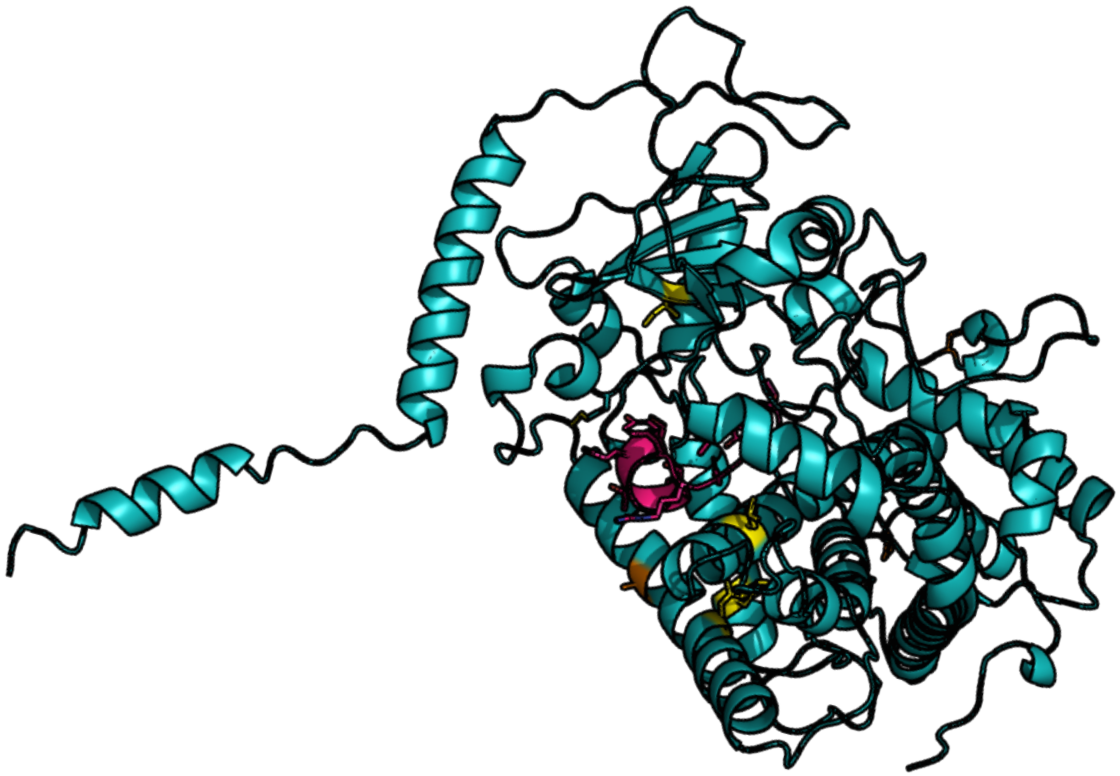
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

DEPARTMENT OF PLANT AND ENVIRONMENTAL SCIENCE



Bachelor Thesis

Jens Sigurd Agger Raabyemagle

Experimentally investigating the use of machine learning as a computational tool for enzyme engineering of CYP79A1.

Supervisors: Assistant Professor Tomas Laursen & Ph.D. Student Kasper Hinz

Submitted on: June 16th, 2023

Preface

This Bachelor Thesis is the culmination of research conducted between February 2023 to June 2023 at the Dynamic Metabolons research group, Section for Plant Biochemistry, Department of Plant and Environmental Science, Faculty of Science, University of Copenhagen. This project concludes the first three years of my academic journey.

First, I would like to thank my primary supervisors Assistant Professor Tomas Laursen and Ph.D. student Kasper Hinz for their guidance and patience with me throughout the entire project. I would also like to thank Ph.D. Pengfei Tian and Professor Wouter Boomsma for their work with, and explanations of the machine learning aspects that lay the foundation for this project. Finally, I would like to thank Assistant Professor Silas Busck Mellor and the rest of the Dynamic Metabolons group for their guidance and inclusion throughout my time at their workplace.

Abstract

Plant specialized metabolites exhibit a remarkable diversity in both their nature and biological functions, making some highly attractive for various industries, including pharmaceuticals, fragrances, and pesticides. The biosynthetic pathways involved in the production of these specialized metabolites often comprise multiple enzymes, among which the cytochrome P450 (P450) enzyme superfamily often plays a prominent role. The central focus of this thesis revolves around exploring the application of artificial intelligence as a powerful computational tool for enzyme and pathway engineering. Specifically, two machine learning models, Cavity and EVE, were employed to propose single amino acid substitutions in a model P450 enzyme, CYP79A1, which is the enzyme catalyzing the committed step in the biosynthesis of the cyanogenic glucoside dhurrin in *Sorghum bicolor*. The aim behind this was to optimize structure and performance as a consequence of the mutations. Mutated variants of CYP79A1 were generated and heterologously expressed in *Saccharomyces cerevisiae* to test the effect of the computationally predicted mutations *in vitro*. Fermentation experiments were conducted to evaluate the differences in metabolite production among the mutant variants, followed by LC-MS analysis to determine the metabolic profiles. The results revealed that almost all of the initial substrate for the pathway had been consumed at two fermentation time points, 24 hours and 48 hours. For this reason, the elucidation of the kinetic properties of the mutated enzymes could not be estimated. The majority of the mutant variants of CYP79A1 were revealed to produce significantly less dhurrin than the wild-type pathway, while three remained unaffected. Employing other experimental strategies might further characterize the true effects of the machine learning suggested mutations.

Contents

1. Introduction

1.1. Plant specialized metabolites and their biosynthesis

Unlike species from the animal kingdom who have the ability to fight or flight when faced with challenges from the environment around them, plants are stationary and therefore have been pressured to evolve their own defense strategies. This has resulted in plants producing an array of 'secondary' or 'specialized' metabolites. These metabolites are not entirely necessary for plant growth and are synthesized from products of primary metabolism, hence the term 'secondary' (Delgoda & Murray, 2017).

Plant specialized metabolites can be divided into three major classifications which are all metabolized from primary metabolism products. The first group is terpenes which are built from 5-carbon units. These often contain toxic properties and are usually produced by plants to deter herbivorous mammals and insects (Taiz et al., 2015). The second group, phenolic compounds, can also act as a deterrent against herbivores, but also provide structural support and protection against harsh environments. This includes involvement in lignin formation and UV radiation screening (Lattanzio, 2013). The third group of specialized metabolites are nitrogen-containing compounds. These metabolites, similarly to terpenes and phenolic compounds, are usually also toxic to herbivores.

Because of the various local environments in which plants have evolved, there is also a huge diversity amongst specialized metabolites. More than 200,000 secondary metabolites have been isolated and more continue to be elucidated (Dixon & Strack, 2003; Turi et al., 2015). Many of these have been utilized by industries to produce fragrance, medicine, pesticides and more (D'Addabbo et al., 2014; Nagegowda & Gupta, 2020; Newman & Cragg, 2020).

The common denominator for all plant specialized metabolites, is that enzymes drive the catalysis of their biosynthesis. Among these are a superfamily consisting of more than 32,000 specifically plant-endogenous enzymes called Cytochrome P450's, that are ubiquitous in land plants (as well as other organisms), and catalyze a large variety of reactions leading to metabolic diversification across different species (Hansen et al., 2021).

1.2. Enzyme engineering to optimize production of high value compounds

Specialized metabolites are synthesized through complex biosynthetic pathways, often consisting of numerous enzymatic reactions, which all require energy and resources. As a result, plants often resort to producing specialized metabolites in very low quantities. Additionally, the biosynthesis is often regulated and might require specific environmental conditions or stress factors (D'Amelia et al., 2021). This poses a problem when manufacturing specialized metabolites for industrial purposes, where

higher yield is more favorable. In a continuously growing population that requires efficient infrastructure, agriculture, and sustainable means of production, it becomes imperative to explore strategies for enhancing the production of specialized metabolites.

Enzymes have a wide range of applications across various industries. Industrial enzymes are extensively employed in the food, beverage, detergent, and biofuel sectors, underscoring their significance to society. Enzymes have evolved naturally to catalyze a diverse array of reactions, and advancements in technology have enabled the acceleration and modification of this evolutionary process through enzyme engineering, which aims to improve already existing enzymes or synthetic enzymes with respects to their structural properties and/or production efficiency (Sharma et al., 2021).

Several strategies can be employed when it comes to engineering of enzymes. One of them is the directed evolution approach. This approach is about generating large libraries of mutated target enzymes, and then conducting high-throughput screening of these to isolate mutations that are favorable with respects to structural properties and production efficiency (Sharma et al., 2021). The libraries are constructed using a variety of strategies including: saturation mutagenesis where each amino acid in a given enzyme is respectively substituted by all other amino acids (Palackal, 2004), truncation where C- or N-terminals of enzymes are removed (Kim et al., 2011), and random mutagenesis where error-prone polymerases are utilized in PCRs to create large numbers of mutant variants of an enzyme (Yokoyama et al., 2010). There is an advantage in using directed evolution for enzyme engineering, since there is no need for in depth understanding of the enzymes of interest – one can just screen for mutations that are more efficient or stable compared to their wildtype (WT). There is however the disadvantage that it is tedious and laborious work to construct and screen libraries consisting of thousands of mutations.

Another more direct approach for enzyme engineering is using rational design. This method requires knowledge of the enzyme beforehand, and employs site-directed mutagenesis to directly substitute residue positions of interest (Sharma et al., 2021). As an example, Baojin Fei and colleagues (Fei et al., 2013) tried to optimize thermostability of *Escherichia coli* enzyme AppA phytase on the basis of research in protein flexibility, surface and salt bridges by employing site directed mutagenesis to substitute target residues. This approach is more time efficient but requires a solid background of information to determine significant enzyme residues. Semi-rational design uses a combination of directed evolution and rational design. Here, prior knowledge of the enzymes or the biochemical properties of these are used to construct libraries that are more targeted and not as comprehensive as in directed evolution approaches (Sharma et al., 2021). Lingyun Rui and colleagues (Rui et al., 2004)

applied this approach by using saturation mutagenesis to directly engineer active-site residues in epoxide hydrolase from *Agrobacterium radiobacter.*

1.3. Artificial Intelligence as a tool for enzyme engineering

Artificial intelligence (AI) has emerged as an ubiquitous and potent tool across multiple industries, with continued development of novel models aimed at tackling distinct academic and industrial challenges. Vast amounts of biological data have been amassed over time and are archived in diverse public databases, such as the Worldwide Protein Data Bank, UniProt, and GenBank. Such voluminous data has served as the foundation for the training of several machine learning (a subset of AI) models, which seek to make predictions on the basis of these. Machine learning has demonstrated its efficacy in predicting beneficial amino acid substitutions in PET hydrolases that are involved in the degradation of plastics, as illustrated by Lu et al. (2022). In this particular study, a machine learning algorithm was trained on over 19,000 protein structures from the Protein Data Bank, allowing it to accurately predict mutations that were subsequently validated through experimentation, thereby enhancing the thermostability of the hydrolase. Additionally, Paik et al. (2023) have successfully applied machine learning algorithms to engineer an optimized *Bst* DNA polymerase. Notably, both of these examples harness the power of machine learning to predict outcomes that would typically be achieved through directed evolution, negating the need for the time-consuming processes such as high-throughput screening. In essence, this methodology simulates saturation mutagenesis and screening, thus offering an efficient and promising alternative to experimental biology.

1.3.1. The Cavity and EVE models

This project employs two machine learning algorithms for the optimization of enzymatic function. The first model is the Cavity Model (unpublished), which has been trained with approximately 16,000 AlphaFold (Jumper et al., 2021; Varadi et al., 2022) predicted three-dimensional protein structures. Like in the two beforementioned machine leaning models, the Cavity model simulates saturation mutagenesis by removing the amino acid side chain of a residue and returns a probability matrix, containing the probability of each amino acid taking that residues position (Figure 1). The probability matrix is generated for each residue position of the enzyme. The results are then run on another algorithm that labels the suggested mutations with their respective impacts on the difference in thermodynamic free energy between folded and non-folded state of the enzyme (ΔΔG). The model weighs heavily on the partial charge and mass of amino acids, with respects to their local microenvironments.

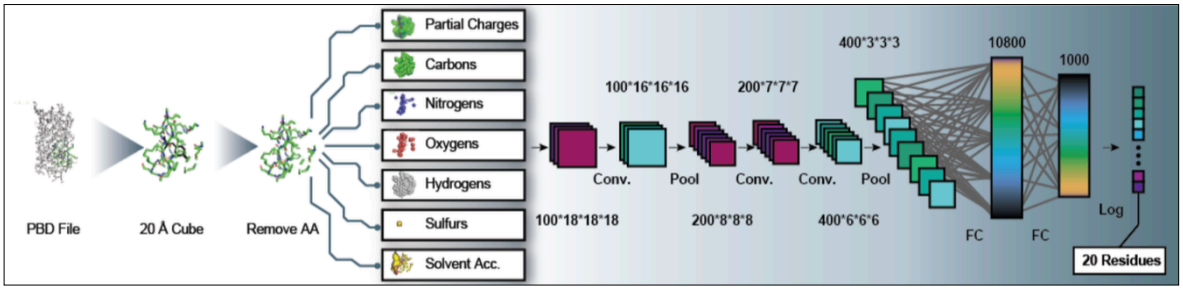


Figure 1 - Concept behind the Cavity model. A PDB file is fed to the model. The model then removes a residue and maps the chemical microenvironment around it in 3D cubes. The data in the cubes are then run through multiple convoluting and pooling layers of a convolutional neural network before returning a probability matrix of all 20 amino acid posibilities for that position (Shroff et al., 2020).

The other machine learning model employed in this study is the Evolutionary model of Variant Effect (EVE), which was developed by Frazer et al. (2021). The primary focus of this model is to predict the pathogenicity of all amino acids situated in all positions of a human protein. In this project, however, the EVE model is adapted to predict amino acid mutations that enhance enzyme fitness. Similar to the aforementioned machine learning model, the EVE model utilizes neural network algorithms but differs in terms of input data. Specifically, the EVE model uses multiple sequence alignments of homologous protein sequences as input data and finds patterns in the evolutionary data, including co-variance, thus generating predictions of fitness optimizing character.

1.4. The dhurrin metabolon

Dhurrin is a type of cyanogenic glucoside that belongs to the subgroup of specialized metabolites containing nitrogen. The biosynthesis of dhurrin in *Sorghum bicolor* is the culmination of a metabolic pathway that involves two key cytochrome P450 (CYP) enzymes, namely CYP79A1 and CYP71E1, as well as a glycosyl-transferase known as UGT85B1. The production of dhurrin through this pathway begins with the conversion of L-tyrosine to *p*-hydroxyphenylacetaldoxime (oxime), catalyzed by CYP79A1. This intermediate is then converted to *p*-hydroxymandelonitrile via the catalyzation of CYP71E1, which is subsequently used as substrate by UGT85B1, resulting in the production of the final metabolite, dhurrin (Figure 2.A) (Kahn et al., 1999). To drive the functions of the two CYP's, an NADPH dependent cytochrome p450 oxidoreductase (POR) donates electrons to the heme-group employed by the respective CYP's (Jensen & Møller, 2010).

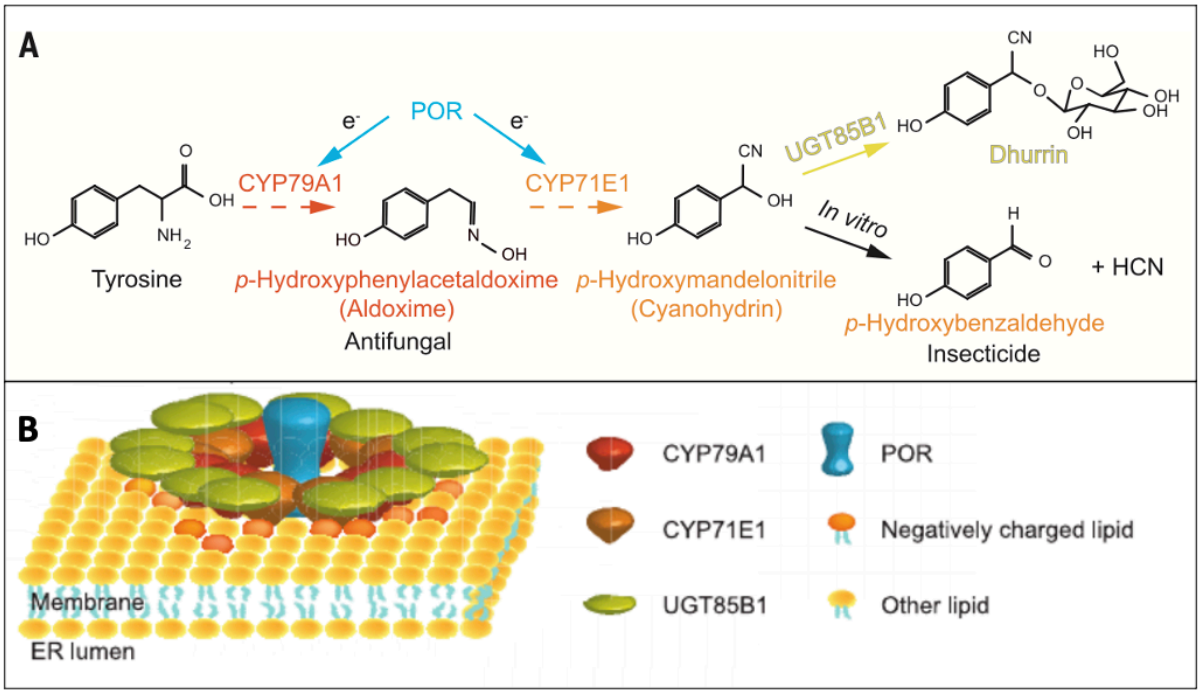


Figure 2 - A.) The biosynthetic pathway for dhurrin in S. bicolor. Dashed arrows indicate more than one step. B.) Visualization of the assembly of the dhurrin metabolon. Figure modified from (Laursen et al, 2016).

The metabolic pathway for dhurrin biosynthesis in *Sorghum bicolor* operates within large enzyme complexes, which are referred to as metabolons and are located in the endoplasmic reticulum (ER) membrane. These metabolons provide several advantages, including enhanced efficiency by channeling pathway intermediates while minimizing metabolic crosstalk, which results in higher yields of dhurrin. Additionally, inclusion of a single membrane-bound POR that supplies electrons to a myriad of membrane-bound CYPs further increases the efficiency of these metabolons (Figure 2.B) (Laursen et al., 2016). The disassembly of the metabolon would result in accumulation of the oxime intermediate, which has been suggested to be applied as a plant defense against fungal infections (Møller, 2010).

Notably, these features make the dhurrin metabolon ideal for investigation into both metabolic and enzyme engineering. Detailed studies of the enzymes involved in the pathway as well as the assembly and disassembly of the metabolon have been conducted, highlighting the different components within.

1.5. Heterologous expression of biosynthetic pathways

Many plants exhibit slow growth rates, leading to limited production of specialized metabolites. To address this challenge, it is preferable to express biosynthetic pathways in more suitable hosts by means of heterologous expression. One such host is *Saccharomyces cerevisiae*, commonly referred to as baker's yeast, which multiplies quickly and therefore can facilitate efficient expression in numerous

individual cells, thereby obtaining much higher titers of specialized metabolites. Yeast is also a suitable model organism for heterologous expression of membrane bound proteins such as CYP's, which cannot be expressed in prokaryotes such as *Escherichia coli,* due to the lack of compartmentalized organelles (Kotopka & Smolke, 2019). Kotopka and Smolke successfully expressed the biosynthetic pathway for dhurrin in yeast, while also further increasing dhurrin titers by overexpressing the genes coding for the enzymes that produce L-tyrosine, which is the initial substrate in the pathway.

DNA can easily be integrated into *S. cerevisiae* chromosomal DNA by homologous recombination. Research has shown that by using an integration platform that takes advantage of this phenomenon, whole biosynthetic pathways consisting of up to eight genes can reliably be heterologously expressed in yeast (Shao et al., 2009). To establish optimal sites for homologous recombination, an expression platform has been developed by Mikkelsen et al. (2012). Specifically, this platform recommends specific integration sites in chromosomes 10, 11, and 12 that are characterized by high expression levels (Figure 3). Moreover, each of these genomic locations is strategically positioned within the yeast genome such that they are flanked by genes that are essential for growth, acting as a control measure ensuring that the integration of foreign DNA does not interfere with the cells' fitness. By leveraging a platform as such, researchers can more precisely control the insertion of genetic material into the yeast genome, maximizing expression whilst reducing the potential effects on cell viability.

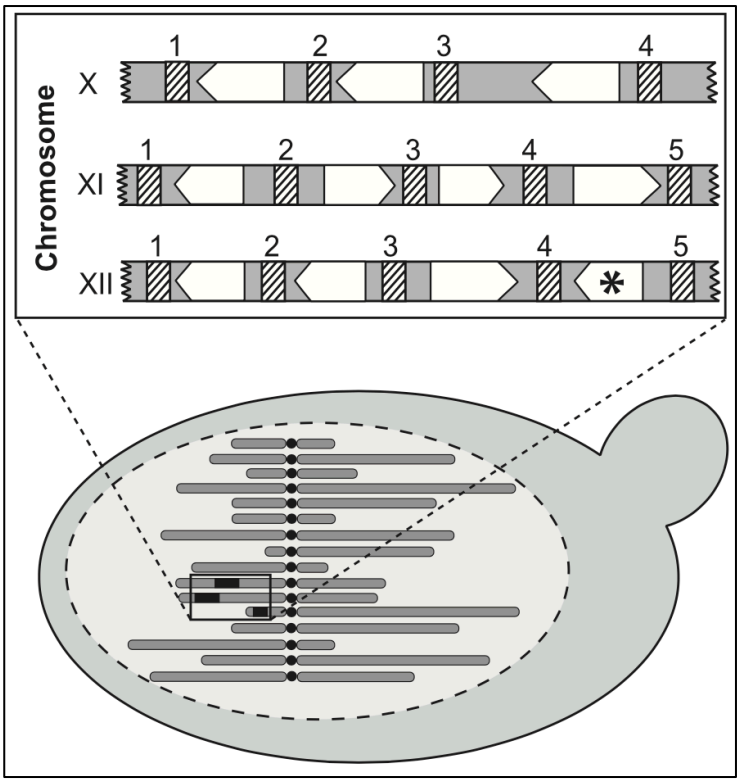


Figure 3 - S. cerevisiae expression platform. 14 integration sites represented by striped boxes are suggested throughout yeast chromosome 10, 11, and 12. These are divided by genes that are crucial for cell viability, represented by white arrows or white arrow with asterisk. Modified from (Mikkelsen et al., 2012).

1.6. Scope of the project and experimental strategy

The aim of this study was to explore the application of machine learning in enzyme and metabolic engineering. Specifically, experiments were applied to elucidate the potential effects of single amino acid mutations in enzymes, suggested by the machine learning models Cavity and EVE. This study focused on a model enzyme, CYP79A1, which catalyzes the committed step in the biosynthetic pathway for the production of dhurrin. The CYP79A1 mutations proposed by the models were evaluated, and eight variants were selected on the basis of their anticipated effects on the structural and performance properties of the enzyme. To achieve this, *S. cerevisiae*, acting as heterologous expression host, was employed to perform batch-fermentations containing strains transformed with the respective mutant variants of CYP79A1 along with the rest of the dhurrin pathway. The quantification of the pathway metabolites was achieved using targeted metabolomics and used to determine the consequences of the mutations.

2. Materials & Methods

2.1. Protein visualization

CYP79A1 protein structure was visualized using the PyMOL (PyMOL, The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) software. UniProt accession number Q43135 was used to acquire the AlphaFold (Jumper et al., 2021; Varadi et al., 2022) .pdb file used in visualizing the protein.

2.2. Generation of mutated CYP79A1 by site directed mutagenesis

During the course of the project, site-directed mutagenesis was employed to introduce mutations suggested by the Cavity and EVE models. Throughout the project, three different approaches were utilized for site-directed mutagenesis. The most successful approach, described later, involved a site-directed mutagenesis protocol that relied on partially overlapping primers (Figure 4) (Liu & Naismith, 2008). The desired mutations were strategically placed in the overlapping 5' ends of the primers. Importantly, the non-overlapping 3' ends of the primers allowed for the generated fragment to serve as a template for subsequent cycles of PCR. This allowed for higher PCR efficiency, whereas this would not be the case, were the primers completely overlapping.

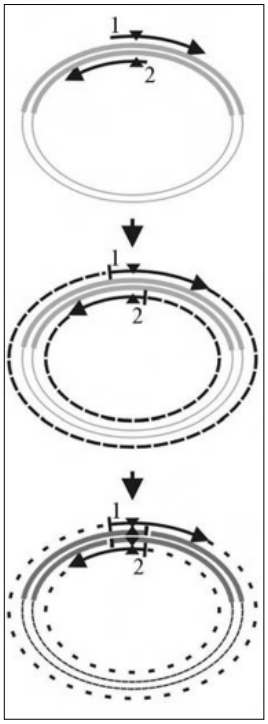


Figure 4 - Site directed mutagenesis with partially overlapping primers. Triangles indicate mutation site. (Liu & Naismith, 2008)

2.2.1. Generation of CYP79A1 mutant strains

Mutant strains of CYP79A1 were generated through site-directed mutagenesis using a PCR-based approach. The mutagenic primers (Appendix 1), designed to partially overlap with each other, were employed for PCR amplification of a pJET1.2 vector containing the gene encoding CYP79A1. PCR reactions were prepared in a volume of 50 µL, containing 10 µL of 10x PCR buffer (Appendix 5), 5 µL of dNTP's (2.5mM), 2.5 µL of each primer (10 µM), 1 µL of template plasmid harboring the CYP79A1 gene (20 ng) (Appendix 8), 0.5 µL of PfuX7 polymerase (10 U/µL) (Nørholm, 2010), and 28.5 µL of deionized water. The PCR cycling conditions were modified from the description of (Liu & Naismith, 2008): initial denaturation at 95°C for 5 minutes, followed by 12 cycles of denaturation at 95°C for 1 minute, annealing at 65°C for 1 minute, and extension at 72°C for 10 minutes. This was succeeded by 3 cycles of annealing at 45°C, 40°C, and 35°C for 1 minute each, and a final extension at 72°C for 30 minutes. Amplicon sizes were verified by staining 5 µL product with GelRed™ at a 1:5 dilution, followed by gel-electrophoresis on a 1%-agarose gel at 130V for 25 min. Bands were visualized using a Bio-Rad ChemiDoc™ UV transilluminator. Subsequently, 20 U of restriction enzyme DpnI was added to each reaction to digest any methylated parental template DNA.

2.3. Creating chemically competent *Escherichia coli* cells

Chemically competent *E. coli* cells were prepared to facilitate various cloning experiments conducted throughout the project. *E. coli* cells of the E. cloni 10G (Lucigen) strain were cultured overnight in 4 mL of Luria-Bertani (LB) broth at 37°C with shaking at a speed of 200 rotations per minute. Subsequently, 1 mL of the overnight culture was transferred to 100 mL of fresh LB broth and grown at 37°C with continuous shaking at 200 rotations per minute until the cell density reached an OD_600 measurement of 0.35. To cool down the cells, they were incubated on ice for 10 minutes. Afterwards, the culture was divided into four ice-cold 50 mL Falcon tubes, each containing 25 mL of the culture. Cells were harvested by centrifugation (4000 x g for 5 min) at a temperature of 6°C. Following centrifugation, the supernatant was carefully discarded, and the cell pellets were resuspended in 2.5 mL of ice-cold 0.1M $CaCl_2$ solution. The resuspended cells were further incubated on ice for 20 minutes. The cells were then again centrifuged at 4000 x g for an additional 5 minutes at 6°C. The supernatant was discarded, and the pellets were resuspended in 1.25 mL of ice-cold 15% glycerol solution supplemented with 0.1M $CaCl_2$. Finally, the cells were dispensed into pre-chilled 1.5 mL Eppendorf tubes, with each tube containing a volume of 50 µL of the prepared cell suspension. Finally, cells were snap-frozen in liquid nitrogen and stored at -70°C.

2.4. Heat-shock transformation of competent *E. coli* cells

A more robust transformation protocol was employed for the efficient transformation of plasmids containing CYP79A1 mutants. In this protocol, 8 µL of DpnI-digested mutant plasmid was mixed with 50 µL of chemically competent E. coli cells and incubated on ice for a duration of 30 minutes. The cells were then subjected to a heat shock in a water bath at 42°C for 1 minute, followed by immediate incubation on ice for 5 minutes. To support cell recovery, 100 µL of Super Optimal Catabolite (SOC) media was added to the transformed cells, which were subsequently allowed to recover at 37°C for 1.5 hours. Finally, 150 µL of the cell mixture was plated on LB-agar plates containing Carbenicillin at a concentration of 100 µg/mL to facilitate the selection of positive transformants.

For the transformation of USER-cloned yeast destination vectors containing the respective desired mutations, the following protocol was used. For this, 7 µL plasmid was mixed with 50 µL chemically competent *E. coli* cells and incubated on ice for 10 min. Cells were subjected to heat shock for 1 min. at 42°C, followed by immediate incubation on ice for 5 minutes. Finally, 50 µL of the cell mixture was plated on LB-agar plates containing Carbenicillin at a concentration of 100 µg/mL to select for positive transformants.

2.5. Sequencing of plasmids

Transformed *E. coli* colonies were gently picked and inoculated in LB broth containing Carbinicillin at a concentration of 100 µg/mL for selection. These were grown overnight at 37°C while shaking at 200 rpm. Plasmids were then purified using an E.Z.N.A.® Plasmid DNA Mini Kit I (omega BIO-TEK), eluding 50 µL 60°C sterilized milliQ water. Samples were then sent for Sanger sequencing using primer 109 or 120 (Appendix 4) (Macrogen Europe; Azenta Life Sciences). Sequencing results were aligned with WT CYP79A1 gene to confirm mutation.

2.6. Generation of plasmid constructs by USER cloning

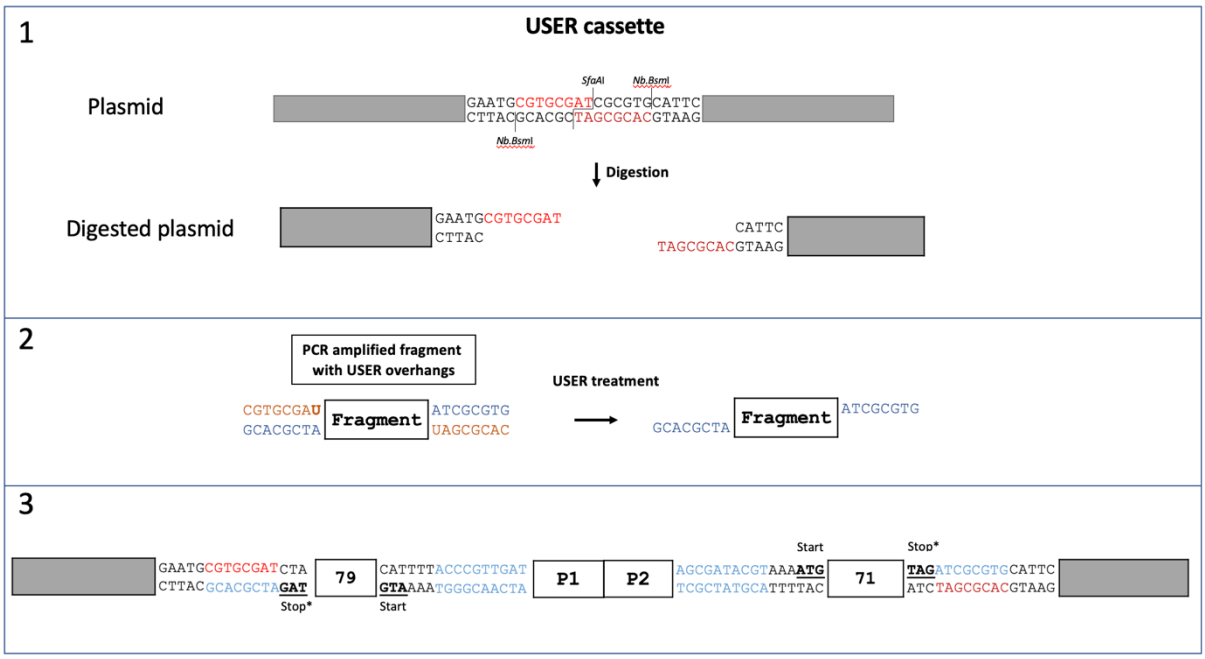


Figure 5 - Workflow of USER cloning. 1.) Restriction digestion and nicking of assembly vector containing USER cassette. 2.) Preparation of USER inserts. 3.) USER fusion and cloning of CYP79A1, P1 (pTDH3), P2 (pSED1), and CYP71E1. Courtesy of Victor Forman, Ph.D.

2.6.1. Restriction digestion and nicking of assembly vector containing USER cassette

Preparation of USER vector was carried out as shown in Figure 5.1 which is done as follows: 5 µg of readily available X-3 Assembler 1 plasmid kindly provided by Victor Forman (Appendix 9) was digested in 50 µL reactions consisting of, 2 µL SfaAI (Thermo Scientific™ 10U/µL), 10x rCutSmart buffer (New England Biolabs™), and sterilized milliQ water. This reaction was incubated at 37°C for 1 hr. 3 µL of the respective digestions were used for gel-electrophoresis to check for adequate linearization of the plasmid backbone. The rest of the samples were cleaned using an E.Z.N.A.® Cycle Pure Kit (omega BIO-TEK), eluding with 60°C sterilized milliQ water. Linearized backbone vector was further nicked by mixing with 10x NEBuffer™ 3.1 (New England Biolabs®), 2 µL Nb.BsmI (New England Biolabs® 10 U/µL), and sterilized milliQ water in 100µL reactions. This mixture was then incubated for 3 hrs. at 65°C,

followed by a clean-up using an E.Z.N.A.® Cycle Pure Kit (omega BIO-TEK), eluding with 50 µL 60°C sterilized milliQ water.

2.6.2. Preparation of inserts

Preparation of inserts were prepared as follows (Figure 5.2): 1 µL of respective mutated CYP79A1 fragments, WT CYP71E1 fragment, and promoter fragment containing the pTDH3 and pSED1 promoters, were PCR amplified using 10x PCR buffer, 5 µL dNTP's (2.5mM), 2.5 µL of the respective USER™ overhang containing primers (Appendix 2), 0.5 µL PfuX7 polymerase, and sterilized milliQ water in a 50 µL touch-down PCR program which is as follows: an initial denaturation of 96°C for 5 min., then 5 times two cycles of 95°C for 30 sec., 68°C for 40 sec., and 72°C for 2 min., where after each of the two cycles, the annealing temperature is decreased by 2°C. Then 24 cycles of 95°C for 30 sec., 58°C for 40 sec, and 72°C for 2 min. Lastly, finishing off with a final extension at 72°C for 10 min. Subsequently, the amplified products were subjected to DNA staining and gel electrophoresis, with 40 µL of each sample loaded into the respective wells. To extract the correctly sized bands, a MicroElute® Gel Extraction Kit (omega BIO-TEK) was utilized, and the elution was performed using 30 µL of sterilized milliQ water at a temperature of 60°C.

2.6.3. USER™ cloning

USER cloning reactions (Figure 5.3) were prepared in reaction volumes of 20 µL, comprising 30 ng of prepared vector backbone, approximately 30 ng of prepared mutated CYP79A1, WT CYP71E1, and promoter fragments. The reactions also included 1 U USER™ enzyme (New England Biolabs®) and 10x rCutSmart buffer (New England Biolabs®). Subsequently, the reactions were incubated at 37°C for 20 minutes, followed by incubation at 25°C for 15 minutes, and finally at 20°C for 5 minutes. USER product was transformed into chemically competent *E. coli* cells as previously described and plated onto LB plates containing 100 µg/mL Carbenicillin. The final X-3 Assembler 1 vector architecture is shown in Appendix 10.

2.7. *E. coli* Colony PCR

Colony PCR was performed on four colonies selected from each transformation. The colonies were delicately picked using a 10 µL pipette tip and transferred to 40 µL of sterile milliQ water, serving as the template for the PCR reaction. PCR reactions were prepared in volumes of 20 µL, comprising 2 µL of 10x PCR buffer, 2 µL of dNTP's (2.5mM), 1 µL each of primers 14 and 120, 1 µL of the template, 0.2 µL of x7 polymerase, and 12.8 µL of sterilized milliQ water. The PCR cycles consisted of an initial denaturation step at 96°C for 5 minutes, followed by 35 cycles of denaturation at 96°C for 30 seconds, annealing at 57°C for 40 seconds, and extension at 72°C for 2 minutes. A final extension step was performed at 72°C for 5 minutes. The resulting amplicons were then verified using DNA staining and

subsequent gel electrophoresis, following the same methodology as previously described. A band at 1373 bp indicated successful cloning.

2.8. Digestion Mapping

To further verify successful insertion of all USER fragments into the yeast integration vector, a digestion map was performed. Two plasmid purified samples from each mutation were subjected to restriction digestion, and a restriction digestion map was generated. The digestions were performed using 20 µL reaction volumes containing 2 µL of 10x rCutSmart buffer, 3 µL of plasmid DNA, 0.2 µL of PstI-HF, 0.2 µL of NdeI, and 14.6 µL of sterilized milliQ water. The digestions were incubated at 37°C for 1 hour. The generated restriction digestion maps were then validated by DNA staining and subsequent gel electrophoresis, following the same procedures as described earlier.

2.9. Transformation of *Saccharomyces cerevisiae*

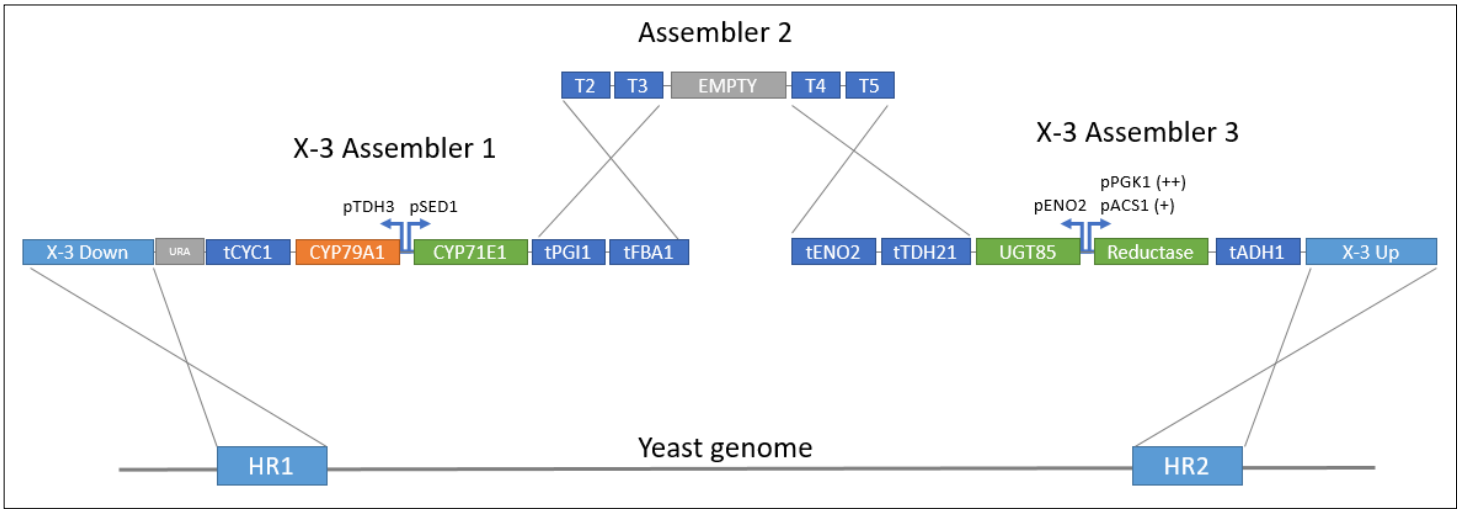


Figure 6: S. cerevisiae assembler system for homologous recombination of DNA. HR1 and -2 are homologous recombination sites in yeast chromosome 10 (courtesy of Victor Forman, Ph.D.)

The expression platform utilized in this study is derived from the work of Victor Forman who designed and generated all plasmids. The integration strategy uses the integration sites described in Mikkelsen et al. (2012). The sequences to be inserted incorporate specific flanking regions that facilitate homologous recombination for the integration into the yeast genome (Figure 6). In particular, the integration site X-3 (Figure 3) corresponds to chromosome 10 of the yeast genome and represents a constitutively transcribed locus within that chromosome.

2.9.1. Linearization of assembler vectors

To prepare the vector DNA for homologous recombination with the yeast chromosomal DNA, it must first be linearized. To do this, 50 µL reactions were made for each mutation, containing 500 ng of X-3 assembler 1, -2, and -3, respectively, 10x rCutSmart buffer, 5 U NotI (New England Biolabs), and sterilized milliQ water. Samples were then incubated for three hours at 37°C.

2.9.2. Transformation of *S. cerevisiae*

S. cerevisiae cells (EUROSCARF, strain Y05210) were first plated onto Yeast Extract Peptone Dextrose (YPD) agar plates and incubated at 28°C for a period of 16 hours. Subsequently, approximately ¼ of a 10 µL inoculation loop containing cells was resuspended in 1 mL of sterile deionized water and subjected to centrifugation at 6000 x g for one minute. The resulting supernatant was discarded, and the cells were then resuspended in 1 mL of 100 mM lithium acetate. Following an incubation at 28°C for 7 minutes, the cells were centrifuged once again at 6000 x g for one minute, and the supernatant was discarded. Subsequently, a sequential addition of the following components took place: 240 µL of polyethylene glycol 3350 50% (w/v), 36 µL of 1M lithium acetate, 50 µL of pre-boiled 2 mg/mL deoxyribonucleic acid from salmon testes (Sigma Aldrich), and 50 µL of linearized DNA insert. The cells were then resuspended in this solution and incubated at 42°C for 60 minutes. Afterward, the cells were subjected to centrifugation at 6000 x g for two minutes, and the supernatant was discarded. The transformed cells were resuspended in 100 µL of sterile deionized water and plated onto yeast nitrogen base (YNB)(VWR®) medium containing 2% glucose, 1% agarose and yeast synthetic Drop-out medium lacking uracil (Y1501; Sigma Aldrich) to select for positive transformants, now containing a gene for production of uracil. Subsequently, the plates were incubated at 28°C for a period of 3 days, allowing the colonies to grow before individual colonies were selected and transferred to fresh plates.

Colony PCR was used to confirm successful homologous recombination of transformed DNA. Four colonies from each mutant strain as well as a control strain known to already contain recombinant WT DNA, as well as a negative control consisting of untransformed yeast colonies. Colonies were gently picked with a pipette tip and suspended in 40 µL 20mM $NaOH$. Suspensions were therefore then subject to alkaline lysis at 96°C for 15 minutes. Un-lysed cells and denatured biomass was then spun down using a table centrifuge for one minute. PCR reactions were prepared in 20 µL reactions consisting of 2x OneTaq Master Mix (New England Biolabs®), 1 µL of primers 903, 904, and 2221 (Appendix 3), and 1 µL of yeast lysate as template. PCR program was as follows: initial denaturation at 95°C for 1 min., then 30 cycles of 95°C for 20 sec., 55°C for 30 sec., and 68°C for 40 sec., followed by a final elongation of 68°C for 5 minutes. Correct amplicon sizes were verified using gel-electrophoresis as previously described, expecting a band at 1059 bp for successful recombination or a band at 1482 bp for unsuccessful recombination.

2.10. Fermentation

Pre-cultures were established by selecting one positive colony for each mutation and transferring it to 1 mL of synthetic complete broth. The pre-cultures were then incubated at 30°C and 250 rpm for 24 hours. Subsequently, fermentations were initiated using 2 mL of synthetic complete broth, and the

pre-culture volumes were added to achieve an initial optical density at 600 nm (OD600) of 0.05. Fermentations were then stopped at time points of 24 and 48 hours. 1500 µL of the fermentation culture was transferred to 1.5 mL Eppendorf tubes and centrifuged at 6000 x g for 2 minutes. Following centrifugation, 200 µL of the supernatant was transferred to new Eppendorf tubes and snap-frozen by submerging them in liquid nitrogen. The remaining supernatant was discarded, and the pellet obtained after centrifugation was also snap-frozen by submerging it in liquid nitrogen. Supernatant and pellet were then stored at -70°C.

2.11. Liquid chromatography-mass spectrometry analysis

2.11.1 Preparation of samples and standards

To prepare samples for liquid chromatography-mass spectrometry (LC-MS) analysis, the supernatant from fermentations were diluted twenty times in water and filter-sterilized as follows: 10 µL of supernatants were diluted with 90 µL sterilized milliQ water and subsequently filtered by centrifugation in a 0.22 µm filter plate at 4000 rpm for 4 min. 70 µL of each sample were transferred to LC-MS vials containing an insert and further diluted by addition of 70 µL filter sterilized milliQ water.

In addition, standards were prepared in different concentrations for calibration purposes. A mixture of L-tyrosine, *p*-hydroxyphenylacetaldoxime, *p*-hydroxyphenylacettaldoxime-glycosylated, *p*-hydroxybenzaldehyde, and dhurrin were prepared in concentrations of 0.005 µM, 0.01 µM, 0.05 µM, 0.1 µM, 0.5 µM, 1 µM, 2.5 µM, 5 µM, 10 µM, and 20 µM.

2.11.2. Triple Quadrupole Mass Spectrometry

Samples were sent for LC-MS analysis at the UCPH DynaMo platform and carried out by Christoph Crocoll, as previously described in Wlodarczyk et al. (2016).

3. Results

The experimental workflow employed in this project can be divided into four key components (Figure 7). The initial phase of the project involved the implementation of the algorithm-suggested single amino acid substitutions in the CYP79A1 gene. To achieve this, a site-directed mutagenesis strategy was employed, which necessitated only a single round of PCR to generate the desired mutant variants. Subsequently, the mutant CYP79A1 genes were fused and cloned using the USER™ method into a destination vector, along with CYP71E1 and their corresponding promoters, constituting the second stage of the project. Once the assembly of the destination vectors was completed, the third stage of the project was initiated, which involved the transformation of *S. cerevisiae* with the destination vectors, along with the genes for the remaining constituents of the dhurrin pathway, namely UGT85B1 and POR2a. Transformants were then fermented and harvested at set time points. Lastly, supernatant from the respective fermentations was sent for metabolomics analysis by LC-MS, to evaluate the effects of the introduced amino acid substitutions in CYP79A1 on the overall flux of L-tyrosine towards *p*-hydroxyphenylacetaldoxime, as well as the downstream products *p*-hydroxybenzaldehyde and dhurrin.

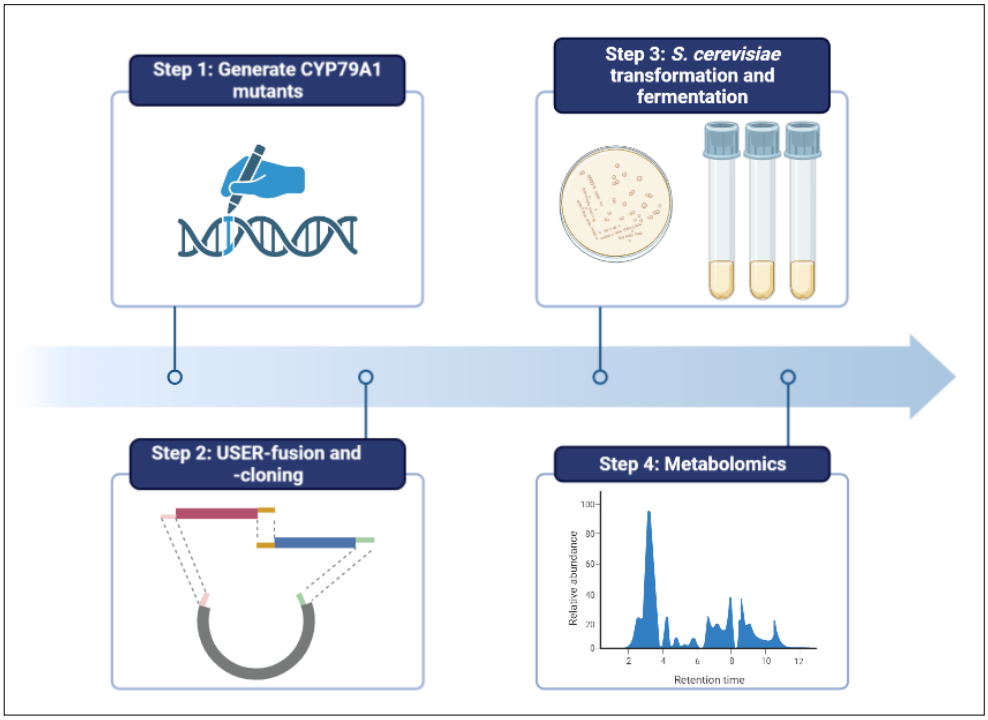


Figure 7: Project workflow

3.1. *In silico* analysis of mutations

Before proceeding with the experimental strategy, it was crucial to ensure that the suggested mutations did not reside within critical substrate recognition sites (SRS) of CYP79A1. To assess this,

the suggested mutations were mapped onto a three-dimensional structure of CYP79A1 using PyMOL software (Figure 8). Specifically, previously identified substrate recognition sites, namely SRS 1 and 6 (Jensen et al., 2011; Vazquez-Albacete et al., 2017), were examined to confirm that they were not targeted for mutation, as the project scope is to evaluate mutations that alter the structural features of the enzyme, which potentially also can optimize the performance. However, it is noteworthy that six of the mutations in this study are in close proximity to the binding pocket of the protein. This observation suggests potential effects on the flexibility, electrostatics and other aspects of the chemical microenvironment surrounding the binding pocket. Furthermore, it was observed that the D281C mutation is situated at a distance of 3.7 Å from Cys267, which is part of the F'-G loop, indicating the possibility of disulfide bridge formation in this specific mutant variant. Additionally, placement of the mutations was assessed with regards to previously determined structures and potential substrate interacting residues (Figure 9). These findings emphasize the importance of assessing the potential structural and functional implications of the identified mutations before proceeding with experimental procedures.

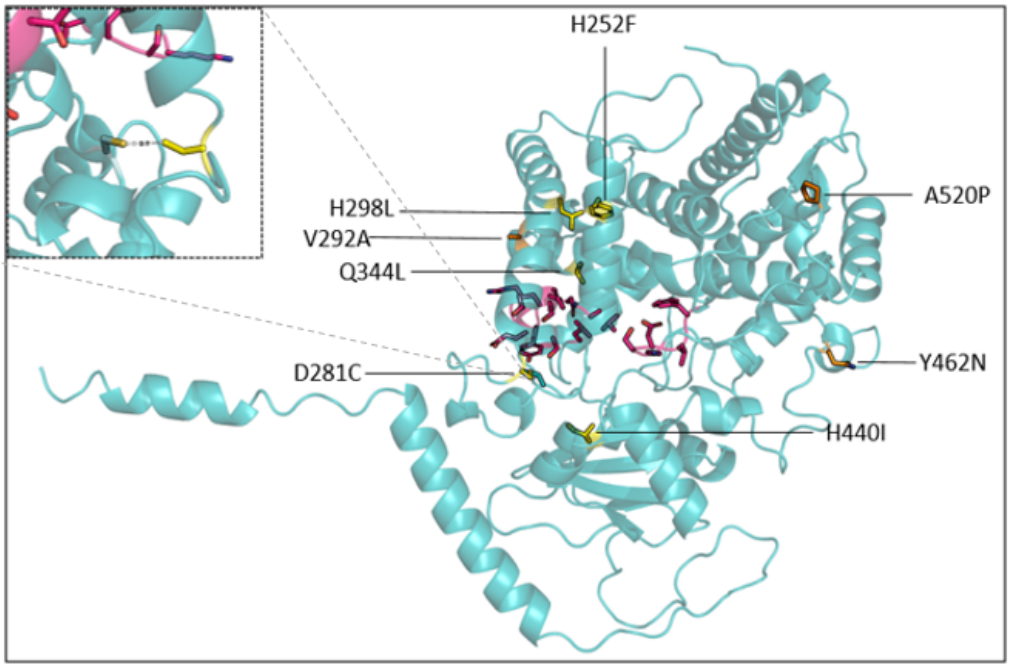


Figure 8 - 3D visualization of CYP79A1 as predicted by AlphaFold. Yellow residues: Cavity model suggested amino acid mutations. Orange residues: Cavity + EVE models suggested mutations. Magenta: Key substrate recognition sites (1 & 6). Top left: close-up of a potential disulfide bridge formation between Cys267 and D281C mutation.

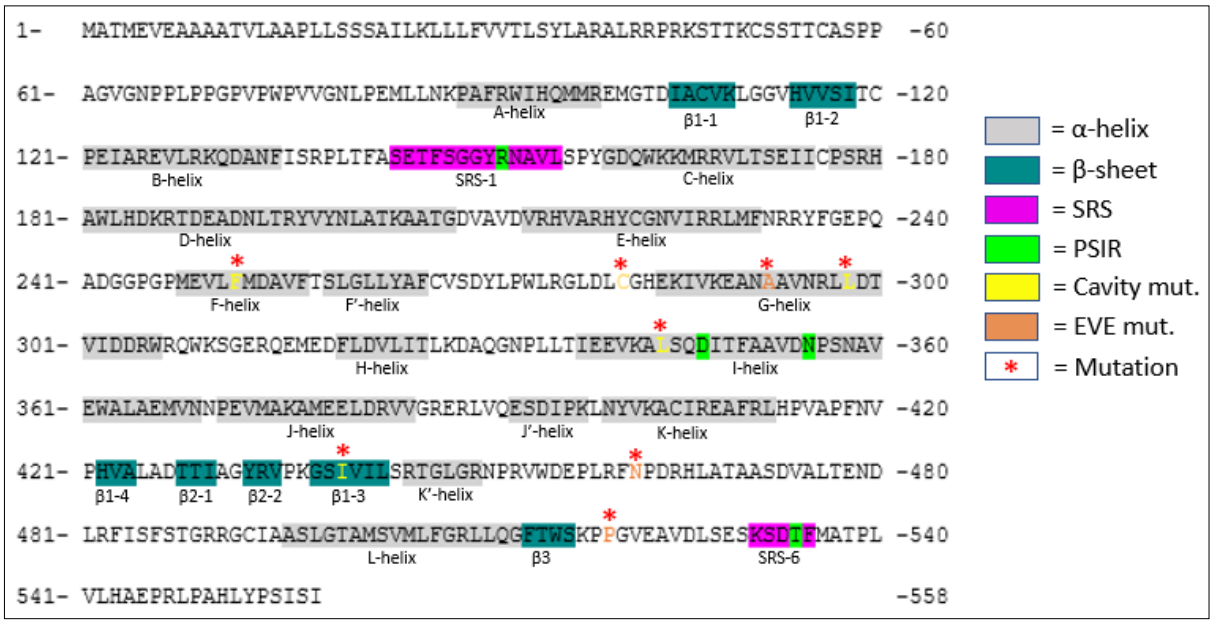


Figure 9 - CYP79A1 amino acid sequence with annotations. α-helices and B-sheets are annotated as in (Jensen et al., 2011). *Substrate recognition sites (SRS) 1 and 6, as well as potential key substrate interacting residues (PSIR) are annotated as in* (Vazquez-Albacete et al., 2017).

3.2. Mutant variants of CYP79A1

The selection of specific single amino acid mutations for experimental investigation was partly guided by the Cavity model (Appendix 6), which provided insight into the changes in change in Gibbs free energy (ΔΔG) associated with each mutation. Among the available options, the five mutations with the most negative ΔΔG values were chosen for further analysis: D281C, Q344L, H298L, H440I, and H256F. These mutations were prioritized based on the significant changes they were predicted to induce in the protein structure and stability.

Additionally, the combination of the Cavity and EVE models (Appendix 7) was employed to identify mutations with low scores (being better) when evaluated using the EVE model. From this analysis, three mutations were selected for experimental investigation: Y462N, V292A, and A520P. These mutations were chosen due to their favorable scores in the EVE model, indicating potential beneficial effects on the protein's fitness.

By combining the insights provided by both the Cavity and EVE models, a set of eight mutations (D281C, Q344L, H298L, H440I, H256F, Y462N, V292A, and A520P) were chosen for further experimental characterization, aiming to elucidate their impact on the structure and fitness of the protein (Table 1)

Amino Acid Substitutions	
Cavity Model	**EVE Model**
D281C	Y462N
Q344L	V292A
H298L	A520P
H440I	
H252F	

Table 1 - Overview of mutations chosen for experimental analysis, and from which model they are suggested by.

To initiate the experimental workflow, mutant variants of CYP79A1 were generated using site-directed mutagenesis. The success of mutagenesis was confirmed by verifying the presence of bands of the expected length, which

in this case was 4651 bp, corresponding to the template pJET1.2 vector containing the CYP79A1 gene (Figure 10). This step ensured that the mutations were successfully introduced into the target gene.

Next, chemically competent *E. coli* cells were transformed with the respective mutated CYP79A1 containing plasmids. The transformed cells were then cultured under appropriate conditions to allow for the amplification of the plasmids and the generation of higher copy numbers. This step aimed to obtain a sufficient quantity of the mutant plasmids for downstream analyses.

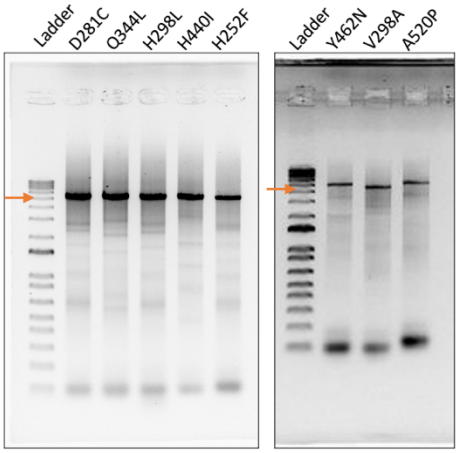


Figure 10 - Visualization of gel-electrophoresis run with PCR samples from site-directed mutagenesis. Orange arrows approximately indicate the expected size of 4651 base-pairs.

To confirm the presence and accuracy of the desired mutations, the plasmids containing the mutant CYP79A1 variants were sent for Sanger sequencing. This sequencing analysis confirmed that the mutagenesis had occurred as intended, validating the generation of the mutant variants (Figure 11). With the successful confirmation of mutagenesis by Sanger sequencing, the experimental workflow could progress to the subsequent steps for further characterization and investigation of the mutant CYP79A1 variants.

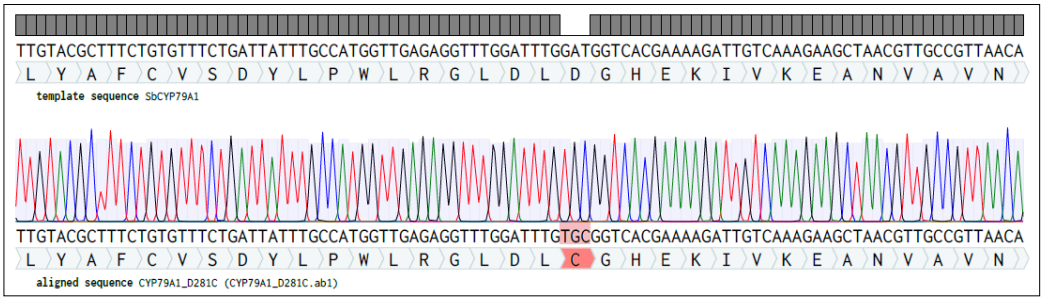


Figure 11 - Exemplary Sanger sequencing read from plasmid containing CYP79A1 D281C aligned with consensus WT CYP79A1. Mutation is highlighted in red, indicating mismatched bases and amino acid with respects to the consensus sequence.

3.3. Assembly of destination vectors

USER™ fusion and cloning was utilized to construct destination vectors (X-3 Assembler 1) that incorporated genes for CYP71E1, promoters for the two CYPs (pSED1 and pTDH3), and the mutated variants of CYP79A1. Once constructs were assembled, they were used to transform *E. coli*, to amplify vector amount. Colonies from the respective *E. coli* strains were then used for colony PCR. This was to select for positive transformants, ensuring the workflow (Figure 12). Primers placed within promoter pTDH3 and CYP79A1, respectively, were used to test if fragments had fused correctly under USER™ treatment, resulting in an amplicon size of 1373 bp. Several colonies from each mutational variant were confirmed to have the correctly sized amplicon.

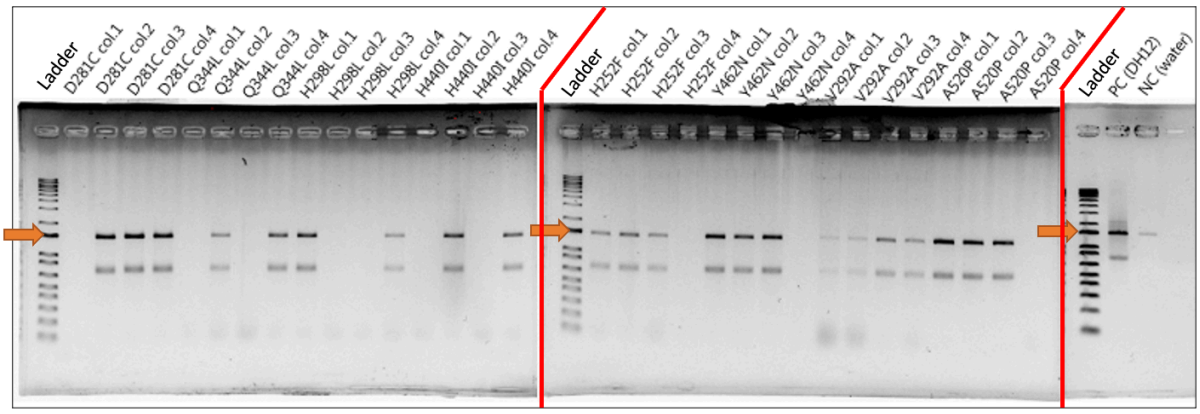


Figure 12 - Visualization of gel-electrophoresis conducted with colony PCR samples. Labels indicate mutation and the E. coli colonies (col) that were picked from each plate and used as template. Expected band size of 1373 bp is indicated by orange arrows at DNA ladders. Individual gels are separated by red lines. Positive control (PC) used readily available DH12 plasmid (Appendix 10), containing the same as the generated plasmids without mutation, as a template. Water was used as template for the negative control (NC).

To further ensure that the vectors contained the correct genes with the correct lengths, a restriction digestion map was designed *in silico* and tested *in vitro* on purified plasmids, derived from colonies that were positive in the colony PCR. Based on the design of the restriction digestion map, eight bands were expected if the vector was assembled correctly. After digestion with PstI and NdeI, the resulting DNA fragments were visualized by gel-electrophoresis (Figure 13). All samples showed the presence of the expected bands, indicating that vectors were assembled correctly and contained genes of proper sizes. Positive and negative controls were included, but later determined to be of too low concentrations for proper visualization.

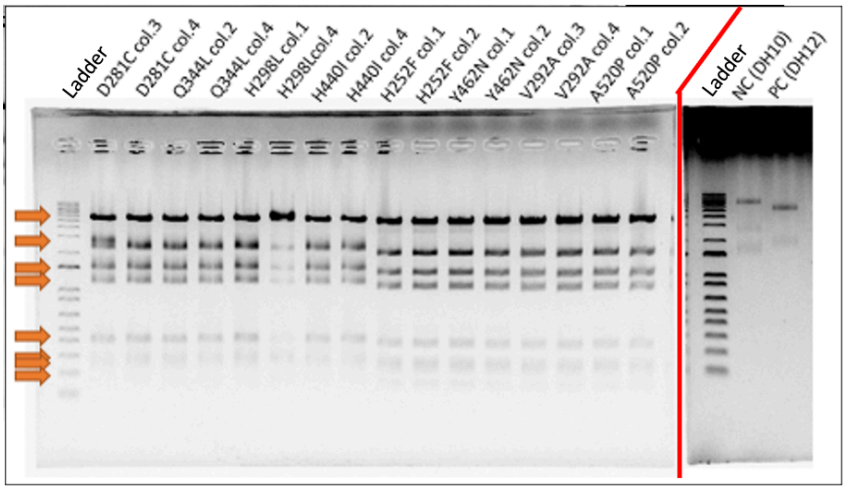


Figure 13 - Restriction digestion map, using PstI and NdeI, of purified plasmids containing the USER-fused mutated CYP79A1's, promoter genes, and CYP71E1. Labels indicate the respective mutations and which E. coli colony was chosen. Individual gels are separated by red line. Orange arrows indicate expected band sizes of appx. 5100, 2100, 1400, 1250, 420, 280, 240, and 150 bp. NC (DH10) is empty vector, and PC (DH12) is non-mutated correctly assembled vector.

3.4. Transformation and fermentation of *S. cerevisiae*

S. cerevisiae was employed as a model organism for metabolomics analysis of the mutated variants of CYP79A1 together with the rest of the dhurrin pathway. Therefore, X-3 assembler vectors were transformed into yeast by heat-shock treatment. To assess proper integration into the yeast genome, a colony PCR was performed. Four colonies from each transformation were examined. The expected band size for successful integration was 1059 bp, and the subsequent gel-analysis confirmed integration of all genes in every colony tested (Figure 14).

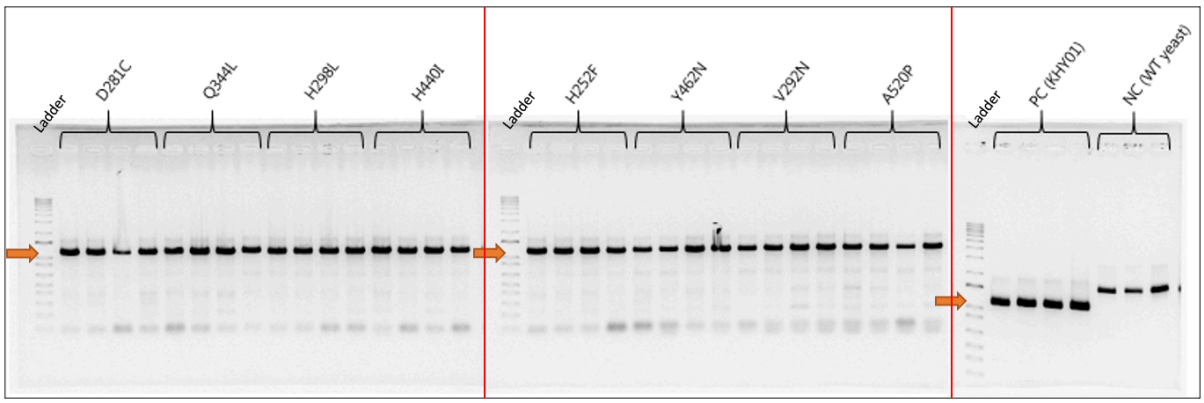


Figure 14 - S. cerevisiae colony PCR. Labels indicate colonies 1-4 from left to right containing the respective mutations. Orange arrows indicate expected size of 1059 bp indicating successful integration of genes into yeast genome. PC (KHY01) is the control containing WT genes for the dhurrin metabolon. NC is WT S. cerevisiae colonies where expected bands at 1382 bp were present.

A fermentation assay was designed such that each strain was subjected to growth for 24 or 48 hours. The assay was designed to investigate potential impact on enzyme stability and overall fitness that the mutations have on CYP79A1. By analyzing the metabolites accumulated in the fermentation broth and assessing the growth by cell density, valuable insights about the performance of the variants can be acquired.

Cell density was measured at 600 nm after the respective time points of fermentation (Figure 15). From this data, it became apparent that growth of the cultures was already reaching a stationary phase after 24 hours, indicating depletion of nutrients in the fermentation broth. Most cultures showed the same growth after 24 hours where $OD_{600} \approx 8$, except for the culture containing CYP79A1 V292A. The same pattern was apparent for fermentations analyzed after 48 hours, where $OD_{600} \approx 8.5$-9, except for V292A again.

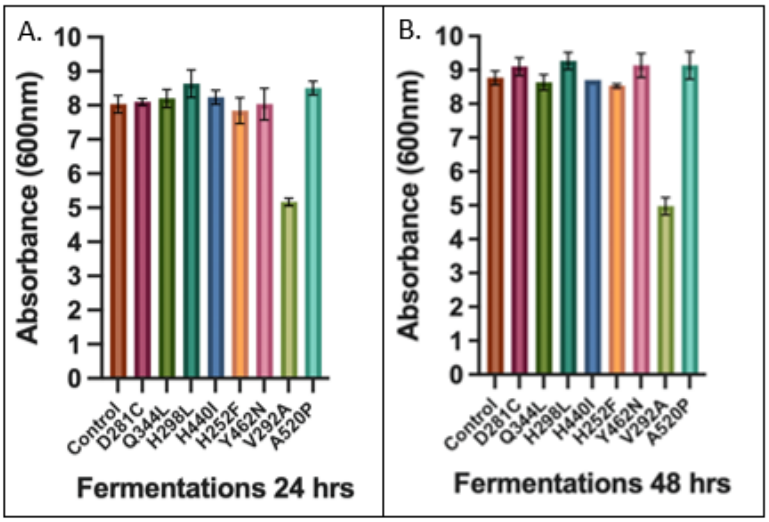


Figure 15 - Cell density measured at 600 nm for A.) Fermentations after 24 hours, and B.) Fermentations after 48 hours. Bars represent mean values of triplicate fermentations ± standard deviation (SD).

Qualitative analysis of the V292A strain demonstrated that it had another phenotype, when compared to all other analyzed strains. Here, yeast cells showed apparent flocculation, meaning that the cells formed larger aggregates in culture (Figure 16). While flocculation is outside the scope of this study, this phenotypic trait is important to note for the following data analysis.

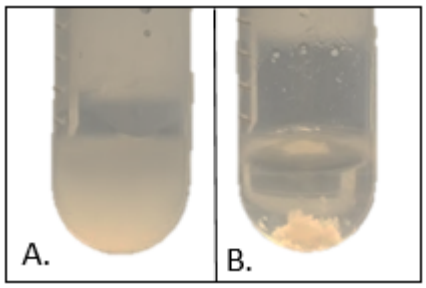


Figure 16 - Fermentations after 48 hours. A.) Fermentation of Y462N strain. B.) Fermentation of V292A strain showing flocculation.

3.5. Metabolomics Analysis by LC-MS

Targeted metabolomics, using triple quadrupole LC-MS was leveraged to quantify the respective accumulated concentrations of L-Tyrosine, oxime, aldehyde, and dhurrin. The data generated through this technique revealed some degree of variance in the efficacy of dhurrin production (Figure 17 C-D). To assess the significance of the differences between the control and the mutant variants, a t-test was conducted. After 24 hours of fermentation, the D281C variant was found to be insignificantly different from the WT control strain. Conversely, the Q344L, H298L, H440I, H252F, Y462, V292A, and A520P strains demonstrated significantly lower levels of dhurrin than the WT counterpart. At the 48-hour timepoint, a slight increase in the concentration of dhurrin was observed for most of the variants, except for the V292A variant, which displayed lower concentrations of dhurrin when compared to the 24-hour timepoint. Notably, this strain was the one showing a flocculation phenotype. Only the Q344L, H298L, H440I, H252F, and V292A strains demonstrated significant differences in dhurrin concentrations when compared to the WT after 48 hours of fermentation. Notably, these strains exhibited a lower concentration of dhurrin than the WT control strain as well.

Measured concentrations of oxime and aldehyde were determined to be of insignificance for further analysis because of their low concentrations. This becomes evident when observing the stacked percentages of each metabolite for each strain (Figure 17 A-B). Here we see that the most part of the metabolites consists of L-Tyrosine, the initial substrate, and dhurrin, the final product. Patterns are consistent for both the 24- and 48-hour timepoints, with very similar distribution of metabolites between the control, D281C, H298L, Y462N, and A520P strains, while there are slightly higher levels of L-Tyrosine in the Q344L, H440I, H252F, and V292A strains.

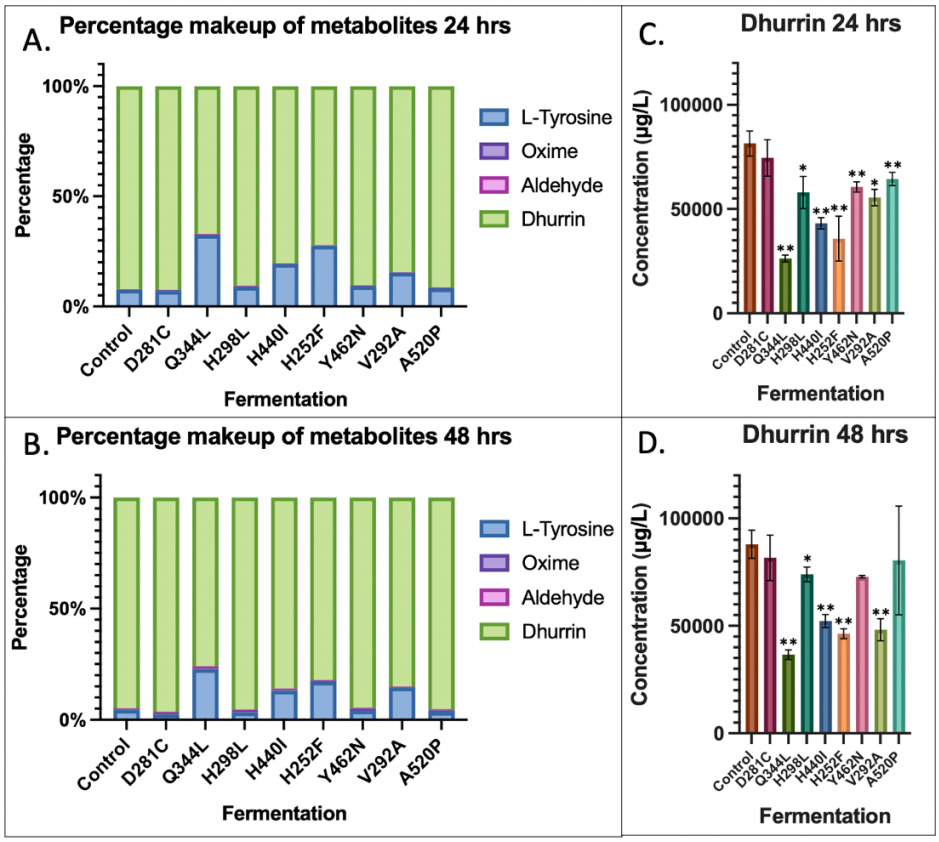


Figure 17 - A & B show the percent makeup of measured metabolites quantified by LC-MS, after 24 and 48 hours, respectively. C & D show the relative concentrations of dhurrin, measured after 24 and 48 hours, respectively. Dhurrin values were considered significantly different from the control, if $p < 0.05$ (, $p < 0.05$; **, $p < 0.01$).*

4. Discussion

4.1. Placement and characteristics of amino acid substitutions

Based on the fermentation results (Figure 17), no amino acid substitutions in CYP79A1 were found to enhance dhurrin production. The mutations suggested by the Cavity model are predicted by only considering their influence on the overall structure of the protein. As mentioned previously, stability of enzymes is often correlational with fitness, and therefore the hypothesis that the mutations would enhance the fitness of the enzyme. However, while enzyme stability can be an indicator of enzyme

fitness, one must also take into consideration that there are other factors that can influence the fitness of an enzyme, especially when situated in a biosynthetic pathway that forms metabolons dynamically.

Previous studies have provided valuable insights into different components of the CYP79A1 protein structure. The Q344L strain exhibited the lowest production of dhurrin across all fermentations, possibly due to the location of the mutation being only three amino acids upstream of a suggested key residue involved in substrate interactions, namely Asp347, which is located on the I-helix of the enzyme (Figure 9) (Vazquez-Albacete et al., 2017). As the helical structure of proteins contains approximately 3.4 amino acids per turn, these two amino acids are in proximity to each other. A closer examination of the 3D AlphaFold model reveals that the side chain of Q344L is in close proximity (4.8 Å) to the hydroxyl group of Asp347, suggesting possible interactions between these two residues as a possible explanation for the reduction in dhurrin production observed in the Q344L strain.

The loop between helices F and G are suggested to have membrane binding functions (Sansen et al., 2007). The D281C substitution, which was the least affected mutation (Figure 17 C-D), is situated in this region of CYP79A1. Hydrophobic and non-polar residues make out the majority of this site, supporting the theory of membrane embedding of this segment. Thus, a substitution from a negatively charged aspartic acid, to a non-polar cysteine could be favorable in this segment. As shown in Figure 8, the D281C mutation is within very close proximity to Cys267, which notably could result in the formation of a disulfide-bridge, which also could help stabilize the protein. However, while the experimental design of this project was not directly designed to evaluate protein stability, it is notable that the best performing mutant variant is the one that most apparently could affect tertiary structure of CYP79A1. Mutation H252F is situated in the F-helix and mutations V292A and H298L are placed within the G-helix. These helices have been proposed to form an opening near the membrane surface, allowing for entrance of substrate (Jensen et al., 2011). Additionally, this area of cytochrome P450 enzymes has also been defined as containing SRS-2 and -3 (Nair et al., 2016). The mutations in these segments might therefore have a negative impact on enzyme function.

Amino acid substitution Y462N was found to be near the positively charged POR interacting residues as also described in Jensen et al. (2011), which are essential for protein-protein interaction with the POR. None of the mutations were found to potentially interfere with heme anchoring. While this is all speculative, it is important to acknowledge and highlight the potential impact that even one amino acid residue can have on enzyme structure, function, and interactions.

4.2. LC-MS results and fermentation strategy

According to the LC-MS results, it is apparent that optimization of the fermentation strategy is possible. The metabolite composition after 24 and 48 hours of fermentation denoted that the fermentations were analyzed at plateau points, with respects to the biosynthesis of dhurrin (Figure 17 A-B). No significant amounts of pathway intermediates were detected, which makes it difficult to make definitive statements about the specific role of the CYP79A1 mutations relative to the entire pathway. L-tyrosine from the synthetic complete media, which constituted the fermentation broth, was mostly depleted in all sample strains (Figure 17. A-B). Due to the dynamic nature of enzymatic reactions, the early stages of metabolism allow for a better understanding of the kinetic parameters of enzymes (Figure 18). Therefore, in retrospect, it would have been beneficial to sample fermentations at earlier time points to facilitate metabolomics analysis. The early sampling would have potentially provided more insightful data of any plausible changes in the efficacy of CYP79A1 that mutations would have caused. Another potential approach to extend the linear phase of catalysis and conduct a stress-testing experiment on the mutated CYP79A1 is through the utilization of fed-batch fermentations. Specifically, L-tyrosine and the nutrients necessary for yeast growth could be consistently supplied at predetermined time points. Such a strategy allows for a prolonged linear phase of catalysis and could elucidate upon the stress tolerance in the mutated strains of CYP79A1.

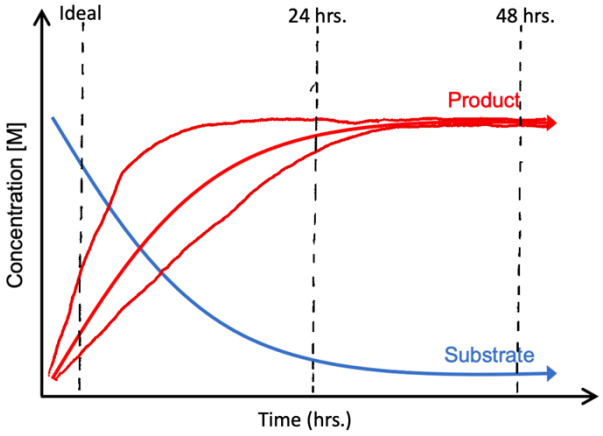


Figure 18 - Graphed representation of the relationship between substrate and product over time. If fermentations are sampled at 24 and 42 hours, not much of a difference would be measured. Ideally, one would sample fermentations in the fast-growing phase.

Furthermore, the biosynthesis of dhurrin has been proven an already highly efficient biological system. The yeast control strain expressing the WT dhurrin pathway had already converted almost all L-Tyrosine to dhurrin, making it impossible to determine if the mutant variants that were not significantly different from the control were performing better. Furthermore, studies have shown that CYP79A1 constitutes only 0.2% of the total amount of microsomal protein in etiolated *S. bicolor* seedlings, while dhurrin makes up to 30% of the seedling tip's dry weight (Sibbesen et al., 1994). This observation of striking performance by the enzyme raises intriguing questions regarding the feasibility of optimizing an enzyme with efficiency of this character.

4.3. Machine learning model characteristics

The machine learning Cavity and EVE models have different characteristics and therefore also limitations. As mentioned earlier, the Cavity model is trained on more than 16,000 random AlphaFold predicted 3D protein models, learning the structural patterns apparent in these. Evolutionarily, enzymes have been pressured to find an optimum between structure and performance, however there is often a trade-off between these, meaning that the increase in stability might lead to decrease of fitness (Hou et al., 2023). With respects to the production of dhurrin, most of the mutations suggested by the Cavity model performed significantly worse than the control strain, potentially demonstrating this trade-off (Figure 17 C-D).

Mutations predicted in combination of the EVE and Cavity models, served as a negation strategy for this trade-off. By selecting mutations that were both predicted to have a strengthening effect on structure as well as fitness, one could expect better results than the WT pathway. The EVE model is trained on evolutionary patterns in protein sequences, thus predicting mutations that should be more specific towards protein fitness. Mutant CYP79A1 strains selected from this pool showed somewhat better production of dhurrin than the mutants suggested by the Cavity model alone. The Y462N and A520P variants were not significantly different from the control in fermentations sampled after 48 hours, meaning that 66% of the EVE mutant variants did not perform worse.

5. Conclusion and Perspectives

The objective of this study was to employ machine learning models, specifically Cavity and EVE, as a computational tool for optimizing the stability and performance of a cytochrome P450 enzyme, CYP79A1. This enzyme plays a pivotal role in catalyzing the committed step of the biosynthetic pathway responsible for the production of the cyanogenic glucoside dhurrin. To assess the effect of amino acid substitutions, eight distinct constructs were generated and subsequently expressed in *Saccharomyces cerevisiae*, along with the other components of the dhurrin pathway. Evaluating the catalytic efficiency of the mutant variants employed fermentations conducted for 24 and 48 hours, followed by targeted metabolomics analysis. The metabolomic profiles revealed that only the D281C, Y462N, and A520P variants did not exhibit a decreased production of dhurrin compared to the wild-type pathway. Additionally, the metabolomics analysis showed that almost all osf the initial substrate for the pathway had been metabolized within 24 hours, implying the necessity for optimizing the fermentation strategy to ascertain the kinetic parameters of the enzyme variants exhibiting unimpaired dhurrin production.

To further elucidate the effects on the variants exhibiting unimpaired production of dhurrin as well as the remaining enzyme variants exhibiting reduced production of dhurrin, additional experimental strategies could be employed. For instance, the thermostability of the mutant CYP79A1 variants could be assessed by examining their melting temperatures using differential scanning calorimetry as done in (Lu et al., 2022). Monitoring protein quantities through techniques such as Western Blotting could provide insights into potential differences in protein degradation among the variants. Furthermore, conducting fermentation experiments solely expressing the CYP79A1 variants could offer valuable information regarding specifically the engineered enzyme. Lastly, considering the previously mentioned benefits of L-tyrosine feeding during fermentations, incorporating this strategy may provide valuable insights into the kinetic parameters of the enzyme variants that did not exhibit compromised dhurrin production. In conclusion, further research is needed to determine the precise effects of using these exciting machine learning models to optimize enzymes and metabolic pathways.

Literature

D'Addabbo, T., Laquale, S., Lovelli, S., Candido, V., & Avato, P. (2014). Biocide plants as a sustainable tool for the control of pests and pathogens in vegetable cropping systems. *Italian Journal of Agronomy*, *9*(4), 137–145. https://doi.org/10.4081/ija.2014.616

D'Amelia, V., Docimo, T., Crocoll, C., & Rigano, M. M. (2021). Specialized metabolites and valuable molecules in crop and medicinal plants: The evolution of their use and strategies for their production. *Genes*, *12*(6). https://doi.org/10.3390/genes12060936

Delgoda, R., & Murray, J. E. (2017). Evolutionary Perspectives on the Role of Plant Secondary Metabolites. I *Pharmacognosy: Fundamentals, Applications and Strategy*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-802104-0.00007-X

Dixon, R. A., & Strack, D. (2003). Phytochemistry meets genome analysis, and beyond..... *Phytochemistry*, *62*(6), 815–816. https://doi.org/10.1016/S0031-9422(02)00712-4

Fei, B., Xu, H., Cao, Y., Ma, S., Guo, H., Song, T., Qiao, D., & Cao, Y. (2013). A multi-factors rational design strategy for enhancing the thermostability of Escherichia coli AppA phytase. *Journal of Industrial Microbiology and Biotechnology*, *40*(5), 457–464. https://doi.org/10.1007/s10295-013-1260-z

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., & Marks, D. S. (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature*, *599*(7883), 91–95. https://doi.org/10.1038/s41586-021-04043-8

Hansen, C. C., Nelson, D. R., Møller, B. L., & Werck-Reichhart, D. (2021). Plant cytochrome P450 plasticity and evolution. *Molecular Plant*, *14*(8), 1244–1265. https://doi.org/10.1016/j.molp.2021.06.028

Hou, Q., Rooman, M., & Pucci, F. (2023). Enzyme Stability-Activity Trade-Off: New Insights from Protein Stability Weaknesses and Evolutionary Conservation. *Journal of Chemical Theory and Computation*. https://doi.org/10.1021/acs.jctc.3c00036

Jensen, K., & Møller, B. L. (2010). Plant NADPH-cytochrome P450 oxidoreductases. *Phytochemistry*, *71*(2–3), 132–141. https://doi.org/10.1016/j.phytochem.2009.10.017

Jensen, K., Osmani, S. A., Hamann, T., Naur, P., & Møller, B. L. (2011). Homology modeling of the three membrane proteins of the dhurrin metabolon: Catalytic sites, membrane surface association and protein-protein interactions. *Phytochemistry*, *72*(17), 2113–2123. https://doi.org/10.1016/j.phytochem.2011.05.001

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, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kahn, R. A., Fahrendorf, T., Halkier, B. A., & Møller, B. L. (1999). Substrate specificity of the cytochrome P450 enzymes CYP79A1 and CYP71E1 involved in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. *Archives of Biochemistry and Biophysics*, *363*(1), 9–18. https://doi.org/10.1006/abbi.1998.1068

Kim, Y. M., Shimizu, R., Nakai, H., Mori, H., Okuyama, M., Kang, M. S., Fujimoto, Z., Funane, K., Kim, D., & Kimura, A. (2011). Truncation of N- and C-terminal regions of Streptococcus mutans

dextranase enhances catalytic activity. *Applied Microbiology and Biotechnology*, *91*(2), 329–339. https://doi.org/10.1007/s00253-011-3201-y

Kotopka, B. J., & Smolke, C. D. (2019). Production of the cyanogenic glycoside dhurrin in yeast. *Metabolic Engineering Communications*, *9*(April), e00092. https://doi.org/10.1016/j.mec.2019.e00092

Lattanzio, V. (2013). Phenolic Compounds: Introduction. I *Natural Products: Phytochemistry, Botany and Metabolism of Alkaloids, Phenolics and Terpenes* (s. 1543–1580). https://doi.org/10.1007/978-3-642-22144-6_57

Laursen, T., Borch, J., Knudsen, C., Bavishi, K., Torta, F., Martens, H. J., Silvestro, D., Hatzakis, N. S., Wenk, M. R., Dafforn, T. R., Olsen, C. E., Motawia, M. S., Hamberger, B., Møller, B. L., & Bassard, J. E. (2016). Characterization of a dynamic metabolon producing the defense compound dhurrin in sorghum. *Science*, *354*(6314), 890–893. https://doi.org/10.1126/science.aag2347

Liu, H., & Naismith, J. H. (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

Lu, H., Diaz, D. J., Czarnecki, N. J., Zhu, C., Kim, W., Shroff, R., Acosta, D. J., Alexander, B. R., Cole, H. O., Zhang, Y., Lynd, N. A., Ellington, A. D., & Alper, H. S. (2022). Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature*, *604*(7907), 662–667. https://doi.org/10.1038/s41586-022-04599-z

Mikkelsen, M. D., Buron, L. D., Salomonsen, B., Olsen, C. E., Hansen, B. G., Mortensen, U. H., & Halkier, B. A. (2012). Microbial production of indolylglucosinolate through engineering of a multi-gene pathway in a versatile yeast expression platform. *Metabolic Engineering*, *14*(2), 104–111. https://doi.org/10.1016/j.ymben.2012.01.006

Møller, B. L. (2010). Dynamic Metabolons. *Science Magazine*, *330*(December), 1328–13. https://doi.org/10.1126/science.1194971

Nagegowda, D. A., & Gupta, P. (2020). Advances in biosynthesis, regulation, and metabolic engineering of plant specialized terpenoids. *Plant Science*, *294*(February), 110457. https://doi.org/10.1016/j.plantsci.2020.110457

Nair, P. C., McKinnon, R. A., & Miners, J. O. (2016). Cytochrome P450 structure–function: insights from molecular dynamics simulations. *Drug Metabolism Reviews*, *48*(3), 434–452. https://doi.org/10.1080/03602532.2016.1178771

Newman, D. J., & Cragg, G. M. (2020). Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products*, *83*(3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285

Nørholm, M. H. H. (2010). A mutant Pfu DNA polymerase designed for advanced uracil-excision DNA engineering. *BMC Biotechnology*, *10*. https://doi.org/10.1186/1472-6750-10-21

Paik, I., Ngo, P. H. T., Shroff, R., Diaz, D. J., Maranhao, A. C., Walker, D. J. F., Bhadra, S., & Ellington, A. D. (2023). Improved Bst DNA Polymerase Variants Derived via a Machine Learning Approach. *Biochemistry*, *62*(2), 410–418. https://doi.org/10.1021/acs.biochem.1c00451

Palackal, N. (2004). An evolutionary route to xylanase process fitness. *Protein Science*, *13*(2), 494–503. https://doi.org/10.1110/ps.03333504

Rui, L., Cao, L., Chen, W., Reardon, K. F., & Wood, T. K. (2004). Active site engineering of the epoxide

hydrolase from Agrobacterium radiobacter AD1 to enhance aerobic mineralization of cis-1,2-dichloroethylene in cells expressing an evolved toluene ortho-monooxygenase. *Journal of Biological Chemistry*, *279*(45), 46810–46817. https://doi.org/10.1074/jbc.M407466200

Sansen, S., Yano, J. K., Reynald, R. L., Schoch, G. A., Griffin, K. J., Stout, C. D., & Johnson, E. F. (2007). Adaptations for the oxidation of polycyclic aromatic hydrocarbons exhibited by the structure of human P450 1A2. *Journal of Biological Chemistry*, *282*(19), 14348–14355. https://doi.org/10.1074/jbc.M611692200

Shao, Z., Zhao, H., & Zhao, H. (2009). DNA assembler, an in vivo genetic method for rapid construction of biochemical pathways. *Nucleic Acids Research*, *37*(2), 1–10. https://doi.org/10.1093/nar/gkn991

Sharma, A., Gupta, G., Ahmad, T., Mansoor, S., & Kaur, B. (2021). Enzyme Engineering: Current Trends and Future Perspectives. *Food Reviews International*, *37*(2), 121–154. https://doi.org/10.1080/87559129.2019.1695835

Shroff, R., Cole, A. W., Diaz, D. J., Morrow, B. R., Donnell, I., Annapareddy, A., Gollihar, J., Ellington, A. D., & Thyer, R. (2020). Discovery of novel gain-of-function mutations guided by structure-based deep learning. *ACS Synthetic Biology*, *9*(11), 2927–2935. https://doi.org/10.1021/acssynbio.0c00345

Sibbesen, O., Koch, B., Halkier, B. A., & Moller, B. L. (1994). Isolation of the heme-thiolate enzyme cytochrome P-450(TYR), which catalyzes the committed step in the biosynthesis of the cyanogenic glucoside dhurrin in Sorghum bicolor (L.) Moench. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(21), 9740–9744. https://doi.org/10.1073/pnas.91.21.9740

Taiz, L., Zeiger, E., Møller, I. M., & Murphy, A. (2015). *Plant Physiology and Development* (Sixth Edit). Sinauer Associates, Inc.

Turi, C. E., Finley, J., Shipley, P. R., Murch, S. J., & Brown, P. N. (2015). Metabolomics for phytochemical discovery: Development of statistical approaches using a cranberry model system. *Journal of Natural Products*, *78*(4), 953–966. https://doi.org/10.1021/np500667z

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Vazquez-Albacete, D., Montefiori, M., Kol, S., Motawia, M. S., Møller, B. L., Olsen, L., & Nørholm, M. H. H. (2017). The CYP79A1 catalyzed conversion of tyrosine to (E)-p-hydroxyphenylacetaldoxime unravelled using an improved method for homology modeling. *Phytochemistry*, *135*, 8–17. https://doi.org/10.1016/j.phytochem.2016.11.013

Wlodarczyk, A., Gnanasekaran, T., Nielsen, A. Z., Zulu, N. N., Mellor, S. B., Luckner, M., Thøfner, J. F. B., Olsen, C. E., Mottawie, M. S., Burow, M., Pribil, M., Feussner, I., Møller, B. L., & Jensen, P. E. (2016). Metabolic engineering of light-driven cytochrome P450 dependent pathways into Synechocystis sp. PCC 6803. *Metabolic Engineering*, *33*, 1–11. https://doi.org/10.1016/j.ymben.2015.10.009

Yokoyama, K., Utsumi, H., Nakamura, T., Ogaya, D., Shimba, N., Suzuki, E., & Taguchi, S. (2010). Screening for improved activity of a transglutaminase from Streptomyces mobaraensis created by a novel rational mutagenesis and random mutagenesis. *Applied Microbiology and Biotechnology*, *87*(6), 2087–2096. https://doi.org/10.1007/s00253-010-2656-6

Appendix

Primer#	Name	Primer Location	Direction	Sequence 5'-3'	nt	Tm	Comment
1	CYP79A1_CAVITY_D281C_Fw	CYP79A1	Forward	GGATTTGTGCGGTCACGAAAAGATTGTCAAAGAAGC	36	65	D281C
2	CYP79A1_CAVITY_D281C_Rv	CYP79A1	Reverse	GTGACCGCACAAATCCAAACCTCTCAACCATGG	33	66	D281C
3	CYP79A1_CAVITY_Q344L_Fw	CYP79A1	Forward	GGCCCTATCTCAAGATATTACTTTTGCTGCCGTTG	35	63,8	Q344L
4	CYP79A1_CAVITY_Q344L_Rv	CYP79A1	Reverse	CTTGAGATAGGGCCTTAACTTCTTCGATGGTCAAC	35	63	Q344L
5	CYP79A1_CAVITY_H298L_Fw	CYP79A1	Forward	GATTACTCGATACCGTTATTGATGATAGATGGAGACAATGGAAGAGTGG	49	65,3	H298L
6	CYP79A1_CAVITY_H298L_Rv	CYP79A1	Reverse	CGGTATCGAGTAATCTGTTAACGGCAACGTTAGCTTC	37	64,7	H298L
7	CYP79A1_CAVITY_H440I_Fw	CYP79A1	Forward	GGGTTCCATTGTTATTTTGTCTAGAACTGGTTTGGGTAG	39	63,1	H440I
8	CYP79A1_CAVITY_H440I_Rv	CYP79A1	Reverse	CAAAATAACAACGGAACCCTTTGGAACTCTATAACCAGC	39	63,6	H440I
9	CYP79A1_CAVITY_H252F_Fw	CYP79A1	Forward	GTTTTATTCATGGATGCTGTTTTCACCTCTTTGGGTTTG	39	63,2	H252F
10	CYP79A1_CAVITY_H252F_Rv	CYP79A1	Reverse	GCATCCATGAATAAAACTTCCATTGGACCTGGACCACCATCAG	43	67,4	H252F
11	CYP79A1_EVE+CAVITY_Y462N_Fw	CYP79A1	Forward	GAGATTTAACCCAGATAGACATTTGGCTACTGCTGC	36	63,6	Y462N
12	CYP79A1_EVE+CAVITY_Y462N_Rv	CYP79A1	Reverse	TCTGGGTTAAATCTCAATGGTTCATCCCAAACTCTTGG	38	64	Y462N
13	CYP79A1_EVE+CAVITY_V292A_Fw	CYP79A1	Forward	GCTAACGCTGCCGTTAACAGATTACACGATACCGTTATTG	40	65,9	V292A
14	CYP79A1_EVE+CAVITY_V292A_Rv	CYP79A1	Reverse	ACGGCAGCGTTAGCTTCTTTGACAATCTTTTCGTG	35	65,8	V292A
15	CYP79A1_EVE+CAVITY_A520P_Fw	CYP79A1	Forward	CTAAACCACCCGGTGTTGAAGCAGTTGATTTGTCT	35	65,1	A520P
16	CYP79A1_EVE+CAVITY_A520P_Rv	CYP79A1	Reverse	ACACCGGGTGGTTTAGACCAAGTGAAACCTTGCAATAATCTAC	43	67	A520P

Appendix 1 - Primers used for site directed mutagenesis. Nucleotides encoding codon for substitutions are in red.

Primer#	Name	Primer Location	Direction	Sequence 5'-3'	nt	Tm	Comment
17	USER_CYP79A1_Fw	CYP79A1	Forward	ATCAACGGGUAAAACAATGGCTACCATGGAAGTTG	35	60,6	USER
18	USER_CYP79A1_Rv	CYP79A1	Reverse	CGTGCGAUTCAGATAGAAATAGATGGGTACAAG	33	62,5	USER
19	USER_pTDH3_p1_Rv	pTDH3	Reverse	ACCCGTTGAUGTTTGTTTATGTGTGTTTATTCGAAACTAAGTTC	44	63,4	USER
20	USER_pSED1_p2_Rv	pSED1	Reverse	ACGTATCGCUCTTAATAGAGCGAACGTATTTTATTTTGCTTG	42	63,2	USER
21	USER_CYP71E1_Fw	CYP71E1	Forward	AGCGATACGUAAAACAATGGCTACTACTGCTACTCC	36	61,8	USER
22	USER_CYP71E1_Rv	CYP71E1	Reverse	CACGCGAUCTATTAAGCAGCTCTTCTATTCTTGTAC	36	61,9	USER

Appendix 2 - Primers used for generating gene fragments containing USER tails. USER tails are marked in red. Green nucleotides are Kozak sequences.

Primer#	Name	Primer Location	Direction	Sequence 5'-3'	nt	Tm	Comment
23	ColonyPCR_pTDH3_Fw	pTDH3	Forward	TGAAATTATTCCCCTACTTGAC	22	49,7	E.coli colony
24	ColonyPCR_CYP79A1_Rv	CYP79A1	Reverse	TTTGGAATATCGGATTCTTGG	21	50	E.coli colony
903	ColonyPCR_yeast_X3_1	Yeast chrom X.3	Forward	TGACGAATCGTTAGGCACAG	20	54,9	yeast colony
904	ColonyPCR_yeast_X3_2	Yeast chrom X.3	Reverse	CCGTGCAATACCAAAATCG	19	47,4	yeast colony
2221	ColonyPCR_yeast_insert	Insert	Reverse	GTTGACACTTCTAAATAAGCGAATTTC	27	53,1	yeast colony

Appendix 3 - Primers used for E. coli colony PCR and S. cerevisiae colony PCR

Primer#	Name	Primer Location	Direction	Sequence 5'-3'	nt	Tm	Comment
109	CYP79A1_pJET_Rev	CYP79A1	Reverse	TCAGATAGAAATAGATGGGTACAAG	25	57	Sequencing
120	CYP79A1_Rv_Seq	CYP79A1	Reverse	TTTGGAATATCGGATTCTTGG	21	52	Sequencing

Appendix 4 - Primers used for sequencing. Primer 109 was used for sequencing mutations close to the C-terminal end of the gene. Primer 120 was used for sequencing of genes in the middle and closer to the N-terminal of the gene.

10x PCR buffer:
0.2M TRIS HCl pH8.8
0.1M KCl
0.06M (NH4)2SO4
0.02M MgSO4
1 mg/mL BSA
1% Triton x 100

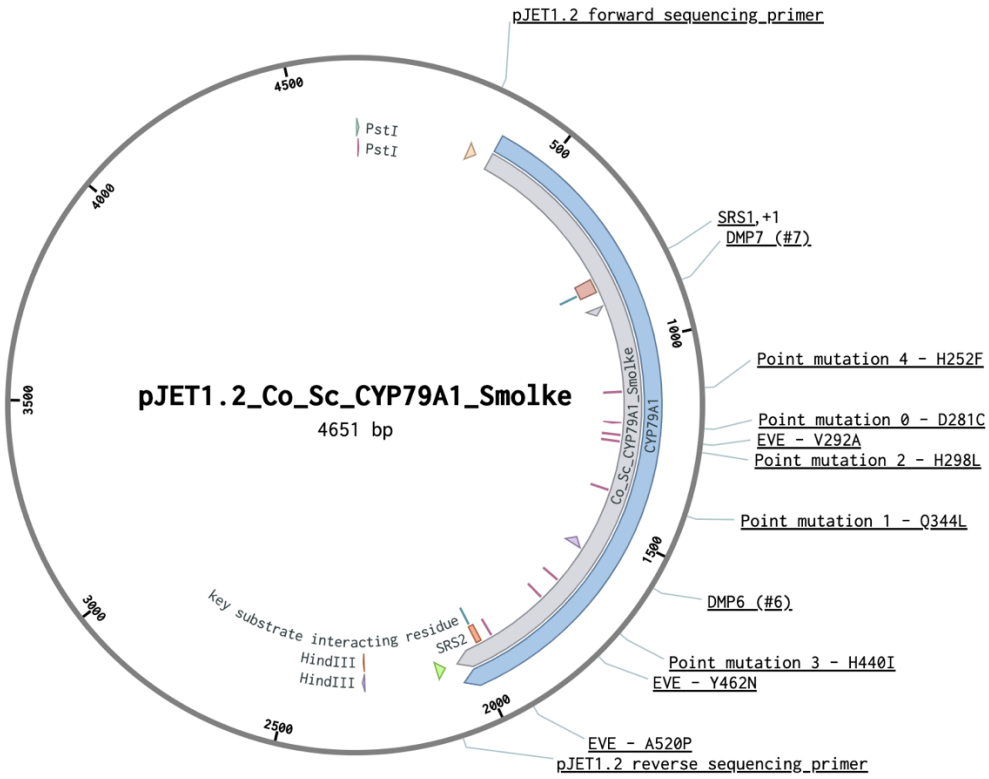
Appendix 5 - 10x PCR buffer composition

	variant	prediction_ddG_cavity_model	protein
0	D281C	-12.552163	Q43135
1	Q344L	-7.097041	Q43135
2	H298L	-4.858373	Q43135
3	H440I	-4.581374	Q43135
4	H252F	-4.486084	Q43135
5	I175F	-4.391475	Q43135
6	H114L	-4.140653	Q43135
7	Q346I	-4.111695	Q43135
8	Y462D	-4.088318	Q43135
9	I286R	-4.046594	Q43135
10	C120D	-4.043237	Q43135
11	N479P	-3.983252	Q43135
12	M502I	-3.866256	Q43135
13	W182E	-3.798245	Q43135
14	G243C	-3.536970	Q43135
15	Q512L	-3.445075	Q43135
16	H283I	-3.436795	Q43135
17	D189E	-3.385868	Q43135
18	T349F	-3.380381	Q43135
19	H180L	-3.120560	Q43135

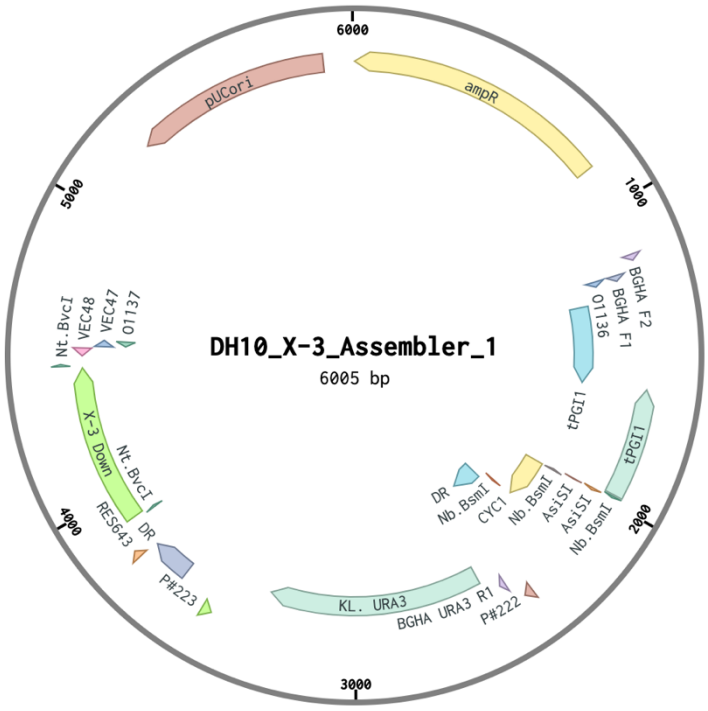
Appendix 6 - Suggested mutations predicted by the Cavity model

protein_name	mutation	prediction_ddG_cavity_model	fitness_score_eve
CYP79A1_Q43135	Y462D	-4.088318	-2.630432
CYP79A1_Q43135	D189E	-3.385868	-2.140930
CYP79A1_Q43135	N134I	-2.434734	-1.703918
CYP79A1_Q43135	V292A	-2.352078	-3.032227
CYP79A1_Q43135	A520P	-2.321203	-2.540100
CYP79A1_Q43135	N134V	-2.196626	-2.519775
CYP79A1_Q43135	Y462N	-2.077211	-4.470947
CYP79A1_Q43135	M248E	-1.978087	-1.516296
CYP79A1_Q43135	D304E	-1.933469	-2.337585
CYP79A1_Q43135	V292K	-1.703001	-3.891907
CYP79A1_Q43135	M374L	-1.637563	-1.375671
CYP79A1_Q43135	N333K	-1.572028	-1.552979
CYP79A1_Q43135	Q315K	-1.533587	-1.571899
CYP79A1_Q43135	V292Q	-1.467377	-1.600647
CYP79A1_Q43135	M317E	-1.347008	-2.088013
CYP79A1_Q43135	C176S	-1.335922	-1.286987
CYP79A1_Q43135	M317Q	-1.255329	-1.397156

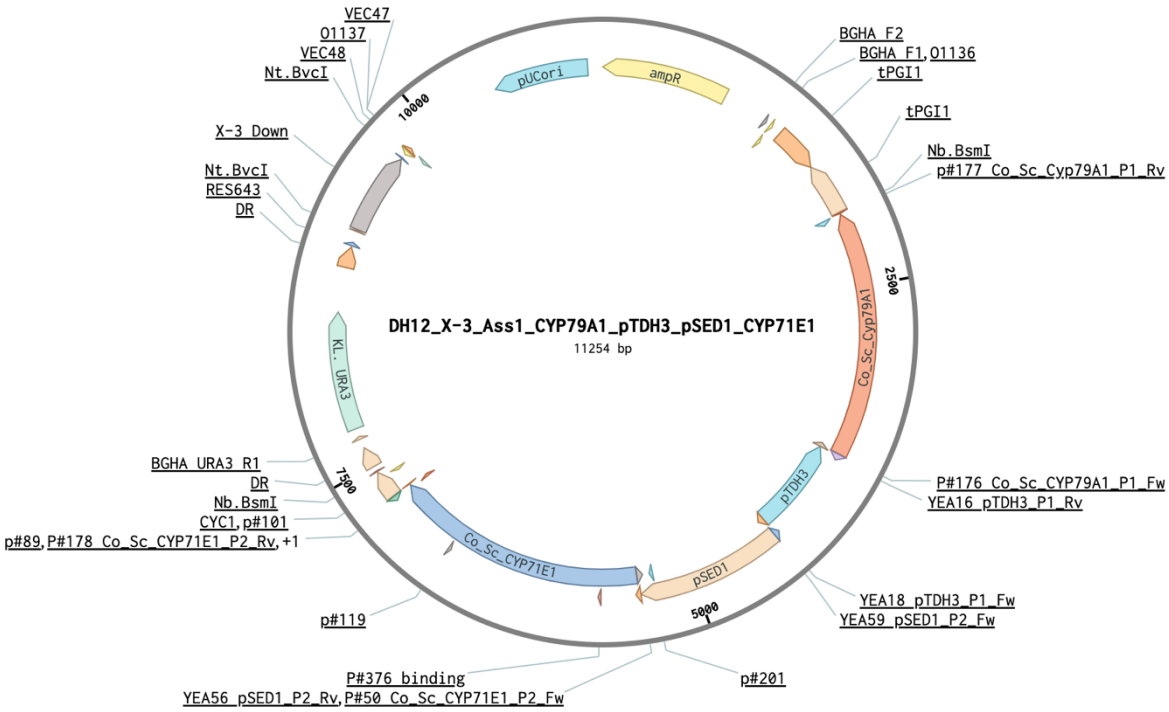
Appendix 7 - Suggested mutations predicted by the combination of the Cavity and EVE models



Appendix 8 - pJET1.2 plasmid containing the CYP79A1 gene used for generating mutant variants through site directed mutagenesis.



Appendix 9 - X-3 Assembler 1 vector before insertion of genes through USER cloning.



Appendix 10 - X-3 assembler 1 vector containing CYP79A1, pTDH3, pSED1, and CYP71E1 genes.