Towards the structure-informed engineering of enzymes in the avenacin biosynthesis pathway

Hans Emile Pfalzgraf

A thesis submitted for the degree of
Doctor of Philosophy to the
School of Biological Sciences,
University of East Anglia

Contents

Abstract

Glycosylated triterpenes represent numerous and diverse plant natural products, but difficult production has limited their applications in health, food and industry.

This thesis describes the structural characterisation of three enzymes from the biosynthetic pathway of avenacin, an antifungal glycosylated triterpene from oat, to enable their rational engineering.

Avena strigosa arabinosyltransferase (AsAAT1) and transglucosidase (AsTG1), involved in the glycosylation of avenacin, were expressed in *Escherichia coli* and purified, but did not crystallise. Several deletion constructs proved insoluble, so molecular models were used to rationalise both the specificity of three AsAAT1 mutants for various sugar donors and the switch from glucosyl hydrolase to transglucosidase activity in AsTG1, which was suggested by a multiple sequence alignment. Molecular dynamics simulations of AsTG1 confirmed its ability to discriminate between analogous substrates.

The membrane-bound *A. strigosa* β-amyrin synthase (AsbAS1), which forms the triterpene scaffold of avenacin, was expressed in *E. coli*. Attempts to purify and crystallise it only led to the high-resolution structure of a contaminant, HPII catalase. AsbAS1 mutants were designed, inspired by a soluble homologue, to simplify this process. With no solubilised protein observed in *E. coli*, AsbAS1 was expressed in the yeast *Pichia pastoris* instead, which resulted in active protein. A homologue, *Euphorbia tirucalli* β-amyrin synthase (EtAS), was expressed in an active form in *E. coli*. These expression methods could be used to produce two different β-amyrin synthases and attempt to obtain the first crystal structure of a plant oxidosqualene cyclase. A multiple sequence alignment and models of other homologues generated with AlphaFold2 enabled the design of four AsbAS1 mutants that may have altered product specificity.

This work shows structural information for three enzymes in the avenacin biosynthesis pathway, leading to the rationalisation of the effect from various amino acids. This can now be tested by expressing mutants using the methods described.

Acknowledgements

This thesis has been possible because of the collaboration and support of many individuals. I would like to express by deepest thanks to Prof Andrew Hemmings for being such an excellent supervisor, to my wife Kira for her endless care and to my parents for the decades of support.

I would like to thank Melissa Salmon, Raquel Fab Rodríguez, Caroline Laurendon, Isabella Maria Acquistapace, Marcus Edwards and the rest of lab 2.30 for teaching me how to make magic happen in the lab and for such a supportive and collaborative environment. I am grateful for the excellent support I received from the NRPDTP team, UEA Student Services, the UEA Medical Services and the UEA Events team.

The lab work presented here was funded by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership grant number BB/M011216/1. It includes contributions from my talented students Rokas Kubilinskas on AsbAS1 solubilisation and Rhys Stanley on AsTG1 and AsAAT1. I would like to thank Andrew Crombie for the GC-MS analysis and Elizabeth Gray for the LC-MS. I am grateful for Yinghong Gu's help with the work on membrane proteins. Outside of UEA, I would like to thank Anastasia Orme, Thomas Louveau, Shingo Kikuchi and the rest of the Osbourn group at JIC for their kind help with the project. I would also like to thank Ralf Flaig for setting up the automated data collection for the apo-GtfC$_{100\text{-}23}$ structure.

Finally, I would like to thank all my housemates and the crew from Sanctuary Sound for being such reliable friends.

Abbreviations

3C	Human rhinovirus (HRV) 3C protease
Å	Ångström (0.1 nm)
Ara	L-Arabinopyranose
AsAAT1	*Avena strigosa* arabinosyltransferase (UGT99D1)
AsbAS1	*A. strigosa* β-amyrin synthase (SAD1)
AsTG1	*A. strigosa* transglucosidase 1 (SAD3/AsGH1)
BDA	Bis-deglucosyl avenacin A1
CBC	Chair-boat-chair
CCC	Chair-chair-chair
cM	centimorgan
CV	Column volumes
CYP	Cytochrome P450
Da	Dalton (1 g/mol)
DDM	n-dodecyl-β-D-maltoside
dH_2O	Deionised water (by reverse osmosis)
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EtAS	*Euphorbia tirucalli* β-amyrin synthase
GC-MS	Gas Chromatography-Mass Spectrometry
GH	Glycosyl hydrolase
Glc	D-Glucopyranose
GlcNAc	N-Acetylglucosamine (D-(acetylamino)-2-deoxy-glucopyranose)
GT	Glycosyltransferase
IEC	Ion-exchange chromatography
IMAC	Immobilised metal ion affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
kb	Kilobase (1,000 nucleotide bases)
kDa	Kilodalton (1,000 g/mol)
LC-MS	Liquid chromatography mass spectrometry
MDA	Mono-deglucosyl avenacin A1
McOSC	*Methylococcus capsulatus* oxidosqualene cyclase

MD	Molecular dynamics
MR	Molecular replacement
MW	Molecular weight
m/z	Mass-to-charge ratio of ion
OS	(*S*)-2,3-oxidosqualene
OSC	Oxidosqualene cyclase
NTA	Nitrilotriacetic acid
PDB	Protein Data Bank
PSPG	Plant secondary product glycosyltransferase
rmsd	Root-mean-square deviation
RNA	Ribonucleic acid
SEC	Size-exclusion chromatography
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SPC	Simple point charge
r.t	Room temperature (20 – 25 °C)
TB	Terrific broth
TG	Transglycosidase
Tris	Tris(hydroxymethyl)aminomethane
UDP	Uridine diphosphate
UGT	UDP-dependent glycosyltransferase
WT	Wild type

Chapter 1: Introduction

Small molecules have been invaluable as drugs and food additives. Unfortunately, chemical synthesis limits their production and diversity,[1, 2] besides suffering from an unfavourable public perception on the grounds of health[3] and sustainability.[4] Natural products offer a trove of varied and versatile alternatives,[1] produced by organisms to increase their survival and competitiveness. These compounds evolved to be ideal for interacting with biological targets.[5] Plants use molecules that attract pollinators or seed-dispersing agents, while making others as defence against herbivores or microbes.[6] A better understanding of natural product biosynthesis would enable an increase in their yield and diversity, whether through metabolic engineering,[7] biocatalysis[8] or the use of heterologous hosts.[9] Additionally, it could allow the tailoring of existing pathways for crop improvement. For example, a plant pathogen's resistance to its host's defence could be overcome by editing the genes of key enzymes in the plant to change the final product.

1.1 Cereals and their diseases

Grasses refer to plants in the *Gramineae* family, part of the monocot group of flowering plants. They are comprised of around 10,000 species that are found all over the world across every kind of habitat, thanks to their ability to grow on many soil types, compete with other plants and survive high levels of predation. Cereals are grasses that are cultivated for food, either directly or through feeding animals, providing humanity with the majority its carbohydrates and covering 70% of the world's croplands.[10] The largest production of cereals comes from maize, rice, wheat, barley, sorghum, millet and oats.[11]

Unfortunately, cereals are susceptible to fungal diseases such as leaf scald, eyespot, ergot or take-all. The latter is caused by the soil-borne pathogen *Gaeumannomyces graminis* var. *tritici* and it mainly affects wheat and barley,[12] causing blackened roots, bleached heads and stunted growth.[13] It is the most damaging wheat root disease worldwide, though its extent has only been reported for the United Kingdom, where it is estimated to affect half of wheat crops, causing 5 to 20% annual yield loss[14] and costing farmers up to £60 million a year in 2006.[15] Take-all worsens when wheat is grown 3 to 5 years in a row so, for lack of resistant wheat cultivars, crop rotation is the main control measure. As for chemical control, it is unreliable for take-all and may give rise to antifungal resistance.[14]

Oats (*Avena spp.*) however, are not generally susceptible to take-all. The determinant of their resistance was identified through a forward genetic screen, as mutants that did not produce the fluorescent natural product avenacin in their roots were susceptible to take-all.[16] The fungicidal activity of this triterpene saponin is thought to result from its ability to insert into the membrane and interact with sterol molecules to form pores, leading to cell lysis.[17]

1.2 Triterpenes and saponins

1.2.1 Functions and applications

Triterpenes may be the largest group of natural products, with over 20,000 varied and often complex compounds reported,[18] mostly from plant, though many were found in bacteria and some were even discovered in marine animals.[19] It is expected that there remains a large undiscovered reservoir, as they are often present in very low concentrations and within complex mixtures.[20] Simple triterpenes make up waxes and membranes on the stems and leaves of plants, but some also act as signalling molecules, like lupeol, which regulates the development of root nodules. Saponins are glycosylated triterpenes or steroids, some of which are known to protect against pests and pathogens. For example, the triterpenoid glycosides avenacin and hederagenin cellobioside confer resistance to take-all disease and to flea beetle, respectively.[21] Though saponins have traditionally been considered secondary metabolites, i.e., not necessary for survival, there is growing evidence of their role in plant development.[22]

Triterpenes and saponins have applications in industrial biotechnology (e.g., in the anaerobic digestion of waste),[23] in the food industry (for example as additives) and in the health sector, both for cosmetics and pharmaceuticals.[18] There are, for example, triterpene hypocholesterolemic, antifungal or phytotoxic drugs.[24] Some saponins are the active ingredients in folk medicines (based on liquorice or ginseng), or provide the sought-after soap-like properties of certain plant extracts.[25, 26] They make good foaming agents and could be used as preservatives or emulsifiers. They can modify flavour or remove cholesterol from dairy products.[25, 27] Some are used as fish poison[28, 29] while others can deliver therapeutics.[30] The presence of saponins in food was originally considered undesirable as they have a bitter taste and were considered anti-nutrient because they inhibited protein digestion, reduced intestinal permeability and showed hemolytic activity. Consequently, much early research focused on their removal. However, there has been

growing evidence of their health benefits such as lowering cholesterol and preventing cancer, as well as their therapeutic applications because of anti-inflammatory, immunomodulatory, antibacterial, antifungal, antiviral or antiparasitic activity.[31–34] The commercial applications of triterpenes are however limited by the difficulty of their production. They remain challenging targets to synthesize chemically, not least because of their numerous stereocenters, while purification from natural sources is limited by their low abundance and a plethora of minor chemical variations.[20] Progress is also hampered by a lack of knowledge about their biosynthesis.[25] The pathway for avenacin, however, has been mostly elucidated.[35]

1.2.2 Biosynthesis

In plants, triterpenes originate from the mevalonate pathway, which generates the two molecules of isopentenyl diphosphate and one molecule of isopentenyl diphosphate that are used by the enzyme farnesyl diphosphate synthase to produce a molecule of farnesyl pyrophosphate (FPP, Fig. 1.1). Squalene synthase then acts on two copies of FPP to produce squalene, which is subsequently oxidised by squalene epoxidase to create 2,3-oxidosqualene (OS), the substrate of oxidosqualene cyclases (OSCs, covered in section 1.2.2.2).[36]

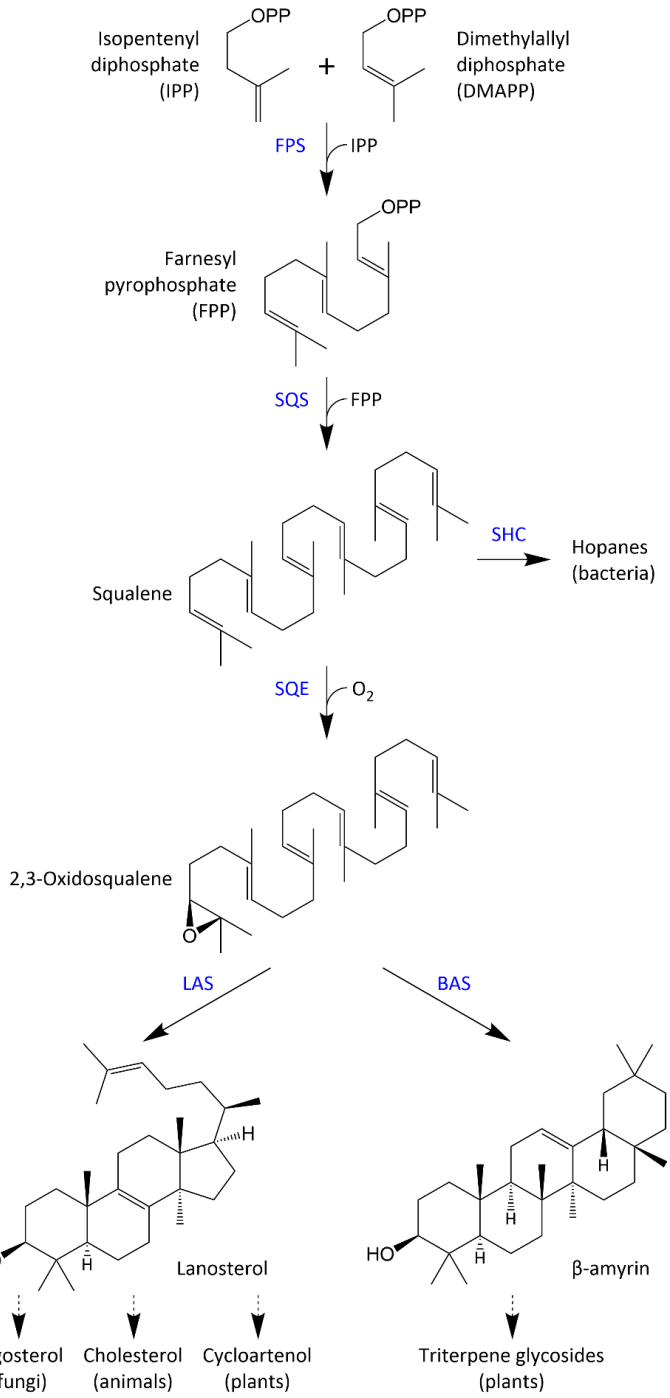


Figure 1.1: Biosynthetic pathway for triterpenes and sterols. The products of the mevalonate pathway (two molecules of IPP and one of DMAPP) are used by farnesyl pyrophosphate synthase (FPS) to produce FPP, two molecules of which are consumed by squalene synthase (SQS) to form squalene, the substrate of bacterial squalene-hopene cyclase (SHC). In eukaryotes, squalene is oxidised by squalene epoxidase (SQE) to generate 2,3-oxidosqualene, the last common precursor of sterols, such as the product of lanosterol synthase (LAS), and non-sterol triterpenes such as β-amyrin made by β-amyrin synthase (BAS). Figure adapted from Thimmappa *et al*[18] using ChemDraw 21.[37]

1.2.2.1 Avenacin

The first committed step in avenacin biosynthesis is the formation of β-amyrin by the OSC called β-amyrin synthase (Fig. 1.2). In *Avena strigosa*, this enzyme (AsbAS1) is encoded by a gene that is clustered with most of those required for the pathway, such as the ones coding for the four cytochrome P450 enzymes (CYPs) that oxidise the resulting triterpene scaffold. This clustering may allow co-adaptation and facilitate the co-regulation of the genes at the chromatin level, helping restrict expression of the pathway enzymes to the epidermal cell layer of the root tip.[35, 38] The biosynthetic gene cluster also contains the two UDP-dependent glycosyltransferases (UGTs, covered in section 1.2.2.3), arabinosyltransferase (AsAAT1) and AsUGT91G16, that add the arabinose and the first glucose group of avenacin, respectively. The resulting avenacin precursor is transported from the cytosol to the vacuole, where the third glycosylation is performed by AsTG1, a glycosyl hydrolase (GH, covered in section 1.2.2.4). The last step, acylation with the fluorescent *N*-methylanthranilate group, happens in the vacuole as well. This spatial segregation is required to protect oat cells from the phytotoxic effects of avenacin and its precursor.[35]

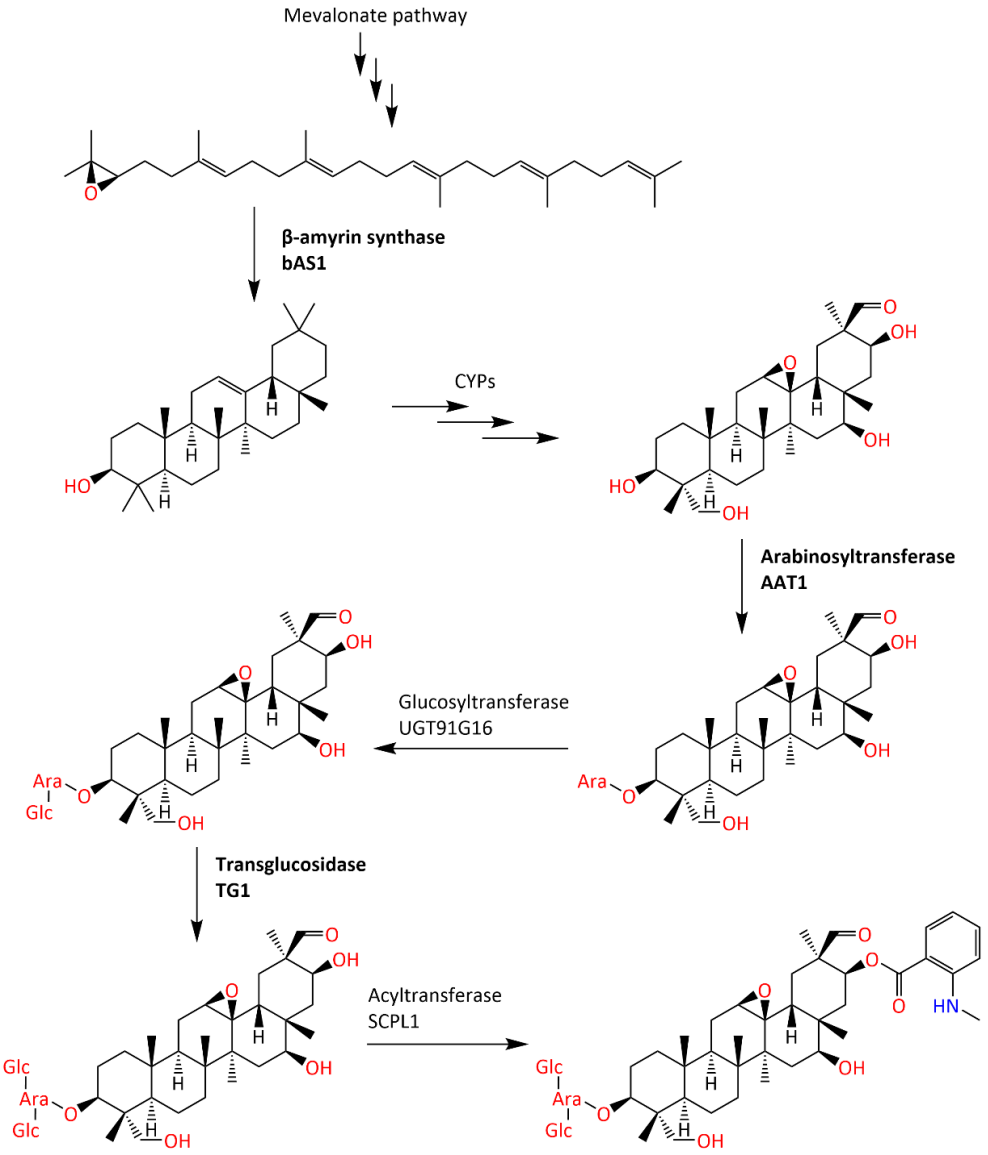


Figure 1.2: Biosynthetic pathway for avenacin A-1. In *A. strigosa*, β-amyrin is formed by AsbAS1 (also known as SAD1) and subsequently oxidised by cytochromes P450 (CYPs), including SAD2 and SAD6. The arabinose and first glucose group are added using UDP-sugar donors by AsAAT1 and AsUGT91G16, respectively, while the last glucose group is transferred from an unknown acyl sugar donor by AsTG1 (SAD3). The pathway ends with acylation, which is catalysed by AsSCPL1 (SAD7) using *N*-methylanthranilate glucoside as a second substrate. The three enzymes written in bold are the targets covered in this thesis. Figure prepared using ChemDraw 21.[37]

AsAAT1 and AsTG1 were chosen as targets because of their unusual glycosylation activity. Being soluble, these enzymes would potentially be low-hanging fruits.[39, 40] AsbAS1, on the other hand, is membrane-bound like most OSCs, making it a high-risk, high-reward target.[18]

1.2.2.2 Oxidosqualene cyclases

OSCs are class II terpene synthases, as they initiate cyclisation by protonation and not by removal of pyrophosphate. They are able to produce a wide range of very diverse compounds from the same linear substrate,[18] nearing a hundred different identified products.[41] In excess of 80 OSCs have been functionally characterized, a third of which make sterols such as the steroid precursors cycloartenol (in plants) and lanosterol (in animals and fungi). Two other common OSC products are the triterpenes lupeol and β-amyrin (Fig. 1.3). Many OSCs make a number of side products, e.g., BARS1 produces baruol 90% of the time, but also makes small amounts of 22 additional products. This is probably accidental and suggests that OSCs may have evolved to actually prevent certain cyclisations, in a type of negative catalysis.[18]

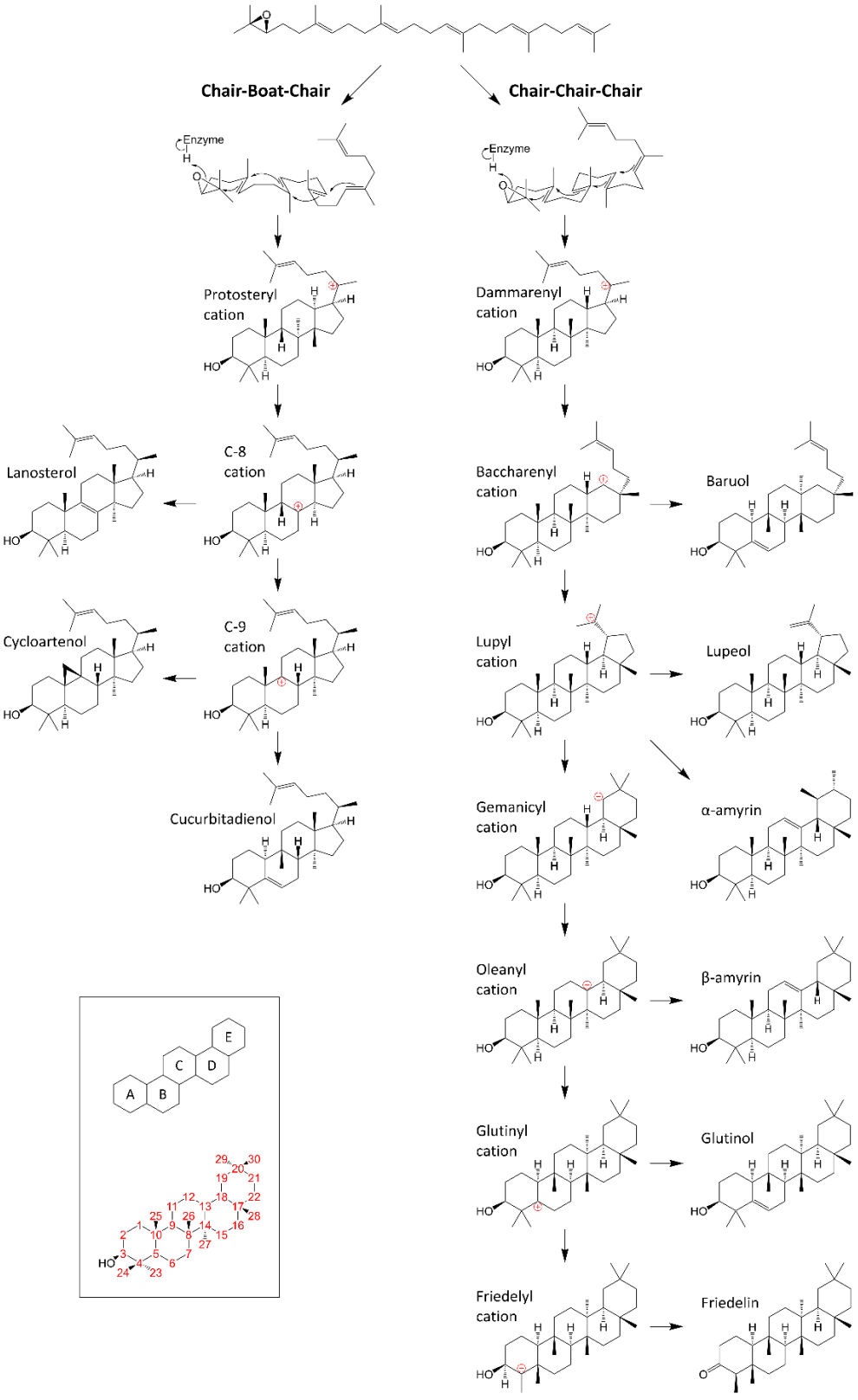


Figure 1.3: Products of oxidosqualene cyclisation. A multitude of products can be obtained through rearrangement of the cation, mostly by methyl and hydride 1,2-shifts. Carbon atoms and rings are labelled as for β-amyrin in the bottom left frame. Figure adapted from Thimmappa *et al*[18] using ChemDraw 21.[37]

Sterols like lanosterol and parkeol are based on the protosteryl cation, which forms when OS assumes a chair-boat-chair (CBC) conformation (Fig. 1.3). On the other hand, a chair-chair-chair (CCC) conformation leads to the dammarenyl cation, which results in triterpenes like lupeol and β-amyrin. This catalysis is achieved with the following steps: substrate folding, epoxide protonation, cyclisation (as well as a potential rearrangement of the cation) and termination, caused either by deprotonation of the cation or by hydroxylation through water capture.[18] Product specificity is achieved by constraining the substrate to the correct conformation, by stabilizing carbocations and by preventing deprotonation or termination by the solvent.[24] For example, the protosteryl cation can form cycloartenol, lanosterol or parkeol depending on which proton is abstracted by their respective OSC.[41] The enzyme is thought to play less of a role in methyl and hydride shifts.[24]

Unfortunately, no molecular or crystals structure of a plant OSC is yet available. Only a bacterial[42] and a human[43] homologue have published structures: squalene-hopene cyclase (SHC) and lanosterol synthase (LAS), respectively. Though they only share around 25% sequence identity, their architectures are very similar: they have two αα barrel domains, as well as a membrane-insertion helix (Fig. 1.4).[18] This anchors the entrance of a non-polar channel into the membrane, allowing the hydrophobic substrates to be recruited from the membrane and channelled to the active site, past a mobile constriction site.[42, 43] Because oxidosqualene cyclisation is highly exothermic (of the order of 200 kJ/mol for squalene to hopene), the released energy would destabilise a typical protein, for which stabilisation energies are much lower (of the order of 20-60 kJ/mol relative to the unfolded state).[44] To accommodate the reaction, OSCs are stabilised by QW-motifs and connected surface α-helices.[42] In SHC and LAS, techniques such as alanine-scanning mutagenesis have allowed identification of residues that are key for cyclisation, ring formation and carbocation stabilisation. Several OSCs from plants have been characterised: for example, cucurbitadienol synthase from *Cucurbita pepo*, α-amyrin synthase from *Olea europaea* and β-amyrin synthase from *Euphorbia tirucalli* (EtAS).[18] However, the mechanistic details of the formation of these different products are still poorly understood.[45] Unlocking this knowledge could allow access to a variety of otherwise intractable triterpene scaffolds.

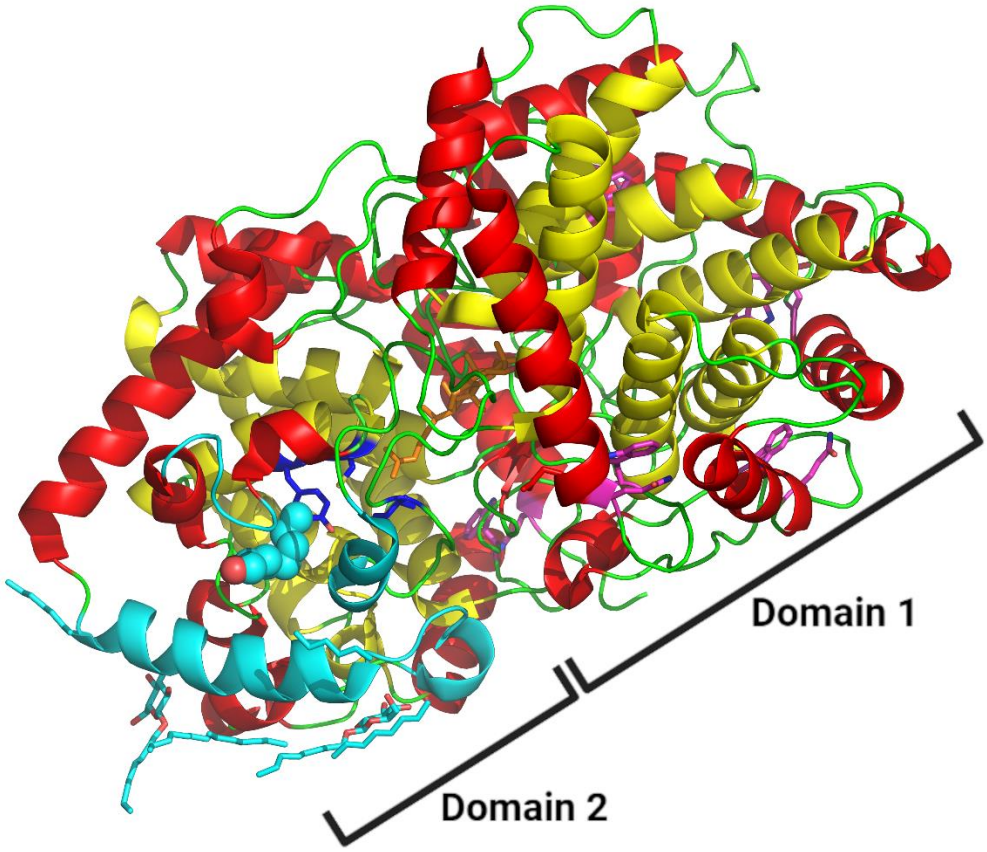


Figure 1.4: Crystal structure of human oxidosqualene cyclase. Each domain is an αα barrel with the outer α-helices shown in red, the inner α-helices in yellow and the loops in green. The membrane-insertion site is in cyan cartoon, surrounded by lipids (cyan sticks). The entrance of the substrate access channel can be seen by an inserted lipid fragment (cyan spheres) below the channel constriction site (dark blue) which is the entrance to the central active site occupied by lanosterol (orange sticks). The glutamine and tryptophan residues of the five QW-motifs are shown as magenta sticks. Figure prepared using PyMOL[46] and Biorender.com using PDB entry 1W6K.[43]

1.2.2.3 UDP-dependent glycosyltransferases

Triterpenes are often glycosylated on their alcohol or carboxyl groups to make saponins, which are more hydrophilic and require this glycosylation for their bioactivity.[25] This usually consists of chains of sugars such as glucose, arabinose, galactose, xylose, rhamnose and glucuronic acid.[18] They are often attached to the C-3 position, as in avenacin, or the C-28 position[25] (see Fig. 1.3 for numbering scheme) by enzymes from the Carbohydrate-Active Enzyme (CAZy) Glycosyltransferase (GT) 1 family.[39, 47] The chains are thought to be made through successive addition by family 1 UGTs, using the appropriate UDP-sugar as a donor. In plants, 12 UGTs are known to glycosylate triterpenes and they

could be used to alter the properties of known compounds,[18] such as bioactivity, reactivity and solubility.[39] Glycosylation also affects cellular localisation, usually because of recognition by active transporters.[25, 48, 49] Even the nature of the sugar can dramatically affect its bioactivity.[50] Thanks to a conserved Plant Secondary Product GT (PSPG) motif,[39] UGTs tend to be much more specific for their sugar donor than for their sugar acceptors, at least *in vitro.*[51] *In planta,* this would be more regulated thanks to compartmentalisation, or perhaps by protein-protein interactions and metabolite channelling.[35] While a UGT is typically promiscuous in terms of sugar acceptor, it tends to show conserved regiospecificity across similar compounds. Most characterized UGTs use the sugar donor UDP-α-D-glucose (UDP-Glc). This is the case for AsUGT91G16 from the avenacin biosynthesis pathway,[40] and for *Medicago truncatula* UGT71G1, the crystal structure of which is available.[52] However, some UGTs transfer the sugar moiety of UDP-activated mannose, xylose, galactose or glucuronic acid. Their sugar donor specificity remains impervious to sequence-based prediction,[53] which is unsurprising considering a single amino acid substitution can change the sugar donor specificity of a UGT.[39, 54] Access to a range of structures would help the endeavour, but structure determination has been hampered by challenging expression, purification and crystallization.[55] Still, over the last 15 years, many plant UGTs have been functionally characterised, 21 of which are able to glycosylate triterpenoids, with the vast majority of these transferring D-glucose. More recently, two triterpenoid arabinosyltransferases, the firsts of their kind, were discovered. AsAAT1, from the avenacin biosynthesis pathway, and GmSSAT1, which is involved in the biosynthesis of soyasaponin Ab (Fig. 1.5).[39] Though two amino acid positions and substitutions have been shown to play a major role in the change of specificity, the structural basis for the effect remains unknown.

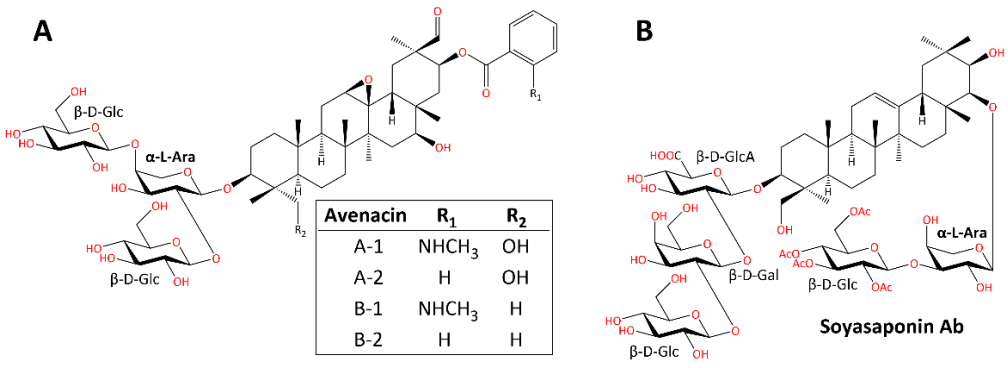


Figure 1.5: Chemical structures of saponins. (A) Avenacins. (**B)** Soyasaponin Ab, another saponin that contains an arabinose group. Figure adapted from Louveau *et al*[39] using ChemDraw 21.[37]

1.2.2.4 Glycosyl hydrolases

GHs can be classified either by their EC number (EC 3.2.1.x) to reflect the reaction they catalyse, or more usefully by the CAZy classification as glycoside hydrolase,[56] of which there are over 100 families that reflect mechanistic features and can provide more sequence-based prediction of structure and specificity.[54] GH families can be grouped into clans (such as GH-A) when there is evidence of common ancestry, e.g., they share their tertiary structure, catalytic residues and catalytic mechanism. There are two general mechanisms for GHs: inversion or retention of the anomeric configuration, with the latter taking two steps because of the formation of a glycosyl-enzyme intermediate, which can allow either hydrolysis or transglycosylation to occur (Fig. 1.6).[35] For example, the transglucosidase AsTG1 adds the second glucose group in avenacin biosynthesis.[40] GHs that use an acceptor aglycone are rare, but have arisen several times among GHs, suggesting a subtle structural change that allows them to catalyse transglycosylation using a structure that is closer to GHs than to transglycosidases (TGs) from other families. With no structures of TGs from this GH family, this structural change remains to be elucidated. Furthermore, new structural data on these unusual TGs could inform a refinement of our current classification, especially at the subfamily level, which is much needed because of an explosion of sequence data and increasing awareness of the importance of glycosylation in biology.[54] Making the classification more robust would improve its predictive power.

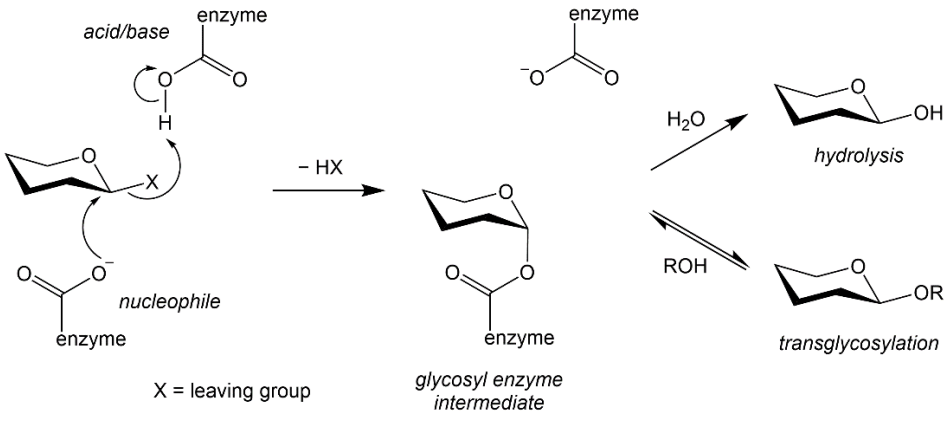


Figure 1.6: Generalized mechanism of a transglycosidase. Enzymatic cleavage of a substrate through a classical Koshland retaining mechanism results in formation of a glycosyl enzyme intermediate. This can partition to react with either water to cause hydrolysis (glycoside hydrolase activity) or to an alternative acceptor, often a sugar, to cause transglycosylation (transglycosidase activity). Figure adapted from cazypedia.org using ChemDraw 21.[37]

1.3 Protein engineering

Enzymes are very attractive catalysts, both for research and industry, because they can allow difficult reactions to be performed quickly, in water, with high substrate specificity and exquisite enantio- or stereoselectivity.[57] While synthetic chemistry is a powerful tool, some reactions are more suited to biocatalysts, e.g., if regiospecificity is challenging, if the reactant or products are labile or if side-products are particularly problematic. Biocatalysts have led to successful industrial-scale production of fructose, acrylamide, aspartame and 6-aminopenicillic acid. With numerous natural products being discovered and their properties studied, biocatalysts that synthesise them and their semi-synthetic analogues will be in increasing demand. Unfortunately, the natural function and requirements of an enzyme are usually very different from what would be ideal for scientists or engineers.[58] Indeed, one may want to optimise its catalytic properties (e.g., rate), its molecular recognition (e.g., selectivity or promiscuity) or its biophysical properties (e.g., thermostability, longevity, co-solvent tolerance).[57–59] So far, optimisation has most often been achieved through directed evolution. This consists in generating a library of variants that are screened for the desired property. Originally, the most common methods were random mutagenesis (e.g., through error-prone PCR) and recombination that produces chimeric mutants (e.g., through DNA shuffling), shown in Fig. 1.7.[57, 60] The former suffers from a high proportion of inactive enzymes while the latter lacks the exploration of original substitutions. Additionally, even with libraries of millions of enzymes, only a small fraction of the possible sequences is sampled. So, rather than overburdening the screening process with ever larger libraries that are unavoidably skewed by the diversity-generating method, researchers are now trying to generate libraries that are smaller but of higher quality. This approach is generally referred to as semi-rational, smart or knowledge-based.[57]

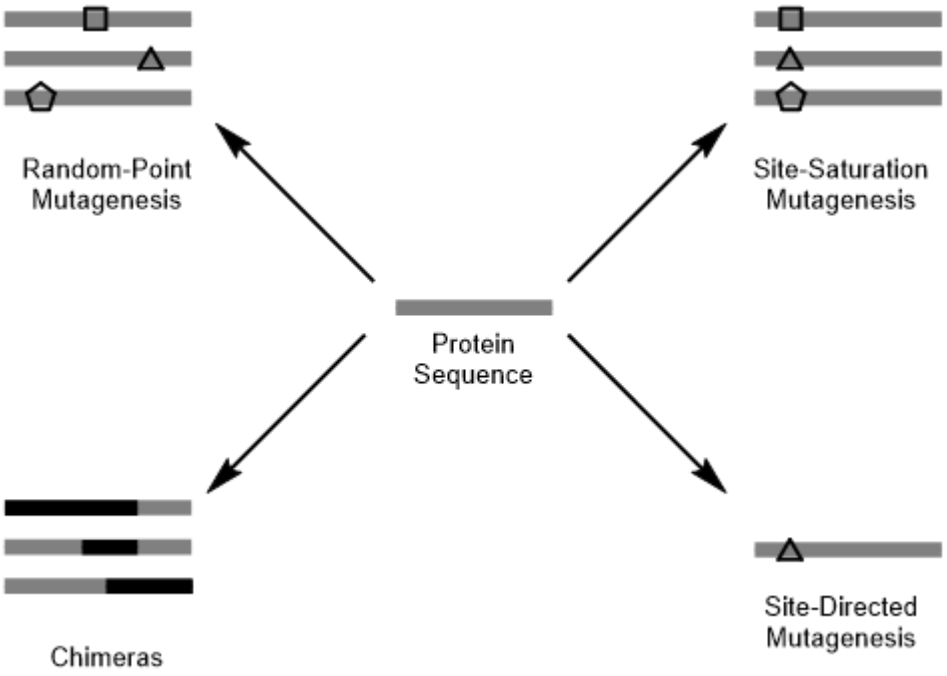


Figure 1.7: Diversity-generating strategies. Schematic of four different methods used to introduce variability within a protein sequence. Polygons represent different point mutations while black bars represent fragments from other proteins. Figure prepared using ChemDraw.[37]

Advantageously, with a reduced library size and number of iterations, one can use more sophisticated methods to assess the desired characteristics, moving away from high-throughput screening assays of easily measured values, which are only mediocre surrogates of the property of interest.[61] In practice, to reduce the number of variants, rather than mutating residues at any position, it is most effective to focus on the sites that are the most promising.[57] These may be near the active site or on key interdomain hinge regions,[62] near allosteric sites or on ones that affect protein dynamics.[63] Positions that are important for function can also be inferred from multiple sequence alignments that show conserved amino acids or from phylogenetic analyses that can suggest the evolutionary history of a set of enzymes.[57] This is the data that is for example used, along with structural information, by the free online server HotSpot Wizard[64] to suggest positions that would be interesting to mutate.

Once positions are identified as promising, rather than using site-saturation mutagenesis which too often leads to loss of activity, fewer substitutions can be chosen based on topological constraints (e.g., size), evolutionary variability (e.g., substitution observed in nature) and mechanistic features (e.g., the need for a general base in the active site).[57] One can also use the mutations that occur spontaneously during synthesis, use

substitution matrices (which predicts which substitutions are most likely) or just base the choices on literature reports of improvements in the property of interest.[61] Substitutions can first be assessed individually, or directly in combinations with each other, in an approach termed combinatorial active site saturation test (CAST).[65] Each substitution can be evaluated in advance to save the *in vitro* labour associated with the ones most likely to have a deleterious impact. This can be done *in silico* using QM and MD simulations as well as increasingly useful machine-learning algorithms.[57, 66] This has shifted the protein engineering approach from being discovery-based to being hypothesis-driven.[57]

In the context of the avenacin biosynthesis pathway, engineering individual enzymes to alter product specificity could lead to the gram-scale production of complex molecules through a platform for recombinant expression in *Nicotiana benthamiana.*[67] Alternatively, the discovered mutations could be introduced into oat by genetic editing to alter the structure of avenacin in a bid to overcome resistance to it, such as that of *G. graminis* var. *avenae.*[68]

1.4 Aims and objectives

The aim of the work reported in this thesis was to obtain structural information on the three targets from the avenacin biosynthesis pathway in *A. strigosa*: AAT1, TG1 and bAS1. This would help rationalise their specificity and enable their structure-informed engineering, leading to mutant enzymes that could produce new molecules against disease or for better food production.

The first objective was the production of large quantities of each enzyme to a high level of purity. This was needed for the second objective of crystallising them for structure determination by X-ray crystallography. In the absence of crystal structure, *in silico* methods would generate structure predictions instead. The last objective was to design mutants that would yield new products which would be difficult to access otherwise.

Chapter 2: General methods

This chapter aims to provide a general introduction and some general methods for the techniques routinely used throughout this thesis.

2.1 Construct design

Determining the structure of a protein by X-ray crystallography is a challenging task with many bottlenecks, the major one being the growth of a crystal suitable for diffraction analysis.[69] Many proteins contain intrinsically disordered regions, often at their termini. The formation of a protein crystal can only tolerate mild disorder or short segments of disorder.[70] Therefore, one would typically design constructs that lack these regions to improve both the chances of crystallisation and the conformational homogeneity of the macromolecules within a crystal.[71, 72] Constructs may contain a tag, be it a short polypeptide or a complete protein, added to facilitate purification, solubilisation and/or detection. Because tags may inhibit crystallisation,[73] it may be desirable to remove them from the protein of interest. This can be achieved by the introduction of a protease cleavage site between the gene of interest and the tag.[74]

For each target protein described in this thesis, the full amino acid sequence (without signal sequence if any was reported or predicted using SignalP5.0[75]) was submitted to the following disorder prediction webservers: DISOPRED,[76] PrDOS,[77] IUPred2A (context-dependent predictions using redox state and protein binding),[78] SPOT-Disorder2[79] and NetSurfP-2.0[80] (both long short-term memory and convolutional neural networks). The results were compared for general agreement, with useful cut-offs being 25% for DISOPRED, 5% false positive for PrDOS, 35% for IUPred2A, a probability of disorder above 5% for SPOT-Disorder2 and NetSurfP-2.0. The DNA sequence for the constructs was then obtained by polymerase chain reaction (PCR) using construct-specific primers (see section 2.2.2).

2.2 PCR

2.2.1 Introduction

Since it was first introduced in 1985 by scientists at Cetus, the polymerase chain reaction (PCR) revolutionised molecular biology by offering many strategies for amplifying,

modifying or detecting DNA. It generally consists in repeated cycles of the same three steps. First, the DNA is denatured with a high temperature (94 to 98 °C) to separate the strands. The temperature is then lowered enough to allow annealing of two different short pieces of DNA, called oligonucleotide primers, on either side and opposite strands of the target DNA. Third, the primers are extended by DNA polymerase, copying the sequence of the target DNA.[81] This cycle of denaturation, annealing and extension is repeated 20 to 40 times, doubling the amount of DNA with the target sequence each time, creating more template for the next cycles.[82] This exponential growth is the reason one can amplify a specific sequence from a minute amount of template, facilitating detection or cloning of the DNA fragment, especially if the primers themselves introduce the means to do so, such as labels or cloning sites.[81]

The ability of primers to anneal despite sequence mismatches allows the introduction of mutations, but also gives rise to one of the key limitations of PCR: the risk of non-specific amplification.[83] This can be mitigated by using longer primers, optimising their sequences or the buffer conditions and annealing at higher temperatures.[84] The original *E. coli* DNA polymerase I fragment was superseded by Taq DNA polymerase from a thermophilic bacterium because it remained active despite the high temperatures. However, the lack of "proof-reading" activity worsened the second main limitation of PCR: the random incorporation of mutations.[83, 85] Thankfully, many DNA polymerases, such as Phusion®, have now been developed to increase fidelity. Furthermore, the decreasing cost of DNA sequencing has democratised its use for sequence verification.[86]

2.2.2 Methods

2.2.2.1 Design of oligonucleotides

Oligonucleotide primers are designed based on the DNA sequence to be amplified, matching the sequence at the 5' end of the gene and the reverse complement at the 3' end. The 5' end of the primer has an additional sequence to introduce cloning sites (for recombination or for restriction digest) and/or protease cleavage sites. The length of the gene-specific region is decided based on several aims:

- A melting temperature between 55 °C and 65 °C according to the nearest neighbour method of Oligocalc.[87]
- Finishing with a G or C.
- Matching the melting temperature for primers used together within 4 °C.[88]

- Low potential of secondary structure and primer-primer annealing (which may require a conservative mutation to be introduced), as detected by the IDT OligoAnalyzer™ tool.

2.2.2.2 Preparation of oligonucleotides

Custom DNA oligonucleotides used as primers for PCR were ordered from Sigma-Aldrich. They are synthesized on a 0.025 µmol scale, purified by desalting only and shipped dry in tubes. Upon arrival, they are dissolved with Buffer EB (Qiagen), which contains 10 mM Tris-Cl, pH 8.5. The alkaline pH prevents acidolysis of DNA during long-term storage of this stock solution.[89] The volume of buffer used is what is recommended by the supplier to reach 100 µM of that particular oligonucleotide. Working solutions at 10 µM are then prepared as needed by diluting the stock with dH_2O.

2.2.2.3 Running PCR

A PCR mix is prepared with the components common to all reactions to be performed, such as the autoclaved dH_2O, concentrated buffer, dNTPs (dATP, dTTP, dCTP, dGTP to a final concentration of 0.2 mM each) and DNA polymerase. DMSO can be added (up to 5% v/v) to help with GC-rich sequences and decrease annealing temperature.[89] It is then dispensed in PCR tubes which usually contain the template DNA (0.1 ng/µL final) and the primers (each 0.5 µM final) to make a 20 µL reaction. A positive control is also set with a PCR reaction that is known to work, and a negative control is set either with no template DNA or with DNA lacking the gene to be amplified (e.g., empty vector). The reactions are then subject to a specific three-step heating programme using a Bio-Rad C1000 Touch thermal cycler.

2.2.2.4 Agarose gel electrophoresis

To analyse the results of PCR reactions, 2 µL samples are mixed with a density reagent and a tracking dye or used as is when the reaction buffer already contains them. Agarose (0.8-1.2% w/w) is dissolved in TAE buffer (40 mM Tris base, 20 mM acetic acid, 1 mM EDTA) by heating in a microwave. Once cooled to below 60 °C, ethidium bromide is added (0.6 µg/mL final) before pouring the solution in the taped tray of an electrophoresis tank to leave the gel to set at r.t. The samples are loaded along with 10 µL of the 1kb DNA ladder (NEB) diluted ten-fold (to 50 µg/mL) with the supplied loading dye in EB buffer (10 mM Tris-Cl, pH 8.5). A voltage of 90 V for 50 mL gels or 140 V for gels larger than 100 mL is

applied until the dye front reaches the ethidium bromide front. The gel is then imaged using a Gbox Chemi XRQ (Syngene) with transillumination at 302 nm.

2.3 Plasmids

2.3.1 Nature and structure

Plasmids are pieces of DNA used in this work in the form of vectors that carry the genes encoding for proteins of interests. Plasmids are naturally present in most bacterial species but can also be found in some archaea and single-celled eukaryotes. Within cells, plasmids exist outside of chromosomes and replicate autonomously, leading to copy numbers anywhere from 1 to thousands. Typically, their size ranges from 1 to 100 thousand bp (base pairs)[90] and they exist in a circular form that is supercoiled, making it more compact than the relaxed topology that can be achieved by introducing a single-strand break (a "nick"). A double-stranded break in a plasmid would linearize the plasmid. Through the action of DNA homologous recombination enzymes, plasmids can form dimers and other oligomers. When analysing a plasmid by agarose gel electrophoresis, the distance it migrates is affected by its size in bp, but also by which form it is in (Fig. 2.1).[91]

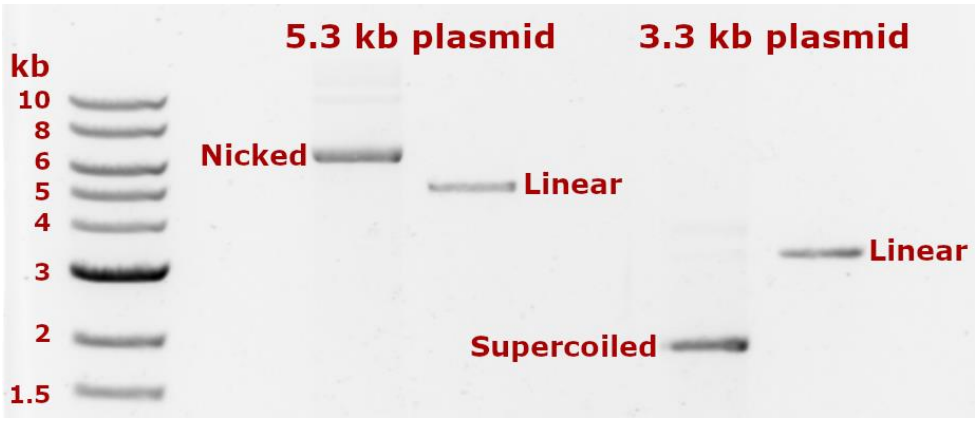


Figure 2.1: Apparent size of plasmids by agarose gel electrophoresis. The sizes in kb of the standard DNA markers are indicated on the left. Plasmids of two different sizes (indicated at the top) were run in either nicked (relaxed), linear or supercoiled form.

2.3.2 Origin of replication

Vectors used for recombinant protein expression are designed and refined according to the needs of the users, but are still based on natural bacterial plasmids. For example, they need an origin of replication (*ori*) that can be used by endogenous

transcription machinery, such as the ubiquitous ColE1-derived *ori*. ColE1 utilizes an RNA species and RNA-RNA interactions to control the number of copies produced.[91] But for higher copy numbers, the gene for the Rom protein that increase the affinity between the complementary RNAs is often removed.[92] This change, along with a point mutation in the RNA duplex, is what leads to the very high copy numbers of pUC-derived plasmids.[93] While a high copy number improves recovery by miniprep[94] and stability, i.e., the likelihood that a daughter cell contains the same plasmid as the mother cell, it can also reduce the growth rate of cells that contain many copies, which can then be outcompeted by cells that have fewer. For recombinant protein expression, cloning the gene in a vector with higher copy numbers does not necessarily increase the yield of protein.[95] The choice of *ori* also matters when several plasmids must be used at the same time. Indeed, if two plasmids using the same *ori* are in the same cell, they can compete for replication and/or partition machinery, often leading to plasmid loss in the daughter cells. Therefore, if a cell already contains a plasmid with one type of *ori*, e.g., p15a (as found in the plasmid pRARE covered in section 2.7.2.2), then one can only co-propagate it with a plasmid that uses a different type of *ori*, e.g., ColE1 (such as pDEST17 and pH9GW, covered in section 2.4.1).[91]

2.3.3 Promoter

The choice of promoter, however, has a major influence on recombinant protein yield as it regulates transcription initiation.[91] A stronger promoter therefore leads to more mRNA transcripts available for protein synthesis. But ideally, it also needs to be tightly regulated to provide precise control over the induction time, as premature recombinant protein expression ("leakiness") can impede cell growth or cause plasmid instability, especially if the protein is toxic.

The first elucidated promoter was in the *E. coli lac* operon, which initiate β-galactosidase synthesis in the presence of lactose, or the analogue isopropyl-β-D-1-thiogalactoside (IPTG) that acts as a gratuitous inducer (an inhibitor that is not a substrate because it is not hydrolysed).[96] This promoter has been engineered into the *lac*UV5 by disrupting the glucose-sensitive repressor to allow protein expression in *E. coli* using rich media.

The T7 promoter is widely used for recombinant protein expression, for example in pET-derived vectors. It is recognised by the bacteriophage T7 DNA-dependant RNA polymerase instead of the endogenous polymerase, providing control over basal

expression. The T7 polymerase can be produced using a copy of its gene integrated into the host chromosome and under the control of the *lac*UV5 promoter to allow induction with IPTG.[91]

2.3.4 Ribosome-binding site

For a messenger RNA transcript to be translated into a protein by a prokaryotic organism, it requires a ribosome-binding site (RBS) that can be recognised by the host's translation machinery. This is achieved by adding a Shine-Dalgarno sequence in-frame with and upstream of the start codon by around 6 bases. Because the promoter is further upstream, this sequence is transcribed into the RNA transcript, allowing the ribosome to bind it to initiate translation of the open-reading frame between the start and stop codons.[91]

2.3.5 Selectable marker

Since each plasmid incurs a metabolic cost, it needs to provide a competitive advantage to be selected for. Otherwise, cells lacking the plasmid would outcompete those that carry it and take over the culture. The selectable marker used most often is an antibiotic-resistance gene carried by the vector. Provided the cells are not already resistant, growth in the presence of that antibiotic will prevent cells lacking the selected plasmid from surviving, while those that carry it can proliferate. Typical antibiotics used for this purpose are ampicillin, kanamycin, chloramphenicol and gentamycin.

Ampicillin inhibits a bacterial transpeptidase involved in cell-wall synthesis.[91] It is therefore particularly effective at stopping growth, thought it may not actively kill cells that are already present. Resistance is conferred by a gene for the β-lactamase enzyme that breaks down ampicillin. A downside of this gene is that the enzyme is secreted, which can cause degradation of the antibiotic in the medium. On solid media for example, this can allow the growth of small satellite colonies of bacteria lacking the resistance gene near a bigger colony that is degrading the antibiotic around it.[97] This issue can be mitigated by using carbenicillin, an ampicillin analogue that is more resistant to degradation.[98]

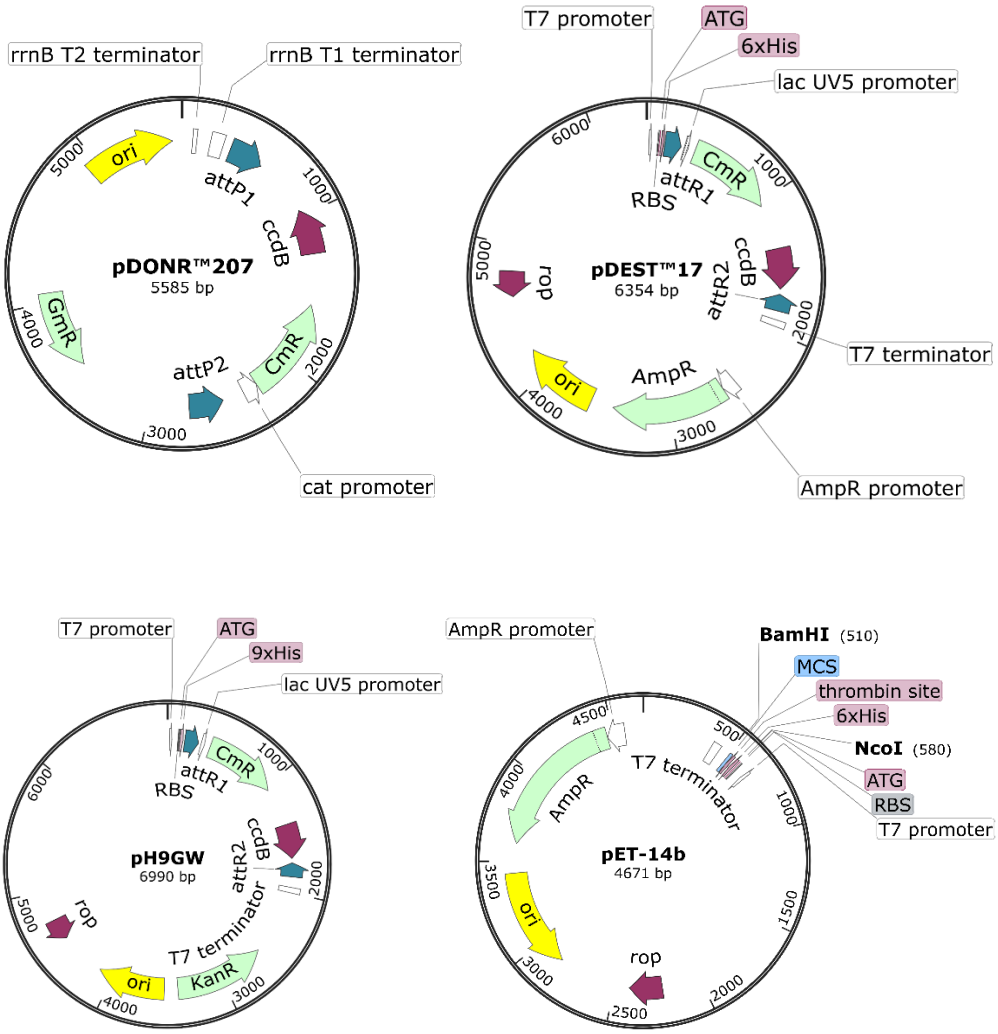
Kanamycin is an aminoglycoside antibiotic that interferes with protein synthesis. Resistance is conferred by a gene for the amino-phosphotransferase that inactivates it after it has been taken up by the cell.[91] Therefore, it does not suffer from the issue of degradation in the media. Chloramphenicol and gentamicin also interfere with protein

synthesis, but can be inactivated by specific acetyltransferases,[99, 100] the genes of which are used in various plasmids.[101]

Zeocin™ is a glycopeptide from the bleomycin family of antibiotics, which shows strong toxicity in both prokaryotes and eukaryotes. It can bind and cleave DNA, leading to cell death.[102] Resistance is conferred by the *Streptoalloteichus hindustanus* bleomycin gene (*Sh ble*) that codes for a protein that stoichiometrically binds this family of antibiotics, inhibiting their DNA cleavage activity. [103]

2.3.6 Plasmid maps

Features of the plasmids used for the work reported in this thesis are indicated on the circular maps in Fig. 2.2.



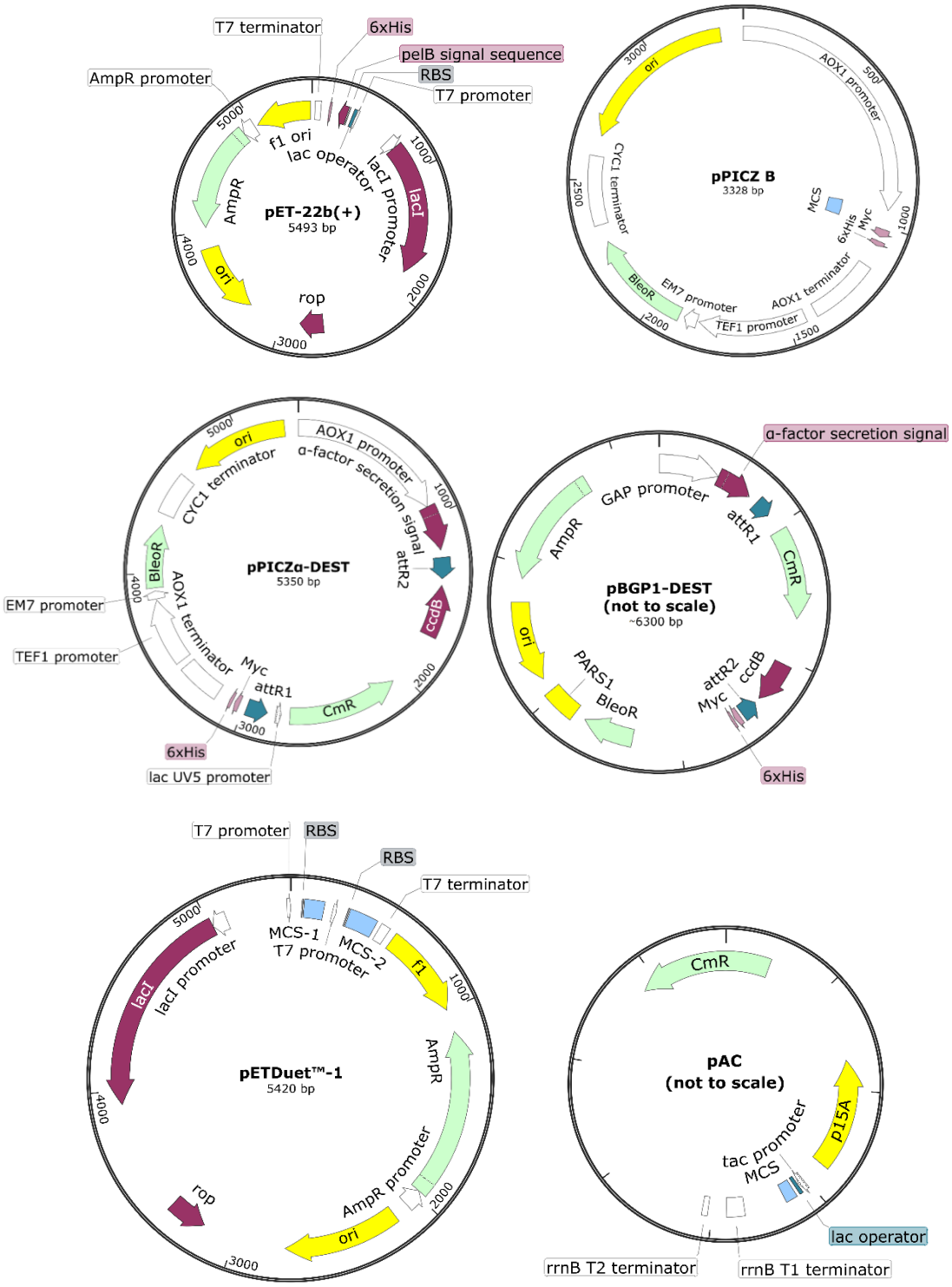


Figure 2.2: Maps of the plasmids used in this work. Features are represented as boxes: white for promoters and terminators, purple for proteins (including the lac and rop repressors), light blue for multiple-cloning sites (MCSs), yellow for origins of replication (ori for the ColE1 type), pink for protein tags, green for antibiotic-resistance genes (Gm = gentamycin, Cm = chloramphenicol, Amp = ampicillin, Kan = kanamycin, Bleo = bleomycin). Figure created using SnapGene 6.1.[104]

2.4 Cloning

2.4.1 Gateway cloning

The Gateway® cloning technology allows the insertion of a gene into a vector based on the mechanism used by the bacteriophage λ to integrate its DNA into the chromosome of *E. coli*. This relies on a highly specific recombination at *att* sites. All of them contain a 25-bp recognition region, which come in four types (*att*B, *att*P, *att*L and *att*R) depending on the presence or absence of "arms" on either side that allow interaction with recombination enzymes. A first mix of enzymes called BP Clonase™ contains the phage's recombination protein integrase and the *E. coli* protein integration host factor. This catalyses *in vitro* recombination between an *att*B-containing DNA fragment (PCR product or expression clone) and an *att*P-containing donor vector. In this thesis, the donor vector pDONR207 (ThermoFisher) was used. The plasmid created by this BP reaction is an *att*L-containing entry clone (Fig. 2.3A). A second enzyme mix called LR Clonase™ contains an additional protein: the phage's excisionase. This catalyses the reverse reaction, allowing recombination of the *att*L-flanked gene from the entry clone with an *att*R-containing destination vector. This generates an *att*B-containing expression clone (Fig. 2.3B).[105] In this thesis, the expression clones are derived from the destination vectors pDEST17 (ThermoFisher) and pH9GW, a Gateway-enabled derivative of the pET-28a vector.[106]

To allow directional cloning, the recognition region of *att* sites has been mutated to form subtypes that only recombine with each other. For example, *att*B1 sites recombine with *att*P1 sites, but not with *att*P2 sites. Therefore, by adding an *att*B1 site at the start of a gene of interest and an *att*B2 site at the end, one can be sure the gene would be cloned in a known orientation into an entry vector that has an *att*P1 and an *att*P2 site, such as pDONR207. Then, when recombining the resulting donor vector with a destination vector that has an *att*R1 site on the promoter side and an *att*R2 site on the terminator site (such as pDEST17), the gene of interest is always cloned in the correct orientation.[105]

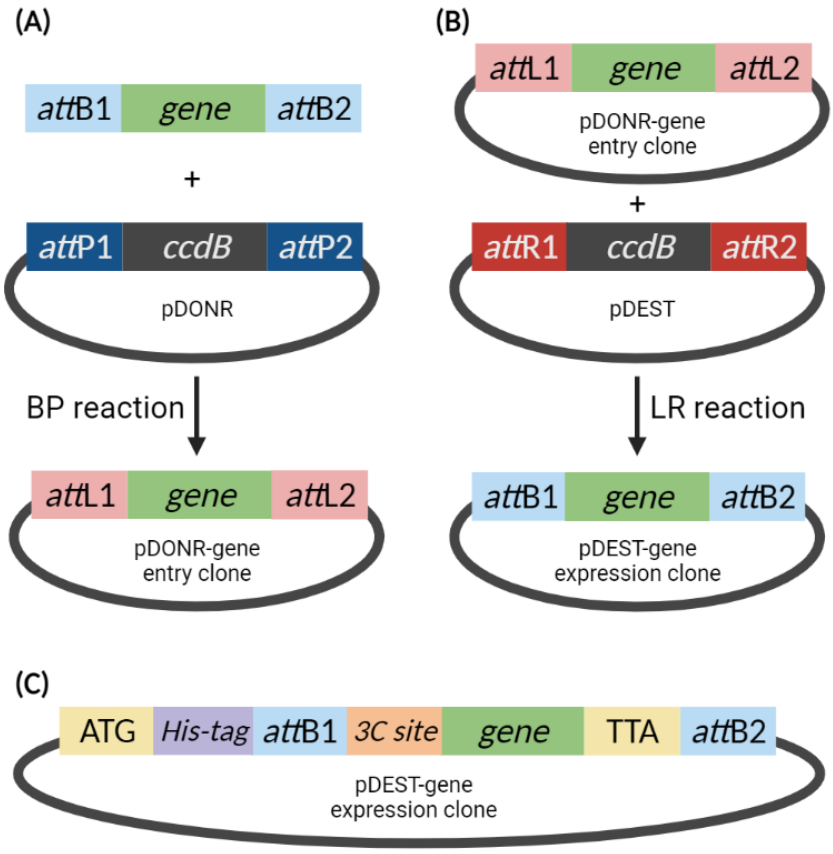


Figure 2.3: Gateway cloning (A) The BP reaction generates an entry clone by recombination of *att*B sites (eg in a PCR product) with corresponding *att*P sites on a donor vector, displacing the toxic *ccdB* gene and creating *att*L sites. **(B)** The LR reaction generates an expression clone by recombination of the *att*L sites of the entry clone with the corresponding *att*R sites on a destination vector, which creates *att*B sites. **(C)** Final construct with a 3C protease cleavage added to the gene. The position of the start (ATG) and stop (TTA) codons are shown, as well as the His-tag gene from the vector. Created with Biorender.com.

The choice of destination vector determines what tags are added to the gene of interest. For example, pDEST17 provides an N-terminal 6xHis-tag while pH9GW adds an N-terminal 9xHis-tag.[106] Since the ability to cleave the tag off the protein of interest is usually desirable for crystallography, a protease cleavage site can be introduced.[74] This can be achieved at the gene amplification stage by adding a DNA sequence coding for the cleavage site into gene-specific primers. For an N-terminal tag, the following nucleotides, which code for a 3C protease cleavage site (in bold), were added on the 5' end of a gene specific forward primers:

5' **CTG GAA GTT CTG TTT CAG GGC CCG** ... ... (gene specific primer) 3'

The reverse primer used is specific for the gene and adds a first fragment of the *att*B2 site (in italics):

5’ *CAA GAA AGC TGG GT*T … … (gene specific primer) 3’

A first stage PCR reaction using these primers yields an amplified product that can be used in a second stage PCR reaction to complete the *att* sites (in italics) using a general *att*B1+3C-site forward primer:

5’ *GGGGACAAGTTTGTACAAAAAAGCAGGCT*TC**CTGGAAGTTCTGTT** 3’

The reverse primer is a general *att*B2 primer:

5’ *GGGGACCACTTTGTACAAGAAAGCTGGGT*T 3’

The PCR product of this second stage reaction thus contains the gene of interest with a 3C-cleavage site directly upstream and flanked by *att*B1 and *att*B2 sites, allowing Gateway cloning into a donor vector then into a destination vector. The protein of interest expressed by the resulting plasmid will contain a 3C protease cleavage site separating it from the N-terminal tag conferred by the vector (Fig. 2.3C).

2.4.2 Restriction enzyme cloning

Restriction enzymes are sequence-specific endonucleases that usually introduce double-stranded breaks. They were discovered because they restricted bacteriophage growth on *E. coli*[107] by preventing infection through cleavage of the phage’s DNA. Thousands of restriction enzymes were discovered and hundreds are available commercially, some engineered to fit the user’s need (such as better tolerance of a common buffer, reducing “star activity” which is non-specific cleavage).[108] This offers a wide variety of recognition sequences that can be cleaved to leave either “sticky ends” (with an overhang) or “blunt ends” (no overhang) as shown in Fig. 2.4.

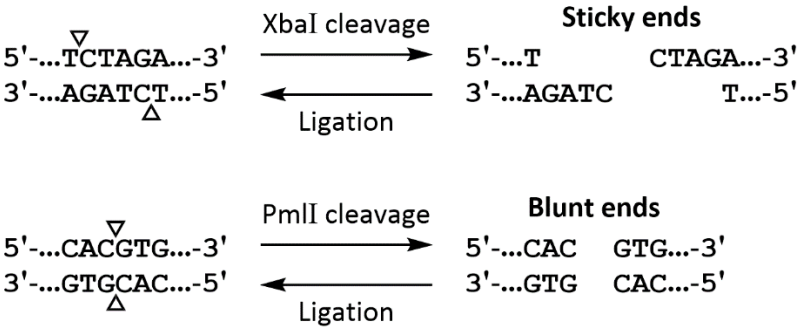


Figure 2.4: Formation and ligation of "sticky" and "blunt" ends. The DNA is cut by the restriction enzymes XbaI and PmlI at the sites marked by triangles on their recognition sequences, forming fragments that can be ligated with compatible ends using DNA ligase. Figure prepared using ChemDraw 21.[37]

Gene-specific primers that also contain a restriction enzyme site of choice on their 5' end are used to amplify the gene of interest by PCR. The amplified fragment and the vector can be digested with the same set of two restriction enzymes, yielding linearized DNA with compatible ends. For example, both the linearised vector and digested fragment may have a blunt end, which is always compatible with the other. On their other end, both could have overhangs of complementary sequences (sticky ends), making them compatible. The small left-over fragment can be purified out so they do not re-ligate in the subsequent ligation reaction, which uses T4 DNA ligase to clone the gene of interest into the chosen vector.[109] DNA fragments with "blunt ends" give lower ligation efficiency than "sticky ends" fragments, but this can be overcome with larger quantities of ligase and longer incubation times.[110]

2.5 Mutagenesis

Site-directed mutagenesis consists in deliberately introducing a specific change in a DNA sequence. This allows the precise tailoring of the amino acid sequence of the protein product, making this technique an invaluable tool in molecular biology. For example, it can be used for studying the function of a protein or for protein engineering. A widely used method is QuickChange™ (Agilent), which relies on PCR using gene-specific complementary primers that both contain the desired mutation (Fig. 2.5A). However, issues often arise because of primer-primer annealing and the scope is limited to single mutations. Several improved methods have been proposed to overcome these issues and improve amplification efficiency.[111] The strategy used in this thesis allows insertions and multiple mutations while keeping straightforward primer design principles. It consists in using gene-

specific primers that still have a complementary region on the 5’ end but also have relatively large non-overlapping 3’ ends. Unlike canonical QuickChange™ primers, these extended primers can use newly-synthesized mutant DNA as template because they bridge the nick introduced in the first cycle (Fig. 2.5B).[112]

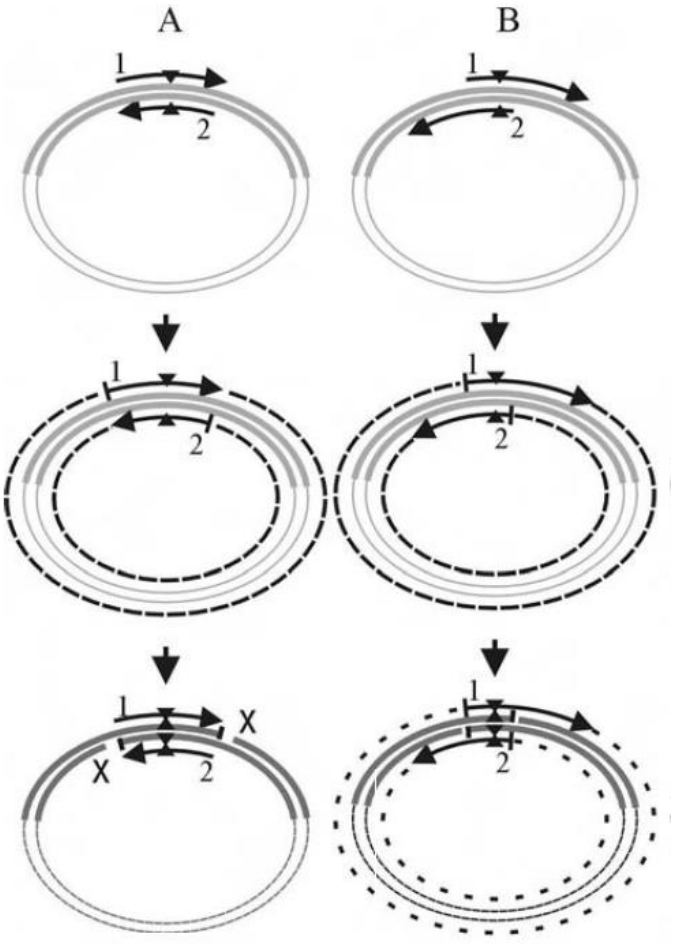


Figure 2.5: Mutagenesis strategies. Schematic representation of two PCR mutagenesis methods. **(A)** The QuickChange™ method uses complementary primers 1 and 2 (circular arrows) that harbour the mutation to be introduced (black triangle). After a first extension (long dashes), the newly synthesized DNA contains nicks (near the two X symbols), which prevents its use as a template for further amplification. **(B)** A modified method uses primers with overlapping 5’ ends and non-overlapping 3’ ends, which can bridge the nicks introduced in the first extension step to allow further cycles of amplification. The two strands of the plasmid template are shown as grey circles. Figure adapted from Liu *et al.*[112]

For the PCR programme, the length of the extension step is proportional to the size of the plasmid, and the temperature of the annealing step is T_m – 5 °C, with T_m being the melting temperature of the gene-specific region of the primer calculated using the

OligoCalc Nearest Neighbour method[87] The presence of a product of the correct size is confirmed by 0.8% (w/w) agarose TAE gel electrophoresis with ethidium bromide staining. While the template DNA (which lacks any mutation) is methylated from its propagation in *E. coli*, the newly synthesised plasmid that carries the mutation is not. Therefore, the template DNA can be selectively digested using the restriction enzyme DpnI, which cleaves DNA at the recognition sequence GATC when it is methylated, for example by the *E. coli* protein called dam.[113]

If agarose gel electrophoresis revealed a pure PCR product, then the remaining reaction components, primers, and the digested template can be purified out using the QIAquick® PCR purification kit (Qiagen). Otherwise, if there are unwanted side-products, they can be removed by running the entire reaction on agarose gel electrophoresis, cutting the band of the right size out and extracting the DNA from it using a QIAquick Gel Extraction Kit (Qiagen). The purified mutagenized DNA is prone to low transformation efficiencies, so XL10-Gold® Ultracompetent Cells (Agilent) are used for propagation. The plasmid is extracted from a transformant culture using the QIAprep Spin Miniprep Kit (Qiagen). To confirm the presence of the right mutation, the DNA is sent for sequencing (Eurofins Genomics) using primers that bind upstream and downstream of the mutated region. For example, the primers can bind the promoter and terminator sequences, or an internal sequence for large genes.

2.6 Transformation

Transformation refers to the uptake of genetic material from the environment through the cell membrane. Any microorganism that is able to undergo transformation is called “competent”. Competency can be triggered by environmental factors, such starvation or DNA damage,[114] but can also be achieved in the lab. The two most frequently-employed methods are chemical competence and electroporation.[115] In bacteria, chemical competence is usually achieved by the use of divalent cations. Though it remains largely unknown, the process may involve ion bridges between the negatively charged phosphate groups of both the DNA and the lipopolysaccharides of the membrane. Once bound to the cell, the DNA enters it during the heat shock step, where the cells are moved from low to high temperature and back.[116]

For the work covered in this thesis, *E. coli* transformation was generally performed according to the following protocol: *E. coli* competent cells were thawed on ice, then

transferred to 1.5 mL microcentrifuge tubes in 10-50 µL aliquots. DNA at 2 or 2.5 ng/µL was added to the cells in aliquots of 0.2 to 5 µL and gently stirred. The cells were left on ice for 30 min before heat-shocking in a 42 °C water bath for 30 s exactly. They were then put back on ice for 1-2 min before adding 100 µL of Super Optimal broth with Catabolite repression (SOC) media (Invitrogen) preheated to 42 °C. For recovery, this was then incubated at 37 °C with shaking at over 400 rpm, though recovery could be skipped when the introduced plasmid conferred ampicillin resistance. The majority of the transformation reaction was then plated on antibiotic selective LB agar before incubation at 37 °C overnight for colony development.

2.6.1 Competent cells

Competent cells were usually prepared in-house except for commercial aliquots of Library Efficiency™ DH5α (Invitrogen™), which was often used for transforming cloning reactions and propagating the resulting plasmids, and XL10 Gold (Agilent™), which were used for transforming mutagenesis reactions. The following protocol was performed under sterile conditions with autoclaved media and solutions. Commercial or in-house cells were inoculated into 10 mL LB with appropriate antibiotics (e.g., chloramphenicol for Rosetta cells) and incubated at 37 °C and 180 rpm overnight.

A 250 mL conical flask with 50 mL of LB was inoculated with 2 mL of overnight culture then grown at 37 °C and 180 rpm to an OD_{600} of 0.3 – 0.4. The culture was then put on ice for 15 min before centrifugation at 2,000 *g* and 4 °C for 5 min. The spent media was then discarded and the cells resuspended in 30 mL of 0.1 M $CaCl_2$ before leaving on ice for 30 min. The cells were then harvested by a second centrifugation step at 2,000 *g* and 4 °C for 5 min before being resuspended in $CaCl_2$ with 30% (v/v) glycerol. They were then aliquoted, flash-frozen in liquid nitrogen and stored at –80 °C before use.

2.6.2 Confirming the presence of the gene of interest

After transforming a cloning reaction into *E. coli*, the resulting transformants are not guaranteed to contain the desired plasmid. For example, with restriction enzyme cloning, it is possible that the transformants only contain the empty vector, which could have been religated. Transforming the BP reaction that aimed to clone a PCR product into a Gateway entry vector can also yield transformants that do not contain the gene. Therefore, the presence of the insert must be confirmed. This can be done with PCR using gene-

specific primers and analysing the results by gel electrophoresis. As the LR reaction is more reliable, the presence and sequence of the gene was not verified in the resulting clones.

2.6.2.1 Colony PCR

A transformant colony could be used to inoculate some liquid media to propagate the transformed plasmid, which would then be extracted and used as template for PCR. However, time and resources can be saved with a colony PCR protocol.

A small amount of several transformant colonies can be picked and resuspended in 25 µL dH_2O with 10 µL of this resuspension kept for culturing if needed, while the remaining 15 µL are incubated at 98 °C for 10 min then centrifuged at 3,220 *g* for 12 min. The supernatant is used as a DNA sample for PCR using gene specific primers. A positive control is set with DNA known to contain the gene of interest, and negative controls can be set with the empty vector (neat and/or from a colony) and/or no DNA. The colony PCR reactions are then analysed by electrophoresis on a 0.8% agarose gel with ethidium bromide staining to detect a product of the correct size.[117]

2.6.2.2 Sequencing

If a colony does have an insert of the correct size that can be amplified with gene-specific primers, the earlier resuspension (10 µL) is inoculated into 10 mL LB with the appropriate selective antibiotic, then incubated at 37 °C and 180 rpm overnight for propagation of the plasmid, which can be extracted from the grown culture and sent for sequencing using either gene-specific primers or sequencing primers. Sanger sequencing is performed by Eurofins Genomics on samples prepared according to their instructions. The results are analysed and aligned to the reference sequence using Geneious 9.0.5.

2.7 Recombinant protein expression in *Escherichia coli*

Escherichia coli is by far the most prolific source of recombinant protein. In the context of structural studies, it is the expression system listed for over 135,000 PDB entries, far more than the second most common system: the insect cells from *Spodoptera frugiperda* that yielded around 6,800 structures.[118] *E. coli* has the advantages of easy transformation and extremely fast growth to a high cell density in cheap media. Since the first heterologous expression of proteins with therapeutic value in the late 1970s,[119] there has been extensive development of expression vectors (see section 2.3), cultivation strategies and engineered strains (covered in this section).[120] Some of these aim to

increase the levels of protein expression while some address the crux of proper protein folding.

2.7.1 Protein folding

The environment of protein synthesis in *E. coli* can be vastly different than that of the original host, for example in terms of pH, redox potential and post-translational processing. This can cause protein instability and aggregation, especially as the exposed hydrophobic regions of overexpressed proteins interact. This leads to the formation of inclusion bodies inside the cells, reducing the yield of soluble protein. While some proteins can be refolded from inclusion bodies *in vitro* using denaturing agents like urea or guanidine hydrochloride,[121] the most wide-spread strategy to promote better protein folding is to induce protein expression at a lower temperature, which reduces hydrophobic interactions.[120] For the projects described in this thesis, most protein expression was carried out at 16 °C.

2.7.2 *E. coli* expression strains

Most *E. coli* cells used for protein expression are derived either from the B strain or from the K-12 strain. The latter was isolated at Stanford University in 1922 from the faeces of a convalescent diphteria patient.[122] The B line, on the other hand, has a more nebulous history prior to its naming by Delbrück and Luria in 1942. It may originate from a strain, then named *Bacillus coli*, which was kept in the Collection of the Institut Pasteur and was used by d'Herelle in 1918. Through various mutagenesis methods, transduction, and selection, *E. coli* strains with desirable genotypes were created from these two cell lines and extensively studied, generating an efficacious toolbox for recombinant protein expression.[123]

2.7.2.1 BL21(DE3)

E. coli BL21 is the most common host for recombinant expression.[124] It is a B line derivative that is deficient in the proteases lon and ompT to reduce degradation of foreign and extracellular proteins. BL21(DE3), like all DE3 lysogen strains, carries a T7 RNA polymerase gene under control of the *lac[UV5]* promoter, so that addition of IPTG allows translation of genes under control of the T7 promoter.[91]

2.7.2.2 Rosetta

Rosetta is a BL21 derivative carrying the pRARE plasmid, which includes a chloramphenicol resistance gene for selection and allows the expression of six tRNAs for codons that are seldom used in *E. coli* coding sequences. Rosetta 2 carries the pRARE2 plasmid that expresses an additional rare codon. This reduces translation issues associated with insufficient availability of tRNAs, such as stalling, termination, frameshifting and amino acid substitutions.[91]

2.7.2.3 Rosetta-gami 2

Expression of recombinant proteins in *E. coli* can be challenging if they contain disulphides bonds and are directed to the reducing environment of the cytoplasm, as they typically are. This is due to the presence of numerous reductases, glutathione and other thiol reductants. While one could attempt to delete the genes for thioredoxin reductase (*trxB*) and for glutathione reductase (*gor*), this is usually lethal to the cell as it cannot reduce essential proteins into their active state. Fortunately, viability can be restored by a mutation in the *ahpC* gene which turns its cytoplasmic peroxidase product into a disulphide reductase.[125] This was done for the strain commercialised under the name Origami 2 by Novagen. Rosetta-gami 2 carries the same pRARE2 plasmid as Rosetta 2, but is derived from Origami 2 (itself a K-12 derivative as opposed to Rosetta 2).

2.7.2.4 SHuffle

While newly synthesized proteins can be directed to the periplasm, cysteines will be oxidised consecutively by the enzyme DsbA. If the recombinant protein needs non-consecutive cysteines to form a bond, it will require the action of a disulphide isomerase to achieve a native fold. This is achieved in the periplasm of *E. coli* by the isomerase DsbC, which has a hydrophobic cleft that can interacts with the exposed hydrophobic residues of a misfolded protein and that may act as a chaperone. It uses its two thioredoxin arms to isomerise the mismatched cysteines of the misfolded protein with its own redox-active cysteines. This can also increase the amount of correctly folded protein in the cytoplasm if DsbC is overexpressed inside of it. That is the strategy adopted for the SHuffle strain, a K-12 derivative that has mutations in *trxB* and *gor* like Origami 2, but also constitutively expresses DsbC in the cytoplasm. The B line derivative equivalent is called SHuffle Express.[126]

2.7.2.5 ArcticExpress (DE3) RP

While growth at lower temperatures is an effective strategy for increasing the recovery of soluble proteins, it can lead *E. coli* chaperonins such as GroES/EL to lose activity. This can paradoxically impede proper folding of recombinant proteins as chaperonins sequester misfolded proteins and assist them in achieving a native fold.[127] To compensate for this, the ArticExpress strain (a B-line derivative) constitutively expresses the Cpn10 and Cpn60 proteins. These are cold-adapted chaperonins from *Oleispira antartica* that are highly active even at 4 °C.[128]

2.7.2.6 SoluBL21

To alleviate the solubility bottleneck of recombinant protein expression, a mutant strain of BL21(DE3) was developed by directed evolution. SoluBL21 is able to produce soluble or partly-soluble recombinant proteins which are completely insoluble or toxic when expressed in BL21.[129]

2.8 Recombinant protein expression in *Pichia pastoris*

While *E. coli* is the most widely-used organism for recombinant expression of bacterial proteins, the success rate for eukaryotic proteins, especially larger ones, is significantly lower.[130] Eukaryotic expression systems can alleviate some of the shortcomings by providing co-translational and post-translational processing, such as disulphide bridge formation, folding, glycosylation and signal sequence processing.[131] Systems based on mammalian cells or insect cells have yielded numerous challenging targets, but they are particularly expensive and time-consuming. A middle-ground can be found with *Pichia pastoris*, a yeast expression system that is widely used for the production of recombinant eukaryotic proteins. It owes its success to its ability to reach high cell densities in aerobic growth, its efficient and tightly-regulated AOX1 promoter, its ease of use and its low cost.[132] It is also well-characterised, in part because of its similarity with the model yeast *Saccharomyces cerevisiae*, and it has seen extensive development, starting in the 1970s, when Phillips Petroleum Company established media and protocol for the production of single-cell protein for animal feed. It was then optimised for recombinant protein expression by the Salk Institute Biotechnology/Industrial Associates. In 1993, the patent was sold to Research Corporation Technologies, availing the system to academic research to this day, currently under a license to sell the components that is held by Life Technologies.

The genomes of both *S. cerevisiae* and *P. pastoris* can undergo homologous recombination with artificially introduced DNA.[131] While *S. cerevisiae* is often used for recombinant protein expression due to the ease of cloning and transformation, especially in higher-throughput projects,[133] *P. pastoris* has seen better success for structural studies. This is in part thanks to higher yields, which are required for crystallisation, and because it is less prone to hyper-glycosylation. Indeed, while yeast can add $(Man)_8$-$(GlcNAc)_2$ cores that are common among higher eukaryotes, it sometimes adds longer chains of sugars that can impede crystallisation. This issue can be further reduced by mutating glycosylation sites,[132] expressing in glycoengineered strains such as KM71H (*OCH1::G418R*) and/or enzymatically deglycosylating the resulting protein.[134]

2.9 Protein purification

2.9.1 Introduction

After a protein has been overexpressed in a recombinant host, it must be extracted. Intracellular expression requires subsequent lysis of the cells, which comes with the major issues of denaturation, proteolysis and contamination of the target protein. Stability can be improved by keeping the temperature low, shortening the preparation time and adding protease inhibitors.[135] In this thesis for example, lysis is performed in a cooled vessel with a commercial protease inhibitor cocktail, while the following steps are carried out as soon as possible and at 4 °C. Contamination is resolved through purification. Because X-ray crystallography requires high levels of purity (>95%), several purification steps are likely to be needed.[136] First, the cell lysate is clarified by centrifugation and/or filtration. The resulting soluble fraction is then purified, for example by selective precipitation (not used in this thesis) or by chromatography, which has largely supplanted preparative electrophoretic techniques.[137]

2.9.2 Chromatography

Chromatography is one of the most common methods for protein purification. It consists in applying a mixture of molecules onto a stationary phase, often in a column, and using a mobile phase to carry the components across it. Because the molecules interact with the stationary phase to various extents, some will move more slowly than others through the chromatography system. The aim is to exploit this difference to successfully separate the protein of interest from its contaminants. There exist a range of

chromatography techniques that can separate molecules based on size, affinity, charge or hydrophobicity.[138]

2.9.2.1 Affinity chromatography

To create the stationary phase, the resin in the column can be functionalized with a ligand that interacts strongly but reversibly with the protein to be purified. The impure sample is passed through the column, allowing binding of the protein of interest. Unfortunately, there may be contaminants that also bind the ligand, or interact with the stationary phase by non-specific interaction (e.g., hydrophobic or ionic). Most can be removed by washing the column with a buffer optimised for minimal disassociation of the target protein and maximal contaminant removal. This can be done by altering pH, ionic strength or concentration of competitive reagent. This reagent should have the ability to bind the immobilised ligand or its binding site on the protein of interest. It is often used in higher concentration for specific elution, where it releases the target protein from the column by competing for binding. Elution can also be achieved by non-specific methods, where changing the conditions (e.g., pH or salt concentration) weakens the binding of the target protein to elute it.[139]

Two common immobilised ligands for affinity chromatography are streptavidin and nickel ions. Streptavidin is a protein that strongly binds an eight-residue peptide called *Strep*-tag, which can be introduced in the protein to be purified. Once the protein is bound and washed, it can be eluted using biotin which competes for streptavidin binding.[140]

Nickel(II) ions can be bound to chelating agents, such as nitrilotriacetic acid (NTA), that are immobilised onto a resin, which can be packed into a column to pass the mobile phase through it or be used as a slurry to be mixed with the mobile phase and recovered by centrifugation.[141] The chelated nickel(II) ions offer binding sites for poly-histidine tags that can be added to the protein's sequence. Elution is usually achieved with imidazole that displaces the histidine residues from the metal ion. Other metal ions, such as cobalt(II), can also be used to optimise protein purity and recovery, even using the same column. This is achieved by stripping the current metal ions using a chelating agent such as EDTA, then recharging with another metal ion.[142]

A widely applicable and effective purification strategy consists in a first step of affinity chromatography, followed by removal of the affinity tag by digestion with a tagged protease. A second step of affinity chromatography then separates the de-tagged protein

(which does not bind the immobilised ligand) from the protease and the tag (which bind the ligand). In some cases, an additional purification step is necessary, usually size-exclusion chromatography or ion-exchange chromatography.[136]

2.9.2.2 Size-exclusion chromatography

Also known as gel-filtration or gel-permeation chromatography, this technique uses a porous inert matrix that can separate molecules based on size. After loading the sample onto the equilibrated column, it is eluted with the mobile phase of choice. Molecules larger than the pore size do not penetrate the gel particles and therefore all elute first at the void volume, i.e., the volume of the interstitial space of the stationary phase. The smaller the molecule relative to the pore size, the more pore volume is available for them to diffuse into, delaying their elution time. Molecules that are small enough to freely diffuse through all the pores elute at the total volume of the stationary phase.[143] This stationary phase can be made of dextran (good selectivity), agarose (good stability), polyacrylamide[138] or a matrix that is a composite of several materials.[144]

2.9.2.3 Ion-exchange chromatography

In ion-exchange chromatography, proteins are separated based on their affinity with a charged matrix.[138] In anion-exchange chromatography, anions adsorbed onto a positively charged matrix are exchanged for negatively charged proteins, while positively charged proteins can be washed out. The most weakly-bound proteins can be selectively eluted by weakening ionic interactions, either by increasing the ionic strength (e.g., with a higher salt concentration) or by decreasing the pH (which decreases the charge on the protein). Gradient or stepwise changes in the mobile phase can therefore separate the target protein from its contaminants. If the target protein is positively charged at the chosen pH, cation-exchange chromatography can be used instead.[145]

2.9.3 SDS-PAGE

2.9.3.1 Introduction

Polyacrylamide gel electrophoresis (PAGE) is a ubiquitous method for separating proteins. Sodium dodecyl sulphate (SDS) is a detergent that allows to run PAGE in denaturing conditions and to separate proteins based on size, as it coats them with a negative charge that correlates with their length.[146] For the work presented in this thesis, SDS-PAGE is routinely used to analyse the results of each purification step.

2.9.3.2 Methods

Loading dye with reductant was prepared and kept at 4 °C for up to one month. It was either made in house (50 mM Tris-HCl pH 6.8, 2% w/v SDS, 0.1% w/v bromophenol blue, 20% v/v glycerol, 100 mM DTT) or from commercial stocks (NuPAGE™ LDS Sample Buffer 4X (ThermoFisher) with added 100 mM DTT or 4X Bolt™ LDS Sample Buffer with added 10X Bolt™ Sample Reducing Agent (InVitrogen)). The samples to be analysed were then added to 3 µL of loading dye with reductant and made up to 12 µL with dH_2O. Typical sample volumes were 9 µL of total culture before induction, 2.5 µL of final induced culture, 9 µL of a 100-fold dilution of insoluble fraction resuspended in total lysate volume with dH_2O, 0.2 µL of soluble fraction, wash and flow-through, 2 µL of elution fraction. These SDS-PAGE samples were then incubated at 95 °C for 10 min before loading 10 µL per well. The precast gels used were either TruPAGE 4-12% 8 x 10 x 0.1 cm x 17 wells (Sigma-Aldrich) with MOPS running buffer (20 mM MOPS, 50 mM Tris, 0.1% w/v SDS, 1 mM EDTA), NuSep Tris-Glycine NB 8% 8.5 x 10 x 0.1 cm x 17 wells (Generon) with Tris-Glycine running buffer (25 mM Tris HCl pH 8.3, 192 mM glycine, 0.1% w/v SDS) or Bolt™ 4-12% Bis-Tris Plus 8 x 8 x 0.1 cm x 12 wells with Bolt™ MES SDS Running buffer (InVitrogen). On a Mini-Protean® Tetra System (Bio-Rad), NuSep gels were run at 200 V and TruPAGE gels were run at 65 V while the dye front is in the resolving gel, then 120 V and 110 mA. Bolt™ gels are run at 165 V for 40 min on a Mini Gel Tank (InVitrogen). For Coomasie stain, the gels were incubated in InstantBlue (Abcam) at r.t overnight before imaging using a G:Box with the GeneSys software (Syngene) with white light.

2.9.3.3 Western Blot

Western blot is an important method for the identification of a specific protein in a complex mixture. Proteins are first separated by size using SDS-PAGE then transferred to a membrane, which is incubated with primary antibodies that specifically bind the protein of interest and are themselves recognised by secondary antibodies that carry a reporter for detection.[147]

For the western blot analyses presented in this thesis, proteins are transferred to a nitrocellulose membrane using an iBlot 2 Gel Transfer device. The membrane is then incubated in a dried-skimmed-milk-based blocking solution to prevent antibodies from binding the membrane non-specifically.[148] The excess primary antibodies are washed off before incubating with secondary antibodies.

2.10 Experimental determination of protein structure

2.10.1 Introduction

There exist numerous experimental methods that aim to elucidate the 3D structure of a protein, but the most common one is X-ray crystallography (87% of the structures in the PDB), followed by NMR spectroscopy (7%) and electron microscopy (5%).[149] The advantage of NMR is that it can provide unique information about the dynamics and interactions of a protein in near-native environments, but the method is only practicable for small proteins (below 40-50 kDa).[150] Cryo-electron microscopy (cryo-EM) enables the structure determination of proteins that are unamenable to the other methods because of their large size, conformational heterogeneity or varied multimeric states.[151] It is less demanding in terms of sample quantity and purity, but has historically yielded structures of much lower resolution. Over the past decade however, significant advances in electron detectors and image processing have catalysed a revolution leading to numerous high-resolution cryo-EM structures.[150] While the lower size limit is constantly being pushed, there are still relatively few structures of proteins below 100 kDa.[151] Given the size of the target proteins covered in this thesis (40-90 kDa), X-ray crystallography was the method chosen for structure determination.

2.10.2 X-ray crystallography

This method requires a crystal, grown with a high concentration of pure target molecule in the right conditions, that diffracts an X-ray beam into a pattern that can be used to determine the size and symmetry of the repeating unit cell within. The intensity of the spots in the diffraction pattern provides information that can used to obtain an electron density map in which to build an atomic model of the target molecule. The model is then iteratively refined to improve agreement with the experimental data while remaining in a thermodynamically favourable conformation. The key stages of this method are preparation of a pure sample (covered in section 2.9), growth of a crystal, diffraction experiments, data processing, phase estimation, model building, refinement and validation.[69]

2.10.2.1 Protein crystallisation

The growth of protein crystals suitable for X-ray crystallography is usually the rate-limiting step as it remains unpredictable and cannot be achieved for most targets.[152] The

most common crystallisation method is vapour diffusion, where a drop of concentrated protein solution is mixed with a precipitant solution, then incubated in a closed system with a reservoir of the precipitant solution. The protein-containing drop then slowly evaporates as its lower concentration of precipitant tends towards the concentration in the larger reservoir. The aim of this reduction in volume is to push the protein concentration past its solubility limit in these conditions, hopefully without causing precipitation. When this supersaturated state is reached, the protein may start to nucleate, forming a point where the crystal can grow. Growth usually continues until the removal of soluble protein brings its concentration down to undersaturation (Fig. 2.6).[153]

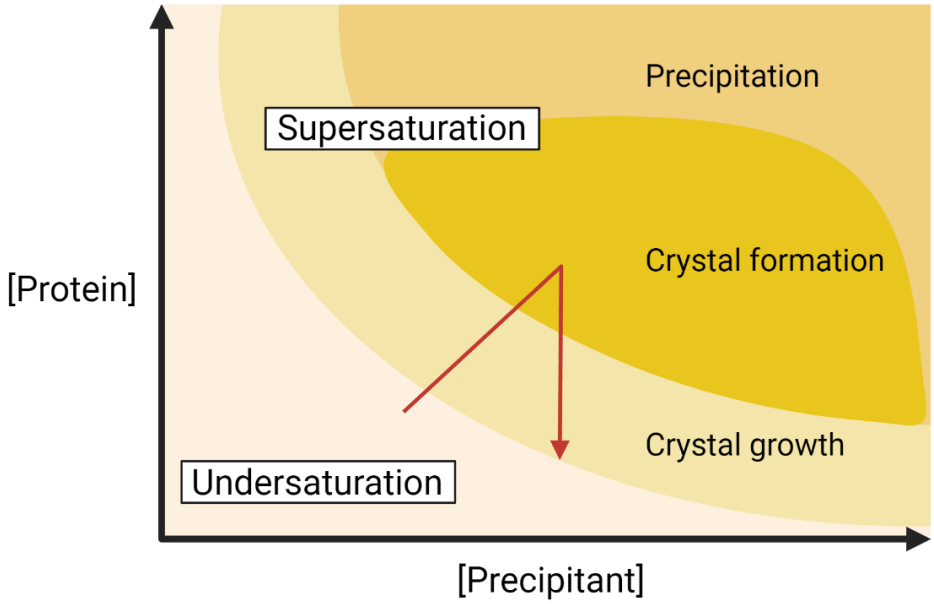


Figure 2.6: Phase diagram for protein crystallisation. The axes represent the concentrations of protein and precipitant. The red arrow shows the progress of a vapour diffusion experiment that successfully forms a protein crystal. Created with Biorender.com based on Chayen.[154]

Protein crystallisation is a rare event that depends on numerous variables, such as temperature, pH, precipitant, protein concentration, additives. A new target will therefore require a broad screening approach. This is typically achieved using specialised liquid handling robots and commercial screens that supply a diverse range of premixed sets of buffers, precipitants and salts that have helped crystallise proteins in the past. Crystallisation screening requires milligrams of pure protein to test thousands of different conditions that are monitored over several weeks using a light microscope. Nearly all the conditions will produce either clear drops (likely undersaturated) or precipitate (likely too supersaturated). A few conditions will yield crystals, though they may be salt crystals, or

protein crystals that are too small or present other issues that prevent the collection of useful diffraction data (such as twinning or disorder). To confirm the identity of the molecule in the crystal, some can be sacrificed by dissolving them and analysing them by SDS-PAGE and/or Western Blot.[69] As protein crystals contain 25% to 90% solvent, they behave in many ways like a gel, making them much more fragile than salt crystals. The best test of crystal quality remains X-ray diffraction.[155] While a good protein crystal may occasionally come from a first screening effort, the promising conditions are usually optimised. This can be done through grid-screening, where two variables (often pH and precipitant concentration) are systematically altered. Varying protein concentration and temperature can also yield better crystals. Crystals that have the wrong shape or size can be used to make seeds to add to a fresh drop.[153]

2.10.2.2 X-ray diffraction

X-ray diffraction experiments usually require the crystal of interest to be extracted from its drop. The most common tool is a mounted loop with an internal diameter close to the size of the crystal. Protein crystals being small and fragile, harvesting them is a delicate and skilful task. To preserve the crystals during transport and X-ray diffraction, most crystals are kept at a temperature near 100 K. To reach it, they are plunged in liquid nitrogen, though this can cause cracking or ice crystals that impacts data quality. To avoid these issues, freshly harvested crystals are immediately bathed in a cryoprotectant, which is similar to the crystallisation solution to avoid degradation of the crystal, but contains a high concentration of a compound that prevents the formation of ice crystals (for example, 30% glycerol).[156]

To obtain a high-quality dataset, the crystals are usually transported at cryogenic temperatures to a synchrotron light source, which can provide a stable high intensity monochromatic X-ray beam. The loop holding the crystal is mounted on a goniometer for alignment and rotations. Crystals are typically checked for diffraction before collecting a full data-set, which is obtained by rotating them under the X-ray beam and recording images of the resulting diffraction (the principles underlying the phenomenon of X-ray diffraction are well described in the literature).[157] When using a single-photon-counting detector, each image is best recorded over about 0.1° of rotation, with a total rotation of 180° or 360°.[158] Some later images may be discarded because of one of the main limiting factors in data collection: radiation damage. X-rays are highly energetic and can cause degradation, for example by heating or by the generation of free-radicals, which affects data quality.[69]

2.10.2.3 Data processing

The diffraction images need to be processed in several ways. First, the diffraction spots, also known as reflections, are identified. Secondly, the reflections are indexed. Each reflection is assigned a set of Miller indices, the integers *h*, *k*, and *l*, that define which plane of the crystal they arise from.[159] At this point, a hypothesis can be made for the unit cell dimensions, the crystal system and the space group. For example, the bigger the unit cell, the closer together the reflections will be. The crystal system depends on the relationships between the length of the three edges of the unit cell (a, b, an c) and between the angles they form (α, β, and γ). The symmetry of the reflections along with systematic absences are used to infer the possible space groups, i.e, how the molecules pack within the unit cell.[160] The third step of data processing is integration. This determines the intensity of the reflections, which is dependent on both the amplitude and the phase of the diffracted X-rays. Finally, the data is scaled and merged. Scaling corrects for image-to-image variations from the diffraction experiment, while merging matches reflections to their spread on other images or their equivalent by symmetry.[161]

2.10.2.4 Data evaluation

When obtained at the Diamond Light Source (Oxfordshire, UK), datasets are automatically processed[162] using a range of different pipelines such as DIALS[163] or Xia2 3dii[164]. The results are evaluated using several metrics. The high-resolution limit is based on the outermost reflections used in data-processing, with a lower number in ångström (Å) being higher-resolution. These reflections give information about the finer details of the structure, but their worse signal-to-noise ratio can decrease the quality of the data. Ideally, the cut-off should be decided based on the resulting structural information, but in practice, each pipeline calculates a reasonable high-resolution limit that can be manually changed to observe its effect on other metrics.[161] I/σ(I), based on the intensity (I) of reflections and their estimated error σ(I), gives an indication of the signal-to-noise ratio, so a higher value is better.[160] R_{meas} represents the agreement between reflections that are equivalent because of their Miller indices or by symmetry, with a lower value being better. $CC_{1/2}$ is the correlation coefficient between the two randomly divided halves of the dataset, with higher values being better. Completeness represents how much of the expected reflections are seen in a dataset, so a higher percentage is better.[165] Processing software also provides these metrics for the high-resolution set of reflections, helping decide whether it is worth including. One typically aims for a completeness above 70% in this outer shell and

93% in total.[166] Data was historically truncated at a resolution shell that has an I/σ(I) value of 2, though it was recently established that small improvements could be achieved from including data to values even below 1.[167] As for $CC_{1/2}$, some useful information can be found in shells that have a value between 0.1 and 0.5.[168]

2.10.2.5 Solving the phase problem

The intensity of a reflection is derived from the amplitude and the phase of the diffracted wave that gave rise to it. While the amplitude can be calculated easily, the phase cannot be measured directly for protein crystals. To be able to solve the structure of the protein, a first estimation of the phase must be obtained. This can be achieved by experimental methods such as isomorphous replacement or anomalous scattering.[69] But in most cases, it is easier and faster to use molecular replacement (MR). This requires a search model, such as a similar structure, ideally solved for a protein with high sequence identity. Alternatively, one can use a homology model (see section 2.11.1), a section of idealised secondary structure, or an ensemble of lower-quality models. This is used as a search model and may need to be edited to remove parts that do not match the structure to be solved. Phasing software can then look for the highest-scoring MR solution by rotating and translating the search model to try to place it as close as possible to the position of the crystallised protein in the unit cell. If successful, this solution is then automatically refined and yields a first estimate of phases, allowing a first electron density map and model to be obtained. Unfortunately, these may be closer to the search model than to the crystallised protein because of phase bias. It therefore needs to be iteratively improved through model-building and refinement.[69]

2.10.2.6 Model-building

Model-building can be carried out with interactive software such as Coot, which displays the current model and electron density map. Coot provides tools for adding and editing residues, ligands and solvent molecules. These can be refined in real-space to automatically match the electron density while avoiding deviations from idealised geometry (e.g., bond lengths and angles).[169] To find areas to focus on, Coot also offers validation tools to analyse fit to density, rotamer probability, Ramachandran plots and others.

2.10.2.7 Refinement

Refinement aims to improve the agreement between the model and the X-ray data, while keeping the model geometry as realistic as possible. Several programs are available for this task, with the most popular ones including *REFMAC*5,[170] *phenix.refine*[171] and *SHELXL*.[172] The fit of the model to the data can be estimated by several reliability factors. The two that are most often used are R and R_{free}, which are commonly expressed as percentages. The R-factor measures the agreement between the amplitudes calculated for the model and those obtained experimentally. It is however prone to overfitting of the data, such as building solvent molecules into the noise of the electron density. To avoid an artificial lowering of R that does not improve the accuracy of the model, R_{free} is much more widely accepted as it is calculated using a subset of reflections that was set aside during data processing so it is never used for model refinement.[173] It is therefore a useful indicator to obtain after each round of model building, which also benefits from the updated map obtained through refinement. These iterations are then repeated until the model is considered final.[170]

2.10.2.8 Validation and deposition

The final structure must undergo validation through the wwPDB OneDep system. This analyses the model to identify clashes between atoms, Ramachandran outliers, rotamer outliers, poor fit to density, unusual ligand geometry and other issues that may warrant further attention. After the final corrections, the structure can be deposited along with descriptions of the contents and experimental methods to obtain a unique PDB code. A wwPDB biocurator must check the submission and approve it to generate an official validation report that should be used for journal submissions. The structure is usually released to the public upon publication of the associated manuscript.[174]

2.11 Structure prediction

In the absence of experimentally determined structures, for example if a protein does not crystallise or for lack of time, the only way of obtaining structural information is through structure prediction. This strategy is based on a fact established in the 1960s: protein structures are uniquely determined by their sequence. Yet, going from sequence to structure has remained a challenge ever since, as it requires many steps, decisions and calculations that are not guaranteed to produce an accurate result.[175] The three main methods currently available are homology modelling (using a template structure with a

similar sequence), threading (using a template structure of similar fold) and *ab initio* structure prediction (no template), with the latter set to supplant the others.

2.11.1 Homology modelling

The unknown structure of a target protein can be predicted with the help of an experimentally determined structure of a homologous protein. This template structure is found through the alignment of the sequence of the target protein with sequences of known structures in the PDB. Their sequence identity should ideally be above 30%, especially for shorter proteins. Multiple sequence alignment can help fine-tune the correspondence of residues in the template structure with residues in the target sequence, allowing the first residues of the model to be built. Gaps and insertions are often found in loops and need to be modelled, either *ab initio* or using template loops. Side chain conformations can be borrowed from the template structure, from other short templates or from single residues in high-resolution structures. The model is then optimised using molecular dynamics methods (see section 2.12).[175]

2.11.2 *Ab initio* structure prediction

In the absence of an available template structure, one can rely on physical interactions (e.g., running simulations as covered in section 2.12) and on evolutionary history. The latter is a more recent approach that uses, for example, insights from multiple sequence alignments to infer pairs of residues that coevolve and therefore ought to be close to each other.[176] Despite continued improvements and growth of available data, *ab initio* methods have historically been the most challenging ones for structure prediction. Their progress can be tracked from the results of the community experiment CASP, the critical assessment of structure prediction. Every two years, participating groups are challenged to use sequence information to compute the structures of around 100 targets for which the experimental structures are not yet public. The predictions are then independently assessed for accuracy.[177] In 2014, the first significant structure was accurately predicted without a template by Rosetta thanks to coevolutionary restraints.[178] As these contact predictions gradually improved, so did the accuracy of the modelling.[179] A significant breakthrough happened in 2018, as AlphaFold successfully used deep neural networks to obtain prediction accuracies far beyond any other methods, especially for difficult targets.[180, 181] The next CASP in 2020 saw another leap in accuracy thanks to

AlphaFold2 (Fig. 2.7), an entirely redesigned version which achieved atomic accuracies that rivalled experimental structures for at least two thirds of the targets.[176, 177]

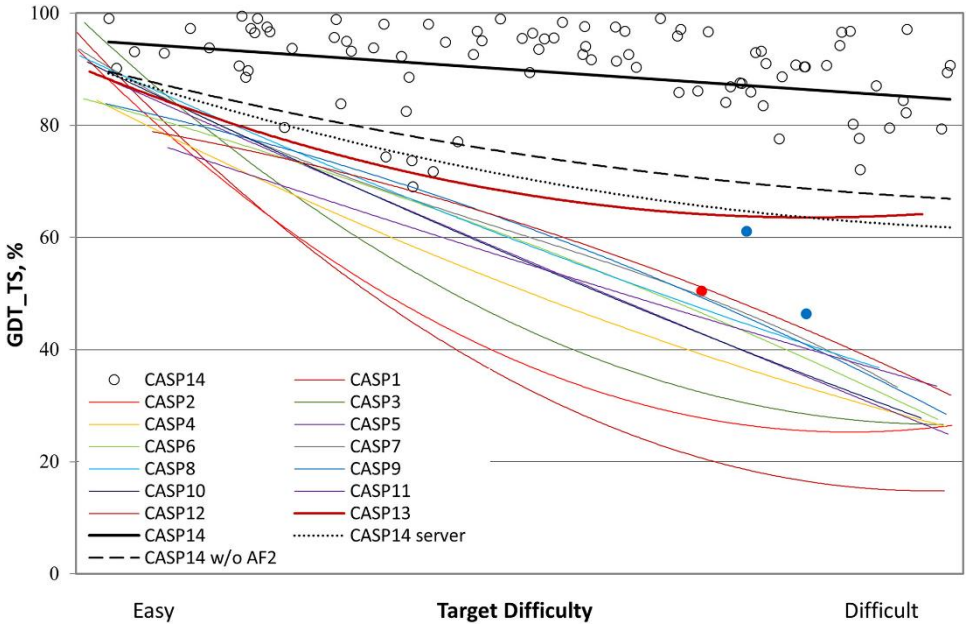


Figure 2.7: Accuracy of the best structure predictions at the CASP experiments. GDT_TS evaluates the agreement between the backbones of the predicted and experimental structures. Values above 50% typically indicate a correct fold prediction while anything above 90% is competitive with experimental structures. Target difficulty is calculated based on similarity to available structures. The best predictions from CASP14 are shown as circles. From the 1st to the 12th round, the trendlines show that the overall improvement in accuracy slowed down. For the 13th and 14th rounds however, AlphaFold and AlphaFold2 were extraordinarily successful in predicting even the most difficult structures to accuracies never achieved before. Figure reproduced from Kryshtafovych *et a.*[177] with permission.

2.12 Molecular dynamics

Molecular dynamics simulation is a wide-spread tool for the theoretical study of macromolecules in a realistic solvent environment,[182] thanks to its ability to model the events at microscopic scales of length and time, allowing prediction or validation of experimental results. It consists in solving the classical equations of motion for each particle (e.g., atom) at each time-step of the simulation. The forces acting on an atom can be calculated by deriving a potential energy function, which is the sum of individual potential energy terms from both bonds and non-bonded interactions.[183] The terms for bonding potential are the potential energy of bonds, bond angles and torsion angles. A bond length (e.g, r_{23} in Fig. 2.8) has a potential energy minimum at an ideal length and a

higher energy when it deviates from it. The same goes for the ideal bond angle (e.g, θ_{234} in Fig. 2.8). Torsion angles (ϕ_{1234} in Fig. 2.8) have higher energy when the end atoms clash (e.g, atoms 1 and 4 are the closest when the ϕ_{1234} angle is 180°) and reach an energy minimum when they are furthest apart (here, an angle of 0°). For non-bonded interactions, one potential energy term is the Van der Waals interaction, which is repulsive at very short distance, attractive at the optimum and then rapidly less favourable as the distance increased. The second essential term for non-bonded interaction aims to model the electrostatic interaction between charged atoms, which remains significant over longer distances.

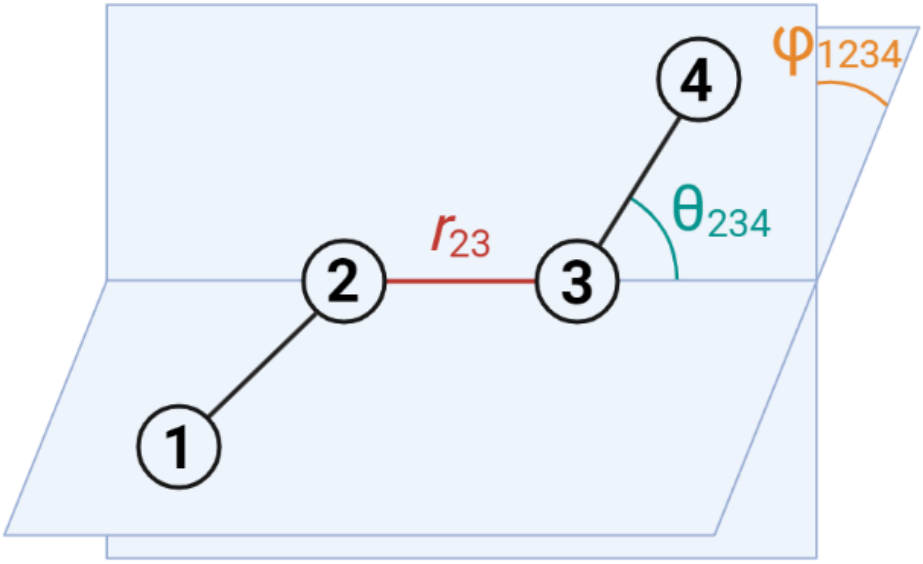


Figure 2.8: Bond lengths and bond angles. Geometry of a 4-atom chain with bond length (r_{23}), bond angle (θ_{234}), and torsion angle (φ_{1234}), the potential energy of which is calculated for MD simulations. Figure adapted from Allen[183] using Biorender.com.

The potential energy function contains many constants, the values of which need to be set in advance to form a set of parameters. These parameters can be derived from more complex and accurate calculations that use quantum mechanics calculations, and can then be validated by comparison of simulation results with experimental ones. Some parameters, such as ideal bond lengths and angles, are obtained by experimental methods that use spectroscopy or diffraction.[184] A set of parameters together with a potential energy function constitute a force-field, several of which are available to researchers, for example GROMOS, CHARMM, OPLS and AMBER.[185] These force-fields also differ in how they model water: GROMOS typically uses Simple Point Charge (SPC) water models based on the tetrahedral geometry seen in ice, while the other three use a Transferrable Interaction Potential 3-Point (TIP3P) model based on the smaller bond angle seen in the gas

phase.[186] They are used to fill the simulated box in which the macromolecule to be studied was placed. No matter the size of the box, a significant proportion of these water models will be near its surface, creating a special environment that is not representative of reality. To solve this issue, one can use periodic boundary conditions, which consists in surrounding the box with exact copies of itself on all sides. A molecule near the edge of the box can therefore interact with the nearby image of a water model that is actually on the opposite side of the box. If that image enters the box, it will be compensated by its corresponding water leaving the box on the other side, which keeps the same number of water models in the box and provides continuity to the solvent environment. It is impossible to calculate the energy of every non-bonded interaction within this infinite array of boxes. One therefore uses cut-off distances to determine which interactions should be computed for each atom.[183]

Chapter 3: AsAAT1

3.1 Introduction

The gene for AsAAT1 is on the extended biosynthetic gene cluster for avenacin, an estimated 0.66 centimorgan (cM) away from the genes that always co-segregate with each other (i.e., 0.0 cM).[39] The purpose of this clustering may be to facilitate co-adaptation and inheritance of the genes necessary for the production of avenacin, as well as co-regulation at the chromatin level to help restrict expression to the root tip.[35] It is there, likely in the cytosol, that AsAAT1 catalyses the first glycosylation step in avenacin biosynthesis: the addition of its arabinose group.[39] Only two other arabinosyltransferases are known to be involved in plant natural product biosynthesis: GmSSAT1 from soybean (*Glycine max*), part of the biosynthetic pathway for soyasaponin Ab,[39] and UGT78D3 from *Arabidopsis thaliana,* which glycosylates flavonols.[187] All three enzymes use a UDP-activated arabinopyranose (UDP-Ara) over other UDP-sugars.[39, 187] AsAAT1 is from the UGT99 family and GmSSAT1 from the UGT73 family, both in group D, where most triterpene UGTs are found. AtUGT78D3, however, is in group F with a flavonoid glucosyltransferase.[39, 188] A common feature of these arabinosyltransferases is the presence of a histidine residue at the C-terminal end of their PSPG motif. Homologous glucosyltransferases, such as AsUGT99A6, the closest relative of AsAAT1, have a glutamine residue at this position.[39] It is not clear how the mutation to histidine, which has arisen independently thrice, causes a change in sugar specificity.

This chapter reports efforts to obtain structural information for AsAAT1 to rationalise this effect. Initially, this involved attempts to solve the enzyme's structure by X-ray crystallography, but as crystallisation proved unsuccessful, the aim was instead achieved through a molecular modelling approach.

3.2 Methods

For the purpose of structure solution by X-ray crystallography, several constructs of AsAAT1 were recombinantly expressed in *E. coli* for crystallisation screening.

3.2.1 Crystallization trials of AsAAT1 with a non-cleavable N-terminal 9xHis-tag (9xHis-AsAAT1)

3.2.1.1 Expression and purification of 9xHis-AsAAT1

Production of AsAAT1 (AsUGT99D1) for crystallisation experiments was initially attempted by following the published strategy.[39] *E. coli* Rosetta 2(DE3) cells transformed with the plasmid pH9GW-AsAAT1 were obtained from Thomas Louveau (John Innes Centre). The vector pH9GW is a derivative of the pET-28a vector (Novagen) which is suitable for Gateway cloning, confers kanamycin resistance and adds nine histidine codons upstream of the cloned gene.[189] A single colony was cultured in 50 mL liquid Lysogeny Broth (LB) media under 50 µg/mL kanamycin selection overnight at 37 °C. Aliquots of 2.5 mL of this preculture were used to inoculate two 1 L baffled conical flasks with 500 mL LB media and 50 µg/mL kanamycin, which were shaken at 37 °C, 200 rpm. After 4 h, the culture reached an OD_{600} of 0.6 and production of recombinant enzyme was induced by addition of IPTG to a final concentration of 0.2 mM before being incubated overnight at 16 °C. Bacterial cells were harvested by centrifugation at 8,950 *g* and resuspended in 15 mL buffer A1 (50 mM Tris HCl pH 7.5, 300 mM NaCl, 20 mM imidazole) with one tablet of AEBSF protease inhibitor (ThermoFisher), 1 mM EDTA, 1 mg/mL lysozyme and DNAse (bovine). The cells were lysed by passage through a French press at 1,000 psi three times using a pressure cell cooled to 4 °C. The cell debris were removed by centrifugation at 19,000 rpm and 4 °C for 30 min. The soluble fraction was enriched for the His-tagged recombinant AsAAT1 by immobilised metal ion affinity chromatography (IMAC) using a nickel-agarose resin column (HisTrap HP 1 mL, GE Healthcare) equilibrated with 10 mL buffer A1 and eluted with increasing imidazole concentration (20 to 500 mM). Fractions containing AsAAT1 as identified by SDS-PAGE were pooled and concentrated to a final volume of 2 mL using a 10 kDa Amicon filter unit (Millipore) at 4 °C. The concentrate was further purified by gel filtration at 4 °C through a HiLoad 16/600 Superdex 75 pg column at a flow rate of 0.4 mL/min with a buffer consisting of 50 mM Tris HCl pH 7.5, 300 mM NaCl and 1 mM dithiothreitol (DTT). Fractions showing pure protein of the predicted MW by SDS-PAGE were pooled and concentrated by centrifugation at 4,000 rpm and 4 °C in a 10 kDa Amicon filter unit (Millipore). Protein concentrations were calculated using absorbance at 280 nm together with the calculated MW and extinction coefficient for AsAAT1 with 9xHis-tag. These were estimated to be 56,152 Da and 59.93 $mM^{-1}cm^{-1}$ using the exact amino acid sequence and the ExPASy ProtParam online server.[190] Once the concentration

reached 5.3 mg/mL, any further reduction in volume would not increase the concentration further and the protein was considered to have reached its solubility limit in these conditions.

3.2.1.2 Crystallisation trials of 9xHis-AsAAT1

Crystallisation screening experiments were carried out at 16 °C with six commercially available screens: Structure Screen™ 1 and 2 Eco Screen,[191] JCSG-plus™ Eco Screen,[192] PACT premier™ Eco Screen,[193] Morpheus® Screen,[194] MIDAS™ Screen,[195] and LMB Crystallization Screen™, all from Molecular Dimensions. The first two screens were also set with dispase, V8 or chymotrypsin spiking at 1:100 and 1:500 (w/w) protease:protein ratio. All protease were obtained from Sigma and diluted to 0.25 mg/mL. Dispase was diluted with 17 mM Tris HCl pH 7.5 and 100 mM NaCl, V8 protease with dH_2O and chymotrypsin with 1 mM HCl. For each protease, either 1 µL (for 1:500) or 5 µL (1:100) were added to 25 µL 9xHis-AsAAT1 (5 mg/mL). The screens were set up in 96-well 2-drop MRC plates sealed with ClearVue Sheets (Molecular Dimensions) employing an OryxNano protein crystallisation robot (Douglas Instruments Ltd.). The sitting drop vapour diffusion technique was used with a drop size of 0.5 µL containing the protein and screen solution at either 1:1 or 1:2 ratio, equilibrated against 50 µL of screen solution per reservoir of the 96-well plate. The screening plates were monitored for crystal formation using a SZX12 Stereo Microscope (Olympus). Crystals were harvested using mounted LithoLoops (Molecular Dimensions) and transferred to drops of well solution with added 30 % v/v glycerol for cryoprotection before flash-freezing in liquid nitrogen. Crystals were sent in a dry shipper to Diamond Light Source, beamline i04, for X-ray diffraction experiments.

3.2.2 Crystallization trials of de-tagged AsAAT1

As 9xHis-AsAAT1 failed to produce diffracting crystals, it was decided to attempt the crystallisation of tag-free recombinant AsAAT1 in case the His-tag was impeding crystallisation.

3.2.2.1 Recloning of AsAAT1 with a cleavable N-terminal His-tag for the production of tag-free enzyme

This strategy involved recloning the AsAAT1 gene into the Gateway vectors pDEST17 and pH9GW, with addition of a HRV 3C protease cleavage site between the gene and the His-tag. To achieve this, plasmid pH9GW-AsAAT1 (obtained from Thomas Louveau,

John Innes Centre) was propagated in *E. coli* DH5α SC (see section 2.6 for transformation protocol) then used as a template to amplify the AsAAT1 cDNA. To add the 3C protease cleavage site and the first part of the *att*B2 Gateway cloning site, the following gene-specific primers were used (gene specific region unformatted, 3C protease cleavage site in italics, overlap with the *att*B2 cloning site in blue):

AT-GW-F 5'-*CTGGAAGTTCTGTTTCAGGGCCCG*ATGGGGAAACCAGCAGC-3'

AT-GW-R 5'-CAAGAAAGCTGGGTTTTAGATGGTGAAGCGTTGGTTGGACGAGGTC-3'

This first stage PCR was carried out using Phire Hot Start II DNA polymerase (Thermo Fisher) with 3% (v/v) DMSO using the following program: 98 °C for 3 min, followed by 32 cycles of 98 °C for 10 s, 58 °C for 20 s and 72 °C for 30 s, followed by 72 °C for 5 min. The presence of a product of the correct size was confirmed by 0.8% (w/w) agarose TAE gel electrophoresis with ethidium bromide staining, and the resulting fragments were purified using a QIAquick® PCR purification kit (Qiagen). This purified PCR product was used as a template for the second stage PCR reaction, which added an *att*B1 site (in red) upstream of the gene and completed the *att*B2 site (in blue) downstream of the gene by using the following general primers:

attB1-3C-F 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTCCTGGAAGTTCTGTT-3'

attB2-R 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTT-3'

This second stage PCR was carried out using the same protocol as for the first stage. The resulting PCR reaction was used as is to perform the BP clonase reaction to generate plasmid pDONR207-3C-AsAAT1. This reaction used 1 µL of BP clonase enzyme mix (Invitrogen) in a 4 µL reaction incubated at 25 °C for 16 h, before termination by addition of 1 µL proteinase K mix (Invitrogen) followed by incubation at 37 °C for 10 min. The resulting BP reaction mix was used to transform *E. coli* DH5α LE cells by heat shock before growing them overnight at 37 °C on selective LB agar plates (20 µg/mL gentamycin). The presence of the gene was confirmed in one of the resulting colonies by colony PCR (see section 2.6.2.1 for the protocol). This colony was grown in 10 mL LB (20 µg/mL gentamycin) at 37 °C overnight and the plasmid pDONR207-3C-AsAAT1 extracted using the QIAprep Spin Miniprep Kit (Qiagen). Sequencing confirmed the presence of the gene with one conservative mutation.

Plasmid pDONR207-3C-AsAAT1 was used as an entry clone to transfer the 3C-AsAAT1 gene into the destination vectors pH9GW and pDEST17 to generate the expression plasmids pH9GW-3C-AsAAT1 and pDEST17-3C-AsAAT1. This was achieved by using 0.5 µL of LR clonase enzyme mix (Invitrogen) in 5 µL reactions incubated at 25 °C for 10 h before termination with 1 µL proteinase K mix (Invitrogen) incubated at 37 °C for 10 min. The resulting LR reactions were used to transform *E. coli* DH5α LE cells, which were plated on selective LB agar (100 µg/ml carbenicillin for which pDEST17 carries a resistance gene or 50 µg/mL kanamycin for pH9GW) and grown at 37 °C overnight. The presence of the AsAAT1 gene was confirmed by colony PCR. Verified transformant colonies were grown in 10 mL selective LB (100 µg/ml carbenicillin for pDEST17 or 50 µg/mL kanamycin for pH9GW) at 37 °C overnight. The plasmids were extracted using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions.

3.2.2.2 Expression screen of 6xHis-3C-AsAAT1 using a pDEST17-derived plasmid

The pDEST17-3C-AsAAT1 plasmid was transformed into the following *E. coli* strains: BL21(DE3), SHuffle® T7 (NEB), SHuffle® T7 Express (NEB), ArcticExpress(DE3) RP (Agilent), and Rosetta-gami™ 2(DE3) (Novagen). The whole volume of each transformation was plated on Lysogeny Broth (LB) agar with carbenicillin (100 µg/mL) to select for the pDEST17-derived plasmid and with added tetracycline (10 µg/mL) for Rosetta-gami™ 2(DE3) or gentamycin (20 µg/mL) for ArcticExpress(DE3) RP. The plates were incubated at 37 °C overnight. A single transformant colony for each strain was used to inoculate 50 mL LB with appropriate antibiotics, which was then incubated at 37 °C and 180 rpm overnight. The resulting pre-cultures were used to inoculate 22 mL each into 500 mL LB (100 µg/mL carbenicillin) in 2 L baffled conical flasks, which were incubated at 37 °C and 200 rpm. When the cultures reached an OD_{600} of 0.5-0.6, they were moved to 16 °C and after 30 min of acclimatisation, recombinant protein expression was induced overnight by adding IPTG to a final concentration of 0.02 or 0.3 mM. The cells were harvested by centrifugation at 8,950 *g* and 16 °C for 35 min.

3.2.2.3 Small-scale purification of 6xHis-3C-AsAAT1 from expression screen

The cells harvested from 5 mL aliquots of culture were resuspended in 1 mL lysis buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 20 mM imidazole, 1 mM EDTA, 10% v/v glycerol, 1% v/v Tween™ 20, 1 mg/mL lysozyme, cOmplete protease inhibitor) and incubated at 30 °C and 220 rpm for 45 min. After adding 37.5 U of Benzonase® to each lysis

reaction, they were incubated a further 20 min. Further lysis was achieved by sonication on ice (10 s on then 10 s off, repeated 3 times). Small-scale purification was carried out at 4 °C. The insoluble fraction was removed by centrifugation at 13,000 rpm for 20 min. Ni-NTA Superflow (Qiagen) resin beads (100 µL) were equilibrated with 200 µL wash buffer (50 mM Tris-HCl pH 8.0, 500 mM NaCl, 20 mM imidazole, 10% v/v glycerol, 1% v/v Tween™ 20) by mixing. The supernatant was removed by centrifugation at 2,500 rpm for 1 min. The resin was then incubated with 1 mL of soluble fraction from the lysate at 800 rpm for 1 h. After centrifugation at 2,500 rpm for 1 min, the supernatant was discarded. The beads were washed twice with lysis buffer then twice with wash buffer (with washing consisting of addition of 400 µL buffer, mixing to evenness, centrifugation at 2,500 rpm for 1 min then discarding 400 µL of supernatant). The bound proteins were eluted by incubating 50 µL elution buffer (wash buffer with 1 M imidazole instead of 20 mM) for 12 min at 800 rpm then centrifugation at 3,000 rpm for 3 min. The supernatant was then analysed by running denatured samples on SDS-PAGE.

3.2.2.4 Expression of 9xHis-3C-AsAAT1 using a pH9GW-derived plasmid

Expression screening with the pDEST17-3C-AsAAT1 plasmid did not yield sufficient protein to warrant scaling-up, but small-scale purification revealed BL21(DE3) to be the most promising expression strain. It was therefore chosen for an attempt at large-scale expression of AsAAT1 with cleavable His-tag using a pH9GW-derived plasmid. The pH9GW-3C-AsAAT1 plasmid was transformed into *E. coli* BL21(DE3) cells. The whole volume of each transformation was plated on Lysogeny Broth (LB) agar with kanamycin (50 µg/mL). The plates were incubated at 37 °C overnight. A single transformant colony was used to inoculate 10 mL LB with kanamycin (50 µg/mL) and incubating at 37 °C and 180 rpm overnight. The resulting culture was used to prepare a 25% (v/v) glycerol stock to be stored at -80 °C. This glycerol stock was used to make a preculture by inoculating it into 100 mL LB with kanamycin (50 µg/mL) to be incubated at 37 °C and 180 rpm overnight. For large-scale expression, 8 x 500 mL LB with kanamycin (50 µg/mL) in 2 L baffled conical flasks were inoculated with 12 mL of preculture each. They were incubated at 37 °C and 200 rpm to an OD_{600} of 0.4 and then moved to 16 °C. After 45 min of acclimatisation, recombinant protein expression was induced overnight by adding IPTG to a final concentration of 0.05 mM. The cells were harvested by centrifugation at 8,950 *g* and 16 °C for 30 min, then either lysed immediately or flash frozen then stored at -80 °C for later use.

3.2.2.5 Purification of de-tagged AsAAT1

The harvested BL21 cells were resuspended in 30 mL AG binding buffer (buffer A1 with 5% (v/v) glycerol to attempt improving protein solubility)[196] with an added 30 mg lysozyme, 30 mg bovine DNAse, 1 tab of cOmplete protease inhibitor, then subjected to three cycles of cell lysis using a French press (Thermo) set to 1,000 psi. The soluble fractions were separated from cell debris by centrifugation at 5,100 *g* and 4 °C for 75 min and further clarified using 0.45 um syringe filters.

The clarified soluble fractions were first purified by nickel ion affinity chromatography using a 1 mL HisTrap HP column (GE Healthcare) at a flow rate of 1.4 mL/min on an ÄKTA Pure chromatography system (GE Healthcare) at 4 °C. The column was pre-equilibrated with 10 CV of AG binding buffer, before loading the soluble fractions, then washing with 20 CV of binding buffer. The bound proteins were eluted with a gradient of 20 to 500 mM imidazole, resulting from the mixing of binding buffer with a proportion of BG elution buffer (B1 with 5% (v/v) glycerol) increasing from 0% to 45% over 40 CV, then to 100% over 10 CV, with a final 10 CV step at 100%. The eluates were collected in 2 mL fractions, which were assessed by running denatured samples on SDS-PAGE.

The fractions of the eluate that contained any protein were pooled. To prepare for tag-cleavage with 3C protease, the imidazole was removed with a HiPrep 26/10 desalting column (GE Healthcare) at a flow rate of 15 mL/min using an ÄKTA Pure chromatography system (GE Healthcare) at 4 °C. Elution with AG binding buffer lacking imidazole was collected in 2 mL fractions, 10 of which were pooled as they showed UV absorbance on the chromatogram and eluted before any significant change in conductivity. The pooled protein fractions were incubated with 16.67 µL of 3 mg/mL His-tagged recombinant HRV 3C protease with slow stirring at 4 °C overnight. The precipitate that formed was removed by centrifugation at 5,100 *g* and 4 °C for 15 min.

To remove the cleaved His-tag, the uncleaved protein, the 3C protease and the nickel-binding contaminants from the first nickel ion affinity chromatography, a second one with the same parameters was run to recover the further purified cleaved protein in the flow-through, the presence of which was confirmed by running denatured samples on SDS-PAGE.

For the third purification step by SEC, the fractions containing the cleaved protein were first pooled and concentrated to 1.5 mL at 4 °C with an Amicon® Ultra 15 mL

Centrifugal Filter (10 kDa MWCO, Merck). Precipitate was removed by centrifugation at 13,000 rpm and 4 °C for 10 min. The sample was then loaded on a HiLoad 16/600 Superdex 75 pg column (GE Healthcare) pre-equilibrated and eluted at a flow rate of 0.4 mL/min with SEC buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 5 mM DTT). The elution was collected in 2 mL fractions and results were assessed by running denatured samples of the peak fractions on SDS-PAGE. The fractions containing pure protein were pooled and concentrated at 4 °C with an Amicon® Ultra 4 mL Centrifugal Filter (10 kDa MWCO, Merck). The protein concentration was measured regularly with a NanoDrop™ Spectrophotometer (Thermo Scientific) using absorbance at 280 nm. When a reduction in volume stopped causing an increase in protein concentration, it was considered to have reached its solubility limit. This was calculated to be at 5.8 mg/mL based on the protein's molecular weight and extinction coefficient estimated by the ExPASy ProtParam tool[190] based on the protein sequence.

3.2.2.6 Extraction of the insoluble fraction of lysate

In an attempt to increase the yield of soluble AsAAT1 from the lysate, a high-salt extraction was attempted. This consisted in resuspending the insoluble fraction in 50 mL high-salt AG buffer (1 M potassium phosphate pH 7.5, 20 mM imidazole, 5% v/v glycerol) before centrifugation at 5,100 *g* and 4 °C for 75 min. The resulting supernatant was then analysed with a NanoDrop™ Spectrophotometer (Thermo Scientific) using absorbance at 280 nm and by SDS-PAGE.

In a detergent extraction approach, a fresh insoluble fraction following 9xHis-3C-AsAAT1 overexpression was resuspended in AG buffer with 1 mini-tab cOmplete EDTA-free and 1% (v/v) Triton™ X-100 using a glass homogeniser. The cell debris was removed by centrifugation at 5,100 *g* and 4 °C for 75 min and the remaining membrane components were removed by ultra-centrifugation at 257,000 *g* and 4 °C for 1 h. The clarified soluble fraction was loaded on a 1 mL HisTrap HP column (GE Healthcare) at a flow rate of 1.4 mL/min on an ÄKTA Prime chromatography system (GE Healthcare) at 4 °C. The column was pre-equilibrated with 10 CV of AGT binding buffer (AG buffer with 0.1% v/v Triton™ X-100), before loading the soluble fractions, then washing with 10 CV of AGT binding buffer. Any bound proteins were eluted with a gradient of 20 to 500 mM imidazole, resulting from the mixing of AGT binding buffer with a proportion of BGT elution buffer (BG with 0.1% v/v Triton™ X-100) increasing from 0% to 100% over 40 CV, with a final 10 CV step at 100%.

3.2.2.7 Crystallisation trials of de-tagged AsAAT1

Crystallisation screening experiments were carried out with tag-free AsAAT1 freshly purified and concentrated to 3.3 mg/mL, as 5 mg/mL yielded excess precipitation in a limited (12-condition) trial. These experiments were performed at 16 °C with 10 commercially available screens: the same six screens described in section 3.2.1.2 as well as PEG/Ion Screen™, Index™, PEGRx™, SaltRx™, all from Hampton Research. Structure Screen™ 1 and 2 Eco Screen, JCSG-plus™ Eco Screen , PACT premier™ Eco Screen and MIDAS™ Screen were also set at 4 °C. All screens were set up in 96-well 2-drop MRC plates (SwisSCI) sealed with ClearVue Sheets (Molecular Dimensions) employing an OryxNano protein crystallisation robot (Douglas Instruments Ltd.). The sitting drop vapour diffusion technique was used with a drop size of 0.5 µL containing the protein and screen solution at either 1:1 or 1:2 ratio, equilibrated against 50 µL of screen solution per reservoir of the 96-well plate. The screening plates were monitored for crystal formation using a SZX12 Stereo Microscope (Olympus).

A microseeding experiment was performed from a crystallisation condition that yielded several small crystals in the hope of obtaining larger ones. Additional screen solution (0.2 M calcium chloride dihydrate, 0.1 M sodium HEPES pH 7.5 and 28% (v/v) PEG 400) was first added to the crystal-containing drop to verify it did not dissolve the crystals. The drop was then transferred to a PTFE Seed Bead™ (Hampton Research) tube containing 50 µL of screen solution. This was shaken at 1000 rpm for a total of 30 s at 4 °C which yielded seeds of the appropriate size, as observed under the microscope. Two successive 10-fold dilutions were made using 50 µL of seed suspension and 450 µL of screen solution. Seeding experiments were set up in triplicate using 0.6 µL drop sizes, the same ratio of 1:2 (v/v) protein to screen solution (including seeding solution) and with either 1:7 or 1:4 (v/v) seeding solution to screen solution. This experiment was also performed with the reservoir diluted by adding 1:2 (v/v) water to screen solution in an attempt to slow crystal growth. To potentially improve the crystals, the same condition was also optimised by screening on a 5 x 5 grid that systematically alters the pH (7.5 ± 0.2 or 0.4) and/or the concentration of PEG 400 (28 ± 2 or 4%), with or without UDP (10:1 molar ratio with AsAAT1).

3.2.3 Production of the shortened construct of AsAAT1 (9xHis-3C-AsAAT1-ΔN)

To improve the chance of crystallisation, the construct AsAAT1-ΔN was designed by removing the 21-residue N-terminal region predicted to be disordered by DISOPRED[76] (as described in section 2.1).

3.2.3.1 Cloning of 9xHis-3C-AsAAT1-ΔN

The plasmid pH9GW-AsAAT1 was used as template to amplify the shortened version of the AsAAT1 gene with AsAAT1-ΔN-specific primers that add the 3C protease cleavage site and the first part of the *att*B2 Gateway cloning site (gene specific region unformatted, 3C protease cleavage site in italics, overlap with the *att*B2 cloning site in blue):

AT-ΔN-GW-F 5'-*CTGGAAGTTCTGTTTCAGGGCCCG*CGTGCGCACTTTGTGTTC-3'

AT-GW-R 5'-CAAGAAAGCTGGGTTTTAGATGGTGAAGCGTTGGTTGGACGAGGTC -3'

This first stage PCR was carried out using Phire Hot Start II DNA polymerase (Thermo Fisher): 98 °C for 3 min, followed by 32 cycles of 98 °C for 10 s, 48 °C for 20 s and 72 °C for 30 s, followed by 72 °C for 5 min. The second PCR reaction, cloning and transformation were performed essentially as for the full-length cleavable construct.

3.2.3.2 Expression and purification of 9xHis-3C-AsAAT1-ΔN using a pH9GW-derived vector

The plasmid pH9GW-3C-AsAAT1-ΔN, which consists of the AsAAT1-ΔN cDNA with a N-terminal 3C protease cleavage site in the pH9GW vector, was transformed into *E. coli* BL21(DE3) cells using the same method as for the full-length cleavable construct. The same expression method was also used apart from the IPTG concentration, which was 0.02 mM for this construct. Purification by IMAC was carried out using the same method as for the first purification step of the full-length construct, but without glycerol in the buffers.

3.2.4 Generation of a predicted structure of AsAAT1 bound to UDP-Ara

As all attempts to crystallise AsAAT1 had proved unsuccessful, it was decided to use an *in silico* approach to yield structural information. A prediction for the structure of AsAAT1 was provided by Thomas Louveau (John Innes Centre). It had been generated with

I-TASSER[197] using the crystal structure of *Medicago truncatula* UGT71G1 complexed with UDP-Glc as a threading template (PDB entry: 2ACW).[52] The resulting model contained a strained loop comprising residues Trp396 to Ser402 due to a 2-residue insertion relative to the template. To identify the most likely conformation for this loop, 20 loop models were generated using the MODELLER[198] plugin to Chimera.[199] The six loop conformations with the best scores in terms of estimated RMSD and overlap were used to generate models for the structure of the complex with UDP-Ara, based on the conformation of UDP-Glc found in PDB entry 2ACW.[52] The resulting draft complexes were relaxed using the molecular dynamics program GROMACS[200] and the force field 53a6.[201] The models were solvated in a cubic periodic box of SPC water molecules and subjected to 100 steps of energy minimization. The necessary forcefield parameters for UDP-Ara were based on those available for uridine, ATP and glucose in the 53a6 forcefield. Following this step, the optimal model was selected for analysis based on having the best QMEAN score[202] and no Ramachandran or rotamer outliers in the remodelled loop according to the structure validation service, MolProbity.[203]

3.3 Results and discussion

3.3.1 Expression, purification and crystallisation screening of AsAAT1

3.3.1.1 AsAAT1 with non-cleavable N-terminal His-tag

Initial attempts to produce sufficient AsAAT1 for crystallisation were based closely on a published strategy and used an existing expression vector.[39] The gene had been cloned into the pH9GW vector, which is a derivative of the pET-28a vector that is compatible with gateway cloning and adds an N-terminal 9xHis-tag. The resulting pH9GW-AsAAT1 plasmid had been transformed into the *E. coli* Rosetta 2 (DE3) protein expression strain (see section 2.7.2.2 for details). A streak of this transformant was obtained and used in an initial attempt to express AsAAT1, using the original protocol.

For the lysis of the cells to be more thorough, it was performed with a French press instead of a sonicator. The purification by immobilised metal ion affinity chromatography (IMAC) with imidazole gradient yielded samples that were heavily contaminated, most likely by endogenous *E. coli* proteins that bind nickel ions or AsAAT1 itself. The two foremost contaminant proteins were of approximately 70 kDa molecular weight as determined by SDS-PAGE analysis (data not shown). These may correspond to GlmS (66.8 kDa) and YfbG (74.2 kDa) as suggested by other authors.[204] This first purification step was optimised by halting the upwards gradient of imidazole concentration at 32 mM, washing the main contaminants out, then continuing the increase in concentration (Fig. 3.1). While the purity of the AsAAT1-containing fraction was thus improved, it was still insufficient for crystallisation screening: a major contaminant seemed to stick to AsAAT1, and several minor contaminants of lower-molecular weights eluted at various imidazole concentrations, as seen by SDS-PAGE (Fig. 3.2).

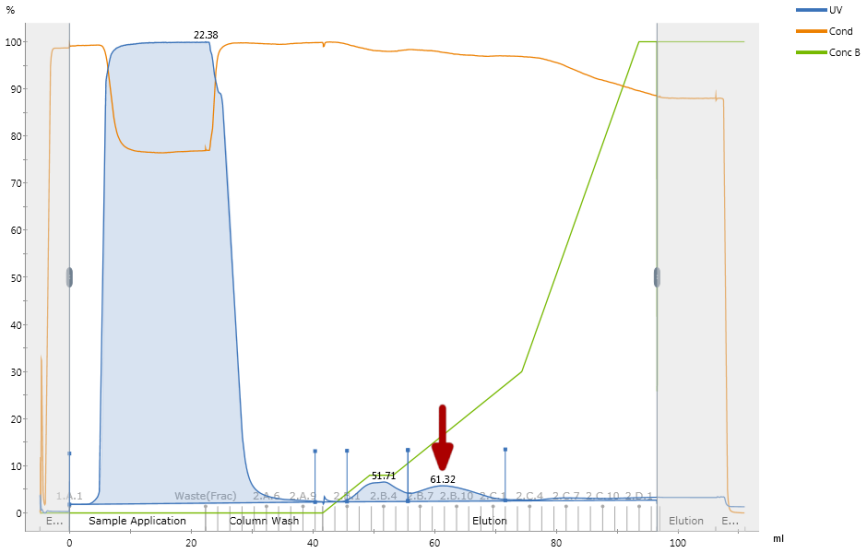


Figure 3.1: Chromatogram of the first stage IMAC purification of 9xHis-AsAAT1. The step inserted within the gradient at 32 mM imidazole helps resolve the two major peaks: a first one from contaminants, then a second one from 9xHis-AsAAT1, labelled with a red arrow. The blue curve represents absorbance at 280 nm. The green line represents the proportion of high-imidazole buffer from 0% to 100%. Fraction names are written in grey on the x-axis. Figure generated by Unicorn 7 (GE Healthcare) and edited with GIMP.

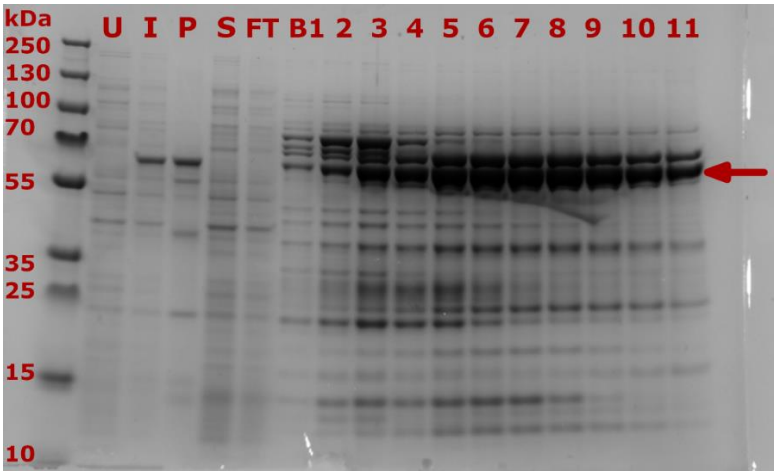


Figure 3.2: SDS-PAGE analysis of expression and first purification step of 9xHis-AsAAT1 (IMAC). 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights, the total protein of the uninduced culture (U) and final induced culture (I) which shows the appearance of the overexpressed protein; pellet of the lysate (P), i.e., the insoluble fraction; soluble fraction of the lysate (S) where any recombinant protein is obscured by background proteins; unbound proteins of the flow-through (FT); elution fraction numbers. The arrow indicates the expected migration distance of 9xHis-AsAAT1.

A second stage purification was therefore performed using SEC with the reductant dithiothreitol in the running buffer. This successfully separated all higher molecular weight contaminants and most of the others out of the AsAAT1-containing fractions, according to SDS-PAGE analysis (Fig. 3.3).

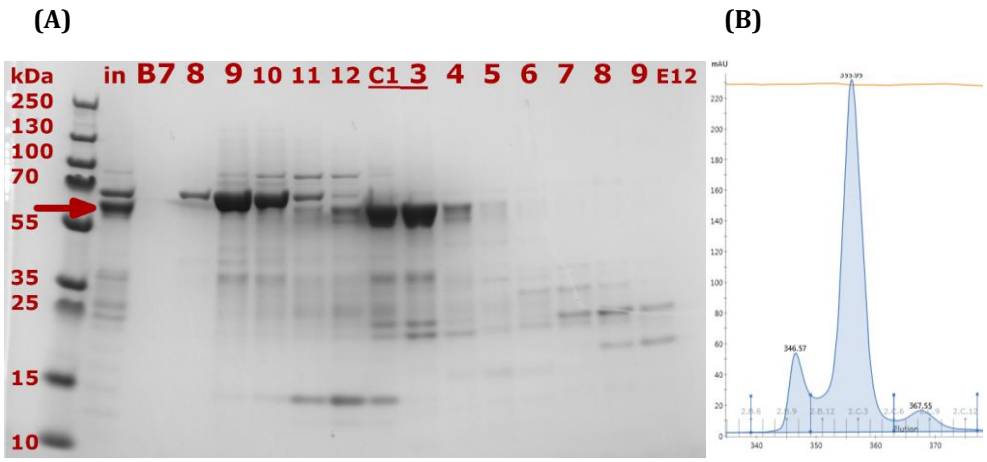


Figure 3.3: SDS-PAGE analysis and chromatogram of the second purification step of 9xHis-AsAAT1 (SEC). (A) 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights, the protein sample injected into the column (in) and the elution fraction numbers (with the pooled fractions underlined). The arrow indicates the expected migration distance of 9xHis-AsAAT1. **(B)** Chromatogram with fractions numbers written in grey on the x-axis and the blue curve representing absorbance at 280 nm. The large second peak corresponds to the elution of 9xHis-AsAAT1.

The purified AsAAT1 with N-terminal 9xHis-tag was concentrated to its maximum solubility at 5.3 mg/mL and subjected to crystallisation trials using six commercial screens, some with spiking with various proteases (V8, chymotrypsin and dispase) in the hope of forming a fragment *in situ* that would crystallise more easily.[205] Unfortunately, of the few crystals that appeared in the 2,304 conditions tested, none diffracted X-rays.

3.3.1.2 AsAAT1 with cleavable His-tag for the purification and crystallisation screening of tag-free enzyme

To increase the chances of crystallisation, it was attempted with tag-free AsAAT1. First, PCR was used to add a 3C protease cleavage site and Gateway adapters to the gene, which was cloned into the storage vector pDONR207. This 3C-AsAAT1 gene was then cloned into the gateway expression vectors pDEST17 and pH9GW, which allow IPTG-

inducible expression of N-terminal-His-tagged enzyme. The previously introduced 3C cleavage site allows the tag to be cleaved off the overexpressed protein.

We set out to screen for expression in all our available *E. coli* protein expression strains at two different IPTG concentrations. After lysis by sonication, the resulting soluble fraction had a strong background of endogenous proteins that obscured bands of recombinant protein by SDS-PAGE. It was therefore decided to perform small-scale nickel-affinity purifications to observe variations in expression levels of soluble His-tagged protein (Fig. 3.4). Overall, the lower IPTG concentration (0.02 mM) gave better results, as well as the strains BL21(DE3) and Arctic Express (DE3) RP.

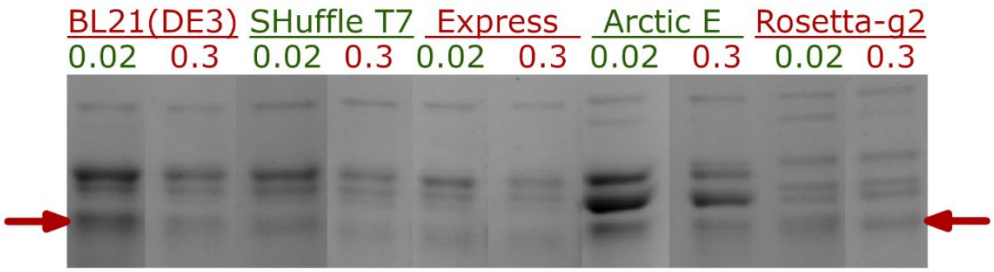


Figure 3.4: SDS-PAGE analysis of the small-scale purification of 6xHis-3C-AsAAT1 from an expression screen. 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. Composite of the 100-50 kDa region of the eluate lanes. From left to right, the 0.02 mM and 0.3 mM IPTG expressions in BL21(DE3) pLysS, SHuffle T7, SHuffle T7 Express, Arctic Express RP and Rosetta-gami 2. The arrows indicate the expected migration distance of 6xHis-AsAAT1.

E. coli BL21(DE3) was chosen for large-scale expression as it was the one of the two most promising strain from small-scale expression purification. The second promising strain, Arctic Express (DE3) RP, was less enticing because of its highly expressed chaperones, one of which (likely to be the 60 kDa Cpn60)[128] was of similar molecular weight as AsAAT1 and co-eluted during IMAC.

Large-scale expression using the pH9GW-3C-AsAAT1 plasmid in BL21(DE3) yielded sufficient soluble protein (Fig. 3.5) to proceed with the next stages of purification. The imidazole was removed with a desalting step to allow tag-cleavage by 3C protease. A second stage of IMAC let the de-tagged protein flow past the nickel ions to be collected in the flow-through. However, it retained the His-tagged protease, the His-tag, any remaining tagged-protein and other nickel-binding contaminants, separating them out (Fig. 3.6). There were a major and a minor contaminant remaining, which were successfully

separated out by SEC with reductant to yield approximately 2 mg of fully purified AsAAT1 (Fig. 3.7).

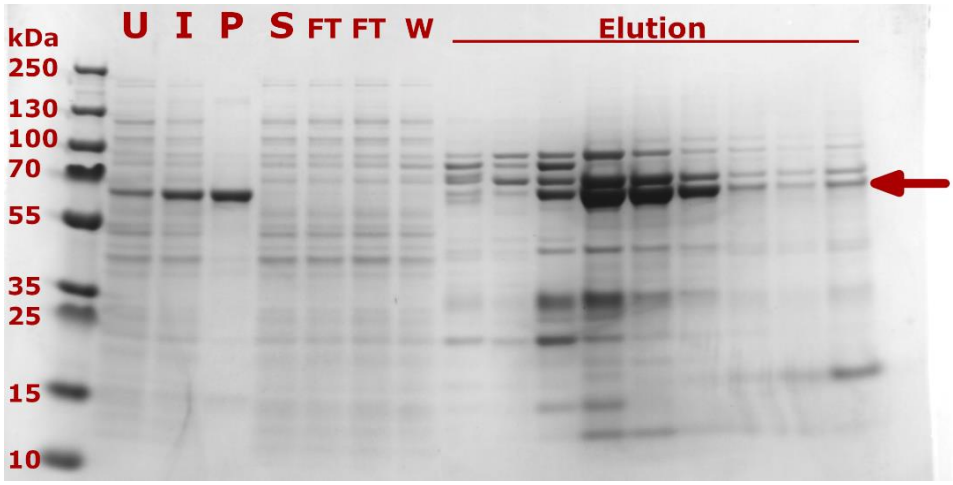


Figure 3.5: SDS-PAGE analysis of the first purification step of 9xHis-3C-AsAAT1 (IMAC1). 4-12% acrylamide SDS-PAGE gel with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated, the total protein of the uninduced culture (U) and final induced culture (I); pellet of the lysate (P), which is the insoluble fraction; soluble fraction of the lysate (S); unbound proteins of the flow-throughs (FT); column wash fraction (W); selected elution fractions. The arrow indicates the expected migration distance of 9xHis-AsAAT1.

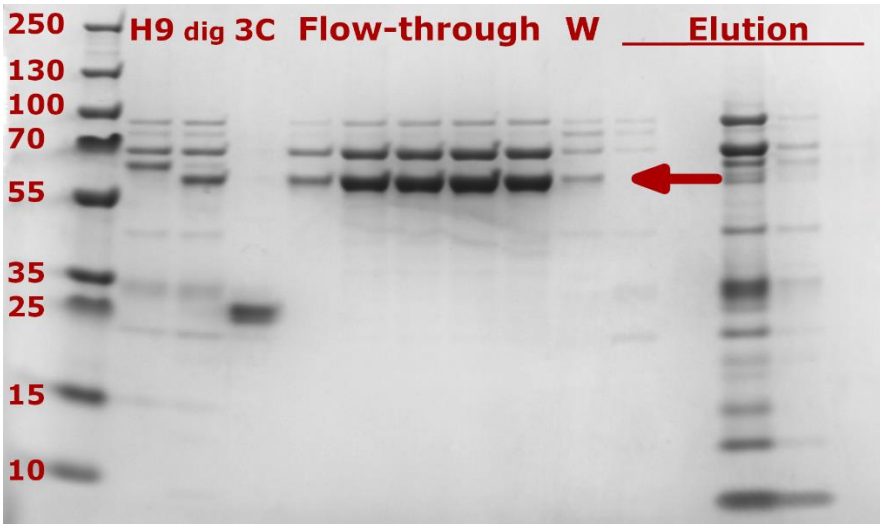


Figure 3.6: SDS-PAGE analysis of the second purification step of 9xHis-3C-AsAAT1 (IMAC2). 4-12% acrylamide SDS-PAGE gel with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated in kDa, the protein sample before (H9) and after (dig) tag-cleavage, the 3C protease, selected flow-through fractions, the subsequent column wash (W), selected elution fractions. The arrow indicates the expected migration distance of de-tagged AsAAT1.

(A) **(B)**

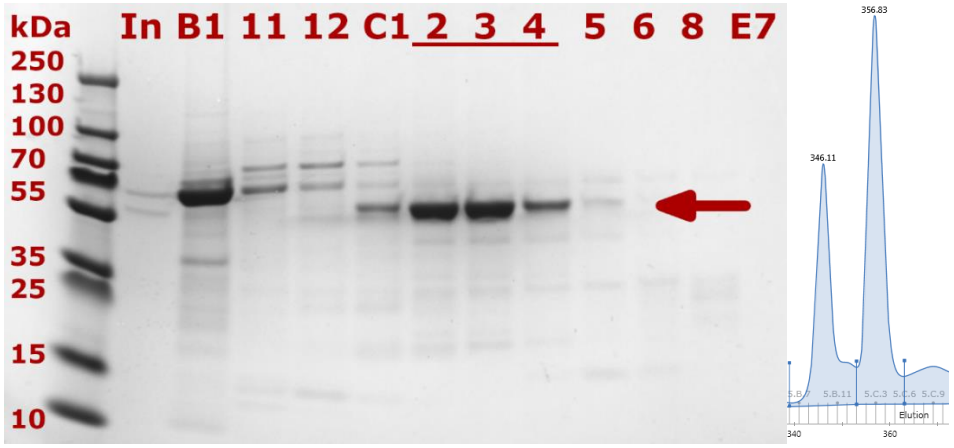


Figure 3.7: SDS-PAGE analysis and chromatogram of the third purification step of 9xHis-3C-AsAAT1 (SEC). (A) 4-12% acrylamide SDS-PAGE gel with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated, the protein sample injected onto the column (In) and the fractions numbers (with the pooled fractions underlined). The arrow indicates the expected migration distance of de-tagged AsAAT1. **(B)** Chromatogram with fractions numbers written in grey on the x-axis and the blue curve representing absorbance at 280 nm. The large second peak corresponds to the elution of de-tagged AsAAT1.

As a way of potentially improving the yield and because the vast majority of the produced protein was visible in the insoluble fraction of the lysate, we attempted to extract some of it back. First, we attempted a high-salt extraction,[206] but while the resulting extract contained hundreds of milligrams of protein, SDS-PAGE analysis revealed it only contained background protein and no visible 9xHis-3C-AsAAT1 (data not shown). For a second extraction strategy, we used the detergent Triton™ X-100, but the resulting soluble fraction did not show any proteins that would bind nickel ions, as seen on the chromatogram of the following IMAC step (Fig. 3.8).

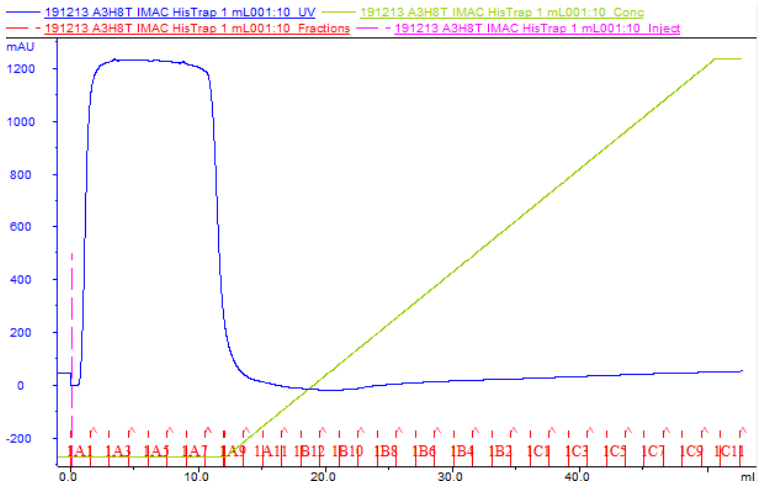


Figure 3.8: Chromatogram of the attempted purification of 9xHis-3C-AsAAT1 from detergent extraction (IMAC). The blue curve represents absorbance at 280 nm. The green line represents the proportion of high-imidazole buffer from 0% to 100%. Fraction names are written in red on the x-axis. Generated by an ÄKTA Prime chromatography system (GE Healthcare).

To ensure the purified protein did not rapidly degrade before crystallisation, the stability of the de-tagged AsAAT1 in our chosen buffer was evaluated by running various time-points from 4 °C or 16 °C storage on SDS-PAGE (Fig. 3.9). This showed that AsAAT1 was remarkably resistant to proteolysis.

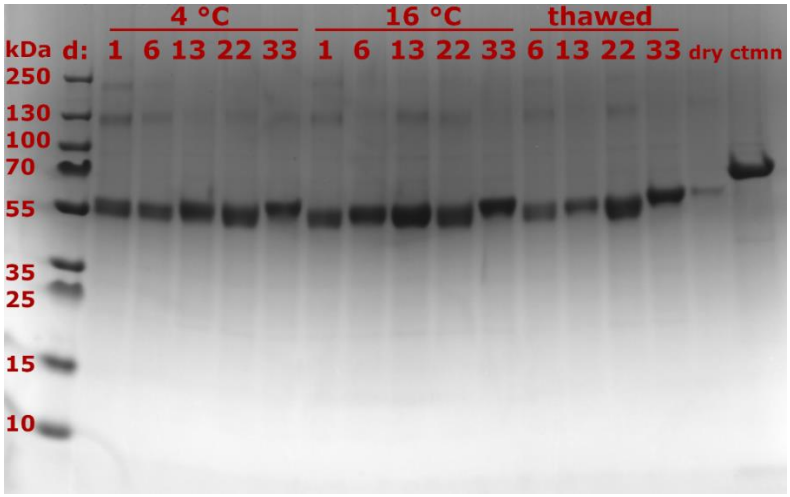


Figure 3.9: Degradation analysis of de-tagged AsAAT1 by SDS-PAGE. 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. The protein was concentrated to 3 mg/mL in SEC buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl). Samples were left for 1 to 33 days at 16 °C (including a dried sample in lane "dry") or at 4 °C with ("thawed") or without a freeze-thaw cycle. "Ctmn" refers to the 57 kDa contaminant separated by SEC and left at 4 °C for 33 days. There is no visible sign of proteolysis.

The purified tag-free AsAAT1 had a maximum solubility of 5.8 mg/mL. A small number of test crystallisation conditions were tried for a concentration of 5 mg/mL, but almost all of them resulted in precipitate. Therefore, a lower concentration of 3.3 mg/mL was used for our larger scale crystallisation screening efforts, which gave a better ratio of precipitate to clear drops (approximately 1:1). None of our 10 commercial crystallisation screens (1,920 conditions) produced crystals that would diffract, so we attempted the first three screens in the presence of the substrate UDP-Ara in 1.5:1 ratio with the enzyme. Unfortunately, it did not yield any useful crystals either. We attempted all 10 commercial screens with UDP present in 10:1 ratio with AsAAT1, but to no avail. We also attempted four crystallisation screens at 4 °C, which yielded one promising condition. This condition was optimised in a 5 x 5 grid of +/- 2 or 4 % (v/v) PEG and +/- 0.2 or 0.4 pH unit, but we did not obtain better crystals. We tried microseeding the conditions with various dilutions of broken crystals, but they did not grow.

3.3.1.3 AsAAT1 with shortened N-terminus (AsAAT-ΔN) and cleavable His-tag

To increase the chances of crystallisation, we designed a shortened construct, AsAAT1-ΔN, which lacked the residues at the N-terminus that were predicted to be disordered (see section 2.1). It was overexpressed using the plasmid pH9GW-3C-AsAAT1-ΔN in the same way as the full-length construct. Unfortunately, only very little soluble protein appeared by SDS-PAGE analysis of the IMAC eluate (Fig. 3.10). Further purification was therefore not carried out as the construct may not be soluble enough.

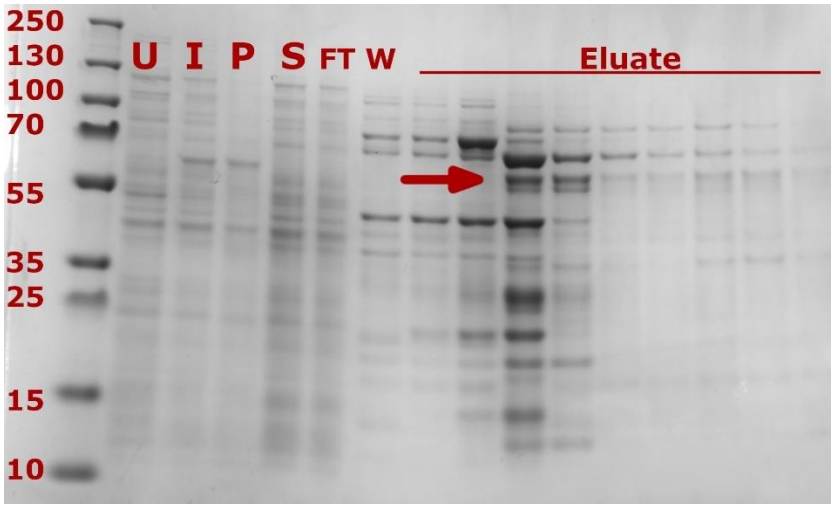


Figure 3.10: SDS-PAGE analysis of expression and first purification step of 6xHis-3C-AsAAT1-ΔN. 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated in kDa, the total protein of the uninduced culture (U) and final induced culture (I) which shows the appearance of the overexpressed protein; pellet of the lysate (P) which is the insoluble fraction; soluble fraction of the lysate (S) where any recombinant protein is obscured by background proteins; unbound proteins in the flow-through (FT); weakly-bound proteins in the wash (W); selected eluate fractions. The arrow indicates the expected migration distance of 6xHis-3C-AsAAT1-ΔN.

3.3.2 Prediction of the structure of AsAAT1 in complex with UDP-Ara

Despite extensive efforts to obtain crystals of AsAAT1 by screening over 8,000 conditions, crystallisation proved impossible. The strategy for obtaining structural information was therefore shifted to *in silico* methods.

A structure prediction had been generated using I-TASSER[197] based on the crystal structure of *Medicago truncatula* UGT71G1 complexed with UDP-Glc (PDB: 2ACW).[52] A significant difference between AsAAT1 and the threading template is the insertion of an arginine and a leucine residue in a loop near the active site (Fig. 3.11). This resulted in an unnaturally strained loop in the I-TASSER model, so it was remodelled using MODELLER (Fig. 3.12).[198]

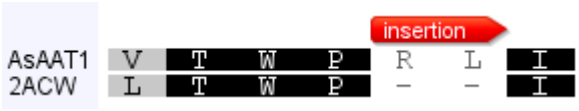


Figure 3.11: Alignment between the amino acid sequence of AsAAT1 and that of the homologue *M. truncatula* UGT71G1 (PDB: 2ACW). This shows the insertion of Arg398 and Leu399. Figure prepared using Geneious 9.0.5

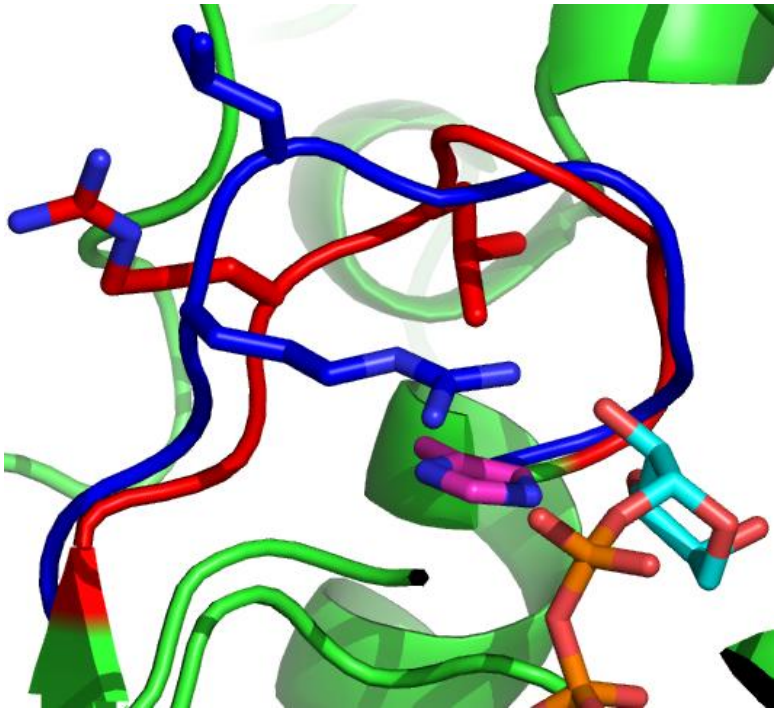


Figure 3.12: The active site loop of AsAAT1 was remodelled. The original I-TASSER[197] model (green cartoon) contained a strained loop (red) near the active site His404 (magenta sticks), so it was remodelled (blue). The two residues of the insertion (Arg398 and Leu399) are shown as sticks. The UDP-Ara ligand is shown as cyan sticks. Figure prepared using PyMOL.[46]

UDP-Ara was inserted into the model of AsAAT1 based on the conformation of UDP-Glc in the template structure. To achieve a more realistic model and resolve steric clashes, this draft complex was relaxed using the molecular dynamics program GROMACS. This required the design of forcefield parameters for UDP-Ara to allow energy minimisation. The resulting model shows that the residues key to sugar specificity, His404 and Pro154, are in close proximity to the arabinose moiety of UDP-Ara (Fig. 3.13).

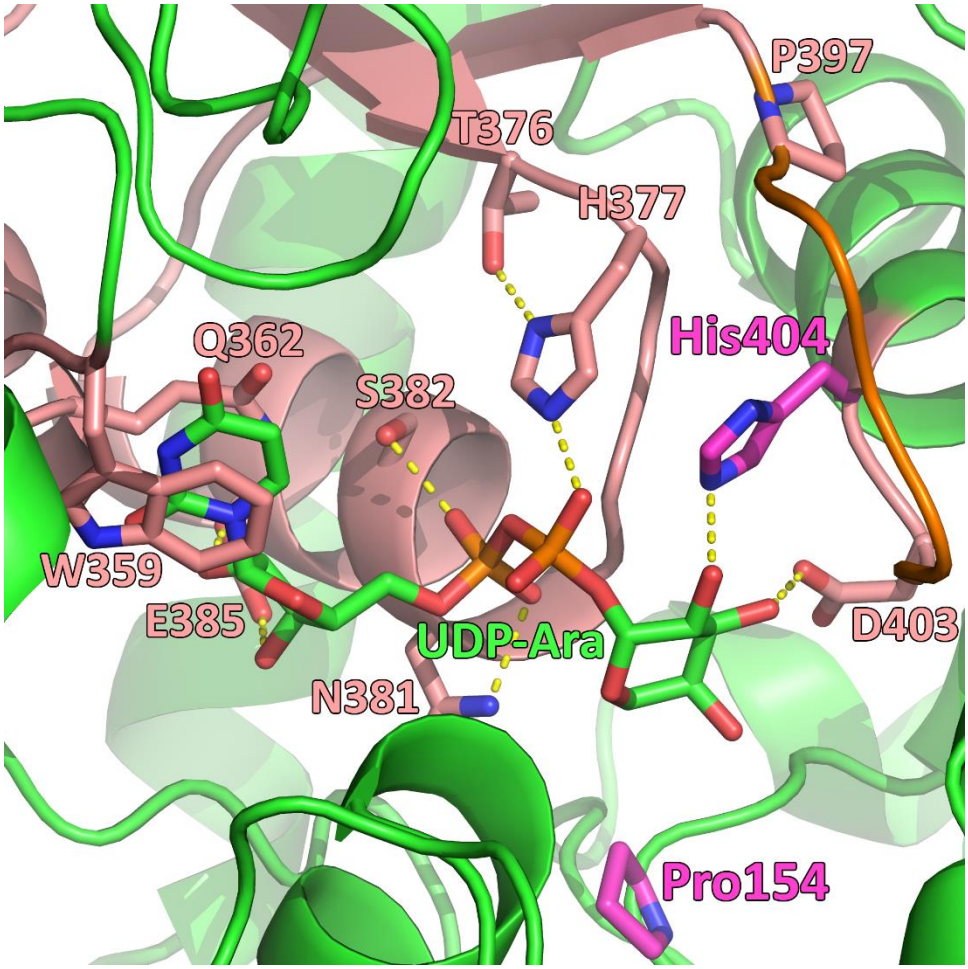


Figure 3.13: Structural model of AsAAT1 with bound UDP-Ara. The protein is represented in green ribbons with the PSPG motif in salmon, including the side chains of highly conserved residues (salmon sticks). The loop in orange was remodelled. His404 and Pro154 are shown as magenta sticks. Hydrogen bond interactions are shown with yellow dots. UDP-Ara is labelled and shown with carbon atoms in green. Figure prepared using PyMOL.[46]

Comparison with other glycosyltransferase structures validates the predicted bound conformation of the UDP moiety in our model. Other instances can be seen in the crystal structures of *Medicago truncatula* UGT71G1 complexed with UDP-Glc (PDB: 2ACW),[52] of *Glycyrrhiza uralensis* UGT73P12 complexed with UDP (PDB: 7C2X)(unpublished) and of *Oryza sativa* Os79 complexed with a UDP-Glc analogue (PDB: 5TMD).[207] In all these structures, the uracil moiety is clamped between an aromatic residue and a glutamine residue, and in the same orientation as in the AsAAT1 model. The hydroxyl groups of the ribose moieties are hydrogen-bonded to a conserved glutamate side chain. The first phosphate group of the UDP moiety accepts a hydrogen bond from a

conserved asparagine and its phosphoester linkage from a conserved histidine. The second phosphate receives a hydrogen-bond from a threonine or a serine. (Fig. 3.14).

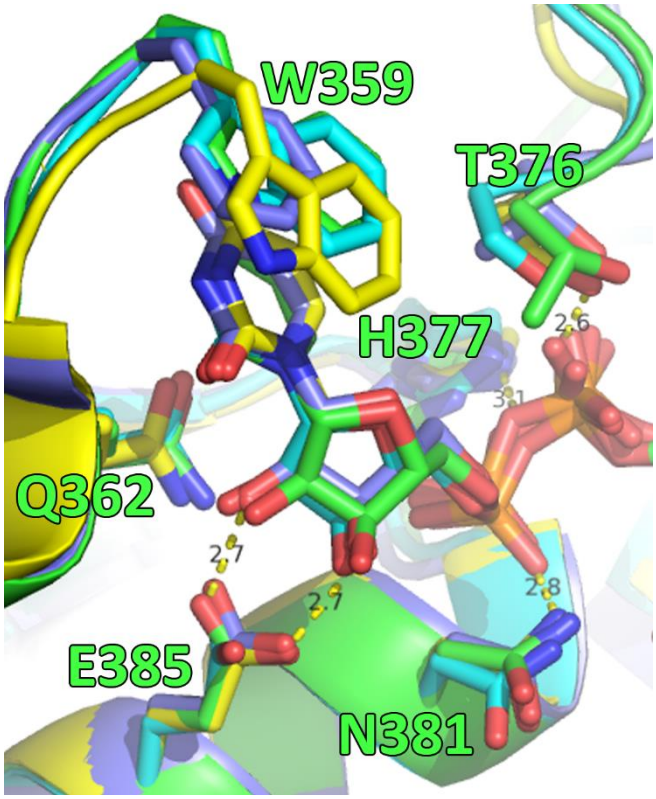


Figure 3.14: UDP conformation in the AsAAT1 model and in homologous structures. The PSPG motif and Thr298 from the AsAAT1 model (green cartoon and sticks) were aligned to their equivalent in three other structures of UGTs complexed with UDP-sugars (PDB: 2ACW, cyan; 7C2X, yellow with hydrogen bonds shown as yellow dots and distance in ångström; 5TMD, teal). The binding mode of the UDP moiety is conserved across structures. Figure prepared using PyMOL.[46]

The AsAAT1 model can help rationalise the effect of active site mutations on sugar specificity. It has been reported in the literature[39] that the His404Gln mutation in AsAAT1 cause a switch in primary specificity from arabinopyranose to xylopyranose sugar donors. These sugars differ only in their stereochemistry around carbon atom C4 (Fig. 3.15), so it could be argued that a histidine residue at this position favours an axial 4-hydroxyl while a glutamine residue favours an equatorial one.

On the other hand, the Pro154Ser mutant has increased activity towards galactose, which is essentially arabinose with a 5-hydroxymethyl substituent. When this mutation is coupled with His404Gln, the resulting mutant has its main activity switched to glucose, once again suggesting that the glutamine residue favours an equatorial 4-hydroxyl group.

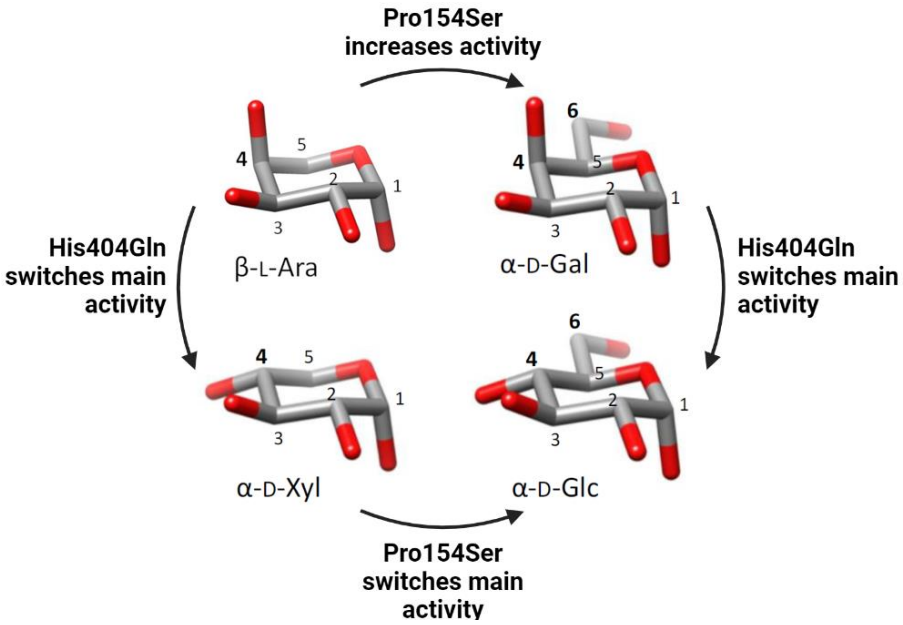


Figure 3.15: Structure of the four sugars for which activity was characterised in AsAAT1. Two single- and one double-mutant AsAAT1 have had their activity against all four corresponding UDP-sugars measured. Adapted from Louveau *et al*[39] using Biorender.com.

While His404 does not directly interact with the 4-hyroxyl group of the sugar according to the model, the side chain of histidine is larger than that of glutamine. This seems to push the sugar further towards Pro164 (Fig. 3.16), and prevents any equatorial 4-hydroxyl group from accepting a hydrogen bond from the backbone nitrogen of a conserved tryptophan residue. It may also cause a steric clash of equatorial 4-hydroxyl groups that is lessened in the more compact sugars that have an axial group instead. This suggests a structural basis for the fact His404Gln increases the preference for equatorial 4-hydroxyl groups. The effects of the Pro154Ser mutation may also be rationalised in terms of gain of hydrogen bond from the side chain to the 5-hydroxymethyl group as seen in the related glycosyltransferases.

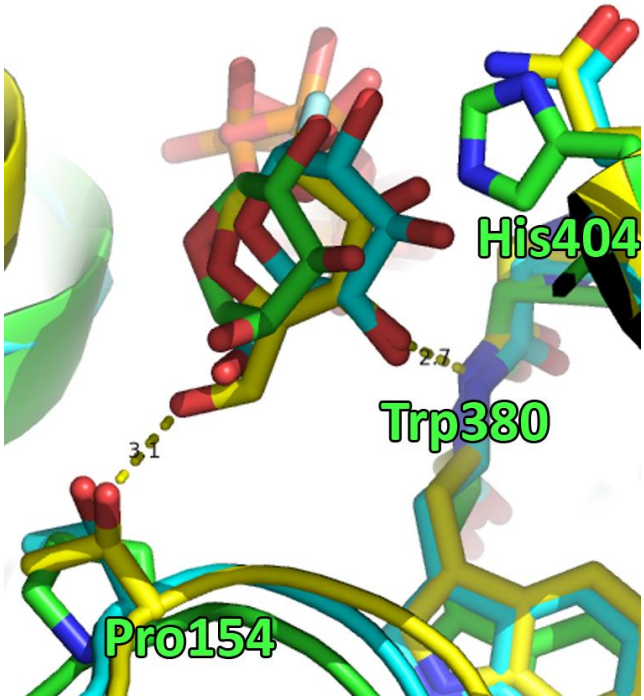


Figure 3.16: Binding pocket of the sugar moiety. The AsAAT1 model is shown in green and the crystal structures of two related plant glycosyltransferases are shown in cyan (PDB: 2ACW) and yellow (PDB: 7C2X) with its hydrogen bonds shown as yellow dashes and their length in ångström indicated. Figure prepared using PyMOL.[46]

3.4 Conclusions and future work

In conclusion, the purification of AsAAT1 with a N-terminal 9xHis-tag, expressed in *E. coli*, was much improved from the published one, but the construct did not crystallise. Tag-free AsAAT1 was also produced this way but, despite extensive crystallisation efforts, would not yield diffracting crystals. The deletion mutant AsAAT1-ΔN proved to be insoluble, so structural information was obtained through *in silico* methods instead. An energy-minimised molecular model of AsAAT1 in complex with UDP-Ara provided information about the binding of its sugar moiety. This helped rationalise the effect of two key active site residues on sugar specificity in terms of loss of hydrogen-bonds relative to glucosyltransferases.

A crystal structure of AsAAT1 is expected to be more accurate than the model presented herein. Though it may be the case that AsAAT1 cannot be crystallised, further attempts could be considered, for example by using nucleants, such as bio-glass,[208] or by cross-seeding with microcrystals of a different protein,[209] such as $GtfC_{100\text{-}23}$, another UGT crystallised as reported in this thesis (see Appendix 1). If a crystallisation method could be found and led to the solution of the crystal structure of AsAAT1, attempts to co-crystallise

and/or soak the ligands UDP-Ara and deglycosyl-avenacin A1 may yield structures of the resulting complexes.[52] To prevent sugar transfer, it may be required to synthesize a non-transferable UDP-Ara analogue (e.g., UDP-2-deoxy-2-fluoro-Ara)[210] or to generate catalytically inactive mutants. The latter may be achieved by mutating the catalytic base His35 (identified by alignment of the predicted structure) to Asn or Gln. Alternatively, mutation of the proton acceptor Asp132 to Asn may be sufficient.

Finally, in the absence of crystals of AsAAT1, the accuracy of the model presented here may be improved by performing docking and energy minimisation on a more accurate model of the enzyme, such as one generated by the AlphaFold2 software that has become available for academic research since the completion of this study.[176]

Chapter 4: AsTG1

4.1 Introduction

The final glycosylation step in the avenacin biosynthesis pathway is the addition of a 1,4-linked D-glucose moiety to the L-arabinose group of the molecule. This is catalysed by AsTG1, which is encoded on the *Sad3* locus identified in an avenacin-deficient oat mutant. At 3.6 cM, *Sad3* is less closely linked to the main gene cluster than the gene for AsAAT1 (0.66 cM). However, its expression is still restricted to the root tip. The enzyme is essential to healthy root development, likely due to the phytotoxicity of its substrate when it accumulates.[35] While AsTG1 is able to glucosylate variants of its natural substrate, control of the product is achieved through compartmentalisation: the precursor is generated in the cytosol, then transported to the vacuole, where AsTG1 is localised thanks to its signal sequence.[40] Unlike the UGTs that catalyse the first two glycosylations of avenacin, AsTG1 is a glycosyl hydrolase. It is part of cluster 6 of the GH1 family. Half of the characterised enzymes in this cluster have shown transglycosidase activity, and so did two *Delphinium grandiflorum* enzymes in cluster 7.[40, 211] The switch from glycosyl hydrolase to transglycosidase has therefore occurred multiple times, but its structural basis remains unknown.[212]

The aim of the work described in this chapter was to understand the determinants of transglycosidase activity based on the molecular structure of AsTG1. To obtain the recombinant enzyme for crystallisation, overexpression of various constructs was attempted in *E. coli*. Unfortunately, they either suffered from low solubility or did not crystallise. This prompted a shift to *in silico* methods, such as homology modelling, multiple-sequence alignment and molecular dynamics simulations.

4.2 Methods

In preparation for structure solution by X-ray crystallography, several constructs of recombinant AsTG1 were heterologously overexpressed in *E. coli*.

4.2.1 Crystallisation trials of AsTG1 with a non-cleavable N-terminal 9xHis-tag (9xHis-AsTG1)

The full-length construct with a non-cleavable N-terminal 9xHis-tag, 9xHis-AsTG1, was used in an initial attempt at crystallisation.

4.2.1.1 Expression of 9xHis-AsTG1

Heterologous overexpression of 9xHis-AsTG1 for the purpose of crystallisation followed the published protocol with minor modifications.[40] Briefly, colonies of *E. coli* Rosetta(DE3) cells transformed with the CDS for AsTG1 (lacking the signal sequence predicted by SignalP)[75] cloned into pH9GW (pH9GW-AsTG1) were obtained from Anastasia Orme (John Innes Centre). A single colony was used to inoculate LB (with 50 µg/ml kanamycin to select for the pH9GW vector and 30 µg/ml chloramphenicol to select for the pRARE plasmid) and was incubated at 37 °C and 180 rpm overnight. A 500 µL aliquot of the resulting culture was thoroughly mixed with 500 µL of sterile 50% glycerol and flash frozen in liquid nitrogen to prepare a glycerol stock for long-term storage at -80 °C.

This glycerol stock was used to prepare a preculture by inoculating it into 2 x 100 mL LB (50 µg/ml kanamycin) and incubating at 37 °C and 180 rpm overnight. For large-scale expression, 8 x 750 mL LB (50 µg/ml kanamycin) in 2 L baffled conical flasks were each inoculated with 25 mL of preculture. These were incubated at 37 °C and 200 rpm to an OD_{600} of 0.35, then moved to 16 °C and 200 rpm. Once the cultures reached an OD_{600} of 0.5, recombinant protein expression was induced overnight by adding 1 M IPTG to a final concentration of 0.05 mM. The cells were harvested by centrifugation at 8,950 *g* and 16 °C for 45 min, resuspended to 45 mL in AG buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 20 mM imidazole, 5% (v/v) glycerol) then flash frozen and stored at -80 °C.

4.2.1.2 Purification of 9xHis-AsTG1

The frozen pellet was thawed then mixed on ice for 15 min with 40 mL of 2X lysis buffer (AG buffer with 0.1 mg/mL DNase from bovine pancreas (Sigma-Aldrich), 2 mg/mL lysozyme from chicken egg white (Sigma-Aldrich), 4 mini-tab cOmplete EDTA-free protease inhibitor cocktail (Roche)). The cells were then lysed by two passages through a cooled cell disruptor (Constant Systems) set to 30,000 psi. The soluble fraction was separated from the cell debris by centrifugation at 5,100 *g* and 4 °C for 75 min and further clarified using 0.45 µm syringe filters.

The clarified soluble fractions were purified by immobilised nickel ion affinity chromatography using a 1 mL HisTrap HP column (GE Healthcare) at a flow rate of 1.4 mL/min on an ÄKTA Pure chromatography system (GE Healthcare) at 4 °C. The column was pre-equilibrated with 5 CV of AG buffer, before loading the soluble fractions, then washing with 25 CV of AG buffer. The bound proteins were eluted at 1 mL/min of a gradient from 20 to 500 mM imidazole, resulting from the mixing of binding buffer with a proportion of BG elution buffer (AG with 500 mM imidazole instead of 20 mM) increasing from 0% to 2.5% over 2.5 CV, maintained over 10 CV to selectively elute contaminants, then increasing to 40% over 25 CV. The eluates were collected in 2 mL fractions, which were assessed by running denatured samples on SDS-PAGE.

For the second and last purification step, the fractions of the eluate that showed the presence of a protein of the correct apparent molecular weight were pooled and concentrated to around 2 mL at 4 °C with an Amicon® Ultra 15 mL Centrifugal Filter (10 kDa MWCO, Merck). Precipitate was removed by centrifugation at 13,000 rpm and 4 °C for 10 min. The sample was then purified by SEC on a HiLoad 16/600 Superdex 75 pg column (GE Healthcare) pre-equilibrated and eluted at a flow rate of 0.4 mL/min with SEC buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 5 mM DTT). The elution was collected in 2 mL fractions and results were assessed by running denatured samples of the peak fractions on SDS-PAGE. The fractions containing pure protein were pooled and concentrated at 4 °C with an Amicon® Ultra 2 mL Centrifugal Filter (10 kDa MWCO, Merck) according to the manufacturer's instructions. The protein concentration was measured regularly with a NanoDrop™ Spectrophotometer (Thermo Scientific) using absorbance at 280 nm. When a reduction in volume stopped causing an increase in protein concentration, it was considered to have reached its solubility limit. This was calculated as 39 mg/mL based on absorbance at 280 nm and the molecular weight and extinction coefficient for AsTG1 with 9xHis-tag. These were estimated to be 58,088 Da and 87.21 $mM^{-1}cm^{-1}$ using the exact amino acid sequence and the ExPASy ProtParam online server.[190] Any precipitate was removed by centrifugation at 13,000 rpm and 4 °C for 10 min.

4.2.1.3 Characterisation of 9xHis-AsTG1 by LC-MS

Purified 9xHis-AsTG1 protein was characterised by LC-MS. To do this, freshly purified 9xHis-AsTG1 was diluted to 50 µM in SEC buffer. A major protein contaminant separated by SEC was analysed as purified. These 50 µL samples were each diluted with 450 µL of 2% acetonitrile, 0.1% formic acid. A volume of 10 µL was injected onto a ProSwift

reversed phase RP-1S column (4.6 x 50 mm; Dionex). The protein was eluted at a flow rate of 0.2 mL/min with a 15 min linear gradient from 2% to 100% acetonitrile resulting from the mixing of solvent A (0.1 % (v/v) formic acid in dH_2O) and solvent B (0.1% (v/v) formic acid in acetonitrile). The eluent was continuously infused into the ESI source of a micrOTOF-QIII mass spectrometer using the HyStar software (Bruker Daltonics). Mass spectra were acquired between 50 and 3000 m/z with the following parameters: dry gas flow of 8 L/min, nebuliser gas pressure of 0.8 bar, dry gas at 240 °C, capillary voltage at 4500 V, offset of 500 V, collision RF at 650 Vpp. The data was deconvoluted over the mass range of 40 to 98 kDa using the software Compass DataAnalysis version 4.1 (Bruker).

4.2.1.4 Crystallization trials of 9xHis-AsTG1

Crystallisation screening experiments were initiated with freshly purified 9xHis-tagged AsTG1 concentrated to 15 mg/mL. The concentration was decided based on manual tests of 10 conditions and an overnight screen of 192 conditions to determine which concentration yielded precipitate in half of them. Three commercially available screens were used: Structure Screen™ 1 and 2 Eco Screen,[191] JCSG-plus™ Screen,[192] and PACT premier™ Screen,[193] all from Molecular Dimensions. The screens were set up at 16 °C in 96-well 2-drop MRC plates sealed with ClearVue Sheets (Molecular Dimensions) employing an OryxNano protein crystallisation robot (Douglas Instruments Ltd.). The sitting drop vapour diffusion technique was used with a total drop size of 0.5 µL containing the protein and screen solution at either a 1:1 or a 1:2 volume ratio, equilibrated against 50 µL of screen solution per reservoir of the 96-well plate. The screening plates were monitored for crystal formation using a SZX9 Stereo Microscope (Olympus).

4.2.2 Cloning of AsTG1 variants with a cleavable N-terminal 9xHis-tag

As the 9xHis-tagged enzyme failed to crystallise, it was decided to pursue tag-free enzyme variants to increase the likelihood of crystallisation. Specifically, deletion mutants were designed to remove the N- and/or C-terminal sections of the polypeptide that were predicted to be disordered (see section 2.1). These mutants and the full-length wild-type protein were engineered to introduce a protease cleavage site between the protein and the 9xHis-tag. The constructs AsTG1-ΔNΔC and AsTG1-ΔC were obtained by PCR amplification and Gateway cloning, but this strategy did not work for the AsTG1 and AsTG1-ΔN constructs. These were therefore obtained by mutagenesis of the pH9GW-AsTG1-WT plasmid once it was acquired (see Fig. 4.1 and section 4.2.4).

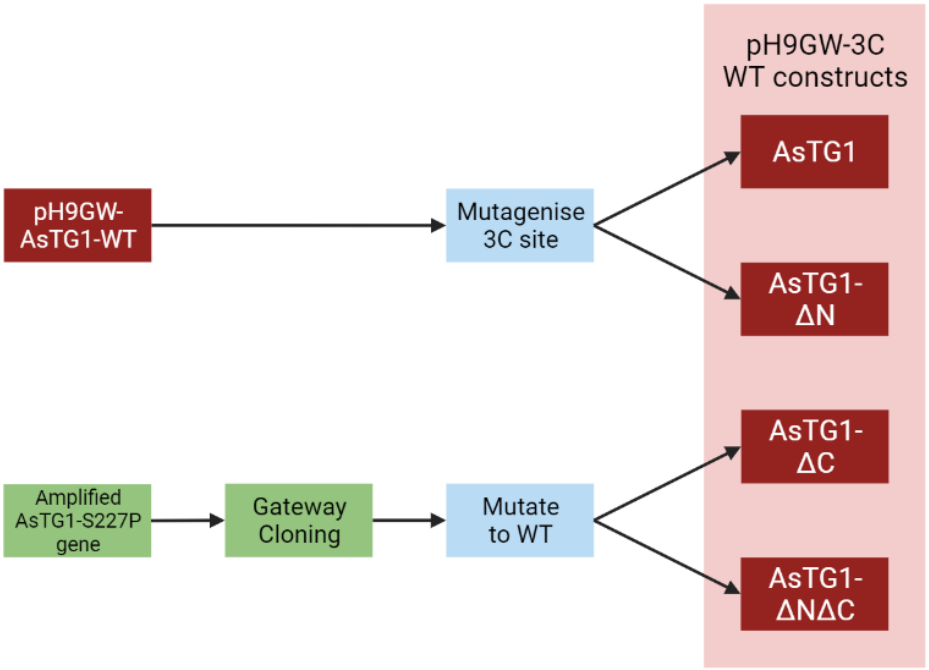


Figure 4.1: Strategy for obtaining expression clones of the four cleavable constructs of AsTG1. Created with Biorender.com.

4.2.2.1 Amplification of AsTG1 construct genes

The AsTG1-pEAQ-Dest1(NoSig) plasmid was obtained from Thomas Louveau (John Innes Centre). It is derived from a Gateway-compatible vector that confers kanamycin resistance and is designed for protein production in plant tissue.[213] It was propagated in *E. coli* DH5α SC (Subcloning Efficiency™, Invitrogen) then used at 2.5 ng/µL as a template for a first stage PCR to amplify the gene using the following gene-specific primers (gene specific region unformatted, 3C protease cleavage site in bold, overlap with the *att*B2 cloning site in blue):

TG-full-3C-F 5'-**CTGGAAGTTCTGTTTCAGGGCCCG**ATGGGAGACGTTGTGGTGGCGG-3'

TG-ΔN-3C-F 5'-**CTGGAAGTTCTGTTTCAGGGCCCG**ATGACCCGCCGTGACTTCCCC-3'

TG-ΔC-attB-R 5'-CAAGAAAGCTGGGTTTTAGTTTAGGAAGCTAGAGTACCATCTG-3'

For AsTG1-ΔC, the primers TG-full-GW-F and TG-ΔC-attB-R were used. For AsTG1-ΔNΔC, the primers TG-ΔN-3C-F and TG-ΔC-attB-R were used. The first stage PCR was carried out with Phire Hot Start II DNA polymerase (Thermo Fisher) with 5% (v/v) DMSO using the following program: 98 °C for 3 min, followed by 32 cycles of 98 °C for 10 s, 48 °C for 20 s, 72 °C for 30 s, followed by 72 °C for 5 min. The presence of a single product of the correct size

was confirmed by 0.8 % (w/w) agarose TAE gel electrophoresis with ethidium bromide staining, and the resulting fragments were purified using a QIAquick® PCR purification kit (Qiagen) with 30 µL autoclaved dH_2O incubated 4 min before elution.

4.2.2.2 Introduction of *att* cloning sites into AsTG1 construct genes

The purified products of the first stage PCR were diluted 10-fold with autoclaved dH_2O and 2 µL aliquots were used as DNA template for the second stage PCR, which added an *att*B1 site (in red) upstream of the 3C protease cleavage site (in bold) and completed the *att*B2 site (in blue) downstream of the genes using the following general primers:

attB1-3C-F 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTC**CTGGAAGTTCTGTT**-3'

attB2-R 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTT-3'

This second stage PCR was carried using the same protocol as above but with 1% (v/v) DMSO and 45 °C for annealing instead of 48 °C.

4.2.2.3 Cloning of AsTG1 constructs into the Gateway entry vector pDONR207

The resulting PCR fragments were used as is in 0.5 µL aliquots to perform BP clonase reactions, each with 1 µL BP clonase mix, 2.5 µL TE buffer (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA), and 1 µL pDONR207 (58 ng/µL). This was incubated at 25 °C for 16 h before adding 0.5 µL aliquots of proteinase K and incubating at 37 °C for 10 min. The resulting reactions were used in 1 µL aliquots to transform DH5α LE (Library Efficiency™, Invitrogen) cells (25 µL each). After recovery growth, the majority of the transformation reaction was plated on LB agar (20 µg/mL gentamycin). The presence of a gene of the correct size was confirmed in a few of the resulting colonies for each construct by colony PCR (see 2.6.2.1 for the protocol) using entry-vector specific flanking primers:

SeqLA 5'-TCGCGTTAACGCTAGCATGGATCTC-3'

SeqLB 5'-GTAACATCAGAGATTTTGAGACACGGG-3'

These colonies were inoculated in 10 mL LB (20 µg/mL gentamycin) each and grown at 37 °C and 180 rpm overnight. The plasmids pDONR207-3C-AsTG-ΔC and pDONR207-3C-AsTG-ΔNΔC were extracted from the cultures using a QIAprep Spin Miniprep Kit (Qiagen) with 30 µL EB for elution. Sequencing (Eurofins Genomics) with the SeqLA and SeqLB primers confirmed the presence of the gene but revealed a non-conservative substitution leading to a P227S mutation.

4.2.2.4 Cloning of AsTG1 constructs into the Gateway expression vector pH9GW

Plasmids pDONR207-3C-AsTG1-ΔC and pDONR207-3C-AsTG1-ΔNΔC were used as entry clones to transfer their inserts into the destination vector pH9GW, generating the expression plasmids pH9GW-3C-AsTG1-ΔC and pH9GW-3C-AsTG1-ΔNΔC. This was achieved by using 0.5 µL of LR clonase enzyme mix (Invitrogen), 1 µL pH9GW (58 ng/µL), 0.2 µL entry clone (between 509 and 684 ng/µL) and made up to 5.5 µL with TE buffer pH 8.0. After incubation at 25 °C for 10 h, the LR reactions were terminated with 1 µL proteinase K mix (Invitrogen) incubated at 37 °C for 10 min. Aliquots of the resulting LR reactions (0.2 µL each) were used to transform *E. coli* DH5α LE cells (15 µL each), which were plated on selective LB agar (50 µg/mL kanamycin). A transformant colony per construct was inoculated into 10 mL LB with appropriate antibiotic for growth at 37 °C and 180 rpm overnight. The expression plasmids were extracted from the resulting cultures using the QIAprep Spin Miniprep Kit (Qiagen) with 30 µL EB incubated 5 min before elution.

4.2.2.5 Mutagenesis of AsTG1 Gateway plasmids to obtain wild-type genes

To repair the P227S mutation and revert to the WT sequence, mutagenesis was performed on all AsTG1 entry and expression vectors. They were diluted to 50 ng/µL to use either 0.2 or 1 µL as DNA template for PCR using Phusion® High-Fidelity DNA Polymerase (NEB) without DMSO using the following primers (mutation in bold, primer-primer overlap in italics):

TG-S227P-F 5'-*CAGGGAG**C**CCTACA*TCGCGGCGCACCAC-3'

TG-S227P-R 5'-*TGTAGG**G**CTCCCTG*GTGGAGTCTCCGGCGG-3'

The thermal cycling protocol was 98 °C for 3 min, followed by 32 cycles of 98 °C for 20 s, 47 °C for 30 s, 72 °C for 4 min, followed by 72 °C for 10 min. The presence of a product of the correct size was confirmed by 0.8% (w/w) agarose TAE gel electrophoresis with ethidium bromide staining. The methylated DNA template was digested using 0.5 µL DpnI (NEB) per reaction, incubated at 37 °C for 3 h. The correct product was purified by 0.8% (w/w) agarose gel electrophoresis, cutting the band of the right size and extracting the DNA with a QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. Aliquots (2 µL) of the resulting DNA samples (6 to 8 ng/µL) were transformed into 10 µL aliquots of XL10-Gold® Ultracompetent Cells (Agilent) by heat shock. A transformant colony of each was inoculated into 10 mL LB (50 µg/mL kanamycin) for growth at 37 °C and 180 rpm overnight. Plasmids were extracted from the cultures using the QIAprep Spin Miniprep

Kit (Qiagen) with 30 µL EB incubated 3 min before elution. Sequencing (Eurofins Genomics) with SeqLA and SeqLB flanking primers confirmed the presence of the gene mutated back to wild-type.

4.2.3 Expression and purification of AsTG1-ΔC and AsTG1-ΔNΔC

AsTG1-ΔC and AsTG1-ΔNΔC are the shortened constructs of AsTG1 that could be cloned with a 3C protease cleavage site using Gateway cloning. After having fixed the mutation carried over from the original plasmid, the expression clones pH9GW-3C-AsTG1-WT-ΔNΔC and pH9GW-3C-AsTG1-WT-ΔC were used to express these constructs with an N-terminal 9xHis-tag in *E. coli*, starting with the shortest one as it was considered to have the best chances of crystallising.

4.2.3.1 Expression and purification of AsTG1-ΔNΔC with a cleavable N-terminal 9xHis-tag

The plasmid pH9GW-3C-AsTG1-WT-ΔNΔC (10 ng) was transformed into 22 µL of BL21(DE3) cells by heat shock. Plating on LB agar (50 µg/mL kanamycin) yielded numerous colonies, one of which was inoculated into 50 mL LB (50 µg/mL kanamycin) for growth at 37 °C and 180 rpm overnight. The resulting culture was used to prepare a 25% (v/v) glycerol stock for long-term storage at -80 °C. This stock was used to inoculate 100 mL LB (50 µg/mL kanamycin) which was incubated at 37 °C and 180 rpm overnight. This pre-culture was then used to inoculate 10 mL aliquots into 8 x 500 mL LB (50 µg/mL kanamycin) incubated at 37 °C and 180 rpm for grow-up. When it reached an OD_{600} of 0.4, it was transferred to 16 °C and 200 rpm. After 45 min of acclimatisation, it reached an OD_{600} of 0.55 and was induced with 1 M IPTG to a total of 0.1 mM for recombinant protein expression overnight. Cells were harvested by centrifugation at 8,950 *g* and 16 °C for 25 min then stored at -80 °C.

The frozen cell pellet was lysed and the soluble fraction purified in a first step of immobilised nickel ion affinity chromatography as described previously for the non-cleavable construct. The resulting fractions that showed the presence of a protein of the right size were pooled. To prepare for tag-cleavage with 3C protease, the imidazole was removed using a HiPrep 26/10 desalting column (GE Healthcare) at a flow rate of 15 mL/min using an ÄKTA Pure chromatography system (GE Heathcare) at 4 °C. Fractions which showed UV absorbance on the chromatogram were pooled to a total of 18 mL. They were estimated to contain 2.4 mg of protein, based on integration of the A280 peak and a protein extinction coefficient calculated using the protein's sequence with the online server

ExPASy ProtParam.[190] Based on the manufacturer's recommendations, our protein sample was incubated with 4 µL of 3 mg/mL recombinant His-tagged HRV 3C protease with slow stirring at 4 °C overnight. The precipitate that formed was removed by centrifugation at 5,100 rpm and 4 °C for 15 min.

A second IMAC with the same parameters was run to recover the further purified protein in the flow-through, the presence of which was confirmed by running denatured samples on SDS-PAGE.

For the third purification step, the fractions containing the cleaved protein were pooled and concentrated to 3 mL at 4 °C with an Amicon® Ultra 15 mL Centrifugal Filter (10 kDa MWCO, Merck). Precipitate was removed by centrifugation at 13,000 rpm and 4 °C for 10 min. The sample was then purified by size-exclusion chromatography on a HiLoad 16/600 Superdex 75 pg column (GE Healthcare) pre-equilibrated and eluted at a flow rate of 0.4 mL/min with SEC buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 5 mM DTT). The elution was collected in 2 mL fractions and results were assessed by running denatured samples on SDS-PAGE.

4.2.3.2 Expression and purification of AsTG1-ΔC with cleavable N-terminal 9xHis-tag

As the construct lacking both disordered termini, AsTG1-ΔNΔC, appeared too insoluble for the production of the large quantities of protein needed for crystallisation screens, the alternative construct AsTG1-ΔC was expressed and purified instead. To do so, the plasmid pH9GW-3C-AsTG1-WT-ΔC was transformed into BL21(DE3) for protein expression as for pH9GW-3C-AsTG1-WT-ΔNΔC. The same two IMAC purification steps were used, with the amount of 3C protease after desalting scaled to 17 µL for the estimated 8 mg of protein. The fractions from the second IMAC were analysed by running denatured samples on SDS-PAGE.

4.2.4 Construction and expression of AsTG1 with cleavable 9xHis-tag and its shortened construct AsTG1-ΔN

The constructs lacking the C-terminal region predicted to be disordered yielded little soluble protein, so it was hoped that the full-length and ΔN constructs that still included that region would have better solubility. As they were unamenable to Gateway cloning with a 3C protease cleavage site, they were constructed by mutagenesis of the plasmid pH9GW-AsTG1.

4.2.4.1 Mutagenesis to obtain the full-length and ΔN constructs of AsTG with cleavable 9xHis-tag

The plasmid pH9GW-AsTG1 was obtained from Anastasia Orme and used as a DNA template. PCR was carried out with Phire Hot Start II DNA polymerase (Thermo Fisher) and either 0% or 5% (v/v) DMSO, using 10, 25 or 50 ng plasmid as template DNA and the following mutagenesis primers (3C protease cleavage site in bold, primer-primer overlap in italics):

TG-3C-mut-F 5'-*AGCAGGCTTAATG**TTG***GAAGTGTTATTTCAGGGCCCG GGAGACGTTGTGGTGGCGG-3'

TG-ΔN-3C-mut-F 5'-*AGCAGGCTTAATG**TTG***GAAGTGTTATTTCAGGGCCCG ACCCGCCGTGACTTCCCC-3'

TG-3C-mut-R 5'-***CAA****CATTAAGCCTGCT*TTTTTGTACAAACTTGTTGATTCGAGACC-3'

For full-length AsTG1, the primers TG-3C-mut-F and TG-3C-mut-R were used. For AsTG-ΔN, the primers TG-ΔN-3C-mut-F and TG-3C-mut-R were used. The thermal cycling protocol was: 98 °C for 3 min, followed by 32 cycles of 98 °C for 20 s, 50 °C for 30 s, 72 °C for 4 min, followed by 72 °C for 10 min. The presence of products of the correct size was confirmed by 0.8% (w/w) agarose TAE gel electrophoresis with ethidium bromide staining. The methylated DNA template was digested using 0.5 µL DpnI (NEB) per reaction, incubated at 37 °C for 3 h. All reactions for each construct were pooled and the correct products were purified by running them on 0.7% (w/w) agarose gel electrophoresis, cutting the bands of the right size and extracting the DNA with a QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. The resulting DNA was diluted with autoclaved dH_2O then transformed in XL10-Gold® Ultracompetent Cells (Agilent) by heat shock. Transformant colonies were inoculated into 10 mL LB (50 µg/mL kanamycin) each for growth at 37 °C and 180 rpm overnight. The propagated plasmids pH9GW-3C-AsTG1-WT and pH9GW-3C-AsTG1-WT-ΔN were extracted from the cultures using the QIAprep Spin Miniprep Kit (Qiagen) with 30 µL EB incubated 5 min before elution. Sequencing (Eurofins Genomics) with a T7 promoter forward primer and with a gene-specific reverse primer (sequence below) confirmed the presence of the 3C-AsTG1 and 3C-AsTG1-ΔN genes.

TG-attB-R 5'-CAAGAAAGCTGGGTTTTACGCAGAGTCGTAATATTGTTTCTTGGG-3'

4.2.4.2 Expression of 9xHis-3C-AsTG1-ΔN

After the construction of plasmids pH9GW-3C-AsTG1-WT-ΔN and pH9GW-3C-AsTG1-WT by mutagenesis, the shortened construct AsTG-ΔN was expressed first, as it was considered to have better chances of crystallising than the full-length protein. Plasmid pH9GW-3C-AsTG1-WT-ΔN (2.5 ng) was used to transform 20 µL of Rosetta(DE3) chemically competent cells (Novagen) by heat shock. A single transformant colony was inoculated 50 mL LB (with 50 µg/mL kanamycin and 30 µg/mL chloramphenicol) and incubated at 37 °C and 180 rpm overnight. The resulting culture was used to prepare a 25% (v/v) glycerol stock that was flash frozen and stored at -80 °C. This frozen glycerol stock was used inoculate 10 mL LB (50 µg/mL kanamycin and 30 µg/mL chloramphenicol) to revive it at 37 °C and 180 rpm overnight. Aliquots of the resulting culture (1 mL) were used to inoculate 2 x 100 mL LB (50 µg/mL kanamycin and 30 µg/mL chloramphenicol) for pre-culture at 37 °C and 200 rpm overnight. The pre-cultures were pooled, and 20 mL aliquots were used to inoculate 8 x 750 mL LB (50 µg/mL kanamycin and no chloramphenicol as it slows growth considerably), which were incubated at 37 °C and 200 rpm for grow-up. When they reached an OD_{600} of 0.35, the temperature was lowered to 16 °C. After 45 min of acclimatisation, the cultures were induced at an OD_{600} of 0.5 using IPTG to a final concentration of 0.05 mM for recombinant protein expression overnight. Cells were harvested by centrifugation at up to 8,950 *g* and 16 °C for 45 min, then resuspended to 45 mL with AG buffer before flash-freezing and storage at -80 °C.

4.2.4.3 Purification of 9xHis-3C-AsTG1-ΔN

The frozen cell pellet was lysed and the soluble fraction purified in a first step of immobilised nickel ion affinity chromatography as described previously for the non-cleavable full-length construct (see section 4.2.1.2). To prepare for tag-cleavage with 3C protease, the imidazole was removed using a HiPrep 26/10 desalting column (GE Healthcare) at a flow rate of 15 mL/min using an ÄKTA Pure chromatography system (GE Heathcare) at 4 °C. Fractions which showed UV absorbance on the chromatogram were pooled to a total of 16 mL. This was incubated with 30 µL of 3 mg/mL His-tagged recombinant HRV 3C protease with slow stirring at 4 °C overnight. The precipitate that formed was removed by centrifugation at 5,100 *g* and 4 °C for 25 min.

To remove the cleaved His-tag, the uncleaved protein, the 3C protease and the nickel-binding contaminants from the first nickel ion affinity chromatography, a second

IMAC with the same parameters was run in the hope of recovering the further purified cleaved protein in the flow-through.

4.2.4.4 Expression and purification of 9xHis-3C-AsTG1

Plasmid pH9GW-3C-AsTG1-WT was used to transform Rosetta(DE3) cells for protein expression as for pH9GW-3C-AsTG1-WT-ΔN (see section 4.2.4.2). A first IMAC purification step followed by desalting was carried out as for 9xHis-3C-AsTG1-ΔN. The resulting fractions that showed UV absorbance on the chromatogram were pooled to a total of 49 mL. This was incubated with 50 µL of 3 mg/mL His-tagged recombinant 3C protease with slow stirring at 4 °C overnight. The precipitate that formed was removed by centrifugation at 5,100 *g* and 4 °C for 20 min.

A second stage nickel ion affinity chromatography was performed using the same method as the first. The desalting step was repeated on the eluate and the resulting fractions were pooled to a total of 15 mL, which was incubated with 200 µL of 3 mg/mL 3C protease with slow stirring at 4 °C overnight. To assess this second tag-cleavage attempt, a third and final nickel ion affinity chromatography was performed using the same method and analysis as the first.

4.2.5 *In silico* studies of AsTG1

In the absence of crystal structure of AsTG1, it was decided to use multiple sequence alignment, homology modelling and MD simulations to better understand the determinants of transglycosidase activity.

4.2.5.1 Multiple sequence alignment for GH1 enzymes

Characterised eukaryotic glycosyl hydrolases from family 1 were identified on the CAZY database,[214] and their UniProt accession number[215] used to obtain their sequence. By manually removing putative enzymes, homologues, hydrolase-like proteins, enzymes with unspecified function or sugar substrate, since-updated or removed sequences, and unnamed proteins, this led to a list of 3 transglucosidase, 81 glucosyl hydrolases, 14 myrosinases and 4 dissacharide-removing hydrolases, which were all labelled as such. The list was submitted to the T-coffee tool[216] of the MPI Bioinformatics Toolkit.[217]

4.2.5.2 AsTG1 homology modelling

The best template for a homology model of AsTG1 was identified using the SWISS-MODEL template search.[218] Out of the six structures with the highest GMQE score (Global Model Quality Estimate), the structure of Os3BGlu6 with the best resolution (PDB: 3GNO)[219] was chosen. The amino acid sequence of AsTG1 (lacking the signal sequence predicted by SignalP)[75] was submitted to RosettaCM on the Robetta protein structure prediction service,[220] specifying 3GNO as the template structure and 10 as the number of models to generate, so as to obtain local error estimates. Of the five resulting homology models that were output, the one with the best active site was identified by averaging the local error estimates for the atoms of the active site residues, taken as residues within 3.5 Å of the covalently-bound glucose group from the aligned structure of the transglucosylation-deficient E178Q mutant (PDB: 3WBE).[221]

4.2.5.3 Molecular dynamics simulations of AsTG1 in complex with mono- or bis-deglucosyl avenacin

Insights into the role of active site residues and the binding of both mono- and bis-deglucosyl avenacin A-1 to AsTG1 were obtained through MD simulations. To generate the glucosylated form of the AsTG1 enzyme that is its presumed reaction intermediate, the catalytic residue Glu381 was modified to give *O*-glucosyl-L-glutamate (EGL). The initial coordinates for this variant were generated by reference to the crystal structure of rice BGlu1 E386G/S334A mutant complexed with cellotetraose (PDB: 3SCV)(unpublished). This GH1 family enzyme has glucosyltransferase activity and the crystal structure of the catalytically inactive mutant, when superimposed onto apo-AsTG1, revealed the bound tetrasaccharide to be positioned such that the C1 atom of the terminal non-reducing sugar was adjacent to the OE2 side chain carboxylate oxygen of Glu381 in the apo-AsTG1 model. The transformed coordinates of the non-reducing D-glucose moiety from the BGlu1 complex were therefore transferred to the model structure of AsTG1 to generate *O*-glucosyl-E381 AsTG1 (EGL-AsTG1).

Unoptimized Cartesian coordinates for the ligands mono- and bis-deglucosyl avenacin A-1 (MDA and BDA, respectively) were obtained from the isomeric SMILES representations of the corresponding entries in the PubChem database[222] transformed to Cartesian space using the electronic Ligand Builder and Optimization Workbench (eLBOW) tool in Phenix.[171] Semiempirical QM-optimized atomic coordinates and partial charges,

together with force field parameters, were obtained from the Automated Topology Builder (ATB) version 3.0.[223] To generate starting coordinates for the complex of glucosylated AsTG1 with MDA (EGL-AsTG1:MDA), the L-arabinopyranose ring of the avenacin A-1 precusor was manually docked again utilizing the crystal structure of rice BGlu1 E386G/S334A mutant complexed with cellotetraose (PDB entry 3SCV). In this process, MDA was positioned so that its axial hydroxyl oxygen at the C4 position of the L-arabinopyranose ring (atom O4) was located 3 Å from carbon C1 of the glucose ring covalently bound to the catalytic residue Glu381. The ligand molecule was then oriented to avoid short van der Waals contacts with residues in the active site cavity. These coordinates were then used to generate a starting model for the complex with BDA (EGL-AsTG1:BDA) by replacing the β-D-glucopyranosyl moiety at the 2-position of the arabinopyranose ring with a hydroxyl group.

Molecular dynamics simulations were performed on models EGL-AsTG1, EGL-AsTG1:MDA and EGL-AsTG1:BDA using the GROMACS 2020.4 molecular dynamics package[182] with the amber99sb-ildn force field.[201] Processing of the structures prior to the MD simulations was performed. The protonation states of histidine, aspartate and glutamate residues were selected with reference to the H++ server.[224] MD simulations in aqueous solution were then performed at a constant temperature of 298 K in a cubic box with 10 Å distance from the centre of the protein to the edge of the box. The box was solvated by the Simple Point Charge (SPC) water model, adding sodium counter ions to ensure neutral charge of the system. Prior to the unrestrained MD simulations, the systems were subjected to a maximum of 10,000 steps of energy minimisation using the conjugate gradient optimization method followed by 20 ps of position-restrained MD in the NVT canonical ensemble with force constants of 1,000 kJ mol^{-1} nm^{-2} on all protein atoms in order to equilibrate the water molecules in the solvation box. The equilibrated system was then subjected to a short production MD runs of 10 ns duration using a 2 fs time step in the NPT isothermal-isobaric ensemble. Note that a weak distance restraint with a force constant of 50 $kJmol^{-1}Å^{-2}$ was applied to maintain a distance of 3 Å between the axial hydroxyl oxygen at the C4 position of the L-arabinopyranose ring (atom O4) of MDA and BDA, and the C1 atom of the glucose ring covalently bound to the catalytic residue Glu381. The purpose of this restraint was to allow the ligands to sample only that conformational space within the active site which was consistent with the ligand being able to act as a sugar acceptor given the known transferase activity of the enzyme.

Analysis of the MD trajectories was carried out using embedded tools in the GROMACS package. Root mean square deviation (RMSD) and root mean square fluctuations (RMSF) of the Ca atoms were calculated with the original model as a reference. Clustering[225] was performed in GROMACS using this atom selection with a cut-off of 1.0 Å.

4.3 Results and discussion

4.3.1 Expression, purification and crystallisation screening of AsTG1

4.3.1.1 Optimised purification of 9xHis-tagged AsTG1 yields highly soluble protein

The first attempt to obtain crystals of AsTG1 followed the published expression and purification strategy as closely as possible.[40] The gene lacking the signal sequence had been cloned into the pH9GW vector, which added a non-cleavable N-terminal 9xHis tag. The pH9GW-AsTG1 plasmid had been transformed into the *E. coli* Rosetta (DE3) protein expression strain (see section 2.7.2.2 for details). A streak of this transformant was obtained and used the first large-scale expression of AsTG1, using the original protocol.

For the lysis of the cells to be more thorough, it was performed with a cell disruptor instead of a sonicator. A first purification attempt by immobilised metal ion affinity chromatography (IMAC) used a straight imidazole gradient. SDS-PAGE analysis revealed a strong band of the right size in the eluate, but also weaker bands of the same size in all protein-containing fractions, including flow-through and wash. This may be due AsTG1 binding non-specifically to other proteins, in a way that makes the His-tag unavailable for nickel binding. Along with the UV chromatogram, analysis by SDS-PAGE also revealed that the two major contaminants (which may be GlmS (66.8 kDa) and YfbG (74.2 kDa) from *E. coli*) started eluting just before the bulk of the protein of the right size started co-eluting (Fig. 4.2A). For better separation in subsequent purification attempts, we halted the upwards gradient of imidazole concentration at 32 mM for 10 mL to wash the contaminants out before continuing the increase in imidazole concentration (Fig. 4.2B).

(A) **(B)**

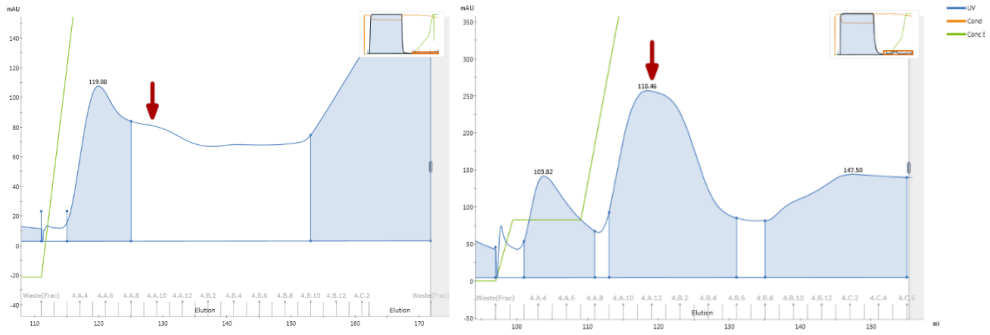


Figure 4.2: Chromatogram of the first stage IMAC purification of 9xHis-AsTG1 with two different gradients. (A) Straight gradient elution or **(B)** a gradient with a step inserted at 32 mM imidazole. The red arrow points to the elution peak of 9xHis-AsTG1 on the blue curve (absorbance at 280 nm). The green line represents the proportion of high-imidazole buffer. Fraction names are written in grey on the x-axis.

While the purity of the AsTG1-containing fraction was thus improved, it was still insufficient for crystallisation screen: a major contaminant seemed to stick to AsTG1, and several minor contaminants of higher and lower molecular weight were still present (Fig. 4.3).

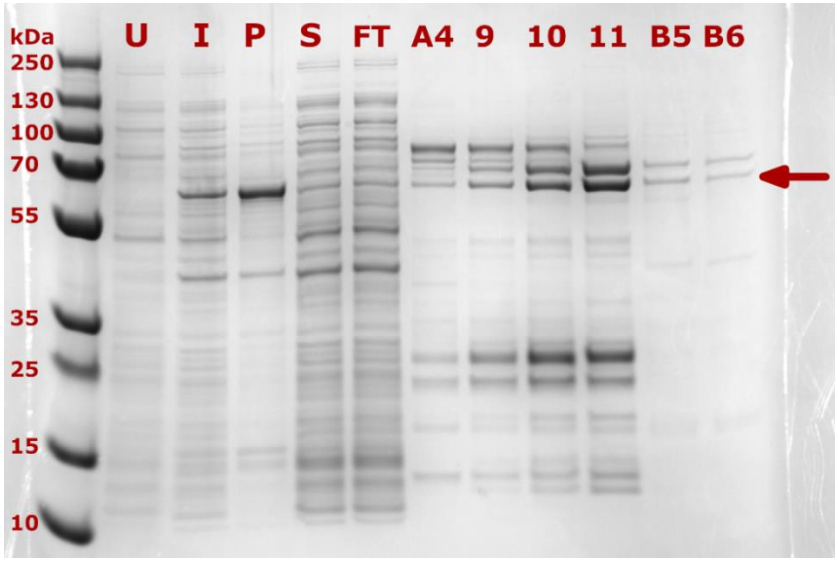


Figure 4.3: SDS-PAGE analysis of expression and first purification step of 9xHis-AsTG1 (IMAC). 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights, the total protein of the uninduced culture (U) and final induced culture (I) which shows the appearance of the overexpressed protein; pellet of the lysate (P) which is the insoluble fraction; soluble fraction of the lysate (S); unbound proteins of the flow-through (FT); fraction numbers from gradient elution with a step. The arrow indicates the expected migration distance of 9xHis-AsTG1.

It was therefore decided to perform a second stage purification using SEC with the reductant dithiothreitol in the running buffer. This successfully separated the main lower and all higher molecular weight contaminants out of the AsTG1-containing fractions, according to SDS-PAGE analysis (Fig. 4.4). The major contaminant was analysed by LC-MS to give a mass of 57,200.08 Da, which could not reliably be assigned to a typical contaminant from IMAC, though the closest match was for Hsp60 (GroEL) at 57.0 kDa[204] and some *E. coli* GroEL sequences (UniProt:[215] C4NV17) have a calculated MW of 57,206 Da.

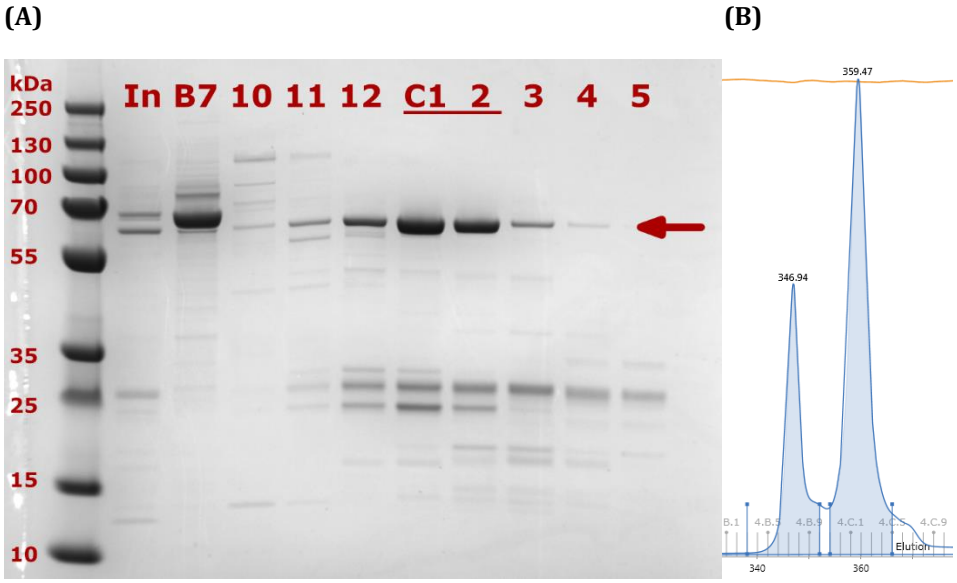


Figure 4.4: SDS-PAGE analysis and chromatogram of the second purification step of 9xHis-AsTG1 (SEC). (A) 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights, the protein sample injected into the column (In) and the elution fraction numbers (with the pooled fractions underlined). The arrow indicates the expected migration distance of 9xHis-AsTG1. **(B)** Chromatogram with fractions numbers written in grey on the x-axis and the blue curve representing absorbance at 280 nm. The large second peak corresponds to the elution of 9xHis-AsTG1.

4.3.1.2 9xHis-tagged AsTG1 is moderately stable to proteolytic degradation

A first grow-up in 750 mL LB yielded 0.2 mg of protein which was concentrated to 4.6 mg/mL and used to assess resistance to proteolysis at 4 °C and 16 °C by taking time-point samples to be denatured and run on SDS-PAGE (Fig. 4.5). This revealed an apparent cleaving of most of the protein by the 21st day at this concentration, leaving a slightly smaller species. The majority of the contaminants in the 10 to 70 kDa range of apparent molecular weight also degrade into species that do not appear on the gel by the 45th day.

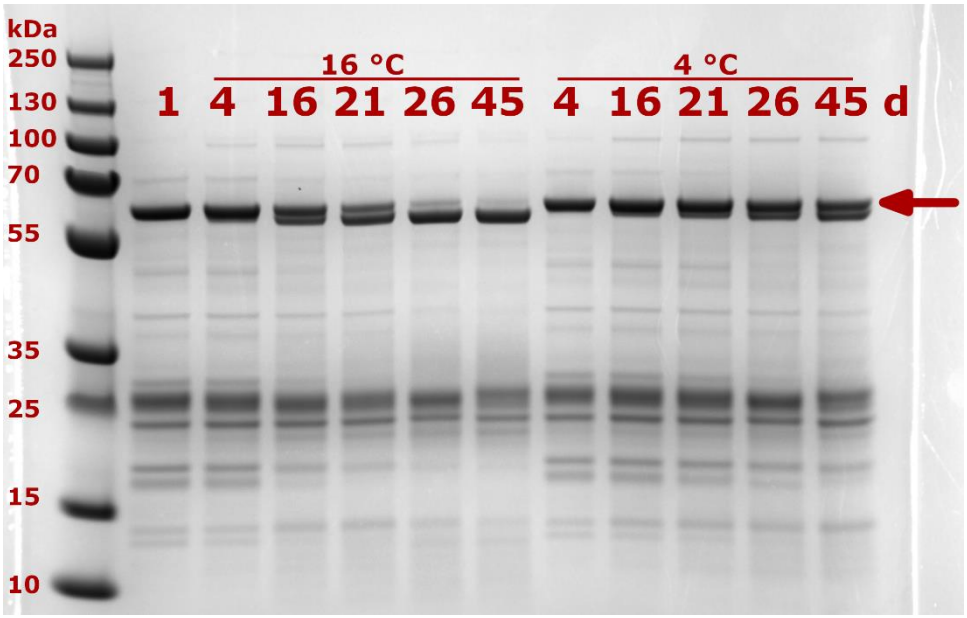


Figure 4.5: Degradation analysis of 9xHis-AsTG1 by SDS-PAGE. 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. The protein was concentrated to 4.6 mg/mL in SEC buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 5 mM DTT). Samples were left for 1 to 45 days at 16 °C or at 4 °C. The arrow indicates the expected migration distance of 9xHis-AsTG. The full-length protein is apparently partly cleaved to exist as a slightly smaller species.

4.3.1.3 Crystallisation trials of 9xHis-AsTG1 are unsuccessful

The 9xHis-AsTG1 purified from 6 L of culture was concentrated to establish its maximum solubility as 39 mg/mL. A first crystallisation screen at 16 °C was attempted at a dilution of 21 mg/mL, but this led to precipitate in most conditions overnight. Subsequent screens were therefore set with 15 mg/mL 9xHis-AsTG1. Six commercial screens were set with 50% or 33% protein in screening solution, but none of the 1,152 conditions screened produced any protein crystals.

4.3.2 Studies of AsTG1 deletion variants with cleavable 9xHis-tag

To improve the chances of crystallisation, four constructs with cleavable His-tagged were designed: 9xHis-3C-AsTG1, 9xHis-3C-AsTG1-ΔN, 9xHis-3C-AsTG1-ΔC and 9xHis-3C-AsTG1-ΔNΔC. In the latter three mutants, either an N-terminal peptide, a C-terminal peptide or both have been removed. These polypeptides correspond to regions that are predicted to be disordered (see section 2.1). This was done by using construct-specific primers for 3C-AsTG1-ΔC and -ΔNΔC then cloning the PCR products into the vector pH9GW. A mutation in the DNA template was unfortunately carried over, so it was corrected by

mutagenesis. For the 3C-AsTG1 and 3C-AsTG1-ΔN constructs, the cloning was unsuccessful despite our best efforts, so they were produced by mutagenesis of the plasmid pH9GW-AsTG1-WT.

4.3.2.1 9xHis-3C-AsTG1-ΔNΔC

The most truncated construct was judged to be most likely to crystallise[226] and was cloned and expressed first. SDS-PAGE analysis of the first purification step revealed a small amount of protein of the right MW in the eluate while the vast majority of the strongly expressed recombinant protein was found in the insoluble fraction (Fig. 4.6). Further purification resulted in decreasing amounts of protein, which were insufficient for crystallisation experiments. This construct was therefore judged to be too insoluble.

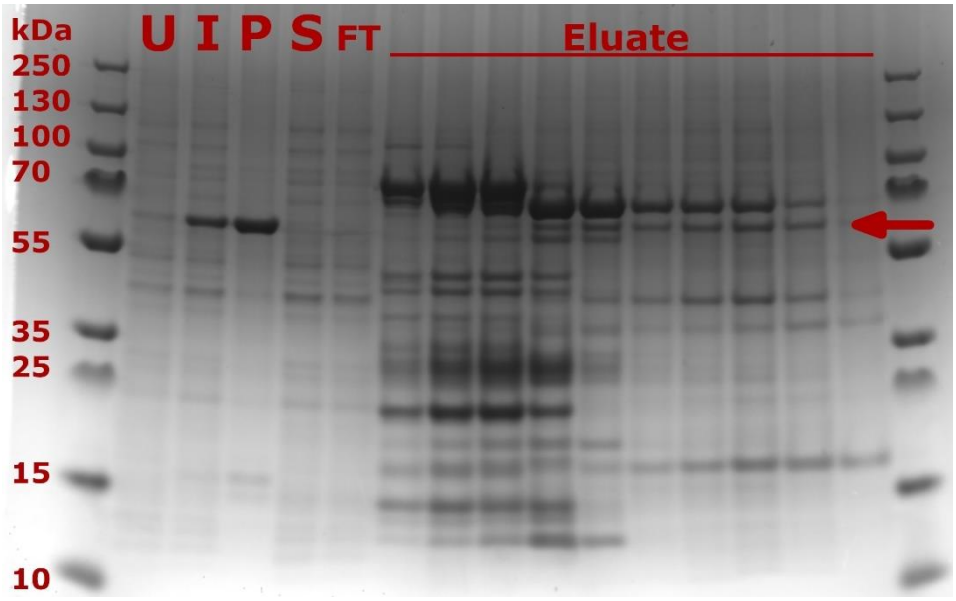


Figure 4.6: SDS-PAGE analysis of expression and first purification step of 9xHis-3C-AsTG1-ΔNΔC (IMAC). 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights, the total protein of the uninduced culture (U) and final induced culture (I) which shows the appearance of the overexpressed protein, the pellet of the lysate (P) which is the insoluble fraction, the soluble fraction of the lysate (S), unbound proteins of the flow-through (FT) and selected eluate fractions. The arrow indicates the expected migration distance of 9xHis-3C-AsTG1-ΔNΔC.

4.3.2.2 9xHis-3C-AsTG1-ΔC

In the hope that the presence of the disordered N-terminus would help with protein solubility, the construct 3C-AsTG1-ΔC was expressed and purified. Unfortunately, it suffered from the same issue of insufficient amount of soluble protein from IMAC (Fig. 4.7).

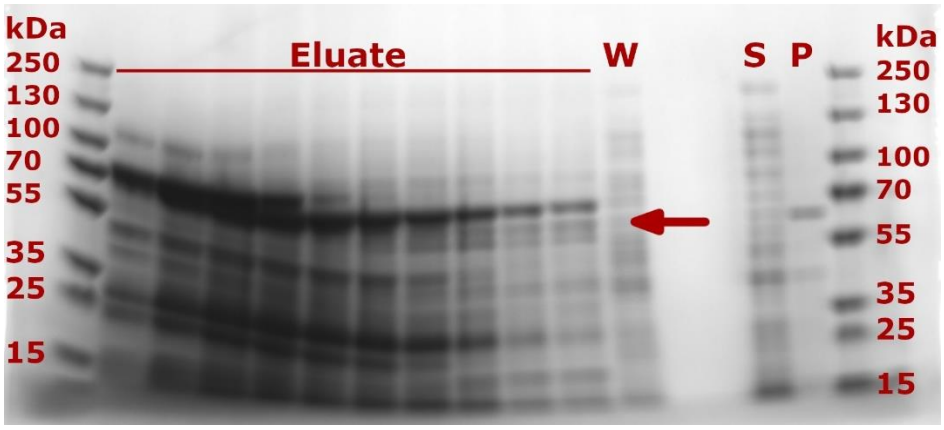


Figure 4.7: SDS-PAGE analysis of expression and first purification step of 9xHis-3C-AsTG1-ΔC (IMAC). 4-12% gradient polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights; consecutive eluate fractions; column wash (W); soluble fraction of the lysate (S); pellet of the lysate (P) which is the insoluble fraction. The arrow indicates the expected migration distance of 9xHis-3C-AsTG1-ΔC.

4.3.2.3 9xHis-3C-AsTG1-ΔN

It may have been the disordered C-terminus that was essential to the solubility of AsTG1. The AsTG1-ΔN was therefore obtained, this time by mutagenesis of the pH9GW-AsTG-WT plasmid to create the truncation mutant and introduce a 3C protease cleavage site. After expression and a first stage IMAC purification, only a very faint SDS-PAGE band of the right apparent MW could be seen in the eluate, which mostly contained a higher-molecular-weight contaminant (Fig. 4.8). This was once again considered insufficient for further studies.

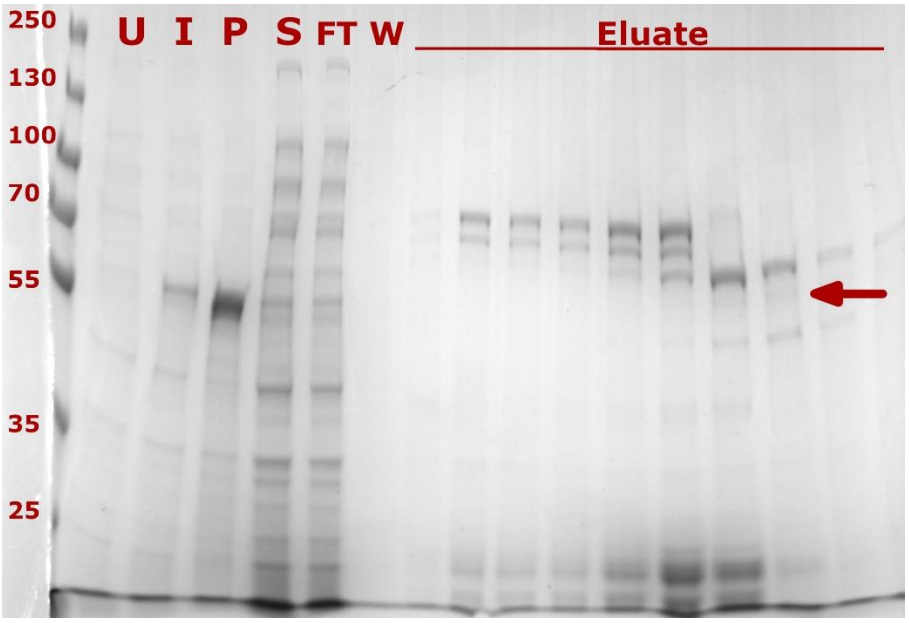


Figure 4.8: SDS-PAGE analysis of expression and first purification step of 9xHis-3C-AsTG1-ΔN (IMAC). 8% polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights in kDa, the total protein of the uninduced culture (U) and final induced culture (I) which shows the appearance of the overexpressed protein, the pellet of the lysate (P) which is the insoluble fraction, the soluble fraction of the lysate (S), unbound proteins of the flow-through (FT), the column wash (W) and selected elution fractions. The arrow indicates its expected migration distance of 9xHis-3C-AsTG1-ΔN.

4.3.2.4 9xHis-3C-AsTG1

With both terminal truncations being deleterious to the solubility of AsTG1, the full-length protein was expressed with a 3C-cleavable His-tag using the pH9GW-3C-AsTG1-WT plasmid, the sequence of which was confirmed by sequencing. Expression following the same method as for the non-cleavable construct did yield a protein of the right size that eluted from a first step of IMAC (Fig. 4.9A). After a tag-cleavage reaction, a second step of IMAC failed to produce the de-tagged protein in the flow-through, but did elute still-tagged protein, as shown by SDS-PAGE and on the chromatogram (Fig. 4.9B). There was no visible appearance of a band of smaller size as is usually seen after protease digest (e.g., in Fig. 3.6). Despite our best efforts, the tag does not appear to be cleavable in this construct, so it was abandoned as it does not confer any advantage over the non-cleavable 9xHis-AsTG1 construct.

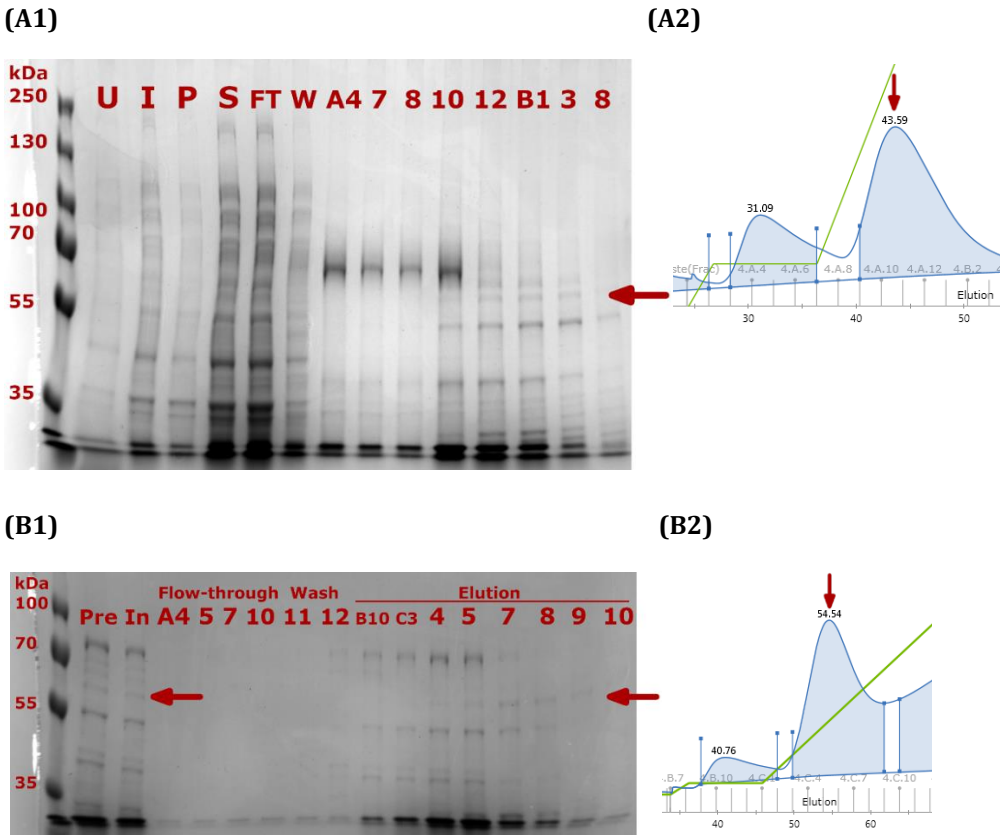


Figure 4.9: SDS-PAGE analyses and chromatograms of the first and last purification step of 9xHis-3C-AsTG1 (IMAC). 8% polyacrylamide SDS-PAGE gels with InstantBlue™ staining with protein standard ladder (molecular weights indicated) and fraction numbers. Chromatograms with the blue curve representing absorbance at 280 nm and the green line representing the proportion of high-imidazole buffer. Fraction names are written in grey on the x-axis. **(A1)** SDS-PAGE analysis of the first IMAC step. From left to right, the lanes show the total protein of the uninduced culture (U) and final induced culture (I), the pellet of the lysate (P) which is the insoluble fraction, the soluble fraction of the lysate (S), unbound proteins of the flow-through (FT), the column wash (W) and the elution fractions. **(A2)** Chromatogram of the first IMAC step. **(B1)** SDS-PAGE analysis of the last IMAC step post-digest. From left to right, the lanes show the protein sample before the tag-cleavage digestion (Pre) and after, as injected into the column (In), then the fraction numbers. **(B2)** Chromatogram of the last IMAC step. The arrows indicate the expected migration distances and elution times of 9xHis-3C-AsTG1.

4.3.3 Homology modelling, sequence alignment and molecular dynamics

In the absence of a crystal structure for AsTG1, a homology model was generated with RosettaCM using the 1.83 Å resolution structure of a rice β-glucosidase (Os3BGlu6, PDB: 3GNO) as a template, as the two enzymes share 45% sequence identity. Alignment of

the homology model with the crystal structure of the glucosylated Os3BGlu6 (PDB: 3WBE) reveals Glu381 as the likely nucleophile of AsTG1. While the preceding residue is almost always a threonine in glycosyl hydrolases, AsTG1 has a histidine in position 380 (Fig. 4.10). This matches three other known GH1-fold transglucosidases: Os9BGlu31, *Dianthus caryophyllus* anthocyanin 5-*O*-glucosyltransferase and *D. grandiflorum* anthocyanin 7-*O*-glucosyltransferase.[212] The Leu244 residue of AsTG1 also stands out as it is present in three other transglucosidases: Os9BGlu31 and two other acyl-glucose-dependent anthocyanin glucosyltransferase from *D. grandiflorum* (UniProt:[215] U6C5K2 and U6C7C6). Other transglucosidases have a tyrosine residue, whereas glycosyl hydrolase tend to have a hydrophilic residue in this position.

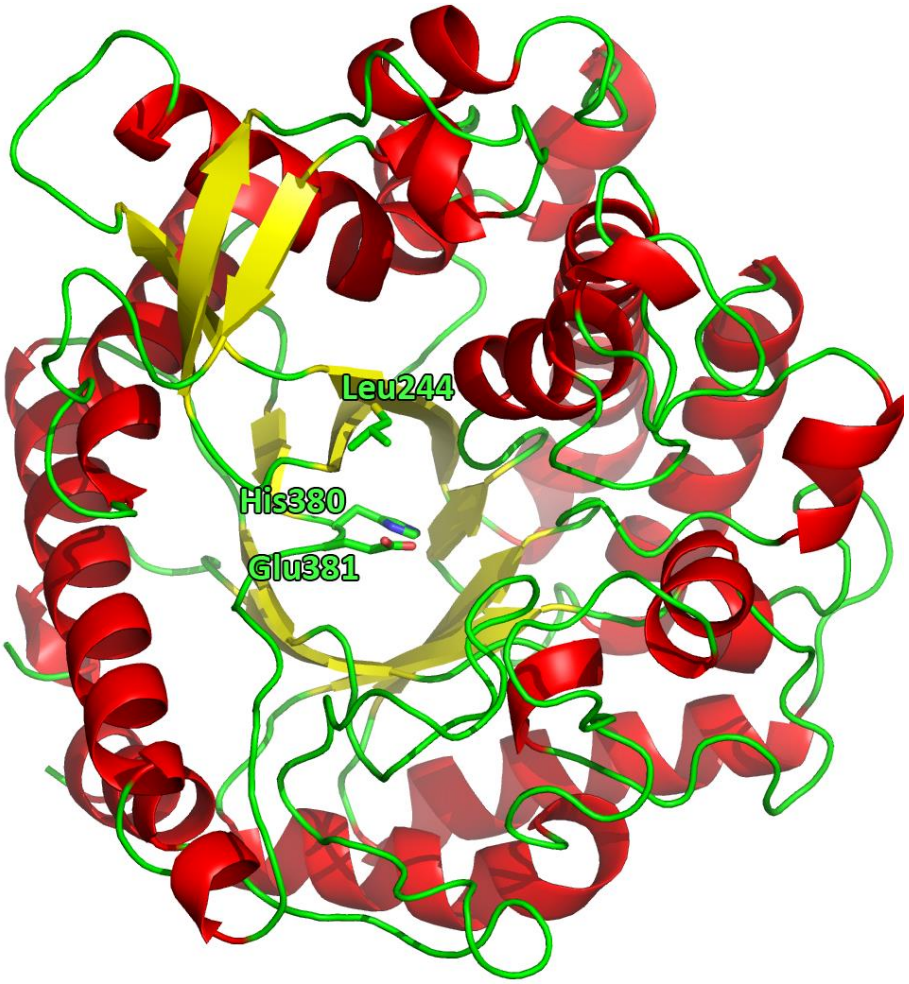


Figure 4.10: Homology model of AsTG1. This was generated with RosettaCM using the structure of Os3BGlu6 (PDB: 3GNO) as a template. The secondary structure is shown in cartoon representation, with α-helices in red, β-sheets in yellow and loops in green. The conserved transglycosidase His and Leu of interest and the nucleophilic Glu are shown as green sticks. Figure prepared using PyMOL[46] and labelled with Photoshop CS2.

To explore the role of His380 and Leu244 in the determinant of transglucosidase over glycosyl hydrolase activity, the glucosylated AsTG1 intermediate was subject to a MD simulation. This results in Leu244 being 6 Å from the anomeric carbon, which may be too far to shield it from water molecules or impede on their deprotonation by the conserved acid-base Glu173. Still, experimental error and the use of a tyrosine residue could mean this mechanism contributes to the avoidance of hydrolysis. Alternatively, hydrophilic residues from glycosyl hydrolases may position Glu173 for effective acid-base catalysis of hydrolysis. During MD, the His380 side chain moves even further from the glucosylated residue, ending at a distance of around 10 Å, so it is not immediately clear how it would prevent hydrolysis or encourage transglycosylation. Upon binding of the sugar acceptor, it may move closer to Glu381 and supply its carbonyl group with a proton, activating the transfer of the sugar.

To better understand this process in AsTG1, its glucosylated intermediate was simulated with its endogenous sugar acceptor, mono-deglucosyl avenacin A1 (MDA). A restraint was applied between the atoms that are known to form a covalent bond in the reaction, so that conformational space could be sampled by the acceptor molecule and productive binding conformations could be identified. The starting position of MDA in the active site was based on that of cellotretraose in its crystal structure in complex with Os3BGlu7, yet as the simulation progresses, MDA rapidly binds to the other side of the active site cleft through hydrophobic interactions (Fig. 4.11A). From this apparently stable bound conformation, the C4 hydroxyl group of the arabinose moiety is likely deprotonated by the Glu173 base catalyst, allowing its nucleophilic attack on the anomeric carbon of the glucose moiety covalently bound to Glu381. This S_N2 mechanism with a sugar acceptor is what allows transglycosylation to generate a 1,4-linked glucose in the product.

Why does this happen *in vitro* with MDA, but not significantly so with the further deglycosylated bis-deglucosyl avenacin A1 (BDA)?[35] A further MD simulation with BDA instead of MDA was run in the same way to understand the effect of the missing 1,2-linked glucose. Instead of binding onto the hydrophobic surface as observed for MDA, BDA would not settle into a specific binding site. Instead, it sampled various conformations compatible with the distance restraint applied to its arabinose O4 atom (Fig. 4.11B). The presumption is that without a defined stable binding mode, BDA is less likely to attain the precise orientation and position needed for transglycosylation.[20]

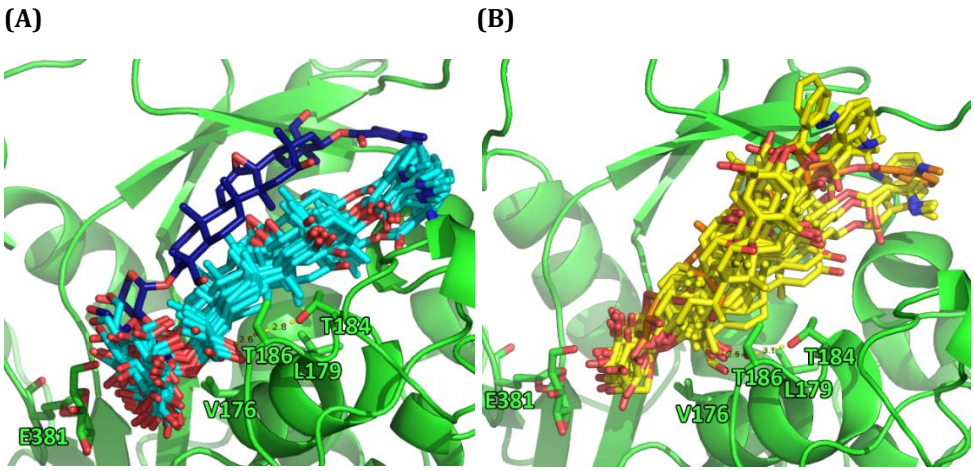


Figure 4.11: MD trajectories of mono- and bis-deglucosyl avenacin A1 in the active site of AsTG1. The E381 residue is glucosylated in the reaction intermediate (green sticks). **(A)** Mono-deglucosyl avenacin A1 is known to be glucosylated by AsTG1. It was positioned into the active site (as dark blue sticks) then subjected to MD simulation, with representative steps shown as cyan sticks. Within 3 ns, the ligand migrates to contact the hydrophobic surface (including the four labelled residues shown as green sticks) and adopts a consistent conformation that is conducive to transglycosylation. **(B)** Bis-deglucosyl avenacin A1 is known not to be significantly glycosylated by AsTG1.[35] It was positioned into the active site (as orange sticks) then subjected to MD simulations (yellow sticks). Without the additional glucose group, the ligand does not remain onto the binding site and becomes conformationally unstable.

4.4 Conclusions and future work

Crystallography-grade AsTG1 with a N-terminal 9xHis-tag was obtained from heterologous expression in *E. coli* Rosetta followed by a 2-step purification, but none of the conditions screened were found to lead to diffracting crystals. To improve the chances of crystallisations, three deletion constructs were designed by excluding polypeptide regions predicted to be disordered. After heterologous expression, purification attempts revealed the constructs to be mostly insoluble. To obtain tag-free full-length enzyme, a 3C protease cleavage site was introduced by mutagenesis, but attempts to remove the 9xHis-tag from AsTG1 proved unsuccessful. To study the structure-function relationship of the enzyme in the absence of crystal structure, a homology model of the enzyme was generated. Along with a multiple sequence alignment, this revealed two residues that may be key to the switch from glycosyl hydrolase to transglucosidase activity. MD simulations of the glucosylated enzyme intermediate suggested a binding site for the mono-deglucosyl

avenacin A1 substrate but not for bis-deglucosyl avenacin A1, which may explain why the latter does not undergo significant transglycosylation *in vitro*.

For future work, as the non-cleavable 9xHis-AsTG1 construct yielded the most soluble enzyme, further crystallisation screening experiments could be attempted with protease spiking (as in section 3.2.1.2) in the hope of forming a smaller unit more likely to crystallise. Alternatively, the natural breakdown product seen by SDS-PAGE (Fig. 4.5) may be isolated and characterised by LC-MS to identify its sequence. This could then be cloned for expression, purification and crystallisation to be attempted.

Because protein solubility was a frequent bottle-neck in this study, it may be worthwhile to use solubility-tags such as MBP, SUMO[120] or Fh8.[227] These protein fusion partners have had significant success in enabling the expression of challenging targets in a soluble form, though crystallisation or tag-removal can be more unreliable.[228] Alternatively, larger quantities of soluble AsTG1 may be obtained from expression in a eukaryotic host such as the yeast *Pichia pastoris*.

His380 was identified as a potential determinant of transglucosidase activity in AsTG1, as it is found in other GH1-fold TGs, while glycosyl hydrolase usually have a threonine residue in this position. This histidine residue is also found in a non-cyanogenic β-glucosidase from *Cicer arietinum* (UniProt:[215] Q700B1), so the enzyme could be assayed for to establish the predictive power of this substitution.

With a method established for MD simulations of glucosylated AsTG1 with sugar acceptors, it could now be possible to screen other potential ligands in the same way to find if they settle into the active site and are therefore more likely to be glycosylated by AsTG1 *in vitro*, offering a route to the enzyme-catalysed synthesis of various glycosides.

Chapter 5: β-amyrin synthase

5.1 Introduction

More than 80 plant OSCs have been functionally characterised,[18] producing around 100 different triterpene skeletons. AsbAS1 is a monofunctional OSC (i.e., it makes a single product)[229] that forms β-amyrin in the first committed step of the avenacin biosynthesis pathway. This membrane-bound enzyme is encoded by the gene *sad1*, which is expressed in the root tip epidermis and the root cap, where avenacin is found, as well as in the root elongation zone.[230] This gene is fully linked with those for some of the other enzymes in the pathway, including two CYPs that oxidise the β-amyrin scaffold.[35, 38] It is also linked, albeit less closely, with the genes for AsAAT1 and AsTG1. This may help the co-regulation of the genes necessary for avenacin biosynthesis.[35] AsbAS1 is an unusual triterpene synthase, as it closely related to sterol synthases,[231] all likely evolved from an ancestral cycloartenol synthase-like gene. The β-amyrin synthase function was selected for in AsbAS1,[38] while in its closest characterised relative, AK070534 from *O. sativa*, it is the production of the tricyclic triterpene achilleol B that was selected for.[231, 232] AsbAS1 is actually more similar to human lanosterol synthase than it is to other plant β-amyrin synthases found in dicots, such as EtAS from pencil cactus (*Euphorbia tirucalli*).[230, 231] While *E. tirucalli* produces large amounts of triterpenes, the biological role of EtAS is unclear,[229] as β-amyrin was not detected in the plant's poisonous latex[233, 234] despite being found in its aerial parts[235] and in the latex of another *Euphorbia* species.[236] Contrary to AsbAS1, the β-amyrin synthase activity of EtAS is well-conserved among its homologs.[237] It is in EtAS that the effect of mutations on product specificity has been most extensively studied.[238–243] However, the structure-function relationships of OSCs in general are complex and poorly understood.[45] High-resolution crystal structure data for a plant OSC would help provide further insights but, as yet, none have had their structure elucidated. The crystal structure of only one oxidosqualene cyclase, human OSC (hOSC),[43] has been solved to date, though structures of a bacterial homologue are available.[42]

The aim of the work described in this chapter was to gain structural insights into the mechanism of AsbAS1 to enable rational engineering of its product specificity. First, extensive attempts were made to express the oat enzyme in both *E. coli* and *P. pastoris* at the levels needed for crystallization, building on previous work by Melissa Salmon.[244] These attempts ultimately proving unsuccessful, attention turned to EtAS, which was

previously shown to be expressed in *E. coli* in functional form. Finally, the publication of the AlphaFold protein structure prediction pipeline[176] allowed prediction of the structures of both enzymes, which were used to design mutations predicted to yield new triterpene scaffolds.

5.2 Methods

5.2.1 Expression of AsbAS1 in *E. coli* BL21

The initial strategy for expression of recombinant AsbAS1 used *E. coli* BL21 as a host and vectors derived from pET-14b and pH9GW.

5.2.1.1 Propagation of the pET-14b-AsbAS1 plasmid

A glycerol stock of *E. coli* BL21(DE3) cells carrying pET-14b-AsbAS1, generated by Melissa Salmon, was plated on LB agar (100 µg/ml carbenicillin) and incubated at 37 °C for 2 d. A resulting colony was restreaked and incubated at 37 °C overnight to avoid satellite colonies. A single colony was inoculated in 50 mL LB (100 µg/ml carbenicillin) for growth at 37 °C and 180 rpm overnight. The plasmid was extracted from 15 mL of the culture using a QIAprep Spin Miniprep Kit (Qiagen) with 30 µL EB incubated 3 min before elution.

5.2.1.2 Cloning of AsbAS1 into the pH9GW vector

Plasmid pDONR207-AsbAS1 (obtained from James Reed, John Innes Centre) was used as an entry clone to transfer the AsbAS1 gene into the destination vector pH9GW and generate the expression plasmid pH9GW-AsbAS1. This was achieved by using 2 µL of LR clonase enzyme mix (Invitrogen) in a 11 µL reaction incubated at 25 °C for 4.5 h. The resulting LR reaction was used immediately to transform *E. coli* DH5α LE cells, which were plated on selective LB agar (50 µg/ml kanamycin, for which pH9GW carries a resistance gene) for growth at 37 °C overnight. Presence of a gene of the right size was confirmed by colony PCR. Verified transformant colonies were grown in 10 mL selective LB (50 µg/ml kanamycin) at 37 °C overnight and the plasmid extracted using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions.

5.2.1.3 Expression of WT AsbAS1 using pH9GW- and pET-14b-derived plasmids

The plasmids pH9GW-AsbAS1 and pET-14b-AsbAS1 were transformed into *E. coli* BL21(DE3) by heat shock and the cells plated on LB agar (50 µg/ml kanamycin or 100 µg/ml carbenicillin, respectively). Transformant colonies were inoculated into 10 mL LB with

appropriate antibiotic and incubated at 37 °C and 180 rpm, until they reached an OD_{600} of 0.6. At this point, half of the cultures were moved to 20 °C 200 rpm and after 10 min of acclimatisation, protein expression was induced overnight in all cultures by addition of IPTG to a final concentration of 1 mM.

Cells were harvested by centrifugation at 3,220 *g* for 10 min, then resuspended in BugBuster® (Novagen) and incubated at 30 °C and 200 rpm for 10 min. After addition of 5 µL of Benzonase® (Sigma Aldrich), the lysis reaction was incubated a further 15 min. The cell debris were separated from the soluble fraction by centrifugation at 18,407 *g* for 15 min. Samples of uninduced culture, final culture, soluble fraction and insoluble fraction for each condition were denatured and analysed by SDS-PAGE.

5.2.2 Periplasmic expression of AsbAS1 in *E. coli* SoluBL21™

In the absence of soluble AsbAS1 following cytoplasmic expression in *E. coli*, it was decided to attempt periplasmic expression. First, the gene was cloned into the pET-22b vector (which adds the pelB secretion signal sequence and a C-terminal His-tag), then expressed in a specialist strain, followed by detergent solubilisation and a three-step purification before crystallisation screening.

5.2.2.1 Cloning of AsbAS1 into the pET-22b vector for periplasmic localisation

To add NcoI and XhoI restriction sites (shown in red and blue, respectively) as well as a C-terminal 8xHis-tag (underlined), the following primers were used (start and stop codon **in bold**):

F-NcoI 5'-GGAAGACCC**ATG**GCGTGGAGGCTAACAATAGGTG-3'

R-XhoI-8xHis 5'-CGTTCCCTCGAG**TTA**GTGATGATGATGATGGTGATG-3'

PCR was carried out using Q5 Hot Start High-Fidelity DNA Polymerase (NEB) with pJH113-AsbAS1 as the template according to the following program: 98 °C for 30 s, followed by 5 cycles of 98 °C for 10 s, 62 °C for 10 s and 72 °C for 90 s, then 30 cycles of 98 °C for 10 s and 72 °C for 100 s, followed by 72 °C for 2 min. A product of the correct size was purified by 0.8% agarose TAE gel electrophoresis with runSafe staining (Cleaver) using a GeneJet Gel Extraction Kit (ThermoFisher) according to the manufacturer's instructions. This purified PCR product and the empty pET-22b vector were each digested using FastDigest™ NcoI and XhoI (ThermoFisher) according to the manufacturer's instructions

then purified using a GeneJet Gel Extraction Kit (ThermoFisher) according to the manufacturer's instructions.

Ligation was set up with 6.5 µL digested pET-22b, 1 µL digested PCR product and 0.5 µL T4 DNA Ligase (Roche) in a 10 µL reaction using the supplied reaction buffer. After incubation at 4 °C overnight, the entire ligation reaction was used to transform into *E. coli* Top10 cells, which were then plated on LB agar (100 µg/mL carbenicillin to select for the pET-22b vector). A transformant was inoculated in 10 mL LB (100 µg/mL carbenicillin) and incubated at 37 °C and 180 rpm overnight. The propagated pET-22b-AsbAS1 plasmid was extracted using GeneJet Plasmid Miniprep Kit (ThermoFisher) and presence of the correct gene with two conservative mutations was confirmed by sequencing using primers specific for the T7 promoter and terminator.

5.2.2.2 Expression of AsbAS1 in *E. coli* SoluBL21 using a pET-22b-derived plasmid

The plasmid pET-22b-AsbAS1 was transformed into *E. coli* SoluBL21™ Chemically Competent cells (Amsbio) by heat shock and the cells plated on TB agar (100 µg/ml carbenicillin). A single transformant colony was inoculated into 10 mL TB (100 µg/ml carbenicillin) and incubated at 37 °C and 180 rpm overnight. The resulting culture was used to prepare a 25% (v/v) glycerol stock for long-term storage at -80 °C. This stock was used to inoculate 500 mL TB (100 µg/ml carbenicillin), which was incubated at 37 °C and 180 rpm overnight. This pre-culture was used to inoculate 40 mL aliquots into 10 x 1 L TB (100 µg/ml carbenicillin) in 2 L baffled conicals and, incubated at 37 °C and 200 rpm until they reached an OD_{600} of 0.4-0.5. At this point, the temperature was reduced to 25 °C. After 30 min of acclimatisation, protein expression was induced at 20 °C for 1 d by addition of IPTG to a final concentration of 0.1 mM. The cells were then harvested by centrifugation at 8,950 *g* and 20 °C for 45 min and stored at -20 °C until needed.

5.2.2.3 Protein purification from SoluBL21

The frozen cells were thawed and resuspended in 150 mL in Buffer 1 (20 mM Tris-HCl pH 8, 300 mM NaCl, 5 mM imidazole) with added 1 mM DTT, DNaseI, 1 mg/mL lysozyme and two tabs of cOmplete EDTA-free protease inhibitor cocktail (Roche). The cells were then lysed by two passages through a cooled cell disruptor (Constant Systems) set to 30,000 psi before addition of Triton™ X-100 to a final concentration of 1% (v/v). The soluble fraction was separated from the cell debris by centrifugation at 5,000 *g* and 4 °C for 10 min and further clarified by centrifugation at 50,000 *g* and 4 °C for 30 min.

The clarified soluble fraction was first purified by immobilised nickel ion affinity chromatography using a 5 mL HisTrap HP column (GE Healthcare) at a flow rate of 5 mL/min on an ÄKTA Prime chromatography system (GE Healthcare) at 4 °C. The column was pre-equilibrated with 5 CV of Buffer 1, before loading the soluble fractions, then washing with 1 CV of Buffer 1, collected as a 5 mL fraction. The bound proteins were eluted with Buffer 3 (Buffer 1 with 300 mM imidazole in total) and collected in 1 mL fractions, which were assessed by running denatured samples on SDS-PAGE. As the eluate showed a band of the right size, the salts were removed for IEC by using a HiPrep 26/10 desalting column (GE Healthcare) at a flow rate of 15 mL/min using an ÄKTA Pure chromatography system (GE Heathcare) at 4 °C. The eluate was collected in 1.5 mL fractions, 9 of which were pooled as they showed UV absorbance on the chromatogram yet did not overlap with a conductance peak from the salts.

A second purification step used anion exchange chromatography on a 5 mL HiTrap Q FF column (GE Healthcare) at a flowrate of 5 mL/min on an ÄKTA Prime chromatography system (GE Healthcare) at 4 °C. The column was pre-equilibrated with 10 CV Low-Salt Buffer (20 mM Tris-HCl pH 7.8, 150 mM NaCl, 1 mM TCEP), before loading the pooled desalted fractions, then eluting with a gradient of 0.15 to 1 M NaCl, resulting from the mixing of Low-Salt Buffer with a proportion of High-salt buffer (20 mM Tris-HCl pH 7.8, 1 M NaCl, 1 mM TCEP) increasing from 0% to 100% over 20 CV002E

The peak fractions were pooled and a third and final purification step was performed by size-exclusion chromatography on a HiLoad 16/600 Superdex 75 pg column (GE Healthcare) pre-equilibrated and eluted at a flow rate of 1 mL/min with Low-salt buffer. The elution was collected in 1.5 mL fractions and results were assessed by running denatured samples of the peak fractions on SDS-PAGE. Fractions containing pure protein were pooled and concentrated to 5.5 mg/mL at 4 °C with an Amicon® Ultra 4 mL Centrifugal Filter (10 kDa MWCO, Merck). The protein concentration was measured with a NanoDrop™ Spectrophotometer (Thermo Scientific) using absorbance at 280 nm and the assumption that 1 mg/mL gives rise to an absorbance of 1 AU.

5.2.2.4 Crystallisation screening of the protein purified from *E. coli* SoluBL21

Crystallisation screening experiments were initiated at 16 °C with five commercially available screens: LMB Crystallization Screen™ from Molecular Dimensions; PEG/Ion Screen™ from Hampton Research; SG1™,[245] MemTrans™ and MemGold2™,[246] all from Protomnis. The screens were set up in 96-well 2-drop MRC plates sealed with ClearVue

Sheets (Molecular Dimensions) employing an OryxNano protein crystallisation robot (Douglas Instruments Ltd.). The sitting drop vapour diffusion technique was used with a drop size of 0.5 µL containing the protein and screen solution at either 3:2 or 1:1 ratio, equilibrated against 60 µL of screen solution per reservoir of the 96-well plate. The screening plates were monitored for crystal formation using a SZX12 Stereo Microscope (Olympus). As many small crystals appeared, further crystallisation screening experiments aimed to obtain a few large crystals by diluting the protein sample to 4 mg/mL in the same Low-salt buffer and a further five crystallisation screens: PACT premier™ Screen[193] and MIDASplus™,[195] both from Molecular Dimensions; Index™ and PEGRx™, and Crystal Screen Cryo™, all from Hampton Research. Crystals were harvested using mounted LithoLoops (Molecular Dimensions) and flash-frozen in liquid nitrogen. See Appendix 2 for the methods of data collection, processing and analysis.

5.2.3 Engineering of solubility in AsbAS1 based on homology models

5.2.3.1 Generation of homology models of AsbAS1 and a soluble homologue for the identification of sites for mutagenesis

As AsbAS1 could be expressed in *E. coli* BL21 using a pET-14b-derived vector, this platform was chosen for a first structure-based engineering project, which aimed to make soluble mutants of the enzyme based on the soluble bacterial homologue *Methylococcus capsulatus* oxidosqualene cyclase (McOSC).[247] In the absence of crystal structures for either enzyme, homology models were generated for both of them using SWISS-MODEL[218] with hOSC (PDB ID: 1W6K)[43] as a template. The two models and the templates were aligned with PyMOL[46] to identify the potential membrane-insertion region of AsbAS1. In this region, hydrophobic residues were selected for mutagenesis when they aligned to hydrophilic residues in the soluble McOSC. A second alignment, using Clustal Omega,[248] gave different results and therefore yielded a different set of mutations. Finally, a mutation introduced a negatively charged residue in an attempt to reduce the protein's affinity for the negatively charged membrane.[249] All of these mutations are reported in Table. 5.1.

5.2.3.2 Mutagenesis of the pET-14b-AsbAS1 plasmid for the generation of a mutant library

Single- and double-site mutagenesis of AsbAS1 was performed with the modified version of the QuikChange method described in section 2.5.[112] The primers were designed based on the AsbAS1 cDNA sequence, choosing a section around the mutation with a T_m of

40-50 °C as calculated by the OligoCalc Nearest Neighbour method.[87] The primer was then extended towards the 3’ end until this non-overlapping region had a T_m 5-10 °C higher than the overlapping region. The resulting design for mutagenesis primers are reported in Table 5.1.

Table 5.1: List of primers used for AsbAS1 solubilisation mutagenesis. The mutations were designed based on homology models, a sequence alignment or a change of charge.

Primer name	Mutation	Origin	Sequence
bAS1.F 1	G325D	Models	5'-CTTATATCTGATTGCCTAACGAAAATTGTGGAGCC-3'
bAS1.R 1	G325D	Models	5'-TAGGCAATCAGATATAAGATCTTGTGCCCGTGAG-3'
bAS1.F 2	A540R	Models	5'-ACATTCCGTTGGTTAGAGGTTCTCAACCCTTCT-3'
bAS1.R 2	A540R	Models	5'-TAACCAACGGAATGTCCGTTTGCATTCGTAGG-3'
bAS1.F 3	W337R & W338R	Align	5'-TTGAATAGGAGGCCAGCAAACAAGCTAAGAGATAGAGC-3'
bAS1.R 3	W337R & W338R	Align	5'-TGGCCTCCTATTCAAAATTGGCTCCACAATTTTCGTT-3'
bAS1.F 4	W337R	Align	5'-TTGAATAGGTGGCCAGCAAACAAGCTAAGAGATAGAGC-3'
bAS1.R 4	W337R	Align	5'-TGGCCACCTATTCAAAATTGGCTCCACAATTTTCGTT-3'
bAS1.F 5	I334R	Models	5'-GGAGCCAAGGTTGAATTGGTGGCCAGC-3'
bAS1.R 5a	I334R	Models	5'-TCAACCTTGGCTCCACAATTTTCGTTAGGCAACCA-3'
bAS1.R 5b	I334R & I330R	Models	5'-TCAACCTTGGCTCCACTCTTTTCGTTAGGCAACCA-3'
bAS1.F 6	P333R	Models	5'-GGAGCGAATTTTGAATTGGTGGCCAGC-3'
bAS1.R 6a	P333R	Models	5'-TCAAAATTCGCTCCACAATTTTCGTTAGGCAACCA-3'
bAS1.R 6b	P333R & I330R	Models	5'-TCAAAATTCGCTCCACTCTTTTCGTTAGGCAACCA-3'
bAS1.F 7a	L322R	Models	5'-CAAGATCGTATATCTGGTTGCCTAACGAAAATTGTGGA-3'
bAS1.F 7b	L322R & I330R	Models	5'-CAAGATCGTATATCTGGTTGCCTAACGAAAAGAGTGGA-3'
bAS1.R 7	L322R	Models	5'-ACCAGATATACGATCTTGTGCCCGTGAGCGT-3'
bAS1.F 8	I330R	Models	5'-ACGAAAAGAGTGGAGCCAATTTTGAATTGGTGGCC-3'
bAS1.R 8	I330R	Models	5'-CTCCACTCTTTTCGTTAGGCAACCAGATATAAGATCTTG-3'
bAS1.F 9	I334E	Charge	5'-GGAGCCAGAGTTGAATTGGTGGCCAGC-3'
bAS1.R 9	I330R & I334E	Mixed	5'-TCAACTCTGGCTCCACCTCTTTCGTTAGGCAACCA-3'

PCR was carried out using Phire Hot Start II DNA polymerase (Thermo Fisher) with 5% DMSO and pDONR207-AsbAS1 as the template according to the following program: 98 °C for 3 min, followed by 30 cycles of 98 °C for 30 s, 50 °C for 1 min and 68 °C for 8 min, followed by 72 °C for 10 min. The presence of a product of the correct size was confirmed by 1% agarose TAE gel electrophoresis with ethidium bromide staining.

To selectively remove any trace of the WT gene, the methylated template DNA was digested by adding 0.5 µL of DpnI (NEB) to each PCR reaction and incubating at 37 °C for 2

h. The DNA was purified into water using a QIAquick® PCR purification kit (Qiagen) and transformed into XL10-Gold® Ultracompetent cells (Agilent) by adding 0.5 µL of purified plasmid to 10 µL of cells, then incubating on ice for 30 min before a 30 s heat shock at 42 °C. After 1-2 min on ice, 90 µL of warm SOC medium was added to the cells, which were then incubated at 37 °C and 600 rpm for pre-growth. The entirety of the transformation reaction was then plated on selective LB agar (100 ug/mL carbenicillin) and grown at 37 °C overnight. Colonies were inoculated into 10 mL LB (100 ug/mL carbenicillin) and grown at 37 °C and 180 rpm overnight. The plasmids were extracted using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions. The mutations were confirmed by sequencing using the following internal sequencing primers:

bAS1.FOR seq 5'-TTTGTTGGGCCTATTAGTCC-3'

bAS1.REV seq 5'-TAGAGTCGGGATAAACTCGC-3'

The plasmid carrying the mutant gene AsbAS1-A540R was used as a template for a second round of mutagenesis using the same method and primer pair 1, 4, 7a or 8, yielding a set of four plasmids bearing double-mutations.

To obtain a plasmid carrying the AsbAS1 gene with its entire main membrane-anchoring helix (S317 to R344) swapped for the equivalent from the soluble McOSC (S212 to R234), mutagenesis was also performed using the same method but a 55 °C annealing temperature and the following primers (helix in italics, bAS1-specific region unformatted):

bAS1.FOR HelSwap 5'-*CGCGTTACGAGCGTCGGCCTTGGAAGGC* GCTAAGAGATAGAGCTTTAACTAAC-3'

bAS1.REV HelSwap 5'-*CCAGAAGATCGTAAACCAGCCTCAGGAC* CCGTGAGCGTGGGTAATGAAGGTC-3'

An additional ligation step was performed before transformation by adding T4 DNA ligase (1 µL) and T4 polynulceotide kinase (1 µL) to 12.5 µL of PCR product in a 20 µL reaction in T4 ligase ATP-containing buffer (NEB).

5.2.3.3 Expression and western blot analysis of AsbAS1 solubilisation mutants

The mutagenized plasmids were transformed into *E. coli* BL21(DE3) as above, using 0.5 µL plasmid for 30 µL cells. Single colonies were inoculated into 10 mL LB (100 ug/mL carbenicillin) and incubated at 37 °C and 180 rpm to an OD_{600} of 0.5-0.6 before adding IPTG to a concentration of 1 mM for induction of protein expression overnight. Cells were

harvested by centrifugation at 3,220 *g* for 10 min and resuspended in 0.4 mL BugBuster® Protein Extraction Reagent (Millipore) with 0.1 mini-tab cOmplete EDTA-free protease inhibitor (Roche) and 0.2 µL Benzonase (Sigma-Aldrich) before incubation at 30 °C and 200 rpm for 30 min. The lysate was spun down at 18,407 *g* and both the soluble and insoluble fraction were denatured and run SDS-PAGE, along samples of uninduced and final culture. The proteins were transferred to a nitrocellulose membrane using an iBlot 2 Gel Transfer Device and Regular Stacks (Invitrogen) with the P0 method (20 V for 1 min, 23 V for 4 min then 25 V for 2 min). The membrane was incubated in blocking solution (5% dried skimmed milk, 20 mM Tris HCl pH 7.5 and 7.5 mM NaCl) at 4 °C overnight. A 1:10,000 dilution of rabbit anti-bAS1 primary antibodies (from Melissa Salmon) in blocking solution was then incubated with the membrane at r.t for 13 min, before washing with TBSTT (20 mM Tris HCl pH 7.5, 0.5 mM NaCl, 0.05% Tween® 20, 0.2% Triton X-100) for 5, 5, then 10 min. A 1:2,000 dilution of goat anti-rabbit,HRP secondary antibodies (Thermo Scientific) in blocking solution was incubated with the washed membrane for 5 min at r.t. For detection, it was then incubated with SuperSignal™ West Pico PLUS working solution for 5 min before imaging with a G:Box imager and GeneSys software (Syngene).

5.2.5 Expression of AsbAS1 *Pichia pastoris*

Attempts were made to use restriction/ligation cloning and the pPICZB vector to clone WT AsbAS1 and a number of variants thereof containing cleavable and non-cleavable N- and C-terminal affinity tags, as well as deletion mutant AsbAS1-ΔC lacking a C-terminal peptide predicted to be disordered. This was supplemented by cloning the WT enzyme using the Gateway method with the secretory vectors pBGP1-DEST and pPICZα-DEST, which respectively remain episomal or integrate the gene into the host's chromosome.

5.2.5.1 Cloning of AsbAS1 constructs into the pPICZB vector

Eight AsbAS1 constructs were designed, including two ΔC constructs lacking the last four residues predicted to be disordered using the standard protocol described earlier (see section 2.1). The constructs are summarised in Table 5.2.

Table 5.2: AsbAS1 constructs designed for expression in *P. pastoris* using the pPICZB vector. The primers used to amplify the AsbAS1 gene in each case are indicated and their sequences are reported below.

Name	Construct name	Primers used
(1) AsbAS1	Native	F1 R1
(2) 6xHis-AsbAS1	N-terminal 6xHis-tag	F2 R1
(3) AsbAS1-6xHis	C-terminal 6xHis-tag	F1 R2
(4) 9xHis-3C-AsbAS1	3C-cleavable N-terminal 9xHis-tag	F3 R1
(5) AsbAS1-9xHis	C-terminal 9xHis-tag	F1 R3
(6) AsbAS1-ΔC-9xHis	C-terminal 9xHis-tag ΔC	F1 R4
(7) 9xHis-3C-AsbAS1-ΔC	3C-cleavable N-terminal 9xHis-tag ΔC	F3 R5
(8) Strep-3C-AsbAS1	3C-cleavable N-terminal strep-II-tag	F4 R1

To add PmlI and XbaI restriction sites (shown in red and blue, respectively), His-tags or strep-II-tag (underlined) and 3C protease cleavage sites (*in italics*), the following primers were used (start and stop codon **in bold**):

F1-native 5'-GTACATCACGTGACC**ATG**TGGAGGCTAACAATAGGTGAGG-3'

F2-6xHis 5'-GTACATCACGTGACC**ATG**CATCACCATCACCATCAC TGGAGGCTAACAATAGGTGAG-3'

F3-9xHis-3C 5'-GTACATCACGTGACC**ATG**CATCACCATCACCATCACCATCACCAT *TTGGAAGTGTTATTTCAGGGCCCG*TGGAGGCTAACAATAGGTGAG-3'

F4-Strep-3C 5'-GTACATCACGTGACC**ATG**TGGAGCCATCCGCAGTTTGAAAAA *TTGGAAGTGTTATTTCAGGGCCCG*TGGAGGCTAACAATAGGTGAG-3'

R1-native 5'-CATGTATCTAGA**TTA**GCTCTTAATCGCAAGAAGTCGACGGC-3'

R2-6xHis 5'-CATGTATCTAGA**TTA**GTGATGGTGATGGTGATG GCTCTTAATCGCAAGAAGTCGACG-3'

R3-9xHis 5'-CATGTATCTAGA**TTA**GTGATGGTGATGGTGATGGTGATGGTG GCTCTTAATCGCAAGAAGTCGACG-3'

R4-9xHis-ΔC 5'-CATGTATCTAGA**TTA**GTGATGGTGATGGTGATGGTGATGGTG AAGAAGTCGACGGCGAAGTTCCC-3'

R5-ΔC 5'-CATGTATCTAGA**TTA**AAGAAGTCGACGGCGAAGTTCCC-3'

PCR was carried out using Phire Hot Start II DNA polymerase (Thermo Fisher) with 5% DMSO and pDONR207-AsbAS1 as the template according to the following program: 98 °C for 1 min, followed by 32 cycles of 98 °C for 10 s, 55 °C for 15 s and 72 °C for 45 s, followed by 72 °C for 2 min. Products of the correct size were observed by 0.8% agarose TAE gel electrophoresis with ethidium bromide staining, and these fragments were purified using a QIAquick® PCR purification kit (Qiagen) with 30 µL EB elution. The PCR products and the empty pPICZB vector were each digested using PmlI and XbaI (NEB) according to the manufacturer's instructions then purified using a QIAquick® PCR purification kit (Qiagen) with 50 µL EB elution.

Ligation was set up with 40 ng digested pPICZB, 80 ng digested PCR product and 1 U T4 DNA Ligase (Invitrogen) in the supplied reaction buffer. After incubation at 14 °C overnight, the ligation reaction was diluted 5-fold to transform it into *E. coli* DH5α LE cells, which were then plated on LB agar (25 µg/mL Zeocin™). Transformants were analysed by colony PCR using AsbAS1-specific primers and empty pPICZB vector transformants as a negative control. The colony showing an insert of the right size was incubated in 10 mL LB (25 µg/mL Zeocin™) overnight and the plasmid extracted. Presence of the correct gene was confirmed only for one of the constructs, pPICZB-AsbAS1-(6), by sequencing (Eurofins Genomic) using primers specific for the AOX1 promoter and terminator:

F-AOX1 5'-GACTGGTTCCAATTGACAAGC-3'

R-AOX1 5'-GCAAATGGCATTCTGACATCC-3'

5.2.5.2 Cloning of the AsbAS1 gene into the pBGP1-DEST and pPICZα-DEST vectors

Plasmid pDONR207-AsbAS1 was used as an entry clone to transfer the AsbAS1 gene into the destination vectors pBGP1-DEST and pPICZα-DEST to generate the secretory expression plasmids pBGP1-DEST-AsbAS1 and pPICZα-DEST-AsbAS1. This was achieved by using 1 µL of LR clonase enzyme mix (Invitrogen), 75 ng destination vector and 75 ng entry vector made up to 4 µL with TE. The reaction was incubated at 25 °C for 8 h before arresting with 0.5 µL proteinase K. A 0.8 µL aliquot was used as is to transform *E. coli* DH5α LE cells (40 µL), which were plated on selective low-salt LB agar (25 µg/mL Zeocin™) for growth at 37 °C overnight. Transformant colonies were grown in 10 mL selective LB (25 µg/mL Zeocin™) at 37 °C overnight and the plasmid extracted using the QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer's instructions.

5.2.5.3 Transformation of *P. pastoris* X-33 and KM71H(*OCH1::G418R*) with the plasmids pPICZB-AsbAS1-(6), pBGP1-DEST-AsbAS1 and pPICZα-DEST-AsbAS1

To prepare competent cells, *P. pastoris* X-33 and KM71H(*OCH1::G418R*) were first revived from glycerol concentrated stocks by inoculating 50 µL into 5 mL YPD media (1% yeast extract, 2% Bacto™ peptone, 2% dextrose) and growing at 30 °C and 200 rpm overnight. The resulting cultures were then used to inoculate 50 mL YPD for growth at 30 °C and 200 rpm overnight. When the cultures reached an OD_{600} of 1, the cells were harvested by centrifugation at 3,220 *g* and r.t for 5 min, washed with 50 mL autoclaved dH_2O, re-spun, washed with 50 mL SED buffer (1 M sorbitol, 20 mM DTT, 25 mM EDTA pH 8.0), re-spun, washed with 20 mL ice-cold 1 M sorbitol, re-spun at 4 °C, then resuspended with 1 mL ice-cold 1 M sorbitol.

The circular plasmids pPICZB-AsbAS1-(6) and pPICZα-DEST-AsbAS1, and the empty pPICZB and pPICZα-DEST vectors were linearised by digesting with SacI-HF (NEB) using 1 µL per 1 µg DNA in CutSmart® buffer and incubating at 37 °C for 15 min. The efficiency of cleavage was verified by running samples on 1% agarose TAE gel by electrophoresis with ethidium bromide staining.

Competent X-33 and KM71H(*OCH1::G418R*) aliquots (80 µL) were transformed with either linearised pPICZB-AsbAS1-(6), circular pBGP1-DEST-AsbAS1, linearised pPICZα-DEST-AsbAS1 or their empty vector equivalents using a MicroPulser electroporator (Bio-Rad) with a single 1.5 kV pulse. The transformed cells were resuspended in 1 mL ice-cold 1 M sorbitol then incubated at 30 °C for 1 h 45 min before plating on selective YPDS agar (YPD with 1 M sorbitol and 2% agar, with selection by 100 µg/mL Zeocin™) and incubating at 30 °C for 4 d.

Transformant colonies of pPICZB-AsbAS1-(6), pPICZα-DEST-AsbAS1 and their empty vector equivalents in X-33 were screened for the Mut+ phenotype by streaking them on MM agar (0.34% w/v yeast nitrogen base, 1% w/v ammonium sulphate, 0.4 µg/mL biotin, 1% v/v methanol, 1.5% agar) and MD agar (i.e., MM agar with the methanol replaced by 2% w/v dextrose) with 100 µg/mL Zeocin™. Colonies that grew well on both MD agar and MM agar were inoculated into 10 mL YPD and incubated at 30 °C and 200 rpm for 4 – 7 d. All cultures, except for pBGP1-DEST-AsbAS1 in X-33 that always failed to grow, were centrifuged at 3,220 *g* and r.t for 5 min. The media was discarded and the pellet used to prepare 25% (v/v) glycerol stocks to be stored at –80 °C

5.2.5.4 Expression of AsbAS1 in *P. pastoris* and evaluation by SDS-PAGE and western blot

The frozen glycerol stocks of X-33 transformed with pPICZα-DEST-AsbAS1 (two transformants) or pPICZB-AsbAS1-(6) (three transformants), of KM71H(*OCH1::G418R*) transformed with pPICZB-AsbAS1-(6) (two transformants), pBGP1-DEST-AsbAS1 (one transformant) or pPICZα-DEST-AsbAS1 (two transformants), and of the no-vector controls for each were used to inoculate 10 µL into 10 mL BMGY (100 mM potassium phosphate pH 6, 1% w/v yeast extract, 2% w/v Bacto-peptone™,1.34% w/v yeast nitrogen base, 0.4 µg/mL biotin, 1% v/v glycerol with selection by 100 µg/mL Zeocin™), which were incubated at 30 °C and 200 rpm for 4 d. The cells were harvested by centrifugation at 3,220 *g* and r.t for 5 min then resuspended into 50 mL BMMY (as for BMGY but with 0.5% v/v methanol instead of 1% v/v glycerol) for incubation at 30 °C and 250 rpm for 3 d, with supplementation of 0.5% v/v methanol every day. The cells were harvested by centrifugation at 5,250 *g* and r.t for 15 min. Samples of all cultures and of the supernatants from pPICZα-DEST-derived transformants were analysed by SDS-PAGE with InstantBlue™ staining. A repeat SDS-PAGE gel was analysed by western blot, essentially as described for the expression of AsbAS1 mutants, but using IRDye® 800 CW goat anti-rat secondary antibodies and imaged using an Odyssey CLX system.

5.2.5.5 Extraction of β-amyrin from *P. pastoris*

The cells were extracted with ethyl acetate for GC-MS analysis to detect any β-amyrin produced by recombinant AsbAS1. First, the water was removed using a CoolSafe freeze-dryer (ScanVac) until it reached a pressure of 0.9 Pa. The dried pellets were crushed and extracted using 5 mL ethyl acetate by incubation at 65 °C and 200 rpm for 1 h. The insoluble fraction was separated by a first centrifugation at 5,250 *g* and r.t for 30 min and a second centrifugation of 1 mL of the resulting supernatant at 13,000 rpm and 4 °C for 10 min. A 750 µL aliquot was taken without disturbing the pellet for each extract, then mixed with 250 µL of ethyl acetate with 200 µg/mL coprostanol (Abcam) as internal standard.

5.2.5.6 GC-MS analysis of culture extracts

Samples were analysed by GC-MS using a Shimadzu GCMS-QP2010s fitted with an AOC-20s autosampler and a Shimadzu SHIM-5MS column (30 m x 0.25 mm, 0.25 µm film thickness). Helium was used as a with a column flow rate of 0.9 mL/min. The instrument was operated in split mode (1:5 split) with a sample injection volume of 1 µl. The inlet was

held at 280 °C, and the interface and ion source at 250 °C and 200 °C, respectively. The oven was programmed to rise from 200 to 310 °C at 10 °C/min, hold for 3 min, rise from 310 to 350 °C at 10 °C/min, and hold at 350 °C for 6 min. Initially, mass spectra of standards were obtained in scan mode operated from m/z 50 – 800. For quantification, selected ion monitoring (SIM) mode was used, detecting ions shown in Table 5.3.

Table 5.3: Compounds identified by GC-MS.

Compound	Retention time (min)	Quantifier ion (m/z)	Qualifier ion (m/z)
Squalene	10.31	69	81, 95, 68, 136
2,3-oxidosqualene	11.19	69	81, 71
Coprostanol	12.93	233	215, 55
β-amyrin	15.88	218	203, 69, 95

Peak integration was performed using Shimadzu software GCMS solution v.2.50 and quantification was based on a four-point calibration curve prepared with 10-fold dilutions of β-amyrin standard (0.2 – 200 µg/mL), or three-point calibration curves for squalene (2.5, 12.5, 25 µg/mL) and 2,3-oxidosqualene (1, 5, 10 µg/mL), with reference to the internal standard (coprostanol). Analytic standards of squalene (product code 442785), 2,3-oxidosqualene (product code 21719-5mg) and β-amyrin (product code 09236-10MG-F) were bought from Sigma-Aldrich.

5.2.6 Expression of EtAS in *E. coli*

To establish an *in vivo* functional assay of a β-amyrin synthase in a readily manipulated host, EtAS was expressed in *E. coli*, with or without genes involved in biosynthesis of its substrate, 2,3-oxidosqualene. In addition, this protein could also serve as an alternative crystallisation target if it expressed at high levels.

5.2.6.1 Preparation of plasmids

Plasmids pETD-EtAS (containing the *Euphorbia tirucalli* β-amyrin synthase gene) and pAC-HpIDI/AtSQS/AtSQE (containing the *Haematococcus pluvialis* isopentenyl diphosphate isomerase gene, *HpIDI*; the *A. thaliana* squalene synthase gene, *AtSQS*; and the *A. thaliana* squalene epoxidase gene, *AtSQE*; respectively) were constructed and kindly provided by Dr. Miho Takemura (Ishikawa Prefectural University, Japan).[250] Freeze-dried plasmids were dissolved in 10 mM Tris-HCl (pH 8.5) buffer and 50 ng of each was transformed into 100 µL *E. coli* Top10 competent cells by heat shock. After 1 h of pre-

growth, the cells were plated onto LB agar (100 µg/mL carbenicillin for pETD-EtAS and 10 µg/mL tetracycline for pAC-HpIDI/AtSQS/AtSQE) and incubated at 37 °C overnight. A single colony for each plasmid was inoculated into 5 mL LB with appropriate antibiotics and incubated at 37 °C with shaking for overnight. The plasmids were extracted using a GeneJET Plasmid Miniprep Kit (ThermoFisher). The 3' sequence of plasmid pETD-EtAS was confirmed by Sanger sequencing with T7 reverse primer at Source BioScience.

5.2.6.2 Production of squalene, 2,3-oxidosqualene and β-amyrin in E. coli BL21(DE3)

E. coli BL21(DE3) cells were transformed with plasmids pAC-HpIDI/AtSQS/AtSQE and pETD-EtAS, either alone and in combination, by heat shock. Transformants inoculated into 10 mL LB supplemented with appropriate antibiotics and incubated at 37 °C and 180 rpm overnight. These pre-cultures were used to inoculate 100 µL into 10 mL 2X YT media (1.6% w/v tryptone, 1% w/v yeast extract, 5% w/v NaCl, pH 7.0) supplemented with appropriate antibiotics and then incubated at 37 °C and 180 rpm until they reached an OD_{600} of 0.8-1.0. At this point, protein expression was induced by the addition of isopropyl β-D-thiogalactopyranoside (IPTG) to a final concentration of 0.05 mM for 2 days at 20 °C. Cells were harvested by centrifugation at 3,220 *g* for 10 min.

5.2.6.3 Extraction and analysis of squalene, 2,3-oxidosqualene and β-amyrin

The cell pellets from duplicate cultures were resuspended in 1 mL STE buffer (10 mM Tris-HCl pH8.0, 1 mM EDTA, 0.1 M NaCl) to be washed and divided into 3 microcentrifuge tubes. After centrifugation at 3,220 *g* for 10 min, the supernatant was discarded and the cell pellets were extracted with 3x200 µL of chloroform-methanol (2:1) by vortexing for 10 min then recovering the organic phase by centrifugation at 3,220 *g* for 10 min. The combined organic phases were evaporated and the resulting extracts were dissolved in 100 µ ethyl acetate (with 50 µg/mL coprostanol as an internal standard) for gas chromatography–mass spectrometry (GC-MS) analysis as described in section 5.2.5.6.

5.2.6.4 Overexpression of EtAS in *E. coli*

E. coli BL21(DE3) and SHuffle® T7 Express *lysY* cells were transformed with plasmids pETD-EtAS by heat shock. Transformant colonies were inoculated into 10 mL LB (100 µg/mL carbenicillin for pETD-EtAS and 30 µg/ml chloramphenicol to select for *lysY*) and incubated at 37 °C and 180 rpm overnight. The pre-cultures were used to inoculate 100 µL into 10 mL LB or 2X YT media (100 µg/mL carbenicillin and 30 µg/ml chloramphenicol) which were incubated at 37 or 30 °C to an OD_{600} of 0.4–1.5. Protein expression was induced for 3 to 4 h

or overnight at 15, 20 or 25 °C by addition of IPTG to a final concentration of 0.05 or 0.1 mM. Cells were harvested by centrifugation at 5,000 *g* for 12 min then stored at -20 °C for further analysis. Cell pellets from 1 mL aliquots of culture were resuspended in 100-200 µL lysis buffer (20 mM Tris-HCl pH 8 and 300 mM NaCl supplemented with cOmplete™, EDTA-free Protease Inhibitor Cocktail) and lysed by sonication using Soniprep 150 Plus Ultrasonic Disintegrator (MSE). Samples were analysed by SDS-PAGE.

5.2.7 Molecular models for plant OSCs and their complex with β-amyrin

Reviewed UniProt entries for four β-amyrin synthases (from *Bruguiera gymnorhiza*, *A. thaliana*, *Pisum sativum* and *Glycyrrhiza glabra*) were found by looking for proteins with over 50% sequence identity with EtAS on UniProt (entry Q401R6).[215] This also led to the entries for the glutinol and friedelin synthases from *Kalanchoe daigremontiana*. Their precomputed AlphaFold2 structure predictions were downloaded and their alignment analysed in PyMOL.[46]

Models of the structures of EtAS and AsbAS1 were generated by submitting their sequence to the AlphaFold Colab notebook.[251] The protonation states of histidine, aspartate and glutamate residues were then modified by reference to the results of prediction by the H++ server.[224] Unoptimized Cartesian coordinates for β-amyrin (bA) were obtained from the isomeric SMILES representations of the corresponding entry in the PubChem database[222] transformed to Cartesian space using the electronic Ligand Builder and Optimization Workbench (eLBOW) tool in Phenix.[171] Semiempirical QM-optimized atomic coordinates and atomic partial charges, together with force field parameters, were obtained from the Automated Topology Builder (ATB) version 3.0.[223] To generate starting coordinates for the complex of AsbAS1 with bA (AsbAS1:bA), AsbAS1 was first superimposed onto the crystal structure of hOSC in complex with lanosterol (PDB: 1W6K). A bA molecule, geometry-optimized in ATB, was then manually positioned and oriented to best overlay its A and B rings with the corresponding rings of lanosterol while avoiding short van der Waals contacts with residues in the AsbAS1 active site cavity. Energy minimization of the AsbAS1:bA complex was performed with the GROMACS 2020.4 molecular dynamics package[182] with the amber99sb-ildn force field.[201] Energy minimization was performed in aqueous solution with the complex at the centre of a cubic box with 10 Å distance from the centre of the protein to the edge of the box. No attempt was made to introduce lipid molecules which would presumably be present. The box was solvated by the Simple Point Charge (SPC) water model, adding sodium counter ions to

ensure neutral charge of the system. The system was then subjected to energy minimisation employing periodic boundary conditions and using the conjugate gradient optimization method until the system energy decrease between steps was less than 1 kcal mol^{-1}. This occurred after 1673 cycles. No interatomic or positional restraints were applied. Analysis of the energy minimized model was carried out using PyMOL.[46]

5.3 Results and discussion

5.3.1 Cytoplasmic expression of AsbAS1 in *E. coli* BL21

A first expression trial in *E. coli* BL21 aimed to reproduce the work of Melissa Salmon, which yielded inclusion bodies of a tag-less construct of WT AsbAS1, while attempting the expression of a His-tagged construct for a potential refolding purification using IMAC.

The AsbAS1 gene was cloned into the pH9GW vector using Gateway cloning to add an N-terminal 9xHis-tag. The gene had also been cloned by Melissa Salmon into the pET-14b vector using the restriction enzymes NcoI and BamHI, which removes the His-tag from the vector. The resulting plasmids pH9GW-AsbAS1 and pET-14b-AsbAS1 were then used for an expression trial in *E. coli* BL21. SDS-PAGE analysis revealed the strongest band of the appropriate apparent MW to be from inducing protein expression at 37 °C using the pET-14b vector, with less protein from expression at 20 °C. Unfortunately, no major band was seen for 9xHis-tagged AsbAS1 from the induced cultures using the pH9GW-derived plasmid (Fig. 5.1).

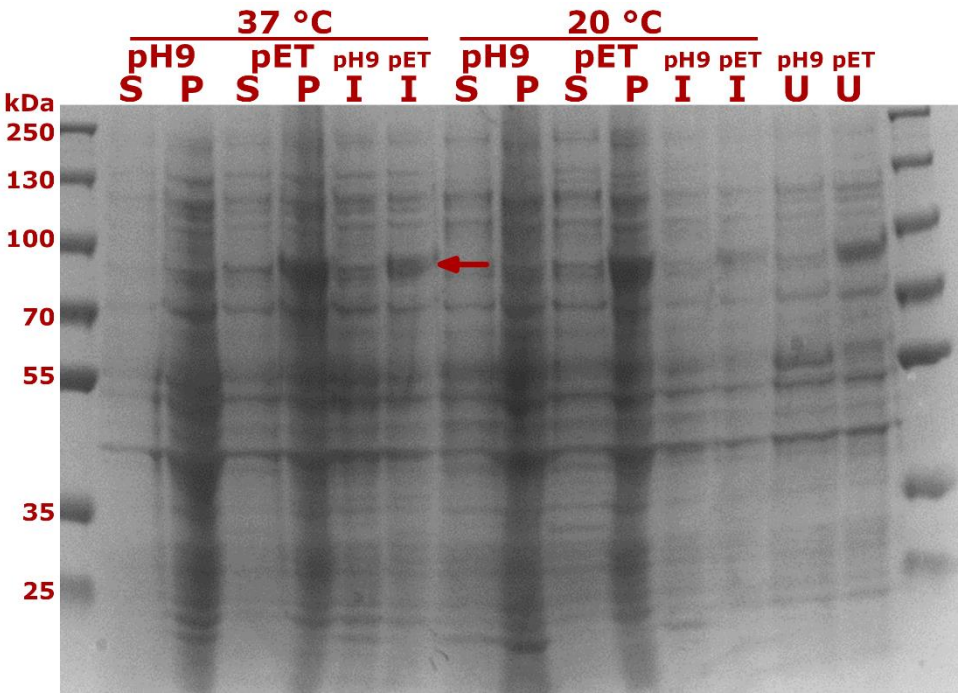


Figure 5.1: SDS-PAGE analysis of the AsbAS1 expression screen in *E. coli* BL21. 10% polyacrylamide SDS-PAGE with InstantBlue™ staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated on the left; then for cultures using the plasmids derived from pH9GW (pH9) and pET-14b (pET), the soluble fraction of the lysate (S); pellet of the lysate (P), i.e., the insoluble fraction; the total protein of the final induced culture (I) and uninduced culture (U). The arrow indicates the expected migration distance of AsbAS1.

The AsbAS1 produced in *E. coli* using pET-14b is known to form inclusion bodies and the lack of affinity tag would render purification difficult.[244] It was therefore decided to find a strategy for the production of folded His-tagged AsbAS1. After an unsuccessful attempt at Ni-NTA bead purification with n-dodecyl-β-D-maltoside (DDM) from *E. coli* BL21 expressing a pJH113-derived plasmid (data not shown), it was decided to target the protein to outside of the cytoplasm.

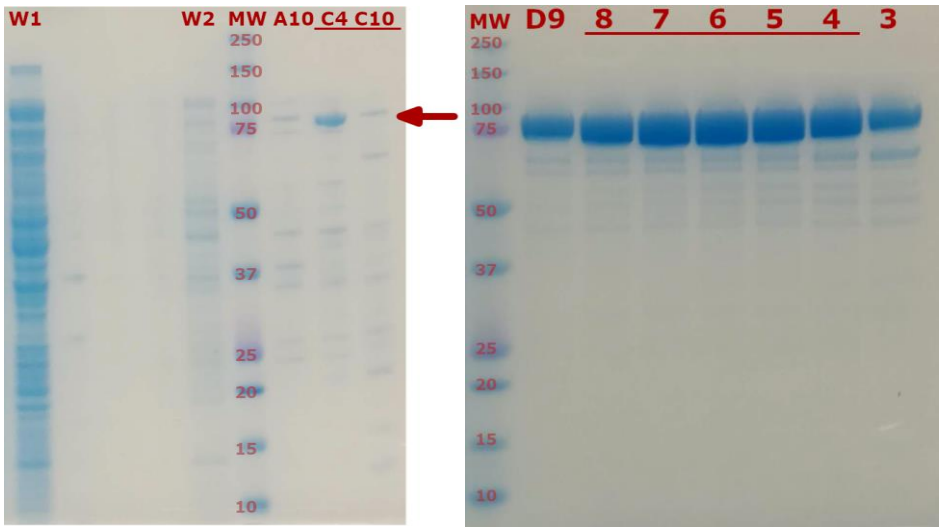
5.3.2 Periplasmic expression of AsbAS1-6xHis in *E. coli* SoluBL21

Recombinant proteins have better chances of folding properly in the periplasm of *E. coli* instead of its cytoplasm. Indeed, periplasmic expression can avoid the formation of inclusion bodies.[252] Furthermore, disulphide bonds cannot form in the *E. coli* cytoplasm, which is a reductive environment due to the action of redoxin enzymes. The periplasm, however, contains the disulphide-forming enzymes of the Dsb system (covered in more detail in section 2.7.2.4).[253] AsbAS1 contains 21 cysteine residues, some of which could be

forming disulphide bonds that are critical to folding (e.g., Cys614 and Cys724, which are close to each other according to the homology model described in section 5.2.3.1). It was therefore decided to target AsbAS1 to the periplasm of *E. coli* SoluBL21™ (strain described in section 2.7.2.6) in an attempt to produce it in a properly folded form.

This consisted in cloning the AsbAS1 gene in the pET-22b vector, which adds the pelB secretion signal sequence upstream of the gene and a C-terminal 6xHis-tag,[252] and inducing expression in *E. coli* SoluBL21™ at 20 °C. The lysate was incubated with Triton™ X-100 to attempt extracting AsbAS1 from the membrane lipids.[254] A first purification step by IMAC yielded an SDS-PAGE band of the right size in the eluate (Fig. 5.2A), so it was further purified by anion exchange chromatography then SEC (Fig. 5.2B).

(A1) **(B1)**



(A2) **(B2)**

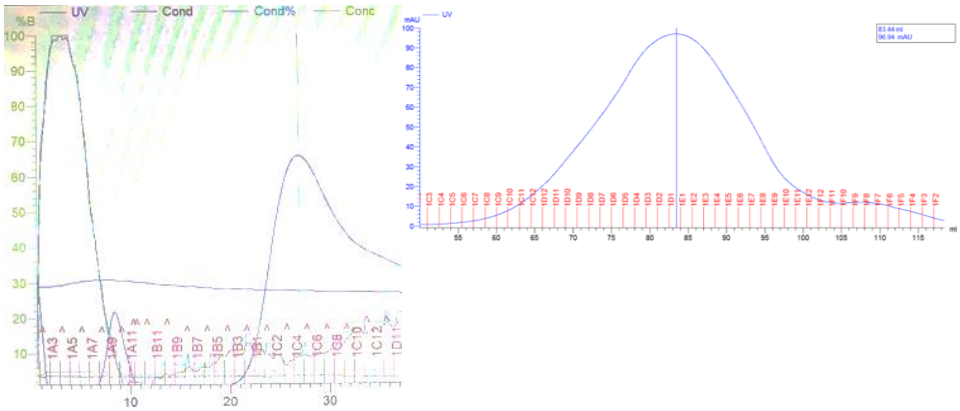


Figure 5.2: SDS-PAGE analyses and chromatograms of the purification of AsbAS1-6xHis. 4-12% gradient polyacrylamide SDS-PAGE with Quick Coomassie staining. Chromatograms with the blue curve representing absorbance at 280 nm. Fraction numbers are written in red on the x-axis. **(A1)** SDS-PAGE analysis of the first purification step (IMAC). From left to right, the lanes show the first (W1) and second wash (W2), the protein standard ladder with molecular weights indicated in kDa, and the eluate fraction numbers with the pooled fractions underlined. The arrow indicates the expected migration distance of AsbAS1-6xHis. **(A2)** Chromatogram of the first purification step (IMAC). **(B1)** SDS-PAGE analysis of the third purification step (SEC). From left to right, the lanes show the protein standard ladder with molecular weights indicated in kDa, and the eluate fraction numbers with the pooled fractions underlined. **(B2)** Chromatogram of the third purification step (SEC).

The purified protein appeared green, which was unexpected for AsbAS1. We therefore analysed the sample by UV-vis spectroscopy and observed a peak at 410 nm (data not shown), which is characteristic of a haem.[255] Crystallisation screens yielded

numerous crystals, some of which were analysed by western blot using anti-AsbAS1 antibodies, which did not show any significant band (data not shown). X-ray diffraction data was collected, and the cell dimensions (±1 Å) were searched on the PDB. A matching structure of the haem-containing *E. coli* HPII catalase was found (PDB: 4BFL) and used for molecular replacement, which was successful. The structure of the HPII catalase contaminant was solved and analysis is reported in Appendix 2.

As a way of making the expression, purification and crystallisation of AsbAS1 more tractable, it was decided to design and express a library of mutants with an altered membrane-interacting surface.

5.3.3 Design and expression of AsbAS1 solubilisation mutants in *E. coli* BL21

A structure-informed engineering project for AsbAS1 attempted to make this membrane-bound protein soluble by taking inspiration from a soluble bacterial homologue, McOSC.[247] This involved an analysis of homology models and sequence alignments to design mutants, which were then expressed in *E. coli* BL21. A soluble AsbAS1 mutant would help crystallisation efforts, as well as having applications for biocatalysis.

A homology model of AsbAS1 was generated using the structure of hOSC[43] as a template. Comparison with the original analysis of the hOSC structure led to inferences about the position of the membrane-insertion region of AsbAS1, including a major α-helix (from Ser317 to Asn336, corresponding to hOSC Ser290 to His306) and a minor α-helix (from Phe539 to Ser548, corresponding to hOSC His510 to Ser528). A homology model of the soluble McOSC was also created in the same way. When aligned to the AsbAS1 model, it reveals hydrophobic to hydrophilic substitutions in the membrane-insertion region (Fig. 5.3). A sequence alignment using Clustal Omega[248] revealed two more substitutions. These were used to design AsbAS1 mutants that could potentially be soluble.

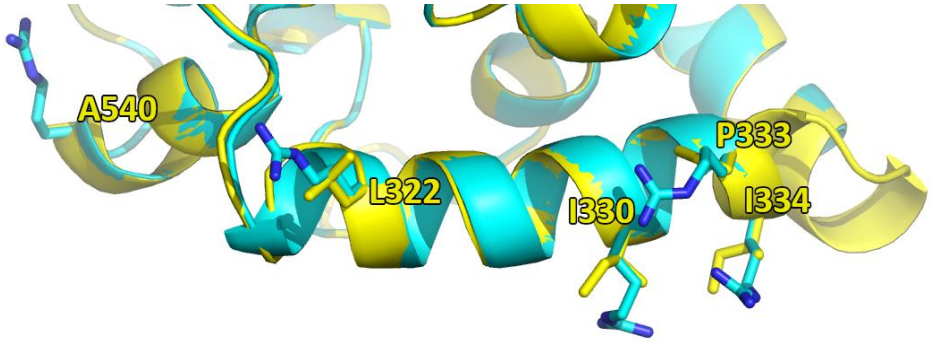


Figure 5.3: Membrane insertion region of AsbAS1 and its equivalent in McOSC. Homology model of AsbAS1 (yellow) aligned with a homology model of McOSC (cyan), both generated using SWISS-MODEL with human oxidosqualene cyclase as a template (PDB ID: 1W6K). Hydrophobic AsbAS1 residues shown as yellow sticks were mutated to arginine residues individually or as pairs.

For mutagenesis, the modified version of the QuikChange method[112] (described in section 2.5) was attempted on the pET-22b- and pET-14b-AsbAS1 plasmids. While the pET-22b-derived plasmid suffered from low success-rate and non-specific amplification (data not shown), the same method on the pET-14b-AsbAS1 plasmid proved to be remarkably effective for generating mutants (Fig. 5.4).

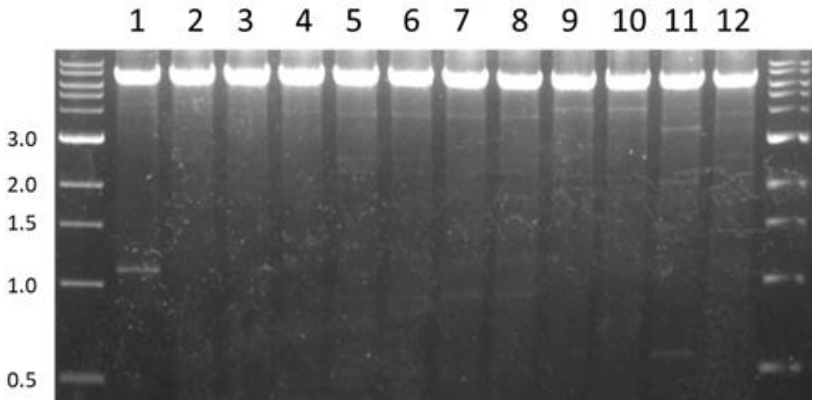


Figure 5.4: Analysis of the AsbAS1 mutagenesis reactions by agarose gel electrophoresis. The mutant numbers (described in table 5.4) are indicated above the lanes. The sizes in kb of the standard DNA markers are indicated on the left. A negative control reaction with DNA template but not primer did not reveal any visible band by the same analysis.

Four plasmids with a double mutation were prepared with the same method by using a single-mutation template and primers covering a different region. An additional mutant was created, in which the entire main membrane-insertion helix of AsbAS1 was swapped for the equivalent helix in McOSC. This resulted in a library of 17 mutants, summarised in table 5.4.

Table 5.4: AsbAS1 mutant library created in an attempt to solubilise this membrane-bound enzyme. For each mutant, this table indicates the name assigned, the primers used in mutagenesis (see table 5.1 for primer sequences) and the mutations introduced.

Mutant name	Primers used	Mutation
1	bAS1.F 5 bAS1.R 5a	I334R
2	bAS1.F 5 bAS1.R 5b	I334R & I330R
3	bAS1.F 6 bAS1.R 6a	P333R
4	bAS1.F 6 bAS1.R 6b	P333R & I330R
5	bAS1.F 9 bAS1.R 9	I330E & I334E
6	bAS1.F 2 bAS1.R 2 bAS1.F 1 bAS1.R 1	A540R & G325D
7	bAS1.F 2 bAS1.R 2 bAS1.F 4 bAS1.R 4	A540R & W337R
8	bAS1.F 2 bAS1.R 2 bAS1.F 7a bAS1.R 7	A540R & L322R
9	bAS1.F 2 bAS1.R 2 bAS1.F 8 bAS1.R 8	A540R & I330R

Mutant name	Primers used	Mutation
9	bAS1.F 2 bAS1.R 2 bAS1.F 8 bAS1.R 8	A540R & I330R
10	bAS1.F 1 bAS1.R 1	G325D
11	bAS1.F 2 bAS1.R 2	A540R
12	bAS1.F 3 bAS1.R 3	W337R & W338R
13	bAS1.F 4 bAS1.R 4	W337R
14	bAS1.F 7a bAS1.R 7	L322R
15	bAS1.F 7b bAS1.R 7	L322R & I330R
Helix swap	bAS1.F HelSwap bAS1.R HelSwap	S317 to R344 swapped for McOSC S212 to R234

After expression of all mutants in *E. coli* BL21, the cells were lysed and the soluble fractions recovered by ultra-centrifugation. They were then analysed by western blot using antibodies raised against AsbAS1. Unfortunately, no significant band was seen in these fractions for any of the mutants (eight of which are shown in Fig. 5.5. This could be because the mutations failed to solubilise AsbAS1 or because this detection method might not be sensitive enough for the amounts of protein present (see section 5.3.4). Alternatively, the issue may reside in the inability of *E. coli* BL21 to fold AsbAS1 properly, resulting in the

formation of inclusion bodies. This is common for recombinant proteins expressed in *E. coli*, but it is sometimes possible to refold them using denaturing agents.[121] Unfortunately, our attempts using urea, DDM and Ni-NTA bead purification of His-tagged AsbAS1 did not succeed (data not shown).

Figure 5.5: Western blot analysis of the pelleted cell debris and soluble fractions of the lysate from *E. coli* expression of AsbAS1 mutants (numbered). 4-12% gradient acrylamide SDS-PAGE transferred to a nitrocellulose membrane, exposed to rabbit anti-AsbAS1 primary antibodies then anti-rabbit,HRP secondary antibodies; developed with SuperSignal™ West Pico Plus Chemiluminescent Substrate. The mutant names are indicated at the top of the lanes. The line indicates the expected migration distance of AsbAS1.

5.3.4 Expression of AsbAS1 in *P. pastoris*

As the prokaryotic expression host *E. coli* once again failed to produce folded AsbAS1, expression of this eukaryotic protein was undertaken in the methylotrophic yeast *Pichia pastoris* (*Komagataella phaffii*),[256] which has an excellent track record for structural studies of membrane-bound proteins.[132] This work built on the results of Melissa Salmon, who demonstrated intracellular expression of functional AsbAS1 in this system, but could not fully purify the recombinant enzyme.[244]

In an attempt to generate the large quantities of purified protein required for X-ray crystallographic analysis, expression was attempted in two strains of *P. pastoris*: the wild-type X-33 strain, whose methanol-utilisation phenotype (Mut^+) allows rapid growth on methanol (pending phenotypic screening after transformation), and the glycoengineered KM71H(*OCH1::G418R*) strain, which, despite growing slowly on methanol (Mut^S

phenotype), has the advantage of being less prone to protein hyperglycosylation. The resulting polysaccharides could otherwise lead to heterogenous protein samples that are less likely to crystallise.[134]

Eight constructs of AsbAS1 were designed with various affinity tags, including deletion mutants lacking the C-terminal region predicted to be disordered, and generated by PCR before attempting to clone them. Unfortunately, despite screening a total of 88 transformants by colony PCR to detect the presence of the AsbAS1 gene, only one of the constructs, AsbAS1-ΔC-9xHis, was successfully cloned into the vector pPICZB. The low success rate may be either due to loss of activity of the restriction enzymes, which could be alleviated by using fresh stocks, or due to the choice of blunt-end instead of sticky-end cloning. Before further attempts at cloning the other constructs, which may facilitate purification or crystallisation, the pPICZB-AsbAS1-ΔC-9xHis clone was taken forward into expression screening to evaluate the suitability of this vector compared to others.

AsbAS1 was also cloned into the gateway-compatible vectors pBGP1-DEST and pPICZα-DEST.[257] These add the prepro-α-factor signal sequence upstream of the cloned gene, which is removed by endogenous proteases but targets the resulting protein to the extracellular space.[258] This protein secretion strategy has been used in the past to express other membrane-bound proteins for crystallisation.[132] The pPICZ vectors need to be linearised and transformation happens through their integration into the host's chromosome by homologous recombination with the inducible AOX1 promoter. This can happen several times, so the resulting transformants may have multiple copies of the cloned gene, which usually leads to higher levels of protein expression.[259] This is why it can be worthwhile to check recombinant protein expression levels in several transformants. The episomal pBPG1-DEST vector, on the other hand, uses the constitutive GAP promoter and can be transformed as a circular plasmid, which will be maintained as an independent genetic element. This vector therefore leads to consistent transcription levels, which can be advantageous, e.g., for detecting changes in activity between mutants *in vivo*.[257] The constitutive promoter also simplifies higher-throughput screening of activity because it removes the induction step that requires a monitoring of the cell density.[260]

P. pastoris X-33 and KM71H(*OCH1::G418R*) transformants were therefore obtained for AsbAS1 cloned in pPICZB, pBGP1-DEST and pPICZα-DEST. Along with their vector-only controls, they were screened for AsbAS1 expression. SDS-PAGE analysis of the total culture and culture medium was inconclusive because a band of the right size is seen by Coomassie

stain even in the absence of the AsbAS1 gene. This probably corresponds to the AOX1 protein, which would obscure AsbAS1 (data not shown). A second detection method consisted in western blot analysis using antibodies raised against AsbAS1 expressed in *E. coli*. Unfortunately, no significant band was observed in any of the cultures (Fig. 5.6).

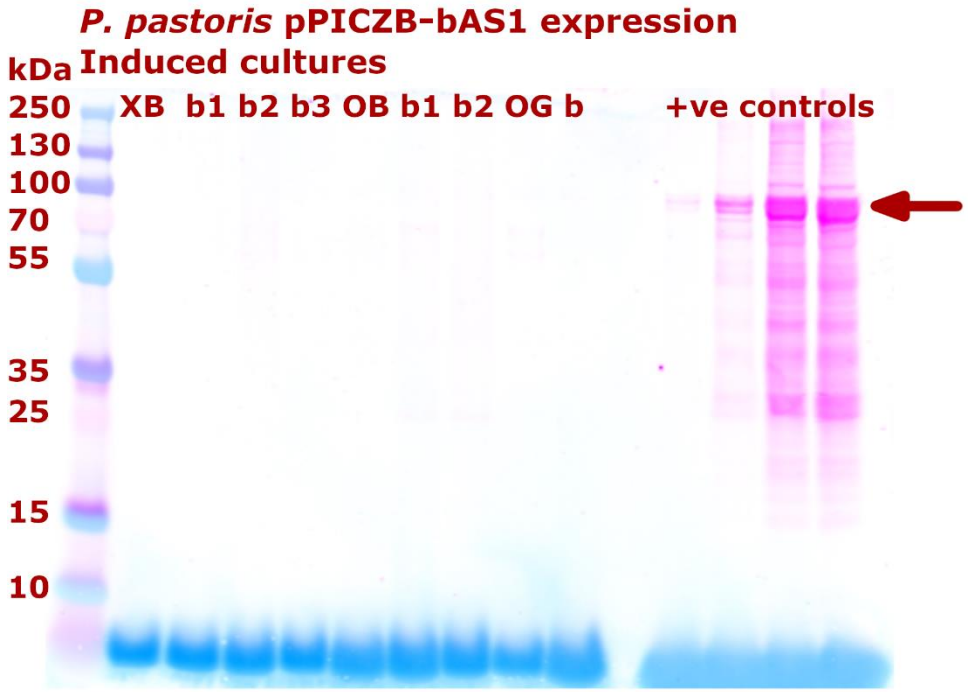


Figure 5.6: Western blot analysis of AsbAS1 expression in *Pichia pastoris*. 4-12% gradient acrylamide SDS-PAGE transferred to a nitrocellulose membrane, exposed to rabbit anti-AsbAS1 primary antibodies then IRDye® 800 CW goat anti-rat secondary antibodies. From left to right, the lanes show the protein standard ladder with molecular weights, the total protein content of the final induced culture of X-33 transformed with the pPICZB empty vector (XB) or pPICZB-AsbAS1 (transformants b1, b2 and b3) and of KM71H(*OCH1::G418R*) transformed with the pPICZB empty vector (OB), pPICZB-AsbAS1 (transformants b1 and b2), pBGP1-DEST empty vector (OG) or pBGP1-DEST-AsbAS1 (b). The positive control is the insoluble fraction of AsbAS1 expressed in *E. coli*. The arrow indicates its expected migration distance.

Given that western blotting failed to detect AsbAS1 expressed from any of the constructs tested in *P. pastoris*, a third and final strategy for the quantification of recombinant AsbAS1 in *P. pastoris* consisted in detecting its β-amyrin product. Indeed, while yeast does not produce this compound, it does make the oxidosqualene precursor as part of the ergosterol biosynthesis pathway.[261] It is therefore possible for active recombinant AsbAS1 to use endogenous oxidosqualene to produce β-amyrin. The dried cells from each culture were extracted using ethyl acetate and the organic fractions

analysed by GC-MS. This revealed the presence of β-amyrin in the X-33 and KM71H(*OCH1::G418R*) cells transformed with pPICZB-AsbAS1 but no other culture (Table 5.5). This may be because the pBGP1-DEST-AsbAS1 and pPICZα-DEST-AsbAS1 plasmids do not produce active protein or because oxidosqualene cannot reach extracellular protein. As X-33 and KM71H(*OCH1::G418R*) were equivalent in AsbAS1 activity, it may be more judicious to use the latter for large-scale expression and purification for the purpose of crystallisation, because of the reduced hyperglycosylation of proteins in this system.

Table 5.5: GC-MS analysis of *P. pastoris* culture extracts. β-amyrin was detected in the pPICZB-AsbAS1 transformants, with the two highest producers shown in bold.

Strain	Vector	Clone	β-amyrin (μg/mL)
KM71H(OCH1::G418R)	pPICZB	Empty vector	0.00
		AsbAS1-A	0.25
		AsbAS1-B	**0.32**
	pBGP1-DEST	Empty vector	0.00
		AsbAS1	0.00
	pPICZα-DEST	Empty vector	0.00
		AsbAS1-A	0.00
		AsbAS1-B	0.00
X-33	pPICZB	Empty vector	0.00
		AsbAS1-A	0.19
		AsbAS1-B	**0.27**
		AsbAS1-C	0.20
	pPICZα-DEST	Empty vector	0.00
		AsbAS1-A	0.00
		AsbAS1-B	0.00

5.3.5 Expression of EtAS in *E. coli*

The results for GC-MS showed that AsbAS1 can be expressed in an active form in *P. pastoris*, albeit at presumably low levels given that it was not detected by western blotting. However, the use of the pPICZB vector requires numerous steps and does not guarantee all transformants to be equivalent. This makes the strategy inconvenient for medium-throughput engineering efforts. It was therefore decided to investigate the capacity of *E. coli* to express active β-amyrin synthase from an episomal vector and analyse the enzyme's product using GC-MS so that the effect of mutations could be observed more conveniently.

Unlike *P. pastoris*, *E. coli* does not produce 2,3-oxidosqualene. It therefore requires a squalene epoxidase (SQE) gene to be introduced. To increase the levels of the squalene precursor, it is possible to co-express genes for squalene synthase (SQS) and for isopentyl diphosphate delta-isomerase (IDI), which contributes earlier in the biosynthetic pathway. A method reported in the literature[250] used genes from *A. thaliana* and the green algae *Haematococcus pluvialis* cloned into a vector with the artificial promoter *tac*, yielding the

plasmid pAC-HpIDI/AtSQS/AtSQE. Because this pAC vector carries a p15a origin of replication,[262] a β-amyrin synthase gene can be co-expressed if it is cloned into a vector with a ColE1 replicon, such as the plasmid pETD-EtAS, which carries the gene for *Euphorbia tirucalli* β-amyrin synthase, a well-characterised OSC. This method was reproduced here: after co-expression of pAC-HpIDI/AtSQS/AtSQE and pETD-EtAS in *E. coli* BL21, cell extracts were analysed by GC-MS. When both plasmids are present, β-amyrin is successfully produced (2.66 μg/mL in the extract); but in controls with either plasmid on their own, none is observed (Table 5.6), which confirms the absence of oxidosqualene and β-amyrin synthase in *E. coli*. Compared to the culture expressing pAC-HpIDI/AtSQS/AtSQE alone, the cells expressing the additional pETD-EtAS plasmid seem to have consumed all the available oxidosqualene. They also had a lower cell density, probably due to the added metabolic burden,[263] which may explain the lower levels of total triterpenes.

Table 5.6: GC-MS analysis of β-amyrin and its precursors in *E. coli* expressing genes of its biosynthetic pathway. The plasmids pAC-HpIDI/AtSQS/AtSQE and pETD-EtAS were expressed alone or in combination in *E. coli* BL21 cells, the organic extracts of which analysed alongside standards by GC-MS to estimate the concentrations of squalene, oxidosqualene and β-amyrin.

Genes expressed	Squalene (μg/mL)	2,3-oxidosqualene (μg/mL)	β-amyrin (μg/mL)
HpIDI/AtSQS/AtSQE	18.41	24.70	0.00
EtAS	0.29	0.00	0.00
HpIDI/AtSQS/AtSQE and *EtAS*	7.21	0.00	2.66

With EtAS being produced in an active form in *E. coli*, it would be worthwhile to attempt its overexpression with a view to obtaining a crystal structure, which would be the first for a plant OSC and could be used as a better template than hOSC for an AsbAS1 homology model.

The pETD-EtAS plasmid was therefore used for a protein expression screen in *E. coli* BL21(DE3) and SHuffle® T7 Express lysY cells, using various induction temperatures and durations. SDS-PAGE analysis did reveal overexpression of a protein, but it was not visible in the soluble fraction of the lysate and its apparent MW was significantly smaller than the predicted MW of 87.6 kDa (Fig. 5.7). DNA sequencing of the 3' end of the gene did not reveal any premature stop codon.

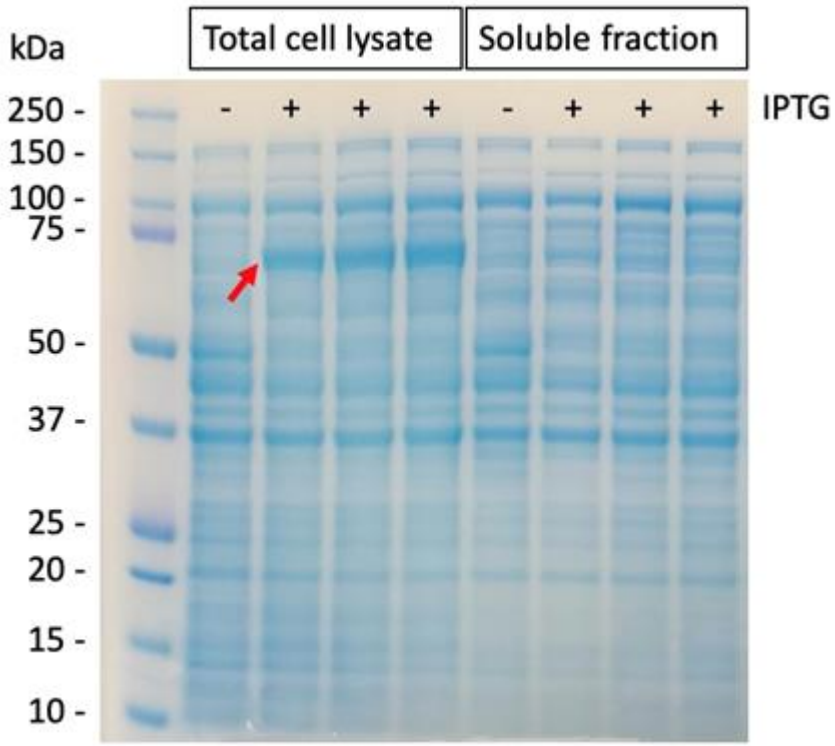


Figure 5.7: SDS-PAGE analysis of EtAS expression in BL21(DE3). 4-12% gradient polyacrylamide SDS-PAGE with Quick Coomassie staining. From left to right, the lanes show the protein standard ladder with molecular weights indicated in kDa, the total proteins of the IPTG-free control (-) and the three IPTG-induced replicates (+) and the soluble fraction of the same lysates. The arrow indicates the induced protein.

It may still be worth attempting a larger-scale expression and purification by IMAC to obtain enough protein for characterisation (e.g., by LC-MS) to identify the overexpressed protein. Alternatively, expression in *P. pastoris* may yield more soluble EtAS.

With the *in vivo* activity assay established for EtAS in *E. coli*, it would now be possible to measure the activity of the 17 solubilisation mutants by transforming their pET-14b clones with the compatible pAC-HpIDI/AtSQS/AtSQE plasmid for expression and GC-MS analysis. Unfortunately, time restraints did not permit these experiments.

5.3.6 Towards the rational engineering of product specificity in AsbAS1

In the absence of crystal structures, molecular models for AsbAS1 and EtAS were created using AlphaFold2. These may be more accurate and less biased than the homology models generated as described in section 5.2.3.1. A further four β-amyrin synthases as well as a friedelin synthase and a glutinol synthase had precomputed AlphaFold models from their reviewed UniProt entries[215] which were used for comparison.

The aligned models revealed AsbAS1 to have quite an unusual active site (Fig. 5.8), likely because it is the only β-amyrin synthase in the set to come from monocots. Indeed, all other β-amyrin synthase models analysed here are from dicots, which evolved separately from the monocots β-amyrin synthases. In monocots, they evolved from sterol OSCs, retaining some of their features (such as Tyr533 in AsbAS1) while acquiring others to coax the OS precursor into a chair-chair-chair conformation to yield triterpenes instead.[244] The dicot β-amyrin synthases, on the other hand, have evolved from a different sterol OSC and use a tryptophan residue in the position of AsbAS1 Tyr533 to achieve the CCC substrate conformation required for triterpene biosynthesis.

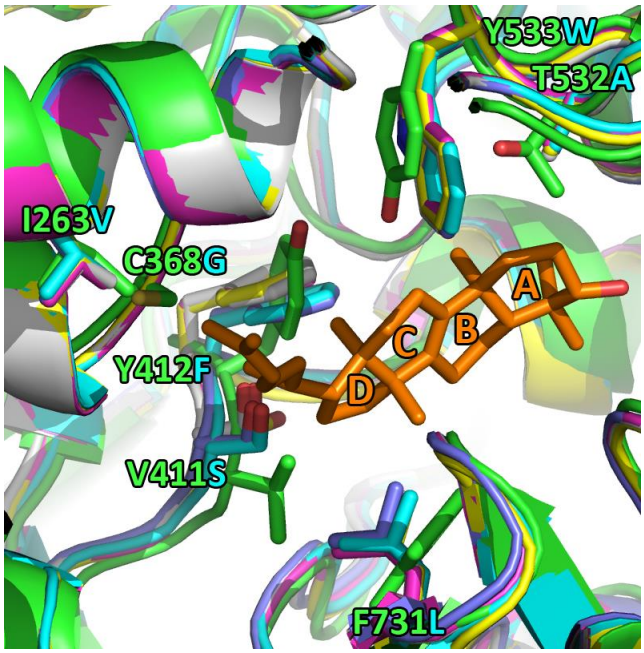


Figure 5.8: Active site of AsbAS1 and four other β-amyrin synthases. The molecular models of the β-amyrin synthases from the monocot *A. strigosa* and the dicots *Euphorbia tirucalli* (cyan), *Bruguiera gymnorhiza* (magenta), *Arabidopsis thaliana* (yellow), *Pisum sativum* (white) and *Glycyrrhiza glabra* (grey-blue) were generated by AlphaFold2. A lanosterol molecule (orange sticks with its rings labelled) was placed based on the alignment with the crystal structure of hOSC (PDB: 1W6K). Residues from AsbAS1 (green sticks and labels) that differ from conserved residues in dicot β-amyrin synthases (coloured sticks and cyan labels) are indicated. Figure prepared using PyMOL[46] and labelled with Photoshop CS2.

However, some residues remain conserved or conservatively substituted, allowing the design of mutations that could alter the product specificity of AsbAS1. Analysis of a multiple sequence alignment of plant OSCs[244] reveals the Phe259 residue (a conserved tyrosine in dicots) to be substituted for a histidine residue in the dozens of characterised

cycloartenol and lanosterol synthases. The Phe259His mutation may therefore change the product of AsbAS1 from β-amyrin to cycloartenol or lanosterol, especially because it already has a Tyr533 residue that is characteristic of these sterol OSCs. Product specificity for cycloartenol may be further enhanced by the mutating the Val482 residue, conserved in all β-amyrin synthases, to isoleucine, which is conserved in all 17 characterised cycloartenol synthases.

Trp257 is conserved in all 19 β-amyrin synthases in the sequence alignment,[244] except in the *P. sativum* triterpene synthase that produces some lupeol as one of its minor products. The latter enzyme and all five lupeol synthases have a leucine residue in this position. The Trp to Leu mutation has been used to make lupeol the major product of a dicot β-amyrin synthase before,[264] so it would be interesting to see if it would have the same effect in the monocot AsbAS1.

Finally, the Val482Leu mutation would take inspiration from the from the glutinol and friedelin synthases from *Kalanchoe daigremontiana*. These make the two triterpenes that result from the intermediates with a carbocation on the A ring, closest to the 3-hydroxyl group. Based on the docking of β-amyrin in AsbAS1, it seems the longer side chain may push the A ring closer to the π-cloud of Trp611, which would potentially stabilise the positive charge (Fig. 5.9).

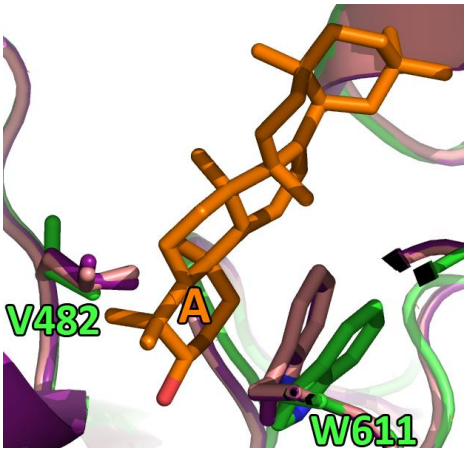


Figure 5.9: The A ring of β-amyrin is sandwiched between Val482 and Trp611. Superposition of the AlphaFold models of glutinol and frieldelin synthases from *Kalanchoe daigremontiana* (purple and salmon, respectively) with the energy-minimised model of AsbAS1 (green) in complex with β-amyrin (orange sticks). Residues of AsbAS1 are labelled. The increased steric bulk of leucine residues in the dicot enzymes in place of Val482 in AsbAS1 would push the A ring (orange label) into closer contact with the π-cloud of Trp611, potentially enhancing the stabilisation of carbocations on the A ring. Figure prepared using PyMOL[46] and labelled with Photoshop CS2.

5.4 Conclusions and future work

The AsbAS1 enzyme was originally overexpressed in *E. coli* BL21 using the pET-14b vector, but to attempt obtaining it His-tagged and folded, expression was attempted with the pET-22b vector, which directs the recombinant protein to the periplasm. Purification of an impurity protein of similar MW led to the crystallisation and structure solution of the *E. coli* HPII catalase (see Appendix 2). This was followed by structure-informed engineering attempts to render the membrane-bound AsbAS1 soluble by mutagenesis of residues in contact with the membrane based on homology models. A library of mutants was expressed in *E. coli* using the pET-14b vector, but western blot analysis of the resulting soluble fractions did not reveal any solubilised AsbAS1. Attention then turned to expression trials of AsbAS1 constructs using the methylotrophic yeast, *P. pastoris*. Of those tested, a C-terminal deletion mutant was shown by GC-MS to be in an active form. The homologous β-amyrin synthase, EtAS, was then expressed in an active form in *E. coli* and it may be possible to purify it. Finally, the release of AlphaFold2 allowed the prediction of the 3-dimensional structures of several plant OSCs. These models, along with a multiple sequence alignment, were interrogated to derive predictions of active site residue

substitutions which should lead to rational changes in product specificity. For example, the Phe259His mutation could change the product of AsbAS1 from β-amyrin to lanosterol, while an additional mutation (Val482Ile) could yield cycloartenol instead. Trp257Leu is likely to lead to the formation of lupeol, while the Val482Leu mutant may produce glutinol and/or friedelin.

In the future, these AsbAS1 activity mutants could be generated and expressed in *P. pastoris*, and their products analysed by GC-MS based on the method presented in this thesis. For higher-throughput screening, AsbAS1 mutants could be expressed and assayed *in vivo* in *E. coli* harbouring the plasmid pAC-HpIDI/AtSQS/AtSQE that leads to high level of the OS precursor. Once identified, mutants that produce new triterpene scaffolds could be expressed along other enzymes in *N. benthamiana* to yield large quantities of triterpenoids that would be difficult to obtain otherwise.[67] This would unlock their potential applications in food, health and industry (see section 1.2.1).

The library of AsbAS1 solubilisation mutant could also be expressed in *P. pastoris* in the hope of producing a water-soluble plant OSC, which would facilitate crystallisation and have potential applications in biocatalysis.

Finally, heterologous expression of EtAS on a small scale was promising, so large-scale expression and purification with detergents may allow crystallisation screening and the potential solution of the first plant OSC crystal structure.

Chapter 6: Conclusion

This thesis describes the structural characterisation of three enzymes involved in the biosynthesis of avenacin, an antimicrobial glycosylated triterpene that makes oat resistant to the fungal disease take-all. This aimed to enable the rational protein engineering that would generate products that are otherwise intractable.

Chapter 3 reports the study of AsAAT1, a UDP-dependent arabinosyltransferase that catalyses the first glycosylation step in the biosynthesis of avenacin. After heterologous expression in *E. coli*, the enzyme was produced both with and without a His-tag, at levels of purity and in quantities high enough for crystallisation screening. Despite extensive efforts, no diffracting crystals could be made. Structural information was therefore obtained through the generation of a molecular model of AsAAT1 in complex with its ligand UDP-Ara using threading and energy minimisation. This gave insights into the structural determinants of the enzyme's specificity for four different sugar donors, depending on which of two mutations are introduced. Further crystallisation or modelling strategies could be attempted to obtain a more accurate and informative structure of this unusual active site.

Chapter 4 covers efforts to understand the structure-function relationship of AsTG1, a glycosyl-hydrolase-fold transglucosidase that catalyses the last glycosylation step in the biosynthesis of avenacin. Recombinant AsTG1 with a N-terminal 9xHis-tag was generated in large quantities and high purity from heterologous expression in *E. coli* followed by a 2-step purification. This allowed crystallisation screening, but no diffracting crystals could be obtained. Three deletion constructs of AsTG1 were designed to improve crystallisability, but suffered from insolubility. A protease cleavage site was introduced into the full-length construct to attempt producing tag-free enzyme, but the removal of the 9xHis-tag proved unsuccessful. To better understand the determinants of transglycosidase activity in the absence of a crystal structure, a homology model of AsTG1 was generated. A multiple sequence alignment of GH1 enzymes revealed that His380 and Leu244 could be responsible for the switch from glycosyl hydrolase to transglycosidase activity. While MD simulations were inconclusive towards this hypothesis, they did reveal a binding site for the substrate mono-deglucosyl avenacin A1, but not for the analogue bis-deglucosyl avenacin A1, which may explain why it does not undergo significant transglycosylation *in vitro*. This

offers an *in silico* method to screen various sugar acceptors for their ability to be glycosylated by AsTG1.

Chapter 5 describes structure-based engineering efforts for AsbAS1, a membrane-bound oxidosqualene cyclase which catalyses the first committed step in avenacin biosynthesis. It was first expressed in *E. coli* BL21, but to attempt purifying folded protein, a His-tagged construct targeted to the periplasm was expressed in SoluBL21. This led to the purification of an impurity protein, *E. coli* HPII catalase, which was crystallised and its high-resolution structure solved. To simplify purification and crystallisation of AsbAS1, potentially soluble mutants were designed based on its homology model and that of a soluble homologue. The resulting library was expressed in *E. coli*, but western blot analysis did not reveal any AsbAS1 in the soluble fractions. The mutants could be expressed in a different system, the yeast *Pichia pastoris*. Indeed, when this was used for expression trials of AsbAS1, it resulted in active His-tagged protein from one of the constructs, as shown by GC-MS analysis. Another β-amyrin synthase, EtAS, was also expressed in an active form, this time in *E. coli*. These two expression methods may therefore be used to produce two different β-amyrin synthases for purification and crystallisation screening, with the view to obtaining the first crystal structure of a plant OSC. Before this could be achieved, preliminary models of plant OSCs were generated with AlphaFold2, and along with a multiple sequence alignment, they enabled the design of four mutants of AsbAS1 that may have altered product specificity: Phe259His for lanosterol, coupled with Val482Ile for cycloartenol; Trp257Leu for lupeol; and Val482Leu for glutinol and/or friedelin. These mutants could be assayed in *P. pastoris* or *E. coli*.

Overall, the work presented in this thesis has generated structural information for three enzymes in the avenacin biosynthesis pathway, which led to the rationalisation or prediction of effects from various amino acids, which can now be tested by expressing mutants using the methods presented herein.

Appendix 2: HPII Catalase

A2.1 Introduction

Catalase is a hydrogen peroxide:hydrogen peroxide oxidoreductase found in almost all aerobic organisms due to its role in protecting against oxidative damage. *E. coli* produces two of these hydroperoxidases, HPI and HPII, both of which are able to catalyse the dismutation of hydrogen peroxide to oxygen and water:[276]

$$2H_2O_2 \rightarrow O_2 + 2H_2O$$[277]

HPII catalase exists as a homotetramer of 84 kDa subunits, each with a haem d in a *cis*-spirolactone form. Interestingly, the haem d is formed from haem b through catalysis by HPII using hydrogen peroxide (Fig. A2.1). This is probably concomitant with the formation of the unique covalent linkage between the Nδ of His392 and the Cβ of the Tyr415 residue that coordinates the haem iron on the proximal side.[278] Other catalases usually have their haem "flipped" 180°, exchanging the positions of rings II and III with that of rings I and IV, respectively.[279]

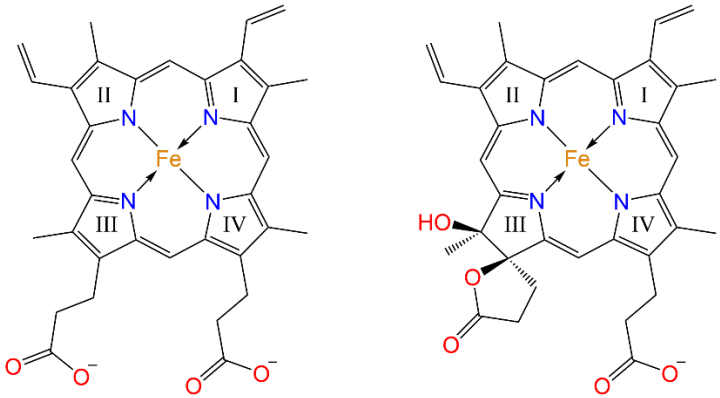


Figure A2.1: Structure of the haem that binds and that forms in the active site of HPII catalase. Haem b binds HPII catalase and is found in catalytically inactive mutants thereof. Haem d formation is catalysed by the enzyme. Figure adapted from Loewen[276] using ChemDraw 21.[37]

A2.2 Methods

Crystals were sent in a dry shipper to Diamond Light Source, beamline i04. Images from a single crystal were collected on a Eiger2 XE 16M (Dectris) detector with 3600 frames

of 0.1° rotation and an exposure time of 0.01 s at an X-ray wavelength of 0.9795 Å. Integration and scaling was performed with DIALS[163] to a maximum resolution of 1.70 Å, before merging the data with AIMLESS[268] to a maximum resolution of 1.79 Å. Molecular replacement with PHASER[269] used chain A of the previously solved 1.64 Å resolution structure of natively expressed HPII catalase (PDB: 4BFL, unpublished) as a search model. This led to a solution that was then subject to several rounds of automatic refinement with REFMAC5[170] and phenix.refine[171] interspersed with manual model adjustment using COOT.[169]

A2.3 Results and discussion

HPII catalase purified from *E. coli* SoluBL21™ was crystallised and its structure solved to 1.79 Å by molecular replacement using the structure of the natively expressed HPII catalase from *E. coli* BL21 (PDB: 4BFL, unpublished). The structure was refined to an R_{free} of 15%. This represented an improvement over the search model which has a reported R_{free} of 20% (Table A2.1).

Table A2.1: Data collection and refinement statistics for the structure of HPII catalase from *E. coli* SoluBL21. Statistics for the highest-resolution shell are shown in parentheses. Generated using Phenix.[171]

Wavelength	0.9795
Resolution range	56.33 - 1.79 (1.854 - 1.79)
Space group	P 1 21 1
Unit cell	73.791 171.643 123.165 90 104.471 90
Total reflections	1780540 (92637)
Unique reflections	266804 (19906)
Multiplicity	6.7 (4.7)
Completeness (%)	95.94 (71.82)
Mean I/sigma(I)	32.22 (6.19)
Wilson B-factor	15.34
R-merge	0.05136 (0.1892)
R-meas	0.0556 (0.2117)
R-pim	0.02109 (0.09232)
CC1/2	0.999 (0.977)
CC*	1 (0.994)
Reflections used in refinement	266671 (19899)
Reflections used for R-free	13240 (1030)
R-work	0.1206 (0.1721)
R-free	0.1540 (0.1675)
CC(work)	0.975 (0.949)
CC(free)	0.974 (0.948)
Number of non-hydrogen atoms	27585
macromolecules	23899
ligands	208
solvent	3478
Protein residues	2984
RMS(bonds)	0.015
RMS(angles)	1.76
Ramachandran favored (%)	97.35
Ramachandran allowed (%)	2.59
Ramachandran outliers (%)	0.07
Rotamer outliers (%)	0.86
Clashscore	2.42
Average B-factor	20.16
macromolecules	18.75
ligands	15.09
solvent	30.17

The asymmetric unit consists of the homotetramer, with the 4 subunits interlinked and each bound to a *cis*-spirolactone haem d (Fig. A2.2), as opposed to ring-opened haem d in the structure from which the search model was derived.

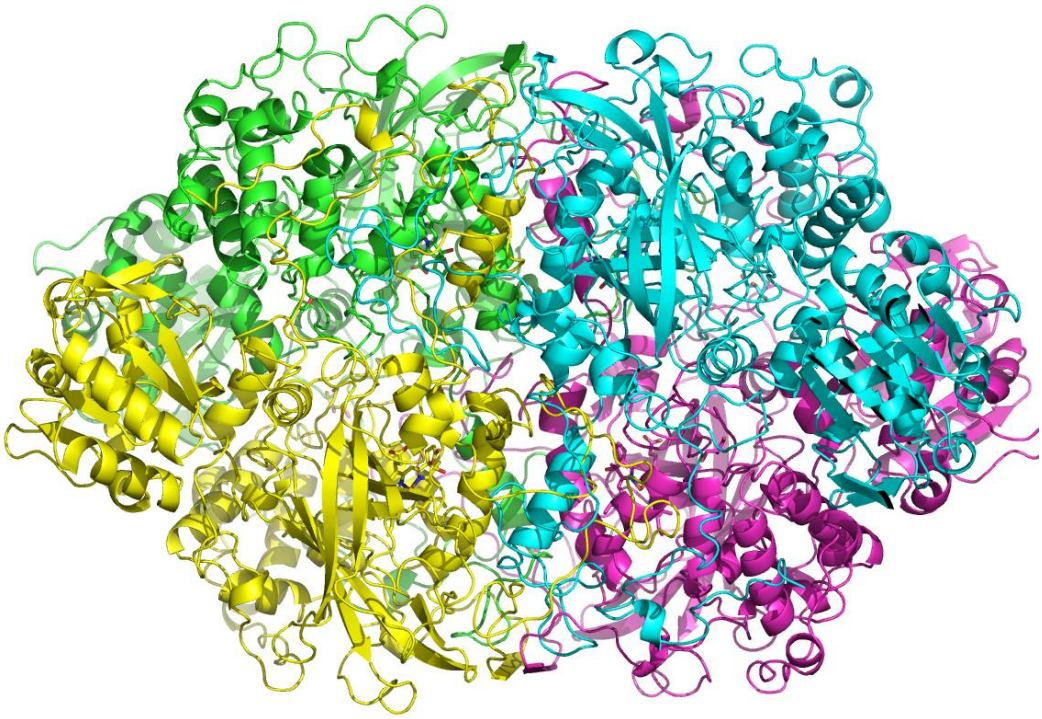


Figure A2.2: Asymmetric unit of the crystal structure HPII catalase reported herein. Each polypeptide chain is coloured differently and shown as a cartoon representation. The ligands are shown as sticks. Figure prepared using PyMOL.[46]

Compared to the search model, another difference in the structure from SoluBL21 is the Ile710Val mutation that is observed in the electron density (Fig. A2.3), likely acquired during the directed evolution process that the strain underwent.

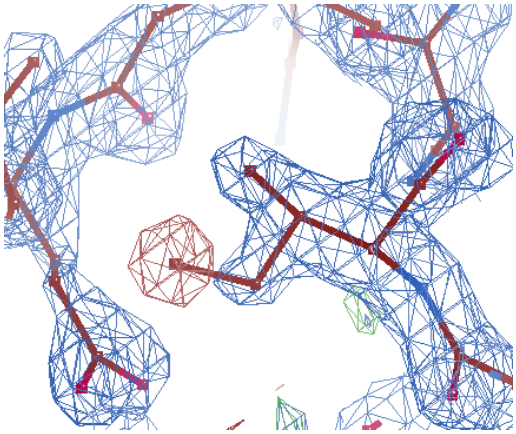


Figure A2.3: Electron density at position 710 after molecular replacement. The blue mesh represents the double-difference Fourier at 1.5 rmsd and the red mesh the negative single-difference Fourier map. Figure prepared using WinCoot 0.9.6.[169]

Furthermore, haem is seen in its spirolactone form and the bond between His392 and Tyr415 is observed to be present with a refined bond length of 1.5 Å (Fig. A2.4).

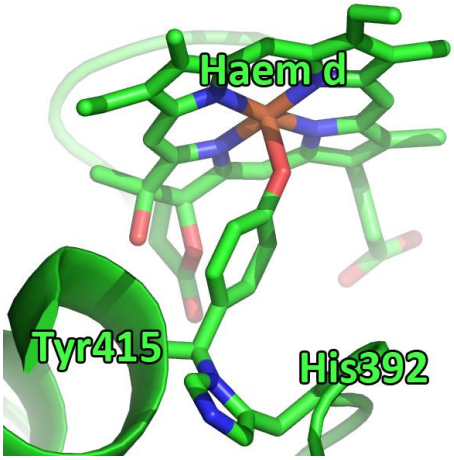


Figure A2.4: Proximal side of the haem d of HPII catalase. In the structure solved herein, the *cis*-spirolactone haem d and the covalently bound side chains of Tyr415 and His392 are shown as green sticks within the active site of HPII catalase (cartoon representation). Figure prepared using PyMOL[46] and labelled with Photoshop CS2.

A notable feature of the structure presented here is the electron density on the distal side of the haem. Indeed, while the search model has a vacant axial coordination site, the refined structure has electron density that has been interpreted as an oxygen atom bonded to the haem iron. More unusual is a larger volume of electron density adjacent to the axial ligand site, which may correspond to an O_2 molecule, the product of HPII catalase activity (Fig. A2.5A).[279] The alternative interpretation of this density as a single water molecule results in significant residual peaks in the mFo-DFc electron density map (Fig.

A2.5B). Molecular oxygen can sometimes be seen in crystal structures of catalases, such as that of mutant catalase-peroxidase from *Burkholderia pseudomallei* (PDB: 5KQK), but to my knowledge has never been reported in a crystal structure of HPII catalase.

(A) **(B)**

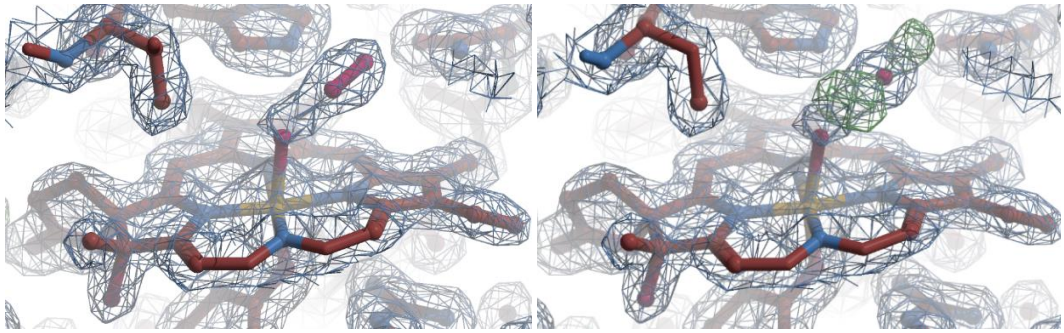


Figure A2.5: Interpretations of the electron density in the active site of HPII catalase. The model is shown as red sticks. The 2mFo-DFc Fourier map electron density map contoured at 1.5σ is shown as a blue mesh and the positive peaks of the mFo-DFc Fourier electron density map contoured at 3.0σ are shown as a green mesh. **(A)** Final structure with molecular oxygen in the active site. **(B)** Alternative interpretation with a water molecule. Figure prepared using WinCoot 0.9.6.[169]

In conclusion, herein is reported the high-resolution crystal structure of HPII catalase from *E. coli* SoluBL21, which reveals an amino acid substitution compared to the BL21 strain and may be the first to show all the products of the reactions catalysed by the enzyme: the covalently bound side chains of His392 and Tyr415, the *cis*-spirolactone form of haem d, and an oxygen molecule in the active site.

Appendix 3: Norwich Science Festival Activity

A3.1 Introduction

A3.1.1 The Norwich Science Festival

Every October, the Norwich Science Festival was held at The Forum to showcase the science carried out in the city through exhibitions, shows and hands-on activities. It was organised with the help of many organisations, such as the ones on the Norwich Research Park, and ran by volunteers, including scientists which the public got a chance to meet. Because it took place during the half-term holidays, many families with kids attended the festival. The content therefore needed to be accessible to people from all ages. An activity was designed to explain the topic of this thesis and ran at all three editions of the Norwich Science Festival that happened during the period of the research project.

A3.1.2 Format of the activity

1. By telling the story of a farmer, the take-all disease and oat's resistance to it are explained to the participants.
2. The participants examine oat and wheat seedlings under UV light to find that the roots of oat fluoresce because of avenacin.
3. Toy bricks are used to explain how an enzyme can build the avenacin molecule
4. A puzzle prompts the participants to engineer the enzyme by changing its shape so that it can build a different molecule.
5. The participants draw a schematic of their engineered enzyme and produced molecule then take a lanyard card home.

A3.1.3 Learning outcomes

The learning outcomes for the participants were to know more about the research covered in this thesis and why it is related to things they know (oat and farming). They saw that an experiment could give clues about the mechanism of resistance to disease. They got an understanding of what an enzyme is, how it can make molecules and why engineering it can lead to new molecules being made.

A3.2 Materials and methods

A3.2.1 List of materials

The following materials were used:

- UV light box (365 nm wavelength);
- Oat and wheat seedlings (see section A.3.2.3 for methods);
- Oat groats and wheat grain in two labelled sacks;
- A3 print-out of “My research” (Fig. A3.1);
- A4 print-out of *G. graminis* (Fig. A3.2);
- Toy brick enzyme and avenacin (Fig. A3.3A);
- Toy brick modified avenacin (Fig. A3.3B);
- Tape;
- Box of spare toy bricks;
- Square-grid paper (ideally 1 cm squares);
- Blue, red and green markers;
- Lanyard cards (Fig. A3.4).

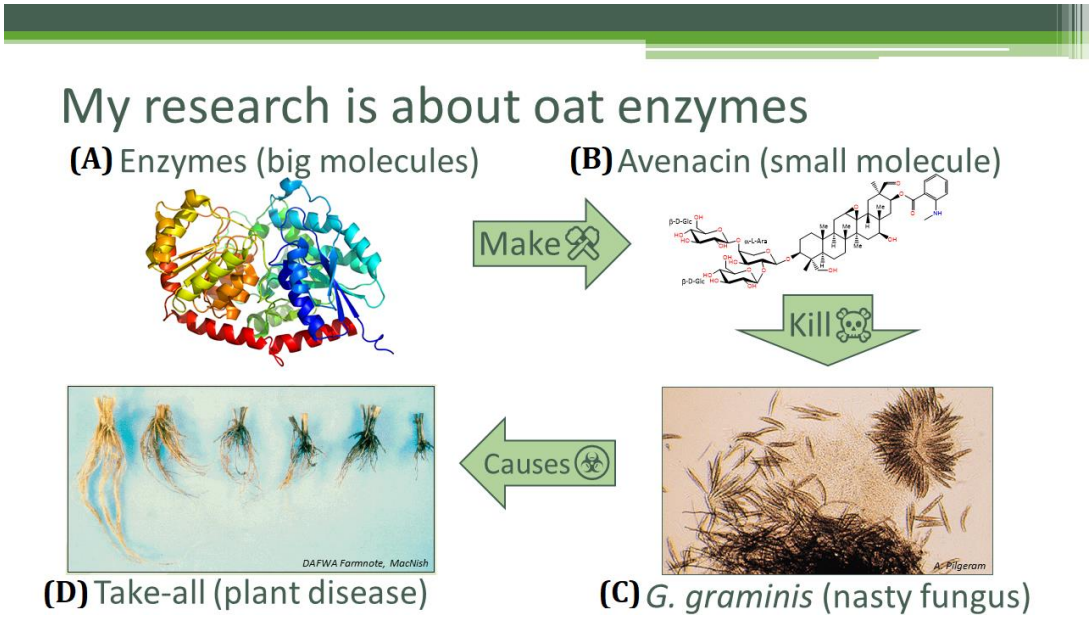
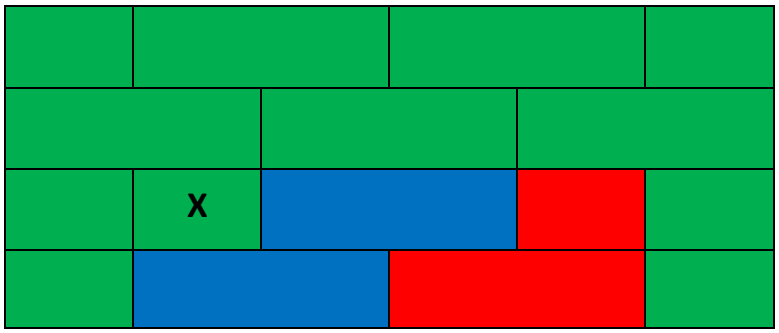


Figure A3.1: Schematic explaining the four key aspects of the project context and their relationship. **(A)** Enzymes make **(B)** avenacin which kills **(C)** *Gaeumannomyces graminis* var. *tritici*, the causative agent of **(D)** the take-all disease. The slide was printed in size A3 to show the effect of take-all on wheat roots. It was also used as an illustration of *G. graminis*, to explain the difference between a macromolecule and a small molecule, or to show features of the avenacin molecule.

Figure A3.2: Petri dish of *Gaeumannomyces gramins* var. *tritici* growing on agar. The photo was printed in size A4 to illustrate that the fungus is like a mould. Credit: Sarah Worsley, School of Biological Sciences, University of East Anglia, UK.

(A)



(B)

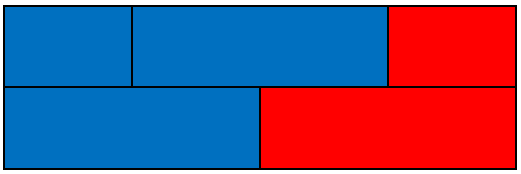
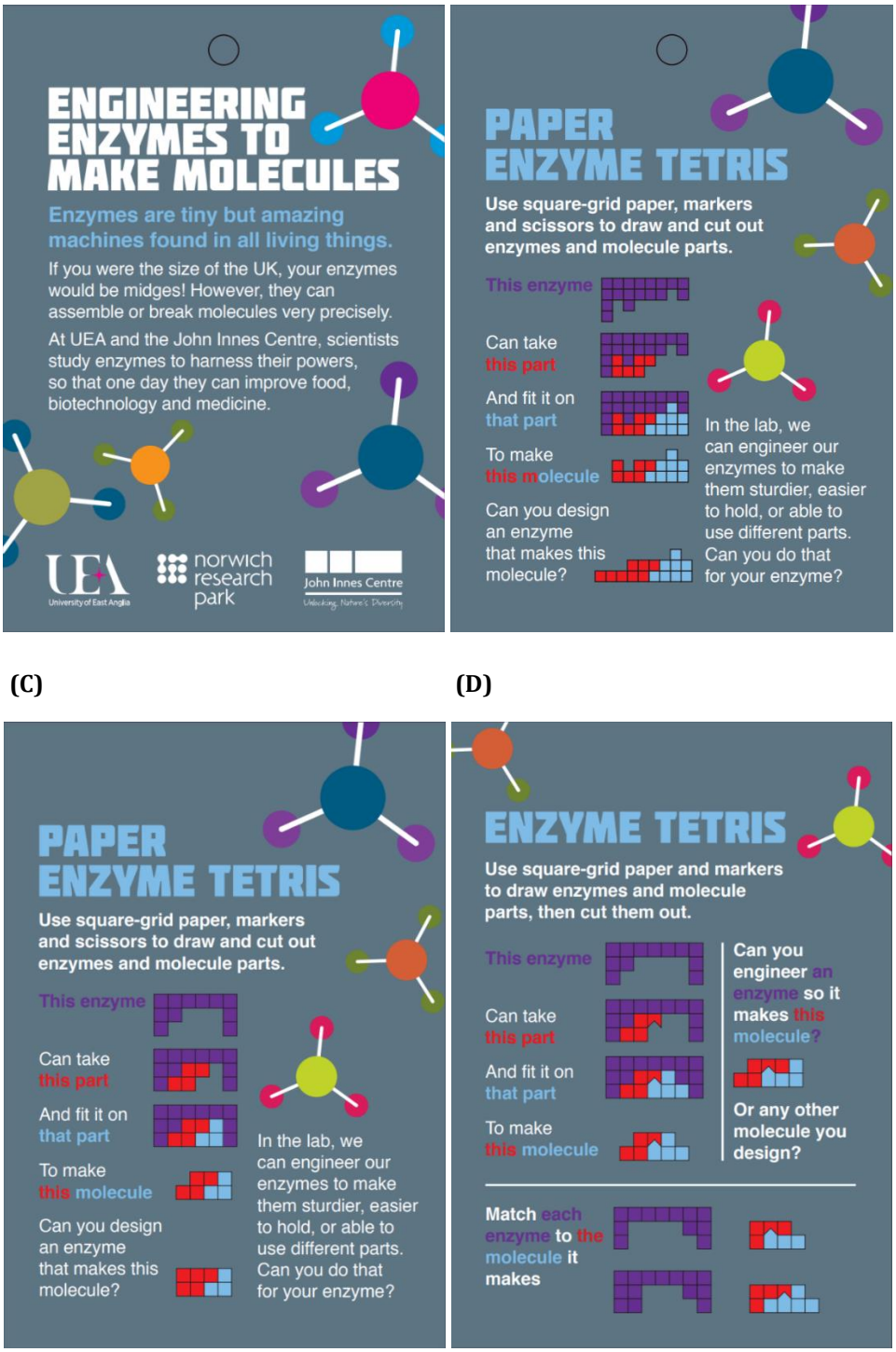


Figure A3.3: Blueprints of the toy-brick props. (A) Enzyme (green, above), and the first (blue, left) and second part (red, right) of the avenacin molecule. The block that needed to be removed to engineer the enzyme is marked with a cross. **(B)** Modified avenacin molecule. Each toy brick enzyme and molecule part was taped on the side facing away from the participants so that it stayed together during the demonstration of catalysis.

A3.2.2 Lanyard card design

A short text describing the research area was written for the front of the lanyard cards given out to participants. A short activity to do at home was designed and updated every year so that the returning participants could collect a new version. These ideas were then sent by the UEA Events team to a designer give them a professional look before printing 600 copies.

(A) **(B)**



(C) **(D)**

Figure A3.4: Design of the lanyard cards handed to participants. (A) Front explaining the research and back giving an activity to do at home for the **(B)** 2018, **(C)** 2019 and **(D)** 2021 edition.

A3.2.3 Methods for seed germination

Seedlings were provided by Rachel Melton (John Innes Centre). To produce them, oat and wheat seeds were first dehusked to improve germination frequency, then sterilised by washing in 0.5% sodium hypochlorite (prepared by 20-fold dilution of a 10% commercial stock (Sigma)). The sterilised seeds were washed three times in dH_2O then placed on 0.8% (w/v) sterile agar in square petri dishes. These were sealed with parafilm then incubated at 4 °C for at least 2 d. Incubation at 22 °C for 2-3 d is usually sufficient for the roots to grow enough for fluorescence to be clearly visible.

A3.2.4 Script for the activity

[*Introduce yourself e.g.,*] Hi, I'm Hans, a scientist at UEA. To explain the context of my research, let me tell you a story.

A farmer sowed oat [*show sack*] and wheat [*show other sack*], not knowing that in their soil was lurking the nasty fungus called *G. graminis*. [*show Fig. A3.2 saying it's like a mould or Fig. A3.1C if they are likely to understand fungus and microscopes*]

When they went to harvest, the wheat looked terrible and its roots were black [*show Fig. A3.1D*]. But the oat was doing great. Why? Let's investigate!

Here are some wheat seedlings and some oat seedlings. Put each petri dish under UV light. Can you tell the difference?

Why do oat roots glow purple? Because of the fluorescence of oat's secret weapon: avenacin [*show toy-brick avenacin*]

This molecule is able to pierce holes into the "take-all" fungus, so it spews its insides out and die.

How does it make that weapon? With enzymes! Do you know what an enzyme is? Enzymes are amazing! They are minuscule machines found in all living things, including you! They make or break molecules, like avenacin.

[*See Fig. A3.5 for an overview of the following steps*] In oat roots, an enzyme [*green bricks*] can fit the first part of avenacin [*blue bricks*] and the second part [*red bricks*] then

bam! [*slam onto the enzyme so the loosely fitted molecule parts attach properly*] It attaches them together! And that makes avenacin.[*pull out the avenacin molecule*]

My job is to figure out what this enzyme looks like, which is tricky because it's too small to see, even with a microscope. Once we know its structure, we can engineer the enzyme! For example, so that it makes this other molecule,[*show toy-brick modified avenacin (Fig. A3.3B)*] which could be very important for making better vaccines or a better chocolate mousse.

How would you change the shape of this enzyme so that it makes this new molecule?[*see Fig. A3.3B*]

Let's see if your enzyme works: it should be able to pick up this part [*blue part in Fig A3.3B*] because it fits it very well, and then pick up this other part [*red part in FigA3.3B*]. It attaches the parts together to make the new molecule! Congratulations, you've successfully engineered the enzyme! Was that interesting? Now you can draw the blueprint of your engineered enzyme and the molecule that it can build.

[*give lanyard card, see Fig A3.4*]

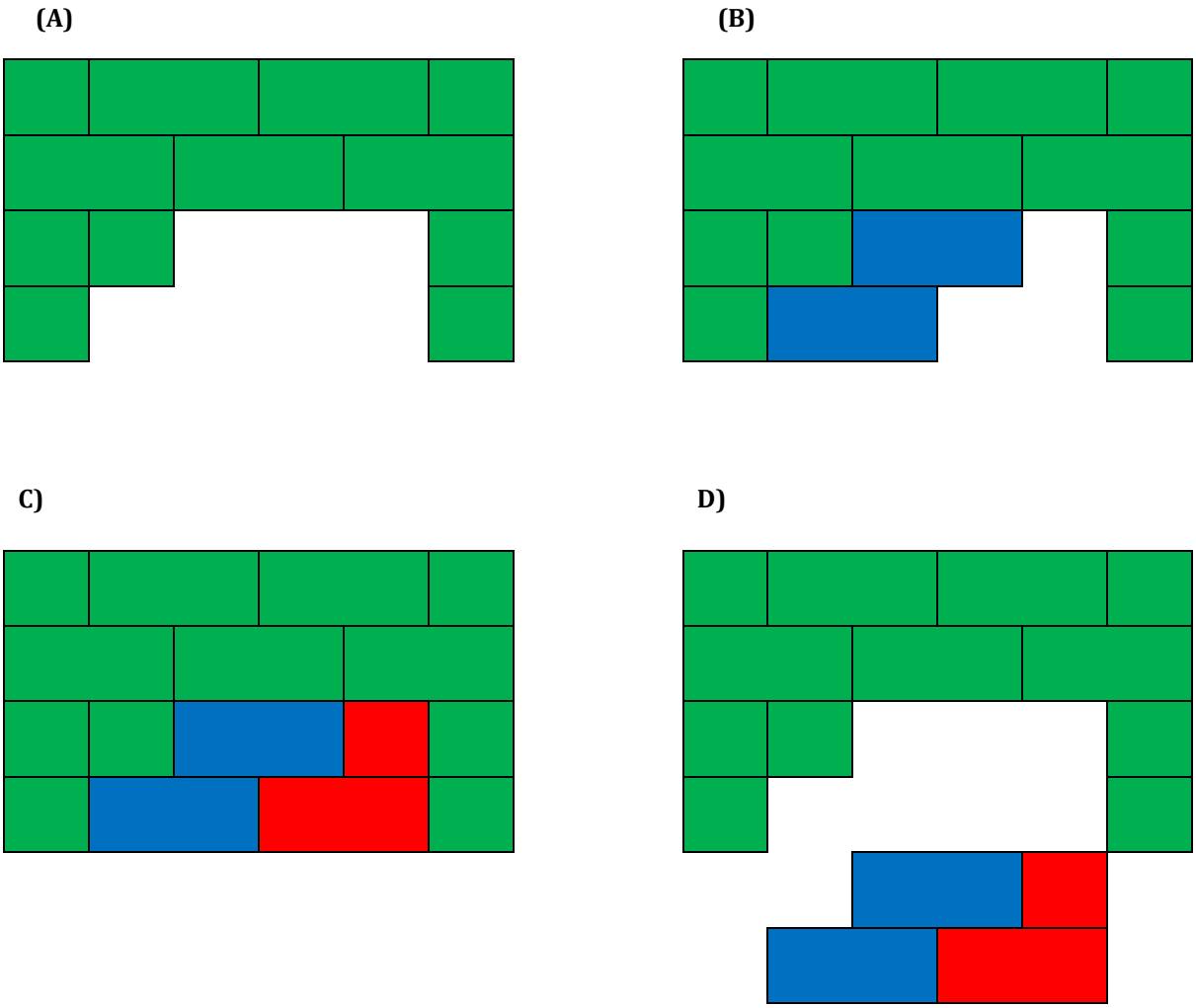


Figure A3.5: The steps of enzyme catalysis explained with toy bricks. (A) The enzyme is unbound, then **(B)** picks up the first part of the avenacin molecule before **(C)** picking up the second part to assemble avenacin. **(D)** Finally, the enzyme releases the avenacin molecule (this is most easily done by pulling the side with the tape first so that no other brick gets detached).

A3.2.5 Evaluation

Evaluation is an essential part of science communication, as it helps improve the activity and demonstrate impact.[280] In the activity presented here, the participants were asked at the end if they thought it was interesting to qualitatively gauge if they were satisfied with it. Because each participant was given a lanyard card, the leftover ones were counted and subtracted from the 600 originally printed to give an estimate of the number of people reached. This was calculated to be between 250 and 450 people depending on the year. Finally, to evaluate understanding, the participants were invited to draw the toy-brick engineered enzyme and molecule parts. Two examples of outstanding understanding are shown in Fig. A3.6.

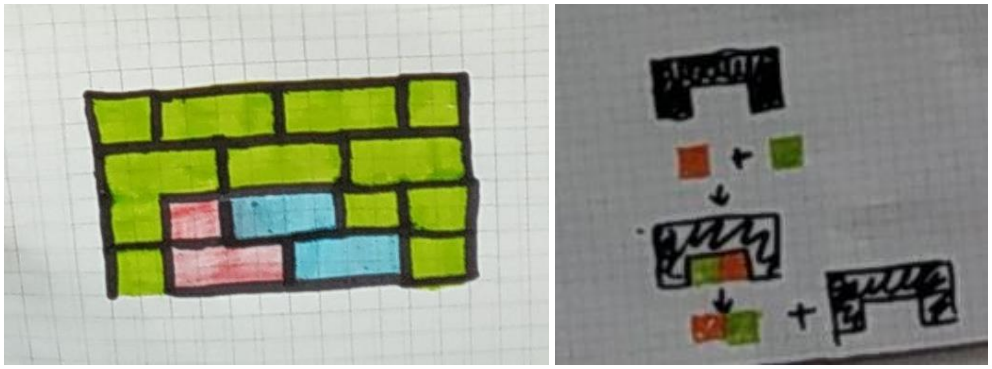


Figure A3.6: Evaluation drawings made by two young participants. They to represent the enzyme they engineered and the molecules it makes.

References

1. Fowler MW (2006) “Plants, medicines and man.” *Journal of the Science of Food and Agriculture*, 86(12):1797–1804. https://doi.org/10.1002/jsfa.2598

2. Lindsay RC (2017) “Food Additives.” *Fennema’s food chemistry*, p801–863. https://www.taylorfrancis.com/chapters/edit/10.1201/9781315372914-14/food-additives-robert-lindsay

3. Van der Schaft P (2015) “Approaches to production of natural flavours.” *Flavour Development, Analysis and Perception in Food and Beverages*, p235–248. https://doi.org/10.1016/B978-1-78242-103-0.00011-4

4. Winterton N (2021) “Chemistry for sustainable technologies: a foundation.” https://books.google.be/books?id=_fsdEAAAQBAJ

5. McChesney JD, Venkataraman SK, Henri JT (2007) “Plant natural products: Back to the future or into extinction?” *Phytochemistry*, 68(14):2015–2022. https://doi.org/10.1016/j.phytochem.2007.04.032

6. Osbourn AE, Lanzotti V (2009) “Preface.” *Plant-derived natural products: synthesis, function, and application*, pV. https://link.springer.com/content/pdf/bfm:978-0-387-85498-4/1

7. Birchfield AS, McIntosh CA (2020) “Metabolic engineering and synthetic biology of plant natural products – A minireview.” *Current Plant Biology*, 24:100163. https://doi.org/10.1016/j.cpb.2020.100163

8. Schrader J, Etschmann MMW, Sell D, Hilmer J-M, Rabenhorst J (2004) “Applied biocatalysis for the synthesis of natural flavour compounds – current industrial processes and future prospects.” *Biotechnology Letters*, 26(6):463–472. https://doi.org/10.1023/B:BILE.0000019576.80594.0e

9. Li Y, Wang J, Li L, Song W, Li M, Hua X, Wang Y, Yuan J, Xue Z (2023) “Natural products of pentacyclic triterpenoids: from discovery to heterologous biosynthesis.” *Natural Product Reports*, Advance Article. https://doi.org/10.1039/D2NP00063F

10. Ingrouille M, Eddie B (2006) “Plants: Evolution and diversity.” *Plants: Evolution and Diversity*, https://doi.org/10.1017/CBO9780511812972

11. (2012) “Cereal production from World Crops Database.” https://world-crops.com/cereal-production/

12. Colhoun J (1971) “Cereals.” *Diseases of Crop Plants*, p181–185.

13. Keller KO, Engel B, Heinrich K (1995) “Specific detection of Gaeumannomyces *graminis* in soil using polymerase chain reaction.” *Mycological Research*, 99(11):1385–1390. https://doi.org/10.1016/S0953-7562(09)81226-4

14. Palma-Guerrero J, Chancellor T, Spong J, Canning G, Hammond J, McMillan VE, Hammond-Kosack KE (2021) “Take-All Disease: New Insights into an Important Wheat Root

Pathogen." *Trends in Plant Science*, 26(8):836–848. https://doi.org/10.1016/j.tplants.2021.02.009

15. Bateman G, Gutteridge R, John Jenkyn (2006) "Take-all in winter wheat- management guidelines." http://adlib.everysite.co.uk/resources/000/202/990/HGCA_take_all_guidelines_autumn_06.pdf

16. Papadopoulou K, Melton RE, Leggett M, Daniels MJ, Osbourn AE (1999) "Compromised disease resistance in saponin-deficient plants." *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12923–12928. https://doi.org/10.1073/pnas.96.22.12923

17. Armah CN, Mackie AR, Roy C, Price K, Osbourn AE, Bowyer P, Ladha S (1999) "The membrane-permeabilizing effect of avenacin A-1 involves the reorganization of bilayer cholesterol." *Biophysical Journal*, 76(1):281–290. https://doi.org/10.1016/S0006-3495(99)77196-1

18. Thimmappa R, Geisler K, Louveau T, O'Maille P, Osbourn A (2014) "Triterpene Biosynthesis in Plants." *Annual Review of Plant Biology*, 65:225–257. https://doi.org/10.1146/annurev-arplant-050312-120229

19. Van Dyck S, Gerbaux P, Flammang P (2010) "Qualitative and quantitative saponin contents in five sea cucumbers from the Indian ocean." *Marine Drugs*, 8(1):173–189. https://doi.org/10.3390/md8010173

20. Geisler K, Hughes RK, Sainsbury F, Lomonossoff GP, Rejzek M, Fairhurst S, Olsen C-E, Motawia MS, Melton RE, Hemmings AM, Bak S, Osbourn A (2013) "Biochemical analysis of a multifunctional cytochrome P450 (CYP51) enzyme required for synthesis of antimicrobial triterpenes in plants." *Proceedings of the National Academy of Sciences*, 110(35):E3360–E3367. https://doi.org/10.1073/pnas.1309157110

21. Nielsen JK, Nagao T, Okabe H, Shinoda T (2010) "Resistance in the plant, Barbarea vulgaris, and counter-adaptations in flea beetles mediated by saponins." *Journal of Chemical Ecology*, 36:277–285. https://doi.org/10.1007/s10886-010-9758-6

22. Moses T, Papadopoulou KK, Osbourn A (2014) "Metabolic and functional diversity of saponins, biosynthetic intermediates and semi-synthetic derivatives." *Critical Reviews in Biochemistry and Molecular Biology*, 49(6):439–462. https://doi.org/10.3109/10409238.2014.953628

23. Miyanishi H, Kimura A, Kasahara S, Ishikawa M (2004) "Upgrading of the Anaerobic Digestion by Addition of Bio Active Reagent (Saponin)." *Japanese Journal of Water Treatment Biology*, 40(2):37–44. https://doi.org/10.2521/jswtb.40.37

24. Schulz-Gasch T, Stahl M (2003) "Mechanistic insights into oxidosqualene cyclizations through homology modeling." *Journal of Computational Chemistry*, 24(6):741–753. https://doi.org/10.1002/jcc.10147

25. Augustin JM, Kuzina V, Andersen SB, Bak S (2011) "Molecular activities, biosynthesis and evolution of triterpenoid saponins." *Phytochemistry*, 72(6):435–457. https://doi.org/10.1016/j.phytochem.2011.01.015

26. Zhu B, Zhang W, Lu Y, Hu S, Gao R, Sun Z, Chen X, Ma J, Guo S, Du S, Li P (2018) "Network pharmacology-based identification of protective mechanism of Panax Notoginseng Saponins on aspirin induced gastrointestinal injury." *Biomedicine and Pharmacotherapy*, 105:159–166. https://doi.org/10.1016/j.biopha.2018.04.054

27. Navarro del Hierro J, Herrera T, García-Risco MR, Fornari T, Reglero G, Martin D (2018) "Ultrasound-assisted extraction and bioaccessibility of saponins from edible seeds: quinoa, lentil, fenugreek, soybean and lupin." *Food Research International*, 109:440–447. https://doi.org/10.1016/j.foodres.2018.04.058

28. Xiang L, Zhang L, Chen X, Xia X, Li R, Zhong J (2018) "Ursane-type triterpenoid saponins from *Elsholtzia bodinieri*." *Natural Product Research*, 33(9):1349–1356. https://doi.org/10.1080/14786419.2018.1477144

29. Francis G, Makkar HPS, Becker K (2001) "Antinutritional factors present in plant-derived alternate fish feed ingredients and their effects in fish." *Aquaculture*, 199(3–4):197–227. https://doi.org/10.1016/S0044-8486(01)00526-9

30. Wang M, Wu B, Shah SN, Lu P, Lu Q (2018) "Saponins as Natural Adjuvant for Antisense Morpholino Oligonucleotides Delivery In Vitro and in mdx Mice." *Molecular Therapy - Nucleic Acids*, 11:192–202. https://doi.org/10.1016/j.omtn.2018.02.004

31. Guclu-Ustundag Ö, Mazza G (2007) "Saponins: Properties, applications and processing." *Critical Reviews in Food Science and Nutrition*, 47(3):258. https://doi.org/10.1080/10408390600698197

32. Gurfinkel DM, Rao AV (2003) "Soyasaponins: The Relationship between Chemical Structure and Colon Anticarcinogenic Activity." *Nutrition and Cancer*, 47(1):24–33. https://doi.org/10.1207/s15327914nc4701_3

33. Yang X-W, Dai Z, Wang B, Liu Y-P, Zhao X-D, Luo X-D (2018) "Antitumor Triterpenoid Saponin from the Fruits of Avicennia *marina*." *Natural Products and Bioprospecting*, 8:347–353. https://doi.org/10.1007/s13659-018-0167-9

34. Yang L, Liu X, Zhuang X, Feng X, Zhong L, Ma T (2018) "Antifungal Effects of Saponin Extract from Rhizomes of Dioscorea panthaica Prain et Burk against Candida albicans." *Evidence-based Complementary and Alternative Medicine*, 2018:ID 6095307. https://doi.org/10.1155/2018/6095307

35. Orme A (2017) "Investigation of oat glucosyltransferases essential for biosynthesis of the antifungal triterpene glycoside avenacin." *University of East Anglia*, Ph.D thesis. https://ueaeprints.uea.ac.uk/id/eprint/67091/1/Orme-Thesis-2017.pdf

36. Wu Y, Zou HD, Cheng H, Zhao CY, Sun LF, Su SZ, Li SP, Yuan YP (2012) "Cloning and characterization of a β-amyrin synthase gene from the medicinal tree *Aralia elata* (Araliaceae)." *Genetics and molecular research : GMR*, 11(3):2301–2314. https://doi.org/10.4238/2012.August.13.4

37. PerkinElmer "ChemDraw."

38. Qi X, Bakht S, Leggett M, Maxwell C, Melton R, Osbourn A (2004) "A gene cluster for secondary metabolism in oat: Implications for the evolution of metabolic diversity in

plants." *Proceedings of the National Academy of Sciences*, 101(21):8233–8238. https://doi.org/10.1073/pnas.0401301101

39. Louveau T, Orme A, Pfalzgraf H, Stephenson MJ, Melton R, Saalbach G, Hemmings AM, Leveau A, Rejzek M, Vickerstaff RJ, Langdon T, Field RA, Osbourn A (2018) "Analysis of two new arabinosyltransferases belonging to the carbohydrate-active enzyme (CAZY) glycosyl transferase family1 provides insights into disease resistance and sugar donor specificity." *Plant Cell*, 30(12):3038–3057. https://doi.org/10.1105/tpc.18.00641

40. Orme A, Louveau T, Stephenson MJ, Appelhagen I, Melton R, Cheema J, Li Y, Zhao Q, Zhang L, Fan D, Tian Q, Vickerstaff RJ, Langdon T, Han B, Osbourn A (2019) "A noncanonical vacuolar sugar transferase required for biosynthesis of antimicrobial defense compounds in oat." *Proceedings of the National Academy of Sciences of the United States of America*, 116(52):27105-27114. https://doi.org/10.1073/pnas.1914652116

41. Hart EA, Hua L, Darr LB, Wilson WK, Pang J, Matsuda SPT (1999) "Directed evolution to investigate steric control of enzymatic oxidosqualene cyclization. An isoleucine-to-valine mutation in cycloartenol synthase allows lanosterol and parkeol biosynthesis." *Journal of the American Chemical Society*, 121(42):9887–9888. https://doi.org/10.1021/ja992589b

42. Wendt KU, Poralla K, Schulz GE (1997) "Structure and function of a squalene cyclase." *Science*, 277(5333):1811–1815. https://doi.org/10.1126/science.277.5333.1811

43. Thoma R, Schulz-Gasen T, D'Arcy B, Benz J, Aebi J, Dehmlow H, Hennig M, Stihle M, Ruf A (2004) "Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase." *Nature*, 432:118–122. https://doi.org/10.1038/nature02993

44. Branden C, Tooze J (1999) "Introduction to Protein Structure." https://books.google.co.uk/books?id=eUYWBAAAQBAJ

45. Salmon M, Thimmappa RB, Minto RE, Melton RE, Hughes RK, O'Maille PE, Hemmings AM, Osbourn A (2016) "A conserved amino acid residue critical for product and substrate specificity in plant triterpene synthases." *Proceedings of the National Academy of Sciences*, 113(30):E4407–E4414. https://doi.org/10.1073/pnas.1605509113

46. Schrödinger LLC "The PyMOL Molecular Graphics System."

47. Cantarel BI, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) "The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics." *Nucleic Acids Research*, 37(suppl_1):D233–D238. https://doi.org/10.1093/nar/gkn663

48. Vetter J (2000) "Plant cyanogenic glycosides." *Toxicon*, 38(1):11–36. https://doi.org/10.1016/S0041-0101(99)00128-2

49. Bowles D, Lim E-K, Poppenberger B, Vaistij FE (2006) "Glycosyltransferases of lipophilic small molecules." *Annual Review of Plant Biology*, 57:567–597. https://doi.org/10.1146/annurev.arplant.57.032905.105429

50. Piochon M, Legault J, Gauthier C, Pichette A (2009) "Synthesis and cytotoxicity evaluation of natural α-bisabolol β-d-fucopyranoside and analogues." *Phytochemistry*, 70(2):228–236. https://doi.org/10.1016/j.phytochem.2008.11.013

51. Vogt T, Jones P (2000) “Glycosyltransferases in plant natural product synthesis: characterization of a supergene family.” *Trends Plant Sci*, 6(9):380–386. https://doi.org/10.1016/S1360-1385(00)01720-9

52. Shao H, He X, Achnine L, Blount JW, Dixon RA, Wang X (2005) “Crystal structures of a multifunctional triterpene/flavonoid glycosyltransferase from Medicago truncatula.” *Plant Cell*, 17(11):3141–3154. https://doi.org/10.1105/tpc.105.035055

53. Osmani SA, Bak S, Møller BL (2009) “Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling.” *Phytochemistry*, 70(3):325–347. https://doi.org/10.1016/j.phytochem.2008.12.009

54. Davies GJ, Sinnott ML (2008) “Sorting the diverse: the sequence-based classifications of carbohydrate-active enzymes.” *Biochemical Journal*, 30(4):26–32. https://doi.org/10.1042/BIO03004026

55. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A (2006) “Structures and mechanisms of glycosyltransferases.” *Glycobiology*, 16(2):29R-37R. https://doi.org/10.1093/glycob/cwj016

56. Henrissat B (1991) “A classification of glycosyl hydrolases based on amino acid sequence similarities.” *Biochemical Journal*, 280(2):309–316. https://doi.org/10.1042/bj2800309

57. Lutz S (2010) “Beyond directed evolution-semi-rational protein engineering and design.” *Current Opinion in Biotechnology*, 21(6):734–743. https://doi.org/10.1016/j.copbio.2010.08.011

58. Powell KA, Ramer SW, Del Cardayr SB, Stemmer WPC, Tobin MB, Longchamp PF, Huisman GW (2001) “Directed evolution and biocatalysis.” *Angewandte Chemie - International Edition*, 40(21):3948–3959. https://doi.org/10.1002/1521-3773(20011105)40:21<3948::AID-ANIE3948>3.0.CO;2-N

59. Whitehouse CJC, Bell SG, Wong L-L (2012) “P450 $_{BM3}$ (CYP102A1): connecting the dots.” *Chem. Soc. Rev.*, 41(3):1218–1260. https://doi.org/10.1039/C1CS15192D

60. Reetz MT, Carballeira JD (2007) “Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes.” *Nature Protocols*, 2:891–903. https://doi.org/10.1038/nprot.2007.72

61. Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J (2007) “Engineering proteinase K using machine learning and synthetic genes.” *BMC Biotechnology*, 7:16. https://doi.org/10.1186/1472-6750-7-16

62. Wu S, Acevedo JP, Reetz MT (2010) “Induced allostery in the directed evolution of an enantioselective Baeyer-Villiger monooxygenase.” *Proceedings of the National Academy of Sciences of the United States of America*, 107(7):2775–2780. https://doi.org/10.1073/pnas.0911656107

63. Saavedra HG, Wrabl JO, Anderson JA, Li J, Hilser VJ (2018) “Dynamic allostery can drive cold adaptation in enzymes.” *Nature*, 558:324–328. https://doi.org/10.1038/s41586-018-0183-2

64. Sumbalova L, Stourac J, Martinek T, Bednar D, Damborsky J (2018) “HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information.” *Nucleic Acids Research*, 46(W1):W356–W362. https://doi.org/10.1093/nar/gky417

65. Reetz MT, Wang LW, Bocola M (2006) “Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space.” *Angewandte Chemie - International Edition*, 45(8):1236–1241. https://doi.org/10.1002/anie.200502746

66. Ehren J, Govindarajan S, Morón B, Minshull J, Khosla C (2008) “Protein engineering of improved prolyl endopeptidases for celiac sprue therapy.” *Protein Engineering, Design and Selection*, 21(12):699-707. https://doi.org/10.1093/protein/gzn050

67. Reed J, Stephenson MJ, Miettinen K, Brouwer B, Leveau A, Brett P, Goss RJM, Goossens A, O’Connell MA, Osbourn A (2017) “A translational synthetic biology platform for rapid access to gram-scale quantities of novel drug-like molecules.” *Metabolic Engineering*, 42:185–193. https://doi.org/10.1016/j.ymben.2017.06.012

68. Osbourn AE, Clarke BR, Dow JM, Daniels MJ (1991) “Partial characterization of avenacinase from *Gaeumannomyces graminis* var. *avenae*.” *Physiological and Molecular Plant Pathology*, 38(4):301–312. https://doi.org/10.1016/S0885-5765(05)80121-3

69. Smyth MS, Martin JHJ (2000) “x Ray crystallography.” *Journal of Clinical Pathology - Molecular Pathology*, 53:8–14. https://doi.org/10.1136/mp.53.1.8

70. Punta M, Simon I, Dosztányi Z (2014) “Prediction and analysis of intrinsically disordered proteins.” *Structural Proteomics: High-Throughput Methods: Second Edition*, 1261:35–39. https://doi.org/10.1007/978-1-4939-2230-7_3

71. Pantazatos D, Kim JS, Klock HE, Stevens RC, Wilson IA, Lesley SA, Woods VL (2004) “Rapid refinement of crystallographic protein construct definition employing enhanced hydrogen/deuterium exchange MS.” *Proceedings of the National Academy of Sciences of the United States of America*, 101(3):751–756. https://doi.org/10.1073/pnas.0307204101

72. Mooij WTM, Mitsiki E, Perrakis A (2009) “ProteinCCD: Enabling the design of protein truncation constructs for expression and crystallization experiments.” *Nucleic Acids Research*, 37(suppl_2):W402–W405. https://doi.org/10.1093/nar/gkp256

73. Smits SHJ, Mueller A, Grieshaber MK, Schmitt L (2008) “Coenzyme- and His-tag-induced crystallization of octopine dehydrogenase.” *Acta Crystallographica Section F Structural Biology and Crystallization Communications*, 64(9):836–839. https://doi.org/10.1107/S1744309108025487

74. Stevens RC (2000) “Design of high-throughput methods of protein production for structural biology.” *Structure*, 8(9):R177–R185. https://doi.org/10.1016/S0969-2126(00)00193-3

75. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, Heijne G von, Nielsen H (2019) “SignalP 5.0 improves signal peptide predictions using deep neural networks.” *Nature Biotechnology*, 37(4):420–423. https://doi.org/10.1038/s41587-019-0036-z

76. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) “Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life.” *Journal of Molecular Biology*, 337(3):635–645. https://doi.org/10.1016/j.jmb.2004.02.002

77. Ishida T, Kinoshita K (2007) “PrDOS: Prediction of disordered protein regions from amino acid sequence.” *Nucleic Acids Research*, 35(suppl_2):W460–W464. https://doi.org/10.1093/nar/gkm363

78. Mészáros B, Erdös G, Dosztányi Z (2018) “IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding.” *Nucleic Acids Research*, 46(W1):W329–W337. https://doi.org/10.1093/nar/gky384

79. Hanson J, Paliwal KK, Litfin T, Zhou Y (2019) “SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning.” *Genomics, Proteomics and Bioinformatics*, 17(6):645–656. https://doi.org/10.1016/j.gpb.2019.01.004

80. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, Marcatili P (2019) “NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning.” *Proteins: Structure, Function and Bioinformatics*, 87(6):520–527. https://doi.org/10.1002/prot.25674

81. Erlich HA, Gelfand D, Sninsky JJ (1991) “Recent advances in the polymerase chain reaction.” *Science*, 252(5013):1643–1651. https://doi.org/10.1126/science.2047872

82. Schochetman G, Ou CY, Jones WK (1988) “Polymerase chain reaction.” *Journal of Infectious Diseases*, 158(6):1154–1157. https://doi.org/10.1093/infdis/158.6.1154

83. Garibyan L, Avashia N (2013) “Polymerase chain reaction.” *Journal of Investigative Dermatology*, 133(3):1–4. https://doi.org/10.1038/jid.2013.1

84. Mullis K, Ferre F, Gibbs R (1994) “The Polymerase Chain Reaction.” https://books.google.co.uk/books?id=woNO4w5HweQC

85. Erlich HA (1989) “Polymerase chain reaction.” *Journal of Clinical Immunology*, 9(6):437–447. https://doi.org/10.1007/BF00918012

86. McInerney P, Adams P, Hadi MZ (2014) “Error Rate Comparison during Polymerase Chain Reaction by DNA Polymerase.” *Molecular Biology International*, 2014:ID 287430. https://doi.org/10.1155/2014/287430

87. Kibbe WA (2007) “OligoCalc: An online oligonucleotide properties calculator.” *Nucleic Acids Research*, 35(suppl_2):W43–W46. https://doi.org/10.1093/nar/gkm234

88. Clontech® Laboratories Inc (2014) “In-Fusion® HD Cloning Kit User Manual.” p6. https://vdocument.in/in-fusion-hd-cloning-kit.html

89. Kim Y, Choi E, Son B, Seo E, Lee E, Ryu J, Ha G, Kim J, Kwon M, Nam J, Kim Y, Lee K (2012) “Effects of Storage Buffer and Temperature on the Integrity of Human DNA.” *Korean Journal of Clinical Laboratory Science*, 44(1):24–30. http://www.kjcls.org/journal/view.html?volume=44&number=1&spage=24

90. Thomas CM, Summers D (2020) “Bacterial Plasmids.” *eLS*, 1:240–250 https://doi.org/10.1002/9780470015902.a0029193

91. Casali N, Preston A (2003) “E. coli Plasmid Vectors : Methods and Applications.” *Methods in Molecular Biology*, 235. https://archive.org/details/springer_10.1385-1592594093/

92. Castagnoli L, Scarpa M, Kokkinidis M, Banner DW, Tsernoglou D, Cesareni G (1989) “Genetic and structural analysis of the ColE1 Rop (Rom) protein.” *EMBO Journal*, 8:621–629. https://doi.org/10.1002/j.1460-2075.1989.tb03417.x

93. Chen D qiang, Zheng X cong, Lu Y jun (2006) “Identification and characterization of novel ColE1-type, high-copy number plasmid mutants in Legionella pneumophila.” *Plasmid*, 56(3):167–178. https://doi.org/10.1016/j.plasmid.2006.05.008

94. O’Sullivan DJ, Klaenhammer TR (1993) “High- and low-copy-number Lactococcus shuttle cloning vectors with features for clone screening.” *Gene*, 137(2):227–231. https://doi.org/10.1016/0378-1119(93)90011-Q

95. Makrides SC (1996) “Strategies for achieving high-level expression of genes in Escherichia coli.” *Microbiological Reviews*, 60(3):512–538. https://doi.org/10.1128/mmbr.60.3.512-538.1996

96. Jayaraman R (2010) “Jacques Monod and the advent of the age of operons.” *Resonance*, 15:1084–1096. https://doi.org/10.1007/s12045-010-0121-6

97. Wright RM, Thompson HL, Freundt E (2017) “Transformation of a Mixed Probiotic Culture and Escherichia coli B with the Antibiotic Resistant Plasmid, pGLO.” *Acta Spartae*, 3(1):13–17. https://doi.org/10.48497/ez7f-nf70

98. Korpimäki T, Kurittu J, Karp M (2003) “Surprisingly fast disappearance of β-lactam selection pressure in cultivation as detected with novel biosensing approaches.” *Journal of Microbiological Methods*, 53(1):37–42. https://doi.org/10.1016/S0167-7012(02)00213-0

99. Shaw WV (1983) “Chloramphenicol acetyltransferase: Enzymology and molecular biology.” *Critical Reviews in Biochemistry and Molecular Biology*, 14(1):1–46. https://doi.org/10.3109/10409238309102789

100. Wilson DN (2014) “Ribosome-targeting antibiotics and mechanisms of bacterial resistance.” *Nature Reviews Microbiology*, 12(1):35–48. https://doi.org/10.1038/nrmicro3155

101. Evangelisti E, Yunusov T, Shenhav L, Schornack S (2019) “N-acetyltransferase AAC(3)-I confers gentamicin resistance to *Phytophthora palmivora* and *Phytophthora infestans*.” *BMC Microbiology*, 19:265. https://doi.org/10.1186/s12866-019-1642-0

102. Chankova SG, Dimova E, Dimitrova M, Bryant PE (2007) “Induction of DNA double-strand breaks by zeocin in *Chlamydomonas reinhardtii* and the role of increased DNA double-strand breaks rejoining in the formation of an adaptive response.” *Radiation and Environmental Biophysics*, 46(4):409–416. https://doi.org/10.1007/s00411-007-0123-2

103. Gatignol A, Durand H, Tiraby G (1988) “Bleomycin resistance conferred by a drug-binding protein.” *FEBS Letters*, 230(1–2):171–175. https://doi.org/10.1016/0014-5793(88)80665-3

104. “SnapGene software by Dotmatics.” https://www.snapgene.com/

105. Reece-Hoyes JS, Walhout AJM (2018) “Gateway recombinational cloning.” *Cold Spring Harbor Protocols*, 2018(1):1–6. https://doi.org/10.1101/pdb.top094912

106. Yu XH, Liu CJ (2006) “Development of an analytical method for genome-wide functional identification of plant acyl-coenzyme A-dependent acyltransferases.” *Analytical Biochemistry*, 358(1):146–148. https://doi.org/10.1016/j.ab.2006.08.012

107. Luria S, Human M (1952) “A Nonhereditary, Host-Induced Variation of Bacterial Viruses.” *Journal of bacteriology*, 64(4):557–569. https://doi.org/10.1128/jb.64.4.557-569.1952

108. Pingoud A, Alves J, Geiger R (1993) “Restriction enzymes.” *Methods in molecular biology*, 16:107–200. https://doi.org/10.1385/0-89603-234-5:107

109. Innis M, Gelfand D, Sninsky J, White T (1990) “PCR Protocols: A Guide to Methods and Applications.” https://books.google.co.uk/books?id=Z5jwZ2rbVe8C

110. Invitrogen (2002) “T4 DNA Ligase.” https://assets.thermofisher.com/TFS-Assets/LSG/manuals/t4dnaligase_1U_man.pdf

111. Zheng L, Baumann U, Reymond JL (2004) “An efficient one-step site-directed and site-saturation mutagenesis protocol.” *Nucleic acids research*, 32(14):e115. https://doi.org/10.1093/nar/gnh110

112. Liu H, Naismith JH (2008) “An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol.” *BMC biotechnology*, 8:91. https://doi.org/10.1186/1472-6750-8-91

113. Vovis GF, Lacks S (1977) “Complementary action of restriction enzymes endo R · DpnI and endo R · DpnII on bacteriophage f1 DNA.” *Journal of Molecular Biology*, 115(3):525–538. https://doi.org/10.1016/0022-2836(77)90169-3

114. Johnston C, Martin B, Fichant G, Polard P, Claverys JP (2014) “Bacterial transformation: Distribution, shared mechanisms and divergent control.” *Nature Reviews Microbiology*, 12(3):181–196. https://doi.org/10.1038/nrmicro3199

115. Acquistapace IM (2018) “An investigation into the structural determinants of the positional specificity of hydrolysis of myo-inositol hexakisphosphate by HP phytases.” *University of East Anglia*, Ph.D thesis. https://ueaeprints.uea.ac.uk/id/eprint/73011/1/IMA_Thesis_2018_PhD_Biomolecular_Sciences.pdf

116. Chang AY, Chau VW, Landas JA, Yvonne (2017) “Preparation of calcium competent Escherichia coli and heat-shock transformation.” *Journal of Experimental Microbiology and Immunology (JEMI)*, 1:22–25. https://ujemi.microbiology.ubc.ca/sites/default/files/Chang%20et%20al%20JEMI-methods%20Vol%201%20pg%2022-25.pdf

117. Bergkessel M, Guthrie C (2013) “Chapter Twenty Five - Colony PCR.” *Methods in Enzymology*, 529:299–309. https://doi.org/10.1016/B978-0-12-418687-3.00025-2

118. RCSB PDB (2022) “PDB Statistics: PDB Data Distribution by Expression Organism (Gene Source).” https://www.rcsb.org/stats/distribution-expression-organism-gene

119. Baeshen NA, Baeshen MN, Sheikh A, Bora RS, Ahmed MMM, Ramadan HAI, Saini KS, Redwan EM (2014) “Cell factories for insulin production.” *Microbial Cell Factories*, 13:141. https://doi.org/10.1186/s12934-014-0141-0

120. Rosano GL, Ceccarelli EA (2014) “Recombinant protein expression in Escherichia coli: Advances and challenges.” *Frontiers in Microbiology*, 5:172. https://doi.org/10.3389/fmicb.2014.00172

121. Rudolph R, Lilie H (1996) “In vitro folding of inclusion body proteins.” *The FASEB Journal*, 10(1):49–56. https://doi.org/10.1096/fasebj.10.1.8566547

122. Lederberg J (2004) “E. coli K-12.” *Microbiology Today*, 31:116. https://socgenmicrobiol.org.uk/pubs/micro_today/pdf/080402.pdf

123. Daegelen P, Studier FW, Lenski RE, Cure S, Kim JF (2009) “Tracing Ancestors and Relatives of Escherichia coli B, and the Derivation of B Strains REL606 and BL21(DE3).” *Journal of Molecular Biology*, 394(4):634–643. https://doi.org/10.1016/j.jmb.2009.09.022

124. Sørensen HP, Mortensen KK (2005) “Advanced genetic strategies for recombinant protein expression in Escherichia coli.” *Journal of Biotechnology*, 115(2):113–128. https://doi.org/10.1016/j.jbiotec.2004.08.004

125. Ritz D, Lim J, Reynolds CM, Poole LB, Beckwith J (2001) “Conversion of a peroxiredoxin into a disulfide reductase by a triplet repeat expansion.” *Science*, 294(5540):158–160. https://doi.org/10.1126/science.1063143

126. Lobstein J, Emrich CA, Jeans C, Faulkner M, Riggs P, Berkmen M (2012) “SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm.” *Microbial Cell Factories*, 11:753. https://doi.org/10.1186/1475-2859-11-56

127. Thirumalai D, Lorimer GH (2001) “Chaperonin-mediated protein folding.” *Annual Review of Biophysics and Biomolecular Structure*, 30:245–269. https://doi.org/10.1146/annurev.biophys.30.1.245

128. Ferrer M, Chernikova TN, Yakimov MM, Golyshin PN, Timmis KN (2003) “Chaperonins govern growth of Escherichia coli at low temperatures.” *Nature Biotechnology*, 21(11):1266–1267. https://doi.org/10.1038/nbt1103-1266

129. Genlantis (2004) “SoluBL21™ Competent E. coli.” https://lib.store.yahoo.net/lib/yhst-131428861332406/SoluBL21-Toxic-Clones.pdf

130. Braun P, LaBaer J (2003) “High throughput protein production for functional proteomics.” *Trends in Biotechnology*, 21(9):383–388. https://doi.org/10.1016/S0167-7799(03)00189-6

131. Cereghino JL, Cregg JM (2000) “Heterologous protein expression in the methylotrophic yeast Pichia pastoris.” *FEMS Microbiology Reviews*, 24(1):45–66. https://doi.org/10.1111/j.1574-6976.2000.tb00532.x

132. Byrne B (2015) “Pichia pastoris as an expression host for membrane protein structural biology.” *Current Opinion in Structural Biology*, 32:9–17. https://doi.org/10.1016/j.sbi.2015.01.005

133. Newstead S, Kim H, Von Heijne G, Iwata S, Drew D (2007) “High-throughput fluorescent-based optimization of eukaryotic membrane protein overexpression and purification in Saccharomyces cerevisiae.” *Proceedings of the National Academy of Sciences of the United States of America*, 104(35):13936–13941. https://doi.org/10.1073/pnas.0704546104

134. Rodríguez RF (2018) “Structure-Function Studies of a Purple Acid Phytase.” *University of East Anglia*, Ph.D thesis. https://ueaeprints.uea.ac.uk/id/eprint/69554/1/2018Faba-RodriguezRFRPhD.pdf

135. Ersson B, Rydén L, Jan-Christer Janson (2011) “Introduction to Protein Purification.” *Protein Purification: Principles, High Resolution Methods, and Applications, Third Edition*, I:1–22. https://doi.org/10.1002/9780470939932

136. Kim Y, Bigelow L, Borovilos M, Dementieva I, Duggan E, Eschenfeldt W, Hatzos C, Joachimiak G, Li H, Maltseva N, Mulligan R, Quartey P, Sather A, Stols L, Volkart L, Wu R, Zhou M, Joachimiak A (2008) “High-Throughput Protein Purification for X-Ray Crystallography and NMR.” *Advances in Protein Chemistry and Structural Biology*, 75:85–105. https://doi.org/10.1016/S0065-3233(07)75003-9

137. Westermeier R (2011) “Electrophoresis in Gels.” *Protein Purification: Principles, High Resolution Methods, and Applications, Third Edition*, IV:363–377. https://doi.org/10.1002/9780470939932

138. Coskun O (2016) “Separation Techniques: Chromatography.” *Northern Clinics of Istanbul*, 3(2):156–160. https://doi.org/10.14744/nci.2016.32757

139. Urh M, Simpson D, Zhao K (2009) “Affinity Chromatography: General Methods.” *Methods in Enzymology*, 463(C):417–438. https://doi.org/10.1016/S0076-6879(09)63026-3

140. Schmidt TGM, Skerra A (2007) “The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins.” *Nature Protocols*, 2:1528–1535. https://doi.org/10.1038/nprot.2007.209

141. Qiagen (2003) “The QIAexpressionist™.” *Qiagen GmbH, Düsseldorf, Germany*, https://www.qiagen.com/us/resources/resourcedetail?id=79ca2f7d-42fe-4d62-8676-4cfa948c9435&lang=en

142. GE Healthcare (2009) “HisTrap affinity columns instructions.” https://webhome.auburn.edu/~duinedu/manuals/HisTrapHP.pdf

143. Mori S, Barth HG (1999) “1.2 Brief Description of Size Exclusion Chromatography.” *Size Exclusion Chromatography*, 1:5. https://books.google.co.uk/books?id=2D3R8qvJgt8C

144. GE Healthcare (2011) “HiLoad Superdex data file.” https://sapientia.ualg.pt/bitstream/10400.1/3116/14/SEC%20Columns.pdf

145. Amersham Biosciences (2002) “Ion Exchange Chromatography Principles and Methods.” https://www.ualberta.ca/biological-sciences/media-library/mbsu/protein-purification-and-fplc/ion-exchange-chrom.pdf

146. Brunelle JL, Green R (2014) “One-dimensional SDS-polyacrylamide gel electrophoresis (1D SDS-PAGE).” *Methods in Enzymology*, 541:151–159. https://doi.org/10.1016/B978-0-12-420119-4.00012-4

147. Hnasko TS, Hnasko RM (2015) “The Western Blot.” *Methods in Molecular Biology*, 1318:87–96. https://doi.org/10.1007/978-1-4939-2742-5_9

148. Mahmood T, Yang PC (2012) “Western blot: Technique, theory, and trouble shooting.” *North American Journal of Medical Sciences*, 4(9):429–434. https://doi.org/10.4103/1947-2714.100998

149. RCSB PDB (2022) “PDB Data Distribution by Experimental Method and Molecular Type.” https://www.rcsb.org/stats/summary

150. Bai X, McMullan G, Scheres SHW (2015) “How cryo-EM is revolutionizing structural biology.” *Trends in Biochemical Sciences*, 40(1):49–57. https://doi.org/10.1016/j.tibs.2014.10.005

151. Wu M, Lander GC (2020) “How low can we go? Structure determination of small biological complexes using single-particle cryo-EM.” *Current Opinion in Structural Biology*, 64:9–16. https://doi.org/10.1016/j.sbi.2020.05.007

152. Derewenda ZS, Vekilov PG (2006) “Entropy and surface engineering in protein crystallization.” *Acta Crystallographica Section D: Biological Crystallography*, 62:116–124. https://doi.org/10.1107/S0907444905035237

153. McPherson A, Gavira JA (2014) “Introduction to protein crystallization.” *Acta Crystallographica Section F:Structural Biology Communications*, 70:2–20. https://doi.org/10.1107/S2053230X13033141

154. Chayen NE (2004) “Turning protein crystallisation from an art into a science.” *Current Opinion in Structural Biology*, 14(5):577–583. https://doi.org/10.1016/j.sbi.2004.08.002

155. Hemmings A (2021) “Personal communication.”

156. Pflugrath JW (2004) “Macromolecular cryocrystallography - Methods for cooling and mounting protein crystals at cryogenic temperatures.” *Methods*, 34(3):415–423. https://doi.org/10.1016/j.ymeth.2004.03.032

157. Elton LRB, Jackson DF (1966) “X-Ray Diffraction and the Bragg Law.” *American Journal of Physics*, 34(11):1036. https://doi.org/10.1119/1.1972439

158. Casanas A, Warshamanage R, Finke AD, Panepucci E, Olieric V, Nöll A, Tampé R, Brandstetter S, Förster A, Mueller M, Schulze-Briese C, Bunk O, Wang M (2016) “EIGER detector: Application in macromolecular crystallography:” *Acta Crystallographica Section D: Structural Biology*, 72(9)https://doi.org/10.1107/S2059798316012304

159. Sun S, Zhang X, Cui J, Liang S (2020) “Identification of the Miller indices of a crystallographic plane: A tutorial and a comprehensive review on fundamental theory,

universal methods based on different case studies and matters needing attention." *Nanoscale*, 12:16657–16677. https://doi.org/10.1039/d0nr03637d

160. Evans PR (2011) "An introduction to data reduction: Space-group determination, scaling and intensity statistics." *Acta Crystallographica Section D: Biological Crystallography*, 67:282–292. https://doi.org/10.1107/S090744491003982X

161. Powell HR (2017) "X-ray data processing." *Bioscience Reports*, 37(5):BSR20170227. https://doi.org/10.1042/BSR20170227

162. Winter G, McAuley KE (2011) "Automated data collection for macromolecular crystallography." *Methods*, 55(1):81–93. https://doi.org/10.1016/j.ymeth.2011.06.010

163. Winter G, Waterman DG, Parkhurst JM, Brewster AS, Gildea RJ, Gerstel M, Fuentes-Montero L, Vollmar M, Michels-Clark T, Young ID, Sauter NK, Evans G (2018) "DIALS: Implementation and evaluation of a new integration package." *Acta Crystallographica Section D: Structural Biology*, 74:85–97. https://doi.org/10.1107/S2059798317017235

164. Winter G (2010) "Xia2: An expert system for macromolecular crystallography data reduction." *Journal of Applied Crystallography*, 43:186–190. https://doi.org/10.1107/S0021889809045701

165. Jeffrey P (2006) "X-ray Data Collection Course." http://xray0.princeton.edu/~phil/Facility/Guides/XrayDataCollection.html

166. International Union of Crystallography (2022) "Data recommended for inclusion in Acta Cryst. F articles." https://journals.iucr.org/f/services/evaluationcriteria.html

167. Dubach VRA, Guskov A (2020) "The resolution in x-ray crystallography and single-particle cryogenic electron microscopy." *Crystals*, 10(7):580. https://doi.org/10.3390/cryst10070580

168. Karplus PA, Diederichs K (2015) "Assessing and maximizing data quality in macromolecular crystallography." *Current Opinion in Structural Biology*, 34:60–68. https://doi.org/10.1016/j.sbi.2015.07.003

169. Emsley P, Cowtan K (2004) "Coot: Model-building tools for molecular graphics." *Acta Crystallographica Section D: Biological Crystallography*, 60:2126–2132. https://doi.org/10.1107/S0907444904019158

170. Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA (2011) "REFMAC5 for the refinement of macromolecular crystal structures." *Acta Crystallographica Section D: Biological Crystallography*, 67(4):355–367. https://doi.org/10.1107/S0907444911001314

171. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, 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 Crystallographica Section D: Biological Crystallography*, 66:213–221. https://doi.org/10.1107/S0907444909052925

172. Sheldrick GM (2008) “A short history of SHELX.” *Acta Crystallographica Section A: Foundations of Crystallography*, 64:112–122. https://doi.org/10.1107/S0108767307043930

173. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) “Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures.” *FEBS Journal*, 275(1):1–21. https://doi.org/10.1111/j.1742-4658.2007.06178.x

174. Young JY, Westbrook JD, Feng Z, Sala R, Peisach E, Oldfield TJ, Sen S, Gutmanas A, Armstrong DR, Berrisford JM, Chen L, Chen M, Di Costanzo L, Dimitropoulos D, Gao G, Ghosh S, Gore S, Guranovic V, Hendrickx PMS, Hudson BP, Igarashi R, Ikegawa Y, Kobayashi N, Lawson CL, Liang Y, Mading S, Mak L, Mir MS, Mukhopadhyay A, Patwardhan A, Persikova I, Rinaldi L, Sanz-Garcia E, Sekharan MR, Shao C, Swaminathan GJ, Tan L, Ulrich EL, Ginkel G van, Yamashita R, Yang H, Zhuravleva MA, Quesada M, Kleywegt GJ, Berman HM, Markley JL, Nakamura H, Velankar S, Burley SK (2017) “OneDep: Unified wwPDB System for Deposition, Biocuration, and Validation of Macromolecular Structures in the PDB Archive.” *Structure*, 25(3):536–545. https://doi.org/10.1016/j.str.2017.01.004

175. Krieger E, Nabuurs SB, Vriend G (2003) “Homology Modeling.” *Structural Bioinformatics*, 44:509–523. https://doi.org/10.1002/0471721204

176. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021) “Highly accurate protein structure prediction with AlphaFold.” *Nature*, 596:583–589. https://doi.org/10.1038/s41586-021-03819-2

177. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2021) “Critical assessment of methods of protein structure prediction (CASP)—Round XIV.” *Proteins: Structure, Function and Bioinformatics*, 89:1607–1617. https://doi.org/10.1002/prot.26237

178. Ovchinnikov S, Kim DE, Wang RYR, Liu Y, Dimaio F, Baker D (2016) “Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta.” *Proteins: Structure, Function and Bioinformatics*, 84(S1):67–75. https://doi.org/10.1002/prot.24974

179. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A (2018) “Critical assessment of methods of protein structure prediction (CASP)—Round XII.” *Proteins: Structure, Function and Bioinformatics*, 86(S1):7–15. https://doi.org/10.1002/prot.25415

180. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D (2019) “Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13).” *Proteins: Structure, Function and Bioinformatics*, 87(12):1141–1148. https://doi.org/10.1002/prot.25834

181. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J (2019) “Critical assessment of methods of protein structure prediction (CASP)—Round XIII.” *Proteins: Structure, Function and Bioinformatics*, 87(12):1011–1020. https://doi.org/10.1002/prot.25823

182. Hess B, Kutzner C, Spoel D Van Der (2008) “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation.” *Journal of Chemical Theory and Computation*, 4(3):435–447. https://doi.org/10.1021/ct700301q

183. Allen MP (2004) “Introduction to Molecular Dynamics Simulation.” *Computational Soft Matter: From Synthetic Polymers to Proteins*, 23:1–28. https://dasher.wustl.edu/chem478/reading/md-intro-2.pdf

184. Betz RM, Walker RC (2015) “Paramfit: Automated optimization of force field parameters for molecular dynamics simulations.” *Journal of Computational Chemistry*, 36(2):79–87. https://doi.org/10.1002/jcc.23775

185. Mu Y, Kosov DS, Stock G (2003) “Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments.” *Journal of Physical Chemistry B*, 107(21):5064–5073. https://doi.org/10.1021/jp022445a

186. Kadaoluwa Pathirannahalage SP, Meftahi N, Elbourne A, Weiss ACG, McConville CF, Padua A, Winkler DA, Costa Gomes M, Greaves TL, Le TC, Besford QA, Christofferson AJ (2021) “Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations.” *Journal of Chemical Information and Modeling*, 61(9):4521–4536. https://doi.org/10.1021/acs.jcim.1c00794

187. Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, Niida R, Watanabe-Takahashi A, Inoue E, Saito K (2008) “Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in Arabidopsis.” *Plant Cell*, 20(8):2160–2176. https://doi.org/10.1105/tpc.108.058040

188. Ross J, Li Y, Lim E-K, Bowles DJ (2001) “Higher plant glycosyltransferases.” *Genome Biology*, 2:reviews3004.1. https://doi.org/10.1186/gb-2001-2-2-reviews3004

189. D’Auria JC, Reichelt M, Luck K, Svatoš A, Gershenzon J (2007) “Identification and characterization of the BAHD acyltransferase malonyl CoA: Anthocyanidin 5-O-glucoside-6″-O-malonyltransferase (At5MAT) in Arabidopsis thaliana.” *FEBS Letters*, 581(5):872–878. https://doi.org/10.1016/j.febslet.2007.01.060

190. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A (2005) “Protein Identification and Analysis Tools on the ExPASy Server.” *The Proteomics Protocols Handbook*, :571–607. https://doi.org/10.1385/1-59259-890-0:571

191. Jancarik J, Kim SH (1991) “Sparse matrix sampling. A screening method for crystallization of proteins.” *Journal of Applied Crystallography*, 24:409–411. https://doi.org/10.1107/S0021889891004430

192. Collins B, Stevens RC, Page R (2005) “Crystallization optimum solubility screening: Using crystallization results to identify the optimal buffer for protein crystal formation.” *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 61:1035–1038. https://doi.org/10.1107/S1744309105035244

193. Newman J, Egan D, Walter TS, Meged R, Berry I, Jelloul MB, Sussman JL, Stuart DI, Perrakis A (2005) “Towards rationalization of crystallization screening for small- To medium-sized academic laboratories: The PACT/JCSG+ strategy.” *Acta Crystallographica*

Section D: Biological Crystallography, 61:1426–1431. https://doi.org/10.1107/S0907444905024984

194. Gorrec F (2009) "The MORPHEUS protein crystallization screen." *Journal of Applied Crystallography*, 42:1035–1042. https://doi.org/10.1107/S0021889809042022

195. Grimm C, Chari A, Reuter K, Fischer U (2010) "A crystallization screen based on alternative polymeric precipitants." *Acta Crystallographica Section D: Biological Crystallography*, 66:685–697. https://doi.org/10.1107/S0907444910009005

196. Gräslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schütz A, Heinemann U, Yokoyama S, Büssow K, Gunsalus KC (2008) "Protein production and purification." *Nature Methods*, 5:135–146. https://doi.org/10.1038/nmeth.f.202

197. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2014) "The I-TASSER suite: Protein structure and function prediction." *Nature Methods*, 12:7–8. https://doi.org/10.1038/nmeth.3213

198. Šali A, Blundell TL (1993) "Comparative protein modelling by satisfaction of spatial restraints." *Journal of Molecular Biology*, 234(3):779–815. https://doi.org/10.1006/jmbi.1993.1626

199. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) "UCSF Chimera - A visualization system for exploratory research and analysis." *Journal of Computational Chemistry*, 25(13):1605–1612. https://doi.org/10.1002/jcc.20084

200. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) "GROMACS: Fast, flexible, and free." *Journal of Computational Chemistry*, 26(16):1701–1718. https://doi.org/10.1002/jcc.20291

201. Oostenbrink C, Villa A, Mark AE, Van Gunsteren WF (2004) "A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6." *Journal of Computational Chemistry*, 25(13):1656–1676. https://doi.org/10.1002/jcc.20090

202. Benkert P, Tosatto SCE, Schomburg D (2008) "QMEAN: A comprehensive scoring function for model quality assessment." *Proteins-Structure Function and Bioinformatics*, 71(1):261–277. https://doi.org/10.1002/prot.21715

203. 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 Crystallographica Section D: Biological Crystallography*, 66:12–21. https://doi.org/10.1107/S0907444909042073

204. Bolanos-Garcia VM, Davies OR (2006) “Structural analysis and classification of native proteins from E. coli commonly co-purified by immobilised metal affinity chromatography.” *Biochimica et Biophysica Acta - General Subjects*, 1760(9):1304–1313. https://doi.org/10.1016/j.bbagen.2006.03.027

205. Dong A, Xu X, Edwards AM, Chang C, Chruszcz M, Cuff M, Cymborowski M, Leo R Di, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov G V., Bochkarev A, Dahlgren LG, Dhe-Paganon S, Dimov S, Dombrovski L, Finerty P, Flodin S, Flores A, Gräslund S, Hammerström M, Herman MD, Hong BS, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park HW, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H (2007) “In situ proteolysis for protein crystallization and structure determination.” *Nature Methods*, 4(12):1019–1021. https://doi.org/10.1038/nmeth1118

206. Fujiyama K, Hino T, Kanadani M, Watanabe B, Jae Lee H, Mizutani M, Nagano S (2019) “Structural insights into a key step of brassinosteroid biosynthesis and its inhibition.” *Nature Plants*, 5:589–594. https://doi.org/10.1038/s41477-019-0436-6

207. Wetterhorn KM, Newmister SA, Caniza RK, Busman M, McCormick SP, Berthiller F, Adam G, Rayment I (2016) “Crystal Structure of Os79 (Os04g0206600) from Oryza sativa: A UDP-glucosyltransferase Involved in the Detoxification of Deoxynivalenol.” *Biochemistry*, 55(44):6175–6186. https://doi.org/10.1021/acs.biochem.6b00709

208. Saridakis E, Chayen NE (2009) “Towards a ‘universal’ nucleant for protein crystallization.” *Trends in Biotechnology*, 27(2):99–106. https://doi.org/10.1016/j.tibtech.2008.10.008

209. Bergfors T (2003) “Seeds to crystals.” *Journal of Structural Biology*, 142(1):66–76. https://doi.org/10.1016/S1047-8477(03)00039-X

210. Gibson RP, Tarling CA, Roberts S, Withers SG, Davies GJ (2004) “The donor subsite of trehalose-6-phosphate synthase: Binary complexes with udp-glucose and udp-2-deoxy-2-fluoro-glucose at 2 Å resolution.” *Journal of Biological Chemistry*, 279(3):1950–1955. https://doi.org/10.1074/jbc.M307643200

211. Opassiri R, Pomthong B, Onkoksoong T, Akiyama T, Esen A, Ketudat Cairns JR (2006) “Analysis of rice glycosyl hydrolase family 1 and expression of Os4bglu12 β-glucosidase.” *BMC Plant Biology*, 6:33. https://doi.org/10.1186/1471-2229-6-33

212. Ketudat Cairns JR, Pengthaisong S, Luang S, Sansenya S, Tankrathok A, Svasti J (2012) “Protein-carbohydrate Interactions Leading to Hydrolysis and Transglycosylation in Plant Glycoside Hydrolase Family 1 Enzymes.” *Journal of Applied Glycoscience*, 59(2):51–62. https://doi.org/10.5458/jag.jag.jag-2011_022

213. Sainsbury F, Thuenemann EC, Lomonossoff GP (2009) “PEAQ: Versatile expression vectors for easy and quick transient expression of heterologous proteins in plants.” *Plant Biotechnology Journal*, 7(7):682–693. https://doi.org/10.1111/j.1467-7652.2009.00434.x

214. CAZy “Glycoside Hydrolase Family 1, Characterized.” http://www.cazy.org/GH1_characterized.html

215. The UniProt Consortium (2021) “UniProt: the universal protein knowledgebase in 2021.” *Nucleic Acids Research*, 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100

216. Notredame C, Higgins DG, Heringa J (2000) “T-coffee: A novel method for fast and accurate multiple sequence alignment.” *Journal of Molecular Biology*, 302(1):205–217. https://doi.org/10.1006/jmbi.2000.4042

217. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V (2018) “A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.” *Journal of Molecular Biology*, 430(15):2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

218. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, De Beer TAP, Rempfer C, Bordoli L, Lepore R, Schwede T (2018) “SWISS-MODEL: Homology modelling of protein structures and complexes.” *Nucleic Acids Research*, 46(W1):W293–W303. https://doi.org/10.1093/nar/gky427

219. Seshadri S, Akiyama T, Opassiri R, Kuaprasert B, Ketudat Cairns J (2009) “Structural and enzymatic characterization of Os3BGlu6, a rice beta;-glucosidase hydrolyzing hydrophobic glycosides and (1→3)- and (1→2)-linked disaccharides.” *Plant Physiology*, 151(1):47–58. https://doi.org/10.1104/pp.109.139436

220. Song Y, Dimaio F, Wang RYR, Kim D, Miles C, Brunette T, Thompson J, Baker D (2013) “High-resolution comparative modeling with RosettaCM.” *Structure*, 21(10):1735–1742. https://doi.org/10.1016/j.str.2013.08.005

221. Hua Y, Sansenya S, Saetang C, Wakuta S, Cairns JRK (2013) “Enzymatic and structural characterization of hydrolysis of gibberellin A4 glucosyl ester by a rice β-D-glucosidase.” *Archives of Biochemistry and Biophysics*, 537(1):39–48. https://doi.org/10.1016/j.abb.2013.06.005

222. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE (2021) “PubChem in 2021: New data content and improved web interfaces.” *Nucleic Acids Research*, 49(D1):D1388–D1395. https://doi.org/10.1093/nar/gkaa971

223. Koziara KB, Stroet M, Malde AK, Mark AE (2014) “Testing and validation of the Automated Topology Builder (ATB) version 2.0: Prediction of hydration free enthalpies.” *Journal of Computer-Aided Molecular Design*, 28:221–233. https://doi.org/10.1007/s10822-014-9713-7

224. Gordon JC, Myers JB, Folta T, Shoja V, Heath LS, Onufriev A (2005) “H++: A server for estimating pKas and adding missing hydrogens to macromolecules.” *Nucleic Acids Research*, 33(suppl_2):W368–W371. https://doi.org/10.1093/nar/gki464

225. Daura X, Gademann K, Jaun B, Seebach D, Gunsteren WF van, Mark AE (2004) “Peptide Folding: When Simulation Meets Experiment.” *Angewandte Chemie International Edition*, 38(1–2):236–240. https://doi.org/10.1002/(sici)1521-3773(19990115)38:1/2<236::aid-anie236>3.3.co;2-d

226. Gräslund S, Sagemark J, Berglund H, Dahlgren LG, Flores A, Hammarström M, Johansson I, Kotenyova T, Nilsson M, Nordlund P, Weigelt J (2008) "The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins." *Protein Expression and Purification*, 58(2):210–221. https://doi.org/10.1016/j.pep.2007.11.008

227. Costa SJ, Almeida A, Castro A, Domingues L, Besir H (2013) "The novel Fh8 and H fusion partners for soluble protein expression in Escherichia coli: A comparison with the traditional gene fusion technology." *Applied Microbiology and Biotechnology*, 97:6779–6791. https://doi.org/10.1007/s00253-012-4559-1

228. Smyth DR, Mrozkiewicz MK, McGrath WJ, Listwan P, Kobe B (2003) "Crystal structures of fusion proteins with large-affinity tags." *Protein Science*, 12(7):1313–1322. https://doi.org/10.1110/ps.0243403

229. Kajikawa M, Yamato KT, Fukuzawa H, Sakai Y, Uchida H, Ohyama K (2005) "Cloning and characterization of a cDNA encoding β-amyrin synthase from petroleum plant *Euphorbia tirucalli* L." *Phytochemistry*, 66(15):1759–1766. https://doi.org/10.1016/j.phytochem.2005.05.021

230. Haralampidis K, Bryan G, Qi X, Papadopoulou K, Bakht S, Melton R, Osbourn A (2001) "A new class of oxidosqualene cyclases directs synthesis of antimicrobial phytoprotectants in monocots." *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13431–13436. https://doi.org/10.1073/pnas.231324698

231. Srivastava G, Sandeep, Garg A, Misra RC, Chanotiya CS, Ghosh S (2020) "Transcriptome analysis and functional characterization of oxidosqualene cyclases of the arjuna triterpene saponin pathway." *Plant Science*, 292:110382. https://doi.org/10.1016/j.plantsci.2019.110382

232. Ito R, Mori K, Hashimoto I, Nakano C, Sato T, Hoshino T (2011) "Triterpene Cyclases from *Oryza sativa* L.: Cycloartenol, Parkeol and Achilleol B Synthases." *Organic Letters*, 13(10):2678–2681. https://doi.org/10.1021/ol200777d

233. Nielsen PE, Nishimura H, Liang Y, Calvin M (1979) "Steroids from Euphorbia and other latex-bearing plants." *Phytochemistry*, 18:103–104. https://doi.org/10.1016/S0031-9422(00)90923-3

234. Mwine J, Van Damme P, Hastilestari BR, Papenbrock J (2013) "*Euphorbia tirucalli* L. (Euphorbiaceae) – The Miracle Tree: Current Status of Knowledge." *African Natural Plant Products Volume II: Discoveries and Challenges in Chemistry, Health, and Nutrition*, 1127:3–17. https://doi.org/10.1021/bk-2013-1127.ch001

235. Uchida H, Ohyama K, Suzuki M, Yamashita H, Muranaka T, Ohyama K (2010) "Triterpenoid levels are reduced during *Euphorbia tirucalli* L. callus formation." *Plant Biotechnology*, 27(1):105–109. https://doi.org/10.5511/plantbiotechnology.27.105

236. Biesboer DD, D'Amour P, Wilson SR, Mahlberg P (1982) "Sterols and triterpenols in latex and cultured tissues of Euphorbia pulcherrima." *Phytochemistry*, 21(5):1115–1118. https://doi.org/10.1016/S0031-9422(00)82427-9

237. Liu Y, Cai Y, Zhao Z, Wang J, Li J, Xin W, Xia G, Xiang F (2009) “Cloning and Functional Analysis of a β-Amyrin Synthase Gene Associated with Oleanolic Acid Biosynthesis in *Gentiana straminea* MAXIM.” *Biological & Pharmaceutical Bulletin*, 32(5):818–824. https://doi.org/10.1248/bpb.32.818

238. Luo Y, Jiang Y, Chen L, Li C, Wang Y (2023) “Applications of protein engineering in the microbial synthesis of plant triterpenoids.” *Synthetic and Systems Biotechnology*, 8(1):20–32. https://doi.org/10.1016/j.synbio.2022.10.001

239. Ito R, Hashimoto I, Masukawa Y, Hoshino T (2013) “Effect of Cation-π Interactions and Steric Bulk on the Catalytic Action of Oxidosqualene Cyclase: A Case Study of Phe728 of β-Amyrin Synthase from *Euphorbia tirucalli* L.” *Chemistry - A European Journal*, 19(50):17150–17158. https://doi.org/10.1002/chem.201301917

240. Ito R, Masukawa Y, Nakada C, Amari K, Nakano C, Hoshino T (2014) “β-Amyrin synthase from *Euphorbia tirucalli*. Steric bulk, not the π-electrons of Phe, at position 474 has a key role in affording the correct folding of the substrate to complete the normal polycyclization cascade.” *Org. Biomol. Chem.*, 12(23):3836–3846. https://doi.org/10.1039/C4OB00064A

241. Ito R, Nakada C, Hoshino T (2017) “β-Amyrin synthase from *Euphorbia tirucalli* L. functional analyses of the highly conserved aromatic residues Phe413, Tyr259 and Trp257 disclose the importance of the appropriate steric bulk, and cation–π and CH–π interactions for the efficient catalytic action of the polyolefin cyclization cascade.” *Organic & Biomolecular Chemistry*, 15(1):177–188. https://doi.org/10.1039/C6OB02539K

242. Hoshino T, Nakagawa K, Aiba Y, Itoh D, Nakada C, Masukawa Y (2017) “*Euphorbia tirucalli* β-Amyrin Synthase: Critical Roles of Steric Sizes at Val483 and Met729 and the CH-π Interaction between Val483 and Trp534 for Catalytic Action.” *ChemBioChem*, 18(21):2145–2155. https://doi.org/10.1002/cbic.201700368

243. Aiba Y, Watanabe T, Terasawa Y, Nakano C, Hoshino T (2018) “Strictly Conserved Residues in *Euphorbia tirucalli* β-Amyrin Cyclase: Trp612 Stabilizes Transient Cation through Cation-π Interaction and CH-π Interaction of Tyr736 with Leu734 Confers Robust Local Protein Architecture.” *ChemBioChem*, 19(5):486–495. https://doi.org/10.1002/cbic.201700572

244. Dokarry M (2010) “β-amyrin synthase: Investigating the structure and function of an oxidosqualene cyclase involved in disease resistance in oats.” *University of East Anglia*, Ph.D thesis. https://ueaeprints.uea.ac.uk/id/eprint/20511/1/2010DokarryMPhD.pdf

245. Fazio VJ, Peat TS, Newman J (2014) “A drunken search in crystallization space.” *Acta Crystallographica Section F: Structural Biology Communications*, 70:1303–1311. https://doi.org/10.1107/S2053230X1401841X

246. Parker JL, Newstead S (2012) “Current trends in α-helical membrane protein crystallization: An update.” *Protein Science*, 21(9):1358–1365. https://doi.org/10.1002/pro.2122

247. Lamb DC, Jackson CJ, Warrilow AGS, Manning NJ, Kelly DE, Kelly SL (2007) “Lanosterol biosynthesis in the prokaryote Methylococcus Capsulatus: Insight into the evolution of

sterol biosynthesis." *Molecular Biology and Evolution*, 24(8):1714–1721. https://doi.org/10.1093/molbev/msm090

248. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." *Molecular Systems Biology*, 7:539. https://doi.org/10.1038/msb.2011.75

249. Gavriljuk K, Itzen A, Goody RS, Gerwert K, Kötting C (2013) "Membrane extraction of Rab proteins by GDP dissociation inhibitor characterized using attenuated total reflection infrared spectroscopy." *Proceedings of the National Academy of Sciences of the United States of America*, 110(33):13380–13385. https://doi.org/10.1073/pnas.1307655110

250. Takemura M, Tanaka R, Misawa N (2017) "Pathway engineering for the production of β-amyrin and cycloartenol in Escherichia coli—a method to biosynthesize plant-derived triterpene skeletons in E. coli." *Applied Microbiology and Biotechnology*, 101:6615–6625. https://doi.org/10.1007/s00253-017-8409-z

251. Google, DeepMind (2021) "AlphaFold Colab." https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb

252. Sockolosky JT, Szoka FC (2013) "Periplasmic production via the pET expression system of soluble, bioactive human growth hormone." *Protein Expression and Purification*, 87(2):129–135. https://doi.org/10.1016/j.pep.2012.11.002

253. Makino T, Skretas G, Georgiou G (2011) "Strain engineering for improved expression of recombinant proteins in bacteria." *Microbial Cell Factories*, 10:32. https://doi.org/10.1186/1475-2859-10-32

254. Das A, Bysack A, Raghuraman H (2021) "Effectiveness of dual-detergent strategy using Triton X-100 in membrane protein purification." *Biochemical and Biophysical Research Communications*, 578:122–128. https://doi.org/10.1016/j.bbrc.2021.09.031

255. Wang M, Nie Y, Wu XL (2021) "Extracellular heme recycling and sharing across species by novel mycomembrane vesicles of a Gram-positive bacterium." *ISME Journal*, 15:605–617. https://doi.org/10.1038/s41396-020-00800-1

256. Heistinger L, Gasser B, Mattanovich D (2020) "Microbe profile: Komagataella phaffii: A methanol devouring biotech yeast formerly known as pichia pastoris." *Microbiology*, 166(7):614–616. https://doi.org/10.1099/mic.0.000958

257. Sasagawa T, Matsui M, Kobayashi Y, Otagiri M, Moriya S, Sakamoto Y, Ito Y, Lee CC, Kitamoto K, Arioka M (2011) "High-throughput recombinant gene expression systems in Pichia pastoris using newly developed plasmid vectors." *Plasmid*, 65(1):65–69. https://doi.org/10.1016/j.plasmid.2010.08.004

258. Waters MG, Evans EA, Blobel G (1988) "Prepro-α-factor has a cleavable signal sequence." *Journal of Biological Chemistry*, 263(13):6209–6214. https://doi.org/10.1016/s0021-9258(18)68773-3

259. Aw R, Polizzi KM (2013) “Can too many copies spoil the broth?” *Microbial Cell Factories*, 12:128. https://doi.org/10.1186/1475-2859-12-128

260. Lee CC, Williams TG, Wong DWS, Robertson GH (2005) “An episomal expression vector for screening mutant gene libraries in Pichia pastoris.” *Plasmid*, 54(1):80–85. https://doi.org/10.1016/j.plasmid.2004.12.001

261. Baumann K, Adelantado N, Lang C, Mattanovich D, Ferrer P (2011) “Protein trafficking, ergosterol biosynthesis and membrane physics impact recombinant protein secretion in Pichia pastoris.” *Microbial Cell Factories*, 10:93. https://doi.org/10.1186/1475-2859-10-93

262. Chang ACY, Cohen SN (1978) “Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the P15A cryptic miniplasmid.” *Journal of Bacteriology*, 134(3):1141–1156. https://doi.org/10.1128/jb.134.3.1141-1156.1978

263. Bentley WE, Mirjalili N, Andersen DC, Davis RH, Kompala DS (1990) “Plasmid-encoded protein: The principal factor in the ‘metabolic burden’ associated with recombinant bacteria.” *Biotechnology and Bioengineering*, 35(7):668–681. https://doi.org/10.1002/bit.260350704

264. Kushiro T, Shibuya M, Masuda K, Ebizuka Y (2000) “Mutational studies on triterpene synthases: Engineering lupeol synthase into β-amyrin synthase.” *Journal of the American Chemical Society*, 122(29):6816–6824. https://doi.org/10.1021/ja0010709

276. Loewen P (1996) “Probing the structure of catalase HPII of Escherichia coli - A review.” *Gene*, 179(1):39–44. https://doi.org/10.1016/S0378-1119(96)00321-6

277. Bravo J, Fita I, Ferrer JC, Ens W, Hillar A, Switala J, Loewen PC (1997) “Identification of a novel bond between a histidine and the essential tyrosine in catalase HPII of Escherichia coli.” *Protein Science*, 6(5):1016–1023. https://doi.org/10.1002/pro.5560060507

278. Maté MJ, Sevinc MS, Hu B, Bujons J, Bravo J, Switala J, Ens W, Loewen PC, Fita I (1999) “Mutants that alter the covalent structure of catalase hydroperoxidase II Escherichia coli.” *Journal of Biological Chemistry*, 274(39):27717–27725. https://doi.org/10.1074/jbc.274.39.27717

279. Nicholls P, Fita I, Loewen PC (2000) “Enzymology and structure of catalases.” *Advances in Inorganic Chemistry*, 51:51–106. https://doi.org/10.1016/s0898-8838(00)51001-0

280. Spicer S (2017) “The nuts and bolts of evaluating science communication activities.” *Seminars in Cell and Developmental Biology*, 70:17–25. https://doi.org/10.1016/j.semcdb.2017.08.026