucleic Acids Res.* 50, D571–D577.
- Duarte, S., Nunes, L., Borges, P.A.V., Nobre, T., 2018. A bridge too far? An integrative framework linking classical protist taxonomy and metabarcoding in lower termites. *Front. Microbiol.* 9, 1–5.
- Ejaz, U., Sohail, M., Ghanemi, A., 2021. Cellulases: From bioactivity to a variety of industrial applications. *Biomimetics* 6.
- Eklöf, J.M., Brumer, H., 2010. The XTH gene family: An update on enzyme structure, function, and phylogeny in xyloglucan remodeling. *Plant Physiol.* 153, 456–466.
- Fasim, A., More, V.S., More, S.S., 2021. Large-scale production of enzymes for biotechnology uses. *Curr. Opin. Biotechnol.* 69, 68–76.
- Floudas, D., Binder, M., Riley, R., Barry, K., Blanchette, R.A., Henrissat, B., Martínez, A.T., Otillar, R., Spatafora, J.W., Yadav, J.S., Aerts, A., Benoit, I., Boyd, A., Carlson, A., Copeland, A., Coutinho, P.M., de Vries, R.P., Ferreira, P., Findley, K., Foster, B., Gaskell, J., Glotzer, D., Górecki, P., Heitman, J., Hesse, C., Hori, C., Igarashi, K., Jurgens, J.A., Kallen, N., Kersten, P., Kohler, A., Kües, U., Kumar, T.K.A., Kuo, A., LaButti, K., Larrondo, L.F., Lindquist, E., Ling, A., Lombard, V., Lucas, S., Lundell, T., Martin, R., McLaughlin, D.J., Morgenstern, I., Morin, E., Murat, C., Nagy, L.G., Nolan, M., Ohm, R.A., Patyshakuliyeva, A., Rokas, A., Ruiz-Dueñas, F.J., Sabat, G., Salamov, A., Samejima, M., Schmutz, J., Slot, J.C., St John, F., Stenlid, J., Sun, H., Sun, S., Syed, K., Tsang, A., Wiebenga, A., Young, D., Pisabarro, A., Eastwood, D.C., Martin, F., Cullen, D., Grigoriev, I. V., Hibbett, D.S., 2012. The Paleozoic

- origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336, 1715–9.
- Forsberg, Z., Vaaje-kolstad, G., Westereng, B., Bunsæ, A.C., Stenstrøm, Y., Mackenzie, A., Sørlie, M., Horn, S.J., Eijsink, V.G.H., 2011. Cleavage of cellulose by a cbm33 protein. *Protein Sci.* 20, 1479–1483.
- Freeze, H., Loomis, W.F., 1977. Isolation and characterization of a component of the surface sheath of *Dictyostelium discoideum*. *J. Biol. Chem.* 252, 820–824.
- Freeze, H., Loomis, W.F., 1978. Chemical analysis of stalk components of *Dictyostelium discoideum*. *BBA - Gen. Subj.* 539, 529–537.
- Gado, J.E., Harrison, B.E., Sandgren, M., Ståhlberg, J., Beckham, G.T., Payne, C.M., 2021. Machine learning reveals sequence-function relationships in family 7 glycoside hydrolases. *J. Biol. Chem.* 297, 100931.
- Galbe, M., Wallberg, O., 2019. Pretreatment for biorefineries: A review of common methods for efficient utilisation of lignocellulosic materials. *Biotechnol. Biofuels* 12, 1–26.
- Gardner, K.H., Blackwell, J., 1974a. The structure of native cellulose. *Biopolymers* 13, 1975–2001.
- Gardner, K.H., Blackwell, J., 1974b. The hydrogen bonding in native cellulose. *BBA - Gen. Subj.* 343, 232–237.
- Gaulin, E., Pel, M.J.C., Camborde, L., San-Clemente, H., Courbier, S., Dupouy, M.A., Lengellé, J., Veysiere, M., Le Ru, A., Grandjean, F., Cordaux, R., Moumen, B., Gilbert, C., Cano, L.M., Aury, J.M., Guy, J., Wincker, P., Bouchez, O., Klopp, C., Dumas, B., 2018. Genomics analysis of *Aphanomyces* spp. identifies a new class of oomycete effector associated with host adaptation. *BMC Biol.* 16, 1–21.
- Geng, A., Cheng, Y., Wang, Y., Zhu, D., Le, Y., Wu, J., Xie, R., Yuan, J.S., Sun, J., 2018. Transcriptome analysis of the digestive system of a wood-feeding termite (*Coptotermes formosanus*) revealed a unique mechanism for effective biomass degradation. *Biotechnol. Biofuels* 11, 1–14.
- Ghelardini, L., Pepori, A.L., Luchi, N., Capretti, P., Santini, A., 2016. Drivers of emerging fungal diseases of forest trees. *For. Ecol. Manage.* 381, 235–246.
- Gírio, F.M., Fonseca, C., Carvalheiro, F., Duarte, L.C., Marques, S., Bogel-Lukasik, R., 2010. Hemicelluloses for fuel ethanol: A review. *Bioresour. Technol.* 101, 4775–4800.
- Gritzali, M., Brown, R.D., 1979. The Cellulase System of *Trichoderma*. pp. 237–260.
- Gruno, M., Våljamäe, P., Pettersson, G., Johansson, G., 2004. Inhibition of the *Trichoderma reesei* cellulases by cellobiose is strongly dependent on the nature of the substrate. *Biotechnol. Bioeng.* 86, 503–511.
- Haddad Momeni, M., Payne, C.M., Hansson, H., Mikkelsen, N.E., Svedberg, J., Engström, A., Sandgren, M., Beckham, G.T., Ståhlberg, J., 2013. Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidion irregulare*. *J. Biol. Chem.* 288, 5861–5872.

- Harris, P. V., Welner, D., McFarland, K.C., Re, E., Navarro Poulsen, J.-C., Brown, K., Salbo, R., Ding, H., Vlasenko, E., Merino, S., Xu, F., Cherry, J., Larsen, S., Lo Leggio, L., 2010. Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family. *Biochemistry* 49, 3305–3316.
- Hatakka, A., Hammel, K.E., 2010. Fungal Biodegradation of Lignocelluloses. In: *Industrial Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 319–340.
- Hedison, T.M., Breslmayr, E., Shanmugam, M., Karnpakdee, K., Heyes, D.J., Green, A.P., Ludwig, R., Scrutton, N.S., Kracher, D., 2021. Insights into the H₂O₂-driven catalytic mechanism of fungal lytic polysaccharide monooxygenases. *FEBS J.* 288, 4115–4128.
- Henriksson, G., Petterssen, G., Johansson, G., 2000. A critical review of cellobiose dehydrogenase. *J. Biotechnol.* 79, 93.
- Hibbett, D., Blanchette, R., Kenrick, P., Mills, B., 2016. Climate, decay, and the death of the coal forests. *Curr. Biol.* 26, R563–R567.
- Himmel, M.E., Ding, S.-Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W., Foust, T.D., 2007. Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 315, 804–7.
- Hoang, T.D., Nghiem, N., 2021. Recent developments and current status of commercial production of fuel ethanol. *Fermentation* 7.
- Hobdey, S.E., Knott, B.C., Momeni, M.H., Taylor, L.E., Borisova, A.S., Podkaminer, K.K., VanderWall, T.A., Himmel, M.E., Decker, S.R., Beckham, G.T., Ståhlberg, J., 2016. Biochemical and structural characterizations of two *Dictyostelium cellobiohydrolases* from the Amoebozoa kingdom reveal a high level of conservation between distant phylogenetic trees of life. *Appl. Environ. Microbiol.* 82, 3395–3409.
- Hu, J., Chandra, R., Arantes, V., Gourelay, K., Van Dyk, J.S., Saddler, J.N., 2015. The addition of accessory enzymes enhances the hydrolytic performance of cellulase enzymes at high solid loadings. *Bioresour. Technol.* 186, 149–153.
- Igarashi, K., Uchihashi, T., Koivula, A., Wada, M., Kimura, S., Okamoto, T., Penttilä, M., Ando, T., Samejima, M., 2011. Traffic jams reduce hydrolytic efficiency of cellulase on cellulose surface. *Science (80-)*. 333, 1279–1282.
- International Energy Agency, 2021. *World Energy Outlook 2021*.
- International Energy Agency, 2022a. *World Energy Outlook 2022*.
- International Energy Agency, 2022b. *World Energy Investment 2022*.
- Jones, S.M., Transue, W.J., Meier, K.K., Kelemen, B., Solomon, E.I., 2020. Kinetic analysis of amino acid radicals formed in H₂O₂-driven CuI LPMO reoxidation implicates dominant homolytic reactivity. *Proc. Natl. Acad. Sci. U. S. A.* 117, 2–8.
- Judelson, H.S., 2017. Metabolic Diversity and Novelties in the Oomycetes. *Annu. Rev. Microbiol.* 71, 21–39.
- Kallioinen, A., Puranen, T., Siika-aho, M., 2014. Mixtures of thermostable enzymes show high performance in biomass saccharification. *Appl. Biochem.*

- Biotechnol. 173, 1038–56.
- Kari, J., Kont, R., Borch, K., Buskov, S., Olsen, J.P., Cruyz-Bagger, N., Våljamäe, P., Westh, P., 2017. Anomeric selectivity and product profile of a processive cellulase. *Biochemistry* 56, 167–178.
- Kari, J., Molina, G.A., Schaller, K.S., Schiano-di-Cola, C., Christensen, S.J., Badino, S.F., Sørensen, T.H., Røjel, N.S., Keller, M.B., Sørensen, N.R., Kolaczowski, B., Olsen, J.P., Krogh, K.B.R.M., Jensen, K., Cavaleiro, A.M., Peters, G.H.J., Spodsberg, N., Borch, K., Westh, P., 2021. Physical constraints and functional plasticity of cellulases. *Nat. Commun.* 12, 1–10.
- Karp, S.G., Medina, J.D.C., Letti, L.A.J., Woiciechowski, A.L., de Carvalho, J.C., Schmitt, C.C., de Oliveira Penha, R., Kumlehn, G.S., Soccol, C.R., 2021. Bioeconomy and biofuels: the case of sugarcane ethanol in Brazil. *Biofuels, Bioprod. Biorefining* 15, 899–912.
- Kern, M., McGeehan, J.E., Streeter, S.D., Martin, R.N.A., Besser, K., Elias, L., Eborall, W., Malyon, G.P., Payne, C.M., Himmel, M.E., Schnorr, K., Beckham, G.T., Cragg, S.M., Bruce, N.C., McQueen-Mason, S.J., 2013. Structural characterization of a unique marine animal family 7 cellobiohydrolase suggests a mechanism of cellulase salt tolerance. *Proc. Natl. Acad. Sci. U. S. A.* 110, 10189–10194.
- Kim, S., Dale, B.E., 2004. Global potential bioethanol production from wasted crops and crop residues. *Biomass and Bioenergy* 26, 361–375.
- Kim, S., Dale, B.E., Jenkins, R., 2009. Life cycle assessment of corn grain and corn stover in the United States. *Int. J. Life Cycle Assess.* 14, 160–174.
- Kleywegt, G.J., Zou, J.Y., Divne, C., Davies, G.J., Sinning, I., Ståhlberg, J., Reinikainen, T., Srisodsuk, M., Teeri, T.T., Jones, T.A., 1997. The crystal structure of the catalytic core domain of endoglucanase I from *trichoderma reesei* at 3.6 Å resolution, and a comparison with related enzymes. *J. Mol. Biol.* 272, 383–397.
- König, H., Li, L., Fröhlich, J., 2013. The cellulolytic system of the termite gut. *Appl. Microbiol. Biotechnol.* 97, 7943–7962.
- Kont, R., Bissaro, B., Eijsink, V.G.H., Våljamäe, P., 2020. Kinetic insights into the peroxygenase activity of cellulose-active lytic polysaccharide monooxygenases (LPMOs). *Nat. Commun.* 11.
- Korányi, T.I., Fridrich, B., Pineda, A., Barta, K., 2020. Development of ‘Lignin-First’ Approaches for the Valorization of Lignocellulosic Biomass. *Molecules* 25.
- Koshland, D.E., 1953. STEREOCHEMISTRY AND THE MECHANISM OF ENZYMIC REACTIONS. *Biol. Rev.* 28, 416–436.
- Kubicki, J.D., Yang, H., Sawada, D., O’Neill, H., Oehme, D., Cosgrove, D., 2018. The Shape of Native Plant Cellulose Microfibrils. *Sci. Rep.* 8, 4–11.
- Kurasin, M., Valjamae, P., 2011. Processivity of Cellobiohydrolases Is Limited by the Substrate. *J. Biol. Chem.* 286, 169–177.
- Kuusk, S., Bissaro, B., Kuusk, P., Forsberg, Z., Eijsink, V.G.H., Sørli, M., Valjamae, P., 2018. Kinetics of H₂O₂-driven degradation of chitin by a

- bacterial lytic polysaccharide monooxygenase. *J. Biol. Chem.* 293, 523–531.
- Langston, J.A., Shaghasi, T., Abbate, E., Xu, F., Vlasenko, E., Sweeney, M.D., 2011. Oxidoreductive cellulose depolymerization by the enzymes cellobiose dehydrogenase and glycoside hydrolase 61. *Appl. Environ. Microbiol.* 77, 7007–7015.
- Lee, D.W., Hwang, I., 2020. A Fight between Plants and Pathogens for the Control of Chloroplasts. *Cell Host Microbe* 28, 351–352.
- Linger, J.G., Taylor, L.E., Baker, J.O., Vander Wall, T., Hobdey, S.E., Podkaminer, K., Himmel, M.E., Decker, S.R., 2015. A constitutive expression system for glycosyl hydrolase family 7 cellobiohydrolases in *Hypocrea jecorina*. *Biotechnol. Biofuels* 8, 45.
- Liu, F., Short, M.D., Alvarez-Gaitan, J.P., Guo, X., Duan, J., Saint, C., Chen, G., Hou, L., 2020. Environmental life cycle assessment of lignocellulosic ethanol-blended fuels: A case study. *J. Clean. Prod.* 245, 118933.
- Lynd, L.R., Weimer, P.J., Zyl, W.H. Van, Pretorius, I.S., 2002. Microbial Cellulose Utilization: Fundamentals and Biotechnology. *Bioresour. Technol.* 66, 506–577.
- Mäkelä, M.R., Donofrio, N., De Vries, R.P., 2014. Plant biomass degradation by fungi. *Fungal Genet. Biol.* 72, 2–9.
- Marinović, M., Aguilar-Pontes, M.V., Zhou, M., Miettinen, O., de Vries, R.P., Mäkelä, M.R., Hildén, K., 2018. Temporal transcriptome analysis of the white-rot fungus *Obba rivulosa* shows expression of a constitutive set of plant cell wall degradation targeted genes during growth on solid spruce wood. *Fungal Genet. Biol.* 112, 47–54.
- Martinez, D., Berka, R.M., Henrissat, B., Saloheimo, M., Arvas, M., Baker, S.E., Chapman, J., Chertkov, O., Coutinho, P.M., Cullen, D., Danchin, E.G.J., Grigoriev, I. V., Harris, P., Jackson, M., Kubicek, C.P., Han, C.S., Ho, I., Larrondo, L.F., de Leon, A.L., Magnuson, J.K., Merino, S., Misra, M., Nelson, B., Putnam, N., Robbertse, B., Salamov, A.A., Schmoll, M., Terry, A., Thayer, N., Westerholm-Parvinen, A., Schoch, C.L., Yao, J., Barabote, R., Nelson, M.A., Detter, C., Bruce, D., Kuske, C.R., Xie, G., Richardson, P., Rokhsar, D.S., Lucas, S.M., Rubin, E.M., Dunn-Coleman, N., Ward, M., Brettin, T.S., 2008. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). *Nat. Biotechnol.* 26, 553–560.
- Martinez, D., Larrondo, L.F., Putnam, N., Gelpke, M.D.S., Huang, K., Chapman, J., Helfenbein, K.G., Ramaiya, P., Detter, J.C., Larimer, F., Coutinho, P.M., Henrissat, B., Berka, R., Cullen, D., Rokhsar, D., 2004. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.* 22, 695–700.
- McCann, H.C., 2020. Skirmish or war: the emergence of agricultural plant pathogens. *Curr. Opin. Plant Biol.* 56, 147–152.
- McNamara, J.T., Morgan, J.L.W., Zimmer, J., 2015. A molecular description of cellulose biosynthesis. *Annu. Rev. Biochem.* 84, 895–921.
- Mélida, H., Sandoval-Sierra, J. V., Diéguez-Uribeondo, J., Bulone, V., 2013.

- Analyses of extracellular carbohydrates in oomycetes unveil the existence of three different cell wall types. *Eukaryot. Cell* 12, 194–203.
- Mendgen, K., Hahn, M., Deising, H., 1996. Morphogenesis and mechanisms of penetration by plant pathogenic fungi. *Annu. Rev. Phytopathol.* 34, 367–386.
- Michaux, S.P., 2021. Assessment of the Extra Capacity Required of Alternative Energy Electrical Power Systems to Completely Replace Fossil Fuels.
- Michel, G., Chantalat, L., Duce, E., Barbeyron, T., Henrissat, B., Kloareg, B., Dideberg, O., 2001. The κ -carrageenase of *P. carrageenovora* Features a Tunnel-Shaped Active Site. *Structure* 9, 513–525.
- Mohanty, S.K., Swain, M.R., 2019. Bioethanol Production From Corn and Wheat: Food, Fuel, and Future. *Bioethanol Prod. from Food Crop.* 45–59.
- Mohnen, D., 2008. Pectin structure and biosynthesis. *Curr. Opin. Plant Biol.* 11, 266–277.
- Monschauer, Y., Lemercier, T., Deng, Y., Luna, L., Höhne, N., Brecha, R., Cantzler, J., Ganti, G., Schaeffer, M., Fuentes Hutfilter, U., 2019. TRANSFORMATION POINTS: ACHIEVING THE SPEED AND SCALE REQUIRED FOR FULL DECARBONISATION, CAT Decarbonisation Series.
- Nakamura, A., Kanazawa, T., Furuta, T., Sakurai, M., Saloheimo, M., Samejima, M., Koivula, A., Igarashi, K., 2021. Role of Tryptophan 38 in Loading Substrate Chain into the Active-site Tunnel of Cellobiohydrolase I from *Trichoderma reesei*. *J. Appl. Glycosci.* 68, 19–29.
- Nakamura, A., Okazaki, K. ichi, Furuta, T., Sakurai, M., Iino, R., 2018. Processive chitinase is Brownian monorail operated by fast catalysis after peeling rail from crystalline chitin. *Nat. Commun.* 9, 3814.
- Nakamura, A., Tsukada, T., Auer, S., Furuta, T., Wada, M., Koivula, A., Igarashi, K., Samejima, M., 2013. The tryptophan residue at the active site tunnel entrance of *Trichoderma reesei* cellobiohydrolase Cel7A is Important for initiation of degradation of crystalline cellulose. *J. Biol. Chem.* 288, 13503–13510.
- Nakashima, K., Azuma, J.I., 2000. Distribution and properties of endo- β -1,4-glucanase from a lower termite, *Coptotermes formosanus* (shiraki). *Biosci. Biotechnol. Biochem.*
- Nakashima, K., Watanabe, H., Azuma, J.I., 2002. Cellulase genes from the parabasalian symbiont *Pseudotrichonympha grassii* in the hindgut of the wood-feeding termite *Coptotermes formosanus*. *Cell. Mol. Life Sci.* 59, 1554–1560.
- Nelsen, M.P., DiMichele, W.A., Peters, S.E., Boyce, C.K., 2016. Delayed fungal evolution did not cause the Paleozoic peak in coal production. *Proc. Natl. Acad. Sci. U. S. A.* 113, 2442–2447.
- Nishimura, Y., Otagiri, M., Yuki, M., Shimizu, M., Inoue, J. ichi, Moriya, S., Ohkuma, M., 2020. Division of functional roles for termite gut protists revealed by single-cell transcriptomes. *ISME J.* 14, 2449–2460.
- Olson, Å., Aerts, A., Asiegbu, F., Belbahri, L., Bouzid, O., Broberg, A., Canbäck,

- B., Coutinho, P.M., Cullen, D., Dalman, K., Deflorio, G., van Diepen, L.T.A., Dunand, C., Duplessis, S., Durling, M., Gonthier, P., Grimwood, J., Fossdal, C.G., Hansson, D., Henrissat, B., Hietala, A., Himmelstrand, K., Hoffmeister, D., Höglberg, N., James, T.Y., Karlsson, M., Kohler, A., Kües, U., Lee, Y.H., Lin, Y.C., Lind, M., Lindquist, E., Lombard, V., Lucas, S., Lundén, K., Morin, E., Murat, C., Park, J., Raffaello, T., Rouzé, P., Salamov, A., Schmutz, J., Solheim, H., Ståhlberg, J., Véléz, H., de Vries, R.P., Wiebenga, A., Woodward, S., Yakovlev, I., Garbelotto, M., Martin, F., Grigoriev, I. V., Stenlid, J., 2012. Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen. *New Phytol.* 194, 1001–1013.
- Payne, C.M., Knott, B.C., Mayes, H.B., Hansson, H., Himmel, M.E., Sandgren, M., Ståhlberg, J., Beckham, G.T., 2015. Fungal Cellulases. *Chem. Rev.* 115, 1308–1448.
- Payne, C.M., Resch, M.G., Chen, L., Crowley, M.F., Himmel, M.E., Taylor, L.E., Sandgren, M., Ståhlberg, J., Stals, I., Tan, Z., Beckham, G.T., 2013. Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc. Natl. Acad. Sci. U. S. A.* 110, 14646–51.
- Penttilä, M., Nevalainen, H., Rättö, M., Salminen, E., Knowles, J., 1987. A versatile transformation system for the cellulolytic filamentous fungus *Trichoderma reesei*. *Gene* 61, 155–164.
- Penttilä, R., Siitonen, J., Kuusinen, M., 2004. Polypore diversity in managed and old-growth boreal *Picea abies* forests in southern Finland 117, 271–283.
- Peterson, R., Nevalainen, H., 2012. *Trichoderma reesei* RUT-C30 - Thirty years of strain improvement. *Microbiology* 158, 58–68.
- Phillips, C.M., Beeson, W.T., Cate, J.H., Marletta, M.A., 2011. Cellobiose dehydrogenase and a copper-dependent polysaccharide monooxygenase potentiate cellulose degradation by *Neurospora crassa*. *ACS Chem. Biol.* 6, 1399–406.
- Rantasalo, A., Landowski, C.P., Kuivanen, J., Korppoo, A., Reuter, L., Koivistoinen, O., Valkonen, M., Penttilä, M., Jäntti, J., Mojzita, D., 2018. A universal gene expression system for fungi. *Nucleic Acids Res.* 46, e111.
- Rantasalo, A., Vitikainen, M., Paasikallio, T., Jäntti, J., Landowski, C.P., 2019. Novel genetic tools that enable highly pure protein production in *Trichoderma reesei*.
- Rastogi, M., Shrivastava, S., 2017. Recent advances in second generation bioethanol production: An insight to pretreatment, saccharification and fermentation processes. *Renew. Sustain. Energy Rev.* 80, 330–340.
- Ravalason, H., Jan, G., Mollé, D., Pasco, M., Coutinho, P.M., Lapierre, C., Pollet, B., Bertaud, F., Petit-Conil, M., Grisel, S., Sigoillot, J.C., Asther, M., Herpœl-Gimbert, I., 2008. Secretome analysis of *Phanerochaete chrysosporium* strain CIRM-BRFM41 grown on softwood. *Appl. Microbiol. Biotechnol.* 80, 719–733.
- Renewable Fuels Association, n.d. Annual Ethanol Production [WWW Document]. URL <https://ethanolrfa.org/markets-and-statistics/annual-ethanol-production>

(accessed 10.22.22).

- Robinson, J.M., 1990. Lignin, land plants, and fungi: Biological evolution affecting Phanerozoic oxygen balance. *Geology* 18, 607.
- Rosales-Calderon, O., Arantes, V., 2019. A review on commercial-scale high-value products that can be produced alongside cellulosic ethanol, *Biotechnology for Biofuels*. BioMed Central.
- Ruiz-Dueñas, F.J., Morales, M., García, E., Miki, Y., Martínez, M.J., Martínez, A.T., 2009. Substrate oxidation sites in versatile peroxidase and other basidiomycete peroxidases. *J. Exp. Bot.* 60, 441–452.
- Rytöja, J., Hildén, K., Yuzon, J., Hatakka, A., de Vries, R.P., Mäkelä, M.R., 2014. Plant-polysaccharide-degrading enzymes from Basidiomycetes. *Microbiol. Mol. Biol. Rev.* 78, 614–49.
- Sabbadin, F., Henrissat, B., Bruce, N.C., McQueen-mason, S.J., 2021a. Lytic polysaccharide monooxygenases as chitin-specific virulence factors in crayfish plague. *Biomolecules* 11, 1–12.
- Sabbadin, F., Urresti, S., Henrissat, B., Avrova, A.O., Welsh, L.R.J., Lindley, P.J., Csukai, M., Squires, J.N., Walton, P.H., Davies, G.J., Bruce, N.C., Whisson, S.C., McQueen-Mason, S.J., 2021b. Secreted pectin monooxygenases drive plant infection by pathogenic oomycetes. *Science* (80-.). 373, 774–779.
- Seidl, V., Seibel, C., Kubicek, C.P., Schmoll, M., 2009. Sexual development in the industrial workhorse *Trichoderma reesei*. *Proc. Natl. Acad. Sci. U. S. A.* 106, 13909–13914.
- Sethi, A., Kovaleva, E.S., Slack, J.M., Brown, S., Buchman, G.W., Scharf, M.E., 2013. A GHF7 cellulase from the protist symbiont community of reticulitermes flavipes enables more efficient lignocellulose processing by host enzymes. *Arch. Insect Biochem. Physiol.* 84, 175–193.
- Singh, A., Taylor, L.E., Vander Wall, T.A., Linger, J., Himmel, M.E., Podkaminer, K., Adney, W.S., Decker, S.R., 2015. Heterologous protein expression in *Hypocrea jecorina*: A historical perspective and new developments. *Biotechnol. Adv.*
- Ståhlberg, J., Johansson, G., Pettersson, G., 1993. *Trichoderma reesei* has no true exo-cellulase: all intact and truncated cellulases produce new reducing end groups on cellulose. *BBA - Gen. Subj.* 1157, 107–113.
- Su, Yujie, Zhang, P., Su, Yuqing, 2015. An overview of biofuels policies and industrialization in the major biofuel producing countries. *Renew. Sustain. Energy Rev.* 50, 991–1003.
- Subramanian, V., Schuster, L.A., Moore, K.T., Taylor, L.E., Baker, J.O., Vander Wall, T.A., Linger, J.G., Himmel, M.E., Decker, S.R., 2017. A versatile 2A peptide-based bicistronic protein expressing platform for the industrial cellulase producing fungus, *Trichoderma reesei*. *Biotechnol. Biofuels* 10, 1–15.
- Sützl, L., Foley, G., Gillam, E.M.J., Bodén, M., Haltrich, D., 2019. The GMC superfamily of oxidoreductases revisited: Analysis and evolution of fungal GMC oxidoreductases. *Biotechnol. Biofuels* 12, 1–18.

- Suzuki, M.R., Hunt, C.G., Houtman, C.J., Dalebroux, Z.D., Hammel, K.E., 2006. Fungal hydroquinones contribute to brown rot of wood. *Environ. Microbiol.* 8, 2214–2223.
- Tan, T.C., Kracher, D., Gandini, R., Sygmund, C., Kittl, R., Haltrich, D., Hällberg, B.M., Ludwig, R., Divne, C., 2015. Structural basis for cellobiose dehydrogenase action during oxidative cellulose degradation. *Nat. Commun.* 6.
- Tan, X., Hu, Y., Jia, Y., Hou, X., Xu, Q., Han, C., Wang, Q., 2020. A Conserved Glycoside Hydrolase Family 7 Cellobiohydrolase PsGH7a of *Phytophthora sojae* Is Required for Full Virulence on Soybean. *Front. Microbiol.* 11, 1–11.
- Todaka, N., Inoue, T., Saita, K., Ohkuma, M., Nalepa, C.A., Lenz, M., Kudo, T., Moriya, S., 2010a. Phylogenetic Analysis of Cellulolytic Enzyme Genes from Representative Lineages of Termites and a Related Cockroach. *PLoS One* 5, e8636.
- Todaka, N., Lopez, C.M., Inoue, T., Saita, K., Maruyama, J.I., Arioka, M., Kitamoto, K., Kudo, T., Moriya, S., 2010b. Heterologous expression and characterization of an endoglucanase from a symbiotic protist of the lower termite, *reticulitermes speratus*. *Appl. Biochem. Biotechnol.* 160, 1168–1178.
- Todaka, N., Moriya, S., Saita, K., Hondo, T., Kiuchi, I., Takasu, H., Ohkuma, M., Piero, C., Hayashizaki, Y., Kudo, T., 2007. Environmental cDNA analysis of the genes involved in lignocellulose digestion in the symbiotic protist community of *Reticulitermes speratus*. *FEMS Microbiol. Ecol.* 59, 592–599.
- Vandhana, T.M., Reyre, J. Lou, Sushmaa, D., Berrin, J.G., Bissaro, B., Madhuprakash, J., 2022. On the expansion of biological functions of lytic polysaccharide monooxygenases. *New Phytol.* 233, 2380–2396.
- Vanholme, R., Demedts, B., Morreel, K., Ralph, J., Boerjan, W., 2010. Lignin biosynthesis and structure. *Plant Physiol.* 153, 895–905.
- Vermaas, J. V., Kont, R., Beckham, G.T., Crowley, M.F., Gudmundsson, M., Sandgren, M., Ståhlberg, J., Våljamäe, P., Knott, B.C., 2019. The dissociation mechanism of processive cellulases. *Proc. Natl. Acad. Sci. U. S. A.* 116, 23061–23067.
- Vogel, S., Alvarez, B., Bässler, C., Müller, J., Thorn, S., 2017. The Red-belted Bracket (*Fomitopsis pinicola*) colonizes spruce trees early after bark beetle attack and persists 27, 182–188.
- Wang, Y., Cheng, M.H., Wright, M.M., 2018. Lifecycle energy consumption and greenhouse gas emissions from corn cob ethanol in China. *Biofuels, Bioprod. Biorefining* 12, 1037–1046.
- Watanabe, H., Nakashima, K., Saito, H., Slaytor, M., 2002. New endo- β -1,4-glucanases from the parabasalian symbionts, *Pseudotrichonympha grassii* and *Holomastigotoides mirabile* of *Coptotermes* termites. *Cell. Mol. Life Sci.* 59, 1983–1992.
- Westereng, B., Ishida, T., Vaaje-Kolstad, G., Wu, M., Eijsink, V.G.H., Igarashi, K., Samejima, M., Ståhlberg, J., Horn, S.J., Sandgren, M., 2011. The Putative Endoglucanase PcGH61D from *Phanerochaete chrysosporium* Is a Metal-

- Dependent Oxidative Enzyme that Cleaves Cellulose. *PLoS One* 6, e27807.
- Wolfenden, R., Snider, M.J., 2001. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* 34, 938–945.
- Woon, J.S.K., King, P.J.H., Mackeen, M.M., Mahadi, N.M., Wan Seman, W.M.K., Broughton, W.J., Abdul Murad, A.M., Abu Bakar, F.D., 2017. Cloning, Production and Characterization of a Glycoside Hydrolase Family 7 Enzyme from the Gut Microbiota of the Termite *Coptotermes curvignathus*. *Mol. Biotechnol.* 59, 271–283.
- Wymelenberg, A. Vanden, Gaskell, J., Mozuch, M., Sabat, G., Ralph, J., Skyba, O., Mansfield, S.D., Blanchette, R.A., Martinez, D., Grigoriev, I., Kersten, P.J., Cullen, D., 2010. Comparative transcriptome and secretome analysis of wood decay fungi *postia placenta* and *phanerochaete chrysosporium*. *Appl. Environ. Microbiol.* 76, 3599–3610.
- Xu, G., Goodell, B., 2001. Mechanisms of wood degradation by brown-rot fungi: Chelator-mediated cellulose degradation and binding of iron by cellulose. *J. Biotechnol.* 87, 43–57.
- Zabed, H., Sahu, J.N., Boyce, A.N., Faruq, G., 2016. Fuel ethanol production from lignocellulosic biomass: An overview on feedstocks and technological approaches. *Renew. Sustain. Energy Rev.* 66, 751–774.
- Zeng, Y., Himmel, M.E., Ding, S.Y., 2017. Visualizing chemical functionality in plant cell walls. *Biotechnol. Biofuels* 10, 1–16.
- Zhang, P., McGlynn, A.C., West, C.M., Loomis, W.F., Blanton, R.L., 2001. Spore coat formation and timely sporulation depend on cellulose in *Dictyostelium*. *Differentiation* 67, 72–79.
- Zhang, W., Kou, Y., Xu, J., Cao, Y., Zhao, G., Shao, J., Wang, H., Wang, Z., Bao, X., Chen, G., Liu, W., 2013. Two major facilitator superfamily sugar transporters from *Trichoderma reesei* and their roles in induction of cellulase biosynthesis. *J. Biol. Chem.* 288, 32861–32872.
- Zhong, R., Ye, Z.H., 2015. Secondary cell walls: Biosynthesis, patterned deposition and transcriptional regulation. *Plant Cell Physiol.* 56, 195–214.

Popular science summary

Enormous amounts of plant material are produced on earth every year. Many organisms have evolved to use this resource as an energy source. To do this effectively requires effective catalytic proteins, enzymes, for turning carbohydrates such as cellulose into smaller sugars that can be used for energy metabolism. Perhaps the most effective organisms utilizing plant material for energy are many fungi, often growing on plant materials such as wood, decomposing it. Nowadays, this ability of fungi to produce plant degrading enzymes is taken advantage of in many industrial processes, where fungal enzymes are added for facilitating conversion of cellulose and other carbohydrates into smaller sugars. One example of a such process is the production of ethanol by fermenting sugars into alcohol. While traditionally ethanol production has mainly been done from starch rich grains, in order to make ethanol biofuels a sustainable and feasible alternative for gasoline, plant parts containing more cellulose should be used. This requires the development of more efficient enzymes as they are a significant component of the costs of the process.

Developing more effective enzymes requires understanding what kind of attributes are required for efficient catalysis. This can be aided by resolving the molecular structures of these enzymes at an atomic level, which allows then understanding which structural features are responsible for certain enzyme activity characteristics. In this thesis work I have focused on a group of cellulose degrading enzymes, cellulases, called glycoside hydrolase family 7, or GH7 in short. This type of cellulase enzymes are probably the most important enzyme fungi use for breaking down cellulose, and thus an important part of recycling the carbon atoms contained in cellulose in nature, but also in industrial processes where cellulose degrading enzymes are used. Two different kinds of cellulases are found in this family. So called

cellobiohydrolases break down long cellulose chains processively by threading cellulose chains into their tunnel shaped cavity, and then progressing along a cellulose chain while cleaving it into short fragments called cellobiose, a sugar containing two glucose sugars in one molecule. So called endoglucanases do not progress on cellulose chains in a processive manner, but make cuts in the cellulose chains in a more random manner, this way also making new entry points for the cellobiohydrolases to thread on and start their processive runs.

Traditionally the activities of cellulase enzymes have often been studied by testing their activities using artificial chemical compounds derived from carbohydrates that allow fast and easy detection of enzyme activity by color. However, over time it has become clear that high activity on these compounds does not necessarily correspond to high activity on natural or treated cellulose. We conducted a study to better understand this discrepancy, and showed that this is mostly due to the synthetic compounds binding into positions on the enzyme where they are not cleaved, and in fact block the cleavage site, thus leading to reduced degradation activity. Understanding this phenomenon helps in understanding what causes the inconsistent enzyme activities on commonly used compounds, and how this relates to activity on more natural cellulose.

While many cellulases have been originally discovered in fungi, genetic studies have revealed that similar enzymes are also found in many different organisms. One example of GH7 enzymes producing organisms are found in wood-eating termites, but not in the termites themselves, but tiny unicellular eukaryotic organisms, called protists, found living in their guts. Symbiosis with these protists is crucial for the termites, as they can not degrade wood material effectively without them. In our study we conducted a detailed study of a GH7 endoglucanase from such protist, solving the first molecular structure of a GH7 cellulase from these organisms. The structure revealed that these protists produce GH7 cellulases which are significantly different from previously known GH7 enzymes, and could offer a great starting point for constructing more effective engineered cellulases.

Many GH7 enzymes contain two distinct parts, so called domains, one of which is responsible for catalyzing the reaction, as in cleaving cellulose chains into shorter pieces, and another for binding the enzyme on cellulose to make sure the enzyme efficiently "finds" cellulose chains to cleave. The details of how these two parts act together on processive enzymes are largely

unknown. We used a sophisticated molecular labeling technique to label the two domains on a GH7 enzyme with specific markers. A microscopy technique called FRET tells detailed information about the distance between two labels, in this case two different fluorescent molecules that we had attached at two different positions on the surface of the enzyme molecules. We use this to gain understanding into how the two domains are positioned and move in relation to each other. In another microscopy technique a gold nanoparticle was bound to the enzyme for imaging cellobiohydrolase runs on cellulose surface at very high resolution. This can provide detailed information into what exactly happens during this processive action, and in the future help in determining which parts of the enzyme are crucial for this motion, and perhaps even ways to improve it. Our study is the first to demonstrate the use of these two advanced techniques for studying cellulases, and will hopefully help in designing future studies for this purpose.

Populärvetenskaplig sammanfattning

Enorma mängder växtmaterial produceras på jorden varje år. Många organismer har utvecklats för att använda denna resurs som närings- och energikälla. För att göra detta effektivt krävs effektiva katalytiska proteiner, enzymer, för att bryta ned kolhydratpolymerer som cellulosa till mindre sockerarter som kan användas för energimetabolism. De kanske mest effektiva organismerna som använder växtmaterial för energi är många svampar som ofta växer på och bryter ned trä och annat växtmaterial. Numera utnyttjas denna förmåga hos svampar att producera växtnedbrytande enzymer i många industriella processer, där svampenzymer tillsätts för att underlätta omvandlingen av cellulosa och andra kolhydrater till mindre sockerarter. Ett exempel på en sådan process är framställning av etanol genom att jäsa socker till alkohol. Medan traditionell etanolproduktion huvudsakligen har gjorts från stärkelsesrika spannmål, bör även växtmaterial som innehåller mer cellulosa användas för att göra etanolbiobränslen till ett hållbart och genomförbart alternativ för bensin. Detta kräver utveckling av effektivare enzymer eftersom de är en betydande del av kostnaderna för processen.

Att utveckla mer effektiva enzymer kräver att man förstår vilken typ av egenskaper som krävs för effektiv nedbrytning. Detta kan underlättas genom att ta reda på de molekylära strukturerna hos dessa enzymer på en atomär nivå, vilket gör det möjligt att sedan förstå vilka strukturella egenskaper som är ansvariga för vissa enzymaktivitetsegenskaper. I detta doktorsarbete har jag fokuserat på en grupp cellulosanedbrytande enzymer, cellulaser, kallade glykosidhydrolas-familj 7, eller kort och gott GH7. Denna typ av cellulaseenzymer är förmodligen de viktigaste enzymer som svampar använder för att bryta ner cellulosa, och de spelar därför en nyckelroll i omsättningen av det kol som binds i cellulosa i naturen, men också i

industriella processer där cellulosednedbrytande enzymer används. Två olika sorters cellulaser finns i denna familj. Så kallade cellobiohydrolaser (CBH) bryter ned långa cellulosedkedjor processivt genom att trä cellulosedkedjor in i deras tunnelformade hållighet och sedan fortskrida längs en cellulosedkedja samtidigt som de från änden klyver bort korta fragment som kallas cellobios, som innehåller två glukossocker i en molekyl. Så kallade endoglukanaser (EG) binder cellulosedkedjan i en klyfta istället för en tunnel och arbetar inte processivt, utan klyver cellulosa på ett mer slumpmässigt sätt, vilket skapar nya ändar som cellobiohydrolaserna kan trä in i sin tunnel och starta sina processiva körningar från.

Traditionellt har aktiviteten hos cellulasezymer ofta studerats genom att mäta deras aktiviteter med hjälp av konstgjorda molekyler där små kolhydrater kopplats till andra ämnen som ger en färgreaktion när de bryts ned, vilket möjliggör snabb och enkel upptäckt och mätning av enzymaktivitet. Men med tiden har det blivit tydligt att hög aktivitet på dessa föreningar inte nödvändigtvis motsvarar hög aktivitet på naturlig eller behandlad cellulosa. Vi genomförde en studie för att bättre förstå denna avvikelse, och visade att detta mestadels beror på att de syntetiska föreningarna binder till positioner på enzymet där de inte klyvs, och faktiskt blockerar klyvningsstället, vilket leder till minskad nedbrytningsaktivitet. Att förstå detta fenomen hjälper till att förstå vad som orsakar de inkonsekventa enzymaktiviteterna på vanliga föreningar, och hur detta relaterar till aktivitet på naturlig cellulosa.

Medan många cellulaser ursprungligen har upptäckts i svampar, har genetiska studier visat att liknande enzymer också finns i många andra vitt skilda organismer. Ett exempel på organismer med GH7-enzym är vedätande termiter, men enzymerna produceras inte av termiterna själva, utan i små encelliga eukaryota organismer, kallade protister, som lever i deras mage. Symbios med dessa protister är avgörande för termiterna, eftersom de inte kan bryta ned trämaterial effektivt utan dem. I vår studie genomförde vi en detaljerad studie av ett GH7-endoglukanas från en sådan protist, och kunde kartlägga den första molekylära strukturen av ett GH7-cellulas från dessa organismer. Strukturen avslöjade att dessa protister producerar GH7-cellulaser som skiljer sig väsentligt från tidigare kända GH7-enzym och kan erbjuda en bra utgångspunkt för att konstruera mer effektiva cellulaser i framtiden.

Många GH7-enzymmer innehåller två distinkta funktionella enheter, så kallade domäner, varav en är ansvarig för att katalysera reaktionen, dvs. att klyva cellulosedjor i kortare bitar, och en annan för att binda enzymet på cellulosa för att säkerställa att enzymet effektivt "hittar" cellulosedjor att klyva. Detaljerna om hur dessa två delar verkar tillsammans på processiva enzymmer är i stort sett okända. Vi använde en sofistikerad molekylär märkningsteknik för att märka de två domänerna på ett GH7-enzym med specifika markörer. En mikroskopiteknik som kallas FRET berättar detaljerad information om avståndet mellan två markörer, i det här fallet två olika fluorescerande molekyler som vi hade fäst vid två olika positioner på ytan av enzymmolekylerna. Vi använder detta för att få förståelse för hur de två domänerna är placerade och rör sig i förhållande till varandra. I en annan mikroskopiteknik bands en guldnanopartikel till enzymet för att se och följa hur cellobiohydrolaser arbetar processivt på cellulosaytan med mycket hög upplösning. Detta kan ge detaljerad information om exakt vad som händer under denna processiva nedrytning, och i framtiden hjälpa till att avgöra vilka delar av enzymet som är avgörande för denna aktivitet, och kanske till och med sätt att förbättra den. Vår studie är den första som visar användningen av dessa två avancerade tekniker för att studera cellulaser, och kommer förhoppningsvis att bana väg för framtida studier som ger mycket detaljerad insikt i dessa enzymers funktion, och hur de kan optimeras för bioteknisk användning.

Acknowledgements

If you're reading this, I must have made it in the end. Wow. Ojoj. Huhhuh. My PhD journey has not been the most straightforward, and frankly at times not easy. Yet, I really couldn't have hoped for better people to be working with. Perhaps my facial expressions and demeanor often ranging from neutral to extra grumpy, and with me at times running away from all social situations, haven't done wonders for showing my appreciation, but I truly feel it's been a huge privilege to be working with everyone in the group, corridor, department, and BioCentrum.

Obviously, most people thank their supervisors. But if I didn't have the supportive, patient and knowledgeable supervisors I've had, I would have burned out and fallen many times over. Thank you Corine for your support and encouragement from the start to finish. Thank you Henrik, for being available whenever I have had the sense to ask for help. Thank you Jan for your insights about scientific matters and about life in academia. Thank you Mats for so many things, but especially for having confidence in me, helping me to finish this project in the first place, and for creating an environment where it is great to come to work every day. Jerry, I don't know how to thank you enough for your patience, encouragement and support, but also enthusiasm and inspiration. I have only managed to capture a fraction of the vast knowledge you have, yet I have learned so much from you.

Nils, the man who keeps everything running and somehow never fails to have a great attitude, it's always been a pleasure working with you. You're a big part of what has made this a great place to work. The same goes for you Johanna, and I also want to thank you for all your help in the lab. Thank you Miao, Micke and Saeid for warmly welcoming me into the group, as well as all your help and advice regarding molecular biology and crystallography. Thank you Bing, not only for teaching me so much about *Pichia* and protein

purification, but also for being a great hiking, cooking, and renovation partner, and a great friend. Cori, I'm grateful for having gotten to know you right in the beginning of my PhD, and hopefully long after. Sumitha, I almost feel like I should apologize for helping you get started with this stuff called lignin, especially when I have experienced what it can do to people. Jokes aside, I always enjoyed your insights, whether it is about science, academia, or life in general, and it was really nice to have a familiar face in a conference, even if it was about lignin. Igor, it's been really nice and uplifting to have you around, and I appreciate all your help with data collection, crystallography and everything else in and out of the lab.




Gustav, I have really enjoyed and appreciated having you as lunch company, a Swedish teacher, but really also a mentor and a friend. A special thank you to Laura and Gisele for tolerating me as an office mate, and for all the fun debates and laughs along the years. I will never admit it, but I will miss you after all this is over. Thank you Jonas and Mikolaj for peer support, and for trying to keep me sane by offering healthy (and not-so-healthy) distractions from work. I'm also grateful to previous office mates Anna, Jule, Benjamin and Mahfuz for your company, help in getting started, and for putting up with me and my questions. Thank you to Alyona and Martin for fun dinners, and for the possibility to participate in the special "How's the kappa going?" -stress resilience building program. Whenever I have dared to venture into the coffee corner, and otherwise as well of course, I have appreciated the awesome company of my PhD student colleagues Pernilla, Yasha, Andreia, Sanjana, Ani, Florence, Hasi, Tomas and Mathilde, and the newcomers who I wish I could have been more available to, Naike and Piera. Likewise I've always thoroughly enjoyed interacting with Volkmar, Sabine and Bettina whenever I've had the chance. Many thanks to Monika with all her help in PhD related matters and her always positive attitude. I would also like to thank Peter and Åse and the entirety of their groups, including Yong, Adrian, Anna, Kerstin, and Li for making the corridor a great place to work. The same goes for the recent arrival Simon, you will fit right in.

I own a great gratitude to Akihiko, whose work and help was instrumental in me actually finishing this thesis, and unfortunately did not get to visit in the end due to circumstances. Thank you to Steve and Venkat for all your help, including valuable advice on *Trichoderma*. I'm also grateful for having gotten to know and to work with Japheth and Christy, and wish there would have been possibilities to do even more. I also want to once more thank all

my co-authors on the model compound paper, with a special thanks to Anu and Gunnar J for helping to make it all happen. Thank you to Carol and Magnus for their help with the *Aphanomyces* work. Thank you to Gregg, Davinia and Gunnar H for their help with all things lignin, even if I ended up changing direction. I'm also grateful having had the chance to supervise and work with two great students, Linus and Terese.

Thank you to my dear friends Tommi and Miika for many welcome breaks from all of this, but also peer support, and of course often a place to stay during my visits. Kata, I will spare you the embarrassment of a public love letter here, but I really don't know where I would be without you. I love you like crazy. Kiitos äidille Seijalle kaikesta tuesta ja ymmärryksestä. Mitä vanhemmaksi tulen, sitä paremmin osaan arvostaa miten onnekas olen ollut saadessani sellaiset vanhemmat kuin olen saanut. Thank you to my brother Eetu, and his family Tiina, Aarni and Ohto, whose company and video calls never fail to make me smile, no matter how tired or stressed I am. Last, I would also like to dedicate a few words to my father Ilkka, who sadly passed away unexpectedly soon after I started this journey. He was always my greatest champion, and I know he would have been immensely proud of this as well.

Enzyme kinetics by GH7 cellobiohydrolases on chromogenic substrates is dictated by non-productive binding: insights from crystal structures and MD simulation

Topi Haataja¹, Japheth E. Gado^{2,3}, Anu Nutt^{4,5}, Nolan T. Anderson², Mikael Nilsson⁶, Majid Haddad Momeni¹, Roland Isaksson⁶, Priit Väljamäe⁵, Gunnar Johansson⁴ , Christina M. Payne²  and Jerry Ståhlberg¹ 

1 Department of Molecular Sciences, Swedish University of Agricultural Sciences, Uppsala, Sweden

2 Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY, USA

3 Renewable Resources and Enabling Sciences Center, National Renewable Energy Laboratory, Golden, CO, USA

4 Department of Chemistry, Uppsala University, Sweden

5 Institute of Molecular and Cell Biology, University of Tartu, Estonia

6 Institute of Chemistry and Biomedical Sciences, Linnaeus University, Kalmar, Sweden

Keywords

Cel7; cellulase; fluorescence; ligand binding; *Phanerochaete chrysosporium*; *Trichoderma reesei*

Correspondence

J. Ståhlberg, Department of Molecular Sciences, Swedish University of Agricultural Sciences, PO box 7015, SE-750 07 Uppsala, Sweden

Tel: +46-18-673182

E-mail: jerry.stahlberg@slu.se

C. M. Payne, Department of Chemical and Materials Engineering, University of Kentucky, Lexington, KY 40506, USA

Tel: +1 703-292-2895

E-mail: christy.payne@uky.edu

G. Johansson, Department of Chemistry, Uppsala University, PO box 576, SE-751 23 Uppsala, Sweden

Tel: +46-703-471349

E-mail: gunnar.johansson@kemi.uu.se

(Received 9 June 2022, revised 30 July 2022, accepted 17 August 2022)

doi:10.1111/febs.16602

Cellobiohydrolases (CBHs) in the glycoside hydrolase family 7 (GH7) (EC3.2.1.176) are the major cellulose degrading enzymes both in industrial settings and in the context of carbon cycling in nature. Small carbohydrate conjugates such as *p*-nitrophenyl- β -D-cellobioside (pNPC), *p*-nitrophenyl- β -D-lactoside (pNPL) and methylumbelliferyl- β -D-cellobioside have commonly been used in colorimetric and fluorometric assays for analysing activity of these enzymes. Despite the similar nature of these compounds the kinetics of their enzymatic hydrolysis vary greatly between the different compounds as well as among different enzymes within the GH7 family. Through enzyme kinetics, crystallographic structure determination, molecular dynamics simulations, and fluorometric binding studies using the closely related compound *o*-nitrophenyl- β -D-cellobioside (oNPC), in this work we examine the different hydrolysis characteristics of these compounds on two model enzymes of this class, TrCel7A from *Trichoderma reesei* and PcCel7D from *Phanerochaete chrysosporium*. Protein crystal structures of the E212Q mutant of TrCel7A with pNPC and pNPL, and the wildtype TrCel7A with oNPC, reveal that non-productive binding at the product site is the dominating binding mode for these compounds. Enzyme kinetics results suggest the strength of non-productive binding is a key determinant for the activity characteristics on these substrates, with PcCel7D consistently showing higher turnover rates (k_{cat}) than TrCel7A, but higher Michaelis–Menten (K_M) constants as well. Furthermore, oNPC turned out to be useful as an active-site probe for fluorometric determination of the dissociation constant for cellobiose on TrCel7A but could not be utilized for the same purpose on PcCel7D, likely due to strong binding to an unknown site outside the active site.

Abbreviations

CBH, cellobiohydrolase; GH7, glycoside hydrolase family 7; MD, molecular dynamics; MUC, methylumbelliferyl- β -D-cellobioside; oNPC, ortho-2-nitrophenyl- β -D-cellobioside; pNP, para-4-nitrophenol; PcCel7D, *Phanerochaete chrysosporium* Cel7D; pNPC, para-4-nitrophenyl- β -D-cellobioside; pNPL, para-4-nitrophenyl- β -D-lactoside; TrCel7A, *Trichoderma reesei* Cel7A.

Introduction

Cellobiohydrolases catalyse the hydrolysis of polymeric cellulose and cellooligosaccharides into cellobiose. They are the major workhorses in cellulose degradation in nature, and therefore play a critical role in the carbon cycle. The most abundant cellulases produced by cellulose degrading fungi are usually cellobiohydrolases from the family 7 of glycoside hydrolases (GH7) (EC3.2.1.176). Due to their effectiveness in cellulose degradation, the GH7 enzymes have been extensively studied and engineered to better understand their action mechanism and to utilize them in various biotechnical applications [1].

The GH7 cellobiohydrolases act on cellulose chains processively from the reducing towards the non-reducing end, with a cellulose chain threading into an active site tunnel containing 9–10 glucose unit binding subsites [1–4]. The processive action is dependent on strong affinity to the product binding sites +1/+2 [5–7]. Consequently, the hydrolysis product, cellobiose, binds to this site with high affinity and is a strong inhibitor for these enzymes, with cellobiose accumulation leading to undesirable product inhibition [8,9].

p- and *o*-nitrophenyl glycosides are widely used artificial chromogenic substrates for kinetic studies of GH7 enzymes [10]. These substrates do in principle allow real-time-monitoring, but the sensitivity is greatly enhanced if a non-continuous assay with alkalinification is employed. The inhibition of nitrophenyl glycoside hydrolysis is often employed to measure equilibrium binding of nondegradable ligands. A yet more sensitive detection of glycosidase activity is achieved by the use of fluorogenic substrates such as methylumbelliferyl glycosides [8,10,11]. However, the use of low molecular weight model substrates for polymer degrading enzymes with long arrays of substrate binding subsites is often complicated, since the model substrates may occupy, or even prefer non-productive positions, sometimes to such an extent that the non-productive binding modes are dominating, meaning that these substances can be characterized as reversibly binding inhibitors [12,13]. This leads to kinetic behaviour where both the apparent turn-over of the hydrolysis (k_{cat}) and binding affinity of the enzyme substrate complex (K_M) are affected, with stronger non-productive binding leading to reduction in both values. Thus the true catalytic turnover of the productive enzyme-substrate complex cannot be determined without the dissociation constant of the non-productively bound enzyme substrate complex [14].

Besides activity measurements, ligands with fluorogenic properties have in many cases been used as a probe for displacement titration to study biologically significant binding of non-fluorescent molecules to proteins [15,16]. For carbohydrate degrading enzymes, the changes in the fluorescence of methylumbelliferyl glycosides, when binding to the protein, have been used as active site probes [17–21]. As an alternative approach, the change of intrinsic protein fluorescence resulting from ligand binding may be monitored. Tryptophan fluorescence is sensitive to its environment and may thus be affected by ligand binding, either by contact effects or, alternatively, by means of radiationless energy transfer if the absorption spectrum of the ligand overlaps favourably with the tryptophan emission range. Indeed, substrate-induced changes in the native fluorescence of proteins have been utilized previously for ligand binding studies [22–25].

TrCel7A from the ascomycete fungus *Trichoderma reesei* and PcCel7D from the basidiomycete *Phanerochaete chrysosporium* are two model enzymes from the GH7 enzyme family. Both enzymes contain seven substrate binding subsites (–7 to –1) and three product binding subsites (+1 to +3), a conserved feature in GH7 cellobiohydrolases. Of the known GH7 cellobiohydrolases PcCel7D stands out with its open active site tunnel architecture, with significantly shorter loops in the so called A1, B2 and B3 loop regions compared to TrCel7A [26–28]. In studies conducted on the processivity and inhibition characteristics of these enzymes PcCel7D has shown weaker inhibition by cellobiose compared to TrCel7A, as well as more frequent dissociation from a cellulose chain, leading to slightly shorter processive runs but also higher overall hydrolysis rates with less enzyme bound unproductively in difficult-to-hydrolyse cellulose regions [13,29]. Both attributes are likely explained by the more open active site tunnel structure.

In this work we explore the binding dynamics of common model compounds *p*-nitrophenyl- β -D-cellobioside (pNPC), *p*-nitrophenyl- β -D-lactoside (pNPL) and methylumbelliferyl- β -D-cellobioside (MUC) to the active sites of TrCel7A and PcCel7D. We use enzyme kinetics measurements, X-ray crystal structures and molecular dynamics (MD) simulations to explore the factors governing the catalytic activity of these enzymes on these model substrates by studying productive and non-productive substrate binding, and to shed light into possible reasons for differences in the kinetics and inhibition behaviour. Furthermore, we explore the usefulness of *o*-nitrophenyl- β -D-cellobioside (oNPC) as an active site probe for these cellobiohydrolases.

Results

Enzyme kinetics

To compare the hydrolysis characteristics of the TrCel7A and PcCel7D cellobiohydrolases, enzyme kinetic measurements were performed with oNPC, pNPC, pNPL and MUC as substrates (Fig. 1A). The determined catalytic rate (k_{cat}) and Michaelis–Menten (K_{M}) constants are shown in Table 1, as well as $k_{\text{cat}}/K_{\text{M}}$, which is a measure of the catalytic efficiency of the enzyme with the substrate.

The kinetic parameters differ significantly between the enzymes as well as between substrates. If we start by comparing the catalytic efficiency, MUC seems to be most efficiently hydrolysed of the substrates, with $k_{\text{cat}}/K_{\text{M}}$ values about an order of magnitude higher than for pNPC and pNPL, which in turn are about an order of magnitude higher than for oNPC. Between the enzymes, $k_{\text{cat}}/K_{\text{M}}$ for MUC is very similar, while for oNPC, pNPC and pNPL the values are somewhat lower with PcCel7D than with TrCel7A. Overall, PcCel7D showed much weaker apparent affinity for the substrate than TrCel7A, as reflected by higher K_{M} , which is mostly compensated for by faster hydrolysis, i.e., higher k_{cat} values.

It is noteworthy that the lactoside substrate (pNPL) gave much higher values for both k_{cat} and K_{M} than the corresponding cellobioside substrate (pNPC), although these substrates only differ by the orientation of the 4-hydroxyl at the non-reducing end of the molecule, being axial in pNPL and equatorial in pNPC. The k_{cat} was about 30-fold higher for pNPL over pNPC with TrCel7A and 4-fold higher with PcCel7D. The activity against oNPC turned out to be significantly lower compared to the other substrates with both enzymes. However, the enzyme kinetics with oNPC were quite different for the two enzymes. With PcCel7D, k_{cat} and K_{M} for oNPC were about 230 and 460 times higher, respectively, than with TrCel7A. And while oNPC bound weaker than pNPC to PcCel7D (2.5 times higher K_{M}) and k_{cat} was about one third, oNPC bound even stronger than pNPC to TrCel7A (3.7-fold lower K_{M}) and k_{cat} was 39-fold lower.

Fluorescence titration with oNPC

The strong binding but slow hydrolysis of oNPC by TrCel7A suggests that it could in practice be utilized as a non-reactive inhibitor for GH7 binding studies. Inspired by this, we set out to explore the possibilities of utilizing oNPC as an active site probe for GH7

CBHs. Fluorescence measurements of TrCel7A and PcCel7D solutions demonstrated that addition of oNPC quenched the intrinsic fluorescence of the enzymes because of the overlap of the absorbance spectrum of oNPC with the fluorescence spectra of the enzymes (Fig. 1B). The decrease in fluorescence was dependent on the amount of added oNPC and followed Langmuir isotherms, which enabled the derivation of dissociation constants, K_{d} , for oNPC by regression analysis of fluorescence titration data (Fig. 1C, Table 2). With TrCel7A WT the K_{d} value for oNPC is very close to the K_{M} from enzyme kinetics experiments (7.4 μM vs. 7.0 μM). However, with PcCel7D we were surprised to find a much lower K_{d} of 110 μM for oNPC, compared to the K_{M} of 3200 μM (Table 2).

Addition of cellobiose results in recovery of fluorescence of TrCel7A. Thus, competitive displacement titration could be used for indirect determination of K_{d} for cellobiose (Fig. 1D, Table 2). The methods also allow for binding measurements with catalytically impaired mutants, and K_{d} for oNPC and cellobiose were determined for TrCel7A WT as well as its E212Q, D214N and E217Q mutants (Table 2). The mutants gave similar K_{d} values, albeit slightly lower with the E212Q and E217Q mutants. In the case of PcCel7D, however, fluorescence was not recovered upon addition of cellobiose up to 1 mM concentration (data not shown), which is about five times higher than the inhibition constant, K_{i} , of cellobiose reported previously for PcCel7D [13].

Inhibition by oNPC and lactose

The binding of oNPC to TrCel7A was further analysed by inhibition assays using MUC as substrate. The activity did indeed decrease in the presence of oNPC. Regression analysis of the enzyme kinetic curves confirmed a competitive mode of inhibition with an inhibition constant K_{i} of $5.6 \pm 0.5 \mu\text{M}$ for oNPC, in good agreement with fluorescence titration and enzyme kinetics results. Furthermore, the binding of lactose was assessed by inhibition assays using pNPL as substrate. Lactose showed competitive inhibition with both TrCel7A and PcCel7D and similar inhibition constants, 180 and 183 μM , respectively (Table 2).

Structures and ligand binding

Four new X-ray crystal structures are presented, of TrCel7A WT with oNPC, and TrCel7A E212Q mutant with pNPC, pNPL or lactose bound (PDB: 4V0Z, 4UWT, 7OC8, 7NYT, respectively). The structures

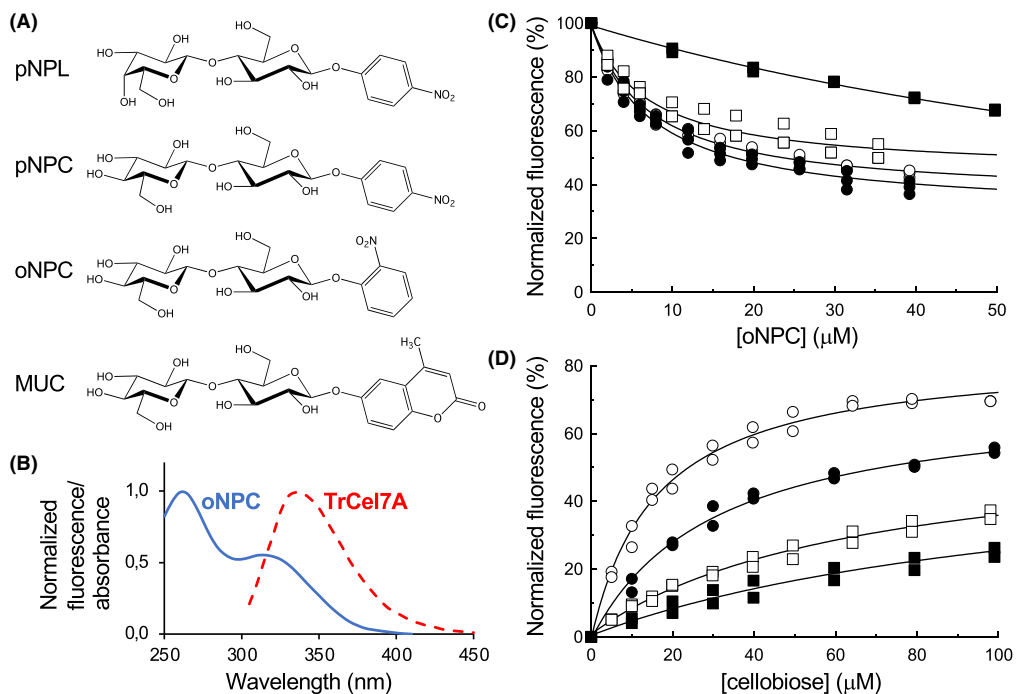


Fig. 1. Substrates and fluorescence titrations. (A) Enzyme kinetics experiments were performed with pNPL, pNPC, oNPC and MUC as substrates. (B) Absorbance spectrum of oNPC (blue) and fluorescence spectrum of TrCel7A wildtype at an excitation wavelength of 295 nm (red). (C) Fluorescence titration of TrCel7A and PcCel7D with oNPC. The concentration of TrCel7A was 0.035 μ M (\square); 0.1 μ M (\bullet) or 0.35 μ M (\circ) and the concentration of PcCel7D was 0.5 μ M (\blacksquare). (D) Displacement titration of TrCel7A with cellobiose at different concentrations of oNPC. The concentration of oNPC was 5 μ M (\circ); 10 μ M (\bullet), 20 μ M (\square) or 40 μ M (\blacksquare).

Table 1. Comparison of kinetic constants of TrCel7A and PcCel7D on different chromogenic substrates. Determined at 25 °C in 50 mM sodium acetate buffer, pH 5.0.

Enzyme	Substrate	k_{cat} (s^{-1})	K_M (μ M)	k_{cat}/K_M ($s^{-1} * \mu M^{-1}$)
TrCel7A	oNPC	$66 \times 10^{-6} \pm 15 \times 10^{-6}$	7.0 ± 4.5	9.5
	pNPC	0.0026 ± 0.0001	26 ± 3	100
	pNPL	0.087 ± 0.002	590 ± 20	147
	MUC	0.013 ± 0.001	12 ± 1	1083
Ratio, TrCel7A	pNPC/oNPC	39	3.7	11
	pNPL/pNPC	33	23	1.5
	MUC/pNPC	5.0	0.46	11
PcCel7D	oNPC	0.015 ± 0.002	3200 ± 100	4.6
	pNPC	0.046 ± 0.0021	1300 ± 160	35
	pNPL	0.17 ± 0.01	5500 ± 400	31
	MUC	0.22 ± 0.01	210 ± 20	1048
Ratio, PcCel7D	pNPC/oNPC	3.1	0.41	7.5
	pNPL/pNPC	3.7	4.2	0.87
	MUC/pNPC	4.8	0.16	30
Ratio, PcCel7D/TrCel7A	oNPC	227	457	0.50
	pNPC	18	50	0.35
	pNPL	2.0	9.3	0.21
	MUC	17	18	0.97

Table 2. Dissociation constants for oNPC and cellobiose binding to TrCel7A WT and catalytic mutants and PcCel7D from fluorescence titration experiments, and inhibition constants for cellobiose^a and lactose.

Enzyme	K_d for oNPC (μM)	K_d for cellobiose (μM)	K_i for cellobiose ^a (μM)	K_i for lactose (μM) ^b
TrCel7A WT	7.4 ± 0.4	23 ± 4	24^c	180 ± 16
TrCel7A D214N	7.1 ± 0.7	8.9 ± 1.1	–	–
TrCel7A E212Q	4.7 ± 0.4	8.1 ± 0.3	–	–
TrCel7A E217Q	3.9 ± 0.4	14 ± 3	–	–
PcCel7D	110 ± 10	–	180^d	183 ± 16

^aPreviously published inhibition constants from [13]; ^bThe error margin represents the 95% confidence interval of the profile likelihood from GRAPHPAD PRISM 8; ^cCompetitive inhibition constant from inhibition experiments with pNPL as substrate at 30 °C, pH 5.0 [13]; ^dMixed-type inhibition constant ($\alpha = 5.7$) estimated from inhibition experiments with CNP-Lac (2-chloro-4-nitrophenyl- β -lactoside) as substrate at 33 °C, pH 5.5 [13].

were refined at 1.7, 1.2, 1.5 and 1.1 Å resolution, respectively, and are nearly identical in terms of backbone structures (pairwise RMSD α -carbon trace values range from 0.10 to 0.28 Å). All the models contain the complete catalytic domain of Cel7A (residues 1

through 434), one N-acetyl glucosamine residue covalently linked to Asn 270 and between two and four Co^{2+} ions. Interestingly, each structure contains two ligand molecules, one at the active site and one on the outside, between neighbouring protein molecules in the crystal (Fig. 2B). In the structure with lactose at the active site, which was obtained by crystal soaking with pNPL, the electron density shows that the ligand molecule on the outside is pNPL rather than lactose. Statistics from diffraction data processing and structure refinement are summarized in Table 3.

The oNPC, pNPC, pNPL and lactose bound in the product binding sites show clear and unambiguous electron density for the sugar units in subsites +1 and +2, and somewhat weaker density for the nitrophenyl group in subsite +3 (Fig. 3). The sugar units of the aryl glycoside substrates bind in the so called “un-primed” binding mode, as designated by Knott *et al.* [30], i.e. with the non-reducing end sugar in subsite +1 close to the catalytic acid/base residue Glu217. This binding position of sugar units in the +1/+2 subsites is found in previous structures of both TrCel7A (PDB: 6CEL, 7CEL, 4C4C & 4C4D) [4,30] and PcCel7D (PDB: 1Z3W) [28]. However, in the lactose complex the disaccharide is slightly tilted away from the

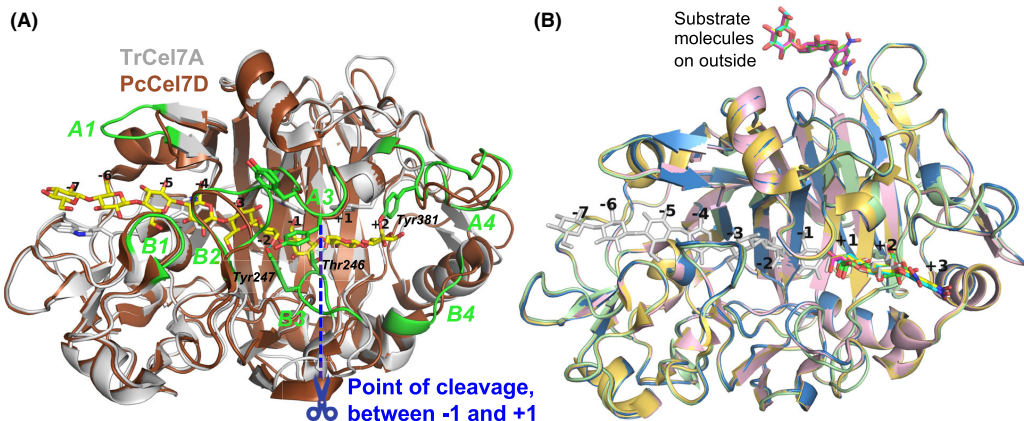


Fig. 2. Overview of protein structures and substrate binding. (A) Crystal structure of the catalytic domain of TrCel7A (light-grey) with cellobiose bound (yellow; PDB: 4C4C) and tunnel-enclosing loops highlighted and labelled in green, superposed with PcCel7D (brown; PDB: 1Z3W). The point of cleavage at the catalytic center is indicated in blue, from which the glucose unit subsites are numbered, with plus-signs towards the reducing end and minus-signs towards the non-reducing end of the sugar polymer. Sidechains are shown of the sugar-binding trophalan platforms at subsites -7, -4, -2 and +1, as well as selected residues involved in substrate binding near the catalytic center. Hydrogen bonds are indicated in cyan between Tyr247 and 6OH at subsite -2 and between Thr246 and 6OH at +1. (B) The four new crystal structures presented here, of TrCel7A showing the binding of the ligands in the product subsites +1 to +3 at the active site, and on the outside of the protein, relative to binding of cellobiose. Ligand/protein colours are as follows: pNPC, yellow/light-yellow (PDB: 4UWT); pNPL, cyan/light-blue (PDB: 7OC8); lactose, green/light-green (PDB: 7NYT); oNPC, magenta/pink (PDB: 4V0Z); Cellobiose, light-grey (PDB: 4C4C). The structure images were created with MACSYMOL [71].

Table 3. Statistics from X-ray diffraction data collection and processing, structure refinement and final model.

	E212Q/pNPC	WT/oNPC	E212Q/pNPL	E212Q/lactose
(A) Diffraction data				
PDB code	4UWT	4V0Z	7OC8	7NYT
Beamline	MAX-lab I911-2	ESRF ID14-3	BioMAX, MAX IV	BioMAX, MAX IV
Wavelength (Å)	1.041	0.931	0.980	0.980
Cell dimensions (Å)	83.30, 81.78, 110.53	83.06, 81.38, 109.94	83.19, 81.51, 109.92	83.54, 82.21, 110.73
Space group	I 2 2 2	I 2 2 2	I 2 2 2	I 2 2 2
Resolution range (Å)	29.2–1.15 (1.25–1.15)	29.1–1.70 (1.73–1.70)	41.6–1.60 (1.63–1.60)	41.8–1.09 (1.13–1.09)
No. of unique reflections ^a	130 094	40 566	49 321	137 896
Completeness (%) ^a	98.1 (87.2)	97.9 (67.6)	99.7 (99.6)	88.5 (41.4)
Multiplicity ^a	7.2 (6.5)	6.9 (5.0)	5.4 (5.5)	6.1 (2.7)
$I/\sigma(I)$ ^a	15.0 (3.2)	26.6 (5.8)	8.1 (2.3)	13.3 (1.3)
R_{merge} ^{a,b}	0.094 (0.53)	0.070 (0.26)	0.15 (0.76)	0.074 (0.68)
(B) Structure refinement				
Resolution used in refinement (Å)	29.19–1.20	14.94–1.70	35.13–1.60	41.77–1.09
No. of reflections, work set	111 569	38 460	47 026	131 092
No. of reflections, test set	5905	2033	2294	6802
R (work set) ^c	0.145	0.129	0.156	0.123
R_{free} ^c	0.159	0.157	0.184	0.143
No. of nonhydrogen atoms				
Protein atoms	3460	3400	3340	7768
Solvent atoms	681	525	405	588
Average B factors (Å ²)				
Overall	8.0	8.7	13.0	12.0
Protein	5.7	7.91	11.9	10.2
Water	18.7	18.63	20.8	21.1
Ligands (in active site)	13.0 (11.6)	20.2 (12.54)	22.0 (29.14)	13.1 (14.0)
RMSD bond lengths (Å)	0.006	0.016	0.010	0.004
RMSD bond angles (°)	1.22	1.76	1.64	1.29
Ramachandran plot outliers ^d	0	0	0	0

^aNumbers in parentheses are for the highest resolution bin; ^b $R_{\text{merge}} = \sum_{hkl} \sum_i |I - \langle I \rangle| / \sum_{hkl} \sum_i I$; ^c $R = \sum ||F_0| - |F_c|| / \sum |F_0|$; the final R-factor is given; ^dwwPDB Validation Service.

catalytic centre towards the exit of the active site, to the “primed” binding position (Fig. 4A), similar to the sugar binding at +1/+2 in the previous TrCel7A structures PDB: 3CEL, 4PLJ and 4D5O, and PcCel7D structures PDB: 1Z3T and 1Z3V [28,31,32]. For the galactose residue of lactose, in subsite +1, there is clear density for the 6-hydroxyl at two different positions and consequently two conformations were modelled of this residue. Also, two alternative conformations are seen in this structure for the B3 loop, from Asp241 to Thr255, as well as for Tyr371 at the tip of the opposing A3 loop.

The pNPC and pNPL substrates bind very similarly, as can be expected, given they only differ at the position of 4OH at the non-reducing end of the molecule (Fig. 4B,C). However, while the 4OH of pNPC makes a favourable hydrogen bond with Glu217, the 4OH of pNPL is instead making a close contact with Gln175 (3.0 Å to NE2 atom) where neither of the atoms are well oriented for hydrogen bonding (Fig. 4B). In the pNPL structure there is a cobalt ion present at the

active site with partial occupancy (0.3), appearing to interact with 3OH and 4OH of the galactose unit in pNPL and with Asp214 and His228. Given that there is electron density for only one binding position of pNPL at the product site, this Co²⁺ ion does not seem to significantly affect the binding position of pNPL. The sugar moieties of oNPC are virtually in the same position as in pNPC and pNPL, but the oNP unit is slightly shifted relative to pNP to avoid a clash between the 2-nitro group and Tyr381 and Pro382 in the A4 loop region (Fig. 4D). There is no sign of electron density in any of the structures for substrate binding at the catalytic centre (i.e. with the disaccharide in subsites –2/–1 and the nitrophenyl in +1) or elsewhere along the active site. Thus, the structures clearly demonstrate that non-productive binding at +1 to +3 is stronger and preferred over productive binding at –2 to +1 for these substrates, at least in the case of TrCel7A.

Superposition of the TrCel7A ligand structures with PcCel7D (lactose complex, PDB: 1Z3V) reveals

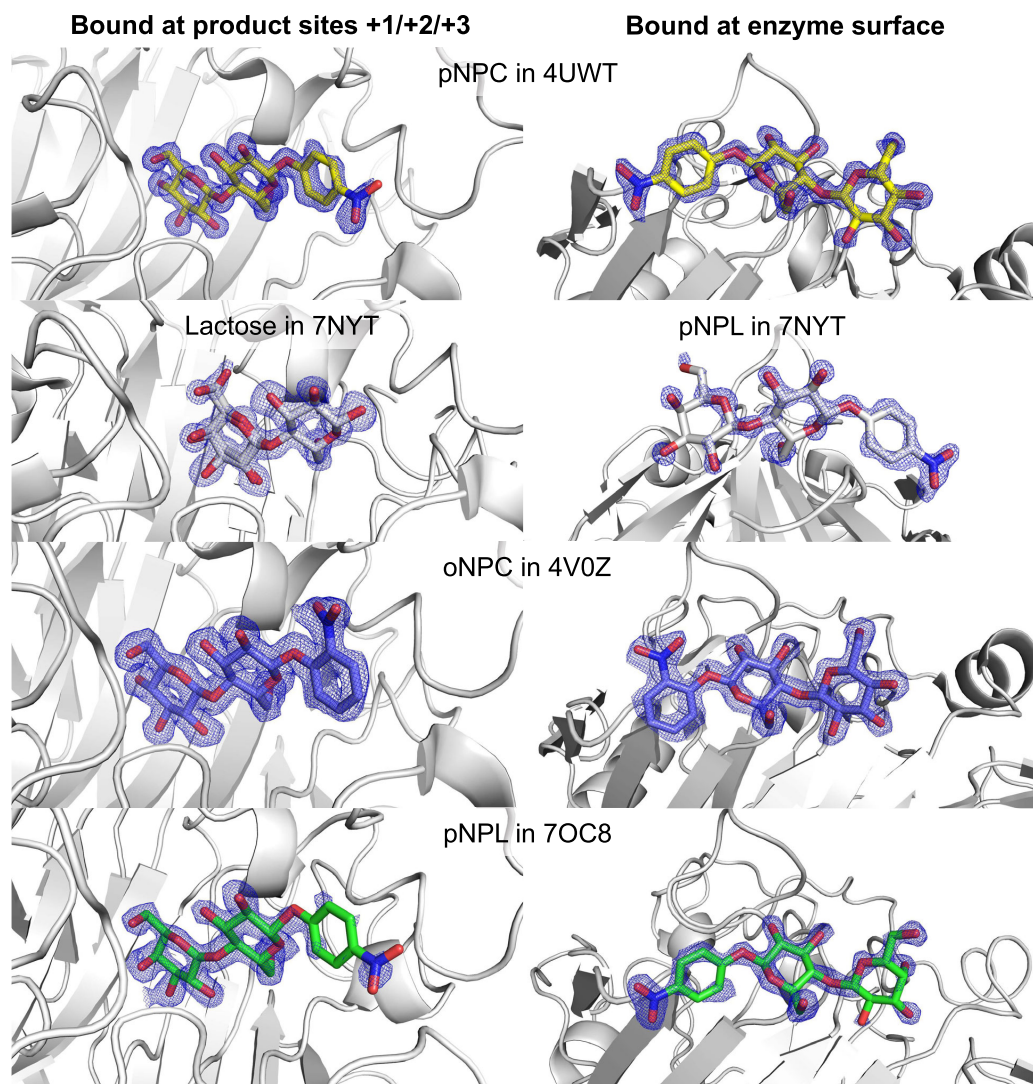


Fig. 3. $2f_o - f_c$ electron density maps for all the sugar ligands bound in the presented TrCel7A structures, those at the product site (left) and those at the surface (right): pNPC in 4UWT (yellow), lactose and pNPL in 7NYT (white), oNPC in 4V0Z (blue), pNPL in 7OC8 (green). Electron density map contour level 1.0σ . The structure images were created with *MACPYMOL* [71].

structural differences that likely forces a different positioning of these substrates in PcCel7D. The +3 subsite is more restricted in PcCel7D due to the insertion of Asp336 in the B4 loop. The carboxylate side chain of Asp336 is pointing towards subsite +3 and overlaps partially with the nitrophenyl groups of oNPC, pNPC and pNPL in the TrCel7A structures, with close

contacts of 1.0, 1.7 and 1.8 Å, respectively (Fig. 5). The same region in TrCel7A contains two glycine residues, allowing more space and possibly leading to stronger binding at the +3 site. Asp at this location is the most common motif among GH7 CBHs, but is missing in TrCel7A and a few closely related CBHs, due to a one-residue deletion in the B4 loop.

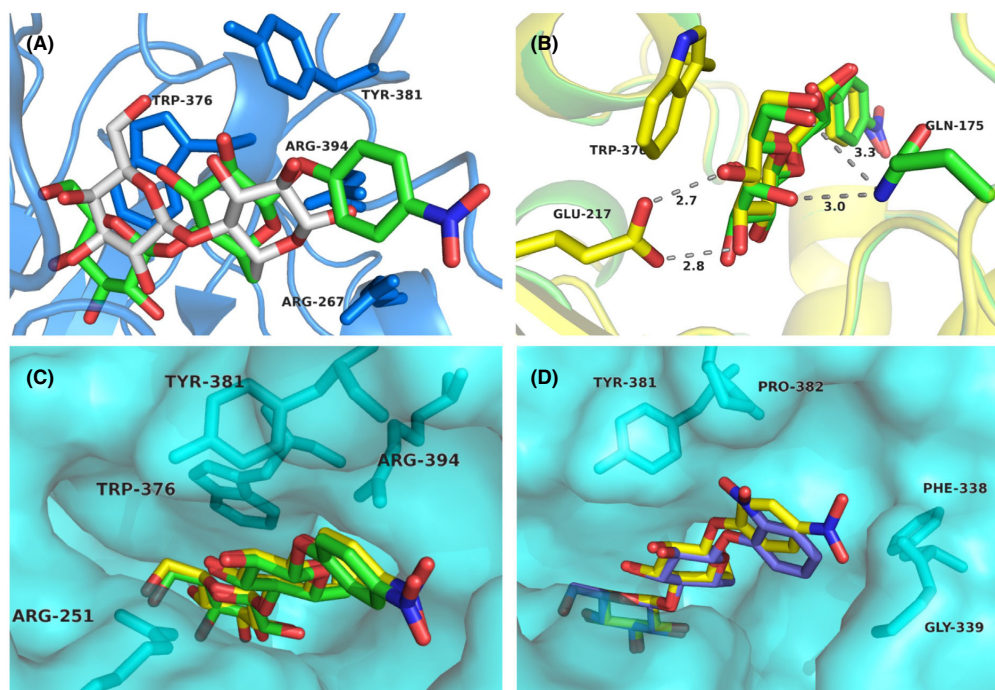


Fig. 4. Comparison of pNPC, pNPL, oNPC and lactose binding at the product subsites of TrCel7A. (A) The pNPL molecule (green) is bound with the sugar units in the “unprimed” position at subsites +1/+2 and with the nitrophenyl group at subsite +3, whereas lactose (white) binds in the “primed” position. The protein backbone and selected residues of the lactose complex (PDB: 7NYT) are shown in blue colour. (B) An overlay of pNPL (green) and pNPC (yellow) viewed from the catalytic center shows common hydrogen bonds between sugar and protein at subsite +1 (3OH to Glu217 and 6OH to Gln175), and the difference in orientation and interactions for 4OH, with Glu217 for the glucose residue of pNPC, and with Gln175 for the galactose residue of pNPL, respectively. (C) The pNPL (green) and pNPC (yellow) ligands, viewed from the active site exit towards the catalytic center, display very similar binding positions. (D) An overlay of pNPC (yellow) and oNPC (blue) shows the difference in binding of the respective nitrophenyl moieties while the cellobiose units overlap closely. In panels (C) and (D) the protein is shown in semitransparent surface representation and selected amino acid residues as sticks. The structure images were created with MACPYMOL [71].

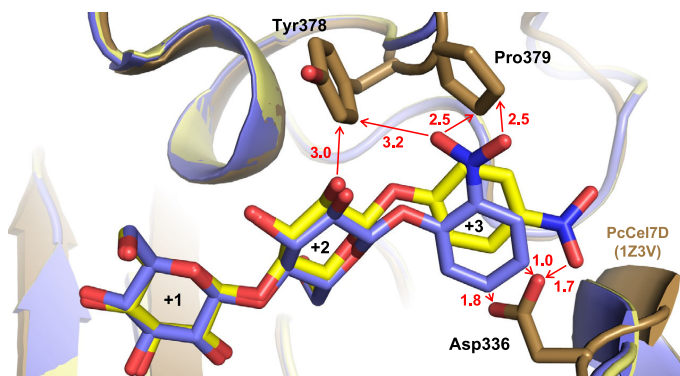
Molecular dynamics simulations

Since enzyme kinetics do not discriminate productive and non-productive binding, and x-ray crystallography only demonstrated the non-productive binding, molecular dynamics (MD) simulations were employed to assess how the substrates would bind in the productive mode (i.e. at subsites –2 to +1). Starting models for productive binding of oNPC, pNPC and pNPL were obtained by taking the glucose residues at subsites –2 and –1 of the TrCel7A Michaelis complex with cellobiose (PDB: 4C4C) and attaching a nitrophenyl group with the glycosidic oxygen in position for protonation by the catalytic acid/base (Glu217), and with the nitrophenyl ring parallel to the Trp376 platform at

subsite +1. For oNPC, two productive binding models were used, pose 1 and pose 2, with the 2-nitro group pointing either “away” or “towards” the catalytic nucleophile (Glu212) and acid/base (Glu217) residues at the catalytic center of the enzyme. As starting models for non-productive binding, we used the crystal structures described herein of TrCel7A in complex with oNPC, pNPC or pNPL bound at subsites +1 to +3. In this case there was only one conformation of oNPC, since only one conformation was seen in the crystal structure. Corresponding models of PcCel7D were obtained by superposition with the crystal structure of PcCel7D in complex with cellobiose (PDB: 1Z3T).

MD simulations were run for 100 ns for TrCel7A and PcCel7D in complex with oNPC, pNPC, and

Fig. 5. The non-productively bound ligands pNPC (yellow; PDB: 4UWT) and oNPC (blue; PDB: 4V0Z) at subsites +1/+2/+3 in TrCel7A superposed on PcCel7D (brown; PDB: 1Z3V) showing the clash between the nitrophenyl group and Asp336 and close contacts with Tyr 378 and Pro379. The structure images were created with MACSYMOL [71].



pNPL in both productive and non-productive binding modes. However, in several complexes, the substrates consistently diffused out of the active site, beginning from 500 ps. Therefore, only the first 500 ps of the production run in the simulations (after equilibration) were used for structural analyses and for computation of ligand-binding free energies with the Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) approach [33]. Fig. 6 shows cluster representations of protein backbone and ligand structures over a 500 ps trajectory for all MD simulations. Snapshots at 500 ps of all the ligand structures are shown in Figs S1 and S2, and binding free energies from the MD simulations in Table 4. Plots of distances between substrate and catalytic amino acids over the entire 100 ns MD runs are shown in Fig. S3.

For the productive binding mode at subsites $-2/-1/+1$, Fig. 7 shows the distances between substrate and catalytic amino acids during the first 1 ns of MD simulations (glycosidic oxygen O1 to the acid/base, and anomeric carbon C1 to the nucleophile, respectively), which indicate larger ligand fluctuations with PcCel7D than with TrCel7A. This is also seen in the cluster representations of backbone and ligand structure in Fig. 6. Snapshots at 500 ps of selected substrates in the productive mode are shown in Fig. 8. For TrCel7A, the sugar units of the substrates remained close to the corresponding glucose units in the Michaelis complex (within 1 Å), and the glucose residue at subsite -1 retained the 1,4B boat conformation. The axial 4OH of pNPL at subsite -2 seems to be readily accommodated without steric hindrance. The nitrophenyl rings lie on the Trp376 platform at subsite +1 and the glycosidic oxygen is within distance to the catalytic acid/base for protonation. However, for oNPC pose 2 the 2-nitro group comes close to and may interfere with the catalytic acid/base Glu217 (Fig. 8A).

Thus, oNPC is less likely to be hydrolyzed when bound in the pose 2 orientation.

With PcCel7D productive mode, the deviation was larger from the starting models and the glucose residues were shifted further upwards from the bottom of the active site (Fig. 8B). In the 500 ps snapshots, the boat conformation of the glucose residue at subsite -1 is only retained in pNPC. In the others it is on the way to a chair in oNPC pose 1, whereas in pNPL and oNPC pose 2 it has flipped from boat and adopts a 4C_1 chair conformation.

In the snapshots of non-productive complexes, the sugar residues overlap at subsites +1 and +2 and the nitrophenyl groups at +3. However, the ligands deviate from each other and from the crystal structures, with up to 2–3 Å distance between corresponding sugar atoms at subsite +1 (TrCel7A/oNPC vs. PcCel7D/pNPL) and up to 5–7 Å between nitrophenyl ring atoms at subsite +3 (TrCel7A/oNPC vs. PcCel7D/pNPC). The MD simulations also show larger flexibility of the protein around the ligands than seen in the crystal structures, such as in the A4 and B4 loop regions that are flanking the +3 subsite (Fig. 6). A tyrosine residue in loop A4 (Tyr381 in TrCel7A, Fig. 4C; Tyr378 in PcCel7D, Fig. 5), which restricts nitrophenyl binding on one side, deviates up to 2 Å at the CA atom and 2.7 Å at OH. The B4 loop on the other side of the nitrophenyl moiety exhibits backbone shifts up to 3.3 Å (TrCel7A Gly339), as well as flexibility of the Asp336 side chain in PcCel7D, although it is consistently pointing towards subsite +3.

The MMPBSA binding free energy calculations do indeed indicate a large difference between the two enzymes (Table 4). As expected, TrCel7A gave more favourable free energy values (-15.0 to -22.0 kcal·mol $^{-1}$) for all substrates and in both binding modes, compared to PcCel7D (-1.6 to -5.0 kcal·mol $^{-1}$), consistent with

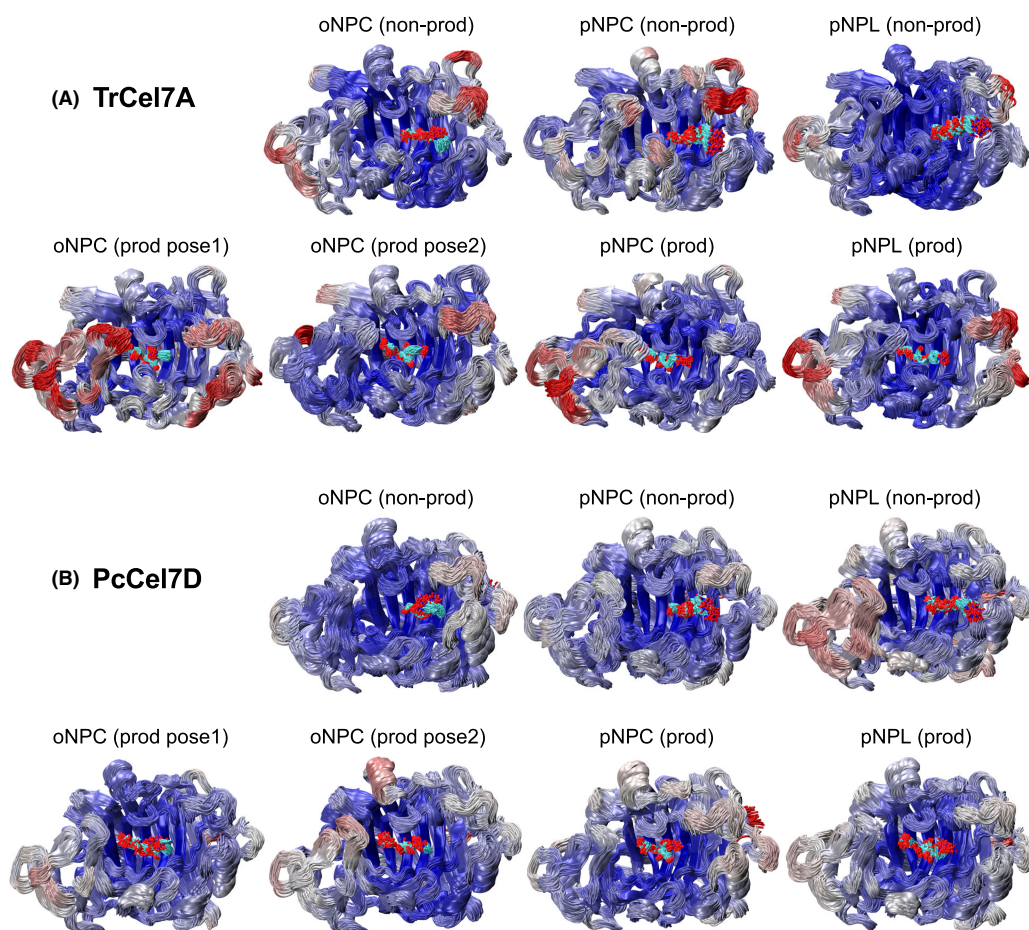


Fig. 6. Cluster representations of (A) TrCel7A and (B) PcCel7D protein backbone and ligand structures shown over a 500 ps MD simulation trajectory. The protein backbones are coloured by RMSF (root mean square fluctuation), where *red* represents the largest fluctuations, and *blue* represents the lowest fluctuations. The structure images were created with VMD [72].

stronger binding of the substrates to TrCel7A than to PcCel7D. Also, binding energies were lower for non-productive than productive binding of pNPC and of pNPL with both enzymes. However, with TrCel7A the binding energies were lower for pNPL than for pNPC, in both binding modes, indicating stronger binding of pNPL than of pNPC, which is contradictory to the results from the enzyme kinetics experiments.

Discussion

Nitrophenyl glycosides are very popular model substrates for glycoside hydrolases, since they provide

good leaving groups and also have favourable spectral properties, making the reactions easy to monitor. Consequently, *p*-nitrophenyl cellobioside and *p*-nitrophenyl lactoside have both found use in cellulase research. It has been found in many cases, though, that the k_{cat} observed has been extremely low, i.e., orders of magnitude lower than that observed for the cleavage of cellulose or cellooligosaccharides. A likely explanation for this phenomenon may be based on the subsite array found in the active site of the enzymes, where model substrates may occupy several alternative positions and unproductive binding may prevent the productive one, leading to a strongly decreased apparent

Table 4. MMPBSA binding free energies for productive and non-productive binding from MD simulation.

Enzyme	Substrate	Productive	Non-productive
		binding in subsites −2 to +1 ($\Delta G_0'$, kcal·Mol ^{−1})	binding in subsites +1 to +3 ($\Delta G_0'$, kcal·Mol ^{−1})
TrCel7A	oNPC	−17.6 ± 0.4/	−14.8 ± 0.4
	(pose1/2)	−18.1 ± 0.4 ^a	
	pNPC	−18.1 ± 0.4	−20.1 ± 0.4
	pNPL	−21.2 ± 0.4	−21.8 ± 0.3
PcCel7D	oNPC	−3.2 ± 0.4/5.2 ± 0.3 ^a	−1.6 ± 0.4
	pNPC	−4.5 ± 0.3	−5.1 ± 0.4
	pNPL	−2.0 ± 0.3	−4.5 ± 0.4

^aTwo values were calculated for oNPC, with the nitro group pointing either “away” (pose 1) or “towards” (pose 2) the catalytic center.

k_{cat} if the non-productive binding is stronger than that at the productive position.

Previous crystal structures have shown that the +1/+2 product binding site is the preferred binding site for cellobiose and lactose in TrCel7A and PcCel7D [28,31,32]. Molecular dynamics simulations of TrCel7A by Knott *et al.* [5] suggested that strong binding to the +1/+2 site is likely an important factor in driving the processive action in the GH7 cellobiohydrolases, making it likely that similar dynamics occur in other enzymes of this class and other processive glycosidases as well [7,34,35]. The crystal structures of TrCel7A with pNPL, pNPC and oNPC bound at the +1/+2/+3 sites we have presented here suggest that non-productive binding is indeed the preferential binding mode for these small model substrates, at least in TrCel7A.

Low apparent K_M -values accompanying the low k_{cat} observed in our experiments and in previous studies also suggest that the slow turnover of these substrates is likely caused by strong non-productive binding, as it is expected to lower both constants by the same factor, with the efficiency constant (k_{cat}/K_M) remaining unaffected (Table 1) [13,14]. For overlapping (competing) binding modes, the apparent kinetic parameters depend on the affinities for productive and non-productive binding as follows (Eqns 1 and 2):

$$k_{\text{cat}}^{\text{app}} = k_{\text{cat}}^{\text{prod}} \times \frac{K_d^{\text{nonprod}}}{K_M^{\text{prod}} + K_d^{\text{nonprod}}} \quad (1)$$

$$K_M^{\text{app}} = K_M^{\text{prod}} \times \frac{K_d^{\text{nonprod}}}{K_M^{\text{prod}} + K_d^{\text{nonprod}}} \quad (2)$$

$k_{\text{cat}}^{\text{app}}$ and K_M^{app} are the apparent catalytic rate and Michaelis–Menten constants, $k_{\text{cat}}^{\text{prod}}$ and K_M^{prod} the intrinsic parameters for productive binding, and K_d^{nonprod} the dissociation constant for non-productive binding. From Eqn (2) follows that K_M^{app} cannot be higher than K_d^{nonprod} (and not K_M^{prod}) if the binding modes overlap (compete). If the two modes have the same binding strength, then $K_M^{\text{app}} = 0.5 * K_d^{\text{nonprod}}$.

The pNPC and pNPL substrates differ only by the orientation of the 4-hydroxyl at the non-reducing end of the molecule, being axial in pNPL and equatorial in pNPC. As expected, they do indeed show about the same k_{cat}/K_M values indicating similar productive binding at −2/−1/+1. This is supported by MD simulations showing that the axial 4OH of pNPL is readily accommodated at subsite −2 without signs of steric hindrance. Yet, both k_{cat} and K_M were much lower for pNPC than for pNPL (about 30-fold and 20-fold, respectively, with TrCel7A, and about 4-fold with PcCel7D). In the case of TrCel7A we now know that both substrates bind preferentially at the non-productive position at subsites +1/+2/+3, which lowers the apparent catalytic constants. Hence, the difference between the substrates is mainly caused by differences in non-productive rather than productive binding. The lower values of k_{cat} and K_M for pNPC show that it is more strongly affected by non-productive binding and binds stronger than pNPL at subsites +1/+2/+3.

In the crystal structures the cellobiose unit of pNPC binds very similarly to cellobiose alone at subsite +1/+2, and the affinity is about the same, as reflected by similar values of K_M for pNPC, and K_d and K_i for cellobiose (26, 23, 24 μM , respectively; Tables 1 and 2). The position of pNPL is practically identical to that of pNPC, except for the orientation of 4OH, which is thus likely the cause of the difference between the substrates. For cellobiose and pNPC, the equatorial 4OH appears to contribute favourably by hydrogen bonding to the catalytic acid/base. With pNPL this hydrogen bond is missing and instead the axial 4OH makes an unfavourably close contact with Gln175 that may rather have a negative effect on the affinity. This is further indicated by the higher apparent K_M for pNPL (590 μM) compared to K_i for lactose (180 μM), which suggests weaker binding of pNPL than of lactose, and by the positional deviation between the lactoside unit of pNPL and lactose alone in the complex structures. The lactose molecule prefers to bind in the “primed” position slightly tilted away from the catalytic centre (Fig. 4A) where 4OH has more space and can make hydrogen bonds with Thr246 and Arg251, whereas pNPL is not tilted to that position, presumably

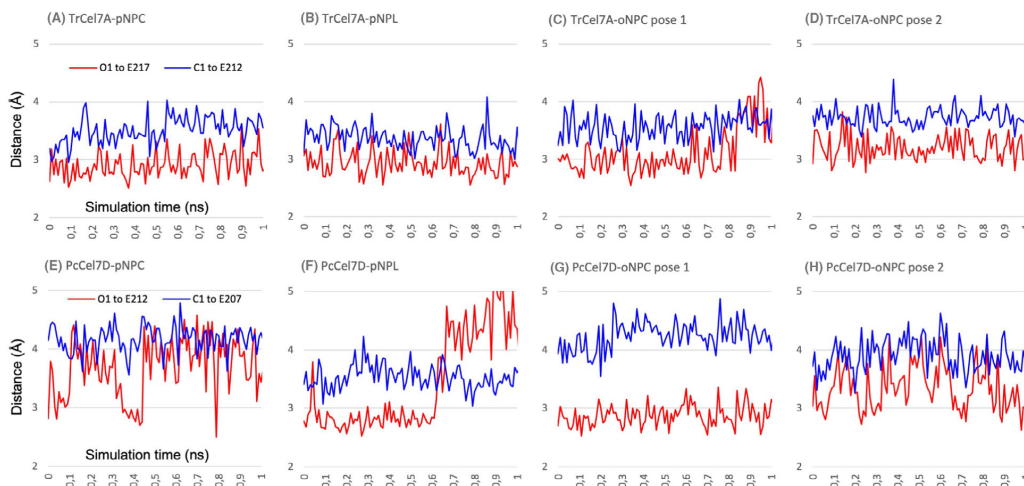


Fig. 7. Distances between substrate and catalytic amino acids during 1 ns of MD simulations of productive binding at subsites $-2/-1/+1$ of pNPC, pNPL and oNPC, in TrCel7A (A–D) and in PcCel7D (E–H). The red line shows the shortest distance from the glycosidic oxygen O1 to the nearest O atom of the catalytic acid/base (Glu/E217 in TrCel7A; Glu/E212 in PcCel7D). The blue line shows the shortest distance from the anomeric carbon C1 to the nearest O atom of the catalytic nucleophile (Glu/E212 in TrCel7A; Glu/E207 in PcCel7D).

hindered by space limitations for the bulky nitrophenyl group at subsite +3.

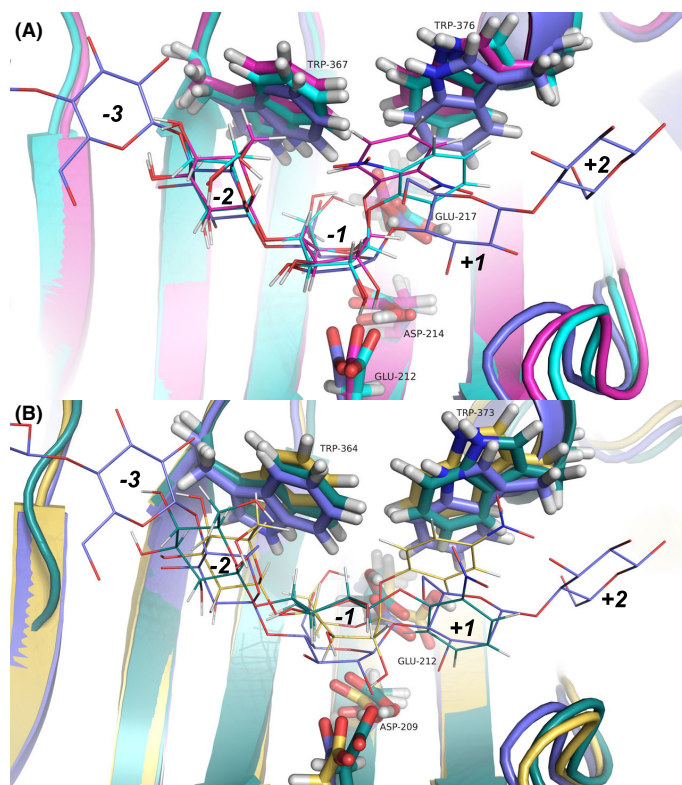
A similar trend is seen with PcCel7D, i.e. similar $k_{\text{cat}}/K_{\text{M}}$ values for pNPC and pNPL suggesting similar productive binding at $-2/-1/+1$, at the same time as the apparent catalytic constants are lower for pNPC, pointing towards stronger non-productive binding than for pNPL. However, PcCel7D shows significantly higher k_{cat} and K_{M} on pNPL than TrCel7A, suggesting much weaker non-productive binding. This could be explained by the different architectures at the +3 site, where Asp336 positioned in the B4-loop of PcCel7D occupies a position where the nitrophenyl-groups in oNPC, pNPC and pNPL are positioned in the TrCel7A structures, likely forcing a different positioning and weaker binding of these substrates in PcCel7D (Fig. 5). The same region in TrCel7A contains two glycine residues, allowing more space and possibly leading to stronger binding at the +3 site. Asp at this location is the most common motif among GH7 CBHs but is missing in TrCel7A and a few closely related CBHs, due to a one-residue deletion in the B4 loop.

Interestingly, TrCel7A mutants with deletions in the B3 loop have shown pNPL hydrolysis kinetics similar to PcCel7D, with higher k_{cat} and K_{M} , and higher K_{i} for cellobiose compared to the wild-type TrCel7A, suggesting this loop plays a role in non-productive

binding and product inhibition [13,36]. Indeed, PcCel7D is missing six residues in this loop region compared to the native TrCel7A (Fig. 2A). At the tip of the B3 loop in TrCel7A, Tyr247 contributes to productive binding through H-bonding with 6OH of the glucose unit at subsite -2 , while Thr246 promotes non-productive binding by H-bonding to 6OH of the sugar unit at subsite +1 (Fig. 2A). The lack of these residues in PcCel7D should result in weaker binding of the substrates, both productive and non-productive binding.

While cellobiose inhibits TrCel7A more strongly than PcCel7D, the results from our lactose inhibition experiments suggest that with lactose the inhibitory effect is more or less equal on both enzymes (Table 2) [13]. In TrCel7A, cellobiose binds strongly at the product sites because of favourable H-bonds with both Glu217 and Thr 246. Lactose binding is weaker due to the clash between the axial 4OH and Gln175, and the loss of the 4OH-Glu217 H-bond when lactose is tilted to the “primed” orientation. In PcCel7D, there is no residue corresponding to Thr246 due to the shorter B3 loop (Fig. 2A), and cellobiose binds weaker in PcCel7D than in TrCel7A. Lactose could be expected to bind even weaker due to the lack of H-bond to the catalytic acid/base Glu217. However, when the lactose molecule is tilted to the “primed” orientation it can make compensating H-bonds with Asp336 in the B4

Fig. 8. Snapshots at 500 ps from MD simulations of productive mode binding of nitrophenyl substrates at subsites $-2/-1/+1$ in TrCel7A and PcCel7D, superposed with the TrCel7A cellobionanase complex PDB: 4C4C (blue). (A) TrCel7A with oNPC pose 1 (cyan) and oNPC pose 2 (magenta). In pose 2 the oNPC is less likely to be hydrolyzed, since the 2-nitro group appears to obstruct protonation of the glycosidic oxygen by the catalytic acid/base Glu217. (B) PcCel7D with pNPC (yellow) and oNPC pose 2 (green). In the oNPC molecule the glucose residue at subsite -1 has flipped from boat to chair conformation. Also, the oNP ring has flipped around so that the 2-nitro group is pointing "away," while it was pointing towards the catalytic center in the starting model. The structure images were created with MACPYMOL [71].



loop. The affinities are actually very similar for cellobiose and lactose with PcCel7D and lactose with TrCel7A, with K_i values of $\sim 180 \mu\text{M}$. Cellobiose binds about one order of magnitude stronger in TrCel7A, probably due to one extra hydrogen bond. The H-bonding partners that differ between enzyme/ligand complexes may be simplified as follows: TrCel7A/cellobiose: Glu217 and Thr246; TrCel7A/lactose: Thr246; PcCel7D/cellobiose: Glu212; PcCel7D/lactose: Asp336.

While the MMPBSA calculations capture the relative differences in binding affinities between TrCel7A and PcCel7D (energy values differ by more than $10.0 \text{ kcal}\cdot\text{mol}^{-1}$), the calculations do not capture the differences between the similar substrates or between productive vs. non-productive binding (some values are within $2.0 \text{ kcal}\cdot\text{mol}^{-1}$). Previous studies have highlighted that MMPBSA may be impractical for comparing ligands with similar affinities due to its low precision [37]. Possibly, more sensitive, computationally intensive estimation methods, such as umbrella sampling, may be needed to evaluate the relative affinities between these substrates.

The structures and MD simulations do not provide straight-forward answers to the question why oNPC is an inferior substrate compared to pNPC (and pNPL). However, the enzyme kinetics, fluorescence titration, and oNPC inhibition studies showed low apparent K_M , K_d and K_i values for oNPC with TrCel7A (7.0 , 7.4 and $5.6 \mu\text{M}$, respectively; Tables 1 and 2), suggesting that strong non-productive binding might at least partially explain the slow hydrolysis in TrCel7A, whereas the higher K_M and K_d values for PcCel7D (3200 and $110 \mu\text{M}$, respectively; Tables 1 and 2) imply that this effect is much less significant in PcCel7D. The fact that PcCel7D fluorescence could not be recovered after oNPC titration by the addition of cellobiose also suggests that there could in fact be another site preferential to the product site for binding oNPC on the enzyme, one which does not have as high affinity for cellobiose as for oNPC. The significantly higher apparent K_M value compared to K_d for oNPC with PcCel7D is in line with this hypothesis, considering that in the case of overlapping productive and non-productive binding modes K_M should not

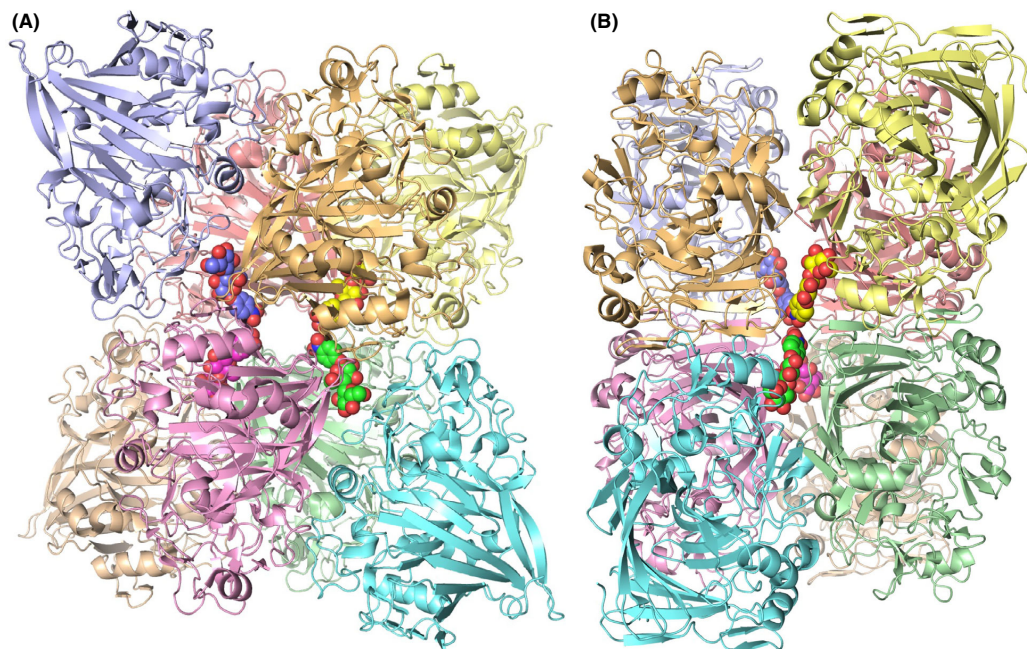


Fig. 9. Crystal packing of protein molecules around the substrate molecules bound at the surface of the TrCel7A enzyme in the crystal structures, viewed along two of the twofold symmetry axes in the crystal. Four substrate molecules bind around the symmetry axes, each making interactions with four surrounding protein molecules. The structure shown is the TrCel7A E212Q/pNPC complex (PDB: 4UWT) with pNPC in space-filling and protein chains in cartoon representation. Colours are arbitrarily chosen for distinction of individual molecules. The structure images were created with MACPYMOL [71].

exceed the K_d value. With TrCel7A the K_M and K_d values for oNPC are roughly equal (instead of $K_M^{app} \leq 0.5 * K_d$ in the case of overlapping binding sites), which implies there is some binding to a location outside the active site in TrCel7A as well, in accordance with previous titration calorimetry results from Colussi *et al.* [38], suggesting stoichiometry of ~ 1.5 binding sites for cellobiose and triose in the E212Q mutant of this enzyme. An additional ligand molecule is indeed seen on the outside of the protein in our crystal structures. However, that binding site is built up by crystal contacts where the ligand is bound by interactions with four neighbouring protein molecules in the crystal lattice (Fig. 9). When the enzyme is free in solution the affinity for this site would most likely be too low to be significant.

When considering the slower hydrolysis of oNPC compared to the other model compounds studied, it also seems clear that there is more space and conformational freedom for a nitro group at the 4-position (as in pNPC), whereas the close proximity of the

2-nitro group in oNPC to the glycosidic oxygen and the catalytic amino acid residues is more likely to interfere with transition-state formation. Such interference would be more pronounced with TrCel7A since the catalytic centre is more enclosed in this enzyme compared to PcCel7D (Fig. 2A).

Another relevant question is why MUC is a much better substrate, as reflected by about an order of magnitude higher catalytic efficiency (k_{cat}/K_M) than pNPC and pNPL. Based on the dynamics of the productively bound substrates at the catalytic centre in the MD simulations, we hypothesize that this is an effect of the larger size of methylumbelliferyl compared to the nitrophenyl group. The methylumbelliferyl aglycone would fill up more of the space available in subsite +1 and will be more firmly bound. That would limit the conformational freedom for the glucose unit at subsite -1 and help to push it towards the catalytic amino acids, thereby increasing the probability to reach and pass the transition state for hydrolysis. Previous computational studies of cellulose hydrolysis and

processivity by TrCel7A have shown that catalytic activation is an essential part of the catalytic mechanism. The enzyme is utilizing binding in the surrounding subsites as handles for bending the cellulose chain so that the glucose unit at subsite -1 will flip from chair to the boat/skew conformational series at the same time as it will be pushed towards the catalytic nucleophile (Glu212 in TrCel7A) [5,30]. The tryptophan residues that serve as sugar binding platforms in the surrounding subsites -2 and $+1$ play an important role by acting as relatively rigid, inelastic surfaces that restrict the conformational freedom of the substrate and promote the glucopyranose ring distortion necessary for catalysis (Fig. 8A). That aromatic-carbohydrate interactions play a role in glucopyranose distortion and TS-stabilization has also been shown in a computational study of the other processive cellobiohydrolase of *T. reesei*, TrCel6A, where aromatic-carbohydrate interactions were examined with molecular simulation [39].

Concluding remarks

In this work we have shown that, at least in the case of TrCel7A, oNPC can be utilized as an active-site probe for fluorometric determination of the dissociation constant for cellobiose and can be used also with catalytically impaired mutants. We have also shown that the enzyme kinetics of GH7 CBHs on the convenient chromogenic substrates pNPC and pNPL is dictated by non-productive binding in the product binding sites rather than productive binding at the catalytic centre. Structural differences distant from the catalytic centre that affect non-productive binding may have large impact on the kinetics, as exemplified by the influence of Asp336 in subsite $+3$ of PcCel7D. Thus, the results of activity assays with these substrates should be interpreted with caution. One-point measurements at a single substrate concentration are still useful for estimation of relative amount or activity of the same enzyme, e.g., to monitor protein purification or pH and temperature dependence, or in protein engineering to improve thermal stability by comparing activity before and after heat treatment (e.g., [40]). However, for comparison of homologues and/or mutants a sufficient range of substrate concentrations is needed so that enzyme kinetics parameters (k_{cat} and K_{M}) can be derived. Very low values of both k_{cat} and K_{M} are indicative of strong product binding, as with TrCel7A. The specificity constant $k_{\text{cat}}/K_{\text{M}}$ is the most instrumental parameter for comparison since it is not affected by non-productive binding but is a measure of the difference in free energy between the substrate in solution and the transition state of hydrolysis.

Materials and methods

Reagents and enzymes

Trichoderma reesei Cel7A and its catalytically inactive mutants E212Q, D214N and E217Q were purified from culture filtrate as described [31,41]. Cel7D from *Phanerochaete chrysosporium* was purified as described in [42]. An additional purification step on Superose 12 gel (Pharmacia) using 0.5 M ammonium sulphate in 50 mM sodium acetate buffer, pH 5.0 was performed for all enzymes used. Catalytic domain of TrCel7A wildtype and E212Q mutant for crystallization were prepared as described [31]. The purity of the enzymes was confirmed by SDS/PAGE. Methylumbelliferyl cellobioside, *o*-nitrophenyl cellobioside, *p*-nitrophenyl cellobioside, *p*-nitrophenyl lactoside, cellobiose and lactose were obtained from Sigma (St. Louis, MO, USA), all other chemicals were of analytical grade.

Ligand binding studies

All experiments were performed in 50 mM sodium acetate buffer, pH 5.0 at 25 °C, unless stated otherwise. Fluorescence of the protein was measured with an Aminco SPF-500 spectrofluorometer. Fluorescence quenching experiments were performed at $\lambda_{\text{ex}} = 280$ nm and $\lambda_{\text{em}} = 340$ nm with excitation band pass at 2 nm and emission band pass at 10 nm. The possible influence of cellobiose to the protein fluorescence was measured. It was found that the presence of cellobiose increases the fluorescence of TrCel7A in a hyperbolic manner with $K_{\text{d}} = 4.5 \pm 1.5$ mM (data not shown). Since our measurements were done at more than 10 times lower concentrations of cellobiose, we can assume that the influence from cellobiose to protein fluorescence is linear (Eqn 4).

Kinetic studies

For kinetic studies with nitrophenyl glycosides as substrates, the enzyme (0.5 or 1 μM) was incubated with substrate at various concentrations, for 2 min with pNPL, 30 min with pNPC, and 18 h (TrCel7A) or 5 h (PcCel7D) with oNPC. The reaction was stopped by the addition of equal volume of 0.1 M sodium hydroxide and the concentration of released pNP or oNP was measured spectrophotometrically at 414 nm, using $\epsilon = 16\,590\text{ M}^{-1}\cdot\text{cm}^{-1}$ for pNP and $\epsilon = 4500\text{ M}^{-1}\cdot\text{cm}^{-1}$ for oNP. For kinetics with methylumbelliferyl cellobioside as a substrate, the enzyme (50 nM TrCel7A or 10 nM PcCel7D) was incubated with various concentrations of the substrate and inhibitor. The reaction was stopped by the addition of equal volume of 0.5 M sodium carbonate. The released product was quantified fluorometrically at $\lambda_{\text{ex}} = 360$ nm and $\lambda_{\text{em}} = 440$ nm using methylumbelliferone as a reference. The kinetic parameters for hydrolysis of oNPC, pNPC, pNPL and MUC were

calculated by nonlinear regression analysis using KYPLOTT software package (KyensLab Inc., Tokyo, Japan).

The inhibition constant of lactose on the TrCel7A and PcCel7D was determined by kinetic measurements using pNPL as a substrate, without and with 0.2 mM lactose as inhibitor. For TrCel7A the pNPL concentrations used were 0.1, 0.2, 0.4, 0.6, 0.8, 1.5, 3, 5, 7 and 10 mM, with an enzyme concentration of 1.5 μM . For PcCel7D the pNPL concentrations used were 1, 2, 3, 4, 5, 8, 10, 14, 17 and 20 mM, with an enzyme concentration of 0.7 μM . The reactions had a volume of 150 μL and were conducted in triplicate in 10 mM sodium acetate buffer pH 5.0 at 20 $^{\circ}\text{C}$ for 30 min and stopped by adding an equal volume of 0.1 M NaOH. Substrate control samples were run with identical pNPL concentrations in triplicate, with enzyme added after the NaOH. The quantity of released pNP was determined as described above. The inhibition constants were determined by non-linear regression using the competitive inhibitor function of GRAPHPAD PRISM 8 (GraphPad Software, San Diego, CA, USA).

Data treatment

Fluorescence quenching data were fitted using nonlinear regression into following equations:

Binding of the quenching ligands to the enzyme (Eqn 3):

$$\frac{F_{\text{oNPC}} - F_0}{F_{\text{sat}} - F_0} = \frac{[\text{oNPC}]}{[\text{oNPC}] + K_{\text{d(oNPC)}}} \quad (3)$$

Here, F_0 is the free protein fluorescence; F_{oNPC} is the observed protein fluorescence at given concentration of the quenching ligand oNPC; F_{sat} represents the fluorescence of the protein-oNPC complex; $[\text{oNPC}]$ is the oNPC concentration; $K_{\text{d(oNPC)}}$ is the dissociation constant for oNPC binding to the protein.

Displacement binding data were fitted into the following equation (Eqn 4):

$$\frac{F_{\text{(oNPC,CB)}} - F_{\text{oNPC}}}{F_0 - F_{\text{oNPC}}} = \frac{[\text{CB}]}{[\text{CB}] + K_{\text{d(CB)}} \times \left(1 + \frac{[\text{oNPC}]}{K_{\text{d(oNPC)}}}\right)} + B \times [\text{CB}] \quad (4)$$

Here, $F_{\text{(oNPC,CB)}}$ is the protein fluorescence at given concentration of cellobiose and oNPC; F_{oNPC} is the protein fluorescence at given concentration of oNPC; $F_0 - F_{\text{oNPC}}$ is the change in fluorescence caused by oNPC; $[\text{CB}]$, concentration of cellobiose; $K_{\text{d(CB)}}$, binding constant of cellobiose to the protein; $[\text{oNPC}]$, oNPC concentration; $K_{\text{d(oNPC)}}$, binding constant for the quenching ligand; $B \times [\text{CB}]$, linear component which takes into account the change of the fluorescence of the protein by adding of cellobiose.

Protein crystallization and structure determination

Crystallization experiments were carried out using the hanging-drop vapour diffusion method [43] by mixing equal amounts of protein (6 $\text{mg}\cdot\text{mL}^{-1}$, in 10 mM sodium acetate, pH 5.0) and reservoir solution [50 or 100 mM morpholinoethane sulphonic acid (Mes), pH 6.0, 21.25% polyethylene glycol 5000 monomethyl ether (m5K), 12.5% glycerol and 5–10 mM cobalt chloride]. Crystals appeared within 1–5 days at room temperature. Ligand soaks with oNPC and pNPC were performed by transferring crystals to hanging drops containing 10 mM oNPC or pNPC, in 0.1 M NaMes, pH 6.0, 25% m5K, 12.5% glycerol and 10 mM CoCl_2 , with a subsequent incubation for 3 h (wild-type with oNPC) or 24 h (E212Q mutant with pNPC) before crystal picking. For pNPL soaking, a few grains of pNPL were added to a drop with TrCel7A E212Q crystals. Individual crystals were picked with 0.1–0.5 mm loops and flash-frozen in liquid nitrogen. Synchrotron x-ray diffraction data were recorded at 100 K. All crystals belong to space group I222 with one protein molecule per asymmetric unit.

X-ray diffraction data for TrCel7A wildtype with oNPC were collected at beamline ID14-3, ESRF, Grenoble, France, and for TrCel7A E212Q with pNPC at beamline I911-2, MAX-lab, Lund, Sweden. The data for wt/oNPC were processed with Denzo and Scalepack [44], and for E212Q/pNPC with Mosflm and Scala [45,46]. Initial phases were obtained from the refined protein coordinates of TrCel7A E212Q in complex with cellobiose (PDB: 3CEL).

Diffraction data for the TrCel7A E212Q structures with lactose and pNPL ligands were collected at the BioMAX beamline at the MAX IV synchrotron in Lund, Sweden, using MXCUBE3 and ISPYB software for data collection and management [47–49]. The data were indexed and integrated through automatic processing with XDS through the EDNA pipeline at MAX IV, and scaled and merged with Aimless either through the EDNA pipeline (TrCel7A E212Q with lactose) or through the CCP4i interface (TrCel7A E212Q with pNPL) [50–54]. This data was used as input reflections into the Dimple-pipeline using the CCP4i2 interface [55]. The peptide chain coordinates from the TrCel7A structure PDB: 4C4C were used as search model input coordinates for Dimple, where a re-indexing of the reflections was performed by Pointless, followed by rigid body refinement and restrained refinement by Refmac5 [56].

All the structures were further refined in several iterative cycles of model building and adjustment in Coot, and restrained refinement in Refmac5 [57]. Statistics from diffraction data processing and structure refinement are summarized in Table 3. The atomic coordinates and experimental structure factor amplitudes have been deposited in the Protein Data Bank with accession codes PDB: 7NYT,

7OC8, 4V0Z, and 4UWT for the lactose, pNPL, oNPC, and pNPC active site ligand structures respectively.

Molecular dynamics simulations and free energy calculations

MD simulations were run for 14 separate systems, which comprised TrCel7A and PcCel7D in complex with oNPC, pNPC, and pNPL, in both productive and non-productive configurations. Two poses were simulated for the oNPC productive complex, with the nitrophenyl group pointing either away from (pose1) or towards (pose2) the catalytic center. Structure models for non-productive binding were derived from the crystal structures presented in this study (TrCel7A models), and by superposition with PcCel7D structure PDB: 1Z3T. The models for productive binding were derived by modifying the TrCel7A/cellononaose Michaelis complex (PDB: 4C4C) [5]. The E217Q mutation in 4C4C was reverted and in all models the wildtype catalytic residues were used. Protonation states of titratable residues were determined by pKa calculations using the H++ webserver with a pH of 5.0 and internal and external dielectrics of 10 and 80 respectively [58,59]. The systems were constructed with CHARMM, a 83 Å × 83 Å × 83 Å water box was added to solvate the system, and sodium ions were added to ensure a net neutral charge. The conformation of protein, carbohydrate, and the nitrophenyl moieties were defined with the CHARMM36 force fields, and water molecules were modelled with the TIP3P force field [60–62].

Minimization, equilibration, and production simulations were conducted with AMBER [63,64]. The CHARMM parameter files were converted to AMBER format with the PARMED package [65]. The minimization routine was performed as follows. First, the protein and ligand were restrained so that only water molecules and ions were minimized for 500 steps. Second, only the protein was fixed, and the ligand and solvent molecules were minimized for 500 steps. Finally, the entire system was minimized without any restraints for 1000 steps. Restraints were achieved by applying a 500 kcal·(mol·Å⁻²)⁻¹ force constant on the desired atoms. In each case, the first 200 steps were performed with the steepest descent method (SD) and the conjugate gradient method was used for the remaining steps. After minimization, the systems were heated from 100 to 300 K over 20 ps with the NVT ensemble and a weak force restraint of 10 kcal·(mol·Å⁻²)⁻¹ on protein and ligand atoms. Subsequently, the systems were equilibrated in the NPT ensemble at 300 K for 200 ps in four equal stages (50 ps each) with gradually decreasing weak restraints. Restraints of 10 and 5 kcal·(mol·Å⁻²)⁻¹ were respectively used in the first two stages on both protein and ligand heavy atoms. In the third stage, restraints of 5 kcal·(mol·Å⁻²)⁻¹ were applied to only ligand heavy atoms, which was followed by a final stage without restraints. The production

run was performed for 100 ns with the NPT ensemble at 300 K. Long-range electrostatics were handled with the Particle mesh Ewald algorithm (PME) [66] and hydrogen distances were fixed with the SHAKE algorithm [67]. MMPBSA calculations were performed with the AMBER-TOOLS package [68] using snapshots from the first 500 ps of the production simulations at an interval of 1 ps. As in previous studies, the ionic strength, external dielectric, and internal dielectric constants were set to 0.15 M, 4.0, and 80.0, respectively [69,70], and the entropy term was excluded in the calculations [70]. Structural visualization and analyses and trajectories were done with PYMOL [71] and VMD [72].

Acknowledgements

We acknowledge MAX IV Laboratory for time on Beamline BioMAX under Proposal 20180025 and thank Uwe Müller and Ana Gonzalez for assistance. Research conducted at MAX IV, a Swedish national user facility, is supported by the Swedish Research council under contract 2018-07152, the Swedish Governmental Agency for Innovation Systems under contract 2018-04969, and Formas under contract 2019-02496. We also thank the staff of the ESRF and EMBL Grenoble for assistance and support in using beamline ID14-3. Funding for the research is gratefully acknowledged from the Swedish Energy Agency (Dnr 2015-009633) and Swedish Natural Science Research Council (NFR). This work was supported in part by the National Science Foundation (NSF) award number 1552355 to CMP in support of JEG. This material is also based upon work supported by (while CMP is serving at) the NSF. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. This work was authored in part by Alliance for Sustainable Energy, LLC, the manager and operator of the National Renewable Energy Laboratory for the U.S. DOE under Contract No. DE-AC36-08GO28308. PV was supported by the Estonian Research Council (Grant PRG1540).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

TH: Solved, refined, validated and deposited protein ligand complex structures (7NYT, 7OC8). Performed lactose inhibition experiments. Analysed protein structures and made protein structure figures. Wrote final

manuscript; JEG: Performed and analysed MD simulations and made corresponding figures. Contributed to final manuscript writing; AN: Performed and analysed enzyme kinetics, inhibition and fluorescence titration experiments and made corresponding figures. Wrote initial manuscript draft (as unpublished manuscript in PhD thesis). Crystallized and solved structures of protein-ligand complexes (4UWT, 4VOZ); NTA: Set up protein structure models and systems for MD simulations; MN: Participated in initial enzyme kinetics and fluorescence titration experiments. MHM: Refined, validated and deposited protein structures (4UWT, 4VOZ); RI: Participated in project planning, experiment design and supervision; PV: Participated in enzyme kinetics and fluorescence titration methods development. Contributed to final manuscript writing; GJ: Initiated and conceptualized the investigation. PhD supervisor for Anu Nutt. Wrote final manuscript; CMP: Initiated and lead the MD simulations and subsequent analyses. JS: Co-initiator of the investigation. Produced and purified proteins. Performed protein crystallization, x-ray data collection and protein structure analysis. Wrote and coordinated the final manuscript.

Data availability statement

The atomic coordinates and structure factors of the TrCel7A structures have been deposited into the Protein Data Bank with accession codes PDB: 4UWT, 4VOZ, 7NYT, and 7OC8, respectively, for the pNPC, pNPL, lactose and oNPC ligand complexes.

References

- Payne CM, Knott BC, Mayes HB, Hansson H, Himmel ME, Sandgren M, et al. Fungal cellulases. *Chem Rev.* 2015;**115**:1308–448.
- Igarashi K, Koivula A, Wada M, Kimura S, Penttilä M, Samejima M. High speed atomic force microscopy visualizes processive movement of *Trichoderma reesei* cellobiohydrolase I on crystalline cellulose. *J Biol Chem.* 2009;**284**:36186–90.
- Lee I, Evans BR, Woodward J. The mechanism of cellulase action on cotton fibers: evidence from atomic force microscopy. *Ultramicroscopy.* 2000;**82**:213–21.
- Divne C, Ståhlberg J, Teeri TT, Jones TA. High-resolution crystal structures reveal how a cellulose chain is bound in the 50 Å long tunnel of cellobiohydrolase I from *Trichoderma reesei*. *J Mol Biol.* 1998;**275**:309–25.
- Knott BC, Crowley MF, Himmel ME, Ståhlberg J, Beckham GT. Carbohydrate-protein interactions that drive processive polysaccharide translocation in enzymes revealed from a computational study of cellobiohydrolase processivity. *J Am Chem Soc.* 2014;**136**:8810–9.
- Olsen JP, Kari J, Windahl MS, Borch K, Westh P. Molecular recognition in the product site of cellobiohydrolase Cel7A regulates processive step length. *Biochem J.* 2020;**477**:99–110.
- Payne CM, Jiang W, Shirts MR, Himmel ME, Crowley MF, Beckham GT. Glycoside hydrolase processivity is directly related to oligosaccharide binding free energy. *J Am Chem Soc.* 2013;**135**:18831–9.
- Claeysens M, Van Tilbeurgh H, Tomme P, Wood TM, McRae SI. Fungal cellulase systems. Comparison of the specificities of the cellobiohydrolases isolated from *Penicillium pinophilum* and *Trichoderma reesei*. *Biochem J.* 1989;**261**:819–25.
- Gruno M, Våljamäe P, Pettersson G, Johansson G. Inhibition of the *Trichoderma reesei* cellulases by cellobiose is strongly dependent on the nature of the substrate. *Biotechnol Bioeng.* 2004;**86**:503–11.
- van Tilbeurgh H, Claeysens M, de Bruyne CK. The use of 4-methylumbelliferyl and other chromophoric glycosides in the study of cellulolytic enzymes. *FEBS Lett.* 1982;**149**:152–6.
- Rabinovich ML, Melnik MS, Herner ML, Voznyi YV, Vasilchenko LG. Predominant nonproductive substrate binding by fungal Cellobiohydrolase I and implications for activity improvement. *Biotechnol J.* 2019;**14**:1–17.
- Becker D, Johnson KSH, Koivula A, Schüle M, Sinnott ML. Hydrolyses of α - and β -cellobiosyl fluorides by Cel6A (cellobiohydrolase II) of *Trichoderma reesei* and *Humicola insolens*. *Biochem J.* 2000;**345**:315–9.
- Von Ossowski I, Ståhlberg J, Koivula A, Piens K, Becker D, Boer H, et al. Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. A comparison with *Phanerochaete chrysosporium* Cel7D. *J Mol Biol.* 2003;**333**:817–29.
- Kuusk S, Våljamäe P. When substrate inhibits and inhibitor activates: implications of β -glucosidases. *Biotechnol Biofuels.* 2017;**10**:1–15.
- Kubala M, Plásek J, Amler E. Fluorescence competition assay for the assessment of ATP binding to an isolated domain of Na⁺, K⁽⁺⁾-ATPase. *Physiol Res.* 2004;**53**:109–13.
- Lin M, Nielsen K. Binding of the *Brucella abortus* lipopolysaccharide O-chain fragment to a monoclonal antibody. Quantitative analysis by fluorescence quenching and polarization. *J Biol Chem.* 1997;**272**:2821–7.
- Zhang S, Irwin DC, Wilson DB. Site-directed mutation of noncatalytic residues of *Thermobifida fusca* exocellulase Cel16B. *Eur J Biochem.* 2000;**267**:3101–15.
- Zhou W, Irwin DC, Escovar-Kousen J, Wilson DB. Kinetic studies of *Thermobifida fusca* Cel9A active site mutant enzymes. *Biochemistry.* 2004;**43**:9655–63.

- 19 van Tilbeurgh H, Loontjens FG, Engelborgs Y, Claeysens M. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414: binding of small ligands to the 1,4- β -glucan cellobiohydrolase II and influence of glucose on their affinity. *Eur J Biochem.* 1989;**184**:553–9.
- 20 van Tilbeurgh H, Pettersson G, Bhikabhai R, Boeck H, Claeysens M. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Reaction specificity and thermodynamics of interactions of small substrates and ligands with the 1,4-beta-glucan cellobiohydrolase II. *Eur J Biochem.* 1985;**148**:329–34.
- 21 Barr BK, Wolfgang DE, Piens K, Claeysens M, Wilson DB. Active-site binding of glycosides by *Thermomonospora fusca* endocellulase E2. *Biochemistry.* 1998;**37**:9220–9.
- 22 Cogswell LP, Raines DE, Parekh S, Jonas O, Maggio JE, Strichartz GR. Development of a novel probe for measuring drug binding to the F1*S variant of human alpha 1-acid glycoprotein. *J Pharm Sci.* 2001;**90**:1407–23.
- 23 Nordenman B, Danielsson Å, Björk I. The binding of low-affinity and high-affinity heparin to antithrombin: fluorescence studies. *Eur J Biochem.* 1978;**90**:1–6.
- 24 Westerlund B, Saarinen M, Person B, Ramaswamy S, Eaker D, Eklund H. Crystallographic investigation of the dependence of calcium and phosphate ions for notexin. *FEBS Lett.* 1997;**403**:51–6.
- 25 Røjel N, Kari J, Sørensen TH, Badino SF, Morth JP, Schaller K, et al. Substrate binding in the processive cellulase Cel7A: transition state of complexation and roles of conserved tryptophan residues. *J Biol Chem.* 2020;**295**:1454–63.
- 26 Haddad Momeni M, Payne CM, Hansson H, Mikkelsen NE, Svedberg J, Engström A, et al. Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidium irregulare*. *J Biol Chem.* 2013;**288**:5861–72.
- 27 Muñoz IG, Ubhayasekera W, Henriksson H, Szabó I, Pettersson G, Johansson G, et al. Family 7 cellobiohydrolases from *Phanerochaete chrysosporium*: crystal structure of the catalytic module of Cel7D (CBH58) at 1.32 Å resolution and homology models of the isozymes. *J Mol Biol.* 2001;**314**:1097–111.
- 28 Ubhayasekera W, Muñoz IG, Vasella A, Ståhlberg J, Mowbray SL. Structures of *Phanerochaete chrysosporium* Cel7D in complex with product and inhibitors. *FEBS J.* 2005;**272**:1952–64.
- 29 Kurasin M, Väljamäe P. Processivity of cellobiohydrolases is limited by the substrate. *J Biol Chem.* 2011;**286**:169–77.
- 30 Knott BC, Haddad Momeni M, Crowley MF, MacKenzie LF, Götz AW, Sandgren M, et al. The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *J Am Chem Soc.* 2014;**136**:321–9.
- 31 Ståhlberg J, Divne C, Koivula A, Piens K, Claeysens M, Teeri TT, et al. Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from *Trichoderma reesei*. *J Mol Biol.* 1996;**264**:337–49.
- 32 Haddad Momeni M, Ubhayasekera W, Sandgren M, Ståhlberg J, Hansson H. Structural insights into the inhibition of cellobiohydrolase Cel7A by xylo-oligosaccharides. *FEBS J.* 2015;**282**:2167–77.
- 33 Kollman PA, Massova I, Reyes C, Kuhn B, Huo S, Chong L, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res.* 2000;**33**:889–97.
- 34 Kurasin M, Kuusk S, Kuusk P, Sørleie M, Väljamäe P. Slow off-rates and strong product binding are required for processivity and efficient degradation of recalcitrant chitin by family 18 chitinases. *J Biol Chem.* 2015;**290**:29074–85.
- 35 Nakamura A, Okazaki KI, Furuta T, Sakurai M, Iino R. Processive chitinase is Brownian monorail operated by fast catalysis after peeling rail from crystalline chitin. *Nat Commun.* 2018;**9**:3814.
- 36 Schiano-di-Cola C, Røjel N, Jensen K, Kari J, Sørensen TH, Borch K, et al. Systematic deletions in the cellobiohydrolase (CBH) Cel7A from the fungus *Trichoderma reesei* reveal flexible loops critical for CBH activity. *J Biol Chem.* 2019;**294**:1807–15.
- 37 Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov.* 2015;**10**:449–61.
- 38 Colussi F, Sorensen TH, Alasepp K, Kari J, Cruys-Bagger N, Windahl MS, et al. Probing substrate interactions in the active tunnel of a catalytically deficient cellobiohydrolase (Cel7). *J Biol Chem.* 2015;**290**:2444–54.
- 39 Payne CM, Bomble YJ, Taylor CB, McCabe C, Himmel ME, Crowley MF, et al. Multiple functions of aromatic-carbohydrate interactions in a processive cellulase examined with molecular simulation. *J Biol Chem.* 2011;**286**:41028–35.
- 40 Goedegebuur F, Dankmeyer L, Gualfetti P, Karkehabadi S, Hansson H, Jana S, et al. Improving the thermal stability of cellobiohydrolase Cel7A from *Hypocrea jecorina* by directed evolution. *J Biol Chem.* 2017;**292**:17418–30.
- 41 Bhikhabhai R, Johansson G, Pettersson G. Isolation of cellulolytic enzymes from *Trichoderma reesei* QM 9414. *J Appl Biochem.* 1984;**6**:336–45.
- 42 Uzcategui E, Raices M, Montesino R, Johansson G, Pettersson G, Eriksson KE. Pilot-scale production and purification of the cellulolytic enzyme system from the

- white-rot fungus *Phanerochaete chrysosporium*. *Biotechnol Appl Biochem*. 1991;**13**:323–34.
- 43 McPherson A. Preparation and analysis of protein crystals. New York, NY: Wiley; 1982.
- 44 Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol*. 1997;**276**:307–26.
- 45 Battice TGG, Kontogiannis L, Johnson O, Powell HR, Leslie AGW. iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr Sect D Biol Crystallogr*. 2011;**67**:271–81.
- 46 Evans P. Scaling and assessment of data quality. *Acta Crystallogr Sect D Biol Crystallogr*. 2006;**62**:72–82.
- 47 Ursby T, Hnberg KA, Appio R, Aurelius O, Barczyk A, Bartalesi A, et al. BioMAX the first macromolecular crystallography beamline at MAX IV laboratory. *J Synchrotron Radiat*. 2020;**27**:1415–29.
- 48 Mueller U, Thunnissen M, Nan J, Eguiraun M, Bolmsten F, Milán-Otero A, et al. MXCuBE3: a new era of MX-beamline control begins. *Synchrotron Radiat News*. 2017;**30**:22–7.
- 49 Delagenière S, Brechereau P, Launer L, Ashton AW, Leal R, Veyrier S, et al. ISPyB: an information management system for synchrotron macromolecular crystallography. *Bioinformatics*. 2011;**27**:3186–92.
- 50 Kabsch W. XDS. *Acta Crystallogr Sect D Biol Crystallogr*. 2010;**66**:125–32.
- 51 Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. *Acta Crystallogr Sect D Biol Crystallogr*. 2011;**67**:235–42.
- 52 Evans PR, Murshudov GN. How good are my data and what is the resolution? *Acta Crystallogr Sect D Biol Crystallogr*. 2013;**69**:1204–14.
- 53 Incardona MF, Bourenkov GP, Levik K, Pieritz RA, Popov AN, Svensson O. EDNA: a framework for plugin-based applications applied to X-ray experiment online data analysis. *J Synchrotron Radiat*. 2009;**16**:872–9.
- 54 Potterton E, Briggs P, Turkenburg M, Dodson E. A graphical user interface to the CCP4 program suite. *Acta Crystallogr Sect D Biol Crystallogr*. 2003;**59**:1131–7.
- 55 Potterton L, Agirre J, Ballard C, Cowtan K, Dodson E, Evans PR, et al. CCP 4 i 2: the new graphical user interface to the CCP 4 program suite. *Acta Crystallogr Sect D Struct Biol*. 2018;**74**:68–84.
- 56 Kovalevskiy O, Nicholls RA, Long F, Carlon A, Murshudov GN. Overview of refinement procedures within REFMAC 5: utilizing data from different sources. *Acta Crystallogr Sect D Struct Biol*. 2018;**74**:215–27.
- 57 Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of coot. *Acta Crystallogr Sect D Biol Crystallogr*. 2010;**66**:486–501.
- 58 Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res*. 2005;**33**:368–71.
- 59 Anandakrishnan R, Aguilar B, Onufriev AV. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res*. 2012;**40**:537–41.
- 60 Best RB, Zhu X, Shim J, Lopes PEM, Mittal J, Feig M, et al. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ , ψ and side-chain χ_1 and χ_2 dihedral angles. *J Chem Theory Comput*. 2012;**8**:3257–73.
- 61 Guvench O, Mallajosyula SS, Raman EP, Hatcher E, Vanommeslaeghe K, Foster TJ, et al. CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate-protein modeling. *J Chem Theory Comput*. 2011;**7**:3162–80.
- 62 Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;**79**:926–35.
- 63 Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, DeBolt S, et al. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput Phys Commun*. 1995;**91**:1–41.
- 64 Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Brozell SR, Cerutti DS, et al. Amber 2021. San Francisco, CA: University of California; 2021. p. 957.
- 65 Shirts MR, Klein C, Swails JM, Yin J, Gilson MK, Mobley DL, et al. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J Comput Aided Mol Des*. 2017;**31**:147–61.
- 66 Darden T, York D, Pedersen L. Particle mesh Ewald: an N-log(N) method for Ewald sums in large systems. *J Chem Phys*. 1993;**98**:10089–92.
- 67 Ryckaert J-P, Ciccoliti G, Berendsen HJ. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J Comput Phys*. 1977;**23**:327–41.
- 68 Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA.py: an efficient program for end-state free energy calculations. *J Chem Theory Comput*. 2012;**8**:3314–21.
- 69 Wang Y, Song X, Zhang S, Li J, Shu Z, He C, et al. Improving the activity of *Trichoderma reesei* cel7B through stabilizing the transition state. *Biotechnol Bioeng*. 2016;**113**:1171–7.
- 70 Hou T, Wang J, Li Y, Wang W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy

calculations based on molecular dynamics simulations. *J Chem Inf Model.* 2011;**51**:69–82.

71 Schrödinger L. The PyMOL molecular graphics system 1.5.0.4. 2010.

72 Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph.* 1996;**14**:33–8.

Supporting information

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

Fig. S1. Snapshots at 500 ps of MD simulation of substrate binding in productive mode at subsites $-2/-1/+1$ in TrCel7A and PcCel7D.

Fig. S2. Snapshots at 500 ps of MD simulation of substrate binding in non-productive mode at subsites $+1/+2/+3$ in TrCel7A and PcCel7D.

Fig. S3. Plots of distances between substrate and catalytic amino acids during 10 ns and 100 ns of MD simulations of productive binding mode at subsites $-2/-1/+1$ of pNPC, pNPL and oNPC, in TrCel7A and in PcCel7D.

ACTA UNIVERSITATIS AGRICULTURAE SUECIAE

DOCTORAL THESIS NO. 2023:9

Enzymes from the glycoside hydrolase family 7 constitute the most abundant components in industrial enzyme cocktails for degradation of cellulosic biomass, and also play an important role in the carbon cycle on earth. In this thesis, a study of commonly used model compounds reveals reasons behind the peculiar kinetics on these enzymes, the first structure of such an enzyme from termite symbiont protozoa is presented, and the use of two novel imaging techniques for studying these cellulases is demonstrated.

Topi Haataja received his graduate education at the Department of Molecular Sciences at SLU, Uppsala. He received his Bachelor's degree in Biochemistry from the University of Oulu, Finland, and his Master's degree in Biotechnology from the University of Helsinki, Finland.

Acta Universitatis Agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

ISSN 1652-6880

ISBN (print version) 978-91-8046-070-5

ISBN (electronic version) 978-91-8046-071-2