

UCLA

UCLA Electronic Theses and Dissertations

Title

Engineering Plant Cytochrome P450s for Self-sufficiency in Escherichia coli

Permalink

<https://escholarship.org/uc/item/51g4c3q1>

Author

Nania, Kane Ray

Publication Date

2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Engineering Plant Cytochrome P450s for
Self-sufficiency in *Escherichia coli*

A thesis submitted in partial satisfaction of the
Requirements for the degree Master of Science
In Chemical Engineering

by

Kane Nania

2017

© Copyright by

Kane Nania

2017

ABSTRACT OF THE THESIS

Engineering Plant Cytochrome P450s for
Self-sufficiency in *Escherichia coli*

by

Kane Nania

Master of Science in Chemical and Biomolecular Engineering

University of California, Los Angeles, 2017

Professor Tatiana Segura, Chair

The objective of this work was to engineer plant Cytochrome P450s to be self-sufficient in *Escherichia coli*, as a proof-of-concept of a novel protein engineering platform, termed ProtoVitro. The P450 reaction of interest was the hydroxylation of limonene for its potential to produce cancer therapeutics.¹ The reaction can be performed naturally by plant P450s but not bacterial ones,² so we created 28 fusion protein variants that each contained a plant P450 heme domain and either a bacterial or eukaryotic reductase domain. The heme domains were selected based on previously observed activity on limonene, and the reductase domains were selected as the maximally informative set from a list of thousands of prokaryotic and eukaryotic sequences. The rapid prototyping of diverse sequences facilitated by ProtoVitro allowed us to identify optimal protein variants in a more robust fashion than alternative protein engineering methods.

This thesis of Kane Nania is approved.

Philip Romero

Dante Simonetti

Yi Tang

Tatiana Segura, Committee Chair

University of California, Los Angeles

2017

Table of Contents

ABSTRACT OF THE THESIS	ii
List of Figures	v
List of Tables	vi
List of Acronyms	vii
INTRODUCTION	1
Protein Engineering.....	1
Natural Products	3
Plant P450s.....	5
MATERIALS AND METHODS.....	7
Materials.....	7
Cloning.....	7
<i>In vivo</i> protein expression	9
<i>In vitro</i> protein expression.....	10
4-Aminoantipyrine assay.....	10
Limonene activity assay	10
GC-MS analysis	11
Quantification of GC-MS results	11
Gel Electrophoresis	12
Sequence Alignment	12
Sequence Similarity Networks	12
RESULTS	13
ProtoVito: a general protein engineering platform	13
Designing a highly informative set of P450 reductase domains	15
Engineering self-sufficient plant cytochrome P450s	17
Cell-free expression of P450	17
Activity of P450 BM3 and P450 J on limonene.....	19
Characterization of P450 fusions.....	23
Hydroxylated products of limonene reaction	26
CONCLUSIONS.....	28
REFERENCES	29

List of Figures

Figure 1: Structures of limonene and its hydroxylated derivatives.	4
Figure 2: Flowchart of ProtoVitro engineering platform.....	14
Figure 3: Example matrix representation X	15
Figure 4: Sequence Similarity Network.....	16
Figure 5: 4-AAP assay on TX-TL expressed P450	18
Figure 6: Effect of hemin on GFP expression using TX-TL	19
Figure 7: Relative percent limonene remaining after limonene reaction of <i>P450 BM3</i> in cell lysate at various reaction time lengths.	20
Figure 8: SDS-PAGE of cell lysates	21
Figure 9: Relative percent limonene remaining after limonene reaction of <i>P450 BM3</i> and <i>P450 J</i> expressed using TX-TL.....	22
Figure 10: SDS-PAGE of TX-TL reaction products.	23
Figure 11: SDS-PAGE of P450 variants expressed in TX-TL.	24
Figure 12: Activity on limonene of all fusion variants	25
Figure 13: Activity on limonene of best hits.	25
Figure 14: Observed mass spectra	27

List of Tables

Table 1: List of reductase and heme domains used in this study.....	8
Table 2: List of variants by number and corresponding domains.....	9

List of Acronyms

TX-TL: transcription-translation; cell-free expression system

NP: natural product

P450 BM3: Cytochrome P450 from *Bacillus megaterium*, CYP102A1

P450 J: Previously engineered Cytochrome P450 from *Bacillus megaterium*

β -glu: β -glucosidase

GFP: green fluorescent protein

NADPH: nicotinamide adenine dinucleotide phosphate

GC-MS: gas chromatography–mass spectrometry

SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis

BLAST: Basic Local Alignment Search Tool

MUSCLE: Multiple Sequence Comparison by Log-Expectation

SSN: Sequence Similarity Network

4-AAP: 4-aminoantipyrine

INTRODUCTION

Protein Engineering

Substantial advancements have been made in the field of protein engineering since the early examples of mutagenesis first described over 30 years ago.³ However, the fundamental challenge that protein engineers face, namely, finding efficient ways to navigate the so-called protein fitness landscape to identify sequences with improved properties, still exists. Because the sequence space containing all possible amino acid variations is incredibly large, it is infeasible to test every point. Even with computational methods, it cannot be searched exhaustively.⁴ Within the sequence space, estimates of functional proteins range from 1 in 10^{11} to 1 in 10^{77} , and functional proteins still have a low chance of having improved target properties.⁵

One solution is directed evolution, which allows for the rapid improvement of protein fitness through iterative rounds of mutation and selection. Directed evolution has produced at least 12 therapeutics,⁶ and is used extensively in metabolic engineering.⁷ Libraries have been used on the industrial scale to screen for dehydrogenases and glycosylhydrolases, among other enzymes.⁸ However, evolutionary engineering methods rely on high-throughput assays to effectively explore protein sequence space. These assays typically involve growth-coupled selections or fluorogenic enzyme substrates, which severely limits the types of properties and enzymes that can be optimized. Most small molecules are not chromophoric, fluorescent, or essential for growth, and cannot be easily transformed into high-throughput assays.⁷ Most properties of industrial or medical importance require more detailed, low-throughput characterization.

Here, we propose a protein engineering framework that overcomes these limitations. First, using computational methods, a maximally informative subset of ten or fewer sequences are

identified from a larger set of thousands of sequences with homology to the protein of interest. Approximately 30 fusion or chimeric variants based on these ten sequences are then assembled, expressed in a cell-free environment, and assayed for the desired property. The modular nature of fusion or chimeric assembly as well as the many advantages of cell-free protein expression allow for protein variants to be rapidly tested and characterized.

In recent years, as an alternative to evolutionary methods, we have seen a shift in some efforts towards smaller, more informed libraries that employ the help of structural knowledge, as in the work done on improving the enantioselectivity of dehalogenases⁹ or the thermostability of ester hydrolases.¹⁰ But even without detailed structural information, we can still identify functional and optimized sequences via chimeric recombination. Chimeras allow for a fair degree of sequence diversity while minimizing structural disruption. They have been used to optimize protein properties and even discover novel enzyme functionalities.^{11,12} Because these chimeras are derived from already-known, functional proteins, there is a high chance that the chimera too will fold properly. Machine learning methods can also be implemented to generate intelligent combinations of a few known sequences, as was done previously to improve Cytochrome P450 thermostability.¹³ Another key advantage of using chimeras or fusion proteins is that many combinations of diverse sequences can be assembled using few initial starting blocks; for instance, eight starting sequences sliced at three locations yields 512 possible variants.

Cell-free protein expression systems are meant to simulate an *in vivo* cellular environment, whereby DNA is added to a crude *Escherichia coli* extract containing necessary transcription-translation (TX-TL) components. They are used here for a number of reasons. First, typical expression systems require DNA to be transformed into a cloning bacterial strain, purified, and re-transformed into an expression strain. With cell-free, once DNA is assembled it

can be expressed directly without a transformation step. High levels of protein can be produced in less than 6 hours.¹⁴ This reduction in time spent waiting for bacteria to grow allows for more efficient testing of many protein variants. Shin et al. demonstrated cell-free production of an equivalent amount of protein as commercial T7-based systems at 2% of the cost.¹⁵ Other advantages of cell-free systems over *in vivo* ones include decreased sensitivity to product toxicity, the ability to express off of linear DNA, suitability for high-throughput because of reduced reaction volumes and times, and the ability to overcome issues with membrane protein production.¹⁶

Natural Products

A natural product (NP) is any primary or secondary metabolite produced by a living organism. They have a variety of functions; in plants, they could be defense against herbivores, pigments to attract pollinators, signaling compounds, or simply waste products. There are many NPs or close derivatives that have medical relevance. Some examples include penicillin from *Penicillium chrysogenum* (antibacterial),¹⁷ lovastatin from *Pleurotus ostreatus* (lowers cholesterol),¹⁸ and artemisinin from *Artemisia annua* (antimalarial).¹⁹ Roughly half of drugs in clinical use today come from natural compounds.²⁰

There are four primary classes of natural products: alkaloids, phenylpropanoids, polyketides, and terpenoids. Terpenoids, also known as isoprenoids, are the largest class with over 50,000 known molecules.^{21,22} They are commonly used as fragrances, flavors, cosmetics, and medicines.²³

Of particular relevance to this work is d-limonene, a cyclic monoterpene found in orange peels and plant oils.¹ In one study, rats with dimethylbenz[a]anthracene-induced cancer were fed a 5% limonene diet for one week, and it was found that tumor latency was prolonged and tumor

multiplicity was halved compared to the negative control.²⁴ Another group found that limonene can cause 80% regression of mammary carcinomas.²⁵ We are interested in hydroxylated derivatives of limonene like perillyl alcohol (Fig. 1), which was actually five times more potent than limonene in preventing tumor formation in rats.¹ Limonene is also a cheap starting material, with prices reaching as low \$0.40 per kg in the past decade.²⁶ Approximately 520,000 tons of limonene are produced from oranges every year, an estimated 87% of which will end up in a landfill to biodegrade.²⁶

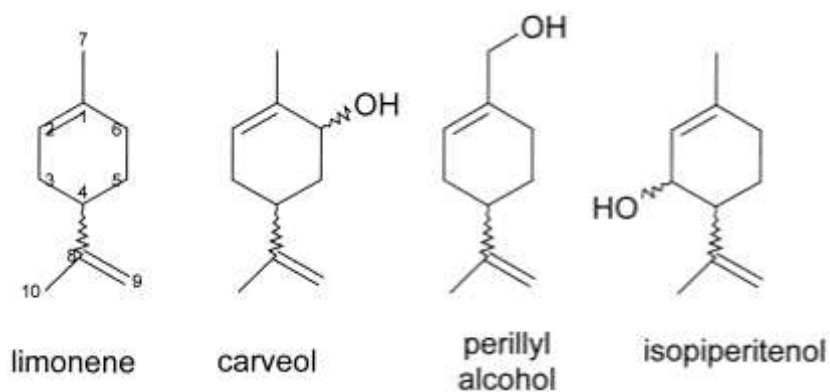


Figure 1: Structures of limonene and its hydroxylated derivatives.

One of the challenges to the pharmaceutical industry is harvesting large amounts of a given NP, as they are typically only produced by nature in small amounts.²³ In addition, they contain impurities, require a lot of resources to produce, and are difficult to scale up. For instance, it takes four mature yew trees to yield 1 g of paclitaxel.²⁷ Due to the complexity of these molecules, chemical synthesis is often difficult and expensive.²³

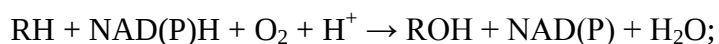
For these reasons, there has been a growing interest in synthesis of natural products in microbial hosts. Just this year, there have been several studies on the production of health-

beneficial natural compounds in *E. coli*, such as tyrosol,²⁸ dimethoxyflavanoids,²⁹ and curcuminoids,³⁰ to name a few. Chang et al. first introduced the mevalonate pathway into *E. coli* to produce isoprenoid precursors dimethylallyl pyrophosphate and isopentenyl pyrophosphate,²³ from which all isoprenoids can be derived.²¹ Wang et al. provides an excellent review of the many metabolic engineering efforts for isoprenoid production since then.²²

Combinatorial chemistry has allowed for improved efficiency in chemical synthesis, but after twenty five years, there hasn't been a dramatic increase in the number of drug candidates.³¹ In terms of drug discovery, biological modification of existing drugs, or using enzymes to generate potential drug variants, may prove more fruitful than chemically synthesizing millions of drug variants. This is because biologically-produced compounds, as compared to combinatorial chemical libraries, have more chiral centers, more rigid structures, and fewer aromatic moieties.³² They often have much simpler structures overall.³² This further necessitates detailed, low throughput protein engineering and characterization.

Plant P450s

Cytochrome P450s are a class of heme-containing monooxygenases found in all domains of life.³³ They perform a wide array of reactions, including hydroxylation, epoxidation, dealkylation, oxidation, dehydration, C-C bond cleavage/formation, and isomerization.³⁴ One key P450 reaction is the activation of molecular oxygen using NAD(P)H,



where RH is a site that can be hydroxylated.³³ This selective functionalization of nonactivated CH groups is generally difficult to do via synthetic chemistry,³⁵ but P450s can naturally do it under mild conditions, making them attractive candidates for use in medical and industrial processes.

Bacterial P450s have been the target of many engineering efforts. The widely studied cytochrome P450 from *Bacillus megaterium* (P450 BM3) has been engineered to hydroxylate or dealkylate numerous non-natural compounds, including terpenes,³⁶ steroids,³⁷ and therapeutics like lovastatin³⁸ and verapamil.³⁹

P450s from plants are of particular interest because of their native ability to synthesize natural products and/or metabolize toxins.²³ However, eukaryotic P450s are often membrane-bound, making them insoluble and difficult to manipulate in the lab.³³ Their application is further precluded by a lack of suitable expression systems, low activity, and the necessity of additional redox-transport enzymes.⁴⁰ Of the thousands of plant P450 sequences available, only a few have been expressed in *E. coli*.²³ Unlike the *B. megaterium* P450, which contains both a heme and reductase domain as part of the same enzyme, plant P450s have heme and reductase functions split across multiple enzymes, resulting in less efficient electron transfer from NADPH.³³

One way to improve heterologous expression of plant P450s in microbial hosts is to simply co-express the additional enzymes, which improved activity of one P450 system by ten-fold.⁴¹ Better than this is to fuse a reductase domain to the plant P450, while also truncating the N-terminal membrane-bound anchor.⁴²⁻⁴⁴ This improves electron transfer and removes the insoluble portion of the enzyme. Zhao et al. found that these modifications resulted in a 4.5-fold increase in activity and over a 70% increase in product formation as opposed to when additional enzymes were co-expressed.⁴³ Another option is to add an N-terminal tag that promotes eukaryotic expression in *E. coli*.^{42,45}

MATERIALS AND METHODS

Materials

Golden Gate Assembly Mix and competent cells were purchased from New England Biolabs (Ipswich, MA). MagicMedia *E. coli* Expression Medium was purchased from ThermoFisher Scientific. Primers and gBlocks and were purchased from Integrated DNA Technologies (Coralville, IA). All other chemicals used in this work were purchased from Sigma-Aldrich. Cell-free expression systems (TX-TL, transcription-translation) were provided by Synvirobio, Inc (San Francisco, CA).

Cloning

Wild type Cytochrome P450 from *B. megaterium* (*P450 BM3*; CYP102A1) was inserted into a pET22b vector using a 3-piece Gibson assembly of the vector backbone and two gBlocks that made up the gene insert. The *P450 J* variant of *P450 BM3*, engineered originally by Fasan et al.,⁴⁶ contains mutations at Y138H, I178V, F205C, S226R, and T295A. *P450 J* was assembled into a plasmid containing a σ_{70} promoter by collaborators at Synvirobio, Inc. A negative control plasmid containing the gene for β -glucosidase in a pET22b vector (*β -glu*) was obtained from Dr. Philip Romero at University of Wisconsin-Madison. Another negative control plasmid containing the gene for green fluorescent protein (*GFP*) was provided by collaborators at Synvirobio, Inc.

The eight P450 reductase domains were assembled onto linear DNA with four P450 heme domains using golden gate cloning to make every possible combination of reductase and heme. Domain names, origin, and background are provided in Table 1. All variations were cloned, but only the ones outlined in Table 2 were successfully expressed and tested analytically.

Table 1: List of reductase and heme domains used in this study.

Name	Organism	Function	Gene or NCBI reference sequence
Reductase domains			
BM3	<i>Bacillus megaterium</i>	bacterial reductase	CYP102A
MA	<i>Mucor ambiguus</i>	eukaryotic reductase	GAN03094
BS	<i>Byssochlamys spectabilis</i>	eukaryotic reductase	GAD98773
CS	<i>Cochliobolus sativus</i>	eukaryotic reductase	XP_007702044
SS	<i>Streptomyces sp.</i>	bacterial reductase	WP_063351614
NM	<i>Nakamurella multipartita</i>	bacterial reductase	WP_015747835
CF	<i>Corynebacterium falsenii</i>	bacterial reductase	WP_065420181
SG	<i>Saprospira grandis</i>	bacterial reductase	WP_002659767
Heme domains			
SM12	<i>Mentha spicata</i>	limonene-6-hydroxylase	CYP71D18
PM17	<i>Mentha x piperita</i>	limonene-3-hydroxylase	CYP71D13
PM2	<i>Mentha x piperita</i>	limonene-3-hydroxylase	CYP71D15
BM3 J	<i>Bacillus megaterium</i>	evolved limonene hydroxylase	CYP102A1 J variant

Table 2: List of variants by number and corresponding domains.

Variant Number	Heme domain	Reductase domain
774	SM12	BM3
775	SM12	MA
777	SM12	CS
778	SM12	SS
780	SM12	CF
781	SM12	SG
782	PM17	BM3
785	PM17	CS
786	PM17	SS
788	PM17	CF
789	PM17	SG
790	PM2	BM3
791	PM2	MA
794	PM2	SS
796	PM2	CF
797	PM2	SG
798	BM3 J	BM3
799	BM3 J	MA
802	BM3 J	SS
804	BM3 J	CF
805	BM3 J	SG

***In vivo* protein expression**

Plasmids containing genes for *P450 BM3*, *P450 J*, and *β -glu* were transformed into BL21 (DE3-) *E. coli* cells. Cells were grown overnight (~18 hr) in MagicMedia, and 1 ml cell pellets were either frozen at -80° C or used immediately. Pellets were resuspended in 100 μ l of Tris-HCl buffer and lysed with an additional 175 μ l of lysis buffer (1 μ l lysonase, 25 μ l BugBuster reagent, 500 μ l of 100 mM Tris-HCl buffer).

***In vitro* protein expression**

In vitro protein expression via TX-TL was carried out in a 100 µl reaction, containing 30 µl of *E. coli* cell extract, 30 µl of TX-TL buffer, and 10 nM of DNA. Hemin, if added, was at a concentration of 10 µM. Expression was carried out for >12 hours and the products were either used immediately or stored at -80° C.

4-Aminoantipyrine assay

In a 96-well plate, a 100 µl reaction containing 80 mM 2-phenoxyethanol, 4 mM H₂O₂, and 20 µl cell lysate was prepared. For this assay, just the heme domain of *P450 BM3* containing a mutation at F87A was used, which allows the enzyme to function as a peroxygenase that uses H₂O₂ instead of a monooxygenase that uses NADPH.⁴⁷ The reaction was carried out for 2 hours at room temperature, and was quenched by the addition of 200 µl of a quench buffer (100 mM NaOH, 4 M urea). To develop the reaction products, 36 µl of 6% 4-aminoantipyrine (4-AAP) was added, the solution was blanked at 500 nm, and 36 µl of 6% potassium persulfate was added. After 2 hours, the absorbance at 500 nm was recorded.

Limonene activity assay

The reaction mixture (500 µl) contained 200 µM s-(-)-limonene, 2 mM nicotinamide adenine dinucleotide phosphate (NADPH), 100 µl of either cell lysate or TX-TL reaction product containing overexpressed enzyme, and 100 mM Tris-HCl buffer (pH 8.2). The reaction was initiated by the addition of NADPH and proceeded for 20 hours, unless otherwise noted. The reaction was quenched by adding 500 µl of 4 M urea solution and placing the samples on ice. Standard curves used in GC-MS analysis were prepared by running the same reaction at various limonene concentrations (0, 25, 50, 100, and 200 µM). For the standards, a negative control

enzyme sample was added, *β -glu* for assays on cell lysate or *GFP* for assays on TX-TL products. All samples were run in duplicate unless otherwise noted. Fresh standards were prepared with every assay run to account for potential variation among different runs.

GC-MS analysis

After quenching the reaction, 50 μ l of 10 mM of internal standard, 1-decene (final concentration 0.48 mM), was added to all reaction products and standards. All samples underwent extraction twice with 300 μ l of chloroform.

Analysis was performed on a Thermo Q Exactive gas chromatography–mass spectrometer (GC-MS) at the UCLA Parasow Mass Spectrometry Laboratory. The GC was equipped with a Zebron ZB-5 column (phenomenex) (length: 60 m, internal diameter: 0.25 mm, thickness: 0.25 μ m). The GC was programmed as follows: 50° C for 3 min, ramp to 300° C at 5° C per min, 300° C for 10 min. The machine was run in splitless mode with an inlet temperature was 250° C, and the filament was turned on after 6 min.

Limonene, 1-decene, and other compounds were verified by comparison of their characteristic mass fragmentation patterns in the National Institute of Standards and Technology mass spectrometry database.

Quantification of GC-MS results

A standard curve was created by plotting the amount of limonene added to the reaction against the ratio of peak areas of limonene to 1-decene for each standard. Peak areas were calculated using Thermo Xcalibur software. Ratios of limonene to 1-decene for enzyme samples were converted to the amount of limonene initially present in the sample using the trend line of the standard curve. This amount of limonene in enzyme samples was divided by the amount in the

negative control to give the percent limonene remaining. In cases where the standard curve couldn't be used, the ratios of limonene to 1-decene in enzyme samples were compared to those in the negative control, resulting in the relative percent limonene consumption.

Gel Electrophoresis

Cell lysates and TX-TL reaction products were mixed with Laemmli sample buffer and heated at 95° C for 10 minutes. Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was performed at 200 V for 20 min, and the gel was stained using Coomassie Blue.

Sequence Alignment

A Basic Local Alignment Search Tool (BLAST) search was performed using the sequence of the heme domain of *P450 BM3*.⁴⁸ After filtering out truncated or redundant sequences, 1,884 viable sequences remained. The online tool Multiple Sequence Comparison by Log-Expectation (MUSCLE) was then used to align the sequences.⁴⁹

Sequence Similarity Networks

A Sequence Similarity Network (SSN) was generated to assess variation among the eight P450 reductase domains. Using BLAST, 1,884 sequences were identified. The data was processed using the online tool EFI-EST and the network was generated using Cytoscape 3.⁵⁰ Alignment histograms and quartile plots were generated in EFI-EST using the 1,884 sequences (e-value = 50, fraction = 1). These plots were used to determine the alignment score (a = 127), a measure of how clustered the resulting SSN would be. The data was then imported into Cytoscape. The percent identity was set to 95, which means that each node on the SSN represents all sequences with 95% or more homology.

RESULTS

ProtoVitro: a general protein engineering platform

ProtoVitro was designed to rapidly characterize a small, but diverse set of protein variants. As outlined in Figure 2, Step 1 is to use BLAST to identify a large set of sequences that are related to the protein of interest. In the second step, three to ten of the most diverse sequences from the larger set are identified using computational methods discussed below. For the enzyme family tested in this paper, Steps 1 and 2 were applied only to the one of the two domains. The sequences for the other domain were selected based on prior research. In Step 3 for this case, every possible combination of the two domains is assembled.

As a more general approach, Steps 1 and 2 could be applied to a full-length protein. The resulting sequences would then be sliced at certain breakpoints, selected randomly or via SCHEMA, to generate different building blocks for the fusion proteins. Step 3 is to assemble every possible fusion variant from these building blocks. For example, if one sliced three sequences at two breakpoints, each fusion would have three parts. Each part could come from any of the three original sequences, resulting in 27 possible fusions. In any case, the number of final sequences should be chosen such that it is feasible to quickly assemble and test all variants (<50).

In Step 4, each variant is expressed *in vitro* using TX-TL. Finally, an enzymatic assay is performed and assessed using appropriate analytical methods. It is also possible to run the assay concurrently with the protein expression step to save time.

Designing a highly informative set of P450 reductase domains

In Step 1 of ProtoVitro for P450 engineering, 1,884 P450 reductase domains were identified using BLAST. The next step was to determine a subset of 8 sequences that were maximally different from each other, thereby comprising the most informative set. The total set of sequences can be represented in matrix form ($\mathbf{X}$), where each row corresponds to a different sequence and each column corresponds to a possible amino acid at each position (Fig. 3). Each sequence is converted to a binary form, where a 1 means the protein contains a given amino acid at a given position, and a 0 means it does not. Thus, the sum of each row of the matrix is equal to the total length of that protein.

	(Position 1)			(Position 2)		(Position 3)		
	<i>leu</i>	<i>ile</i>	<i>lys</i>	<i>ala</i>	<i>pro</i>	<i>ala</i>	<i>gly</i>	<i>thr</i>

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Sequence 1 = *leu-ala-ala*

Sequence 2 = *leu-pro-gly*

Sequence 3 = *ile-pro-thr*

Sequence 4 = *lys-ala-thr*

Figure 3: Example matrix representation $\mathbf{X}$ of a set of 4 protein sequences each with 3 amino acids.

The covariance of the sequence matrix is defined as $\mathbf{X}\mathbf{X}^T$. The determinant of the covariance is proportional to the multivariate Gaussian entropy (Equation 1), which is effectively a measure of mutual information, or how similar a set of sequences are to one another.

$$\det(\text{covar}(\mathbf{X})) = \det(\mathbf{X}\mathbf{X}^T) \propto \text{entropy of Gaussian} \quad (1)$$

The submodularity property of entropy is analogous to convexity, which allows for an efficient method to find the optimum. A greedy algorithm can be used to optimize the submodularity to find the minimum entropy. Starting from the *P450 BM3* reductase sequence as the first entry to a subset of the larger set of 1,884 sequences, additional sequences that minimized the entropy were added iteratively to the subset. For example, the second sequence of the subset was chosen such that it minimized entropy with *P450 BM3*, the third sequence minimized entropy with both *P450 BM3* and the second sequence, and so on, until 8 sequences were obtained. The result can be visualized on a Sequence Similarity Network, indicating that the final 8 proteins are indeed different from each other (Fig. 4).

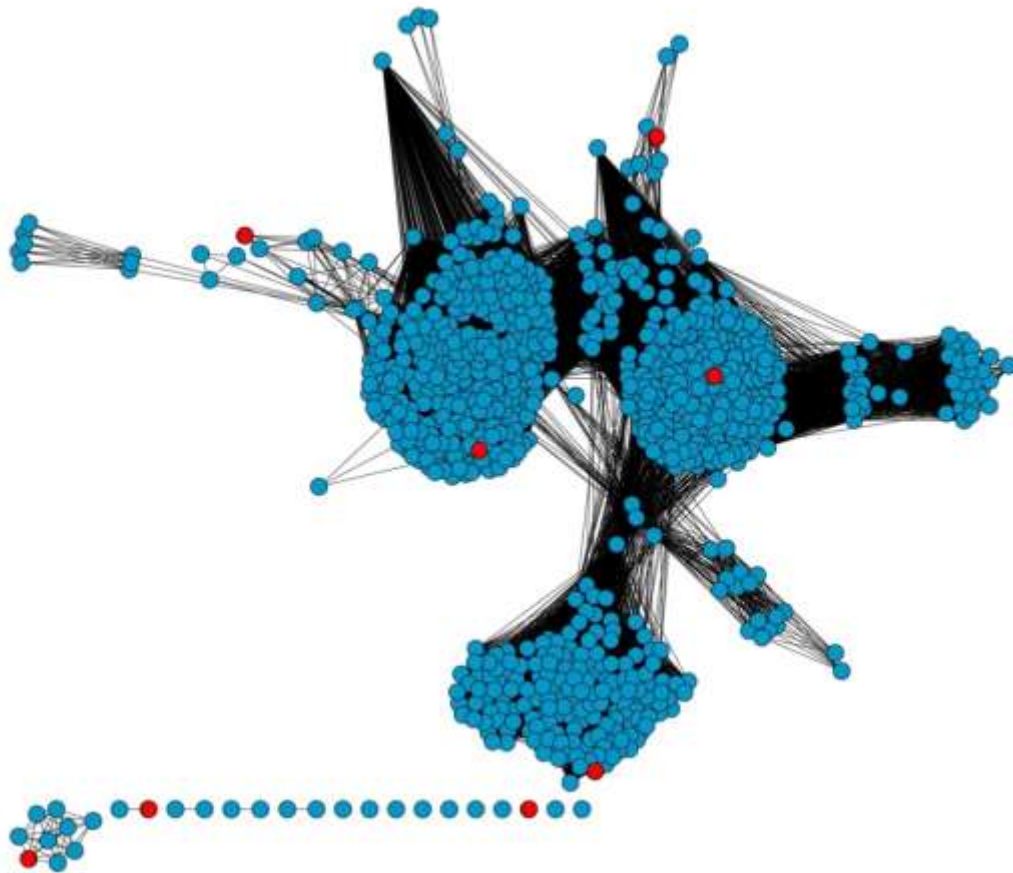


Figure 4: Sequence Similarity Network

SSN for the set of 1,884 P450 reductase domains. Each node represents all sequences with 95% identity. Red nodes are the eight variants that comprise the maximally informative set.

Engineering self-sufficient plant cytochrome P450s

Four heme P450 domains and eight P450 reductase domains were identified and each possible heme and reductase fusion pair was assembled. The heme domains were chosen based on previously shown limonene activity. P450 heme domains from CYP71D18 of *Mentha spicata* (Spearmint), CYP71D13 of *Mentha piperita* (Peppermint; *Mentha aquatica* × *Mentha spicata*), and CYP71D15 of *Mentha piperita* were chosen based on limonene activity shown by Lupien et al. and others,⁵¹ along with *P450 J* from Fasan et al.'s work, which shows enhanced activity on many terpenes including limonene.⁴⁶ The eight reductase domains were selected by the computational methods described above. There were four bacterial reductase domains and three eukaryotic ones.

Cell-free expression of P450

Preliminary experiments confirmed that *P450 BM3* expressed using a cell-free system was active using a 4-aminoantipyrine (4-AAP) assay (Fig. 5). In this case, the gene encoding P450 was under control of a T-7 promoter and supplemental DNA containing T-7 RNA Polymerase was added to the reaction. Previously, Kwon et al. found that to observe proper formation of the heme-containing P450 *in vitro*, either exogenous hemin had to be added to the reaction or a less-dialyzed cell extract (S12) that retained cofactors necessary for heme production had to be used.⁵² Hemin appears to be inhibitory to protein expression using TX-TL, as shown in Figure 6. However, at high concentrations of hemin, the reduction in *GFP* expression seems to plateau at

roughly a third of the expression when no hemin is added. Our results in Figure 5 indicated that the addition of heme precursor hemin to the TX-TL reaction did not make a dramatic difference in activity of *P450 BM3*, and that the S12 extract we used here was sufficient for P450 synthesis.

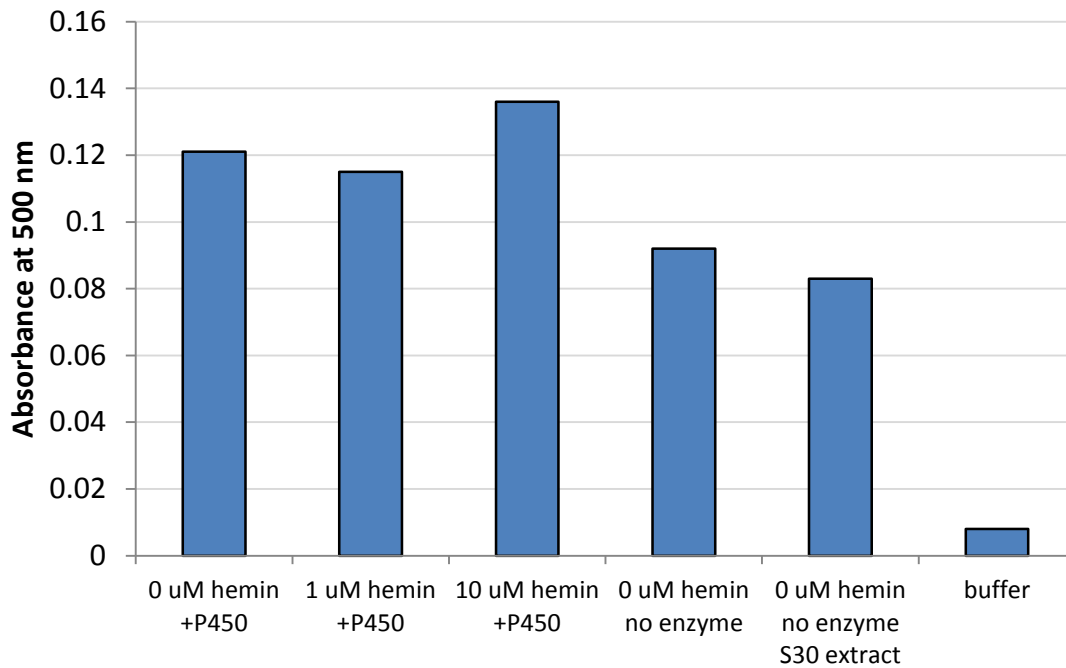


Figure 5: 4-AAP assay on TX-TL expressed P450

Activity of P450 BM3 heme domain expressed with TX-TL using 4-AAP assay at varying concentrations of hemin. All samples were run using the S12 extract except where otherwise noted. The "buffer" negative control was the absorbance of only Tris-HCl buffer in the plate well. n=1.

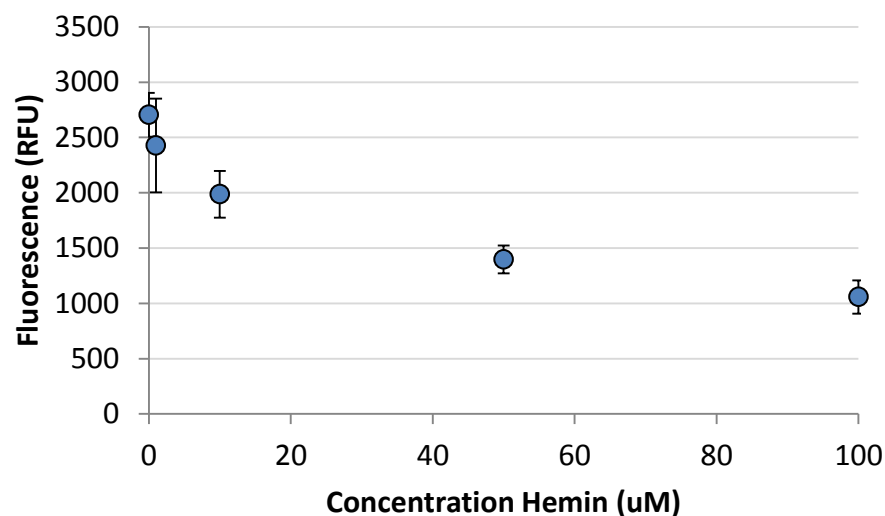


Figure 6: Effect of hemin on GFP expression using TX-TL

GFP was expressed in TX-TL for 3 hours at 0, 1, 10, 50, or 100 μ M hemin and the relative fluorescence was measured. $n=2$.

Activity of P450 BM3 and P450 J on limonene

Wild-type P450 BM3 was previously found to have some minor activity on limonene.⁴⁶ This was confirmed first by expressing the enzyme *in vivo* and running a limonene assay on the cell lysate (Fig. 7). A peak of the gas chromatograph corresponding to limonene was observed at 8.39 min, and a peak corresponding to the internal standard, 1-decene, was observed at 7.32 min. For the 20 hour reaction, 66% of the limonene was consumed by the full-length P450 BM3. The motive for examining substrate depletion, as opposed to the formation of products, was because the P450s are expected to hydroxylate the substrate at various positions. For quantitative results, it is more accurate to look at depletion. In addition, with the fusion proteins it is possible that novel, unexpected products are formed. Expression of P450s was confirmed with SDS-PAGE, with the

expected band seen at 118 kDa (Fig. 8). The *P450 J* variant was also tested here but it appears that the expression was unsuccessful, as there was no band at the correct size on the gel and no activity was observed (data not shown).

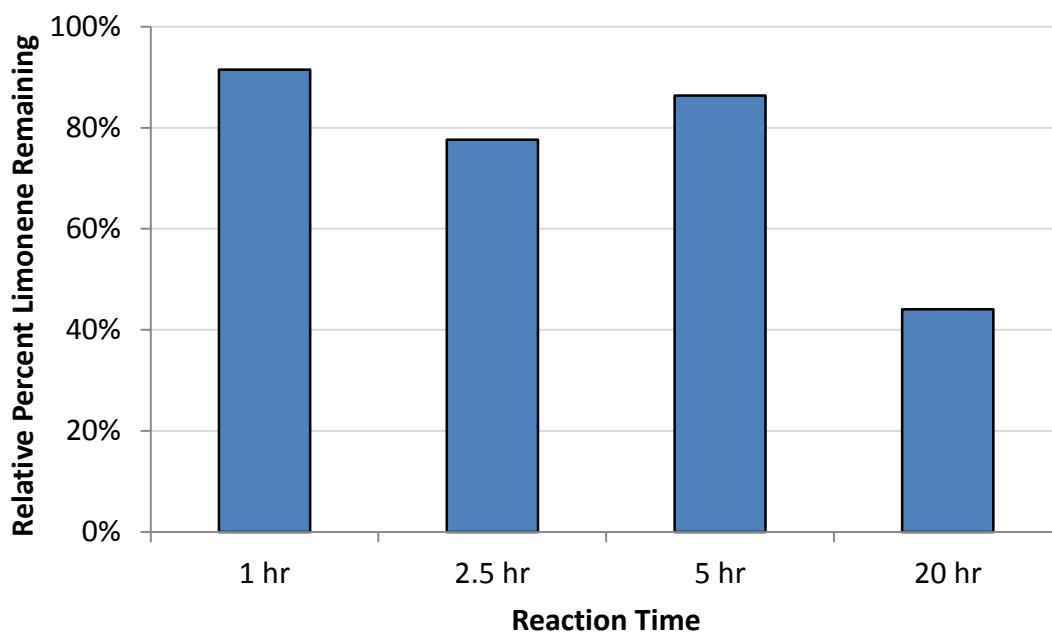


Figure 7: Relative percent limonene remaining after limonene reaction of P450 BM3 in cell lysate at various reaction time lengths.

Cell lysates containing either expressed full-length P450 BM3 or β -glu were allowed to react with limonene for varying times. Results shown indicate the ratios of limonene to 1-decene (internal standard) mass spectra peak areas for P450 samples relative to those of β -glu negative control samples. $n=1$.

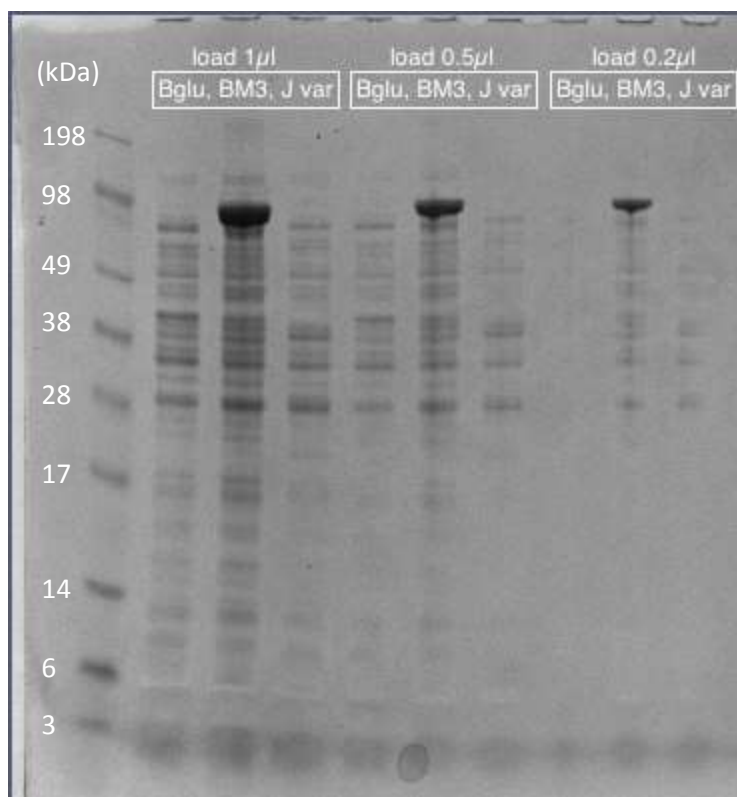


Figure 8: SDS-PAGE of cell lysates expressing β -glu, P450 BM3, or P450 J (J var) at various well loading volumes. Expected bands at 118 kDa.

P450 BM3 and P450 J were further tested after *in vitro* TX-TL expression. Figure 9 indicates that when exogenous hemin was added to the TX-TL reaction, there was 81% of the limonene originally added to the reaction remaining in the P450 BM3 sample compared to the GFP negative control, and $39 \pm 7\%$ for P450 J. When hemin was not added, no limonene consumption was observed. Expression was verified with SDS-PAGE and the correct bands at 118 kDa are observed regardless of whether or not hemin was added (Fig. 10). It appears that P450s can be expressed using TX-TL, but the enzyme will only fold properly, and therefore be active, if the heme produced from hemin is incorporated into P450.

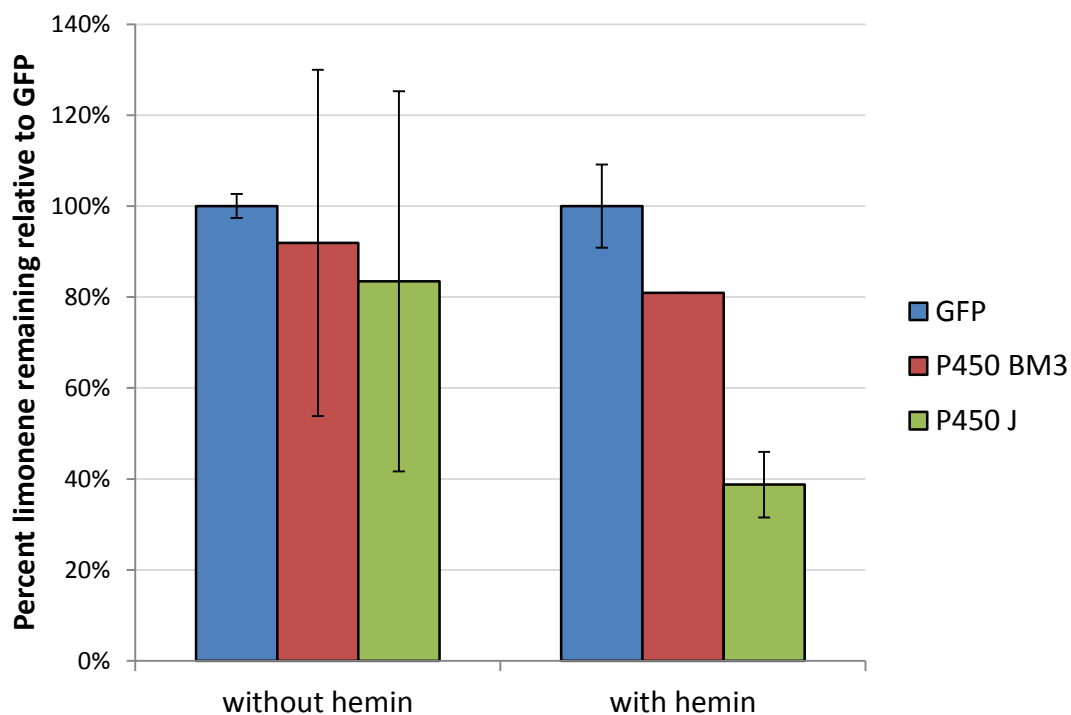


Figure 9: Relative percent limonene remaining after limonene reaction of P450 BM3 and P450 J expressed using TX-TL.

TX-TL reaction mixtures containing either expressed P450 BM3, P450 J, or GFP reacted with limonene. Either 0 or 10 μ M hemin was added to the TX-TL reaction. Results shown indicate the ratios of limonene to 1-decene (internal standard) mass spectra peak areas for all samples relative to those of GFP negative control samples. n=2.

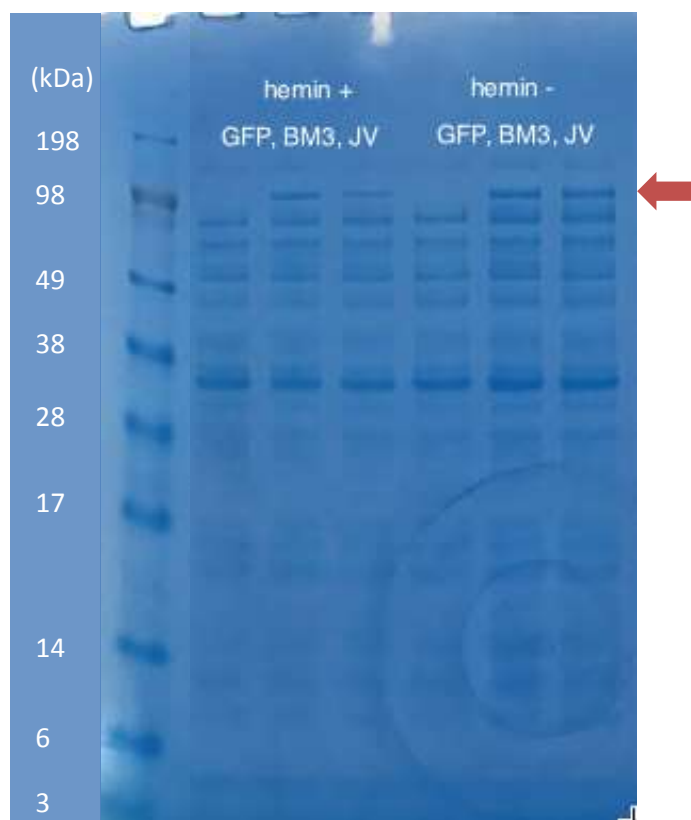


Figure 10: SDS-PAGE of TX-TL reaction products.

Samples are GFP control, P450 BM3, and P450 J variant (JV). Proteins were expressed either with or without 10 μ M hemin added. Expected bands at 118 kDa.

Characterization of P450 fusions

All P450 fusions were assembled in the form of linear DNA. TX-TL expression of the variants was confirmed with SDS-PAGE, with bands present at the expected 118 kDa (Fig. 11). P450 BM3, P450 J, GFP, and variant 774 were also cloned and expressed in plasmid form. The limonene assay was performed and the consumption, as measured by the ratio of limonene to internal standard 1-decene, was plotted in Figure 12. By comparison to the GFP negative

controls, it appears that the linear DNA version of the positive control *P450 J* performed as expected, and most of the variants showed some level of limonene consumption. However, there seemed to be a high degree of variability in the measurements, as indicated by the fact that variants 781, 786, and 797, as well as two of the plasmid samples, showed more apparent limonene consumption than the negative control.

The five best potential variants were selected for further analysis, and limonene consumption was observed for all but one (Fig. 13). Variant 791, containing a peppermint heme domain and eukaryotic reductase domain from *Mucor ambiguus*, showed 59% consumption of the limonene relative to the negative control. Variant 782, containing a different peppermint heme domain and the *B. megaterium* reductase, consumed nearly the same amount. The linear DNA *P450 J* positive control in this case was only at 39%. These results are surely promising, but further analysis will be necessary to demonstrate that they are repeatable. As evidenced by the differences in *GFP* linear DNA and plasmid samples, there may be variability in the GC-MS measurements that affected our results.

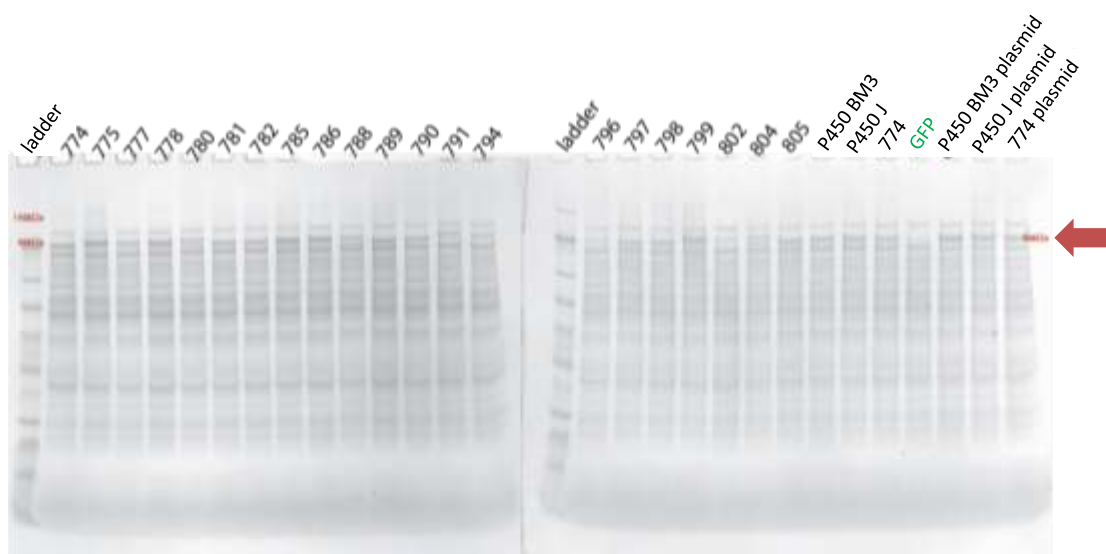


Figure 11: Expression SDS-PAGE of *P450* variants expressed in TX-TL off of linear DNA.

P450 BM3, P450 J, and 774 were also expressed off plasmid DNA. Expected bands at 118 kDa.

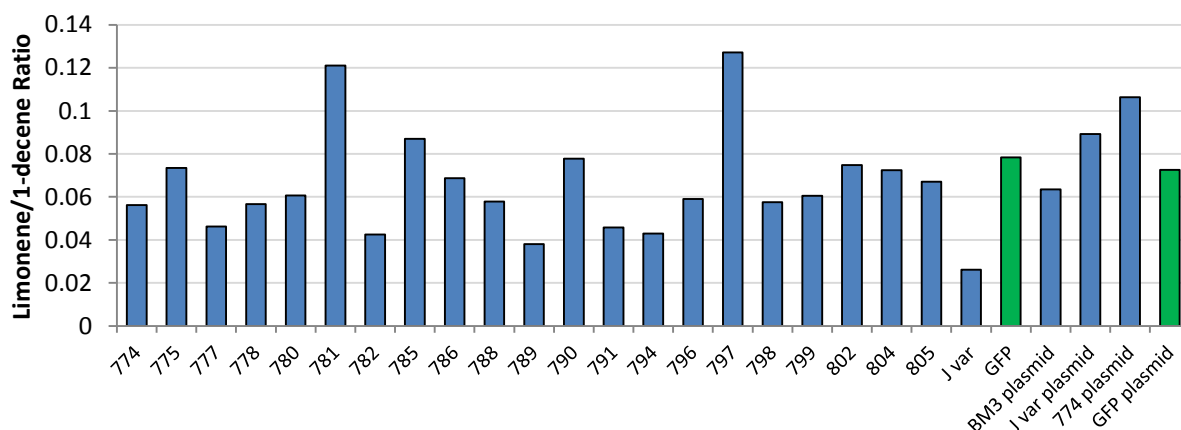


Figure 12: Activity on limonene of all fusion variants, as indicated by the ratio of limonene to 1-decene remaining in the sample after the reaction with limonene. Variants were expressed in TX-TL off of linear DNA. GFP negative controls shown in green. P450 BM3, P450 J (J var), variant 774, and GFP were additionally expressed off of plasmid DNA.

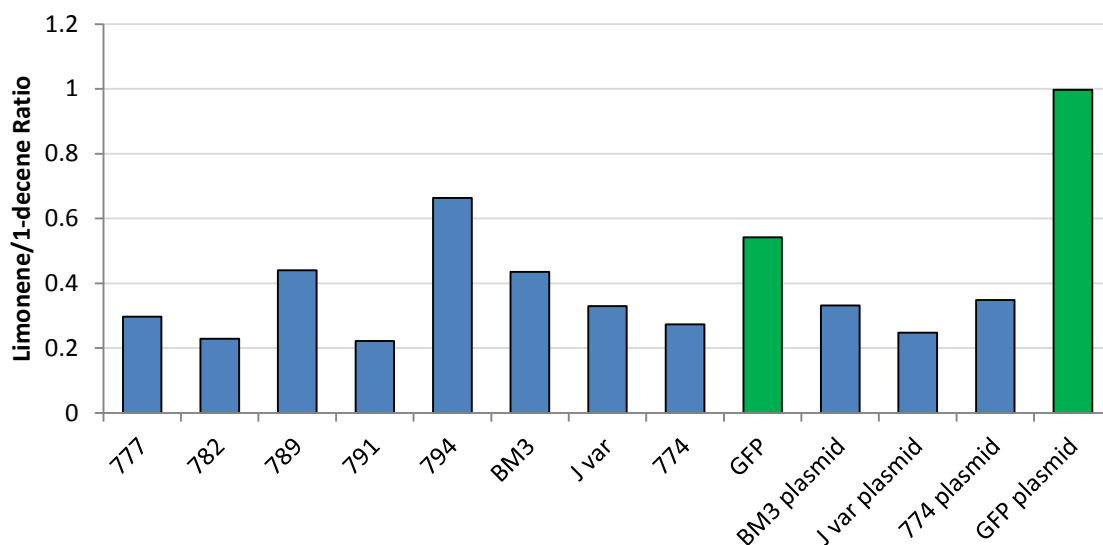
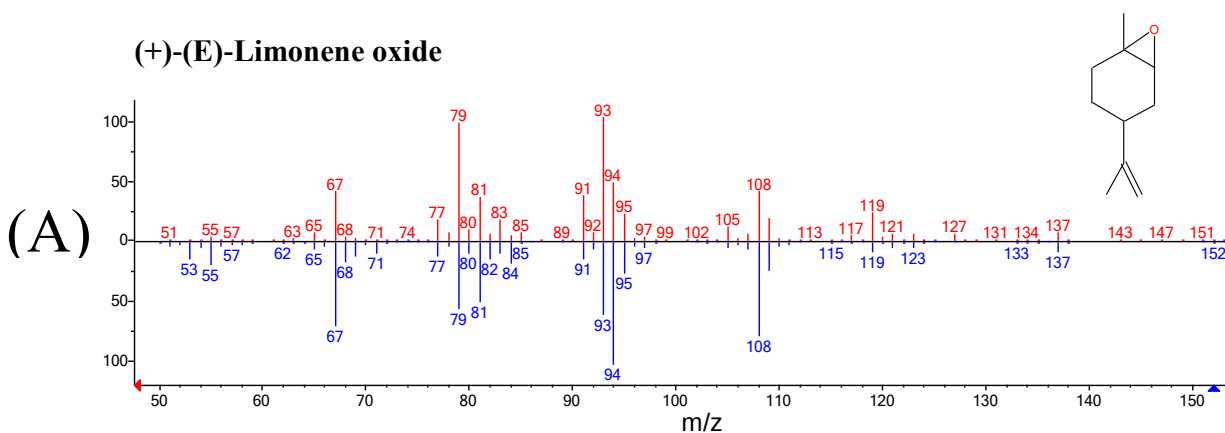


Figure 13: Activity on limonene of best hits from previous run, as indicated by the ratio of limonene to 1-decene remaining in the sample after the reaction with limonene. Variants

expressed in TX-TL off of linear DNA. GFP negative controls shown in green. P450 BM3, P450 J (J var), variant 774, and GFP were additionally expressed off of plasmid DNA.

Hydroxylated products of limonene reaction

A couple of interesting products of the P450 reaction on limonene were observed by GC-MS analysis, including (+)-(E)-Limonene oxide and trans-p-Mentha-2,8-dien-1-ol (Fig. 14). It appears that in our experiment, the P450 hydroxylated or oxidized the substrate at different locations compared mint P450s previously tested, which hydroxylate at positions C-3, C-6, or C-7. It is also possible that the software inappropriately identified the compounds, as even when known compounds are injected into the GC-MS, the percent certainty that the compound was the one we expected was fairly low. However, it is clear that P450 is indeed reacting with limonene, and it is expected that the enzyme will hydroxylate the substrate at various positions.



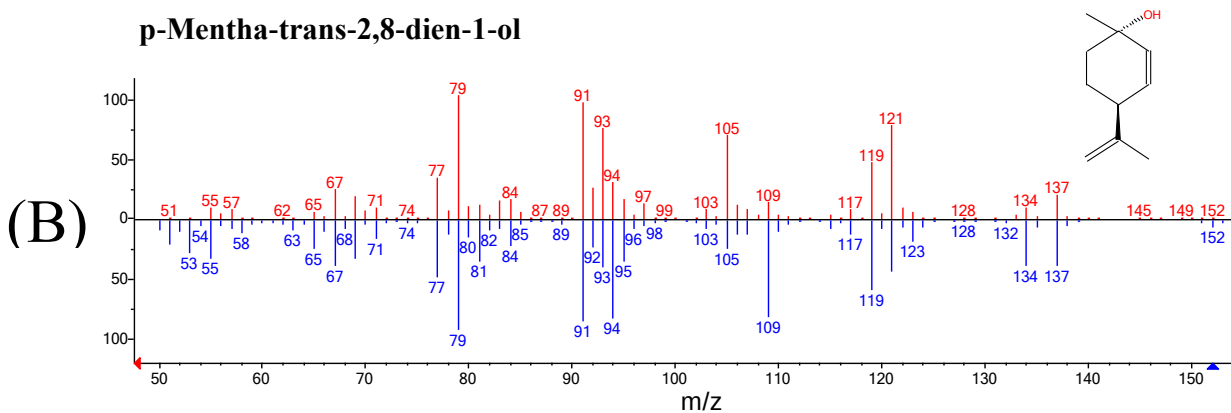


Figure 14: Observed mass spectra for (A) (+)-(E)-Limonene oxide at 11.55 min and (B) p-Mentha-trans-2,8-dien-1-ol at 13.25 min. Red peaks correspond to the mass fragmentation patterns of the observed products and blue peaks correspond to those in the NIST database.

CONCLUSIONS

ProtoVitro allows for the rapid characterization of protein variants in situations where high throughput assays aren't available. It leverages computational methods to provide maximally informative sequences that are likely to fold properly, as well as a modular assembly framework to minimize costs of DNA synthesis. Further efficiencies are obtained via cell-free expression systems, which obviate transformation steps and allow for expression of linear DNA.

In this work, we demonstrated the application of ProtoVitro to the engineering of self-sufficient plant Cytochrome P450s in *E. coli*. Our best variant (variant 791), a fusion of a peppermint heme domain with a eukaryotic reductase domain, showed improved limonene hydroxylation activity over the positive control *P450 J*, an engineered bacterial enzyme. Variant 791 consumed 59% of the limonene relative to the negative control, whereas the *P450 J* only consumed 39%.

One interesting avenue for future research is to increase automation of the platform. Using liquid handling robots or even a Cloud Lab, a fully-automated facility to which users can outsource their experiments and analytics, the routine tasks of DNA assembly, expression, and assaying can be programmed for and ran. The machine learning techniques described by Romero et al. could also be applied to iteratively test the next most informative sequence to efficiently progress towards the optimum.⁵ Due to its modular nature, ProtoVitro is particularly suitable for this type of testing.

REFERENCES

1. Crowell, P. L. *et al.* Chemoprevention of mammary carcinogenesis by hydroxylated derivatives of d-limonene. *Carcinogenesis* **13**, 1261–1264 (1992).
2. Mau, C., Karp, F., Ito, M., Honda, G. & Croteau, R. B. A candidate cDNA clone for (–)-limonene-7-hydroxylase from *Perilla frutescens*. *Phytochemistry* **71**, 373–379 (2010).
3. Chica, R. A. Protein engineering in the 21st century. *Protein Sci.* **24**, 431–433 (2015).
4. Pierce, N. A. & Winfree, E. Protein design is NP-hard. *Protein Eng.* **15**, 779–782 (2002).
5. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl. Acad. Sci.* **110**, E193–E201 (2013).
6. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **10**, 188–195 (2011).
7. Dietrich, J. A., McKee, A. E. & Keasling, J. D. High-throughput metabolic engineering: advances in small-molecule screening and selection. *Annu. Rev. Biochem.* **79**, 563–590 (2010).
8. Lorenz, P. & Eck, J. Metagenomics and industrial applications. *Nat Rev Micro* **3**, 510–516 (2005).
9. van Leeuwen, J. G., Wijma, H. J., Floor, R. J., van der Laan, J. M. & Janssen, D. B. Directed evolution strategies for enantiocomplementary haloalkane dehalogenases: from chemical waste to enantiopure building blocks. *Chembiochem* **13**, 137–148 (2012).
10. Blum, J., Ricketts, M. & Bommarius, A. Improved thermostability of AEH by combining

- B-FIT analysis and structure-guided consensus method. *J. Biotechnol.* **160**, 214–221 (2012).
11. Otey, C. R. *et al.* Functional evolution and structural conservation in chimeric cytochromes p450: calibrating a structure-guided approach. *Chem. Biol.* **11**, 309–318 (2004).
 12. Otey, C. R. *et al.* Structure-guided recombination creates an artificial family of cytochromes P450. *PLoS Biol.* **4**, e112 (2006).
 13. Romero, P., Krause, A. & Arnold, F. Navigating the protein fitness landscape with Gaussian processes. *Proc Natl. Acad Sci* **110**, E193–E201 (2013).
 14. Sun, Z. Z. *et al.* Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology. *J. Vis. Exp.* e50762 (2013). doi:10.3791/50762
 15. Shin, J. & Noireaux, V. Efficient cell-free expression with the endogenous E. Coli RNA polymerase and sigma factor 70. *J. Biol. Eng.* **4**, 8 (2010).
 16. Katzen, F., Chang, G. & Kudlicki, W. The past, present and future of cell-free protein synthesis. *Trends Biotechnol.* **23**, 150–156 (2005).
 17. Zaffiri, L., Gardner, J. & Toledo-Pereyra, L. H. History of Antibiotics. From Salvarsan to Cephalosporins. *J. Investig. Surg.* **25**, 67–77 (2012).
 18. Dewick, P. M. *Medicinal Natural Products*. (John Wiley & Sons, Ltd, 2009). doi:10.1002/9780470742761
 19. Kano, S. Artemisinin-based combination therapies and their introduction in Japan.

- Kansenshogaku Zasshi*. **88**, 18–25 (2014).
20. Paterson, I. & Anderson, E. A. CHEMISTRY: The Renaissance of Natural Products as Drug Candidates. *Science* (80-.). **310**, 451–453 (2005).
 21. Kumar, A. *Metabolic Engineering in Plants*. (springer, 2015). doi:10.1007/978-81-322-2283-5_26
 22. Wang, C., Zada, B., Wei, G. & Kim, S.-W. W. Metabolic engineering and synthetic biology approaches driving isoprenoid production in *Escherichia coli*. *Bioresour. Technol.* **241**, 430–438 (2017).
 23. Chang, M. C. Y., Eachus, R. A., Trieu, W., Ro, D.-K. & Keasling, J. D. Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. *Nat Chem Biol* **3**, 274–277 (2007).
 24. Elson, C. E., Maltzman, T. H., Boston, J. L., Tanner, M. A. & Gould, M. N. Anti-carcinogenic activity of d-limonene during the initiation and promotion/progression stages of DMBA-induced rat mammary carcinogenesis. *Carcinogenesis* **9**, 331–2 (1988).
 25. Elegbede, J. A., Elson, C. E., Tanner, M. A., Qureshi, A. & Gould, M. N. Regression of rat primary mammary tumors following dietary d-limonene. *J. Natl. Cancer Inst.* **76**, 323–5 (1986).
 26. Ciriminna, R. *et al.* Limonene: a versatile chemical of the bioeconomy. *Chem. Commun.* **50**, 15288–15296 (2014).
 27. Liu, T. & Khosla, C. A Balancing Act for Taxol Precursor Pathways in *E. coli*. *Science* (80-.). **330**, 44–45 (2010).

28. Xue, Y. *et al.* Engineering *Escherichia coli* for Enhanced Tyrosol Production. *J. Agric. Food Chem.* **65**, 4708–4714 (2017).
29. Lee, D., Park, H. L., Lee, S.-W. W., Bhoo, S. H. & Cho, M.-H. H. Biotechnological Production of Dimethoxyflavonoids Using a Fusion Flavonoid O-Methyltransferase Possessing Both 3'- and 7-O-Methyltransferase Activities. *J. Nat. Prod.* **80**, 1467–1474 (2017).
30. Kim, E. J., Cha, M. N., Kim, B.-G. G. & Ahn, J.-H. H. Production of Curcuminoids in Engineered *Escherichia coli*. *J. Microbiol. Biotechnol.* **27**, 975–982 (2017).
31. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Last 25 Years. *J. Nat. Prod.* **70**, 461–477 (2007).
32. Feher, M. & Schmidt, J. M. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *J. Chem. Inf. Comp. Sci.* **43**, 218–227 (2003).
33. de Montellano, P. R. Hydrocarbon Hydroxylation by Cytochrome P450 Enzymes. *Chem. Rev.* **110**, 932–948 (2010).
34. Sono, M., Roach, M. P., Coulter, E. D. & Dawson, J. H. Heme-Containing Oxygenases. *Chem. Rev.* **96**, 2841–2888 (1996).
35. Jung, S., Lauchli, R. & Arnold, F. H. Cytochrome P450: taming a wild type enzyme. *Curr. Opin. Biotech.* **22**, 809–817 (2011).
36. Seifert, A. *et al.* Rational Design of a Minimal and Highly Enriched CYP102A1 Mutant Library with Improved Regio-, Stereo- and Chemoselectivity. *ChemBioChem* **10**, 853–861

(2009).

37. Lewis, J. C. *et al.* Combinatorial Alanine Substitution Enables Rapid Optimization of Cytochrome P450BM3 for Selective Hydroxylation of Large Substrates. *ChemBioChem* **11**, 2502–2505 (2010).
38. Kim, D.-H., Ahn, T., Jung, H.-C., Pan, J.-G. & Yun, C.-H. Generation of the Human Metabolite Piceatannol from the Anticancer-Preventive Agent Resveratrol by Bacterial Cytochrome P450 BM3. *Drug Metab. Dispos.* **37**, 932–936 (2009).
39. Sawayama, A. M. *et al.* A Panel of Cytochrome P450 BM3 Variants To Produce Drug Metabolites and Diversify Lead Compounds. *Chemistry (Easton)*. **18**, 1199–1216 (2009).
40. Schückel, J., Rylott, E. L., Grogan, G. & Bruce, N. C. A Gene-Fusion Approach to Enabling Plant Cytochromes P450 for Biocatalysis. *Chembiochem* **13**, 2758–2763 (2012).
41. Ro, D.-K., Ehlting, J. & Douglas, C. J. Cloning, Functional Expression, and Subcellular Localization of Multiple NADPH-Cytochrome P450 Reductases from Hybrid Poplar. *Plant Physiol* **130**, 1837–1851 (2002).
42. Lamb, D. C., Kelly, D. E., Hanley, S. Z., Mehmood, Z. & Kelly, S. L. Glyphosate Is an Inhibitor of Plant Cytochrome P450: Functional Expression of *Thlaspi arvensaen* Cytochrome P45071B1/Reductase Fusion Protein in *Escherichia coli*. *Biochem Biophys Res Co* **244**, 110–114 (1998).
43. Zhao, F. *et al.* Optimization of a cytochrome P450 oxidation system for enhancing protopanaxadiol production in *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **113**, 1787–1795 (2016).

44. Zhu, S., Wu, J., Du, G., Zhou, J. & Chen, J. Efficient Synthesis of Eriodictyol from l-Tyrosine in *Escherichia coli*. *Appl Env. Microb* **80**, 3072–3080 (2014).
45. Vazquez-Albacete, D. *et al.* An expression tag toolbox for microbial production of membrane bound plant cytochromes P450. *Biotechnol. Bioeng.* **114**, 751–760 (2017).
46. Fasan, R., Meharena, Y. T., Snow, C. D., Poulos, T. L. & Arnold, F. H. Evolutionary History of a Specialized P450 Propane Monooxygenase. *J. Mol. Biol.* **383**, 1069–1080 (2008).
47. Cirino, P. C. & Arnold, F. H. Regioselectivity and Activity of Cytochrome P450 BM-3 and Mutant F87A in Reactions Driven by Hydrogen Peroxide. *Adv. Synth. Catal.* **344**, 932–937 (2002).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
50. Atkinson, H. J., Morris, J. H., Ferrin, T. E., Babbitt, P. C. & Madden, T. Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. *PLoS One* **4**, e4345 (2009).
51. Lupien, S., Karp, F., Wildung, M. & Croteau, R. Regiospecific Cytochrome P450 Limonene Hydroxylases from Mint (*Mentha*) Species: cDNA Isolation, Characterization, and Functional Expression of (–)-4S-Limonene-3-hydroxylase and (–)-4S-Limonene-6-hydroxylase. *Arch. Biochem. Biophys.* **368**, 181–192 (1999).

52. Kwon, Y.-C. *et al.* Integrating cell-free biosyntheses of heme prosthetic group and apoenzyme for the synthesis of functional P450 monooxygenase. *Biotechnol. Bioeng.* **110**, 1193–1200 (2013).