PUBLICATIONS OF
THE UNIVERSITY OF EASTERN FINLAND


**Dissertations in Forestry and Natural Sciences**

**JOHAN PÄÄKKÖNEN**

# Metal and ligand binding properties of two enzymes in carbohydrate metabolism

*Escherichia coli* D-2-deoxyribose-5-phosphate aldolase and
*Caulobacter crescentus* xylonolactonase

PUBLICATIONS OF THE UNIVERSITY OF EASTERN FINLAND
DISSERTATIONS IN FORESTRY AND NATURAL SCIENCES

N:o 443

*Johan Pääkkönen*

# METAL AND LIGAND BINDING PROPERTIES OF TWO ENZYMES IN CARBOHYDRATE METABOLISM

## *ESCHERICHIA COLI* D-2-DEOXYRIBOSE-5-PHOSPHATE ALDOLASE AND *CAULOBACTER CRESCENTUS* XYLONOLACTONASE

ACADEMIC DISSERTATION

To be presented by the permission of the Faculty of Science and Forestry for public examination in the Auditorium N100 in Natura Building at the University of Eastern Finland, Joensuu, on 30th November, 2021, at 12 o'clock.

University of Eastern Finland
Department of Chemistry
Joensuu 2021

PunaMusta Oy
Joensuu, 2021
Editors: Pertti Pasanen, Nina Hakulinen

Distribution:
University of Eastern Finland Library / Sales of publications
julkaisumyynti@uef.fi
https://www.uef.fi/kirjasto

ISBN: 978-952-61-4386-6 (print)
ISSNL: 1798-5668
ISSN: 1798-5668
ISBN: 978-952-61-4387-3 (pdf)
ISSNL: 1798-5668
ISSN: 1798-5676

Author's address: Johan Pääkkönen
University of Eastern Finland
Department of Chemistry
P.O. Box 111
80101 JOENSUU, FINLAND
Email: johan.paakkonen@uef.fi

Supervisors: Professor Juha Rouvinen
University of Eastern Finland
Department of Chemistry
P.O. Box 111
80101 JOENSUU, FINLAND
Email: juha.rouvinen@uef.fi

Professor Janne Jänis
University of Eastern Finland
Department of Chemistry
P.O. Box 111
80101 JOENSUU, FINLAND
Email: janne.janis@uef.fi

Docent Nina Hakulinen
University of Eastern Finland
Department of Chemistry
P.O. Box 111
80101 JOENSUU, FINLAND
Email: nina.hakulinen@uef.fi

Reviewers: Professor Tiina A. Salminen
Åbo Akademi
Faculty of Science and Engineering
Structural Bioinformatics Laboratory
Tykistökatu 6 A
20520 TURKU, FINLAND
Email: tiina.salminen@abo.fi

Head of the group, Doctor Petr Novák
Institute of Microbiology of the CAS, v. v. i.
Division BIOCEV
Laboratory of Structural Biology and Cell Signaling
Průmyslová 595
252 50 VESTEC, CZECH REPUBLIC
Email: pnovak@biomed.cas.cz

Opponent: Professor Lari Lehtiö
University of Oulu
Faculty of Biochemistry and Molecular Medicine
P.O. Box 5400
90014 OULUN YLIOPISTO, FINLAND
Email: lari.lehtio@oulu.fi

Johan Pääkkönen
Metal and ligand binding properties of two enzymes in carbohydrate metabolism: *Escherichia coli* D-2-deoxyribose-5-phosphate aldolase and *Caulobacter crescentus* xylonolactonase
Joensuu: University of Eastern Finland, 2021
Publications of the University of Eastern Finland
Dissertations in Forestry and Natural Sciences

## ABSTRACT

In this work, two enzymes involving reactions of carbohydrate derivatives were investigated: the D-2-deoxyribose-5-phosphate aldolase from *Escherichia coli* (*Ec*DERA, EC 4.1.2.4) and the xylonolactonase from *Caulobacter crescentus* (*Cc*XylC, EC 3.1.1.68). The *Ec*DERA catalyses the reversible aldol reaction from acetaldehyde and glyceraldehyde-3-phosphate to D-2-deoxyribose-5-phosphate, and the *Cc*XylC catalyses the hydrolysis of D-xylonolactone to D-xylonic acid. The *Cc*XylC was characterised using high-resolution mass spectrometry, and three-dimensional complex structures of three *Ec*DERA mutants and the *Cc*XylC were determined using X-ray crystallography. In addition, a simulation tool for visualising simple protein association equilibria was developed.

The activity of the *Ec*DERA was known to be affected by directed mutations, and for the three investigated mutants, the mutations could be verified and their probable effects deduced from the crystal structures. Remarkably, unlike in any published DERA crystal structure thus far, acetaldehyde and glyceraldehyde-3-phosphate were seen in the active site and modelled with high certainty in one of the complex structures. These results shed light on why the *Ec*DERA mutants have drastically different activities compared to the wild-type.

The *Cc*XylC was discovered to bind iron(II) specifically and with high affinity and no other metal ions in significant amounts, except for some nonspecific copper. Analysis of reaction kinetics with D-xylono- and D-gluconolactone showed that the *Cc*XylC requires the iron(II) ion to be active, but the presence of only iron(II) ions speeds up the nonenzymatic reactions significantly as well. The complex structures of the *Cc*XylC showed that a substrate with a six-membered ring is bound to the active site more strongly and more specifically than a five-membered ring. Reaction mechanisms for the lactone form interconversion and the enzymatic reaction were proposed based on the data. How the *Cc*XylC would benefit from directed mutations remains unknown until new research is done.

These discoveries are a few small steps towards the potential industrial exploitation of these enzymes for production of reagents, materials and biofuels. The *Ec*DERA can be used for building carbon backbones with the aldol reaction, and the *Cc*XylC can be used in metabolic pathways that use D-xylose, a very abundant sugar, as starting material. Both will be useful in the future when perfected and made available in industrial scale.

***Universal Decimal Classification:*** *543.51, 543.645.4, 543.442.3*
***Library of Congress Subject Headings:*** *Enzymes, Catalysis, Mass spectrometry, X-ray crystallography, Proteins, Thermodynamics, Crystallization, Structure*
***Keywords:*** *Crystal structure, Enzyme catalysis, Native mass spectrometry, Reaction thermodynamics, Structural biology, X-ray crystallography,* D*-2-Deoxyribose-5-phosphate aldolase, Xylonolactonase*

## ACKNOWLEDGEMENTS

This work began in 2016 when I started the practical work of my Master's thesis and was finalised in 2021 during the COVID-19 pandemic. I am thankful to the Department of Chemistry and the University of Eastern Finland for providing the opportunity to study chemistry and to participate in academic research. I also appreciate the funding for the projects provided by the Doctoral Programme in Science, Technology and Computing (SCITECO) and the Academy of Finland.

I am deeply grateful to my main supervisor, Prof. Juha Rouvinen, for accepting me to his academic projects, for arranging most of the funding and for always giving me guidelines and advice when I needed them. I thank my other supervisors, Prof. Janne Jänis and Doc. Nina Hakulinen, for always being supportive in all practical and literal work and for all the more or less academic conversations over the years, as well as Dr. Leena Penttinen, Markus Eronen and Joonas Rautio for directly contributing to the work. Also, great thanks to Dr. Martina Andberg, Dr. Harry Boer, Dr. Anu Koivula, Dr. Hannu Maaheimo and others at VTT for providing the protein samples and for participating in the preparation of Publications **I**–**III**.

I appreciate the reviewers of this thesis, Prof. Tiina A. Salminen and Dr. Petr Novák, for reviewing the manuscript on a tight schedule, and my opponent, Prof. Lari Lehtiö, for taking on the task on an even tighter one. I really wish that I could have given you and myself more time to work with, but this tight schedule was effectively forced by circumstances independent on our actions. Thank you all for your effort and patience.

The Department has been a pleasant working environment, and I have enjoyed every year of working with other students and staff. In particular, I am glad to have been able to work and share knowledge with Dr. Timo Kekäläinen, Dr. Mikko Laitaoja, Dr. Marko Mäkinen, Dr. Merja Niemi, Dr. Tarja Parkkinen, Dr. Mohammad Mubinur Rahman, Dr. Chiara Rutanen, Dr. Senthil Kumar Thangaraj and Veikko Eronen. Everyone else not mentioned have also been a part of making the years at the University the best time of my life so far. I will never forget them and all of you.

Finally, I want to thank my family for all the support over the years. My father Pertti, my mother Kirsi and my sisters Jemina and Julia have been invaluable guides and companions throughout my life and my studies. Regarding this work, I especially appreciate all the academic discussions that I have had with my father, which I believe have helped me to become a decent scientist. Last but not least, I thank my girlfriend Mia-Maria for unconditional love, for emotional support and patience during the writing of this thesis and for being a welcome distraction so that I would not burn myself out.

Joensuu, 30th November, 2021

*Johan Pääkkönen*

## ABBREVIATIONS

| | |
|---|---|
| 4H2PD | (R)-4-Hydroxy-2-pyrrolidone |
| ASU | Asymmetric unit |
| *Cc* | *Caulobacter crescentus* |
| DERA | D-2-Deoxyribose-5-phosphate aldolase |
| DLS | Diamond Light Source |
| DNA | Deoxyribonucleic acid |
| DR | D-2-Deoxyribose |
| DRP | D-2-Deoxyribose-5-phosphate |
| EC | Enzyme Commission number |
| *Ec* | *Escherichia coli* |
| ESI | Electrospray ionisation |
| ESRF | European Synchrotron Radiation Facility |
| FT-ICR | Fourier transform ion cyclotron resonance |
| G3H | Glyceraldehyde-3-phosphate |
| MS | Mass spectrometry / Mass spectrometer |
| PEG | Polyethylene glycol |
| QIT | Quadrupole ion trap |
| RF | Radio frequency |
| RMSD | Root mean square deviation |
| SMP30 | Senescence marker protein 30 |
| TIM | Triosephosphate isomerase |
| VTT | Technical Research Centre of Finland Ltd |
| XylC | Xylonolactonase |

## LIST OF PUBLICATIONS

This thesis consists of the present review of the author's work in the field of biological chemistry and the following selection of the author's publications as well as two unpublished protein crystal structures:

**I** S. Voutilainen, M. Heinonen, M. Andberg, E. Jokinen, H. Maaheimo, J. Pääkkönen, N. Hakulinen, J. Rouvinen, H. Lähdesmäki, S. Kaski, J. Rousu, M. Penttilä, and A. Koivula, "Substrate specificity of 2-deoxy-D-ribose 5-phosphate aldolase (DERA) assessed by different protein engineering and machine learning methods," *Appl. Microbiol. Biotechnol.* **104**, 10515–10529 (2020).

**II** J. Pääkkönen, L. Penttinen, M. Andberg, A. Koivula, N. Hakulinen, J. Rouvinen, and J. Jänis, "Xylonolactonase from *Caulobacter crescentus* is a mononuclear nonheme iron hydrolase," *Biochemistry* **60**, 3046–3049 (2021).

**III** J. Pääkkönen, N. Hakulinen, M. Andberg, A. Koivula, and J. Rouvinen, "Three-dimensional structure of xylonolactonase from *Caulobacter crescentus*: a mononuclear iron enzyme of the 6-bladed β-propeller hydrolase family," *Protein Sci.*, doi: 10.1002/pro.4229 (2021).

**IV** J. Pääkkönen, J. Jänis, and J. Rouvinen, "Simulation of binding in protein studies," submitted for publication.

Throughout the overview, these papers will be referred to by Roman numerals.

## AUTHOR'S CONTRIBUTION

**I** The author did all crystallisation experiments and crystal structure determination in supervision of N. Hakulinen and J. Rouvinen and participated in the preparation of the corresponding parts of the publication.

**II** The author did all the experimental work in collaboration with L. Penttinen and in supervision of N. Hakulinen, J. Rouvinen and J. Jänis and participated in the preparation of the publication.

**III** The author did all the experimental work in supervision of N. Hakulinen and J. Rouvinen and participated in the preparation of the publication.

**IV** The author did all software development in collaboration with J. Rouvinen and participated in the preparation of the publication.

# TABLE OF CONTENTS

# 1 INTRODUCTION

## 1.1 CARBOHYDRATES

Carbohydrates and their derivatives are among the most common biomolecules. They have various functions in nature, including energy storage (D-glucose, glycogen and starch), supporting structures (cellulose and chitin) [1] and nucleic acids (D-ribose in ribonucleic acid (RNA) and D-2-deoxyribose (DR) in deoxyribonucleic acid (DNA)) [2]. They are formed most notably by photosynthesis in chloroplasts, which turns carbon dioxide and water into carbohydrate and oxygen: [3]

$$n\,CO_2 + n\,H_2O \xrightarrow{h\nu} \longrightarrow (CH_2O)_n + n\,O_2. \tag{1.1}$$

Polymers of carbohydrates, or polysaccharides, are the most abundant biopolymers in nature. Lignocellulose (plant dry matter) consisting of cellulose, hemicelluloses and lignin is the most abundant biomaterial and readily available nearly everywhere as raw material and waste material from wood and crop industry. The contemporary industrial focus is on developing methods for converting components of lignocellulose to biofuels, other chemicals and polymers, which are currently dependent on fossil sources [4].

Cellulose is the main component of lignocellulose, and it consists of repeating glucose units connected by β-1,4-glycosidic bonds to form long linear chains. Hemicelluloses are more heterogeneous: other monosaccharides (including xylose, mannose and galactose) and their oxidised acid forms are also possible, and the chains are generally shorter and more branched. Lignin consists of not carbohydrates, but aromatic monolignols interlinked in an irregular manner. All these form together a strong, solid medium stabilised by numerous hydrogen bonds and covalent bonds between lignin and hemicellulose [5].

After D-glucose, the second most abundant monosaccharide in lignocellulose from most sources is D-xylose [6, 7]. It is similar in structure and stereochemistry to D-glucose, but the sixth carbon is absent. Structural formulae visualising this are shown in Figure 1.1. D-Xylose is present in various hemicelluloses, including xylan. Because hemicelluloses are short and amorphous, they can easily be hydrolysed using acids, bases or hemicellulase enzymes and extracted from lignocellulose [5]. A notable derivative of D-xylose is xylitol, a polyalcoholic sweetener, which is formed by reducing the aldehyde group.

As carbohydrates are so abundant and various, it is expected that there are various enzymes acting on them in nature. Carbohydrates are used as precursors to many necessary substances: for example, L-ascorbic acid (vitamin C) is synthesised from D-glucose via a nine-step pathway in most animals but not humans [8, 9]. Both enzymes considered in this work belong to such metabolic pathways. Also, notably, oligo- and polysaccharides are commonly cleaved by various carbohydrate-active enzymes [10].

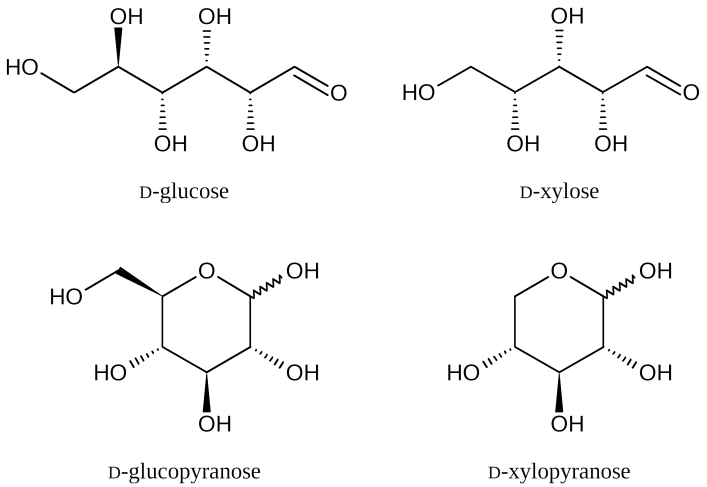



**Figure 1.1:** The linear forms and the cyclic pyranose forms of D-glucose and D-xylose, the two most abundant monosaccharides. All the stereogenic centres that are present in both are identical. The curly bond indicates that the stereogenic centre can be in either configuration (α- and β-anomers).

## 1.2 D-2-DEOXYRIBOSE-5-PHOSPHATE ALDOLASE

The D-2-deoxyribose-5-phosphate aldolase (DERA) catalyses the aldol reaction (Figure 1.2) between acetaldehyde (donor substrate) and glyceraldehyde-3-phosphate (G3H) (acceptor substrate) in nature [11]. The product is D-2-deoxyribose-5-phosphate (DRP), the furanose form of which is further converted into a deoxyribonucleotide, a monomeric unit of DNA, by a subsequent action of three enzymes, though this is not the preferred way of synthesising deoxyribonucleotides in nature [12]. The reaction is an equilibrium reaction which the DERA catalyses in both directions: the DRP can also be cleaved into acetaldehyde and G3H if they are not present in solution. The cleavage of nonphosphorylated DR is catalysed by the DERA as well (Publication **I**).

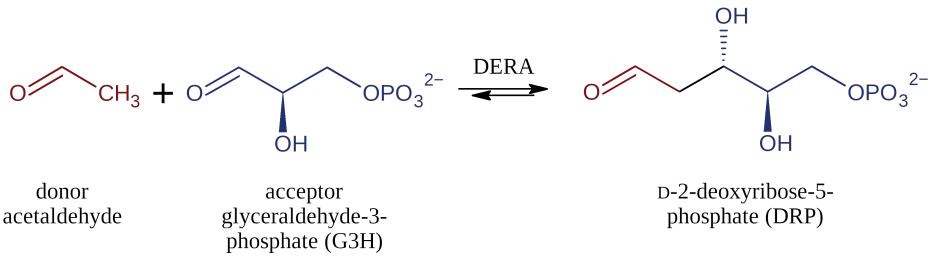



**Figure 1.2:** The aldol reaction catalysed by the DERA enzyme in nature [11]. The DRP is shown in the open-chain form.

The DERA is a class I aldolase. In general, the donor substrate binds to a catalytic lysine at the active site forming a Schiff base intermediate (Figure 1.3) [13]. The acceptor substrate binds to the active site for the aldol reaction. In contrast to other such aldolases, the DERA is remarkably efficient and accepts aldehydes as both donor and acceptor substrates [11] and can even chain multiple acetaldehydes [14].

The first structures of the *Escherichia coli* DERA (*Ec*DERA, Enzyme Commission number: EC 4.1.2.4) were presented by Heine *et al.* in 2001 [11], entries 1JCJ [15] and 1JCL [16] in the Protein Data Bank (PDB) [17, 18]. The tertiary structure is a triosephosphate isomerase (TIM) $(\alpha/\beta)_8$ barrel fold (Figure 1.4) where a barrel of eight parallel β-strands is surrounded by eight α-helices, and there is also a ninth α-helix at the N-terminus [19]. The active site is located at the central cavity in the centre of the barrel. The catalytic lysine is K167, and the side chain is located at the bottom of the cavity.

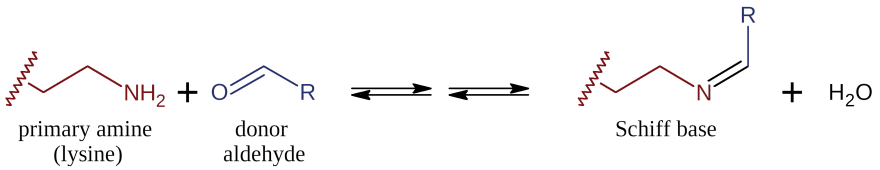



**Figure 1.3:** The net reaction of the Schiff base formation [20]. In the total reaction, there are five intermediates which are not shown here. The primary amine in this case is the catalytic lysine side chain of the DERA, and the aldehyde is acetaldehyde ($R = CH_3$).

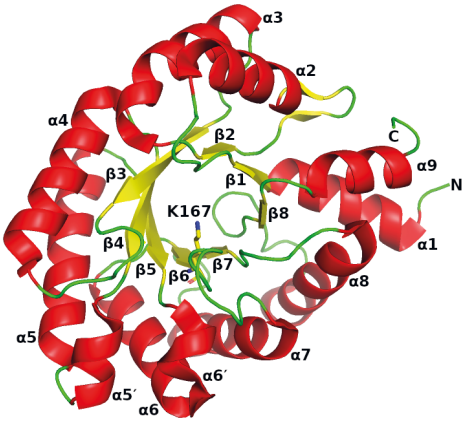



**Figure 1.4:** The tertiary structure of *E. coli* DERA (PDB entry 1JCL [11, 16]). The α-helices and β-strands of the TIM $(\alpha/\beta)_8$ barrel fold are labelled α1–α9 and β1–β8 respectively, in the order in which they appear in the amino acid sequence. The catalytic lysine K167 is also shown as a stick model.

Certain mutations to the *Ec*DERA have been discovered to significantly affect the enzymatic activity. In Publication **I**, Voutilainen *et al.* have tested several one-point mutations near the active site and two- and three-point combinations thereof. One-point mutations were selected by their locations, and successive mutagenesis was

done by manual selecton, saturation mutagenesis and prediction of substrate specificity using machine learning. Three mutants (N21K, T18Q, C47V/G204A/S239D) were deemed interesting enough that their three-dimensional structures were chosen to be determined by X-ray crystallography. Measured enzymatic activities of these mutations on DRP cleavage (the reverse aldol reaction), DR cleavage and acetaldehyde addition are shown in Table 1.1. The N21K mutation caused the DRP cleavage activity to be halved, and the T18Q mutation caused it to disappear. The C47V/G204A/S239D triple mutant had no DRP cleavage activity, negligible DR cleavage activity and more than double acetaldehyde addition activity compared to the wild-type.

**Table 1.1:** Approximate relative activities of *Ec*DERA mutations on DRP cleavage, DR cleavage and acetaldehyde addition. The values are relative to the activity of the wild-type, approximated from Figures 3 and 6 in Publication **I**.

| ***Ec*DERA mutant** | **DRP activity** | **DR activity** | **Acetaldehyde activity** |
|---|---|---|---|
| Wild-type | 1 | 1 | 1 |
| T18Q | 0 | 0.4 | 0.4 |
| N21K | 0.5 | 1.1 | 0.9 |
| C47V | 0.7 | 0.2 | 1.1 |
| G204A | 0 | 0 | 1.7 |
| S239D | 0.3 | 1.3 | 0.7 |
| C47V/G204A/S239D | 0 | 0.1 | 2.7 |

## 1.3 XYLONOLACTONASE

The xylonolactonase (XylC) catalyses the hydrolysis of D-xylonolactone to D-xylonic acid (Figure 1.5). The reaction belongs to a metabolic pathway from D-xylose to D-xylonic acid [21]. Extending this pathway to produce 1,2,4-butanetriol [22, 23], α-ketoglutaric acid [24] or glycolaldehyde and pyruvic acid [25] has been investigated and experimented. These pathways would be effective methods of producing various substances from D-xylose, the second most abundant monosaccharide. 1,2,4-Butanetriol is used as a reagent for the production of, for instance, polyurethane foam, medicine and explosives [22], and it could be used as a fuel as well. α-Ketoglutarate is an intermediate in the citric acid cycle and a precursor of glutamine and glutamic acid. [26]. Glycolaldehyde can be used as a precursor to substances such as ethylene glycol and glycolic acid, and pyruvic acid can be converted to lactic acid and further into poly(lactic acid) (PLA) [27].

The XylC is related to more thoroughly researched gluconolactonases and belongs to the senescence marker protein 30 (SMP30) protein family. The SMP30 were originally believed to be aging markers but were later discovered to catalyse the hydrolysis of D-gluconolactone [28]. It has been reported that in addition to D-gluconono-1,5-lactone, rat SMP30 also catalyses the hydrolysis of L-glucono-1,5-lactone, both D- and L-gulono-1,4-lactone and both D- and L-gulono-1,4-lactone but not D-ribono-1,4-lactone, D-mannono-1,4-lactone or D-glucoheptono-1,4-lactone [8]. Therefore, it is not very substrate-specific, and the XylC likely is not either.

The first structure of a SMP30 gluconolactonase from *Xanthomonas campestris* was reported by Chen *et al.* in 2008 [29] (PDB entry 3DR2 [30]). It was identified as belonging to the family by sequence analysis. The closest homologues to the XylC are the mammalian SMP30 gluconolactonases with about 30 % sequence identity, the first structures of which were reported by Chakraborti *et al.* [28] in 2010 (PDB entries 3G4E [31] and 3G4H [32]). The tertiary structure is a β-propeller consisting of six twisted antiparallel β-sheets arranged in a cylinder (Figure 1.6) [29]. There is a cavity at one end of the cylinder, and at the bottom of the cavity is the active site and a metal binding site. The SMP30 gluconolactonases have been modelled with either $Ca^{2+}$ or $Zn^{2+}$ in the metal binding site, coordinated by asparagine, aspartate and glutamate side chains [28, 33]. However, in the *Xanthomonas campestris* gluconolactonase structure, the modelled $Ca^{2+}$ is also coordinated by another asparagine side chain [29]. In addition to $Ca^{2+}$ and $Zn^{2+}$, $Mg^{2+}$ and $Mn^{2+}$ have also been shown to accelerate the enzymatic reaction and concluded to be able to bind to the metal binding site [28].

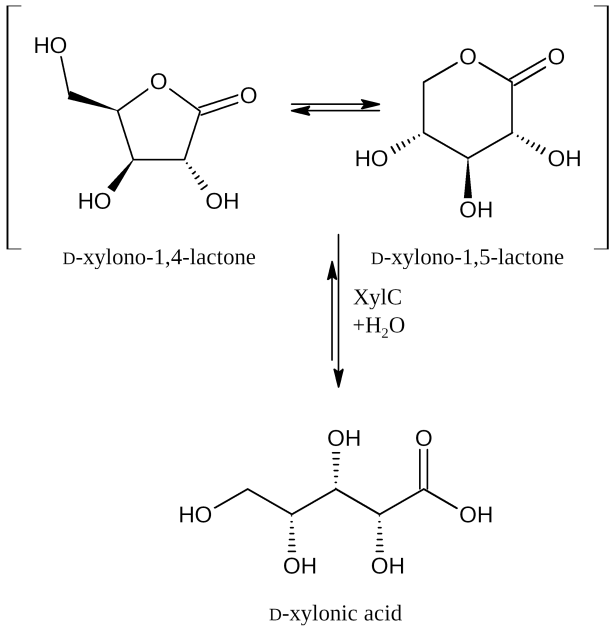



**Figure 1.5:** The lactone hydrolysis catalysed by the xylonolactonase (XylC) [21]. D-Xylonolactone exists as two forms in solution, of which the 1,4-lactone is more abundant [34]. It has been reported that with D-gluconolactone, only the 1,5-lactone is hydrolysed to D-gluconic acid in solution [35], and this may apply to D-xylonolactone as well. However, D-1,4-xylonolactone is reported as the natural substrate of the enzyme, and therefore both lactone forms are considered as possible substrates here.

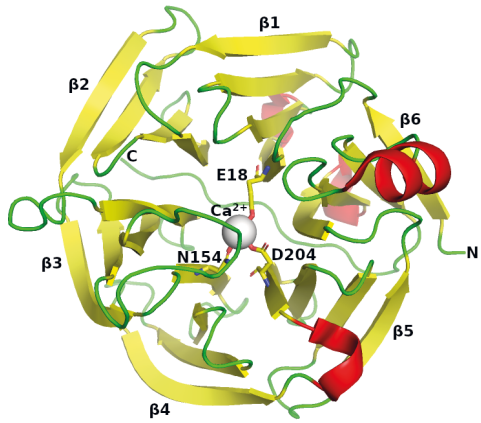



**Figure 1.6:** The tertiary structure of mouse SMP30 gluconolactonase (PDB entry 4GN7 [33, 36]), a homologue of the xylonolactonase. The antiparallel β-sheets of the six-bladed β-propeller fold are labelled β1–β6 in the order in which the sheets appear in the amino acid sequence. As an exception, the first β-strand in the sequence is in sheet β6. The modelled $Ca^{2+}$ ion is shown as a white sphere, and the amino acid residues (E18, N154, D204) coordinating to the $Ca^{2+}$ are also shown.

## 1.4 AIMS OF THE STUDY

The aims of the study were to characterise the xylonolactonase from *Caulobacter crescentus* (*Cc*XylC, EC 3.1.1.68) using mass spectrometry (Publication **II**) and to solve the three-dimensional structures of the *Cc*XylC (Publication **III**) and the three *Ec*DERA mutants (N21K, T18Q, C47V/G204A/S239D) using X-ray crystallography (Publication **I**). Complex structures would be desirable if obtainable so that the ligand binding could be studied. Also, once the *Cc*XylC was observed to bind iron(II), the effect of presence of iron(II) on the lactone hydrolysis was to be investigated. The kinetics of the lactone hydrolysis reactions would also be analysed with mass spectrometry.

Another goal was to develop a simulation program for visualising simple binding equilibria common with proteins (Publication **IV**). The hope is that such a simulation would be useful in education, experiment planning and result validation. Equilibrium concentrations do not behave intuitively, which leads to misconceptions and errors even in published results. With a simple and publicly available simulation, it would be easy to check whether, say, the experimental behaviour of the dimerisation of a protein matches the theoretical behaviour calculated using the determined dissociation constant.

# 2 THERMODYNAMICS OF PROTEIN ASSOCIATION

Like other chemical species, proteins form complexes by self-association (homodimerisation) or by binding to other proteins, small molecules or ions. The binding can occur either by weak interactions (hydrophobic interactions or hydrogen bonding) or by chemical bonding (covalent or ionic bonding). The complex formation is often vital for biological function: for instance, the D-xylonate dehydratase from *Caulobacter crescentus* is inactive unless complexed with a $Mg^{2+}$ ion [37, 38].

The association reaction is an equilibrium between free species A and B and their complex AB:

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{AB}. \tag{2.1}$$

The association constant $K_\mathrm{A}$ is the equilibrium constant of this reaction, and it is defined in terms of the equilibrium concentrations:

$$K_\mathrm{A} = \frac{[\mathrm{AB}]}{[\mathrm{A}]\,[\mathrm{B}]}. \tag{2.2}$$

The more often used constant is the dissociation constant $K_\mathrm{D}$, the equilibrium constant of the reverse reaction. It is practical because it has the same unit as the concentration, and the value is physically meaningful.

$$K_\mathrm{D} = \frac{1}{K_\mathrm{A}} = \frac{[\mathrm{A}]\,[\mathrm{B}]}{[\mathrm{AB}]} \tag{2.3}$$

The equilibrium constants are in connection with the Gibbs free energy of the reaction:

$$\Delta G = -RT \ln\left(K_\mathrm{A} \cdot c^\circ\right) = RT \ln \frac{K_\mathrm{D}}{c^\circ}, \tag{2.4}$$

where $R$ is the molar gas constant, $T$ is temperature and $c^\circ = 1\ \mathrm{mol\,l^{-1}}$. The equilibrium constants are formally unitless and expressed in molar units, so the values are taken in litres per mole ($K_\mathrm{A}$) or moles per litre ($K_\mathrm{D}$) and used as dimensionless numbers as the numerus of the logarithm [39].

## 2.1 SELF-ASSOCIATION (HOMODIMERISATION)

In self-association, a protein (P) forms a dimer with itself. The reaction equation is

$$\mathrm{P} + \mathrm{P} \rightleftharpoons \mathrm{P}_2. \quad (K_\mathrm{D}) \tag{2.5}$$

This is always an equilibrium reaction: neither monomer nor dimer can be isolated, and both are always present in solution. When the total concentration of protein is equal to $K_\mathrm{D}$ or, alternatively, the concentration of free protein monomer is equal to $\frac{1}{2}K_\mathrm{D}$, exactly a half of the protein is in dimeric form. No protein is categorically dimeric since the degree of dimerisation depends on protein concentration. For example, if the $K_\mathrm{D}$ of bovine β-lactoglobulin is 5.0 μmol $\mathrm{l^{-1}}$ (from

$K_A = 2.0 \cdot 10^5$ l mol$^{-1}$ [40]), the protein concentration must be at least 230 µmol l$^{-1}$ (4.3 mg/ml) so that more than 90 % of the protein is in dimeric form (Figure 2.1).

More detailed mathematical consideration of this and the other following cases is presented in Publication **IV**.

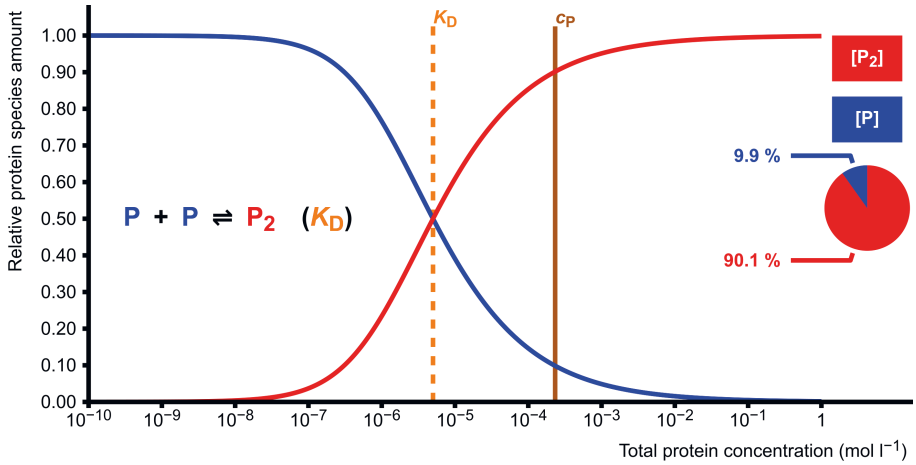



**Figure 2.1:** The proportional amounts of protein monomer (P, blue) and dimer ($P_2$, red) as functions of total protein concentration. The dissociation constant $K_D$ is set at $5.01 \cdot 10^{-6}$ mol l$^{-1}$ (close to the experimental $K_D$ of bovine β-lactoglobulin [40]), and the pie diagram shows the proportional amounts at the set value of total protein concentration $c_P = 2.33 \cdot 10^{-4}$ mol l$^{-1}$. The diagram has been created with the homodimerisation simulation applet described later in Section 2.5.

## 2.2 COMPLEX OR HETERODIMER FORMATION

In complex formation, a protein (P) and a ligand (L) form a complex PL. In heterodimerisation, L is another protein. The reaction equation is

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{PL}. \quad (K_D) \tag{2.6}$$

This is also an equilibrium reaction: there are always some concentrations of complex, free protein and free ligand present in solution. If the total protein concentration is much less than the dissociation constant and the total ligand concentration equals the dissociation constant, the equilibrium concentrations of P and PL are equal. Alternatively, when the free ligand concentration equals the dissociation constant, the concentrations are also equal, meaning that a half of the protein is in complexed form. For example, if the human SMP30 gluconolactonase has dissociation constants 82 µmol l$^{-1}$ and 566 µmol l$^{-1}$ with $Mg^{2+}$ and $Ca^{2+}$ respectively [28] and if the free concentrations of $Mg^{2+}$ and $Ca^{2+}$ are 500 µmol l$^{-1}$ and 1 µmol l$^{-1}$ respectively [28, 41–43], the enzyme is about 86 % saturated with $Mg^{2+}$ but only 0.2 % saturated with $Ca^{2+}$ (Figures 2.2 and 2.3).

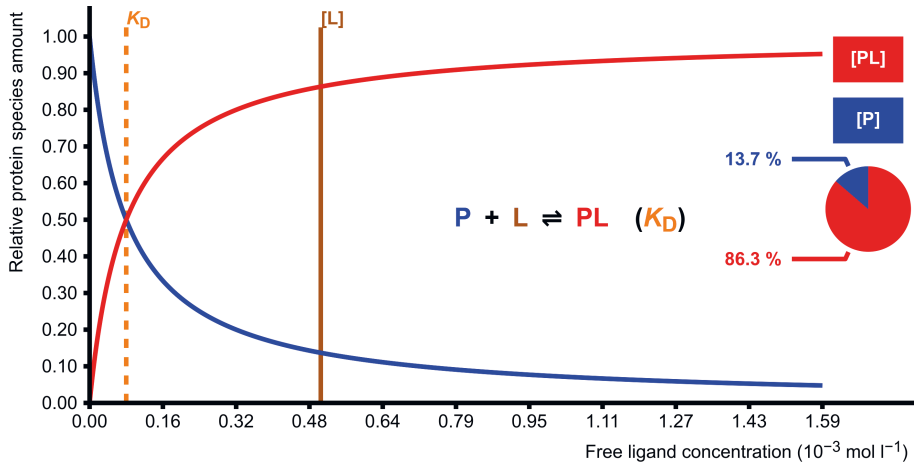



**Figure 2.2:** The proportional amounts of free protein (P, blue) and metal ion complex (PL, red) as functions of free metal ion concentration. The dissociation constant $K_D$ is set at $7.94 \cdot 10^{-5}$ mol $l^{-1}$, close to the experimental value for the SMP30–$Mg^{2+}$ complex, and the pie diagram shows the proportional amounts at the set value of $[L] = 5.01 \cdot 10^{-4}$ mol $l^{-1}$, the reported free $Mg^{2+}$ concentration in cellular medium [28]. In these conditions, the protein is 86.3 % saturated, and so this complex would exist in nature as well.

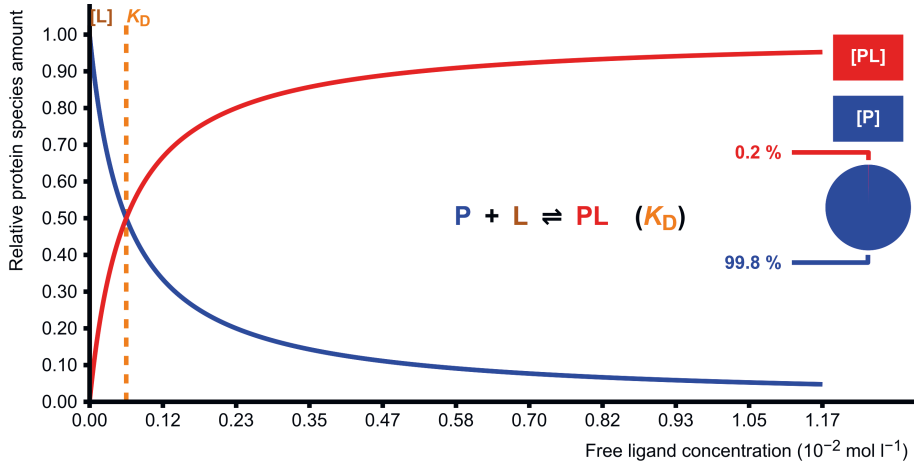



**Figure 2.3:** The proportional amounts of free protein (P, blue) and metal ion complex (PL, red) as functions of free metal ion concentration. The dissociation constant $K_D$ is set at $5.84 \cdot 10^{-4}$ mol $l^{-1}$, close to the experimental value for the SMP30–$Ca^{2+}$ complex, and the pie diagram shows the proportional amounts at the set value of $[L] = 1.00 \cdot 10^{-6}$ mol $l^{-1}$, the upper bound of the reported range of free $Ca^{2+}$ concentration in cellular medium [28]. In these conditions, the protein is only 0.2 % saturated, indicating that this complex is not present in significant amounts in nature.

## 2.3 COMPETITIVE BINDING OF A LIGAND TO TWO RECEPTORS

When a ligand (L) can bind competitively to two receptors (P and P′) which can be either different proteins or different binding sites in the same protein, the are two equilibrium reactions:

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{PL}, \quad (K_\mathrm{D}) \tag{2.7}$$

$$\mathrm{P}' + \mathrm{L} \rightleftharpoons \mathrm{P}'\mathrm{L}. \quad (K'_\mathrm{D}) \tag{2.8}$$

In the case that P and P′ are in the same protein, the receptors are assumed to be independent and successive reactions are disregarded. This model can be used to describe, for instance, how a ligand would bind competitively to the active sites of two enzymes or how metal ions would bind into a protein specifically and nonspecifically. The receptor specificity constant, as defined by Eaton *et al.* [44], describes how much complex PL exists compared to P′L:

$$\alpha_s = \frac{[\mathrm{PL}]}{[\mathrm{P}'\mathrm{L}]}. \tag{2.9}$$

For example, the β-synuclein has two binding sites for copper(II): primary (P) with $K_\mathrm{D} = 0.20\ \mu\mathrm{mol\,l^{-1}}$ and secondary (P′) with $K'_\mathrm{D} = 60\ \mu\mathrm{mol\,l^{-1}}$ [45]. In the model used by Binolfi *et al.* [45], the binding sites have equal concentrations, and let that be $c_\mathrm{P} = c_{\mathrm{P}'} = 300\ \mu\mathrm{mol\,l^{-1}}$, the concentration used in the experiments. Assuming that the total $Cu^{2+}$ concentration is $c_\mathrm{L} = 13\ \mu\mathrm{mol\,l^{-1}}$ (from $0.85\ \mu\mathrm{g\,ml^{-1}}$ observed in the blood of deceased humans [46]) and that the two equilibrium reactions are independent, the specificity constant is about 280 (Figure 2.4). However, P is only 4.3 % saturated and P′ is less than 0.02 % saturated, indicating that copper is slightly bound to the primary receptor in nature and only in insignificant amounts to the secondary receptor. (Figure 2.5).

## 2.4 COMPETITIVE BINDING OF TWO LIGANDS TO A RECEPTOR

In this case, two ligands L and L′ bind competitively to a single receptor P. The situation resembles the previous one, and the same thermodynamic laws apply, only the protein species are replaced with ligand species and vice versa. The reaction equations are

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{PL}, \quad (K_\mathrm{D}) \tag{2.10}$$

$$\mathrm{P} + \mathrm{L}' \rightleftharpoons \mathrm{PL}'. \quad (K'_\mathrm{D}) \tag{2.11}$$

This model can be used to describe how two ligands would compete over a binding site. For example, consider that carbon monoxide and oxygen bind to hemoglobin with dissociation constants $K_\mathrm{D} = 1.3 \cdot 10^{-9}\ \mathrm{mol\,l^{-1}}$ and $K'_\mathrm{D} = 3.0 \cdot 10^{-7}\ \mathrm{mol\,l^{-1}}$ respectively (calculated from the reported kinetic parameters [47, 48]) and that the hemoglobin concentration in the red blood cell is $c_\mathrm{P} = 2.0 \cdot 10^{-2}\ \mathrm{mol\,l^{-1}}$ [47, 49]. With the ligand binding simulation, it can be determined that in order to reach a healthy 95 % oxygen saturation, the concentration of oxygen must be $c_{\mathrm{L}'} \approx 0.95 c_\mathrm{P} = 1.9 \cdot 10^{-2}\ \mathrm{mol\,l^{-1}}$ (Figure 2.6). Competing binding simulation shows that the oxygen saturation decreases to 90 %, the limit of hypoxemia, when the carbon monoxide concentration $c_\mathrm{L}$ rises to $2.0 \cdot 10^{-3}\ \mathrm{mol\,l^{-1}}$, approximately 10 % of the oxygen concentration (Figure 2.7).

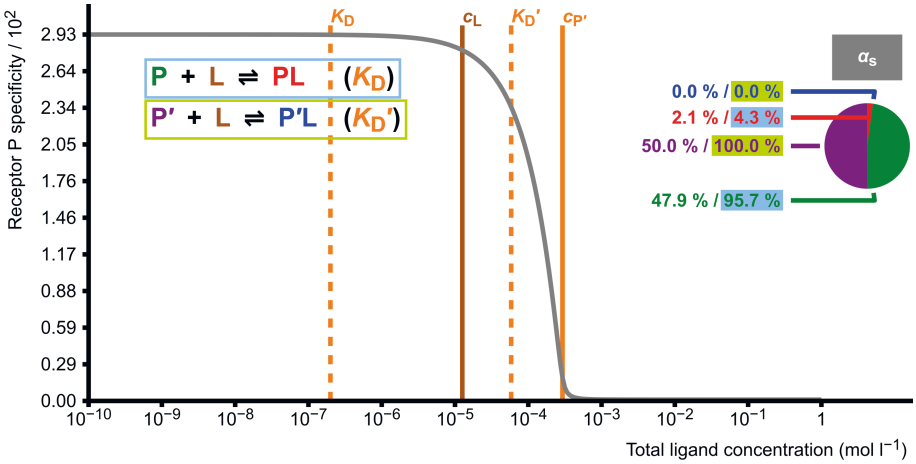



**Figure 2.4:** The specificity of receptor P over P′, defined in Equation (2.9), as a function of total ligand L concentration. Receptor concentrations are $c_{\mathrm{P}} = c_{\mathrm{P}'} = 2.93 \cdot 10^{-4}$ mol l$^{-1}$, and the dissociation constants are $K_{\mathrm{D}} = 2.00 \cdot 10^{-7}$ mol l$^{-1}$ and $K'_{\mathrm{D}} = 5.84 \cdot 10^{-5}$ mol l$^{-1}$, all close to the experimental values by Binolfi *et al.* [45]. The pie diagram shows the proportional amounts of the protein species at $c_{\mathrm{L}} = 1.26 \cdot 10^{-5}$ mol l$^{-1}$, close to the estimated concentration of copper in human blood [46]. The specificity is at the highest when $c_{\mathrm{L}}$ is small, approaching $K'_{\mathrm{D}}/K_{\mathrm{D}} \approx 293$ when $c_{\mathrm{L}}$ approaches zero, and decreases to unity when $c_{\mathrm{L}}$ is much greater than $c_{\mathrm{P}}$ and $c_{\mathrm{P}'}$.

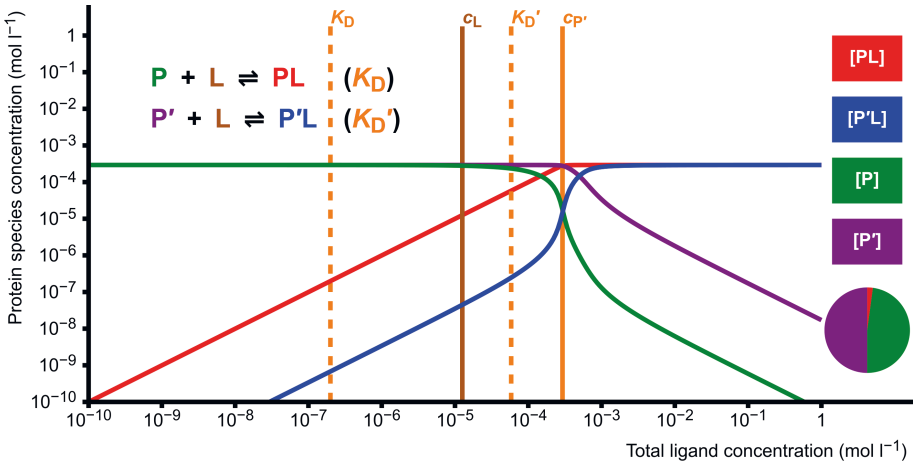



**Figure 2.5:** The absolute concentrations of the protein species as functions of total ligand L concentration in a competing receptors situation. The initial concentrations and dissociation constants are the same as in Figure 2.4, and the pie diagram shows the same values as well. The experimental $c_{\mathrm{L}}$ is low enough that only small amounts of the complexes are present.

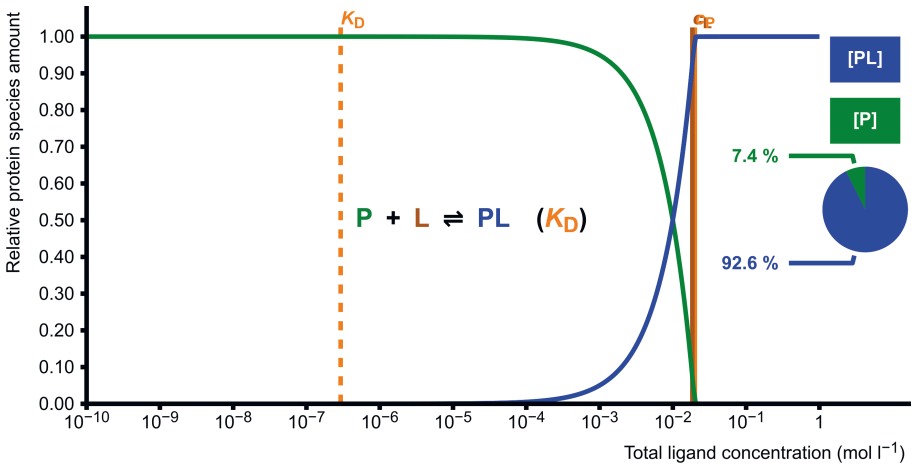



**Figure 2.6:** The proportions of free protein (P, green) and complex (PL, blue) as functions of total ligand L concentration. Total protein concentration is $c_P = 2.00 \cdot 10^{-2}$ mol l$^{-1}$, close to the typical hemoglobin concentration in the red blood cell [47, 49]. The dissociation constant is $K_D = 2.93 \cdot 10^{-7}$ mol l$^{-1}$, close to the value for the hemoglobin–oxygen complex [47, 48]. The pie diagram shows the proportions at the set value of $c_L = 1.85 \cdot 10^{-2}$ mol l$^{-1}$, and the proportion of PL represents the oxygen saturation in blood. Due to the granularity of the sliders in the simulation applet, the desired value of 95 % was not obtainable, but since $[PL] \approx c_L$ when $c_L < c_P$, the 95 % saturation must be reached at $c_L \approx 0.95 c_P$.

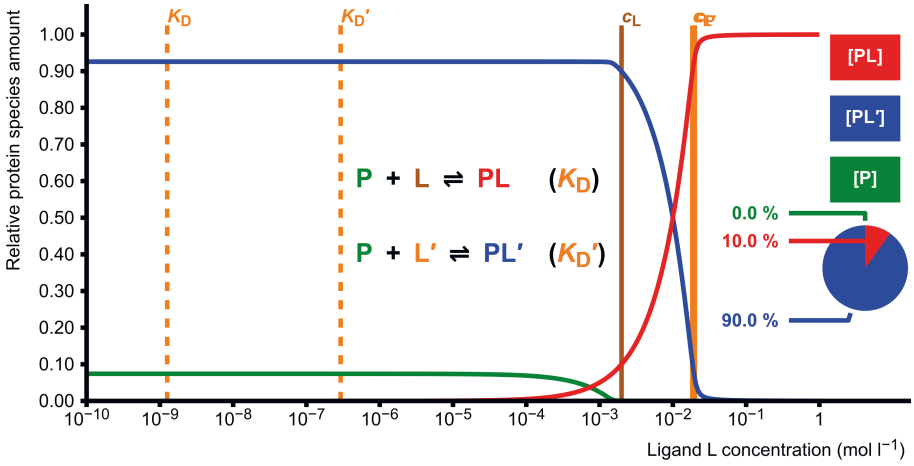



**Figure 2.7:** The proportions of free protein P (green) and two complexes PL (red) and PL′ (blue) as functions of total ligand L concentration. The values for $c_P$, $c_{L'}$ and $K_D'$ are taken from $c_P$, $c_L$ and $K_D$ in Figure 2.6 respectively. The dissociation constant $K_D$ is set at $1.26 \cdot 10^{-9}$ mol l$^{-1}$, close to the value for the hemoglobin–carbon monoxide complex [47, 48]. The pie diagram shows the proportions at the set value of $c_L = 2.00 \cdot 10^{-3}$ mol l$^{-1}$, and the proportion of PL′ represents the oxygen saturation in blood. The oxygen saturation stays at the high level at low carbon monoxide concentrations, but decreases to 90 % at $c_L$ and falls rapidly thereafter.

## 2.5 SIMULATION APPLETS

The aforementioned reaction cases have been programmed into simulation applets that run in a web browser. A screenshot of the ligand binding simulation is shown in Figure 2.8. In each applet, the initial concentrations and dissociation constants (parameters) are set with sliders, and changing them triggers JavaScript code to update a graph of the equilibrium concentrations. The graph is implemented as a Scalable Vector Graphics (SVG) object that can be scaled indefinitely or extracted as a vector graphic file. In addition, instead of total concentration, the equilibrium concentration of free protein or free ligand can be selected as the adjustable parameter in the homodimerisation or the ligand binding simulation respectively. Association constants and Gibbs free energies of association at 25 °C corresponding to the set dissociation constants are also shown. Molar masses of the proteins and ligands can be set, and they will be ignored everywhere except in the calculation of mass concentrations.

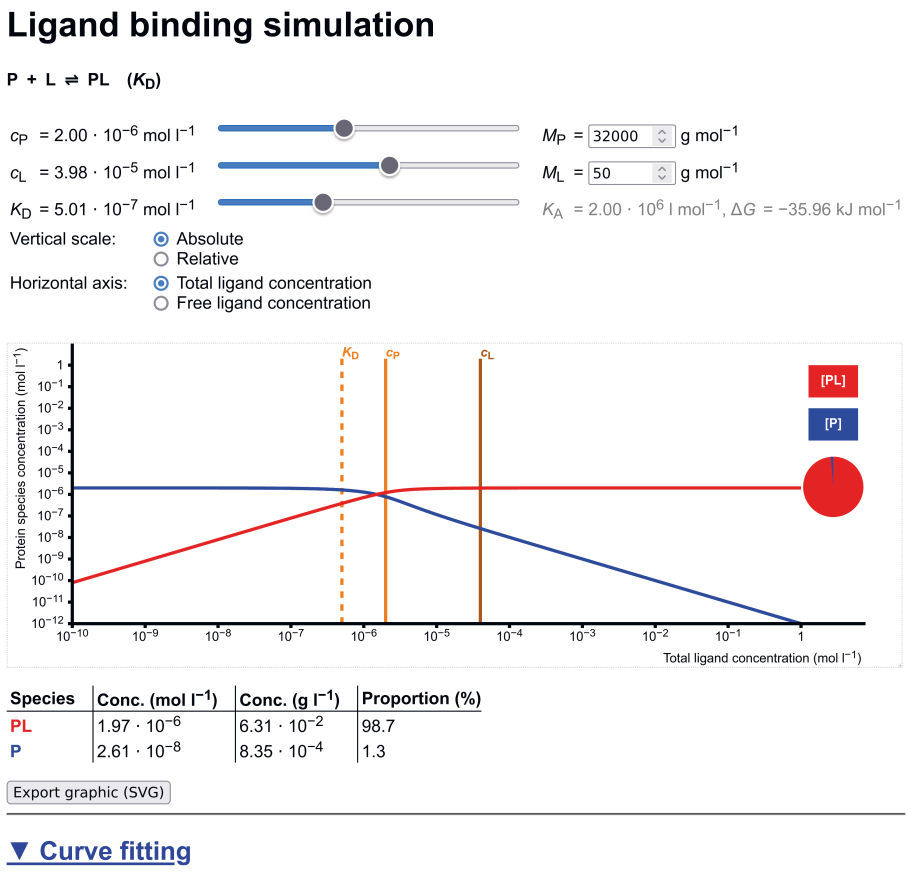



**Figure 2.8:** Screenshot of the ligand binding simulation applet.

The graph displays the equilibrium concentrations of all protein species as functions (curves) of one of the adjustable concentration parameters. In the homodimerisation, ligand binding and competing receptors simulations, either total or equilibrium concentrations of protein, ligand or ligand respectively can be set as the $x$-axis. In the competing ligands simulation, the $x$-axis can be the total concentration of either protein P or ligand L. The $y$-axis can be either logarithmic or linear: the logarithmic scale will show the curves in a fixed scale from $10^{-10}$ to 1, and the linear scale will show the curves in an automatically adjusted scale in the competing receptors simulation or the relative amounts of protein species in the others. It was deemed that the relative scale from 0 to 1 would be useful when there is only one protein species. However, in the competing receptors simulations, where there are two different protein species, using the absolute scale was decided to be preferable. In addition, in the competing receptors simulation, the receptor P specificity $\alpha_s$ can be plotted instead of the equilibrium concentrations, and it is always drawn in a linear scale from 0 to the maximum value.

The slider setting of the parameter on the $x$-axis defines an inspection point. It is marked in the graph as a brown vertical line whereas all other parameters are marked as orange lines. At this point, the equilibrium concentrations of each species are calculated separately from the graph creation and displayed in a table below the graph. Also displayed are the corresponding mass concentrations in grams per litre and the relative amounts of each species. The relative amounts are also shown in the pie diagram in the graph. In the competing receptors simulation, the receptor specificity at this point is shown as well.

There is also a curve fitting tool available in each simulation. Data can be input as pairs of $x$ and $y$ values, and any of the displayed curves can be chosen to be least squares fitted to the data. One or more of the parameters, excluding the one set as the $x$-axis, must be set as free, and after clicking "Calculate", the program finds the optimal values for the free parameters that result in the minimal sum of squared residuals. The other parameters are left at the values set by the user. The parameters are restricted to the possible discrete values of the sliders, so the solution will always be imprecise, and if the best solution is outside the range of some parameter, the returned value will be at one of the extreme ends of the range. In addition to the results being imprecise, the uncertainties of the values are not calculated, and therefore the results should be taken as rough approximations.

The best fit can be searched in three ways: by searching the free parameter space in two successive passes (coarse and fine), by searching the entire space in a single pass or by searching the vicinity of the set values (initial guess). The first method is fast but can potentially fail in exotic cases where the coarse search finds an incorrect minimum and does the fine search in an incorrect part of the free parameter space. The second method is more reliable, but since it checks a much larger number of combinations, it is slow when there are many free parameters. The third method is fast and reliable when the optimal values are close to the initial values, but if not, it can potentially converge to an incorrect point. The two-pass method is the default option, and it should work without fail in realistic cases.

The simulation applets are shared online. All the source files are available in a GitHub repository [50], and the pages can be accessed via a GitHub page (url: `https://protsim.github.io/protsim`) [51] using a web browser. The source files can be freely used and modified by anyone as long as the terms of the GNU General Public License [52] are complied with.

# 3 MATERIALS AND METHODS

## 3.1 SAMPLE PREPARATION

### 3.1.1 D-2-Deoxyribose-5-phosphate aldolase

The mutants of the *Ec*DERA were prepared at VTT Technical Research Centre of Finland Ltd as described in Publication **I**. Briefly, the encoding gene was cloned and expressed in *E. coli*, appended with a hexa histidine tag and transferred into *E. coli* cells with a pBAT4 vector. Mutations were done using a mutagenesis kit, and the enzymes were expressed in *E. coli* strains and extracted using a HisTrap column.

The *Ec*DERA mutants were received from VTT as various solutions: N21K as 1.2 mg/ml in 50 mM Tris buffer (pH 8.0), T18Q as 1.1 mg/ml in 50 mM Tris buffer (pH 8.0) and C47V/G204A/S239D as 1.3 mg/ml in 50 mM sodium phosphate buffer (pH 7.5). The N21K and T18Q mutants were used in crystallisation without any buffer exchange or concentration. However, because the sodium phosphate was deemed incompatible with organic buffers, the C47V/G204A/S239D mutant was transferred to a 50 mM Tris buffer (pH 8.0) using a PD-10 desalting column (GE Healthcare, Little Chalfont, England) and concentrated using a Vivaspin 2 concentrator (10 kDa molecular weight cut-off) (GE Healthcare, Little Chalfont, England). The concentration of this solution was determined to be 3.2 mg/ml using 280 nm ultraviolet absorbance in an Eppendorf BioPhotometer Model #6131 spectrophotometer (Eppendorf AG, Hamburg, Germany), and this concentrate was used in crystallisation.

### 3.1.2 Xylonolactonase

The *Cc*XylC was prepared at VTT as described by Boer *et al.* [27]. Briefly, the encoding gene was purchased as appended with a Strep-tag II, codon optimised in a pBAT4 vector, expressed in *E. coli* cells and extracted with a DEAE FF 16/10 ion exchange column. Two batches of *Cc*XylC were received from VTT: the first was 1.6 mg/ml in 50 mM Tris buffer (pH 8.0), and the second was 5.1 mg/ml in 50 mM Tris buffer (pH 7.5). Quick mass spectrometry tests resulted in identical spectra for both, so no difference was made between them.

For the mass spectrometry measurements, the *Cc*XylC was transferred to 10 mM ammonium acetate solution (ultrapure in HPLC quality water) using a PD-10 desalting column. The concentration of this solution was determined using 280 nm ultraviolet absorbance, and the solution was diluted to approximately 1 μM (0.03 mg/ml). The first native mass spectrum was measured directly from this solution, and a sample for the denatured mass spectrum was prepared by mixing this solution, acetonitrile and acetic acid in final proportions 49.5 %/49.5 %/1.0 % ($v/v$). After bound iron was observed in the native mass spectrum (discussed in Section 4.1.1), new apo-enzyme solutions were prepared similarly, but at least 20 molar equivalents of EDTA was added to the protein at least 15 minutes before desalting to remove any metals. Apo-enzyme–metal samples were prepared by adding metal chloride to the apo-enzyme solution at least 30 minutes before measurement.

For crystallisation, small amounts at a time of the *Cc*XylC stocks were turned into holo-form by adding 100 molar equivalents of $FeSO_4$ and keeping the mixture overnight. Then, they were desalted and eluated with 50 mM Tris buffer (pH 8.0) in a PD-10 desalting column and concentrated using a Vivaspin 2 concentrator in a centrifuge, resulting in various concentrations 5–11 mg/ml. The colour of the solution always remained yellow, which indicated that the iron had remained in the solution bound to the enzyme. These solutions were used as they were or sometimes diluted for some experiments. The crystallisation experiments that ultimately resulted in crystal structures were done with the second batch at 7 mg/ml (rectangular crystal with D-xylose) and with the first batch at 8 mg/ml (the other three).

### 3.1.3 Lactone hydrolysis

The lactone hydrolysis samples were prepared by first mixing stock solutions of xylitol, metal chloride (optional) and apo-enzyme (optional). Solvent was added so that in the end the concentrations would be 1.0 mM, 10 μM and 0.5 μM respectively. When both metal chloride and apo-enzyme were used, this mixture was kept still for at least 30 minutes so that any complex formation would reach equilibrium. Xylitol was used as an inert internal standard that would have a similar mass to the reaction products and in such a concentration that it and the product acid would have similar intensities in the mass spectra. Immediately before measurements, freshly dissolved D-xylono-1,4-lactone or D-glucono-1,5-lactone (isomers as reported in reagent containers) was added so that it would be diluted to 0.25 mM.

## 3.2 MASS SPECTROMETRY

### 3.2.1 Overview

Two mass spectrometers were used in the work: Esquire 3000plus quadrupole ion trap (QIT) and Solarix XR Fourier transform ion cyclotron resonance (FT-ICR), both by Bruker Daltonik GmbH, Bremen, Germany. Both instruments were equipped with similar electrospray ionisation (ESI) ion sources. The sample solutions were injected into the ion source using a glass syringe in a syringe pump. Measurement parameters are presented in the Supporting Information of Publication **II**.

Mass spectrometry (MS) is an analytical method of separating and measuring ions by their mass-to-charge ratios ($m/z$). It is a gas-phase method, so any sample – typically in liquid phase – must be evaporated and ionised before being transferred into the mass spectrometer and detected. Ion separation and detection must be done in vacuum to prevent the ions from colliding with gases [53]. There are various methods of ionisation: some, mainly electron ionisation (EI), typically cause the sample molecules to fragment, whereas some, such as chemical ionisation (CI), electrospray ionisation (ESI) and matrix-assisted laser desorption ionisation (MALDI), exhibit little or no fragmentation. Depending on the ion source, the ions can be radical ions, fragmented ionic molecules, deprotonated molecules or molecules with ionic adducts. There are also various mass analysers which separate and detect ions of different mass-to-charge ratios by how they behave in electric or magnetic fields or both [54].

In ESI, liquid sample is pressed through a thin capillary, evaporated and ionised in an electric field. One of the electrodes is at the spray nozzle, and thus the electric field causes the ejected droplets to be charged. As the solvent evaporates from the droplets, analyte molecules are left charged by deprotonation, protonation or sometimes adduction of sodium or potassium ions, of which protonation is the most common. Some ions will travel through an opening in the counterelectrode and enter the mass spectrometer through another capillary and a skimmer. ESI does not cause fragmentation to the analyte, and even weak interactions, as in protein structures, are preserved. Since the evaporation of the solvent does not require a high temperature, thermally labile analytes can be analysed. Therefore, ESI is a suitable method for ionising various polar compounds that can be protonated or deprotonated, including carbohydrates and proteins. [55]

A QIT mass analyser traps the ions in a three-dimensional electric field created by three electrodes. Ions are transmitted through a buffer gas, typically helium, at low pressure to cool them down. In the QIT, the ions end up in characteristic trajectories dependent on their mass-to-charge ratios. Setting a radio frequency (RF) voltage across the ends of the trap and increasing it causes the trajectories to gradually become unstable and the ions to be ejected. As the RF amplitude is increased, ions with gradually higher mass-to-charge ratios are ejected and are thus separated in RF amplitude and time. Ions ejected through an opening in the far end electrode are detected by an ion detector. [56]

An FT-ICR mass analyser also traps the ions but in a cylindrical cell with a combination of electric and magnetic fields. High-velocity ions are decelerated with a trapping electric field into a strong uniform magnetic field where they enter circular orbits with cyclotron frequencies dependent on their mass-to-charge ratios. Then, the ions are excited with an excitation pulse between excitation plates, which causes ions in the same orbit to bunch up. As the ions of the same $m/z$ are in the same phase, they will incude an sinusoidal voltage between receiver plates which is detected and recorded. Multiple ions with different $m/z$ values induce a signal containing multiple frequencies with amplitudes proportional to the amounts of the ions. The frequencies of the recorded signal are extracted with a Fourier transform, resulting in a spectrum which is converted from frequency space to $m/z$ space. Thus, all ions in the sample are detected simultaneously. This method has high sensitivity and high mass resolution, which makes it suitable for measuring very small concentrations or very large molecules, including whole proteins. [57]

### 3.2.2 Xylonolactonase

The denatured and native mass spectra of the *Cc*XylC were measured with the FT-ICR MS using the solutions described in Section 3.1.2. The tests with apo-*Cc*XylC and metal ions were done in roughly 20 μM metal ion concentration. The affinity of $Fe^{2+}$ to the *Cc*XylC was quantified by measuring a series with differing $Fe^{2+}$ concentrations. The enzyme concentration was kept constant at approximately 1 μM, and the $Fe^{2+}$ concentration was varied from 7.8 nM to 16 μM so that, diluting from 16 μM, it was halved each time. The holo-form could not be observed below 125 nM $Fe^{2+}$, so there ended up being eight data points for the determination of the dissociation constant.

### 3.2.3 Lactone hydrolysis

The lactone hydrolysis reactions were followed with the QIT MS using the solutions described in Section 3.1.3. The measurements were done in chromatogram mode so that a line spectrum (a set of peak positions and intensities) was measured approximately once per second. Profile spectrum (full spectrum, set of ion counts as a function of $m/z$) snapshots from measurements with xylono- and gluconolactone are shown in Figure 3.1. The intensities of xylono- or gluconolactone, the corresponding acid and xylitol over time were extracted from the raw data, and the times were adjusted so that the starting points of the reactions were at zero. The intensities of product acid were divided by the intensities of xylitol, resulting in raw values over time that were proportional to the concentration of the product acid. Theoretical curves of the reaction progress were fitted to the data points in *GNU Octave* [58], yielding the rate coefficients of the reactions. The mathematics are described in detail in the Supporting Information of Publication **II**.

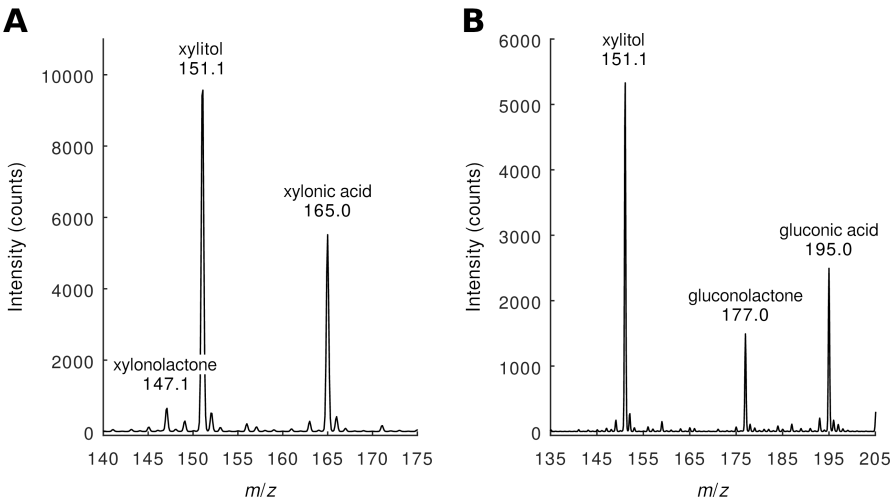



**Figure 3.1:** Representative profile spectra from the enzyme kinetics measurements with **(a)** xylonolactone and **(b)** gluconolactone. The spectra were measured in negative polarisation mode, so all the species are deprotonated and their mass-to-charge ratios appear ∼1.0 Da smaller than their neutral masses.

## 3.3 PROTEIN CRYSTALLISATION

### 3.3.1 Overview

All the crystallisation experiments were done using hanging drop vapour diffusion. For each individual trial, 500 µl of precipitant solution (the reservoir) was first set in a well of a 24-well crystallisation plate. A crystallisation droplet was prepared by mixing equal volumes of protein solution and the precipitant solution on a thin glass disc. The volumes of the droplets were $2 + 2$ µl for the *Ec*DERA and $1 + 1$ µl for the *Cc*XylC. The glass disc was set upside down on top of the well, and the interface was sealed with high vacuum stopcock grease. The progress of crystallisation was monitored by visual observation with a microscope.

In vapour diffusion, the solvent (water) of these two volumes of liquid will evaporate and condense so that the vapour pressure is reached and the difference in concentration is eliminated. Water will effectively travel from the crystallisation droplet to the reservoir until the droplet is concentrated to the same concentration as the reservoir. Since the reservoir is much larger than the droplet, its concentration does not decrease significantly. Over a long period of time, water vapour will also slowly escape through the sealed interface, causing the droplet and the reservoir to dry out. In the process, the concentration of protein also increases, and it becomes insoluble. In suitable conditions, the solution ends up in a metastable state where the protein begins to form crystals instead of precipitating. Formation of crystals requires nucleation, or formation of critical nuclei, which can occur spontaneously or be induced by introducing suitable solid matter into the crystallisation medium. Crystal growth can also be initiated by seeding, or introducing microcrystals of protein that will grow in slightly metastable conditions. [59]

### 3.3.2 D-2-Deoxyribose-5-phosphate aldolase

It was known beforehand that the wild-type *Ec*DERA could be crystallised using polyethylene glycol (PEG) 3350, magnesium formate and buffer at pH 8.0 [60]. After trying different PEGs and buffers and optimising the conditions, the N21K mutant was directly crystallised using the 1.2 mg/ml protein solution and 18–20 % ($w/v$) PEG 4000, 0.2 M magnesium formate and 0.1 M Tris buffer (pH 8.0) as the crystallisation solution. The crystals were two-dimensional plates with largest dimensions up to 300 µm (Figure 3.2a).

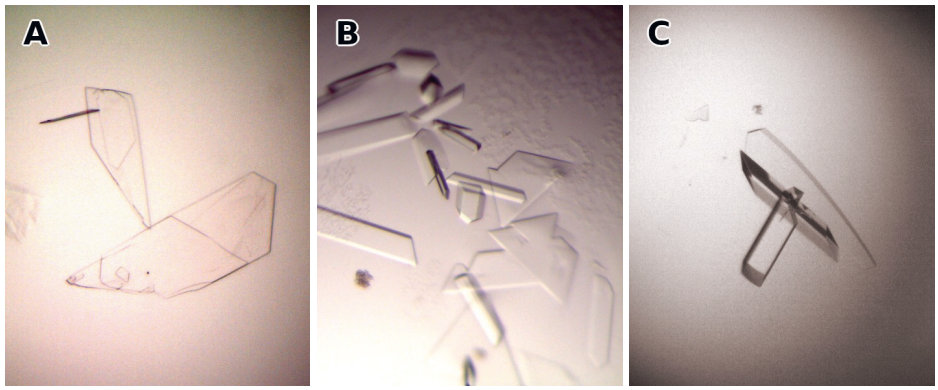



**Figure 3.2:** *Ec*DERA crystals in crystallisation droplets. **(a)** N21K mutant crystals. Crystallisation solution: 20 % ($w/v$) PEG 4000, 0.2 M magnesium formate, 0.1 M Tris buffer (pH 8.0). Protein concentration was 1.2 mg/ml. **(b)** T18Q mutant crystals. Crystallisation solution: 21 % ($w/v$) PEG 4000, 0.2 M magnesium formate, 0.1 M Tris buffer (pH 8.0). Protein concentration was 1.1 mg/ml, and the crystals were seeded by streak seeding. **(c)** C47V/G204A/S239D mutant crystals. Crystallisation solution: 18 % ($w/v$) PEG 3350, 0.2 M magnesium formate, 0.1 M Bis-tris methane buffer (pH 6.5). Protein concentration was 3.2 mg/ml, and the crystals were seeded by streak seeding.

The T18Q mutant required a larger PEG 4000 concentration (at least 23 %) to crystallise, however the crystals then growed too fast and became disordered. The crystals were broken and used in streak seeding: new droplets were prepared with the same protein concentration and lower PEG 4000 concentrations, and they were seeded by stirring the broken crystals with a dog hair and immediately streaking the hair through the new droplets. In the same conditions but with 20–22 % PEG 4000, clean crystals were obtained, some two-dimensional and some three-dimensional, with largest dimensions up to 200 µm. (Figure 3.2b).

The C47V/G204A/S239D mutant, in the 3.2 mg/ml solution in Tris buffer, crystallised as large needles in 21 % ($w/v$) PEG 3350, 0.2 M magnesium formate and 0.1 M Bis-tris methane buffer (pH 6.5). When these were broken and used for streak seeding as described above, three-dimensional crystals with largest dimensions up to 200 µm were obtained in the same conditions but with 17–18 % ($w/v$) PEG 3350 (Figure 3.2c).

### 3.3.3 Xylonolactonase

Suitable crystallisation conditions for the *Cc*XylC were searched with Crystal Screen by Hampton Research (HR2-110), and it yielded immediate hits. Needle crystals were observed in condition 4 (2.0 M ammonium sulfate, 0.1 M Tris buffer, pH 8.5), and clusters of crystals were observed in conditions 16 (1.5 M lithium sulfate, 0.1 M HEPES buffer, pH 7.5) and 39 (2.0 M ammonium sulfate, 2 % ($v/v$) PEG 400, 0.1 M HEPES buffer, pH 7.5). Optimisation of the conditions started from these, and eventually small disordered crystals were obtained in 1.6 M ammonium sulfate and 0.1 M HEPES buffer (pH 7.5) by streak seeding. The protein concentration was approximately 7 mg/ml. The droplets were prepared in the cold room at 4 °C and moved to storage at 20 °C after seeding. These crystals were used for streak seeding in another experiment in otherwise the same conditions but varying ammonium sulfate concentration 1.6–1.8 M, which yielded tiny needle crystals. Streak seeding again with these in the same conditions but 1.5 M ammonium sulfate yielded two-dimensional rectangular crystals with largest dimensions up to 200 µm. Also, similar crystals were obtained by microseeding: in the preparation of the droplet, instead of adding 1 µl of crystallisation solution, previous droplets with tiny needle crystals were diluted 1:320 with the crystallisation solution and used instead. Cocrystallisation with substrate analogues was tried but had no visible effect on crystallisation or the shape of the crystals, except that no crystallisation occurred in presence of D-xylonolactone or D-xylonic acid. A picture of crystals similar to these is shown in Figure 3.3a.

A different crystal form was also observed: in presence of PEG 400, the protein crystallised as parallelogram-shaped two-dimensional plates. Direct crystallisation in 1.6 M ammonium sulfate, 2 % ($v/v$) PEG 400, 0.05 M sodium malonate and 0.1 M HEPES buffer (pH 7.5) yielded severely twinned crystals. Using these for streak seeding and crystallising in the same conditions yielded clean parallelogram-shaped crystals which were small but useable, with largest dimensions up to 150 µm. A picture of these crystals is shown in Figure 3.3b. The protein concentration was approximately 8 mg/ml, and these were also kept at 4 °C during preparation.

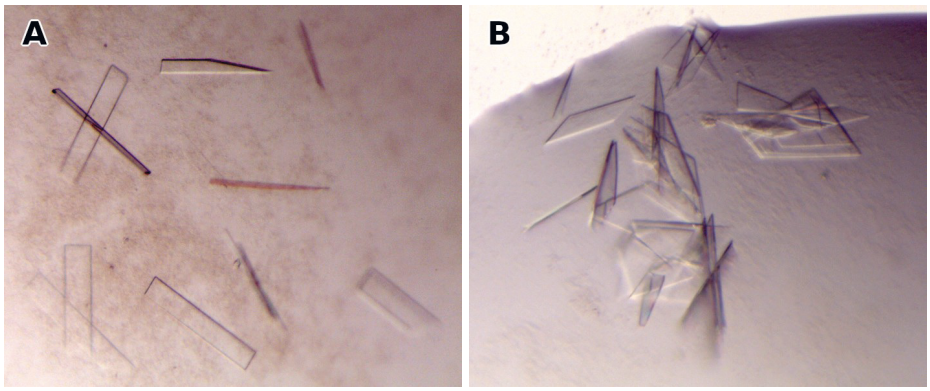



**Figure 3.3:** *Cc*XylC crystals in crystallisation droplets. **(a)** Rectangular crystals. Crystallisation solution: 1.5 M ammonium sulfate, 0.025 M (s)-5-hydroxymethyl-2-pyrrolidone, 0.1 M HEPES buffer (pH 7.5). Protein concentration was 7 mg/ml, and the crystals were seeded by microseeding. **(b)** Parallelogram-shaped crystals. Crystallisation solution: 1.6 M ammonium sulfate, 2 % ($v/v$) PEG 400, 0.05 M sodium malonate, 0.1 M HEPES buffer (pH 7.5). Protein concentration was 8 mg/ml, and the crystals were seeded by streak seeding.

## 3.4 X-RAY CRYSTALLOGRAPHY

### 3.4.1 Overview

When a coherent X-ray beam with a wavelength typically close to 1 Å ($10^{-10}$ m) is shot into crystalline material, it is diffracted. While most of the radiation passes through the crystal, planes of atoms reflect fractions of it in a way characteristic of the contents of the repeating unit cell. These reflections are detected with a two-dimensional detector, and by repeating the irradiation in a range of angles, a three-dimensional picture of the diffraction pattern of the crystal is built. This is known as the reciprocal lattice which is reciprocally connected to the real lattice of the crystal. [61]

The periodic real lattice is described as a periodic electron density function which is written as a Fourier series or, in other words, a sum of complex sinusoidal functions (structure factors) with different amplitudes, frequencies and phases. The Fourier series is truncated at some high frequency based on the quality of the data, and the corresponding wavelength is known as the high resolution limit or just resolution [61]. The quality is measured with parameters including data set completeness, signal-to-noise-ratio $I/\sigma_I$ and correlation coefficient $CC_{1/2}$ between random halves of the data [62]. The reciprocal lattice is the Fourier transform of this function, and so in theory the structural factors would be obtained by performing an inverse Fourier transform on the reciprocal lattice [61]. In practice, the phases of the reflections are impossible to measure, so they must be obtained indirectly. Phasing in macromolecular crystallography is commonly done using molecular replacement, where the phases are estimated using a preexisting homologous structure and are improved as the structure is built to match the electron density [63].

The structure is refined manually in a graphical editor and computationally. By visual inspection of the electron density, the structure can be modified to fit into it, and ligands and water molecules can be added if appropriate electron density peaks

are visible. Computational refinement methods modify the model to fit the electron density better and to have the structure factors match the observed amplitudes, meanwhile keeping restraints on the geometry (including bond lengths, bond angles and torsion angles) of the model. [64]

The correlation between the structure factors and the observed amplitudes is described by the *R*-factor which should be small, with values typically less than 0.2 with good-quality data. There are two *R*-factors in common use: $R_{\text{free}}$ is calculated using a small fraction of the reflections that are ignored in structure refinement, and $R_{\text{work}}$ is calculated using all the rest. $R_{\text{work}}$ will be improved if the model is modified to fit the data better, but $R_{\text{free}}$ will cease to improve once the model becomes overdetailed. Therefore, $R_{\text{free}}$ is a useful measure of the quality of the model with respect to the data. [65]

There are also other measures of model quality. The uncertainty of atomic positions is described by the *B*-factor or the displacement parameter, and large values indicate large displacement or disorder. The Ramachandran plot [66] checks whether the pair of torsion angles in the protein backbone are in certain favourable regions. Nearly all should be, and a lot of outliers indicates that there are errors in the model geometry. [65]

### 3.4.2 D-2-Deoxyribose-5-phosphate aldolase

The *Ec*DERA mutant crystals were soaked in solutions containing substrates or substrate analogues. Using a nylon loop, they were transferred to fresh droplets of the respective crystallisation solution plus 0.1 M ligand and kept there for approximately two days. The tested ligands were acetaldehyde, ethylene glycol, propylene glycol, glycerol, DR and DRP, all shown in Figure 3.4.

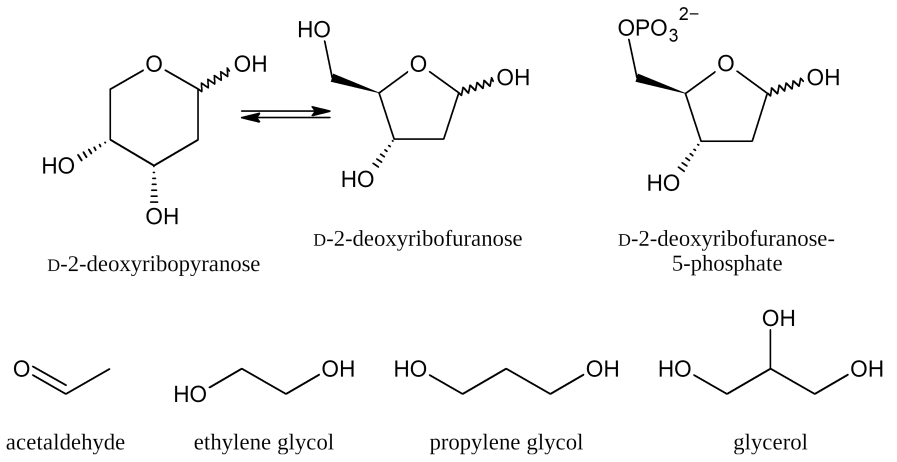



**Figure 3.4:** All the ligands that were used in the soaking experiments with the *Ec*DERA crystals. DR is shown in the pyranose and furanose forms which dominate in aqueous solution. DRP is shown in the furanose form in which it appears in solution, nucleotides and DNA.

After the two-day soaking, the crystals were transferred to cryoprotective solutions and stored in liquid nitrogen. The cryoprotective solutions were prepared by replicating the crystallisation conditions but increasing the PEG concentration to 40 % $(w/v)$. The crystals were removed from the soaking droplets with a nylon loop, set in droplets of cryoprotective solution and immediately picked up in the loops again to be stored in sample pucks in liquid nitrogen. The pucks were then shipped to European synchrotrons.

Crystals of the N21K and T18Q mutants were measured at the ID30A-1 beamline at the European Synchrotron Radiation Facility (ERSF) in Grenoble, France. The samples were set up and measured automatically, and the images and autoprocessed data were returned. The files created by *EDNA_proc* [62, 67, 68] were selected as the data files for the structure calculations.

A crystal of the N21K mutant with DRP was also measured using the home X-ray diffractometer at the Department of Chemistry. The instrument comprised a Nonius FR591 rotating anode X-ray source by Bruker, a mar345dtb goniometer and a mar345 detector by X-ray Research (currently marXperts). The sample was kept at 100 K in a flow of air cooled down with liquid nitrogen. The measurement was set up and the images were collected using the bundled *mar345dtb* software, and the images were processed using *XDS* program package [68].

Crystals of the C47V/G204A/S239D mutant were measured at the i24 beamline at Diamond Light Source (DLS) in Oxford, England. The samples were sent there and measured remotely. The autoprocessed files were downloaded, and those created by *xia2-3dii* [62, 67, 69–71] were selected to be used.

All structure calculations were done in *Phenix* software suite [72]. Phasing was done using molecular replacement with *Phaser* [73], using a structure of wild-type *Ec*DERA (PDB entry 1KTN [74]) as the starting model. Structure refinement was done with *phenix.refine* [64] which used *MolProbity* [75] for structure validation. Visual observation and manipulation of the structure was done in *Coot* [76] and partially in *PyMOL* [77]. Water molecules were added with the "Update waters" option and later edited manually, and the weight optimisations were used in the final refinement cycles. Geometry restraints for the ligands were calculated using *eLBOW* [78].

### 3.4.3 Xylonolactonase

The *Cc*XylC crystals were first measured without ligands. Using conventional cryoprotectants – ethylene glycol, glycerol and sodium malonate – did not work well because the crystals would only diffract to low resolution. A rectangular crystal cryoprotected with a solution containing 3.25 M sodium malonate, 1.0 M ammonium sulfate and 0.1 M HEPES buffer (pH 7.5) was transferred into a nylon loop, stored in liquid nitrogen like the *Ec*DERA crystals and successfully measured on the ID30A-3 beamline at ESRF. Decreasing the ammonium sulfate concentration to 1.0 M was necessary because it would precipitate when cooled down in liquid nitrogen. The crystal diffracted to a resolution of 3.0 Å, enough for building a preliminary structure while awaiting better data. The autoprocessing was unreliable due to an incorrect parameter, so the images had to be processed with *XDS*.

After this, *Cc*XylC complex crystals were prepared by using possible ligands as cryoprotectants. D-Xylonolactone and a number of substrate analogues – D-xylose, 2-piperidone, 2-pyrrolidone, γ-thiobutyrolactone, (S)-5-hydroxymethyl-2-pyrrolidone and (R)-4-hydroxy-2-pyrrolidone (4H2PD) – were tested, all shown in Figure 3.5. The

cryoprotective solutions consisted of 30 % $(w/v)$ cryoprotectant, 1.0 M ammonium sulfate and 0.1 M HEPES buffer (pH 7.5), and for the parallelogram-shaped crystals, 2 % $(v/v)$ PEG 400 was also added. The crystals were transferred to droplets of cryoprotective solution, kept there for a few seconds, picked up in nylon loops and stored in liquid nitrogen. Already during the few seconds, it was noticed that the crystals were unstable in presence of some ligands, including D-xylonolactone.

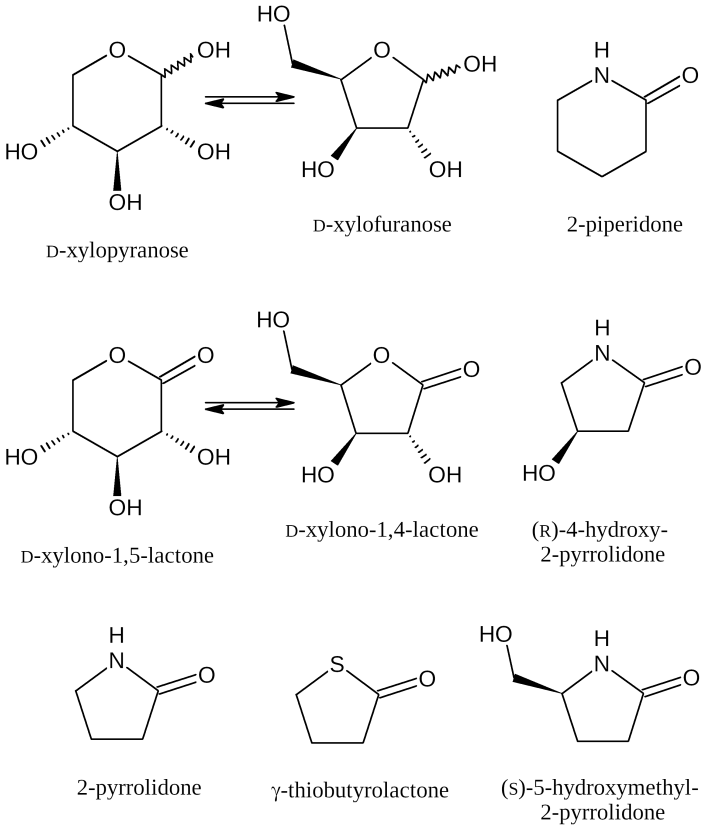



**Figure 3.5:** All the ligands that were tried as cryoprotectants with the *Cc*XylC crystals. D-Xylose is shown in the pyranose and furanose forms, and D-xylonolactone is shown the 1,5- and 1,4-lactone forms, which dominate in aqueous solution.

The complex crystals were measured on the i04-1 beamline at DLS. Useable data were obtained for both crystal forms in presence of D-xylose and the rectangular crystals in presence of 4H2PD. Of the autoprocessed files, ones with the best statistics were selected. For the rectangular crystal with D-xylose, files from *xia2-dials* [62, 67, 69–71], and for the other ones, files from *autoPROC* [62, 67, 68, 71, 79, 80] were used for structure calculations.

The preliminary structure was calculated in *Phenix*. Molecular replacement was attempted with the homologous mouse SMP30 gluconolactonase (PDB entry 4GN7 [33, 36]) in *Phaser*, but it could not find a solution. Thereafter, *MR-Rosetta* [81] was tried, and it also failed, but a partial solution was opened in *Coot* and edited manually to correct all mutations and to place the backbone to the calculated electron density. Molecular replacement with this model in *Phaser* was successful, and after

further editing and refining with *phenix.refine*, this preliminary structure could be used for the molecular replacement for the higher-resolution structures.

For the complex structures, molecular replacement was done with *Phaser*, and refinement was done with *phenix.refine* as with the *Ec*DERA structures. The geometry restraints of the anomers of D-xylopyranose were already in the internal library, whereas for 4H2PD they had to be calculated with *eLBOW*.

# 4 RESULTS AND DISCUSSION

## 4.1 MASS SPECTROMETRY

### 4.1.1 Xylonolactonase

The denatured mass spectrum of the *Cc*XylC (Figure 4.1) featured a typical charge state distribution for denatured protein, and charge state deconvolution (Figure 4.2) yielded the most abundant isotopic mass (31524.89 $\pm$ 0.10) Da (mean $\pm$ standard deviation). This matches the theoretical mass of the apo-enzyme calculated using the amino acid sequence, omitting M1. The native mass spectrum (Figure 4.3) featured four observable charge states, and in addition to the apo-enzyme another significant species was observed. The apparent mass difference, $\sim$53 Da, corresponds to the binding of a $Fe^{3+}$ ion ($\sim$56 Da) and three fewer protons. Charge state deconvolution of the native mass spectrum is shown in Figure 4.4.

The subsequent interest was whether iron and other metals would bind to the apo-enzyme. The solution of EDTA-treated *Cc*XylC was measured, and as expected, only the apo-form was observed. Adding roughly 20 µM $Fe^{2+}$ as iron(II) chloride to the 1 µM enzyme caused the apo-form to be almost completely converted to the complex and its signal to almost disappear (Figure 4.5). This was not, however, observed with any other metal ion: with 20 µM $Mg^{2+}$, $Ca^{2+}$, $Fe^{3+}$, $Co^{2+}$, $Ni^{2+}$ or $Zn^{2+}$, the mass spectrum was indistinguishable from the spectrum of the apo-enzyme, and with 20 µM $Cu^{2+}$, the spectrum became more complicated (Figure 4.6).

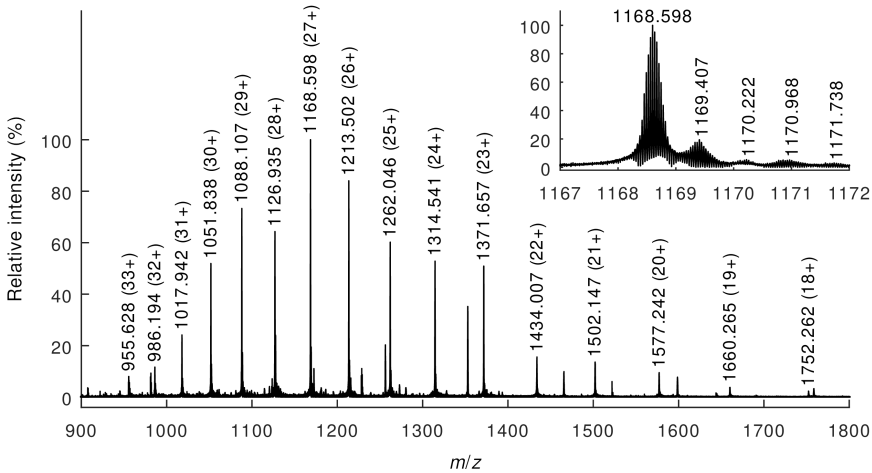



**Figure 4.1:** The denatured mass spectrum of the original *Cc*XylC sample. The inset shows the charge state 27+ magnified.

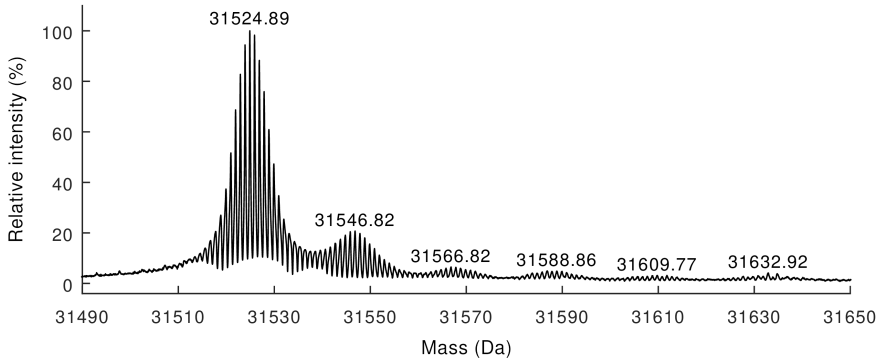



**Figure 4.2:** Charge state deconvolution of the denatured mass spectrum (Figure 4.1) of the *Cc*XylC.

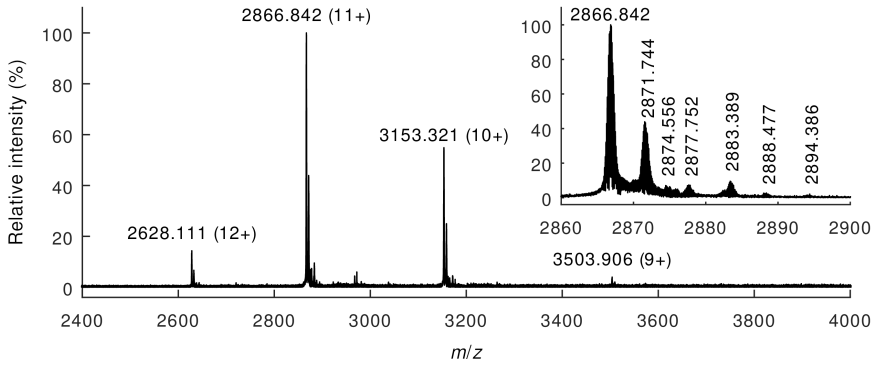



**Figure 4.3:** The native mass spectrum of the original *Cc*XylC sample. The inset shows the charge state 11+ magnified. Mass-to-charge ratios 2866.842 and 2871.744 correspond to the apo- and holo-forms of the enzyme respectively.

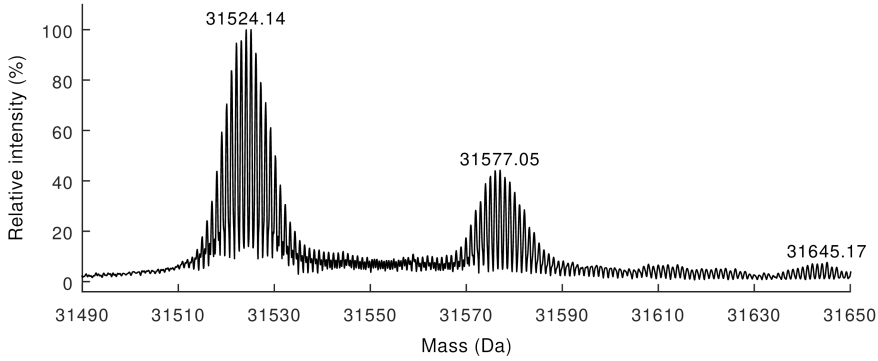



**Figure 4.4:** Charge state deconvolution of the native mass spectrum (Figure 4.3) of the *Cc*XylC. Masses 31524.14 Da and 31577.05 Da correspond to the apo- and holo-forms of the enzyme respectively.

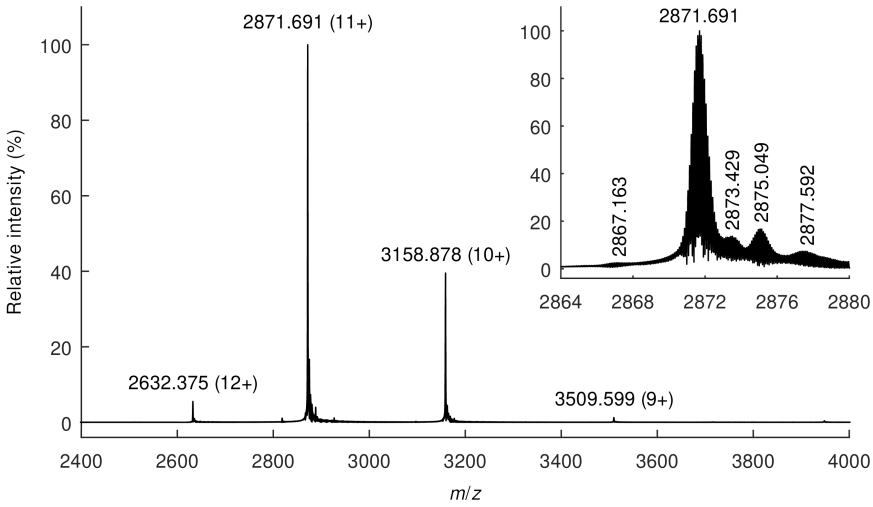



**Figure 4.5:** Native mass spectrum of the experiment with roughly 1 µM *Cc*XylC and roughly 20 µM $Fe^{2+}$. The inset shows the charge state 11+ magnified. The mass-to-charge ratios 2871.691, 2873.429 and 2875.049 correspond to the holo-form of the enzyme, holo-form plus water and holo-form plus two waters respectively.

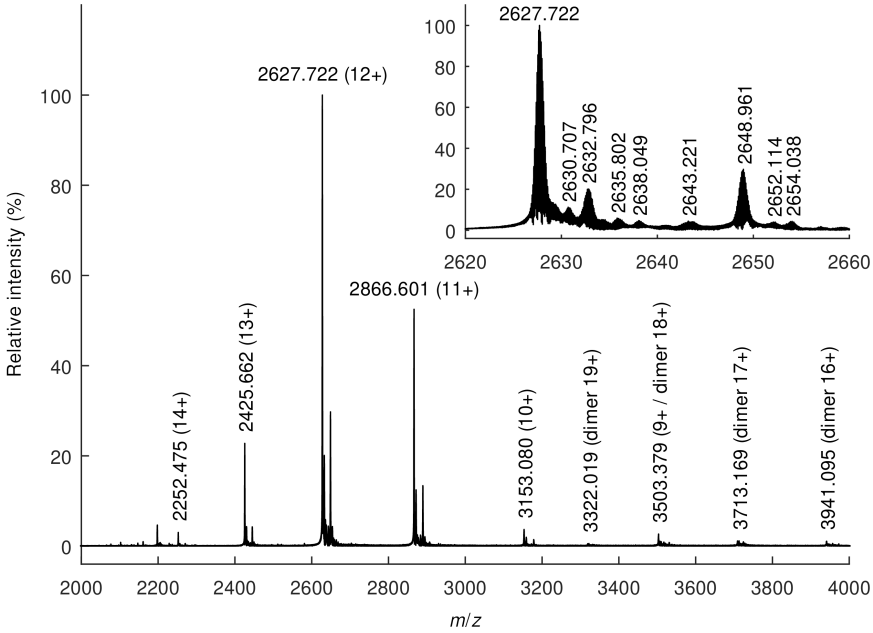



**Figure 4.6:** Native mass spectrum of the experiment with roughly 1 µM *Cc*XylC and roughly 20 µM $Cu^{2+}$. The inset shows the charge state 12+ magnified. The mass-to-charge ratios 2627.722, 2632.796 and 2648.961 likely correspond to the apo-form of the enzyme, apo-form plus copper and apo-form plus two coppers respectively. This is discussed in more detail in the text.

With 20 μM $Cu^{2+}$, new signals of protein species appeared in the spectrum, and dimeric charge states could be observed (Figure 4.6). In the inset of the figure, the neutral mass of the signal probably corresponding to the apo-enzyme ($m/z$ 2627.722, neutral mass 31520.58 Da) is ∼4 Da lower than expected. The second strong signal at $m/z$ 2632.796 (neutral mass 31581.46 Da) matches the expected apparent mass increase of ∼61 Da when a $Cu^{2+}$ ion is bound (∼63 Da minus two fewer protons), but the third strong signal at $m/z$ 2648.961 (neutral mass 31775.44 Da) does not exactly match the binding of four $Cu^{2+}$ ions, which would have a neutral mass of ∼31764 Da. There are four cysteines in a protein monomer, and the copper ions were likely bound to them. This was not investigated further as it was deemed an uninteresting result.

In the spectrum of the *Cc*XylC and 20 μM $Fe^{2+}$, other interesting forms of the holo-enzyme were observed. In the inset of Figure 4.5, the signals at $m/z$ 2873.429 and 2875.049 correspond to the apo-form plus a $Fe^{2+}$ ion and one and two water molecules respectively. It would be expected that all water molecules are extracted from the enzyme in the ionisation process and in high vacuum, but in this case they are coordinated to the iron strongly enough that they remain partially bonded. Similar behaviour has been reported before, for instance by Borchers *et al.* [82], where a water molecule remains bonded to the *E. coli* cytidine deaminase. Since $Fe^{3+}$ ions in solution did not bind to the enzyme, the conclusion is that the iron exists as $Fe^{2+}$ in the structure and is oxidised into $Fe^{3+}$ in the mass spectrometer if both water molecules are extracted. Alternatively, if the $Fe^{2+}$ is oxidised into $Fe^{3+}$ in the ionisation process, it loses the coordinated waters but remains bonded to the enzyme in gas-phase.

For each of the eight native mass spectra of the *Cc*XylC–$Fe^{2+}$ samples (Figure 4.7), the equilibrium concentrations of apo- and holo-forms were calculated. The intensities of all apo- and holo-form signals (including the signals with bound water) were calculated using *DataAnalysis* 5.0 software by Bruker and summed. Assuming that the intensities were proportional to concentrations, the fractional saturations $B_C$ (relative amounts of holo-enzyme) were calculated. Using them and the total concentrations of enzyme and $Fe^{2+}$, concentrations of free $Fe^{2+}$ were also calculated. The titration curve (Figure 4.8) was created by plotting the fractional saturations over the concentrations of free $Fe^{2+}$.

This complex formation is the reaction described in Section 2.2. The model used in this calculation, modified from the simple form used in the simulation, accounts for the maximum saturation level $B_{max}$ which would be unity in the ideal case. The equation is

$$B_C \equiv \frac{[PL]}{c_P} = \frac{B_{max}\,[L]}{K_D + [L]}, \tag{4.1}$$

where [PL] and [L] are the equilibrium concentrations of the holo-enzyme and free iron respectively, $c_P$ is total enzyme concentration and $K_D$ is the dissociation constant. Unweighted orthogonal regression in *Origin Pro 2018* software [83] yielded $K_D = (5.0 \pm 1.3) \cdot 10^{-7}$ mol $l^{-1}$ and $B_{max} = 0.966 \pm 0.009$ with a 95 % level of confidence.

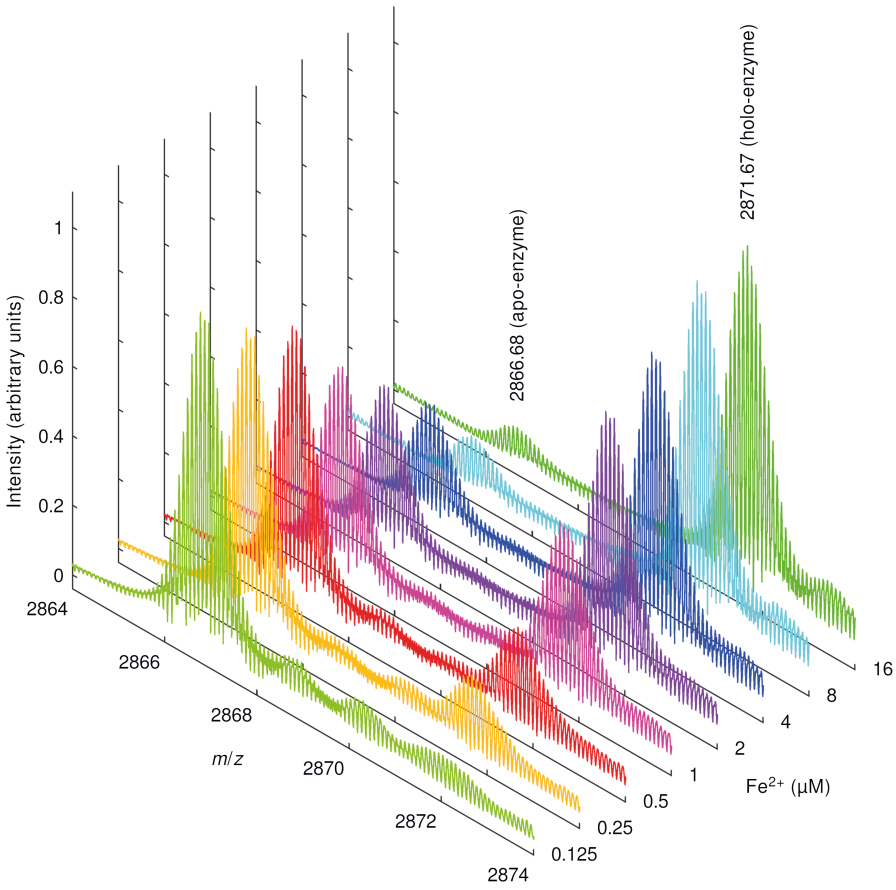



**Figure 4.7:** Charge states 11+ of the native mass spectra of the titration of *Cc*XylC with $Fe^{2+}$. This shows the trend that the equilibrium of apo- and holo-enzymes shifts to the holo-enzyme when the amount of iron in solution is increased. The $Fe^{2+}$ concentrations in the bottom-right are the approximate total concentrations of iron chloride added to each sample. The enzyme concentration was equal in each case and approximately 1 μM.

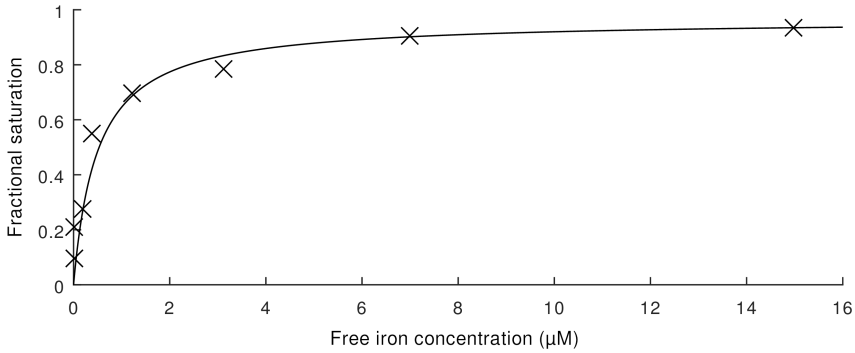



**Figure 4.8:** The titration of the *Cc*XylC with $Fe^{2+}$. The crosses are the data points calculated from the mass spectra shown in Figure 4.7. The line is the fitted theoretical curve of Equation (4.1) with parameters $K_D = 5.0 \cdot 10^{-7}$ mol $l^{-1}$ and $B_{max} = 0.966$.

### 4.1.2 Lactone hydrolysis

When the lactone hydrolysis reaction progressions were followed while measuring them, it was immediately noticed that presence of apo-enzyme did not accelerate the reaction and that both metal ions and holo-enzyme did accelerate it significantly. The calculated rate coefficients of the reactions are shown in Table 4.1. Also shown are reaction half-lives calculated from the rate coefficients.

**Table 4.1:** Measured pseudo-first-order rate coefficients and half-lives of the lactone hydrolysis reactions. The values of the rate coefficients and the half-lives are confidence intervals with a 95 % level of confidence. The substrates are the isomers both reported in the reagent containers and known to be the dominant forms in aqueous solution [34, 84].

| Substrate | *Cc*XylC (μM) | $Fe^{2+}$ (μM) | Rate coefficient ($s^{-1}$) | Half-life (min) |
|---|---|---|---|---|
| D-xylono-1,4-lactone | — | — | $(3.5 \pm 0.5) \cdot 10^{-5}$ | $330 \pm 40$ |
| D-xylono-1,4-lactone | — | 10 | $(8.3 \pm 0.5) \cdot 10^{-5}$ | $138 \pm 8$ |
| D-xylono-1,4-lactone | 0.5 | 10 | $(2.7 \pm 0.2) \cdot 10^{-3}$ | $4.3 \pm 0.3$ |
| D-glucono-1,5-lactone | — | — | $(2.6 \pm 0.3) \cdot 10^{-4}$ | $45 \pm 5$ |
| D-glucono-1,5-lactone | — | 10 | $(4.0 \pm 0.6) \cdot 10^{-4}$ | $29 \pm 4$ |
| D-glucono-1,5-lactone | 0.5 | 10 | $(2.4 \pm 0.2) \cdot 10^{-3}$ | $4.8 \pm 0.4$ |

The hydrolysis of xylonolactone without any metals occurred slowly with a half-life of 330 min. Presence of 10 μM $Fe^{2+}$ accelerated the reaction approximately two-fold to a half-life of 138 min, indicating that the metal ion catalyses some step in the reaction. Presence of 0.5 μM holo-enzyme accelerated the reaction almost 80-fold to a half-life of 4.3 min, indicative of enzymatic catalysis. With the holo-enzyme, there is also approximately 9.5 μM free $Fe^{2+}$ in solution, so the effect of free $Fe^{2+}$ is present as well. Likewise, the hydrolysis of gluconolactone had a half-life of 45 min without any metals, and presence of 10 μM $Fe^{2+}$ or 0.5 μM holo-enzyme accelerated the reaction 1.5-fold and almost 10-fold to half-lives of 29 min and 4.8 min respectively. It is remarkable that whereas the nonenzymatic reactions of gluconolactone

were much faster, the enzymatic reactions of both xylono- and gluconolactone were nearly equally fast. From the determined rate coefficients of the enzymatic reactions, estimates of enzyme specificity constants ($k_{cat}/K_m$) [85] were calculated, and they were $(5400 \pm 900)\ M^{-1}\ s^{-1}$ and $(4900 \pm 900)\ M^{-1}\ s^{-1}$ for the reactions with xylonolactone and gluconolactone respectively.

The hypothesis arising from these data is that the bare metal ion affects the interconversion of the lactone forms. There are two lactone forms, 1,4-lactone and 1,5-lactone. It has been reported that glucono-1,5-lactone is hydrolysed to gluconic acid in aqueous solution whereas glucono-1,4-lactone is not, and the lactone forms interconvert without a gluconic acid intermediate [35]. This is assumed to apply for the forms of xylonolactone as well. Also, the 1,5-lactone, as a six-membered ring, appears to fit better to the active site of the enzyme (discussed in Section 4.2.2), and so it should be the primary substrate.

The proposition is that the interconversion occurs via a bicyclic intermediate that is stabilised by the metal ion. Xylonolactone exists primarily as 1,4-lactone [34], and so it must be converted to 1,5-lactone, which is then hydrolysed to xylonic acid either enzymatically or nonenzymatically. Presence of metal ions accelerates the interconversion. In contrast, gluconolactone exists primarily as 1,5-lactone [84], so the hydrolysis occurs faster, but even then the product is formed faster when the lactone form interconversion is accelerated. The reaction equation is presented in Figure 4.9.

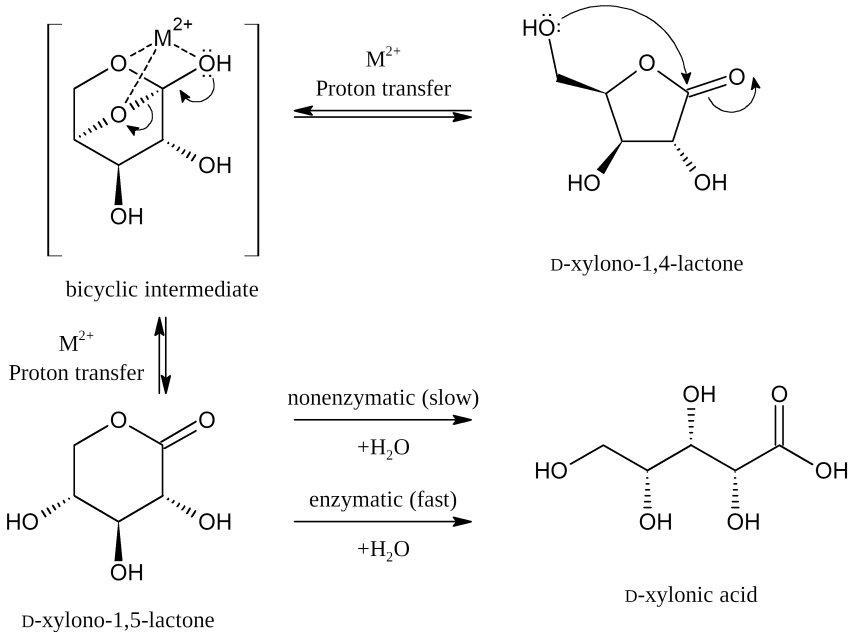



**Figure 4.9:** Proposed reaction mechanism of the hydrolysis of D-xylonolactone. D-Xylono-1,4-lactone and D-xylono-1,5-lactone interconvert via a bicyclic intermediate stabilised by a bivalent metal ion ($M^{2+}$, here $Fe^{2+}$), and D-xylono-1,5-lactone is hydrolysed to D-xylonic acid either enzymatically or nonenzymatically. The hydrolysis of D-gluconolactone to D-gluconic acid is analogous.

## 4.2 X-RAY CRYSTALLOGRAPHY

### 4.2.1 D-2-Deoxyribose-5-phosphate aldolase

Several crystals of the *Ec*DERA diffracted to a resolution less than 2 Å. The structures that were deemed to be interesting enough for publication (Publication **I**) were N21K mutant with ethylene glycol (best resolution of all), N21K mutant with DRP and C47V/G204A/S239D mutant with DRP. Also, a structure of T18Q mutant with acetaldehyde is considered here. Data of these structures are presented in Table 4.2.

All the *Ec*DERA structures have the same space group $P\ 1\ 2_1\ 1$ and nearly the same unit cell dimensions, and the crystal packings are the same. The asymmetric unit (ASU) contains two protein chains, and the tertiary structure is the expected TIM $(\alpha/\beta)_8$ barrel fold. The protein backbones and the amino acid side chains have strong electron density, except for some side chains on the surface.

**Table 4.2:** Data of the considered *Ec*DERA crystal structures. Numbers in parentheses refer to the diffraction shell of the highest resolution.

| ***Ec*DERA mutant** | **N21K** | **N21K** | **T18Q** | **C47V/G204A/S239D** |
|---|---|---|---|---|
| **PDB entry** | **6Z9J** [86] | **6Z9I** [87] | **—** | **6Z9H** [88] |
| Ligand | ethylene glycol | DRP | acetaldehyde | DRP |
| Beamline | ESRF ID30A-1 | home source* | ESRF ID30A-1 | DLS i24 |
| Wavelength (Å) | 0.9660 | 1.5418 | 0.9660 | 0.9688 |
| Space group | $P\ 1\ 2_1\ 1$ | $P\ 1\ 2_1\ 1$ | $P\ 1\ 2_1\ 1$ | $P\ 1\ 2_1\ 1$ |
| $a, b, c$ (Å) | 60.5, 52.8, 80.7 | 61.9, 53.3, 81.2 | 61.8, 53.4, 80.8 | 62.2, 53.4, 80.8 |
| $\alpha, \beta, \gamma$ (°) | 90, 111.2, 90 | 90, 110.2, 90 | 90, 110.7, 90 | 90, 110.7, 90 |
| Resolution (Å) | 31.31–1.50 (1.55–1.50) | 19.27–1.86 (1.97–1.86) | 38.48–1.64 (1.70–1.64) | 39.41–1.72 (1.78–1.72) |
| Observations | 136475 (12275) | 142080 (18431) | 111495 (10585) | 89923 (9206) |
| Unique observations | 74834 (7258) | 39746 (5999) | 59167 (5691) | 51138 (5131) |
| Completeness | 98.3 % (96.0 %) | 94.9 % (90.2 %) | 97.6 % (94.7 %) | 96.9 % (97.8 %) |
| $I/\sigma_I$ | 10.0 (1.1) | 10.5 (2.3) | 8.9 (1.5) | 8.5 (2.1) |
| $CC_{1/2}$ | 0.999 (0.637) | 0.994 (0.710) | 0.997 (0.641) | 0.983 (0.503) |
| $R_{\text{merge}}$ | 0.035 (0.540) | 0.106 (0.550) | 0.046 (0.442) | 0.087 (0.462) |
| $R_{\text{meas}}$ | 0.050 (0.764) | 0.125 (0.664) | 0.065 (0.625) | 0.123 (0.653) |
| $R_{\text{work}}$ | 0.1594 | 0.1611 | 0.1574 | 0.1818 |
| $R_{\text{free}}$ | 0.1867 | 0.2087 | 0.1977 | 0.2266 |
| Mean $B$ factor (Å$^2$) | 24.32 | 17.14 | 23.31 | 23.58 |
| Monomers per ASU | 2 | 2 | 2 | 2 |
| Protein atoms | 3864 | 3833 | 3827 | 3775 |
| Water molecules | 569 | 765 | 703 | 577 |
| Ligand atoms | 9 | 16 | 4 | 8 |
| Ramachandran — | | | | |
| favourable | 97.37 % | 97.56 % | 97.36 % | 97.76 % |
| outliers | 0 % | 0 % | 0 % | 0 % |
| RMSD† of bond — | | | | |
| lengths (Å) | 0.007 | 0.006 | 0.006 | 0.008 |
| angles (°) | 0.931 | 0.792 | 0.801 | 0.944 |

∗ Nonius FR591 rotating anode X-ray source, mar345dtb goniometer, mar345 detector

† Root mean square deviation

The N21K mutant structure with ethylene glycol has no trace of ligands at the active site. The mutation is well visible in the electron density, and its location is at the mouth of the central cavity, indicating that it has a role in regulating the accessibility of the active site (Figure 4.10). A surface model of the cavity is shown in Figure 4.11a. Since the N21K mutant has a decreased DRP cleavage activity, the asparagine in the wild-type probably forms a hydrogen bond to a phosphate oxygen, stabilising the transition state and speeding up the reaction. This would have little effect on DR cleavage and acetaldehyde addition activities, which is the case according to the activity data (Table 1.1).

In the N21K mutant structure with DRP, the reaction products of DRP cleavage – acetaldehyde and G3H – are observed in the active site (Figure 4.11a–c). The acetaldehyde is bound to the amino group of the catalytic K167 as a Schiff base, resembling an ethylidene group (=CH−$CH_3$). The G3H is bound to the side chains of the active site via water molecules and several hydrogen bonds. The mutated N21K side chain is too far to form bonds with the G3H. In the crystal structure, the G3H is observed in only one of two protein chains, and in the other one only the acetaldehyde is visible.

An interesting detail are bound magnesium ions in the crystal structures. While building the models, blobs of electron density in octahedral coordination were noticed at the interface between protein chains. These could not be water molecules due to the geometry, and so they were deduced to be magnesium ions from the magnesium formate that was used in crystallisation. An example of a histidine coordinated to such a magnesium ion is shown in Figure 4.11d.

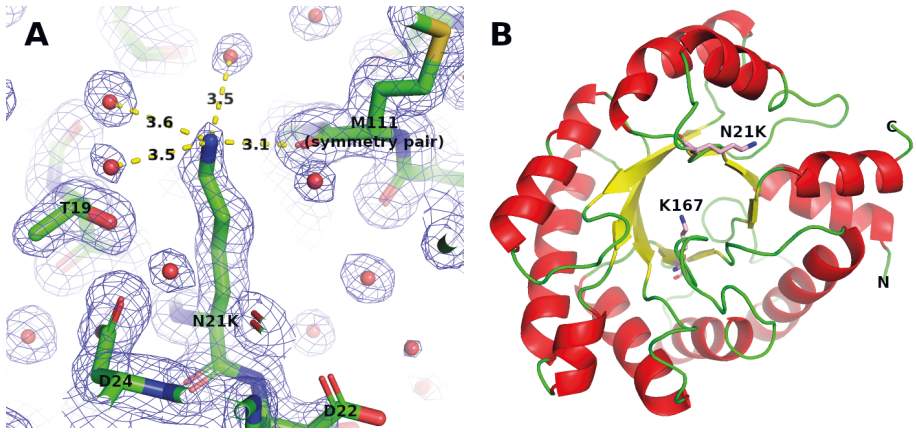



**Figure 4.10:** The *Ec*DERA N21K mutation. **(a)** The electron density of the N21K side chain. Hydrogen bonds are formed from the amino group to water molecules and a peptide oxygen in a symmetry pair. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(b)** The location of the N21K mutation in the tertiary structure. The view is down the central cavity, and the catalytic K167 side chain is also shown. N21K is located at the mouth of the cavity.

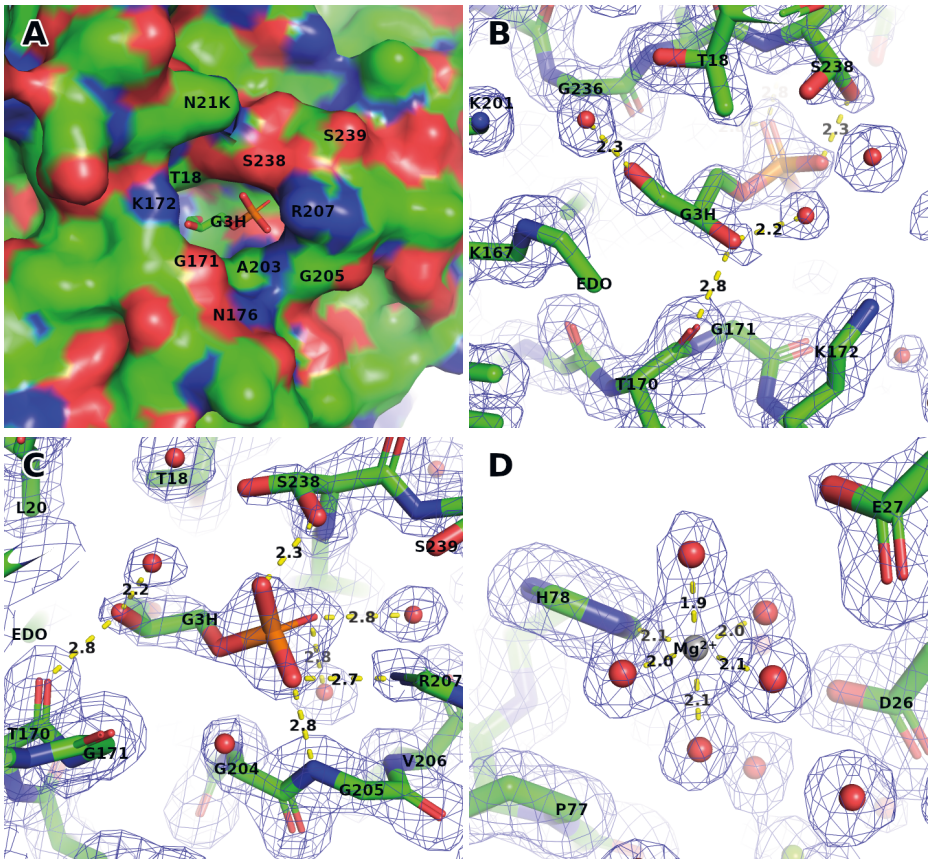



**Figure 4.11:** **(a)** The G3H bound to the surface of the active site of the *Ec*DERA N21K mutant. **(b)** The G3H and the Schiff base form of acetaldehyde (labelled EDO as in the structure in PDB [87]) bound to the active site. View from the end of the aldehyde group and the catalytic K167. The G3H and EDO have refined occupancies of 72 % and 56 % respectively, and S238 has two alternative conformations, of which the one closer to the phosphate group has an occupancy of 63 %. Lengths of the bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(c)** Same, view from the end of the phosphate group. **(d)** A magnesium ion bound to a histidine on the interface of two neighbouring protein chains in the crystal. The model is of the N21K mutant without ligands.

The mutation of the T18Q mutant is also well visible in the electron density (Figure 4.12). It is located at the active site, and the glutamine side chain is larger than threonine, which is visible in Figure 4.11b. Since the T18Q mutant has a severely lower activity compared to the wild-type, the larger side chain of the glutamine probably hinders the binding of ligands to the catalytic K167. At the active site, there is visible electron density of a Schiff base form of acetaldehyde (Figure 4.13a). Because the structure was decided to not be published, the model was not refined to full detail and the Schiff base was not built into the model.

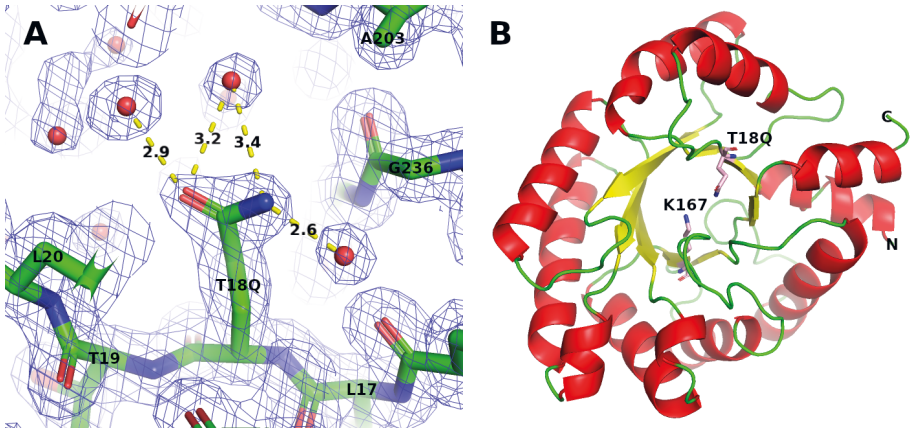



**Figure 4.12:** The *Ec*DERA T18Q mutation. **(a)** The electron density of the T18Q side chain. Hydrogen bonds are formed from the amino group to water molecules nearby. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(b)** The location of the T18Q mutation in the tertiary structure. The view is down the central cavity, and the catalytic K167 side chain is also shown. T18Q is located at the active site close to K167.

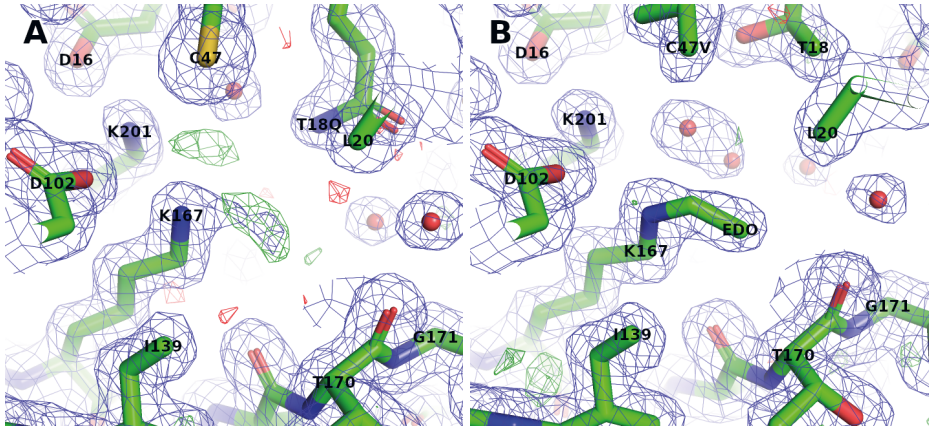



**Figure 4.13:** Acetaldehyde bound as a Schiff base to the catalytic K167 of the *Ec*DERA. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. The difference map is the $F_O - F_C$ map, contoured at $+3\sigma$ level (green) and $-3\sigma$ level (red). **(a)** In the model of the T18Q mutant. The ligand has not been modelled, but electron density corresponding to two carbon atoms is visible. **(b)** In the model of the C47V/G204A/S239D mutant. The modelled Schiff base (labelled EDO as in the structure in PDB [88]) has a refined occupancy of 90 % and strong electron density.

An acetaldehyde is also visible in the active site of the C47V/G204A/S239D mutant (Figure 4.13b), and the mutations are also well visible (Figure 4.14). However, since the mutant is supposed to be inactive towards DRP, there should be little aceteldehyde in solution, but the electron density is unambiguous. Perhaps some acetaldehyde has been formed by nonenzymatic cleavage of DRP over the soak-

ing time and been able to enter the active site. The G204A mutation is located at the mouth of the central cavity, and the larger hydrophobic side chain and a small change in the loop conformation probably prevent DRP and DR from entering the cavity, which explains why the DRP and DR cleavage activities were observed to disappear. The S239D mutation is located farther away from the mouth of the cavity but affects the conformation of the loop and introduces more negative charge near the binding site of the phosphate group of DRP, thereby hindering the DRP binding.

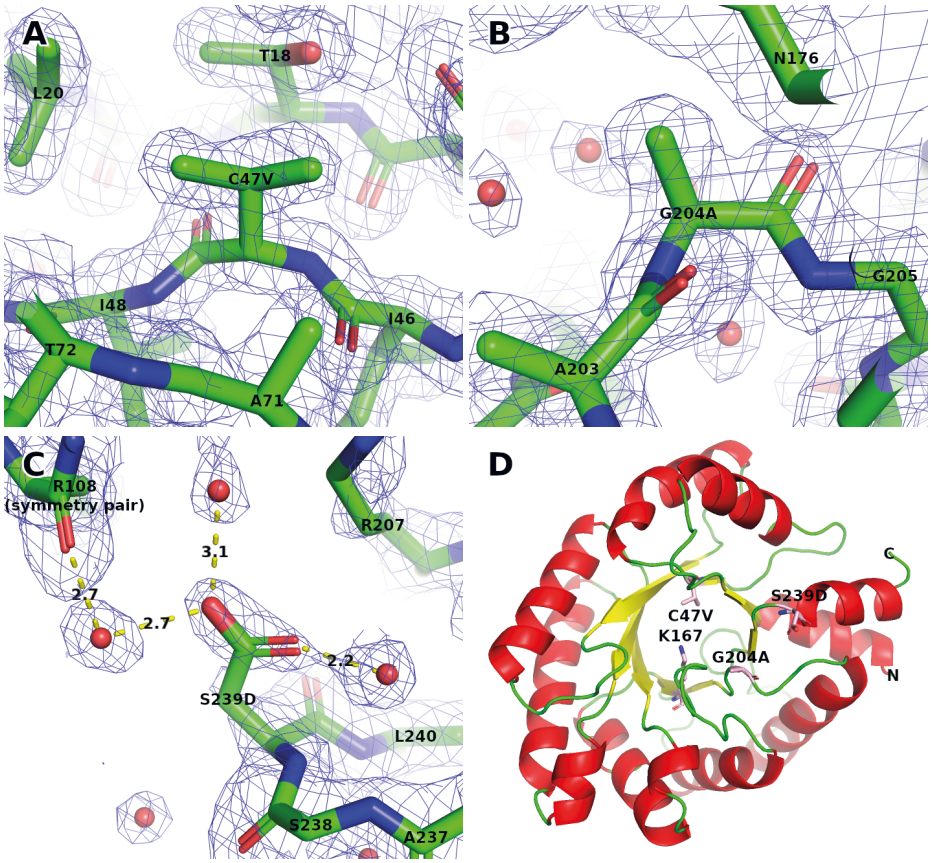



**Figure 4.14:** The *Ec*DERA C47V/G204A/S239D mutations. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(a)** The electron density of the C47V side chain. **(b)** The electron density of the G204A side chain. **(c)** The electron density of the S239D side chain. **(d)** The locations of the mutations in the tertiary structure. The view is down the central cavity, and the catalytic K167 side chain is also shown. C47V and G204A are located at the active site, and S239D is located at the mouth of the cavity.

The C47V mutation is located near the active site. In the structure of the N21K mutant with DRP, the Schiff base of acetaldehyde is in such an orientation that the C–H of the $sp^2$ carbon could form a weak hydrogen bond with the cysteine. However, the carbon–sulfur distance is 4.1 Å in chain A and 4.0 Å in chain B (Figure 4.15a–b), so the bond is very weak. It is shorter than the carbon–carbon distance of the hydrophobic interaction with valine in the C47V/G204A/S239D mutant structure though (Figure 4.15c–d), so the cysteine may stabilise the Schiff base slightly.

No electron density for a ligand bound between the cysteine and the catalytic lysine, as reported by Dick *et al.* [89], is observed. In addition, since valine has a larger side chain than cysteine, the mutation should make the cavity of the active site smaller in volume. Solvent-accessible volumes and mouth areas of the cavities were calculated using the CASTp 3.0 web server [90], and the values (Table 4.3) indicate that the C47V/G204A/S239D mutant has a smaller cavity than the other considered mutants and the wild-type on average. However, the deviation between individual chains is large and explained by variations in loop conformations rather than mutated side chains. Cavity surfaces are visualised in Figure 4.16. These findings do not explain why this C47V mutation causes the DR cleavage activity to decrease significantly. Perhaps the C47 participates in the enzymatic reaction in a way not indicated by these data.

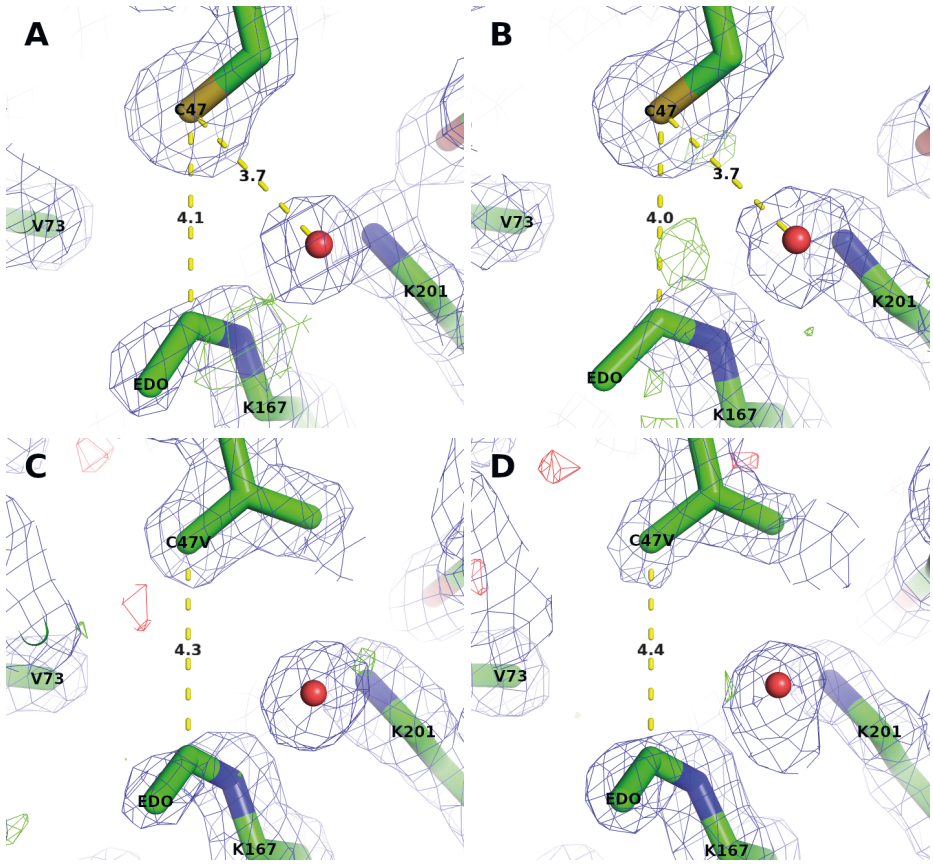



**Figure 4.15:** Comparison of C47 and C47V in the active site of the *Ec*DERA. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. The difference map is the $F_O - F_C$ map, contoured at $+3\sigma$ level (green) and $-3\sigma$ level (red). **(a)** N21K mutant with DRP, chain A. **(b)** N21K mutant with DRP, chain B. **(c)** C47V/G204A/S239D mutant with DRP, chain A. **(d)** C47V/G204A/S239D mutant with DRP, chain B. The sulfur of the cysteine can form a hydrogen bond to a water molecule or a weak hydrogen bond to the Schiff base, though the bond lengths are so long that the hydrogen bonds are very weak. In comparison, the hydrophobic interactions with the mutated valine have a longer equivalent distance.

**Table 4.3:** Solvent-accessible volumes and mouth areas of the central cavities of each chain of the solved *Ec*DERA mutant structures and two wild-type structures. The C47V/G204A/S239D mutant has a smaller cavity than the others on average, but with such a small sample size the difference is not statistically significant. The mouth areas, logically dependent on conformations at the mouth, also seem to be only weakly correlated with the cavity volumes, indicating that there is significant random variation. The values have been calculated with the CASTp 3.0 web server [90] with a probe radius of 1.0 Å.

| ***Ec*DERA mutant** | **N21K** | | **N21K** | |
|---|---|---|---|---|
| **PDB entry** | **6Z9J** [86] | | **6Z9I** [87] | |
| **Protein chain** | **A** | **B** | **A** | **B** |
| Cavity volume (Å$^3$) | 146.007 | 174.846 | 227.878 | 168.759 |
| Mouth area (Å$^2$) | 37.104 | 48.210 | 36.257 | 34.771 |
| ***Ec*DERA mutant** | **T18Q** | | **C47V/G204A/S239D** | |
| **PDB entry** | **—** | | **6Z9H** [88] | |
| **Protein chain** | **A** | **B** | **A** | **B** |
| Cavity volume (Å$^3$) | 199.786 | 125.152 | 144.444 | 94.370 |
| Mouth area (Å$^2$) | 36.847 | 23.689 | 33.184 | 26.934 |
| ***Ec*DERA mutant** | **Wild-type** | | **Wild-type** | |
| **PDB entry** | **1JCL** [11, 16] | | **1P1X** [19, 91] | |
| **Protein chain** | **A** | **B** | **A** | **B** |
| Cavity volume (Å$^3$) | 125.251 | 162.348 | 274.400 | 155.507 |
| Mouth area (Å$^2$) | 28.887 | 32.607 | 59.498 | 50.719 |

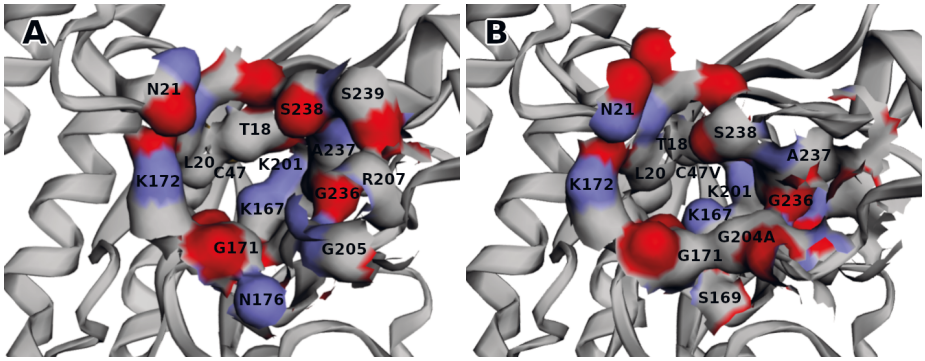



**Figure 4.16:** Surface models of **(a)** the largest and **(b)** the smallest cavity of the *Ec*DERA structures considered in Table 4.3. **(a)** Wild-type (PDB: 1P1X [19, 91]), solvent-accessible cavity volume 274.400 Å$^3$. **(b)** C47V/G204A/S239D mutant, solvent-accessible cavity volume 94.370 Å$^3$. The latter one has a much smaller cavity volume not only because of mutated side chains, but changes in loop and side chain conformations. The pictures have been taken from the model viewer of the CASTp 3.0 web server [90].

### 4.2.2 Xylonolactonase

Complex structures of three *Cc*XylC crystals are considered: rectangular crystal with D-xylose, parallelogram-shaped crystal with D-xylose and rectangular crystal with 4H2PD. Data of these structures, and also the low-resolution preliminary structure, are presented in Table 4.4.

The two visibly different crystal forms also exhibit different space groups, unit cells and packings. Disregarding the preliminary structure, which may have been processed to unnecessarily low symmetry due to low data resolution, the rectangular and parallelogram-shaped crystals have space groups $C\ 2\ 2\ 2_1$ and $P\ 1\ 2_1\ 1$ as well as orthorhombic and monoclinic unit cells respectively. The protein monomers are packed differently in the two different lattices as shown in Figures 4.17 and 4.18. The expected six-bladed β-propeller fold is present, and there is no signifcant difference in backbone conformation between the two crystal packings.

**Table 4.4:** Data of the considered *Cc*XylC crystal structures and the unpublished preliminary structure. Numbers in parentheses refer to the diffraction shell of the highest resolution.

| **Structure / PDB entry** | **Preliminary** | **7PLB** [92] | **7PLC** [93] | **7PLD** [94] |
|---|---|---|---|---|
| Crystal form | rectangle | rectangle | parallelogram | rectangle |
| Ligand | none | D-xylopyranose | D-xylopyranose | 4H2PD |
| Beamline | ESRF ID30A-3 | DLS i04-1 | DLS i04-1 | DLS i04-1 |
| Wavelength (Å) | 0.96770 | 0.91587 | 0.91587 | 0.91587 |
| Space group | $P\ 1\ 2_1\ 1$ | $C\ 2\ 2\ 2_1$ | $P\ 1\ 2_1\ 1$ | $C\ 2\ 2\ 2_1$ |
| $a, b, c$ (Å) | 87.2, 79.6, 96.5 | 85.8, 171.0, 79.1 | 45.8, 82.2, 159.2 | 87.7, 171.7, 79.2 |
| $\alpha, \beta, \gamma$ (°) | 90, 116.9, 90 | 90, 90, 90 | 90, 97.7, 90 | 90, 90, 90 |
| Resolution (Å) | 47.92–3.00 (3.18–3.00) | 85.58–1.73 (1.79–1.73) | 158.24–2.15 (2.23–2.15) | 85.97–1.70 (1.76–1.70) |
| Observations | 50983 (8155) | 121864 (11959) | 115401 (11849) | 131593 (12969) |
| Unique observations | 21347 (3491) | 61016 (6025) | 59670 (6048) | 65883 (6493) |
| Completeness | 89.1 % (91.1 %) | 100.0 % (99.9 %) | 93.5 % (95.2 %) | 99.7 % (98.9 %) |
| $I/\sigma_I$ | 7.0 (1.7) | 9.2 (1.1) | 3.9 (1.0) | 9.2 (1.0) |
| $CC_{1/2}$ | 0.986 (0.613) | 0.998 (0.564) | 0.991 (0.660) | 0.999 (0.538) |
| $R_{\text{merge}}$ | 0.119 (0.620) | 0.049 (0.719) | 0.100 (0.551) | 0.048 (0.784) |
| $R_{\text{meas}}$ | 0.149 (0.768) | 0.069 (1.017) | 0.141 (0.779) | 0.068 (1.109) |
| $R_{\text{work}}$ | 0.2164 | 0.1807 | 0.1978 | 0.1936 |
| $R_{\text{free}}$ | 0.2696 | 0.2178 | 0.2502 | 0.2343 |
| Mean $B$ factor (Å$^2$) | 37.39 | 28.34 | 28.58 | 27.30 |
| Monomers per ASU | 4 | 2 | 4 | 2 |
| Protein atoms | 8845 | 4504 | 8829 | 4457 |
| Water molecules | 63 | 553 | 611 | 579 |
| Ligand atoms | 0 | 147 | 89 | 86 |
| Ramachandran — | | | | |
| favourable | 88.33 % | 97.73 % | 97.71 % | 97.37 % |
| outliers | 2.54 % | 0 % | 0 % | 0 % |
| RMSD of bond — | | | | |
| lengths (Å) | 0.008 | 0.003 | 0.003 | 0.006 |
| angles (°) | 1.418 | 0.692 | 0.630 | 0.815 |

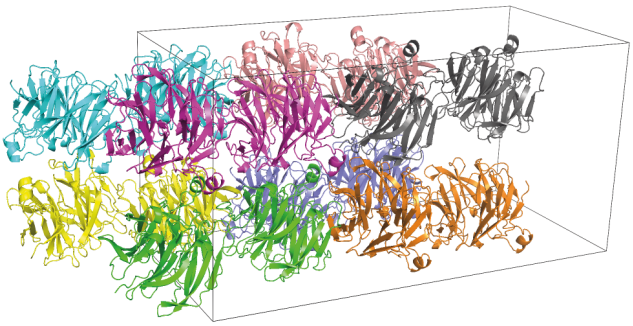


**Figure 4.17:** The unit cell of the rectangular crystal, space group $C\ 2\ 2\ 2_1$. There are eight copies of the asymmetric unit, each coloured differently. The protein monomers are arranged in orthogonal vertical planes that are packed together with a horizontal offset of half a monomer's size.

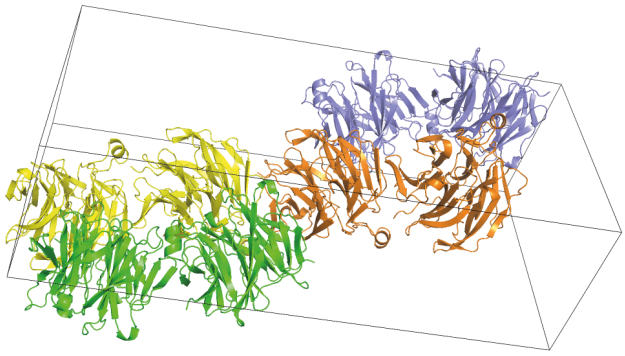


**Figure 4.18:** The unit cell of the parallelogram-shaped crystal, space group $P\ 1\ 2_1\ 1$. There are two copies of the asymmetric unit, each coloured with two different colours to aid legibility. The asymmetric unit of the model file contains the green and yellow chains or, by symmetry, the orange and purple chains. All the displayed chains are roughly in the same plane. Compared to the packing of the rectangular crystal (Figure 4.17), the chains are not orthogonally ordered in the planes, and the offset between planes is not a fraction of monomer size but more coincidental.

The bound iron is observed in the active site, and it is bound to the side chains of E18, N146 and D196. Also, two water molecules and a ligand or a third water molecule are bound to the iron. The coordination geometry is octahedral with little distortion. The occupancy of each modelled iron ion was refined, and the occupancies ranged from 47 % to 93 %. The presence of iron was verified by dissolving crystals in water, desalting, transferring to ammonium acetate solution and measuring a native mass spectrum (Figure 4.19) as described earlier.

The two structures with D-xylose contain several D-xylopyranose molecules both on the surface and in the active sites. In all six active sites, a β-D-xylopyranose is coordinated to the $Fe^{2+}$ via the number 1 hydroxyl group and in a specific orientation (Figure 4.20). The electron densities are strong enough that they could be modelled with high certainty and 100 % occupancy. On the surface, both α- and β-anomers

are observed, and they also have strong and specific electron densities (Figure 4.21). The β-anomer in the active site is also shown in Figure 4.22a as a surface model.

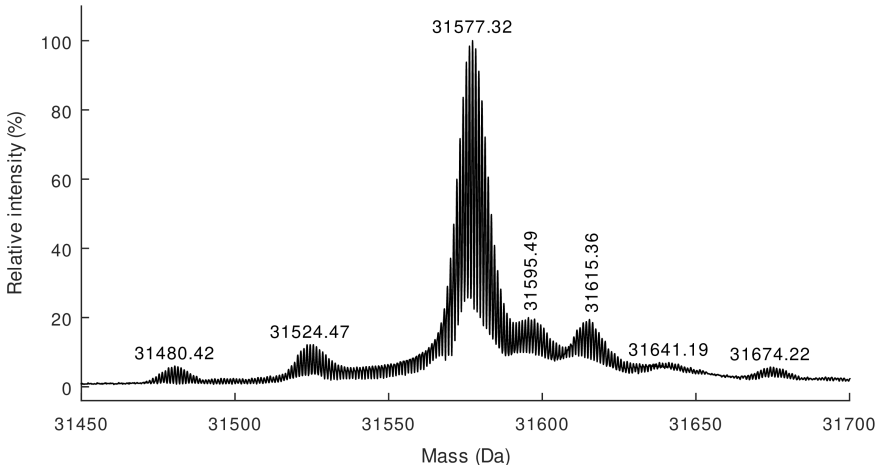



**Figure 4.19:** Charge state deconvolution of the native mass spectrum measured from dissolved *Cc*XylC crystals. The species with the most abundant masses 31524.47 Da, 31577.32 Da, 31595.49 Da and 31615.36 Da are the apo-enzyme, holo-enzyme, holo-enzyme plus water and holo-enzyme plus two waters respectively. The proportional amount of holo-enzyme is more than 90 %.

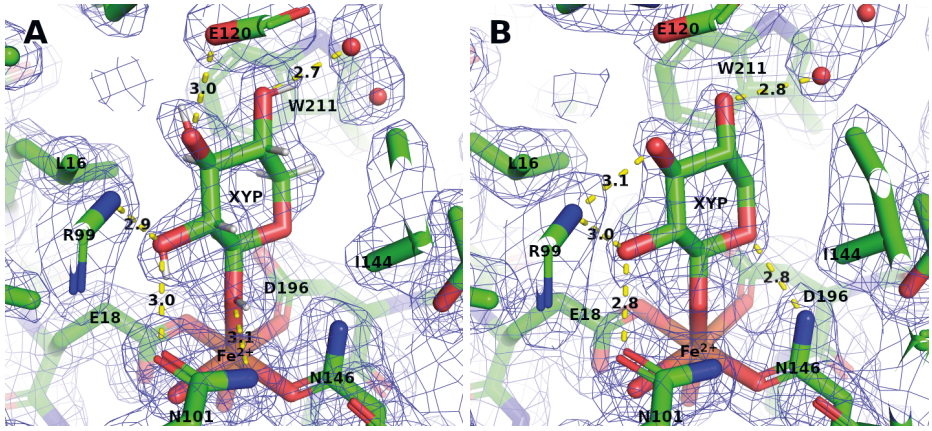



**Figure 4.20:** The β-D-xylopyranose (XYP) bound to the active centre. The number 1 hydroxy group is coordinated to the $Fe^{2+}$ ion, and hydrogen bonds are formed to neighbouring waters and side chains. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(a)** In the structure of the rectangular crystal, resolution 1.73 Å. **(b)** In the structure of the parallelogram-shaped crystal, resolution 2.15 Å. The hydrogen atoms have not been modelled in the lower-resolution structure.

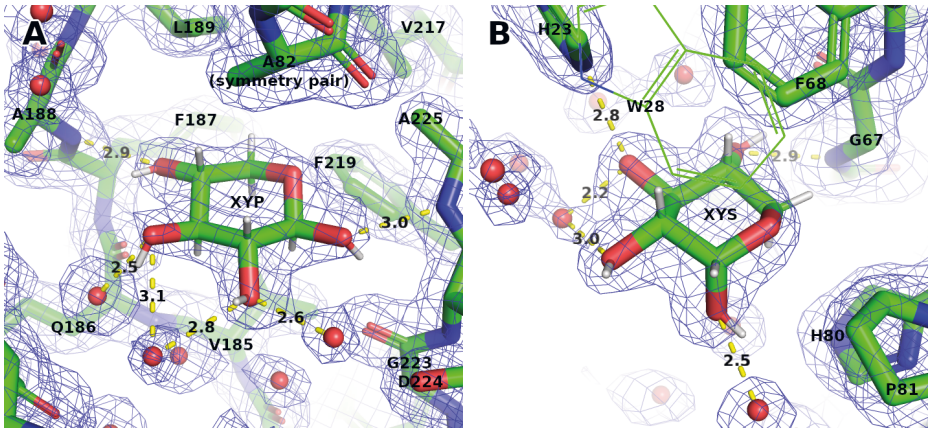



**Figure 4.21:** The D-xylopyranose anomers bound to the surface of the *Cc*XylC in the crystal structure. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. **(a)** β-D-Xylopyranose (XYP) and hydrogen bonds to neighbouring waters and the protein molecule. **(b)** α-D-Xylopyranose (XYS) and hydrogen bonds to neighbouring waters and the protein molecule. For clarity, the W28 in the foreground is shown as a line model and without the electron density.

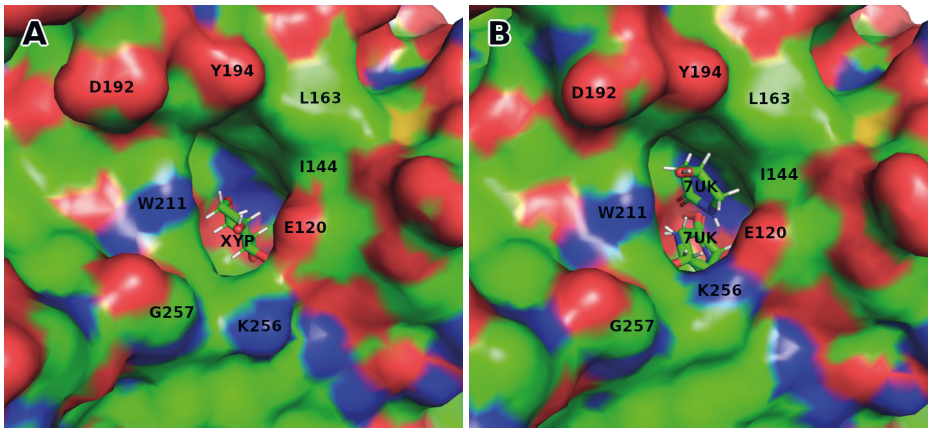



**Figure 4.22:** Surface models of the *Cc*XylC complex structures. **(a)** The β-D-xylopyranose (XYP) bound to the active site, as shown in Figure 4.20a. **(b)** Two 4H2PD molecules (labelled 7UK as in the structure in PDB [94]) bound to the active site and to the mouth of the cavity, as shown in Figure 4.23.

In the 4H2PD structure, five 4H2PD molecules were modelled: one in the active site of one of the two chains, two at the mouths of the central cavities (one per chain) and two on the surface of the enzyme. A surface model of the ones in the cavity is shown in Figure 4.22b. The 4H2PD in the active site was deduced to be coordinated to the $Fe^{2+}$ ion via the carbonyl oxygen even though the electron density of only half of the molecule could be observed (Figure 4.23a). The 4H2PD was modelled in two orientations as separate conformations, and their refined occupancies are 26 % and 34 %. This nonspecific orientation suggests that the five-membered ring

may not be as favourable a substrate as the six-membered ring of β-D-xylopyranose, and thus the six-membered 1,5-lactones would be preferred substrates over the five-membered 1,4-lactones. The electron density in the active site of the other chain was insufficient for modelling more than a water molecule coordinated to the $Fe^{2+}$ ion. The 4H2PD molecules at the mouths of the central cavities are packed against the ring system of W211 and likely stabilised by hydrophobic interactions (Figure 4.23b). These ones as well as the 4H2PD molecules on the surface have strong electron densities and occupancies of 100 % in the model.

This information on substrate binding allowed hypothesising the mechanism of the enzymatic reaction. The assumption is that D-xylonolactone is coordinated to the iron in the same orientation as the D-xylose in the crystal structure, except the carbonyl system will have slightly different geometry. The reaction must be acid- or base-catalysed like the nonenzymatic reaction, and a possible base would be the slightly basic amide group of N101. In the proposed mechanism (Figure 4.24), the coordination to the iron makes the carbonyl carbon more electrophilic, and a water molecule attacks it from the more exposed side near N101. The opposite side is obscured by hydrophobic side chains. The base extracts a proton from the water molecule, leaving a hydroxyl group. Then, the C−O bond of the ester group is opened, leaving a negative charge in the now terminal oxygen. By a proton transfer, the charge is transferred to the carboxyl group, and the carboxyl group extracts the proton from the base, resulting in D-xylonic acid which can again leave the active site. The reverse reaction is possible, but the equilibrium favours the acyclic acid.

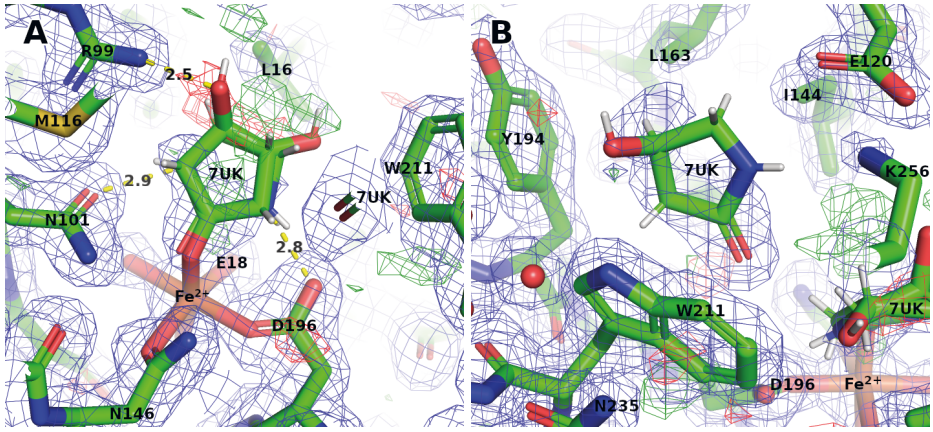



**Figure 4.23:** 4H2PD molecules (labelled 7UK as in the structure in PDB [94]) bound to the *Cc*XylC. Lengths of the hydrogen bonds are in ångströms. The electron density (blue) is the $2F_O - F_C$ map, contoured at $1\sigma$ level. The difference map is the $F_O - F_C$ map, contoured at $+3\sigma$ level (green) and $-3\sigma$ level (red). **(a)** The 4H2PD modelled in the active centre. It is coordinated to the $Fe^{2+}$ ion via the carbonyl oxygen in two orientations that have been modelled as alternative conformations. The electron density is so weak that only the planar system around the carbonyl group can be seen reliably. The orientation which forms a hydrogen bond to R99 has an occupancy of 34 %, and the other one has an occupancy of 26 %. **(b)** The 4H2PD packed against the ring system of W211 at the mouth of the active site. The electron density is strong enough that the 4H2PD could be modelled in a specific orientation and with an occupancy of 100 %.

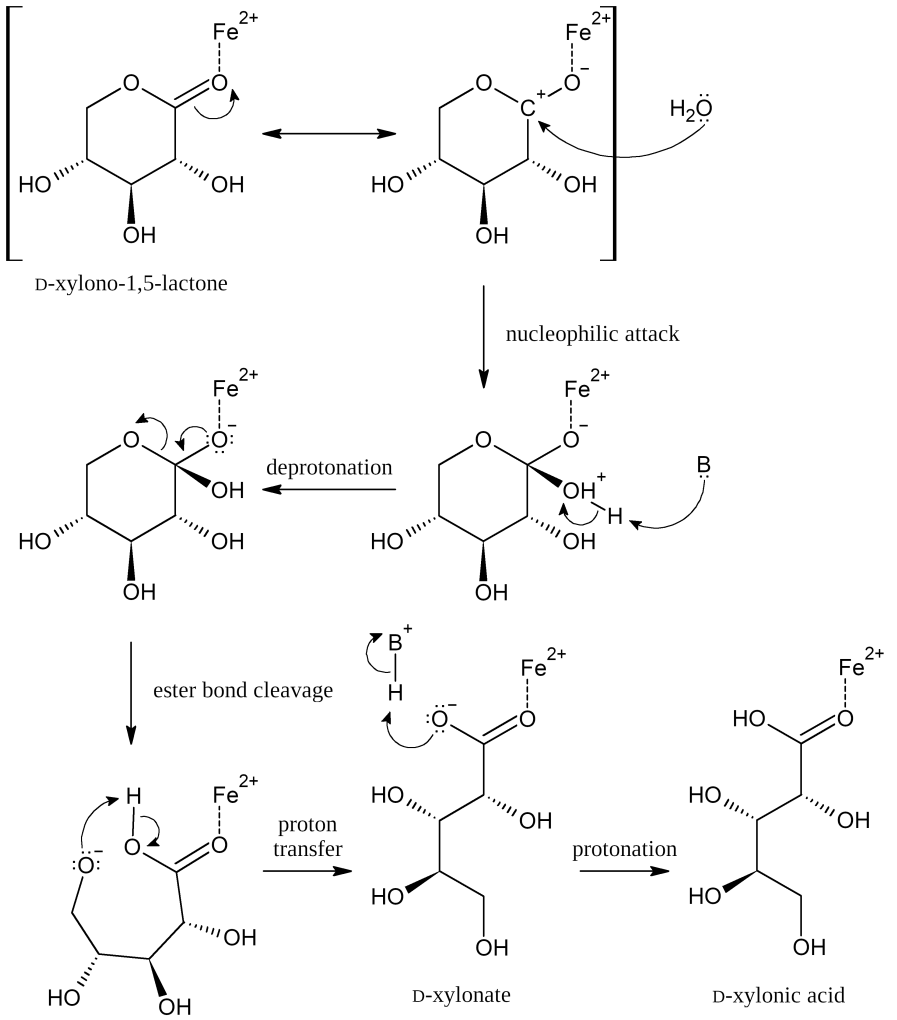



**Figure 4.24:** The proposed mechanism for the enzymatic D-xylono-1,5-lactone hydrolysis. The lactone is coordinated to the iron in the active centre via the carbonyl oxygen, which makes the carbonyl carbon more susceptible to a nucleophilic attack by a water molecule. A base (B), perhaps the amide oxygen of N101, then deprotonates the intermediate, and the ester bond is cleaved. Thereafter, the negative charge is transferred to the carboxyl group, the base returns the proton, and D-xylonic acid is formed. The hydrolysis of D-glucono-1,5-lactone to D-gluconic acid is homologous.

# 5 CONCLUSIONS

The crystal structures of the three *Ec*DERA mutants displayed subtle differences in geometry and chemical environment, which gave some insight to why the mutations affect the enzymatic activity. In the structure of the N21K mutant with DRP, the reaction products of the cleavage of DRP were observed, acetaldehyde bound into the catalytic lysine as a Schiff base and G3H hydrogen bonded to the central cavity. To date, this is the only published structure in PDB where these reaction products have been modelled. Bound acetaldehyde was also present in the structures of the T18Q and C47V/G204A/S239D mutants. These results will hopefully be useful for yet improving the efficiency of the enzyme by further planned mutagenesis and for potentially utilising the *Ec*DERA in industrial synthesis. An effective way of catalysing carbon–carbon bond formation would be invaluable in synthesis of numerous molecules with carbon backbones, for example 3-hydroxybutyric acid, the monomer of polyhydroxybutyrate (PHB), a biodegradable polyester [95].

The *Cc*XylC was observed to bind iron as $Fe^{2+}$ specifically, that is, there is a single metal binding site. The binding affinity was high, with a dissociation constant of 0.50 μM. $Fe^{3+}$ and other metals did not bind in observable amounts, except for $Cu^{2+}$ which was most probably bound nonspecifically on the surface of the protein. The metal binding site was well observed in the crystal structures of the *Cc*XylC, and the presence of iron was shown by measuring a native mass spectrum of the crystals. It would be interesting to test the homologous SMP30 gluconolactonases with native mass spectrometry and to see whether they bind metals similarly. It would be expected since the metal binding site is conserved, however, quite opposite results have been reported. The dissociation constants for $Mg^{2+}$, $Ca^{2+}$, $Mn^{2+}$ and $Zn^{2+}$ that Chakraborti *et al.* [28] have reported were determined by enzymatic assay, an indirect analysis method. Native mass spectrometry, as a direct method, would show the bound metal ions unambiguously.

The complex crystal structures showed that β-D-xylopyranose and 4H2PD bound to the active site. Since the β-D-xylopyranose had strong electron density, 100 % occupancy, specific orientation and several hydrogen bonds, it is a better substrate analogue than the 4H2PD which had weaker electron density, two modelled orientations and fewer hydrogen bonds. Thus, the six-membered ring binds better to the active site than the five-membered ring, and the enzyme probably catalyses the hydrolysis of 1,5-lactone rather than 1,4-lactone. This is in agreement with previous reports of how D-gluconolactone is hydrolysed by acid–base catalysis.

Kinetic analysis of the lactone hydrolysis reaction indicated that presence of $Fe^{2+}$ ions accelerates the nonenzymatic reaction and that the enzymatic reaction is nearly equally fast with both D-xylono- and D-gluconolactone. The determined $k_{cat}/K_M$ ratios were sensible and in about the same order of magnitude as those reported by Chakraborti *et al.* [28] for the SMP30 gluconolactonase in presence of $Mg^{2+}$, $Ca^{2+}$, $Mn^{2+}$ or $Zn^{2+}$, again determined by enzymatic assay. The suggested mechanism of lactone form interconversion and 1,5-lactone hydrolysis explains these findings. However, without detecting the intermediate by mass spectrometry or spectroscopic

methods, the mechanism and the structure of the intermediate remain uncertain. Also, the proposed mechanism of the enzymatic reaction is only speculative, but it could be tested by mutating the N101 and other amino acids at the active site and repeating the experiments with the mutants.

While the results on the *Cc*XylC were in part contradictory to previously published research, they give new valuable information on the enzymatic function. Knowing how the substrates bind to the active site, it is possible to do directed mutagenesis experiments, as has been done with the *Ec*DERA, to potentially increase enzymatic activity or alter the substrate specificity. On the other hand, as the homologous SMP30 gluconolactonases reportedly accept various carbohydrate lactones as substrates, the *Cc*XylC could potentially be used to hydrolyse other carbohydrate lactones as well. This was not investigated in this work, however, with anything else than D-xylono- and D-gluconolactone. These results and possible subsequent new research will undoubtedly help with eventual industrial-scale implementation of the metabolic pathways from D-xylose to various useful substances and production of new biomass-based reagents, materials and possibly fuels.

# BIBLIOGRAPHY


[1] H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour, *Principles of Biochemistry, Carbohydrates*, 2nd ed., International edition (Prentice-Hall, New Jersey, 1996), pp. 215–242.

[2] H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour, *Principles of Biochemistry, Nucleotides*, 2nd ed., International edition (Prentice-Hall, New Jersey, 1996), pp. 243–260.

[3] H. R. Horton, L. A. Moran, R. S. Ochs, J. D. Rawn, and K. G. Scrimgeour, *Principles of Biochemistry, Photosynthesis*, 2nd ed., International edition (Prentice-Hall, New Jersey, 1996), pp. 435–458.

[4] F. H. Isikgor and C. R. Becer, "Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers," *Polym. Chem.* **6**, 4497–4559 (2015).

[5] A. Zoghlami and G. Paës, "Lignocellulosic biomass: understanding recalcitrance and predicting hydrolysis," *Front. Chem.* **7**, 874 (2019).

[6] C. Schädel, A. Blöchl, A. Richter, and G. Hoch, "Quantification and monosaccharide composition of hemicelluloses from different plant functional types," *Plant Physiol. Biochem.* **48**, 1–8 (2010).

[7] N. Xu, W. Zhang, S. Ren, F. Liu, C. Zhao, H. Liao, Z. Xu, J. Huang, Q. Li, Y. Tu, B. Yu, Y. Wang, J. Jiang, J. Qin, and L. Peng, "Hemicelluloses negatively affect lignocellulose crystallinity for high biomass digestibility under NaOH and $H_2SO_4$ pretreatments in *Miscanthus*," *Biotechnol. Biofuels* **5**, 58 (2012).

[8] Y. Kondo, Y. Inai, Y. Sato, S. Handa, S. Kubo, K. Shimokado, S. Goto, M. Nishikimi, N. Maruyama, and A. Ishigami, "Senescence marker protein 30 functions as gluconolactonase in L-ascorbic acid biosynthesis, and its knockout mice are prone to scurvy," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5723–5728 (2006).

[9] M. Nishikimi, T. Kawai, and K. Yagi, "Guinea pigs possess a highly mutated gene for L-gulono-$\gamma$-lactone oxidase, the key enzyme for L-ascorbic acid biosynthesis missing in this species," *J. Biol. Chem.* **267**, 21967–21972 (1992).

[10] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, "The carbohydrate-active enzymes database (CAZy) in 2013," *Nucleic Acids Res.* **42**, D490–D495 (2014).

[11] A. Heine, G. DeSantis, J. G. Luz, M. Mitchell, C.-H. Wong, and I. A. Wilson, "Observation of covalent intermediates in an enzyme mechanism at atomic resolution," *Science* **294**, 369–374 (2001).

[12] A. M. Poole, N. Horinouchi, R. J. Catchpole, D. Si, M. Hibi, K. Tanaka, and J. Ogawa, "The case for an early biological origin of DNA," *J. Mol. Evol.* **79**, 204–212 (2014).

[13] S. Jennewein, M. Schürmann, M. Wolberg, I. Hilker, R. Luiten, M. Wubbolts, and D. Mink, "Directed evolution of an industrial biocatalyst: 2-deoxy-D-ribose 5-phosphate aldolase," *Biotechnol. J.* **1**, 537–548 (2006).

[14] H. J. M. Gijsen and C.-H. Wong, "Unprecedented asymmetric aldol reactions with three aldehyde substrates catalyzed by 2-deoxyribose-5-phosphate aldolase," *J. Am. Chem. Soc.* **116**, 8422–8423 (1994).

[15] A. Heine, G. DeSantis, J. G. Luz, M. Mitchell, C.-H. Wong, and I. A. Wilson, *Observation of covalent intermediates in an enzyme mechanism at atomic resolution*, doi: 10.2210/pdb1jcj/pdb, 2001.

[16] A. Heine, G. DeSantis, J. G. Luz, M. Mitchell, C.-H. Wong, and I. A. Wilson, *Observation of covalent intermediates in an enzyme mechanism at atomic resolution*, doi: 10.2210/pdb1jcl/pdb, 2001.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res.* **28**, 235–242 (2000).

[18] *Protein Data Bank (PDB)*, `https://www.rcsb.org`.

[19] A. Heine, J. G. Luz, C.-H. Wong, and I. A. Wilson, "Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99 Å resolution," *J. Mol. Biol.* **343**, 1019–1034 (2004).

[20] M. Jones Jr and S. A. Fleming, *Organic chemistry, Addition reactions of nitrogen bases: imine and enamine formation*, 4th ed., International student edition (W. W. Norton, New York, 2010), pp. 790–797.

[21] Y. Cao, M. Xian, H. Zou, and H. Zhang, "Metabolic engineering of *Escherichia coli* for the production of xylonate," *PLOS ONE* **8**, e67305 (2013).

[22] Y. Cao, W. Niu, J. Guo, M. Xian, and H. Liu, "Biotechnological production of 1,2,4-butanetriol: an efficient process to synthesize energetic material precursor from renewable biomass," *Sci. Rep.* **5**, 18149 (2016).

[23] T. Bamba, T. Yukawa, G. Guirimand, K. Inokuma, K. Sasaki, T. Hasunuma, and A. Kondo, "Production of 1,2,4-butanetriol from xylose by *Saccharomyces cerevisiae* through Fe metabolic engineering," *Metab. Eng.* **56**, 17–27 (2019).

[24] R. Weimberg, "Pentose oxidation by *Pseudomonas fragi*," *J. Biol. Chem.* **236**, 629–635 (1961).

[25] A. S. Dahms, "3-Deoxy-D-pentulosonic acid aldolase and its role in a new pathway of D-xylose degradation," *Biochem. Biophys. Res. Commun.* **60**, 1433–1439 (1974).

[26] N. Wu, M. Yang, U. Gaur, H. Xu, Y. Yao, and D. Li, "Alpha-ketoglutarate: physiological functions and applications," *Biomol. Ther. (Seoul)* **24**, 1–8 (2016).

[27] H. Boer, M. Andberg, R. Pylkkänen, H. Maaheimo, and A. Koivula, "In vitro reconstitution and characterisation of the oxidative D-xylose pathway for production of organic acids and alcohols," *AMB Expr.* **9**, 48 (2019).

[28] S. Chakraborti and B. J. Bahnson, "Crystal structure of human senescence marker protein 30: insights linking structural, enzymatic, and physiological functions," *Biochemistry* **49**, 3436–3444 (2010).

[29] C.-N. Chen, K.-H. Chin, A. H.-J. Wang, and S.-H. Chou, "The first crystal structure of gluconolactonase important in the glucose secondary metabolic pathways," *J. Mol. Biol.* **384**, 604–614 (2008).

[30] C.-N. Chen, K.-H. Chin, A. H.-J. Wang, and S.-H. Chou, *Structural and functional analyses of XC5397 from Xanthomonas campestris: a gluconolactonase important in glucose secondary metabolic pathways*, doi: 10.2210/pdb3dr2/pdb, 2008.

[31] S. Chakraborti and B. J. Bahnson, *Crystal structure of human senescence marker protein-30(SMP30)(Calcium bound)*, doi: 10.2210/pdb3g4e/pdb, 2010.

[32] S. Chakraborti and B. J. Bahnson, *Crystal structure of human senescence marker protein-30 (Zinc bound)*, doi: 10.2210/pdb3g4h/pdb, 2010.

[33] S. Aizawa, M. Senda, A. Harada, N. Maruyama, T. Ishida, T. Aigaki, A. Ishigami, and T. Senda, "Structural basis of the γ-lactone-ring formation in ascorbic acid biosynthesis by the senescence marker protein-30/gluconolactonase," *PLOS ONE* **8**, e53706 (2013).

[34] M. Hummel, M. Leppikallio, S. Heikkinen, K. Niemelä, and H. Sixta, "Acidity and lactonization of xylonic acid: a nuclear magnetic resonance study," *J. Carbohydr. Chem.* **29**, 416–428 (2010).

[35] K. Shimahara and T. Takahashi, "An infrared spectrophotometric study on the interconversion and hydrolysis of D-glucono-γ- and -δ-lactone in deuterium oxide," *Biochim. Biophys. Acta* **201**, 410–415 (1970).

[36] S. Aizawa, M. Senda, A. Harada, N. Maruyama, T. Ishida, T. Aigaki, A. Ishigami, and T. Senda, *Mouse SMP30/GNL*, doi: 10.2210/pdb4gn7/pdb, 2013.

[37] M. M. Rahman, "Structure and function of iron-sulfur cluster containing pentonate dehydratases," PhD thesis (University of Eastern Finland, 2017), p. 10.

[38] M. Andberg, N. Aro-Kärkkäinen, P. Carlson, M. Oja, S. Bozonnet, M. Toivari, N. Hakulinen, M. O'Donohue, M. Penttilä, and A. Koivula, "Characterization and mutagenesis of two novel iron–sulphur cluster pentonate dehydratases," *Appl. Microbiol. Biotechnol.* **100**, 7549–7563 (2016).

[39] C. P. Woodbury, *Introduction to Macromolecular Binding Equilibria, Notation for binding constants*, 1st ed. (CRC Press, Boca Raton, Florida, 2007), p. 51.

[40] K. Sakurai and Y. Goto, "Manipulating monomer-dimer equilibrium of bovine β-lactoglobulin by amino acid substitution," *J. Biol. Chem.* **277**, 25735–25740 (2002).

[41] N.-E. L. Saris, E. Mervaala, H. Karppanen, J. A. Khawaja, and A. Lewenstam, "Magnesium: an update on physiological, clinical and analytical aspects," *Clin. Chim. Acta* **294**, 1–26 (2000).

[42] T. Fujita, *Calcium homeostasis and signaling in aging, Calcium homeostasis and signaling in aging*, edited by M. P. Mattson, 1st ed. (Elsevier, Amsterdam, The Netherlands, 2002), pp. 13–26.

[43] M. Yamaguchi, *Calcium homeostasis and signaling in aging, Impact of aging on calcium channels and pumps*, edited by M. P. Mattson, 1st ed. (Elsevier, Amsterdam, The Netherlands, 2002), pp. 47–65.

[44] B. E. Eaton, L. Gold, and D. A. Zichi, "Let's get specific: the relationship between specificity and affinity," *Chem. Biol.* **2**, 633–638 (1995).

[45] A. Binolfi, G. R. Lamberto, R. Duran, L. Quintanar, C. W. Bertoncini, J. M. Souza, C. Cerveñansky, M. Zweckstetter, C. Griesinger, and C. O. Fernández, "Site-specific interactions of Cu(II) with α and β-synuclein: bridging the molecular gap between metal binding and aggregation," *J. Am. Chem. Soc.* **130**, 11801–11812 (2008).

[46] T. Lech and J. K. Sadlik, "Copper concentration in body tissues and fluids in normal subjects of Southern Poland," *Biol. Trace Elem. Res.* **118**, 10–15 (2007).

[47] S. Chakraborty, V. Balakotaiah, and A. Bidani, "Diffusing capacity reexamined: relative roles of diffusion and chemical reaction in red cell uptake of $O_2$, CO, $CO_2$, and NO," *J. Appl. Physiol.* **97**, 2284–2302 (2004).

[48] J. S. Olson, E. W. Foley, D. H. Maillett, and E. V. Paster, *Hemoglobin Disorders: Molecular Methods and Protocols, Measurement of rate constants for reactions of $O_2$, CO, and NO with hemoglobin*, edited by R. L. Nagel, 1st ed., vol. 82 (Humana Press, Totowa, New Jersey, 2003), pp. 65–81.

[49] K. Groebe and G. Thews, "Effects of red cell spacing and red cell movement upon oxygen release under conditions of maximally working skeletal muscle," *Adv. Exp. Med. Biol.* **248**, 175–185 (1989).

[50] J. Pääkkönen and J. Rouvinen, *Protein thermodynamics simulations, GitHub repository*, `https://github.com/protsim/protsim`, 2021.

[51] J. Pääkkönen and J. Rouvinen, *Protein thermodynamics simulations, Index page*, `https://protsim.github.io/protsim`, 2021.

[52] Free Software Foundation, Inc., *GNU General Public License, version 2*, `https://www.gnu.org/licenses/old-licenses/gpl-2.0.html`, 1991.

[53] J. T. Watson and O. D. Sparkman, *Introduction to mass spectrometry: Instrumentation, applications and strategies for data interpretation, Introduction*, 4th ed. (John Wiley & Sons, West Sussex, England, 2007), pp. 3–9.

[54] J. T. Watson and O. D. Sparkman, *Introduction to mass spectrometry: Instrumentation, applications and strategies for data interpretation, Types of $m/z$ analyzers*, 4th ed. (John Wiley & Sons, West Sussex, England, 2007), p. 61.

[55] J. T. Watson and O. D. Sparkman, *Introduction to mass spectrometry: Instrumentation, applications and strategies for data interpretation, Electrospray ionization*, 4th ed. (John Wiley & Sons, West Sussex, England, 2007), pp. 485–518.

[56] J. T. Watson and O. D. Sparkman, *Introduction to mass spectrometry: Instrumentation, applications and strategies for data interpretation, Quadrupole ion traps*, 4th ed. (John Wiley & Sons, West Sussex, England, 2007), pp. 82–103.

[57] J. T. Watson and O. D. Sparkman, *Introduction to mass spectrometry: Instrumentation, applications and strategies for data interpretation, FTICR-MS*, 4th ed. (John Wiley & Sons, West Sussex, England, 2007), pp. 122–128.

[58] J. W. Eaton, D. Bateman, S. Hauberg, and R. Wehbring, *GNU Octave version 5.2.0 manual: a high-level interactive language for numerical computations*, `https://www.gnu.org/software/octave/doc/v5.2.0/` (2020).

[59] A. McPherson and J. A. Gavira, "Introduction to protein crystallization," *Acta Cryst.* **F70**, 2–20 (2014).

[60] M. Mariam, "Crystallization of galactarolactone cycloisomerase (Atu3139) and 2-deoxyribose-5-phosphate aldolase (DERA)," Master's thesis (University of Eastern Finland, 2016), pp. 32–36.

[61] G. Rhodes, *Crystallography made crystal clear: A guide for users of macromolecular models, An overview of protein crystallography*, 1st ed. (Academic Press, San Diego, California, 1993), pp. 5–27.

[62] P. R. Evans and G. N. Murshudov, "How good are my data and what is the resolution?" *Acta Cryst.* **D69**, 1204–1214 (2013).

[63] G. Rhodes, *Crystallography made crystal clear: A guide for users of macromolecular models, Molecular replacement: related proteins as phasing models*, 1st ed. (Academic Press, San Diego, California, 1993), pp. 125–129.

[64] P. V. Afonine, R. W. Grosse-Kunstleve, N. Echols, J. J. Headd, N. W. Moriarty, M. Mustyakimov, T. C. Terwilliger, A. Urzhumtsev, P. H. Zwart, and P. D. Adams, "Towards automated crystallographic structure refinement with *phenix.refine*," *Acta Cryst.* **D68**, 352–367 (2012).

[65] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures," *FEBS J.* **275**, 1–21 (2008).

[66] C. Ramakrishnan and G. N. Ramachandran, "Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units," *Biophys. J.* **5**, 909–933 (1965).

[67] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. W. Leslie, A. McCoy, S. J. McNicholas, G. N. Murshudov, N. S. Pannu, E. A. Potterton, H. R. Powell, R. J. Read, A. Vagin, and K. S. Wilson, "Overview of the *CCP*4 suite and current developments," *Acta Cryst.* **D67**, 235–242 (2011).

[68] W. Kabsch, "*XDS*," *Acta Cryst.* **D66**, 125–132 (2010).

[69] G. Winter, "*xia2*: an expert system for macromolecular crystallography data reduction," *J. Appl. Cryst.* **43**, 186–190 (2010).

[70] G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, and G. Evans, "*DIALS*: implementation and evaluation of a new integration package," *Acta Cryst.* **D74**, 85–97 (2018).

[71] P. Evans, "Scaling and assessment of data quality," *Acta Cryst.* **D62**, 72–82 (2006).

[72] D. Liebschner, P. V. Afonine, M. L. Baker, G. Bunkóczi, V. B. Chen, T. I. Croll, B. Hintze, L.-W. Hung, S. Jain, A. J. McCoy, N. W. Moriarty, R. D. Oeffner, B. K. Poon, M. G. Prisant, R. J. Read, J. S. Richardson, D. C. Richardson, M. D. Sammito, O. V. Sobolev, D. H. Stockwell, T. C. Terwilliger, A. G. Urzhumtsev, L. L. Videau, C. J. Williams, and P. D. Adams, "Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*," *Acta Cryst.* **D75**, 861–877 (2019).

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

[74] R. Zhang, A. Joachimiak, A. Edwards, T. Skarina, E. Evdokimova, and A. Savchenko, *Structural Genomics, Protein EC1535*, doi: 10.2210/pdb1ktn/pdb, 2002.

[75] C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma, D. A. Keedy, B. J. Hintze, V. B. Chen, S. Jain, S. M. Lewis, W. B. Arendall III, J. Snoeyink, P. D. Adams, S. C. Lovell, J. S. Richardson, and D. C. Richardson, "MolProbity: more and better reference data for improved all-atom structure validation," *Protein Sci.* **27**, 293–315 (2018).

[76] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of *Coot*," *Acta Cryst.* **D66**, 486–501 (2010).

[77] *Open-Source PyMOL version 1.8.6*, Schrödinger, LLC, `https://github.com/schrodinger/pymol-open-source`, 2017.

[78] N. W. Moriarty, R. W. Grosse-Kunstleve, and P. D. Adams, "*electronic Ligand Builder and Optimization Workbench* (*eLBOW*): a tool for ligand coordinate and restraint generation," *Acta Cryst.* **D65**, 1074–1080 (2009).

[79] C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, and G. Bricogne, "Data processing and analysis with the *autoPROC* toolbox," *Acta Cryst.* **D67**, 293–302 (2011).

[80] I. J. Tickle, C. Flensburg, P. Keller, W. Paciorek, A. Sharff, C. Vonrhein, and G. Bricogne, *STARANISO*, Global Phasing Ltd., Cambridge, United Kingdom, 2018.

[81] F. DiMaio, T. C. Terwilliger, R. J. Read, A. Wlodawer, G. Oberdorfer, U. Wagner, E. Valkov, A. Alon, D. Fass, H. L. Axelrod, D. Das, S. M. Vorobiev, H. Iwaï, P. R. Pokkuluri, and D. Baker, "Improved molecular replacement by density- and energy-guided protein structure optimization," *Nature* **473**, 540–543 (2011).

[82] C. H. Borchers, V. E. Marquez, G. K. Schroeder, S. A. Short, M. J. Snider, J. P. Speir, and R. Wolfenden, "Fourier transform ion cyclotron resonance MS reveals the presence of a water molecule in an enzyme transition-state analogue complex," *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15341–15345 (2004).

[83] *Origin Pro*, OriginLab Corporation, Northampton, Massachusetts, 2018.

[84] Z. Zhang, P. Gibson, S. B. Clark, G. Tian, P. L. Zanonato, and L. Rao, "Lactonization and protonation of gluconic acid: a thermodynamic and kinetic study by potentiometry, NMR and ESI-MS," *J. Solution Chem.* **36**, 1187–1200 (2007).

[85] A. G. Marangoni, *Enzyme kinetics: A modern approach, Characterization of enzyme activity*, 1st ed. (John Wiley & Sons, Hoboken, New Jersey, 2003), pp. 44–60.

[86] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Escherichia coli D-2-deoxyribose-5-phosphate aldolase – N21K mutant*, doi: 10.2210/pdb6z9j/pdb, 2020.

[87] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Escherichia coli D-2-deoxyribose-5-phosphate aldolase – N21K mutant complex with reaction products*, doi: 10.2210/pdb6z9i/pdb, 2020.

[88] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Escherichia coli D-2-deoxyribose-5-phosphate aldolase – C47V/G204A/S239D mutant*, doi: 10.2210/pdb6z9h/pdb, 2020.

[89] M. Dick, R. Hartmann, O. H. Weiergräber, C. Bisterfeld, T. Classen, M. Schwarten, P. Neudecker, D. Willbold, and J. Pietruszka, "Mechanism-based inhibition of an aldolase at high concentrations of its natural substrate acetaldehyde: structural insights and protective strategies," *Chem. Sci.* **7**, 4492–4502 (2016).

[90] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "CASTp 3.0: computed atlas of surface topography of proteins," *Nucleic Acids Res.* **46**, W363–W367 (2018).

[91] A. Heine, J. G. Luz, C. H. Wong, and I. A. Wilson, *Comparison of class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase determined at 0.99 Angstrom resolution*, doi: 10.2210/pdb1p1x/pdb, 2004.

[92] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Caulobacter crescentus xylonolactonase with D-xylose*, doi: 10.2210/pdb7plb/pdb, 2021.

[93] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Caulobacter crescentus xylonolactonase with D-xylose, P21 space group*, doi: 10.2210/pdb7plc/pdb, 2021.

[94] J. Pääkkönen, N. Hakulinen, and J. Rouvinen, *Caulobacter crescentus xylonolactonase with (R)-4-hydroxy-2-pyrrolidone*, doi: 10.2210/pdb7pld/pdb, 2021.

[95] R. Andler, V. Pino, F. Moya, E. Soto, C. Valdés, and C. Andreeßen, "Synthesis of poly-3-hydroxybutyrate (PHB) by *Bacillus cereus* using grape residues as sole carbon source," *Int. J. Biobased Plast.* **3**, 98–111 (2021).

# Publication **I**

S. Voutilainen, M. Heinonen, M. Andberg,
E. Jokinen, H. Maaheimo, J. Pääkkönen,
N. Hakulinen, J. Rouvinen, H. Lähdesmäki,
S. Kaski, J. Rousu, M. Penttilä, and A. Koivula

"Substrate specificity of 2-deoxy-D-ribose 5-phosphate aldolase (DERA) assessed by different protein engineering and machine learning methods"

*Applied Microbiology and Biotechnology* **104**,

pp. 10515–10529, 2020.

Applied Microbiology and Biotechnology (2020) 104:10515–10529
https://doi.org/10.1007/s00253-020-10960-x

BIOTECHNOLOGICALLY RELEVANT ENZYMES AND PROTEINS

# Substrate specificity of 2-deoxy-D-ribose 5-phosphate aldolase (DERA) assessed by different protein engineering and machine learning methods

Sanni Voutilainen[1] · Markus Heinonen[2,3] · Martina Andberg[1] · Emmi Jokinen[2] · Hannu Maaheimo[1] · Johan Pääkkönen[4] · Nina Hakulinen[4] · Juha Rouvinen[4] · Harri Lähdesmäki[2] · Samuel Kaski[2,3] · Juho Rousu[2,3] · Merja Penttilä[1] · Anu Koivula[1]

Received: 22 June 2020 / Revised: 1 October 2020 / Accepted: 12 October 2020 / Published online: 4 November 2020


## Abstract

In this work, deoxyribose-5-phosphate aldolase (*Ec* DERA, EC 4.1.2.4) from *Escherichia coli* was chosen as the protein engineering target for improving the substrate preference towards smaller, non-phosphorylated aldehyde donor substrates, in particular towards acetaldehyde. The initial broad set of mutations was directed to 24 amino acid positions in the active site or in the close vicinity, based on the 3D complex structure of the *E. coli* DERA wild-type aldolase. The specific activity of the DERA variants containing one to three amino acid mutations was characterised using three different substrates. A novel machine learning (ML) model utilising Gaussian processes and feature learning was applied for the 3rd mutagenesis round to predict new beneficial mutant combinations. This led to the most clear-cut (two- to threefold) improvement in acetaldehyde (C2) addition capability with the concomitant abolishment of the activity towards the natural donor molecule glyceraldehyde-3-phosphate (C3P) as well as the non-phosphorylated equivalent (C3). The *Ec* DERA variants were also tested on aldol reaction utilising formaldehyde (C1) as the donor. *Ec* DERA wild-type was shown to be able to carry out this reaction, and furthermore, some of the improved variants on acetaldehyde addition reaction turned out to have also improved activity on formaldehyde.

## Key points

- *DERA aldolases are promiscuous enzymes.*
- *Synthetic utility of DERA aldolase was improved by protein engineering approaches.*
- *Machine learning methods aid the protein engineering of DERA.*

**Keywords** DERA · Aldolase · Protein engineering · Machine learning · Crystal structure determination · C–C bond formation · Biocatalysis

## Introduction

Aldolases are enablers of industrial biocatalysis as they can promote carbon-carbon (C–C) bond formation, which is one of the essential reactions in synthetic chemistry. Aldol reaction can be catalysed by the lyase (EC4) or transferase class (EC2) of enzymes, found in the metabolic pathways in all three domains of life (archaea, bacteria, eukarya). Aldolases catalyse the reversible formation of C–C bonds by the aldol addition of a nucleophilic donor, typically a ketone enolate, onto an electrophilic aldehyde acceptor (Scheme 1).

Aldolase type of enzymes have been found to be promiscuous capable of using a broad range of aldehydes as acceptors, whereas the donor compound is often structurally

**Supplementary Information** The online version of this article (https://doi.org/10.1007/s00253-020-10960-x) contains supplementary material, which is available to authorized users.

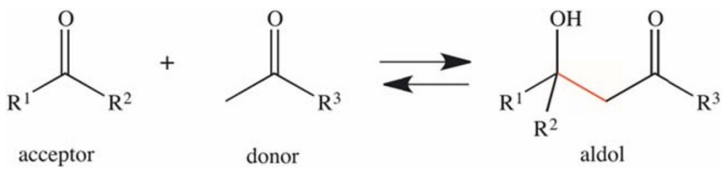
✉ Sanni Voutilainen
sanni.voutilainen@vtt.fi

[1] VTT Technical Research Centre of Finland Ltd, P.O. Box 1000, FI-02044 VTT, Espoo, Finland

[2] Department of Computer Science, Aalto University, Espoo, Finland

[3] Helsinki Institute for Information Technology, Espoo, Finland

[4] Department of Chemistry, University of Eastern Finland, PO Box 111, FI-80101 Joensuu, Finland




**Scheme 1** Aldolases catalyse the reversible formation of C–C bonds by the aldol addition of a nucleophilic donor, typically a ketone enolate, onto anelectrophilic aldehyde acceptor

invariable. Hence, aldolases can be classified according to their donor specificity to, e.g. acetaldehyde-dependent aldolases. Another way to classify these enzymes relates to the different catalytic mechanisms to activate the nucleophilic component. The class I aldolases do not require any cofactor, but exhibit a conserved lysine residue in the active site which forms a Schiff base intermediate with the donor compound to generate an enamine nucleophile.

Deoxyribose-5-phosphate aldolases (DERA, EC 4.1.2.4) are a class I aldolase that catalyses in vivo the reversible addition of the donor molecule, acetaldehyde (C2), to the acceptor molecule, glyceraldehyde 3-phosphate (C3P). DERA is known to be promiscuous in its substrate range as it accepts a wide variety of different acceptor molecules (Chen et al. 1992; Gijsen and Wong 1994). In addition, at the donor site of DERA enzyme acetone, fluoroacetone and propionaldehyde (propanal) have been reported to function (with strongly reduced reaction rates) besides the acetaldehyde (Barbas et al. 1990; Chen et al. 1992). An interesting and unique feature, particularly in terms of synthesis reactions, is the ability of DERA to catalyse sequential acetaldehyde addition (Gijsen and Wong 1994). Here, the first aldol addition reaction creates an aldehyde product, which functions as an acceptor for the subsequent DERA-catalysed stereoselective aldol reaction to add another aldehyde donor substrate. This cascade reaction has been utilised for addition of two equivalents of acetaldehyde to one equivalent of chloroacetaldehyde in the preparation of (3R,5S)-6-chloro-2,4,6-trideoxyhexose, a chiral precursor for the sidechain of the statin drugs (Oslaj et al. 2013). The capacity of wild-type DERA to catalyse aldol addition is, however, rather low. To increase the affinity of the *E. coli* aldolase for chloroacetaldehyde acceptor and stability against high acetaldehyde concentrations, researchers at DSM applied a directed evolution approach (Jennewein et al. 2006). By screening for and combining beneficial mutations, they succeeded in identifying an improved variant with 10-fold improved productivity in *E. coli* under industrially relevant conditions.

Advances in computational chemistry combined with protein engineering strategies have most recently opened new possibilities for more efficient design of enzymatic catalysts for chemical reactions (Linder 2012; Kiss et al. 2013; Mak and Siegel 2014; Jindal et al. 2019). The aim of this work was to engineer DERA aldolase to accept smaller non-phosphorylated acceptor substrates in the aldol addition reaction with acetaldehyde. The work included evaluating first the most suitable DERA aldolase for the protein engineering work, by expressing, purifying and characterising a set of DERA enzymes of different origin. After that, we used different types of mutagenesis approaches in combination with a novel machine learning model to create DERA variants in three mutagenesis rounds, which were screened using a panel of different substrates. Crystal structures of some of the most interesting DERA variants were also solved to provide more insight.

## Materials and methods

### Cloning, protein expression and purification of different DERAs

Seven DERA encoding genes from different organisms (1. DERA [UniProt P0A6L0]-coding gene [deoC] from *E. coli*, 2. DERA [NCBI Reference Sequence: WP_047758083.1] -coding gene from *Geobacillus*, 3. DERA [SwissProt Q5SJ28] -coding gene from *Thermus thermophilus*, 4. DERA [UniProt E0CX06] -coding gene from *Coccidioides immitis*, 5. DERA [UniProt Q03Q50]-coding gene from *Lactobacillus brevis*, 6. DERA [GenBank CRG82919.1]-coding gene from *Talaromyces islandicus* and 7. DERA [NCBI Reference Sequence: XP_003188826.1]-coding gene from *Aspergillus niger*) were cloned and expressed in *E. coli.* All of them were codon optimised for *E. coli* and synthesised with an N-terminal 6x His-tag by Integrated DNA Technologies as so called G-blocks. Histidine tag was added to the N-terminus as it has been shown that the C-terminal tail of *Ec*DERA has a role in the catalytic activity (Schulte et al. 2018). For the codon-optimised nucleotide sequences, see "GenBank accession numbers" section. The synthetic DNA blocks were cloned into the pBAT4 vector (Peränen et al. 1996) linearised with *Nco*I and *Xho*I restriction enzymes. The synthesised insert contained 60 bp overlapping regions in both 5′ and 3′ ends to the vector to allow cloning with Gibson assembly method (Gibson et al. 2009) using Gibson assembly® master mix (New England Biolabs). After assembly, the mixtures were transformed into chemical competent XL1-blue *E. coli* cells.

Single point mutations of *E. coli* DERA (*Ec* DERA) were made with Q5®site-directed mutagenesis kit (New England Biolabs) (list of primers is shown in Table S1) and verified by sequencing (Source BioScience or Microsynth). Some amino acid positions were mutated by using degenerated primers to

generate a selection of amino acid mutations in one PCR reaction. For example, position L20 was mutagenised using forward primer 5′-GTTGATGGACSDNACCACTC TGAACG-3′ to generate mutations L20R, Q, E, H, V, D, G simultaneously. Nucleotide code S stands for G or C and D stands for A, G or T.

Double mutants (containing two point mutations/gene) and some of the triple mutants (containing three point mutations/gene) of *Ec* DERA were generated in a similar manner as for the single mutants by using already existing single mutant in question as the template in the PCR mutagenesis. The triple mutant variants suggested by machine learning were ordered as synthetic DNA blocks from Integrated DNA Technologies, similarly to the wild-type DERA genes described above. Saturation mutagenesis was done by PCR using so called 22c-trick (Kille et al. 2013). The method reduces codon redundancy from often-used saturation mutagenesis method using codon NNK.

DERA variants were expressed in *E. coli* BL21(DE3) strain in LB-medium containing 100 μg/ml ampicillin. DERA wild-type enzymes from different organisms and *Ec* DERA mutants N21K and triple mutant C47V/G204A/S239D, which were crystallised for 3D structure determination, were cultivated in 50-ml scale and all other *Ec* DERA mutants were cultivated in 3-ml or 10-ml scale. The expression strains were cultivated for 6–8 h in 37 °C after which the expression was induced with 0.5 mM IPTG, and after 16-h incubation at 30 °C, cells were harvested (10 min at 4000×*g*). For cell lysis, the cells were re-suspended in B-PER Bacterial Protein Extraction Reagent (Thermo Scientific) supplemented with protease inhibitor (cOmplete mini, EDTA-free, Roche), lysozyme (Sigma Aldrich), and DNAse (Roche). After incubation (1 h, RT) and centrifugation (10 min at 4000×g), the supernatant (crude cell extract) was loaded to a column for purification. The samples from the 50-ml cultivations were loaded on to a HisTrap FF Crude 1-ml column (GE Healthcare) equilibrated with 20 mM sodium phosphate, 0.5 M NaCl, 10 mM imidazole, pH 7.4. The column was washed with the equilibration buffer, and bound protein was eluted with a linear gradient from 10 to 500 mM imidazole. DERA containing fractions were pooled and the buffer was changed to 50 mM Tris-HCl pH 7.5 by EconoPac (BioRad) desalting columns. The protein purity was verified with SDS-PAGE.

The *Ec* DERA mutants were purified from the small-scale cultivations (3 ml and 10 ml) with 0.2-ml HisPur™ Ni-NTA Spin Columns (Thermo Scientific) according to manufacturer's instructions and the buffer was exchanged with PD-10 (GE Healthcare) desalting columns. The protein concentrations were determined by measuring the absorbance at 280 nm and calculated using the theoretical epsilon based on the amino acid sequence (monomer).

## Cloning, protein expression and purification of Klebsiella pneumoniae 1,3-propanediol oxidoreductase (Kp PDOR)

*Klebsiella pneumoniae* 1,3-propanediol oxidoreductase (*Kp* PDOR; UniProt Q59477) encoding gene *dhaT* with a N-terminal 6× His-tag was codon optimised for *E. coli* and synthesised by Integrated DNA Technologies and cloned into the pBAT4 vector in a similar manner as described above for DERA. For the codon-optimised nucleotide sequence, see "GenBank accession numbers" section. Expression of *Kp* PDOR was done in *E. coli* BL21(DE3) by cultivating the *Kp* PDOR expression vector containing strain in 50-ml volume in 250-ml shake flasks in LB-medium (100 μg ampicillin/ml) similarly as for DERA described above. Purification of *Kp* PDOR was also done in the same way as for DERA. The purified fractions of *Kp* PDOR were pooled and the buffer was exchanged with EconoPac desalting columns to 50 mM Tris-HCl pH 7.5, 2 mM DTT, 1 mM $MnCl_2$.

## Following DRP and DR cleavage reactions by DERA enzymes

The cleavage of the natural DERA substrate, deoxyribose 5-phosphate (DRP), was measured using 2-deoxyribose 5-phosphate sodium salt (Sigma-Aldrich) as a substrate in a coupled enzyme system with triosephosphate isomerase (TPI) and glycerol 3-phosphate dehydrogenase (GPD) from rabbit muscle (Sigma-Aldrich) in ambient temperature. DERA activity on DRP liberates glyceraldehyde-3-phosphate, which is reduced to glycerol-3-phosphate by the supplementary enzymes TPI and GDH. The latter reaction consumes NADH, which can be detected by spectrophotometer. The reaction mixture contained 0.1 μM purified DERA wild-type or variant, 5 mM DRP, 3 units of TPI, 2 units of GPD and 0.3 mM NADH in 50 mM Tris-HCl, pH 7.5, supplemented with 5 mM $MgCl_2$. The reaction was initiated by addition of DRP and followed by measuring the decrease of absorbance at 340 nm using a Varioskan microtiter plate reader (Thermo). DERA activity on non-phosphorylated substrate 2-deoxy-D-ribose (DR, Sigma-Aldrich) was assayed similarly in a coupled enzyme system with 4 units of alcohol dehydrogenase (ADH) from *Saccharomyces cerevisiae* (Sigma-Aldrich) using 50 mM DR and 2 μM purified DERA wild-type enzyme or variant, and 0.3 mM NADH in 50 mM Tris-HCl, pH 7.5, supplemented with 5 mM $MgCl_2$. Cleavage of DR by DERA liberates acetaldehyde, which is converted to ethanol by ADH in NADH consuming reaction.

## Sequential aldol addition reaction of acetaldehyde by DERA enzymes

The sequential aldol addition of acetaldehyde was monitored by incubating 5 μM DERA with different amounts (10–50

mM) of acetaldehyde in 50 mM Tris-HCl buffer pH 7.5 in ambient temperature for 20 h. The reactions were stopped by addition of acetonitrile (20 μl of reaction mixture + 80 μl of acetonitrile), clarified by centrifugation and analysed with a UPLC system (Waters) equipped with photodiode array detector. An Acquity BEH Amide column (2.1 × 100 mm, 1.7 μm, Waters) was used in 40 °C with 0.6 ml/min flow rate. The solvents used in the UPLC were eluent A: 50% acetonitrile/50% $H_2O$ and 10 mM ammonium acetate pH 9 and eluent B: 95% acetonitrile/5% $H_2O$ and 10 mM ammonium acetate pH 9. The column was equilibrated with 99.9% B, and 5 μl of sample was injected and eluted with program as follows: 0–0.4 min isocratic 99.9% B, 0.4–0.5 min gradient to 60% B, 0.5–2.0 min gradient 30% B and 2–5 min isocratic 99.9% B. The acetaldehyde concentration was followed by adsorption at 285 nm, and the formation of the aldol addition product by adsorption at 217 nm.

## Aldol addition of formaldehyde and acetaldehyde by DERA enzymes

The addition reaction of formaldehyde and acetaldehyde was monitored by incubating 5 μM DERA with 2 mM formaldehyde and 2 mM acetaldehyde in 50 mM Tris-HCl buffer pH 7.5 in ambient temperature. The reaction was stopped by addition of 2,4,-dinitrophenylhydrazine (2,4-DNPH) and acetonitrile, which also initiated the derivatisation reaction (Allen 1930). For derivatisation, typically 25 μl of the DERA reaction mixture was transferred to an Eppendorf tube containing 70 μl of 2,4-DNPH mixture (5 μl of saturated 2,4-DNPH, 65 μl acetonitrile, and 30 mM phosphoric acid). The derivatisation reaction was allowed to proceed at 22 °C for 1 h or overnight in + 4 °C. After derivatisation, the samples were clarified by centrifugation and injected to an Acquity BEH UPLC C18 column (2.1 mm × 50 mm, 1.7 μm, Waters) equilibrated with 70% $H_2O$, 30% acetonitrile and eluted with isocratic elution with the equilibration buffer using 0.5 ml/min flow rate. The derivatised aldehydes were detected by measuring the absorbance at 360 nm (Allen 1930).

## Identification of the aldol addition product of formaldehyde and acetaldehyde

Propanediol oxidoreductase is an NAD-dependent enzyme that oxidises 1,3-propanediol (1,3-PD) to generate 3-hydroxypropionaldehyde (3-HPA). *Kp* PDOR was incubated with 30 mM 1,3-PD, 5 mM NAD, in 50 mM Tris-HCl buffer, containing 2 mM DTT and 1 mM $MnCl_2$, pH 7.5 in ambient temperature. The aldehydes in the reaction were detected by reversed-phase UPLC after derivatisation with 2,4-DNPH as described above.

The mass of the product of DERA catalysed addition reaction of 10 mM formaldehyde and 10 mM acetaldehyde after 2-h incubation at 22 °C was analysed by LC-MS using a C18 column. After derivatisation with 2,4-DNPH, the reaction was separated using 75:25 water plus 1% formic acid/acetonitrile.

## Analysis of DERA catalysed aldol addition products by NMR spectroscopy

NMR experiments were carried out at 22 °C in 50 mM Na-phosphate buffer, pH 6.8, containing 10% of $D_2O$ (Aldrich). Bruker Avance III NMR spectrometer equipped with a QCI H-P/C/N-D cryoprobe was used. In 1D $^1H$ experiments, the water signal was suppressed by 4-s-long volume selective presaturation (so-called NOESY presaturation) using Bruker's pulse program *noesygppr1D*. For 2D COSY, TOCSY, HSQC and HMBC standard Bruker pulse programs with water signal presaturation were used. In TOCSY, the mixing time (DIPSI2) was 80 or 120 ms, and in HSQC, adiabatic inversion pulses were used and the $^1H$ decoupling was achieved by adiabatic CHIRP sequence. The long range $^1H,^{13}C$ coupling constant in HMBC was set to 8 Hz. The chemical shifts were referenced to internal TSP (3-propionic-2,2,3,3-d4 acid sodium salt, Aldrich). The spectra were processed with Topspin 3.5, pl 7 software (Bruker).

## Circular dichroism spectroscopy to determine the thermostability

Temperature-induced unfolding of the purified DERA proteins from different organisms was measured by circular dichroism (CD) spectroscopy. The purified DERAs were diluted in 10 mM Tris-HCl buffer, pH 7.5 to 3 μM concentration. CD spectra were recorded from 240 to 190 nm using a 1 mm cell and a bandwidth of 1 nm with Chirascan CD spectrophotometer (Applied Photophysics, UK) at 20 °C. The unfolding curves were measured at 222 nm by gradually increasing the sample temperature with a gradient of 2 °C/min until a temperature of 90 °C was reached.

## Development of machine learning (ML) models for DERA mutant screening

A novel ML model was used to automatically predict substrate specificities of DERA mutants based on Gaussian processes, as summarised in Fig. 1. See the Supplementary material Text S1 for a more detailed description of the ML model.

## X-ray crystallography of Ec DERA variants

The three-dimensional structures for two *Ec* DERA variants were determined by X-ray crystallography. The N21K mutant, as a 1.2-mg/ml solution in 50 mM Tris-HCl buffer, pH 8.0, could be crystallised directly from the buffer. The C47V/G204A/S239D mutant, as 1.3-mg/ml solution in

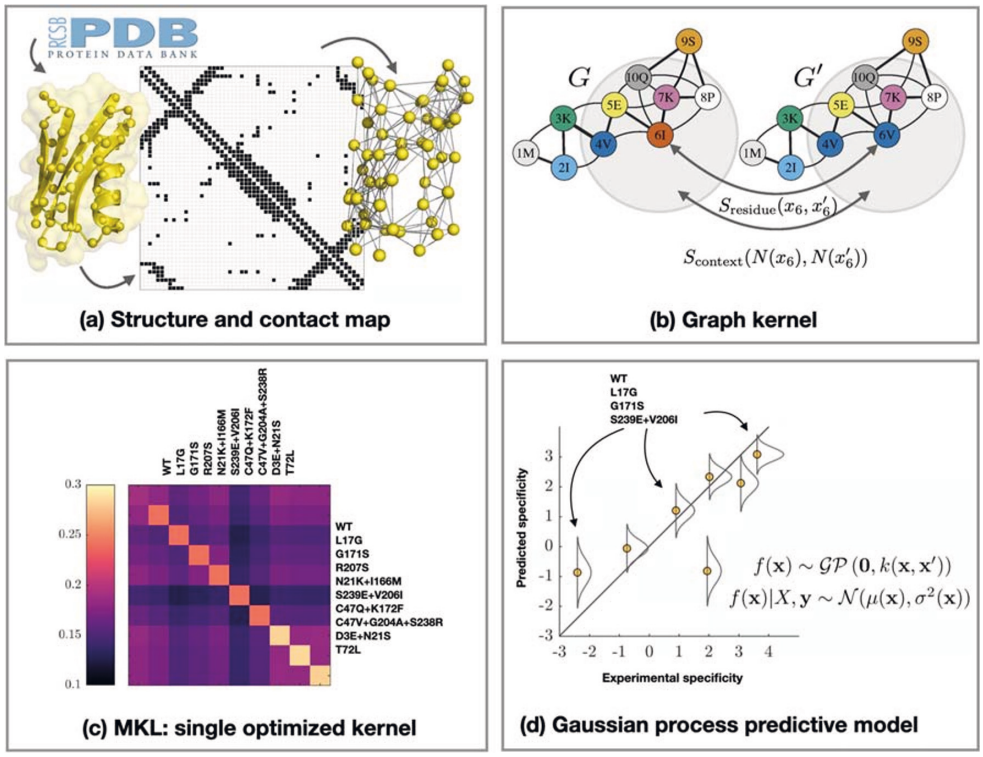



**Fig. 1** Illustration on the machine learning (ML) framework used in this work. The DERA protein structure is encoded as a contact map (**a**), which is combined with multiple node and edge substitution matrices to compute a graph kernel (**b**). We performed multiple kernel learning (MKL) (**c**) to find an optimised kernel to be used in Gaussian process predictive model (**d**). The kernel matrix measures variant similarities informative for substrate specificity. The substrate specificity predictive model is trained using experimental data

50 mM sodium phosphate buffer, pH 7.5, precipitated immediately in all conditions, and therefore, it was transferred to 50 mM Tris-HCl, pH 8.0 buffer for crystallisation using a PD-10 desalting column and concentrated to 3.2 mg/ml. Concentration was done in a centrifuge using a Vivaspin 2 column with 10 kDa molecular weight cut-off.

Both protein samples were crystallised using hanging-drop vapour diffusion. The N21K mutant was crystallised directly from a crystallisation solution with 18–20% PEG4000, 0.2 M magnesium formate and 0.1 M Tris-HCl pH 8.0. The C47V/G204A/S239D mutant formed large, rugged needles in 21% PEG3350, 0.2 M magnesium formate and 0.1 M Bis-Tris, pH 6.5, which were crushed and used in streak seeding, which produced good crystals in the same conditions but 17–18% PEG3350. Prior to X-ray diffraction measurements, crystals were soaked in solutions equivalent to their respective crystallisation solutions plus 0.1 M ligand, D-2-deoxyribose-5-phosphate (DRP), for 2 days.

Crystals were measured at European synchrotrons. They were mounted in nylon loops in cryoprotectant solutions equivalent to their respective crystallisation solutions but with 40% PEG, stored in liquid nitrogen and sent to the synchrotrons for remote measurement. Crystals of the N21K mutant were measured on the ID30A-1 beamline at the European Synchrotron Radiation Facility (ESRF), and crystals of the C47V/G204A/S239D mutant were measured on the I24 beamline at Diamond Light Source (DLS). Autoprocessed MTZ files generated by EDNA_proc and xia2-3dii programs were selected for structure determination for crystals measured at ESRF and DLS, respectively.

In addition, a large crystal of the N21K mutant was measured using the home X-ray diffractometer: Nonius FR591 rotating anode X-ray source by Bruker, mar345dtb goniometer system and mar345 image plate detector by X-ray Research (marXperts). The crystal was transferred to a cryoprotectant solution with 40% PEG4000, 0.2 M magnesium formate and 0.1 M Tris-HCl, pH 8.0 and placed into the sample holder in a nylon loop, where it was cooled down to constant 100 K in cold nitrogen stream. The detector was set to the minimum allowed distance of 150 mm, equivalent to 1.86 Å resolution limit. A data set was collected, and the crystal diffracted beyond

the resolution limit. Images were processed with XDS program package (Version November 1, 2016).

All structure determination calculations were done with PHENIX software suite (Moriarty et al. 2009; Chen et al. 2010; Adams et al. 2010). Phasing was done using phenix.phaser (McCoy et al. 2007) molecular replacement, and previously published wild-type DERA structure (PDB entry 1KTN, no article published) was used as initial model. Mutations were done manually in Coot (Emsley et al. 2010) after molecular replacement. Structures were then refined using phenix.refine (Afonine et al. 2012). Water molecules were first added using the "Update waters" option and later checked and corrected manually. Appropriate ligands were placed in the active sites, and the aldehydes bound as Schiff base were connected to the amino group of the catalytic lysine with appropriate geometry restraints. Presence of partial DRP in the N21K mutant structures was further confirmed by calculating Polder maps (Liebschner et al. 2017). For the final refinement rounds, weight optimisation options were enabled.

### GenBank accession numbers

The nucleotide sequences of codon-optimised DERA and PDOR-coding genes used in this study can be found in the GenBank with the following accession numbers: MT702750 for DERA-coding gene from *E. coli*, MT702753 for DERA-coding gene from *Geobacillus*, MT702754 for DERA-coding gene from *Thermus thermophilus*, MT702748 for DERA-coding gene from *Coccidioides immitis*, MT702749 for DERA-coding gene from *Lactobacillus brevis*, MT702752 for DERA-coding gene from *Talaromyces islandicus*, MT702751 for DERA-coding gene from *Aspergillus niger* and MT682136 for *Klebsiella pneumoniae dhaT* gene.

## Results

### Selecting the most suitable DERA enzyme for the protein engineering work

Several known DERA enzymes of bacterial and fungal origin were initially considered as a target enzyme for the protein engineering work. For the proper comparison, we decided to express in *E. coli* the DERA aldolases from *E. coli*, *Geobacillus* sp., *Thermus thermophilus*, *Lactobacillus brevis*, *Coccidioides immitis*, *Aspergillus niger* and *Talaromyces islandicus*. The characterisation data of the purified enzymes is shown in Table 1. All seven DERAs were shown to be promiscuous, accepting both the natural substrate DRP and the non-phosphorylated version, DR. Moreover, acetaldehyde was also shown to be an acceptor substrate for all the characterised DERAs (as measured in the addition reaction). As high expression level and good thermostability were also considered to be relevant properties, *E. coli (Ec)* DERA was chosen as the target for our mutagenesis work.

### Setting up the analytics for the DERA-catalysed reactions

The activity measurements for DERA wild-type and mutants in the cleavage direction, e.g. cleavage of the natural substrate DRP and the non-phosphorylated substrate DR, were carried out based on the methodology described in the literature. These methods are applicable also with crude cell extracts (data not shown), but we decided to purify the enzymes to be able to accurately compare the specific activities of the different DERA variants (including different DERAs and *Ec* DERA mutants).

The assay for acetaldehyde addition activity was set up for LC using an amide column in alkaline conditions. It was noticed that during the course of the DERA reaction, a product peak appeared (Fig. S1), which could be detected by absorption

**Table 1** Characterisation of seven purified DERA enzymes, expressed in *E. coli* and purified with a His-tag

| Microbial source of the DERA enzyme | Yield of purified protein from 50-ml cultivation | Relative activity on 5 mM DRP[a] | Relative activity on 50 mM DR[a] | Relative activity on 30 mM acetaldehyde[a] | $T_m$ (°C)[b] |
|---|---|---|---|---|---|
| *E. coli* | 12 mg | 1 | 1 | 1 | 65 ± 1 |
| *Aspergillus niger* | 2.1 mg | 0.3 | 1.0 | 0.4 | 48 ± 1 |
| *Talaromyces islandicus* | 2.5 mg | 0.3 | 0.6 | 0.9 | 47 ± 1 |
| *Geobacillus* sp. | 3.8 mg | 0.4 | 1.4 | 1.4 | 75 ± 1 |
| *Thermus thermophilus* | ~ 1 mg | 0.1 | 0.7 | 0.7 | ≥ 90 |
| *Coccidioides immitis* | 1.4 mg | 0.3 | 0.8 | 1.1 | 39 ± 1 |
| *Lactobacillus brevis* | 1.7 mg | 0.6 | 0.3 | 1.2 | 38 ± 1 |

*DRP* deoxyribose-5-phosphate, *DR* deoxyribose

[a] Activities are presented relative to *E. coli* DERA activities

[b] Thermostability ($T_m$) of the purified protein was determined with CD spectroscopy

at 217 nm. The main product of DERA activity on acetaldehyde has been shown to be an aldol addition product of three acetaldehyde molecules in a sequential reaction, where DERA first adds two acetaldehydes to form an C4 aldehyde (3-hydroxybutanal), which is then once more coupled with acetaldehyde into a C6 product, 2,4,6-trideoxyhexose. This product cyclises spontaneously to a hemiacetal and is thus removed from the reaction (Gijsen and Wong 1994). The C6 product formation from the DERA-catalysed reactions could not be quantified in the LC assay due to the fact that 2,4,6-trideoxyhexose is not commercially available as a standard. However, the identification of the formed cyclic trideoxyhexose product was verified by NMR. In the NMR experiments, several products from the enzymatic reaction were observed. After identifying the different spin systems from TOCSY spectra, their structures were determined by standard 2D NMR methods and the chemical shifts were compared with those published in Dick et al. (2016). The main products were the first aldol addition product 3-hydroxybutanal (17%) and the second aldol addition product that has undergone a spontaneous cyclisation to two anomers of pyranose rings (73% and 10%). In addition, a very small amount of crotonaldehyde was detected, a condensation product from two acetaldehyde molecules, that has been reported to be a side-reaction product of DERA (Dick et al. 2016).

Acetaldehyde addition reaction was carried out using acetaldehyde as a sole substrate, i.e. acetaldehyde acts both as acceptor and donor substrate. In the literature, very high acetaldehyde concentrations (e.g. 300–500 mM) are often used to demonstrate the synthesis of 2,4,6-trideoxyhexose with DERA. However, high acetaldehyde concentrations have also been shown to inhibit the DERA activity (Jennewein et al. 2006; Dick et al. 2016; Bramski et al. 2017). The conditions for the enzymatic acetaldehyde addition reaction were thus chosen so that roughly saturating substrate concentration of 30 mM was used. The acetaldehyde standards were reproducible in our LC method; however, the acetaldehyde concentrations in the presence of DERA, measured before or after incubation, were noisier. This was assumed to be because of covalent binding of the acetaldehyde to the enzyme protein (Dick et al. 2016).

In order to analyse DERA-catalysed addition reaction of acetaldehyde with another aldehyde, activity assay based on derivatisation of the aldehydes with 2,4-DNPH in acidic conditions, followed by a reversed-phase LC separation with UV detection, was set up. This method allowed detection of the aldehyde substrates as well as formed aldol addition products present in the DERA reactions.

## About the DERA protein engineering approaches in this work

Engineering of the substrate specificity is often attempted through rational mutagenesis, targeted near the active site or the substrate binding area. Even though the 3D structure of *Ec* DERA is available in high resolution and also in complex with the natural substrate DRP (Heine et al. 2001), it was challenging to rationally design mutations towards improved activity on small non-phosphorylated aldehydes. On the other hand, directed evolution including random mutagenesis to the whole gene, or even random mutagenesis to a targeted area, puts a stress on screening of the activities in high throughput manner. In the present study, we used targeted mutagenesis during the 1st round to create single amino acid mutants of *Ec* DERA. The most beneficial mutants were then combined based on activity data to create double and triple mutants. Furthermore, saturation mutagenesis at certain amino acid spots was also carried out. Finally, during the 3rd round, double or triple point mutants were created using machine learning algorithm predictions, explained in more details in Supplementary section and below. Altogether, roughly 150 *Ec* DERA mutants were characterised during the course of the work.

## Site-directed mutagenesis to make single amino acid mutations to Ec DERA (1st round)

Altogether, 69 single *Ec* DERA mutants, targeting 24 amino acid positions, were created with site-directed mutagenesis, expressed in *E. coli* as His-tagged proteins, and purified with Ni-NTA spin columns. The amino acid positions to be mutagenised were chosen mainly by examining the high-resolution crystal structure in complex with the native *Ec* DERA substrate, DRP (Heine et al. 2001; PDB id. 1JCL). In addition, sequence alignments of DERAs from different origins were utilised to find conserved amino acids and possible targets for consensus mutations. Moreover, the online tool HotSpot-Wizard (Bendl et al. 2016) and literature were utilised when selecting the residues to be mutated. Most of the mutations were targeted in close vicinity (maximum 4 Å distance) of the active site and the substrate binding pocket (Fig. 2). The activities of the purified *Ec* DERA variants were measured with three different substrates: DRP and DR for the cleavage reaction and acetaldehyde for the aldol addition reaction. Altogether, 30 variants having single-point mutations had clearly reduced specific activity (20% or less) towards the natural substrate DRP as compared with the wild-type enzyme (Figs. 3 and 4 and S2). Of these 30 *Ec* DERA variants, five showed additionally improved activity on acetaldehyde aldol addition, i.e. the *Ec* DERA mutants G204A, S239E, L17G, G171A and G171S (Fig. S2).

## Ec DERA variants containing two or three amino acid mutations (2nd round)

Altogether, 62 double or triple *Ec* DERA mutants were created by (a) manually combining the most interesting

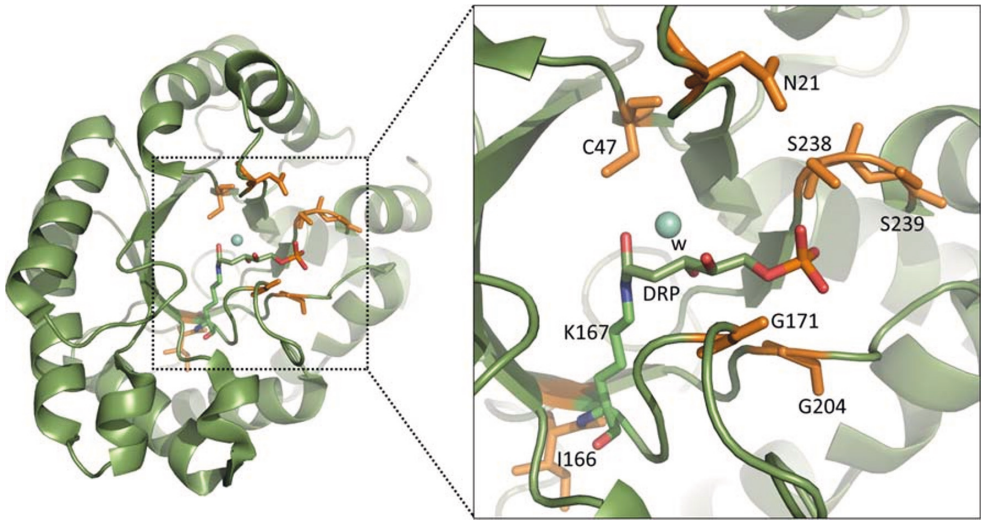



**Fig. 2** The 3D structure of covalent complex of *E. coli* DERA with DRP (PDB id. 1JCL Heine et al. 2001). Mechanistically relevant water molecule is shown as light-blue sphere. Residues N21, C47, I166, G171, G204, S238 and S239, which had the most beneficial effects, mutated in the study are shown in orange. The Schiff-base forming lysine (K167) is shown in green

single mutations and (b) by using saturation mutagenesis on selected spots, as described in more details under Materials and methods section. In the manual combination, one of the selection criteria used was to pick mutations that affected the acetaldehyde addition activity. The *Ec* DERA variants included 34 double or triple mutants, which were combined from the single mutants based on the activity data. In addition, 28 double mutants were created using saturation mutagenesis on amino acids C47, I166 and S238. In each case, these mutants were made on top of the *Ec* DERA variant N21K. The rational in picking the spots for the saturation mutagenesis was that these three amino acid positions C47, I166 and S238 were found to affect favourably to the acetaldehyde addition reaction (Fig. 3). Moreover, C47 has been previously shown to be a target for inactivation by aldehydes (Dick et al. 2016) and based on the *Ec* DERA wild-type complex structure. In addition, S238 is hydrogen bonded to the phosphate group of the natural substrate DRP. Overall, several interesting variants having low activity on both DRP and DR substrates and improved acetaldehyde addition activity were discovered among all the screened double and triple *Ec* DERA variants in the 2nd mutagenesis round, e.g. N21K/C47V, N21K/C47L, N21K/C47F, N21K/C47S, and N21K/S238G. Additionally, the *Ec* DERA variants S239/V206I/L17H, S239/V206I/I166M, and S239D/V206I showed reduced activity on DRP and improved acetaldehyde addition activity (Figs. 4 and S3). It should be also noted that in most cases, no additivity effect of point mutations in terms of activity data could be detected.

## Ec DERA variants created using machine learning (3rd round)

A novel machine learning (ML) model to automatically predict substrate specificities of DERA mutants based on Gaussian processes was developed. Our goals were to train the specificity prediction functions from specificity observations and use these subsequently for screening new potential *Ec* DERA variants. See the Supplementary material Text S1 for a technical description of the development and training of the ML model.

A machine learning model was trained using the data from the available *Ec* DERA mutants, consisting of (i) the 69 single point mutants and (ii) the 62 double or triple point mutants, in total 131 mutants. Each mutant had measured data on DRP, DR and acetaldehyde specificity. The ML model was trained to predict all three substrate specificities. In Fig. S4, the trained combinations of substitution models are shown. The DRP model ended up using six different substitution models whereas the DR model used five, with amino acid interaction features having most predictive information. The acetaldehyde model only used four substitution models with contact energy and packing features having highest weights.

The predicted and measured activity data of the *Ec* DERA mutants are shown in Fig. S5. The ML model achieved cross-validation test correlation on the first two rounds of 0.57 for DRP, 0.81 for DR and 0.54 for acetaldehyde, respectively. The 131 data points were sufficient for the ML model to explain the substrate specificity of each of them.

After this, we screened in silico all possible *Ec* DERA variants with 1–3 amino acid mutations using the ML model,

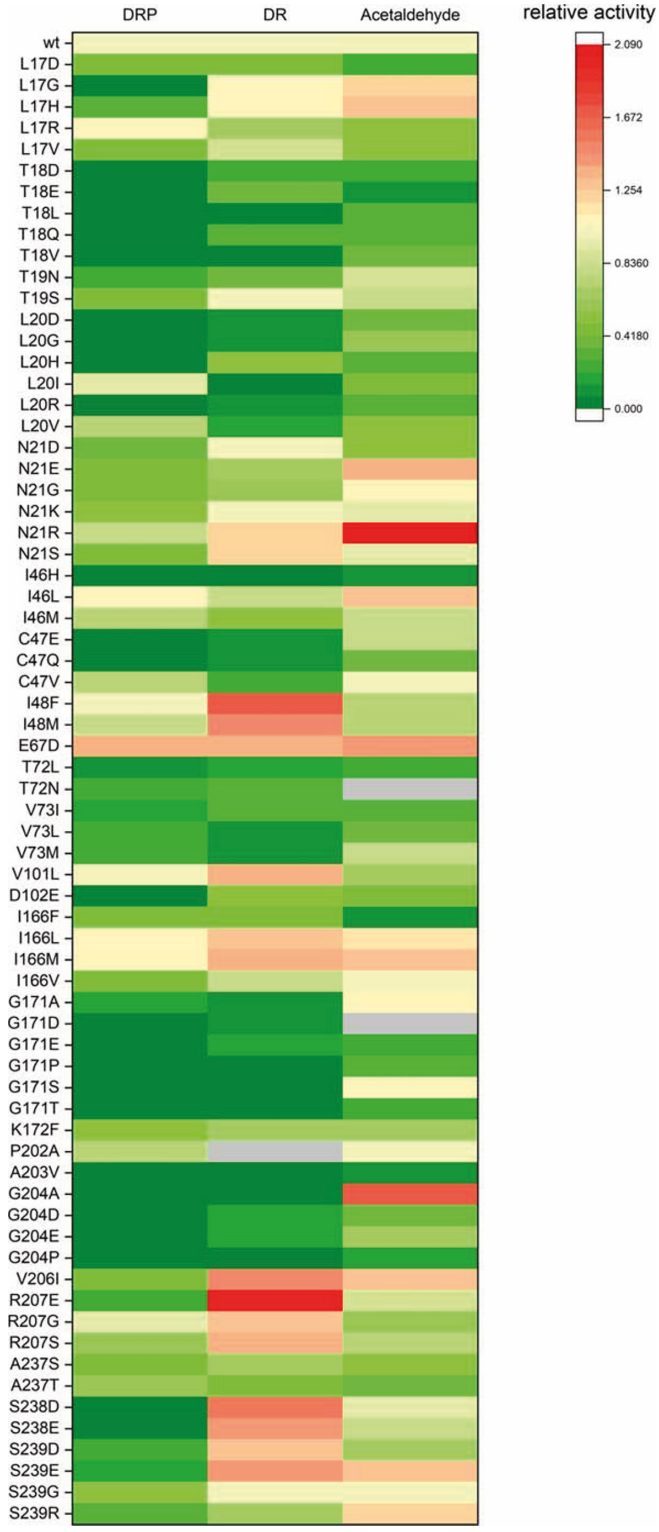



**Fig. 3** Heatmap displaying the 2-deoxyribose 5-phosphate (DRP) and 2-deoxyribose (DR) cleaving activities and acetaldehyde addition activities of all *Ec* DERA single point relative to the wild-type enzyme (wt)

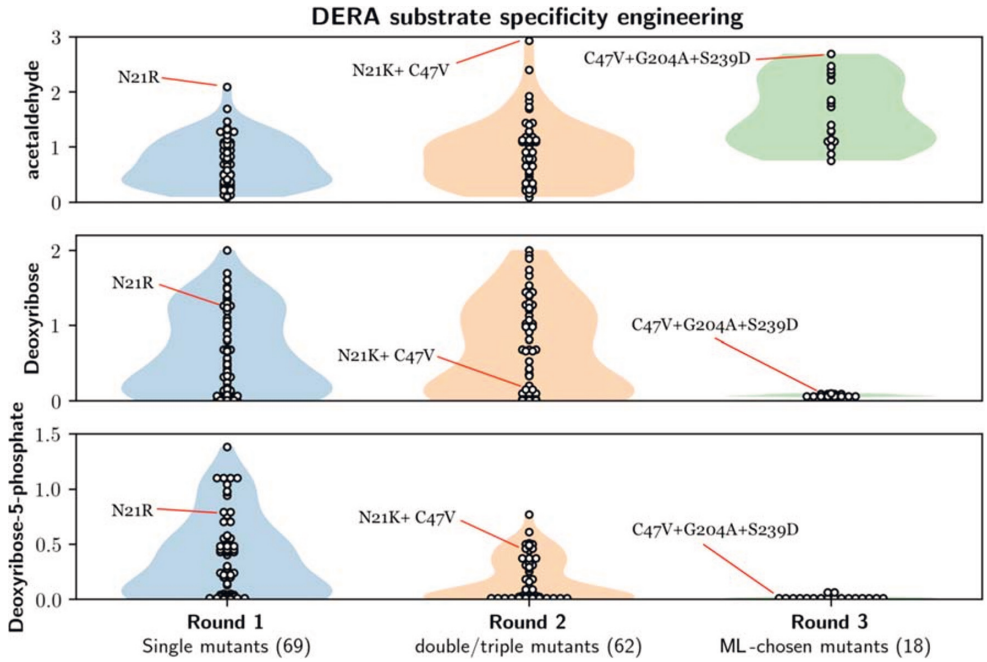



**Fig. 4** Summary of the substrate specificities of *Ec* DERA variants on deoxyribose-5-phosphate (DRP), deoxyribose (DR) and acetaldehyde (rows) over three mutagenesis rounds (columns), demonstrating that the ML-optimised mutants have abolished activities on DRP and DR, and increased acetaldehyde specificity. The white circles indicate specificities of individual DERA variants, and the shaded violin plots indicate smoothed vertical histograms (i.e. overall number as percentages from the total number of mutants)

in total 48,000 new variants. These were sorted based on the predicted acetaldehyde specificity, subtracted with DRP specificity to find the most probable candidates with high acetaldehyde but low DRP specificity. From the top 50 best estimated variants, 18 mutants were manually chosen for in vitro mutagenesis experiments.

The results of these 18 *Ec* DERA variants (containing two or three point mutations) showed that all except three had improved activity on acetaldehyde addition, and five of the variants had more than two-fold improved acetaldehyde addition activity (Figs. 4 and 6). In addition, for all 18 *Ec* DERA mutants, the DR and DRP activity was almost completely abolished (Figs. 4 and 6). The Fig. S6 shows the correlations of the 18 final variants, which indicate a 0.99 correlation for DRP and DR due to their successful specificity removal and a 0.96 correlation between estimated and measured acetaldehyde specificity. We note that one should not directly compare these correlations to the cross-validation correlations due to active selection procedure of the 3rd round variant.

## Crystal structure analysis

The crystal structures of two variants of *Ec* DERA were determined, in order to elucidate mutation-induced changes in the 3D protein structure. The crystal structure of *Ec* DERA N21K variant, which had lowered specific activity on DRP, and wild-type like activity on DR and acetaldehyde, was determined with and without the ligand, D-2-deoxyribose-5-phosphate (DRP). In addition, the crystal structure of one of the best *Ec* DERA variants, C47V/G204A/S239D, having improved activity towards acetaldehyde and basically no activity on DRP or DR, derived from the 3rd mutagenesis around, was determined from the crystal soaked with DRP. Diffraction resolutions for all three structures were high (1.5 to 1.9 Å) and the crystallographic R-factors were very low (Table S2), indicative of high quality 3D structures in each case. The determined three 3D structures were superimposed with the crystal structures of the *Ec* DERA wild-type without (pdb code 1p1x, 1.0 Å resolution) and with a ligand complex (1-hydroxy-pentane-3,4-diol-5-phosphate)(1jcl, 1.1 Å resolution), to analyze mutant-induced changes in the 3D structures.

The complex structure of *Ec* DERA N21K mutant revealed binding of the reaction product from DRP cleavage, i.e. glyceraldehyde-3-phosphate (C3P), to the acceptor site in one of the two protein molecules in the asymmetric unit (Fig. 5a). A similar conformational change in the loop around residue S238 in both the Ec DERA N21K mutant and wild-type structures can be seen upon glyceraldehyde-3-phosphate binding, when uncomplexed and complexed crystal structures are compared. In addition, in both N21K crystal structures,

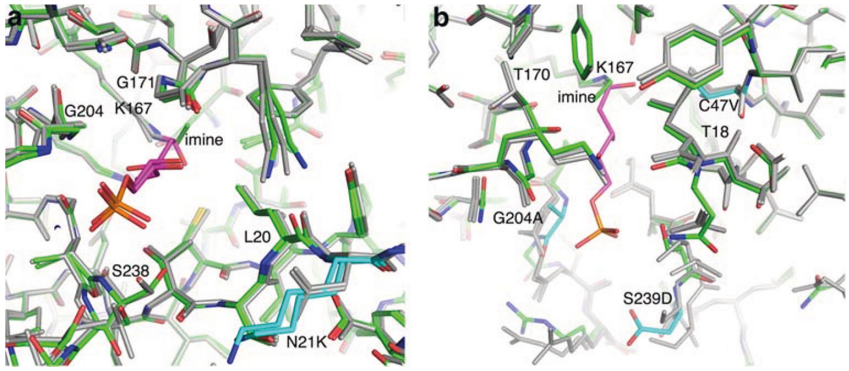



**Fig. 5** The superimposition of the crystal structures of the *Ec* DERA wild-type and two variants around the active site as stick models. *Ec* DERA wild-type structures as uncomplexed and complexed form are shown in grey. The covalently bound 1-hydroxy-pentane-3,4-diol-5-phosphate is in purple. **a** The superimposition with the uncomplexed and complexed form of *Ec* DERA N21K variant. Variant structures in uncomplexed and complexed forms are shown in green and bound ligand, glyceraldehyde-3-phosphate, in purple. **b** The superimposition with the uncomplexed form of *Ec* DERA C47V/G204A/S239D variant. The variant structure is shown in green and the mutated residues are in cyan.

there are slight (about 0.2 to 0.4 Å) movements in the positions of altogether three active site loops containing residues G171, G204 and L20, respectively, when compared with the wild-type DERA crystal structure. The N21K mutation is located in the third loop. These small movements slightly narrow the access to the active site of the N21K variant and may thus contribute to the decreased activity against DRP in many of the N21K containing double mutants tested in this work.

The structure of *Ec* DERA C47V/G204A/S239D variant was determined from crystals after soaking with DRP. The crystal structure showed a presence of a very small covalent adduct to the catalytic amino acid K167 in both molecules in the asymmetric unit. The electron density map showed elongated electron density corresponding to approximately two carbon atoms. In the refinement, it was modelled as an imine group representing a reaction intermediate. No binding of DRP or glyceraldehyde-3-phosphate could be detected. This could imply that binding of the DRP is clearly diminished in the C47V/G204A/S239D variant (Fig. 5b). The conformational differences in this variant as compared with the *Ec* DERA wild-type structure are clearly detectable. The conformation of the loop containing mutation G204A has pushed the position of its Cα atom by 1.0 Å towards active site, thus narrowing it. In addition, there are clear changes in the loop structure containing the S239D mutation. Because the residues of this loop participate in phosphate binding (of glyceraldehyde-3-phosphate acceptor), the S239D mutation would probably decrease binding of DRP both by introducing a negatively charged amino acid residue (phosphate being also negatively charged) and by narrowing the active site entrance. On the other hand, the C47V mutation, located relatively close to the catalytic K167 residue (about 5 Å), has caused only minimal alteration in the position of this amino acid residue, as it is located in the middle of β-strand.

## Further substrate promiscuity testing of Ec DERA variants

The goal of the protein engineering work was to improve the substrate specificity towards smaller, non-phosphorylated aldehydes over glyceraldehyde-3-phosphate. Acetaldehyde is a two-carbon aldehyde, and the only shorter aldehyde is formaldehyde, which is the simplest existing aldehyde. We decided to test also formaldehyde as acceptor in the DERA reaction catalysed by *Ec* DERA variants. The reference test with the wild-type *Ec* DERA enzyme indicated that this was already able to add formaldehyde to the acetaldehyde donor. The verification of the aldehyde addition reaction was carried out using the reversed-phase chromatography after derivatisation with 2,4-DNPH. Here, the product by *Ec* DERA catalysis was found to elute with an identical retention time to the product of *Kp* PDOR catalysed oxidation of 1,3-propanediol (Fig. S7), which is known to be 3-hydroxypropionaldehyde (3-HPA) (Johnson and Lin 1987). Furthermore, the mass m/z 255.0729, which is identical to 2,4-DNPH derivatised 3-HPA, was detected in the reaction, also suggesting 3-HPA to be formed by *Ec* DERA from formaldehyde and acetaldehyde. Moreover, the DERA products of formaldehyde with acetaldehyde were identified as 3-hydroxypropionaldehyde (30%) and the corresponding hydrate (70%) by 1D and 2D NMR experiments. The $^{1}$H NMR spectrum of the products and the product structures are shown in Fig. S8. Interestingly, the DERA addition reaction products of acetaldehyde with itself were not observed here at all.

After this, altogether, 140 *Ec* DERA variants were assayed for addition activity on formaldehyde and acetaldehyde. Of these, six *Ec* DERA variants were found to have improved activity as compared with the wild-type enzyme (Fig. S9). Two of the mutants had single mutations, two were double

and two triple mutants. The relatively small number of *Ec* DERA variants with higher activity on formaldehyde addition to acetaldehyde as compared with the wild-type enzyme was not surprising, as both the screening and the machine learning prediction algorithm were set up to maximise the acetaldehyde addition reaction.

## Discussion

Deoxyribose-5-phosphate aldolases (DERAs) are acetaldehyde-dependent, Class I aldolases catalysing in nature a reversible aldol reaction between an acetaldehyde donor (C2 compound) and glyceraldehyde-3-phosphate acceptor (C3 compound, C3P) to generate deoxyribose-5-phosphate, DRP (C5 compound). The interesting feature is the substrate promiscuity as DERA enzymes have been shown to accept a wide range of aldehydes as acceptor molecules, thus offering a biocatalytic alternative for a (stereo)selective synthesis of C–C bonds. Substrate specificity on the donor substrate side is stricter but recent biodiversity screen has revealed that some DERAs also display nucleophile substrate promiscuity (Hernández et al. 2018; Chambre et al. 2019). Furthermore, DERA enzymes can carry out a tandem reaction with acetaldehyde as a sole substrate, leading to formation of a C6 product, 2,4,6-trideoxyhexose, which cyclises spontaneously and is removed from the reaction. DERA enzymes have been utilised in large-scale synthesis of pharmaceutical intermediates, i.e. statin precursors and pyranoid building blocks, as well as preparation of different types of deoxysugars, deoxyketoses and deoxy-sialic acid (Haridas et al. 2018). Despite of being promising enzymes for application purposes, there are still some challenges related to their usage. In particular, enzyme inactivation under synthesis conditions represents a major obstacles (Jennewein et al. 2006; Dick et al. 2016; Bramski et al. 2017), and thus, variants having improved substrate binding towards non-natural substrates and/or resistance towards high aldehyde concentrations are desired.

The 3D structures of several DERA enzymes have been solved, and there are also a few complex structures available. These have revealed that despite of relatively low sequence identity, all DERA enzymes have the ubiquitous TIM $(\alpha/\beta)_8$-barrel fold where the catalytic amino acids as well as other amino acid residues around the active site seem to be relatively well conserved (Heine et al. 2001; Haridas et al. 2018). The DERA reaction proceeds via Schiff base formation between an active site lysine residue (K167 in *Ec* DERA) and the donor acetaldehyde substrate. The active site of DERA is located in a deep binding cleft, where the donor acetaldehyde binds to the bottom of the cleft and the acceptor glyceraldehyde-3-phosphate to the upper part, near the cleft entrance. Binding of the substrates to the active site cleft is mediated through hydrogen bonds, either directly or through water molecules. Concerning the acceptor site, the hydrogen bonds are particularly directed to the phosphate group of the native substrate (glyceraldehyde-3-phosphate) (Heine et al. 2001; Heine et al. 2004). Based on the published *Ec* DERA complex structures, residues specific for phosphate-binding are the backbone amide groups of S238 and G205 and via a water molecule the residues G204, V206, S239 and G171. Residue K172 provides both a counter-charge and forms a hydrogen bond (via a water molecule) with the phosphate group (Heine et al. 2001).

In this work, we wanted to study the substrate specificity of DERA with various protein engineering approaches, including also machine learning methods. We set as our goal to improve the overall performance of DERA on utilising non-phosphorylated short aldehydes (C3 and C2). Initially, we characterised seven different purified DERA wild-type enzymes of bacterial and fungal origin and found that several of them were promiscuous and could also accept non-phosphorylated aldehydes as the acceptor substrate (Table 1). *E. coli* (*Ec*) DERA was chosen for the protein engineering work as it showed relatively good promiscuous activity towards the desired reactions and had high expression level and good thermostability (Table 1). Additionally, high-resolution complex structures of *Ec* DERA exist both with and without bound substrate (Heine et al. 2001; Heine et al. 2004), and some mutagenesis studies to change its substrate preference have also been carried out, thus providing a good starting point for semi-rational mutagenesis approaches.

DERA mutagenesis to alter the substrate specificity was targeted to the active site of the enzyme (Fig. 2). Initially, Web server HotSpot-Wizard, 3D structures of *Ec* DERA (Heine et al. 2001; PDB id. 1JCL) and literature were utilised when selecting the amino acid residues to be mutated. First mutagenesis round DERA variants were made as single amino acid mutants, and in the 2nd mutagenesis round, beneficial mutations were combined. Furthermore, saturation mutagenesis at certain amino acid spots was also carried out. The 3rd round of mutagenesis was carried out using machine learning (ML)-guided approach. The specific activities of the purified *Ec* DERA variants were measured initially with three different substrates: (1) DRP (cleavage reaction), (2) DR (cleavage reaction), (3) acetaldehyde (aldol addition reaction using acetaldehyde both as the donor and acceptor substrate). Altogether, roughly 150 purified *Ec* DERA mutants, having one to three point mutations, were characterised during the work. Several of the *Ec* DERA variants showed clear change in their substrate spectra (Figs. 3, 4, 6 and S4–S5). The most promising variants had substantially reduced, or completely abolished activity particularly towards the natural substrate (DRP), while showing activity on acetaldehyde addition reaction. Interestingly, we also discovered that most of the tested DERA wild-type enzymes could also accept, besides aldehyde, formaldehyde (C1 aldehyde) as the acceptor molecule.

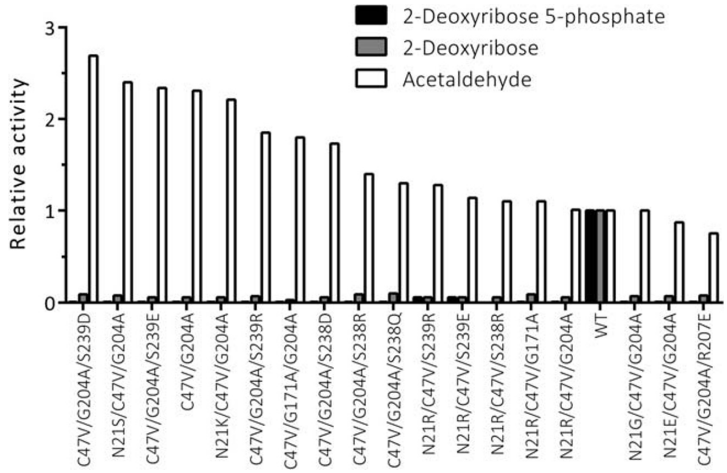



**Fig. 6** The characterisation data of the 18 *Ec* DERA variants containing two or three mutations, created by optimisation with machine learning algorithm. Cleaving activities on 2-deoxyribose 5-phosphate (DRP) and 2-deoxyribose (DR) and on acetaldehyde addition activity are shown as relative activities compared with the wild-type *Ec* DERA (WT)

This prompted us to test the *Ec* DERA variants on formaldehyde (C1) utilisation (i.e. aldol addition reaction between acetaldehyde and formaldehyde). The results demonstrated that some of the mutants had also altered preference towards formaldehyde (C1 aldehyde), the following six variants being the most potent: L17H, N21D, G171T, N21R/R207E, V206I/S239R, N21R/D206I/S239R and E67D/N206I/S239R (Fig. S9). Interestingly, the acetaldehyde and formaldehyde activities did not always correlate, and none of the 3rd round variants (from the ML-guided mutagenesis) that had clear preference for acetaldehyde did not perform particularly well when formaldehyde was offered as an acceptor aldehyde. On the other hand, as discussed more thoroughly below, this also demonstrates the power of the ML-guided mutagenesis combined with screening.

During the work, a sample-efficient machine learning method for designing the optimal mutation strategy for protein engineering was developed. The proposed Gaussian process model is a principled statistical model that excels in moderate to low data settings. The model learns the posterior distribution of specificity functions from observations, while simultaneously performing feature learning for added interpretability of substitution model choices. The Gaussian process model excelled at interpolation and moderate extrapolation of the DERA structure. The results from the 3rd round of *Ec* DERA mutagenesis guided with machine learning (ML) algorithm aiming at high activity on acetaldehyde addition reaction, and low activity on DRP and DR substrates is shown in Fig. 6. As can be seen, all 18 variants tested had basically abolished activity towards the native DRP (i.e. glyceraldehyde-3P acceptor) as well as towards the non-phosphorylated DR substrate (i.e. toward glyceraldehyde acceptor). Moreover, 15 out of 18 variants tested had increased target specificity towards the C2 acceptor aldehyde. Thus, our ML model was able to successfully extrapolate from the characterisation data novel mutant combinations with desired specificity. The usage of the ML model enhanced the mutagenesis work by a dramatic improvement over the conventional mutant selection procedures and these type of methods are clearly useful to speed up the protein engineering work.

The five best *Ec* DERA variants having clearly improved substrate specificity towards acetaldehyde (Fig. 6) were (1) C47V/G204A/S239D, (2) N21S/C47E/G204A, (3) C47V/G204A/S239E, (4) C47V/G204A and (5) N21K/C47V/G204A. By examining all our mutant data, we conclude that mutations to the amino acid residues N21, C47, I166, G171, G204, S238 and S239 seemed to have the most beneficial effects on changing the substrate specificity (Figs. 3, 6, S2-S3). Most of these residues (i.e. G171, G204, S238 and S239, N21) are according to the structural data (by others and us, see also below) involved in binding the phosphate group of the glyceraldehyde-3-phosphate acceptor substrate. Residue I166 is located more distant to the active site, in the loop underlying the active site lysine residue K167. The reason for I166 mutations affecting to the substrate specificity is not clear; however, others have also noted that mutations to this spot affect the substrate binding (Jennewein et al. 2006). Furthermore, the conserved residue C47, which is located at the bottom of the active site cleft, seems to be able to form in some circumstances covalent adducts with aldehydes leading to enzyme inactivation (Dick et al. 2016; Bramski et al. 2017). All our best five mutants on acetaldehyde addition reaction contained a mutation C47V and it is plausible that this mutation promoted the enzymatic reaction by relieving the aldehyde inhibition. Interestingly also, the DERA from the hyperthermophile Archaea *Aeropyrum pernix* has naturally a Valine residue at this position (C47 in *Ec* DERA) (Sakuraba et al. 2003).

In order to further rationalise our protein engineering results, we determined the crystal structures of the two *Ec*

DERA variants C47V/G204A/S239D and N21K as ligand complexes and compared these to the solved *Ec* DERA wild-type structures, as explained in more details under Results section (Fig. 5 and Table S2). We conclude that when aiming for improved binding of small non-phosphorylated aldehydes, mutations that narrow the substrate binding cleft near the entrance and/or affect to the binding of the substrate, in particular the phosphate group seem to be important. However, as evident from our variant data, purely rational design of the mutants remains challenging, and the substrate specificity improvement clearly benefited from saturation mutagenesis combined with ML-guided mutagenesis approaches. Overall, we could demonstrate that the synthetic utility of DERA enzyme may be substantially increased using protein engineering approaches.

**Acknowledgements** We thank Arja Kiema and Kirsi Kiiveri for excellent technical assistance. Ulla Lahtinen is thanked for the LC-MS analysis. We kindly thank ESRF, Grenoble and Diamond Light Source, Oxfordshire for providing the synchrotron facilities.

**Authors' contributions** SV, MA, AK, MP planned the work together with MH and JR. SV and MA conducted the laboratory work, except the crystallisation and structure determination, which was done by JP, NH and JR(Rouvinen). HM planned and performed the NMR experiments. MH, EJ, HL, SK and JR(Rousu) developed the machine learning methods. Everybody contributed in discussing of the results and writing the manuscript.

**Funding** Open access funding provided by Technical Research Centre of Finland (VTT). This work was supported by Business Finland (former Finnish Funding Agency for Technology and Innovation) through LIVING FACTORIES: Synthetic Biology for a Sustainable Bioeconomy (LiF; project number 40128/14) and by Academy of Finland through SA-ENGBIOCAT (decision numbers 288677 and 287241), and by Academy of Finland grant 299915.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethics declarations** This article does not contain any studies with human participants or animals performed by any of the authors.



## References

Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung L-W, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ, Moriarty NW, Oeffner R, Read RJ, Richardson DC, Richardson JS, Terwilliger TC, Zwart PH (2010) PHENIX : a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr Sect D Biol Crystallogr 66:213–221. https://doi.org/10.1107/S0907444909052925

Afonine PV, Grosse-Kunstleve RW, Echols N, Headd JJ, Moriarty NW, Mustyakimov M, Terwilliger TC, Urzhumtsev A, Zwart PH, Adams PD (2012) Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr Sect D Biol Crystallogr 68: 352–367. https://doi.org/10.1107/S0907444912001308

Allen CFH (1930) The identification of carbonyl compounds by use of 2, 4-dinitrophenylhydrazine. J Am Chem Soc 52:2955–2959. https://doi.org/10.1021/ja01370a058

Barbas CF, Wang YF, Wong CH (1990) Deoxyribose-5-phosphate aldolase as a synthetic catalyst. J Am Chem Soc 112:2013–2014. https://doi.org/10.1021/ja00161a064

Bendl J, Stourac J, Sebestova E, Vavra O, Musil M, Brezovsky J, Damborsky J (2016) HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. Nucleic Acids Res 44:W479–W487. https://doi.org/10.1093/nar/gkw416

Bramski J, Dick M, Pietruszka J, Classen T (2017) Probing the acetaldehyde-sensitivity of 2-deoxy-ribose-5-phosphate aldolase (DERA) leads to resistant variants. J Biotechnol 258:56–58. https://doi.org/10.1016/j.jbiotec.2017.03.024

Chambre D, Guérard-Hélaine C, Darii E, Mariage A, Petit J-L, Salanoubat M, de Berardinis V, Lemaire M, Hélaine V (2019) 2-Deoxyribose-5-phosphate aldolase, a remarkably tolerant aldolase towards nucleophile substrates. Chem Commun 55:7498–7501. https://doi.org/10.1039/C9CC03361K

Chen L, Dumas DP, Wong CH (1992) Deoxyribose 5-phosphate aldolase as a catalyst in asymmetric aldol condensation. J Am Chem Soc 114: 741–748. https://doi.org/10.1021/ja00028a050

Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity : all-atom structure validation for macromolecular crystallography. Acta Crystallogr Sect D Biol Crystallogr 66:12–21. https://doi.org/10.1107/S0907444909042073

Dick M, Hartmann R, Weiergräber OH, Bisterfeld C, Classen T, Schwarten M, Neudecker P, Willbold D, Pietruszka J (2016) Mechanism-based inhibition of an aldolase at high concentrations of its natural substrate acetaldehyde: structural insights and protective strategies. Chem Sci 7:4492–4502. https://doi.org/10.1039/C5SC04574F

Emsley P, Lohkamp B, Scott WG, Cowtan K (2010) Features and development of Coot. Acta Crystallogr Sect D Biol Crystallogr 66:486–501. https://doi.org/10.1107/S0907444910007493

Gibson D, Young L, Chuang R-Y, Venter J, Hutchison C III, Smith H (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343–345

Gijsen HJM, Wong C-H (1994) Unprecedented asymmetric aldol reactions with three aldehyde substrates catalyzed by 2-deoxyribose-5-phosphate aldolase. J Am Chem Soc 116:8422–8423. https://doi.org/10.1021/ja00097a082

Haridas M, Abdelraheem EMM, Hanefeld U (2018) 2-Deoxy-d-ribose-5-phosphate aldolase (DERA): applications and modifications. Appl Microbiol Biotechnol 102:9959–9971. https://doi.org/10.1007/s00253-018-9392-8

Heine A, Lerner RA, Barbas CF, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2001) Observation of covalent

intermediates in an enzyme mechanism at atomic resolution. Science 294(80):369–374. https://doi.org/10.1126/science.1063601

Heine A, Luz JG, Wong C-H, Wilson IA (2004) Analysis of the class I aldolase binding site architecture based on the crystal structure of 2-deoxyribose-5-phosphate aldolase at 0.99Å resolution. J Mol Biol 343:1019–1034. https://doi.org/10.1016/j.jmb.2004.08.066

Hernández K, Szekrenyi A, Clapés P (2018) Nucleophile promiscuity of natural and engineered aldolases. Chembiochem:1353–1358. https://doi.org/10.1002/cbic.201800135

Jennewein S, Schürmann M, Wolberg M, Hilker I, Luiten R, Wubbolts M, Mink D (2006) Directed evolution of an industrial biocatalyst: 2-deoxy-D-ribose 5-phosphate aldolase. Biotechnol J 1:537–548. https://doi.org/10.1002/biot.200600020

Jindal G, Slanska K, Kolev V, Damborsky J, Prokop Z, Warshel A (2019) Exploring the challenges of computational enzyme design by re-building the active site of a dehalogenase. Proc Natl Acad Sci 116: 389–394. https://doi.org/10.1073/pnas.1804979115

Johnson EA, Lin ECC (1987) *Klebsiella pneumoniae* 1,3-propanediol: NAD+ oxidoreductase. J Bacteriol 169:2050–2054

Kille S, Acevedo-Rocha CG, Parra LP, Zhang ZG, Opperman DJ, Reetz MT, Acevedo JP (2013) Reducing codon redundancy and screening effort of combinatorial protein libraries created by saturation mutagenesis. ACS Synth Biol 2:83–92. https://doi.org/10.1021/sb300037w

Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. Angew Chem Int Ed 52:5700–5725. https://doi.org/10.1002/anie.201204077

Liebschner D, Afonine PV, Moriarty NW, Poon BK, Sobolev OV, Terwilliger TC, Adams PD (2017) Polder maps: improving OMIT maps by excluding bulk solvent. Acta Crystallogr Sect D Struct Biol 73:148–157. https://doi.org/10.1107/S2059798316018210

Linder M (2012) Computational enzyme design: advances, hurdles and possible ways forward. Comput Struct Biotechnol J 2:e201209009. https://doi.org/10.5936/csbj.201209009

Mak WS, Siegel JB (2014) Computational enzyme design: transitioning from catalytic proteins to enzymes. Curr Opin Struct Biol 27:87–94. https://doi.org/10.1016/j.sbi.2014.05.010

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ (2007) Phaser crystallographic software. J Appl Crystallogr 40:658–674. https://doi.org/10.1107/S0021889807021206

Moriarty NW, Grosse-Kunstleve RW, Adams PD (2009) electronic Ligand Builder and Optimization Workbench ( eLBOW ): a tool for ligand coordinate and restraint generation. Acta Crystallogr Sect D Biol Crystallogr 65:1074–1080. https://doi.org/10.1107/S0907444909029436

Oslaj M, Cluzeau J, Orkic D, Kopitar G, Mrak P, Casar Z (2013) A highly productive, whole-cell DERA chemoenzymatic process for production of key lactonized side-chain intermediates in statin synthesis. PLoS One 8. https://doi.org/10.1371/journal.pone.0062250

Peränen J, Rikkonen M, Hyvönen M, Kääriäinen L (1996) T7 vectors with a modified T7 lac promoter for expression of proteins in *Escherichia coli*. Anal Biochem 236:371–373. https://doi.org/10.1006/abio.1996.0187

Sakuraba H, Tsuge H, Shimoya I, Kawakami R, Goda S, Kawarabayasi Y, Katunuma N, Ago H, Miyano M, Ohshima T (2003) The first crystal structure of Archaeal aldolase. Unique tetrameric structure of 2-deoxy-D-ribose-5-phosphate aldolase from the hyperthermophilic Archaea *Aeropyrum pernix*. J Biol Chem 278:10799–10806. https://doi.org/10.1074/jbc.M212449200

Schulte M, Petrović D, Neudecker P, Hartmann R, Pietruszka J, Willbold S, Willbold D, Panwalkar V (2018) Conformational sampling of the intrinsically disordered C-terminal tail of DERA is important for enzyme catalysis. ACS Catal 8:3971–3984. https://doi.org/10.1021/acscatal.7b04408

# Publication **II**

Biochemistry

pubs.acs.org/biochemistry Communication

# Xylonolactonase from *Caulobacter crescentus* Is a Mononuclear Nonheme Iron Hydrolase

Johan Pääkkönen, Leena Penttinen, Martina Andberg, Anu Koivula, Nina Hakulinen, Juha Rouvinen, and Janne Jänis*

Cite This: *Biochemistry* 2021, 60, 3046−3049 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** *Caulobacter crescentus* xylonolactonase (*Cc* XylC, EC 3.1.1.68) catalyzes an intramolecular ester bond hydrolysis over a nonenzymatic acid/base catalysis. *Cc* XylC is a member of the SMP30 protein family, whose members have previously been reported to be active in the presence of bivalent metal ions, such as $Ca^{2+}$, $Zn^{2+}$, and $Mg^{2+}$. By native mass spectrometry, we studied the binding of several bivalent metal ions to *Cc* XylC and observed that it binds only one of them, namely, the $Fe^{2+}$ cation, specifically and with a high affinity ($K_d$ = 0.5 $\mu$M), pointing out that *Cc* XylC is a mononuclear iron protein. We propose that bivalent metal cations also promote the reaction nonenzymatically by stabilizing a short-lived bicyclic intermediate on the lactone isomerization reaction. An analysis of the reaction kinetics showed that *Cc* XylC complexed with $Fe^{2+}$ can speed up the hydrolysis of D-xylono-1,4-lactone by 100-fold and that of D-glucono-1,5-lactone by 10-fold as compared to the nonenzymatic reaction. To our knowledge, this is the first discovery of a nonheme mononuclear iron-binding enzyme that catalyzes an ester bond hydrolysis reaction.

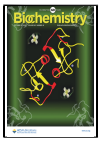
Metal cations are essential for the catalytic activity of several enzymes; thus, their accurate identification, binding affinity determination, and coordination characteristics are essential in the understanding of enzyme function.[1] High-resolution native mass spectrometry (MS) is a powerful method to characterize metal ion binding to folded proteins with a high accuracy.[2] When we used native MS to characterize the xylonolactonase from *Caulobacter crescentus* (*Cc* XylC), we observed unexpectedly that it binds only the $Fe^{2+}$ cation with a high affinity and specificity, suggesting that the previous understanding of the metal ion binding to this enzyme is inadequate.

*Cc* XylC uses D-xylonolactone as a substrate and produces D-xylonic acid.[3] D-Xylonolactone exists as two isomers, 1,4- and 1,5-lactones, which can interconvert via a short-lived bicyclic intermediate.[4,5] Therefore, it is difficult to estimate whether 1,4- or 1,5-lactone would be a preferable substrate for *Cc* XylC. However, the crystal structures of the homologous SMP30 protein have shown well-ordered electron densities for the six-membered ring ligands, suggesting that the binding of 1,5-lactone would be preferable also for *Cc* XylC.[6] Also, Jermyn has found that 1,5-lactone is a true substrate for the homologous gluconolactonase from *Pseudomonas fluorescens*.[4]

This lactonase-catalyzed reaction is the second step in the oxidative, nonphosphorylative D-xylose (Dahms or Weimberg) pathway in bacteria. The Dahms pathway can also be utilized for the production of several platform chemicals such as ethylene glycol, glycolic acid, lactic acid, and 1,4-butanediol starting from xylose-rich biomass fractions.[3] On the basis of the amino acid sequence homology, *Cc* XylC is a member of the senescence marker protein 30 (SMP30) protein family, which includes several aldonolactonases. The amino acid sequence search within the Protein Data Bank (PDB) results in a few enzyme structures, which possess homologous sequences. These include, for example, SMP30 gluconolactonases (31−34% identity), luciferin-regenerating enzyme (33% identity), *Xanthomonas campestris* gluconolactonase (30% identity), and *Linaria vulgaris* diisopropylfluorophosphatase (26% identity). All solved SMP30 family members share an overall tertiary structure consisting of a six-blade $\beta$-propeller with a central channel, where a bivalent metal ion is tri- or tetracoordinated by asparagine, aspartate, or glutamate side chains. $Ca^{2+}$, $Zn^{2+}$, and $Mg^{2+}$ metal ions have been used in the crystal structure refinements.[6−10] Thus, the previous studies suggest that this enzyme family is capable of binding different bivalent metal cations. To obtain further information about the metal ion binding of *Cc* XylC, we chose to use native MS, since it allows the metal binding stoichiometry, affinity, and specificity to be directly observed.

The mass spectra of *Cc* XylC were measured by using a high-resolution Fourier transform ion cyclotron resonance (FT-ICR) instrument (Bruker Solarix XR), equipped with an electrospray ionization (ESI) source. The elemental formula obtained from the amino acid sequence of *Cc* XylC (omitting the initial methionine residue) is $C_{1422}H_{2158}N_{378}O_{422}S_7$, and the corresponding theoretical most abundant isotopic mass is 31 524.76 Da. The mass spectrum of the denatured *Cc* XylC (Figure S1) gave the most abundant mass of 31 524.89 ± 0.10

**Received:** April 12, 2021
**Revised:** September 16, 2021
**Published:** October 11, 2021







https://doi.org/10.1021/acs.biochem.1c00249
*Biochemistry* 2021, 60, 3046−3049

Da (mean ± standard deviation), averaged over the observed charge state distribution, consistent with the theoretical value. In the denatured state, no metal ion binding to *Cc* XylC was observed, as expected. In contrast, when *Cc* XylC was measured in the native state, an additional signal was surprisingly observed at 31 577.49 Da in the deconvoluted mass spectrum. The mass difference of ∼53 Da corresponds to the binding of a single $Fe^{3+}$ ion (theoretical mass 31 577.67 Da, assuming a formal removal of three hydrogens upon iron ion binding). This was an unexpected observation, since no iron binding has been suggested for the other SMP30 family lactonases in any previous studies. To obtain a spectrum of the apoprotein, the protein sample was treated with ethylenediaminetetraacetic acid (EDTA) (at least a 20-fold molar excess) to chelate any metal ions before sample desalting. Following the EDTA treatment, only the 31 524.56 Da signal remained in the spectrum (Figure 1A).

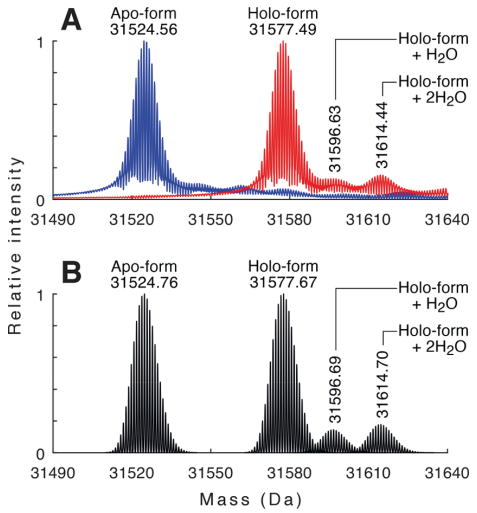



**Figure 1.** (A) Deconvoluted native mass spectra of *Cc* XylC in apo-form (blue) and holo-form (red). (B) Calculated mass spectra of the corresponding forms of *Cc* XylC. The most abundant isotopic masses are indicated.

To study whether the iron binding to *Cc* XylC was specific, the binding of six other metals was also tested. The initial metal binding tests were done by adding either $Mg^{2+}$, $Ca^{2+}$, $Fe^{2+}$, $Fe^{3+}$, $Co^{2+}$, $Ni^{2+}$, $Cu^{2+}$, or $Zn^{2+}$ at 10 μM to apo-*Cc* XylC (EDTA-treated) at 1 μM. All metal ions were added as their analytical grade chloride salts, dissolved in 10 mM ammonium acetate solution. Native MS measurements indicated that only $Fe^{2+}$ was bound strongly and specifically to *Cc* XylC, while the other metals did not show even weak binding, except $Cu^{2+}$, for which the native MS showed that up to four $Cu^{2+}$ ions were able to bind to the protein. The $Cu^{2+}$ binding most likely occurs through the four free cysteine residues and not to the active site of the enzyme.

Also, two minor signals corresponding to the additional binding of one and two water molecules to the iron-complexed holo-form of *Cc* XylC were observed (Figure 1A). These waters are likely coordinated to the metal ion and remain bound during the ionization process. Such water molecules are rarely detected in the gas phase (except in some small metal coordination complexes), due to their low enthalpy of dehydration, but have been previously detected, for example, for an *Escherichia coli* deaminase transition-state analogue complex.[11] As the bound iron ion was observed as $Fe^{3+}$, even though $Fe^{2+}$ was originally added to the solution, we also tested if $Fe^{3+}$ was capable of binding to the protein. Since no binding was observed with $Fe^{3+}$, it is clear that iron binds to *Cc* XylC only in the oxidation state +2 but undergoes a rapid oxidation to +3 in the electrospray process. A similar behavior has been observed earlier with the heme-binding proteins.[12] It was also observed that the iron oxidation during the electrospray process does not occur when water molecules are coordinated to the $Fe^{2+}$ cation (Figure 1A).

To determine the iron binding affinity of *Cc* XylC, 1 μM apoenzyme was titrated with $Fe^{2+}$ (up to 16 μM), and the fractional saturation was monitored. The fitting of the data to the specific, one-site binding model yielded a $K_d$ value of (5.0 ± 1.3) × $10^{-7}$ M and $B_{max}$ of 0.966 ± 0.009 with a 95% level of confidence (Figure 2).

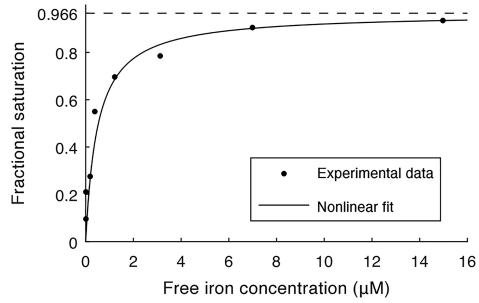



**Figure 2.** Titration of *Cc* XylC with $Fe^{2+}$.

The influence of $Fe^{2+}$ for enzymatic activity was also tested by following the lactone hydrolysis reaction (at the initial pH 6.9) with an ion trap mass spectrometer. Probably due to difficulties in the isolation of lactone isomers, only D-xylono-1,4-lactone and D-glucono-1,5-lactone were commercially available and used as substrates at the initial 0.25 mM concentration. Without an enzyme or metal present (non-enzymatic reaction), both xylono- and gluconolactones were hydrolyzed with the half-lives of 330 ± 40 and 45 ± 5 min (confidence intervals with a 95% level of confidence), respectively. The results are in agreement with the results of Jermyn,[4] who found that the nonenzymatic hydrolysis of 1,5-lactone is faster as compared to 1,4-lactone and suggested that 1,4-lactone is isomerized via a bicyclic intermediate to 1,5-lactone, which is then hydrolyzed. This isomerization reaction is reversible and fast.[4,5]

The addition of 10 μM $Fe^{2+}$ in the absence of enzyme slightly accelerated the hydrolysis of both substrates, suggesting that the bare metal ions can also promote the nonenzymatic hydrolysis reactions. Because the acceleration was greater for 1,4-lactone, this would suggest that metal ion could catalyze the isomerization reaction between 1,4-lactone and 1,5-lactone (Figure 3).

We then performed the hydrolysis reactions in the presence of 10 μM $Fe^{2+}$ and 0.5 μM *Cc* XylC, which resulted in a considerable increase in the hydrolysis rates for both lactones. The half-lives of xylono- and gluconolactones were reduced to 4.3 ± 0.3 and 4.8 ± 0.4 min, respectively, corresponding to the


https://doi.org/10.1021/acs.biochem.1c00249
*Biochemistry* 2021, 60, 3046−3049

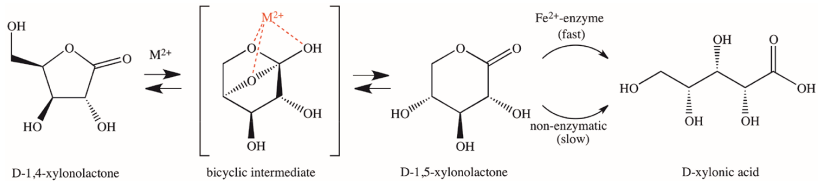



**Figure 3.** Suggested total reaction of D-xylonolactone hydrolysis.

**Table 1. Reaction Rate Constants and Half-Lives for *Cc* XylC-Catalyzed Hydrolysis of Xylono- and Gluconolactones**

| substrate (main isomer) | *Cc*XylC ($\mu$M) | $Fe^{2+}$ ($\mu$M) | rate constant[a,b] $k$ ($s^{-1}$) | half-life[b] $t_{1/2}$ (min) |
|---|---|---|---|---|
| D-xylono-1,4-lactone | 0 | 0 | $(3.5 \pm 0.5) \times 10^{-5}$ | 330 ± 40 |
| D-xylono-1,4-lactone | 0 | 10 | $(8.3 \pm 0.5) \times 10^{-5}$ | 138 ± 8 |
| D-xylono-1,4-lactone | 0.5 | 10 | $(2.7 \pm 0.2) \times 10^{-3}$ | 4.3 ± 0.3 |
| D-glucono-1,5-lactone | 0 | 0 | $(2.6 \pm 0.3) \times 10^{-4}$ | 45 ± 5 |
| D-glucono-1,5-lactone | 0 | 10 | $(4.0 \pm 0.6) \times 10^{-4}$ | 29 ± 4 |
| D-glucono-1,5-lactone | 0.5 | 10 | $(2.4 \pm 0.2) \times 10^{-3}$ | 4.8 ± 0.4 |

[a]Pseudo-first-order. [b]Confidence intervals with a 95% level of confidence.

estimated $k_{cat}/K_M$ values of $(5.4 \pm 0.9) \times 10^3$ and $(4.9 \pm 0.9) \times 10^3$ $M^{-1}$ $s^{-1}$. Since the values are of the same magnitude, it can be concluded that the enzyme catalyzes the hydrolysis of both lactones equally and that the enzyme is not specific to either lactone (xylono- or gluconolactone). The reaction rate over the nonenzymatic hydrolysis was larger for D-xylono-1,4-lactone (100-fold) than for D-glucono-1,5-lactone (10-fold). The apparent reaction rate constants (pseudo-first-order) and the reaction half-lives are summarized in Table 1.

Boer et al. have observed earlier[3] that $Ca^{2+}$ and $Zn^{2+}$ increase the rate of D-xylonolactone hydrolysis. Because we did not observe the binding of these metal cations to *Cc* XylC, they must also speed up the reaction nonenzymatically by binding to the above-mentioned bicyclic reaction intermediate. This interaction is less specific for different bivalent metal ions. If *Cc* XylC utilizes preferably 1,5-lactone as the substrate, the presence of metal ions hastens the formation of 1,5-lactone and consequently also the enzymatic hydrolysis (Figure 3).

Studying the enzyme-catalyzed hydrolysis of an intramolecular ester bond in lactones is challenging because lactones are also nonenzymatically hydrolyzed to sugar acids by utilizing acid or base catalysis in a water medium. There is also evidence in the earlier studies that metal ions, especially bivalent metal ions, are able to weakly catalyze the ester bond hydrolysis.[13,14] Therefore, careful experimental analyses are needed to dissect the actual roles of metal ions in catalysis. In this respect, the use of native mass spectrometry in analyzing the metal ion binding to *Cc* XylC has proven to be very essential. The results show that the enzyme is highly specific for $Fe^{2+}$, and the affinity to other bivalent metal cations is very low. This result suggests that *Cc* XylC is a mononuclear nonheme iron enzyme. To the best of our knowledge, the other known examples of mononuclear iron enzymes are all oxidases, typically utilizing molecular dioxygen.[15,16] In consequence, this study suggests that $Fe^{2+}$ may exist as a catalytic metal also in hydrolytic enzymes. Further studies are needed to clarify if $Fe^{2+}$ would be a catalytic metal ion also among other members of the SMP30 protein family.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.biochem.1c00249.

Experimental procedures, additional mass spectra, iron affinity determination, hydrolysis reaction progress curves (PDF)

### Accession Codes

*Caulobacter crescentus* xylonolactonase C; UniProtKB accession code: A0A0H3C6P8 (A0A0H3C6P8_CAUVN)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Janne Jänis** − *Department of Chemistry, University of Eastern Finland, FI-80101 Joensuu, Finland;* orcid.org/0000-0002-8446-4704; Email: janne.janis@uef.fi

### Authors

**Johan Pääkkönen** − *Department of Chemistry, University of Eastern Finland, FI-80101 Joensuu, Finland*

**Leena Penttinen** − *Department of Chemistry, University of Eastern Finland, FI-80101 Joensuu, Finland*

**Martina Andberg** − *VTT Technical Research Centre of Finland Ltd, FI-02044 VTT Espoo, Finland*

**Anu Koivula** − *VTT Technical Research Centre of Finland Ltd, FI-02044 VTT Espoo, Finland*

**Nina Hakulinen** − *Department of Chemistry, University of Eastern Finland, FI-80101 Joensuu, Finland;* orcid.org/0000-0003-4471-7188

**Juha Rouvinen** − *Department of Chemistry, University of Eastern Finland, FI-80101 Joensuu, Finland;* orcid.org/0000-0003-1843-5718

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.biochem.1c00249


### Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

### Notes

The authors declare no competing financial interest.



https://doi.org/10.1021/acs.biochem.1c00249
*Biochemistry* 2021, 60, 3046−3049

## ■ ACKNOWLEDGMENTS

This work received support from the Academy of Finland through the SA-ENGBIOCAT (Grant Nos. 287241 and 288677) and PENTOX (Grant No. 322610) projects. The FT-ICR facility is supported by Biocenter Finland (FIN-Struct), Biocenter Kuopio, the European Regional Development Fund (Grant No. A70135), and the European Union's Horizon 2020 Research and Innovation Programme (EU FT-ICR MS project; Grant No. 731077). We thank A. Kiema for technical assistance (VTT).

## ■ ABBREVIATIONS

*Cc* XylC, *Caulobacter crescentus* xylonolactonase C; ESI, electrospray ionization; FT-ICR, Fourier transform ion cyclotron resonance; SMP30, senescence marker protein 30.

## ■ REFERENCES

(1) Dudev, T.; Lim, C. Competition among metal ions for protein binding sites: determinants of metal ion selectivity in proteins. *Chem. Rev.* **2014**, *114*, 538−556.

(2) Boeri Erba, E.; Petosa, C. The emerging role of native mass spectrometry in characterizing the structure and dynamics of macromolecular complexes. *Protein Sci.* **2015**, *24*, 1176−1192.

(3) Boer, H.; Andberg, M.; Pylkkänen, R.; Maaheimo, H.; Koivula, A. In vitro reconstitution and characterisation of the oxidative D-xylose pathway for production of organic acids and alcohols. *AMB Express* **2019**, *9*, 48.

(4) Jermyn, M. A. Studies on the glucono-delta-lactonase of *Pseudomonas fluorescens*. *Biochim. Biophys. Acta* **1960**, *37*, 78−92.

(5) Bierenstiel, M.; Schlaf, M. δ-Galactonolactone: synthesis, isolation, and comparative structure and stability analysis of an elusive sugar derivative. *Eur. J. Org. Chem.* **2004**, *2004*, 1474−1481.

(6) Aizawa, S.; Senda, M.; Harada, A.; Maruyama, N.; Ishida, T.; Aigaki, T.; Ishigami, A.; Senda, T. Structural basis of the γ-lactone-ring formation in ascorbic acid biosynthesis by the senescence marker protein-30/gluconolactonase. *PLoS One* **2013**, *8*, No. E53706.

(7) Chakraborti, S.; Bahnson, B. J. Crystal structure of human senescence marker protein 30: insights linking structural, enzymatic, and physiological functions. *Biochemistry* **2010**, *49*, 3436−3444.

(8) Yamashita, K.; Pan, D.; Okuda, T.; Sugahara, M.; Kodan, A.; Yamaguchi, T.; Murai, T.; Gomi, K.; Kajiyama, N.; Mizohata, E.; Suzuki, M.; Nango, E.; Tono, K.; Joti, Y.; Kameshima, T.; Park, J.; Song, C.; Hatsui, T.; Yabashi, M.; Iwata, S.; Kato, H.; Ago, H.; Yamamoto, M.; Nakatsu, T. An isomorphous replacement method for efficient *de novo* phasing for serial femtosecond crystallography. *Sci. Rep.* **2015**, *5*, 14017.

(9) Chen, C.-N.; Chin, K.-H.; Wang, A. H.-J.; Chou, S.-H. The first crystal structure of gluconolactonase important in the glucose secondary metabolic pathways. *J. Mol. Biol.* **2008**, *384*, 604−614.

(10) Koepke, J.; Scharff, E. I.; Lücke, C.; Rüterjans, H.; Fritzsch, G. Statistical analysis of crystallographic data obtained from squid ganglion DFPase at 0.85 Å resolution. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2003**, *D59*, 1744−1754.

(11) Borchers, C. H.; Marquez, V. E.; Schroeder, G. K.; Short, S. A.; Snider, M. J.; Speir, J. P.; Wolfenden, R. Fourier transform ion cyclotron resonance MS reveals the presence of a water molecule in an enzyme transition-state analogue complex. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15341−15345.

(12) He, F.; Hendrickson, C. L.; Marshall, A. G. Unequivocal determination of metal atom oxidation state in naked heme proteins: Fe(III)myoglobin, Fe(III)cytochrome c, Fe(III)cytochrome b5, and Fe(III)cytochrome b5 L47R. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 120−126.

(13) Fife, T. H.; Przystas, T. J. Divalent metal ion catalysis in the hydrolysis of esters of picolinic acid. Metal ion promoted hydroxide ion and water catalyzed reactions. *J. Am. Chem. Soc.* **1985**, *107*, 1041−1047.

(14) Fife, T. H.; Pujari, M. P. Metal ion catalysis in the hydrolysis of esters of 2-hydroxy-1,10-phenanthroline: the effects of metal ions on intramolecular carboxyl group participation. *Bioorg. Chem.* **2000**, *28*, 357−373.

(15) Que, L., Jr.; Ho, R. Y. N. Dioxygen activation by enzymes with mononuclear non-heme iron active sites. *Chem. Rev.* **1996**, *96*, 2607−2624.

(16) Bruijnincx, P. C. A.; van Koten, G.; Klein Gebbink, R. J. M. Mononuclear non-heme iron enzymes with the 2-His-1-carboxylate facial triad: recent developments in enzymology and modeling studies. *Chem. Soc. Rev.* **2008**, *37*, 2716−2744.



https://doi.org/10.1021/acs.biochem.1c00249
*Biochemistry* 2021, 60, 3046−3049

Publication **III**

J. Pääkkönen, N. Hakulinen, M. Andberg,
A. Koivula, and J. Rouvinen

"Three-dimensional structure of xylonolactonase
from *Caulobacter crescentus*: a mononuclear iron enzyme
of the 6-bladed β-propeller hydrolase family"

*Protein Science*,

doi: 10.1002/pro.4229, 2021.

Publication **IV**

J. Pääkkönen, J. Jänis, and J. Rouvinen
"Simulation of binding in protein studies"
Submitted for publication.

**JOHAN PÄÄKKÖNEN**

Biomass, as starting material for production of various chemicals, is a promising and sustainable alternative to fossil sources. Nature manipulates carbohydrates in biomass using metabolic pathways in which most reactions are catalysed by enzymes. In this thesis, properties of two such enzymes are researched by mass spectrometry and X-ray crystallography. The new discoveries are small steps towards the eventual implementation of these enzymes in industrial-scale syntheses.

UNIVERSITY OF
EASTERN FINLAND

**uef.fi**

**PUBLICATIONS OF**
**THE UNIVERSITY OF EASTERN FINLAND**
Dissertations in Forestry and Natural Sciences

ISBN 978-952-61-4386-6
ISSN 1798-5668