Université de Montréal

Accélération de l'exploration de l'espace chimique du cytochrome P450 BM3 par des méthodes de criblage à haut débit et bio-informatiques

par Olivier Rousseau

Département de Chimie
Faculté des Arts et des Sciences

Mémoire présenté à la Faculté des études supérieures et postdoctorales
en vue de l'obtention du grade de maître ès sciences (M.Sc.) en chimie

Août, 2018

Résumé

L'application de la chimie organique est prépondérante dans plusieurs secteurs industriels tels que le pharmaceutique, le cosmétique, l'alimentaire ou les produits ménagers. La chimie organique évolue sans cesse pour rendre les synthèses plus efficaces et moins dispendieuses. Cependant, l'amélioration de voies de synthèse se fait souvent sans égard à leur effet sur l'environnement. Ceci a entrainé l'émergence de la chimie verte pour réduire leur impact environnemental. À cette fin, l'une des stratégies employées est la biocatalyse : l'utilisation d'enzymes due à leurs propriétés 'vertes' et leur grande efficacité réactionnelle.

Les oxydases de type cytochrome P450 sont reconnues pour leur grande promiscuité, liée à leur rôle dans la détoxification de composés xénobiotiques et de molécules endogènes dans le corps. Cette superfamille d'enzymes accomplit une diversité de réactions d'oxydations en une seule étape telles que l'hydroxylation, l'époxidation, la désamination, la déshalogénation et la cyclopropanation. La P450 BM3 (P450 BM3) de *Bacillus megaterium* possède l'une des plus grandes activités catalytiques de sa famille, justifiant l'intérêt grandissant de ses applications industrielles. Ce mémoire se penche sur la nécessité d'améliorer les essais afin d'augmenter la capacité de criblage pour obtenir une plus grande proportion de variants capable de catalyser une réaction désirée.

Il a été proposé que la synthèse de l'indigo par les variants de P450 BM3 soit un indicateur de promiscuité envers de nouveaux substrats. Nous avons exploré l'espace mutationnel de P450 BM3 pour la synthèse de l'indigo en comparant différentes méthodes de criblage de librairies de variants. Dans un premier temps, des librairies de variants furent créées par mutagénèse par saturation de site. Cela permet d'obtenir une représentation uniforme des variants à chaque position sélectionnée plus rapidement et à un coût modique. Ensuite, des méthodes de criblage furent optimisées pour quantifier les variants de P450 BM3 et les cribler de deux façons. Nous avons comparé un essai colorimétrique direct grâce à l'absorbance de l'indigo avec un essai général indirect combinant la fluorescence du cofacteur NADPH et la spectrométrie de masse (LC-MS, GC-MS et LC préparative). Cette deuxième approche permet de cribler les variants indépendamment du substrat. Afin d'augmenter le débit du criblage, nous avons optimisé les méthodes en utilisant des plaques 96 puits avec une station de pipetage

automatisée. Nous avons découvert de multiples nouveaux variants de P450 BM3 synthétisant l'indigo dont une forte fraction de ces derniers s'est révélée aptes à synthétiser une molécule de valeur élevée en industrie : la cétone de framboise.

La corrélation de l'activité de P450 BM3 avec l'indigo et la cétone de framboise justifie l'exploration plus profonde de l'espace mutationnel de P450 BM3. La compréhension de cette corrélation permettrait de guider d'avantage l'ingénierie de P450 BM3 vers de nouveaux substrats. À cette fin, nous tentons d'expliquer la relation entre la structure, le dynamisme et l'activité des variants de P450 BM3. Les méthodes de criblage conventionnelles ne sont pas assez efficaces pour générer de grandes séries de données (« *Big Data* »). Nous appliquons le séquençage de nouvelle génération afin d'identifier $>10^4$ variants simultanément pour relier le génotype (ADN) au phénotype (synthèse d'indigo). L'analyse de ces données nécessite l'utilisation d'outils statistiques et informatiques tels que la modélisation, la dynamique moléculaire et l'analyse de composantes indépendantes du temps et de la structure (tICA). Nous observons une tendance commune des variants actifs de P450 BM3 représenté par un changement d'équilibre entre ses deux conformations majeures. Ce changement conformationnel pourrait expliquer la propension de P450 BM3 à synthétiser l'indigo.

Ce mémoire présente ainsi des méthodes permettant d'accélérer la découverte de variants actifs de P450 BM3 pour diverses réactions d'intérêt en industrie. Les méthodologies ci-développées pourront ensuite guider l'ingénierie d'autres systèmes enzymatiques.

Mots-clés : Analyse de composantes indépendantes du temps et de la structure, analyse statistique, automatisation, biocatalyse, cétone de framboise, criblage à haut débit, cytochrome P450 BM3, dynamique moléculaire, indigo, ingénierie de protéines, modélisation, mutagénèse, oxydation, séquençage nouvelle génération.

Abstract

Organic chemistry plays a major role in the pharmaceutical, cosmetic, food and household product industries. Industrial organic synthesis has constantly evolved to be more efficient and less cost-effective. However, those improvements were often made regardless of their impact on the environment, spawning the creation of the discipline of Green Chemistry. One among many strategies to make chemistry more innocuous to the environment is to apply biocatalysis: the use of enzymes for their 'green' properties and catalytic efficacy.

Cytochrome P450 oxidases are well known for their substrate promiscuity in detoxification of xenobiotics and endogenous molecules in the body. This superfamily of enzymes catalyzes a diversity of oxidation reactions in a single step, including hydroxylations, epoxidations, deaminations, dehalogenations and cyclopropanations. Cytochrome P450 BM3 (P450 BM3) from *Bacillus megaterium* is one of the most efficient of its family, justifying the growing interest for its industrial applications. This thesis considers the necessity to improve assays in order to increase the throughput of enzyme library screening and to improve the ratio of active variants.

Indigo synthesis has been proposed to be an efficient predictor of promiscuity towards novel substrates of P450 BM3. We explore the mutational space of P450 BM3 for indigo synthesis and compare screening methods. We made variant libraries using site-directed saturation mutagenesis to obtain a uniform distribution of mutations at each position and reduce cost and time. We optimized a method to quantify P450 BM3 variants and two different screening methods. We compared a direct colorimetric assay using indigo absorbance and an indirect assay using NADPH fluorescence coupled with mass spectrometry (LC-MS, GC-MS and LC prep). This second approach circumvents the need to develop an assay dependent of the substrate. To increase screening throughput, we optimized methods in a 96-well plate format and use an automated pipetting station. We confirm the discovery of new indigo-producing variants. We further demonstrate that indigo can serve as a strong predictor of raspberry ketone production, a high value industrial compound.

The correlation of the activity of P450 BM3 with indigo and raspberry ketone justifies the exploration of the fitness landscape P450 BM3. Understanding this correlation would accelerate engineering of P450 BM3 for reaction with novel substrates. We attempt to explain the relationship between the structure, dynamism and activity through deep-mutational scanning of P450 BM3. Conventional screening methods are not sufficient to produce large datasets (Big Data); we apply next-generation sequencing (NGS) to identify $>10^4$ variants simultaneously and to link the genotype (DNA) and the phenotype (indigo production). Analysis of NGS for enzyme engineering required the development of statistical and bioinformatics tools such as modeling, molecular dynamics and time-structure independent component analysis (tICA). Those methods allowed us to observe conformational perturbations that were shared between the active variants represented as a shift of equilibrium between two major conformations. This shift may explain the propensity of diverse P450 BM3 variants to synthesize indigo.

This master's thesis successfully demonstrates the accelerated discovery of engineered variants of P450 BM3 with activity for industrially relevant reactions. The methodologies we have developed contribute to knowledge on enzyme engineering and have the potential to be applied more broadly in other enzyme systems.

Keywords : Automation, biocatalysis, cytochrome P450 BM3, high-throughput screening, indigo, modeling, molecular dynamics, mutagenesis, next-generation sequencing, oxidation, protein engineering, raspberry ketone, statistical analysis, time-structure independent component analysis.

Table des matières

Liste des tableaux

Liste des figures

Liste des sigles et abréviations

4-PB.....4-phénylbutan-2-one
AMBER..... Assistant model building with Energy refinement
AMP .....Ampicilline
APCI.....Atmospheric-pressure chemical ionization
API ..... Ingrédients pharmaceutiques actifs
CFC .....Chlorofluorocarbone
CO .....Monoxyde de carbone
Da ..... Dalton
DDE.....Dichlorodiphényldichloroéthylène
DDT.....Dichlorodiphényltrichloroéthane
E.coli .....*Escherichia coli*
ESI..... Electrospray ionization
FACS ..... Fluorescence-activated cell sorting
FAD..... Flavin adenine dinucleotide
FMN ..... Flavin mononucleotide
GROMACS.....Groningen machine for chemical simulations
EE.....Excès énantiomérique
GC-MS .....Chromatographie gazeuse couplée à la spectrométrie de masse
IPTG.....Isopropyl β-D-1-thiogalactopyranoside
K_{cat}.....Constante catalytique
K_M .....Constante de Michaelis-Menten
LB.....Luria-Bertani
LC-MS..... Chromatographie liquide couplée à la spectrométrie de masse
m/z ..... Ration masse sur charge
MD .....Dynamique moléculaire
MM..... Mécanique classiques
MS ..... Spectrométrie de masse
NADPH .....Nicotinamide adénine dinucléotide phosphate
NAPQI.....*N*-acétyl-*p*-benzoquinone imine

NGS ..... Séquençage nouvelle génération
OD ..... Densité optique
P450 BM3 ..... Cytochrome P450 BM3
PAGE ..... Polyacrylamide gel electrophoresis
PCR ..... Réaction de Polymérisation en chaîne
PDB ..... Protein data bank
RK ..... Raspberry ketone
RMN ..... Résonance magnétique nucléaire
SDS ..... Sodium dodecyl sulfate
SF ..... Substrate-free
SB ..... Substrate-bound
QM-MM ..... Mécanique quantique et mécanique classique
TB ..... Terrific broth
tICA ..... time-structure Independent Component Analysis
TTN ..... Total Turnover Number
WGS ..... Whole genome sequencing
WT ..... Wild-type

"Happiness is like a butterfly; the more you chase it, the more it will elude you, but if you turn your attention to other things, it will come and sit softly on your shoulder."

-Henry David Thoreau

Remerciements

Ce mémoire représente évidemment tout le travail que j'ai accompli durant ces 6 ans et demi de baccalauréat et de maitrise, mais il témoigne aussi de toute la progression que j'ai faite au niveau personnel. Être expressif n'a jamais été l'un de mes points forts et je ne peux exprimer toute la gratitude pour les gens qui m'ont entouré durant ces années sans lesquelles je n'aurais JAMAIS pu accomplir tous ce travail.

Je sais exactement où commencer, avec mes plus grands supporteurs: mes parents. Vous avez toujours été là pour me conseiller et m'appuyer dans toutes les décisions que j'avais à prendre. Maman, même si je ne venais pas souvent à la maison, je sais que tu voyais tous les efforts que je mettais. Si j'avais la moindre question ou besoin, tu étais là pour y répondre. Papa, tu étais toujours là pour me remettre dans le droit chemin et garder des habitudes de vie saine. Vous m'avez permis de passer au travers des moments les plus difficiles et de me recentrer sur moi-même. Je sais que c'est grâce à votre fierté, amour et confiance que je suis rendu ici maintenant dans la vie. Émile, mon petit frère, es-tu surpris que je te mentionne aussi?! Je sais qu'on n'a pas passé beaucoup de temps ensemble durant ces années, mais chaque moment passé avec toi m'a permis de me ressourcer! Papi et Mamie, vous avez toujours suivi de près mes aventures académiques et je suis très reconnaissant pour votre support moral et financier sans lesquels j'aurais eu beaucoup de difficulté à me dévouer à mes ambitions. Merci à vous tous et toute la famille!

Simon, mon ami d'enfance, ma source d'inspiration! Toutes ces discussions, tous ces projets, tous ces moments passés avec toi m'ont permis de me rendre maintenant où je suis dans la vie. Le nombre de défis que l'on s'est donné et accompli… je ne peux même pas les compter tellement il y en a! Je sais que l'on va continuer dans notre perpétuelle quête d'aventure et d'exploration. Grâce à toi, j'ai fini ma maitrise, maintenant à toi de finir la tienne! BUSHIDO! Merci!

Kevin, mon ami d'enfance, ma source de relaxation! Je sais que depuis l'université je n'ai pas pu être aussi présent que je le voulais et je suis sûr que tu comprends. Malgré de longues périodes sans se voir et l'éloignement en distance, je sais qu'on ne s'est pas éloigné du tout! Tu ne t'en es peut-être pas rendu compte, mais les moments qu'on se voyait, étaient les moments que j'en

avais le plus besoin, qui me permettait de décrocher, prendre une pause et retourner aux sources. Cela m'a permis d'achever cette épreuve. Allo Vaness! Merci à vous deux!

Maintenant passons du côté universitaire et une bonne transition serait de parler de ceux qui ont transité dans la sphère personnelle de ma vie.

Max, tout a commencé en étant ton stagiaire et dès ce moment, ça l'a dérapé! Le nombre de défis que l'on s'est lancé autant stupides (bouffe à …, œufs, Soylent, ghost pepper) qu'intéressant (demi-ironman, bières, écoles d'été, Virtuenzyme) nous ont vraiment rapproché! Je suis vraiment content que notre symbiose professionnelle ait pu se transformer en amitié véritable et que l'on va continuer nos projets. Durant ces années, tu as été un modèle pour moi et à plusieurs occasions, tu as su me pousser hors de ma zone de confort pour sortir le meilleur de moi-même. Tu dois voir quelque chose en moi. Moi j'avais de la difficulté à le voir, ça doit être parce que je suis daltonien, vous m'avez diagnostiqué! En parlant de daltonisme, Jacynthe m'a fait réaliser que je n'avais pas vu la couleur orange dans mon CV qui était là depuis 10 ans… Cela me mène à Jacynthe. Tu as été d'une aide inestimable au laboratoire pour m'expliquer comment la biochimie fonctionne et que les bactéries ne volent pas. Tu as pu remplacer mes yeux d'homme qui ne trouvait jamais rien et ma mémoire de poisson rouge. Le meilleur conseil que j'ai retenu est qu'il faut toujours modifier un peu la recette, ça fonctionne tout le temps! On a toujours pu lâcher notre fou ensemble et avoir plein de conversations intéressantes, des sérieuses et d'autres un peu moins sérieuses! Désolé de ne pas avoir mangé les gummy bear aux betteraves, je sais que tu vas m'en vouloir toute la vie… Merci énormément à vous deux pour avoir rendu la vie à l'intérieur et à l'extérieur du laboratoire aussi extraordinaire!

Qu'est-ce que j'aurais pu faire sans la vie universitaire. J'étais très calme au début du BAC, jusqu'à ce que je rencontre, ce qui allait devenir, pour le meilleur et pour le pire, les gringos (Simon Laporte, François Fournel, Marc-Antoine Vaudreuil, Simon Forest et Martin Dufresne). Grâce à vous, je me suis ouvert à la vie universitaire et j'ai pu hériter du titre de Chouchou. Je n'oublierai jamais (en tout cas, celle que je me rappelle) toutes les soirées passées ensemble à la maison ou ailleurs. Ça m'a permis de vivre pleinement ma vie universitaire et j'ai vraiment trippé avec vous autres! J'ai aussi eu l'occasion de faire partie de l'AEDCUM en tant que secrétaire et faire plein de bénévolat pour les activités de notre association et de la FAECUM

qui m'a fait réaliser à quel point une communauté peut être solide. Merci à tous de m'avoir fait vivre cette expérience inoubliable, cela a grandement contribué à ma progression dans la vie.

Retournant au laboratoire, j'aimerais remercier tous les autres membres qui m'ont permis d'en apprendre autant et ont contribué à l'obtention de ce diplôme. Merci Sarah Abraham, même si tu as fini plus d'une année avant moi, j'étais très content d'avoir une autre chimiste à la maitrise avec moi dans ce domaine obscure. Un merci spécial à Daniela, pour avoir trouvé le temps entre ses articles et le MBA de m'aider dans la planification et la rédaction d'articles. Merci Saathanan Iathu… Iyattur… Iyathurai! de m'avoir supporté (dans le sens de m'endurer) en tant que ton superviseur de stage et d'avoir contribué à la recherche que je faisais. J'espère que tu arrives maintenant à 9h à ton travail. Merci à tous les autres membres que j'ai côtoyé et m'ont aidé dans mon parcours: Lorea Alejaldre, Sophie Gobeil, Nathalie Rachel, Sarah Ouadhi et Joaquim Guzman. Beaucoup d'autres à l'extérieur du laboratoire m'ont grandement aidé tel que le laboratoire de spectrométrie de masse. Merci à Louiza et Marie-Christine qui m'ont tout montré en MS et m'ont permis d'utiliser leurs instruments privés. Je te l'avais dit Louiza que ça fonctionnerait! Un grand merci à l'atelier sans qui les instruments que j'avais besoin n'auraient jamais fonctionnés. Un très grand merci à PROTEO qui a grandement contribué à ma carrière en me permettant de réaliser différents projets tels que les écoles d'été, un stage en Californie, Virtuenzyme et d'obtenir d'innombrables bourses.

J'aimerais remercier les collaborateurs qui ont contribué à mon projet : Prof. Sebastian Pechmann et Musa Ozboyaci pour leur très grande contribution sur mon second article. J'aimerais tout de même remercier Prof. Nicolas Moitessier en tant qu'ancien collaborateur pour m'avoir donné mon baptême de feu lors de ma première présentation de projet et m'avoir appris une réalité dans le domaine scientifique.

J'aimerais remercier ma directrice de recherche, Prof. Joelle Pelletier de premièrement m'avoir choisi comme stagiaire et ensuite m'avoir fait confiance pour la maitrise. Tu es un exemple de persévérance, assiduité et efficacité qui transparait sur tous les étudiants que tu formes. Tu m'as permis et encouragé de saisir toutes les opportunités qui se présentaient à moi. Je crois que tu as été témoin de mon développement personnel et professionnel au cours de ces années et tu as joué un rôle majeur dans cet accomplissement. Je sais que tu aurais aimé que je continue au doctorat, mais je continue, grâce à toi, vers de nouveaux défis!

Au final, il s'est passé énormément de choses dans ces 6 ans et demi et j'ai savouré chacun des moments que j'ai passé avec chacun d'entre vous. C'était une étape charnière de ma vie m'ayant permis de me découvrir et de m'améliorer en tant que professionnel et individu. J'en sors très fier, comblé et plus déterminé que jamais!

À tout le monde, je vous remercie du fond du cœur!

Chapitre 1: Introduction

1.1 La chimie verte au sein de la chimie fine et pharmaceutique

La chimie fine occupe une place prépondérante dans nos vies quotidiennes. Elle regroupe des molécules complexes utilisées dans la synthèse de médicaments, additifs alimentaires, fragrances, colorants, produits ménagers et plusieurs autres. Ces molécules ont souvent des structures complexes et sont obtenues par une synthèse difficile ne pouvant être produites qu'en petites quantités. Malgré ces désavantages, il y a un intérêt grandissant de la part de l'industrie à améliorer la production de ces derniers, car il y a de plus en plus de contraintes environnementales et légales.[1] Les chimistes ne commencent que depuis récemment à prendre conscience de l'impact de leurs synthèses sur l'environnement.

Auparavant, le manque d'informations sur la toxicologie d'une molécule et le manque de preuves concrètes et directes des effets néfastes qu'ils peuvent engendrer suggéraient une absence de problème. C'est le cas de l'utilisation du dichlorodiphényltrichloroéthane (DDT) comme pesticide depuis la Seconde Guerre Mondiale. Ce pesticide a joué un rôle majeur à cette époque dans l'éradication du typhus et malaria dans certains pays. Dû à son efficacité, il a continué à être utilisé de façon massive en agriculture jusqu'aux années 1970 pour atteindre une production de 2 millions de tonnes.[2] Toutefois, dû à l'augmentation d'observations et d'investigations sur la persistance et la toxicité du DDT et de son métabolite principal le dichlorodiphényldichloroéthylène (DDE), il a été banni dans la majorité des pays. En effet, ces polluants organiques persistants ont été détectés dans la chaine alimentaire, dans le lait maternel et même détectés dans la poussière et l'air des maisons.[3]

Les inquiétudes exprimées face aux niveaux du réchauffement climatique et de la pollution des différentes sphères (géosphère, hydrosphère, biosphère et atmosphère) sont encore plus préoccupantes. Par exemple, il a été démontré que le réchauffement climatique est, entre autres, engendré par la production de gaz à effet de serre.[4] Leur production ayant augmenté de façon flagrante depuis la révolution industrielle [5], cela affecte l'agriculture, la santé humaine, les écosystèmes et plus encore.[6] Une multitude de gaz y contribuent, dont les composés organiques halogénés. La persistance atmosphérique et l'effet sur le réchauffement climatique

de ces gaz ayant des fonctions commerciales ou industrielles n'ont été identifiés que trop tard, après une large utilisation. C'est le cas des chlorofluorocarbones (CFCs) utilisés comme réfrigérants depuis les années 30. Ce n'est qu'en 1970 qu'on a établi une corrélation entre la destruction de la couche d'ozone et les CFCs; le Protocole de Montréal a été ratifié en 1989 pour éliminer leur utilisation.[7] Grâce à ces initiatives, la production de CFCs est proche de zéro depuis 2006. Cependant, des chercheurs ont démontrés l'évidence de l'augmentation d'émissions de CFCs provenant de sources non déclarées ralentissant une fois de plus la régénération de la couche d'ozone.[8]

En tant que chimistes, il est de notre devoir de réévaluer et de mieux orienter nos choix lors d'une synthèse chimique. La résolution de cette problématique a amené à l'avènement de la chimie verte, car elle considère l'impact environnemental à la fois des produits chimiques finaux, ainsi qu'aux procédés pour arriver à ces derniers. Cette technologie durable a pris forme au début des années 90 aux États-Unis suite à la ratification du « *Pollution Prevention Act* ». Par la suite, le Royaume-Uni, l'Italie et le Japon ont lancé leur propre initiative en chimie verte.[9] Pour sa part, le Canada a débuté ses activités de recherche sur la chimie verte en 2000. T.H. (Bill) Chan, de l'université de McGill, a lancé la branche canadienne du « *Green Chemistry Institute* » qui fut rapidement succédé par la création du « *Canadian Green Chemistry Network* » en 2002 regroupant plusieurs chercheurs canadiens sur la chimie verte.[10]

Paul T. Anastas et Mary M. Kirchhoff ont été les premiers à cerner le terme chimie verte en donnant la définition suivante : Des technologies utilisant l'énergie et les matières premières (renouvelables de préférence) de façon efficace et réduire, jusqu'à éliminer, la production de déchets en évitant l'utilisation de réactifs et solvants toxiques.[11] Pour établir cette mentalité, les pionniers de la chimie verte ont défini 12 principes pour atteindre ces buts (Table 1-1).[12]

Table 1-1. Les 12 principes de la chimie verte

1	Prévention des déchets
2	Économie d'atomes
3	Conception de méthodes de synthèses moins dangereuses
4	Conception de produits chimiques plus surs
5	Solvants et auxiliaires moins polluants
6	Recherche de rendement énergétique
7	Utilisation de ressources renouvelables
8	Réductions du nombre de dérivés
9	Catalyse
10	Conception de produits en vue de leur dégradation
11	Observation en temps réel en vue de prévenir la pollution
12	Une chimie fondamentalement plus fiable

L'application de la chimie verte a donc le potentiel d'affecter de multiples étapes dans le processus de synthèse chimique. Premièrement, l'utilisation de solvants de moins en moins toxiques est une nécessité puisque le solvant constitue la principale composante en termes de masse dans une réaction. En effet, en industrie pharmaceutique, les solvants représentent plus de 80% du mélange réactionnel ainsi que 56% de l'énergie consommée lors de la synthèse d'ingrédients pharmaceutiques actifs (API). Ils sont également responsables de 50% des émissions de gaz post-traitement.[13-14] Plusieurs listes de classification ont été émises pour proposer des alternatives plus vertes ayant des efficacités semblables (Figure 1-1).[15-16] De plus, des méthodes de recyclage doit être priorisées.[17]

Suggérés	Utilisables	Indésirables
Eau	Cyclohexane	Pentane
Acétone	Heptane	Hexane(s)
Éthanol	Toluène	Diisopropyle éther
2-Propanol	Méthylcyclohexane	Éther diéthylique
1-Propanol	Méthyl *t*-butyl éther	Dichlorométhane
Acétate d'éthyle	Isooctane	Dichloroéthane
Acétate d'isopropyle	Acétonitrile	Chloroforme
Méthanol	2-Méthyltétrahydrofurane	N, N-Diméthylformamide
Butan-2-one	Xylènes	N-Méthylpyrrolidinone
1-Butanol	Diméthylsulfoxide	Pyridine
t-Butanol	Acide acétique	Acétate de diméthyle
	Éthylène glycol	Dioxane
		Diméthoxyéthane
		Benzène
		Tétrachlorométhane

Figure 1-1. Solvants priorisés par Pfizer lors de synthèses chimiques.
Figure adaptée de [16].

Deuxièmement, le rôle de la chimie verte n'est pas seulement de développer des alternatives vertes, mais aussi de réduire les déchets produits. Les principes d'économie d'atome [18] et de « *E(nvironmental) factor* » [19] ont donc été instaurés pour optimiser les procédés de synthèse en réduisant la quantité de matériel de départ et en augmentant les rendements pour ultimement réduire les déchets produits. Le « *E factor* » est une note allant de 0 à 100 (0 étant le meilleur) permettant de standardiser l'efficacité environnementale en calculant la quantité de déchets produits par unité de produit formé. Il tient compte de la quantité ainsi que de la nature des réactifs et solvants choisis permettant de guider les chimistes dans leurs décisions. Cette approche est souhaitable pour l'industrie, car elle permet de réduire le coût des matériaux de départ ainsi que les coûts de gestion des déchets découlant de leur séparation, traitement et destruction. En plus, l'industrie a la responsabilité de répondre aux pressions exercées par le marché.[20] Par exemple, l'ingrédient actif pharmaceutique Sertraline, contenu dans le Zoloft, est manufacturé par Pfizer et a généré 3 milliards de dollars canadiens en 2005. La révision de sa synthèse selon les principes de la chimie verte a permis d'éviter la production de 20 millions de kilogrammes de déchets pour réduire le E factor à 8. Cela leur a valu le prix du « *US EPA Green Presidential Award* » en 2002.[21]

Troisièmement, les rendements peuvent entre autres être améliorés par l'utilisation de catalyseurs métalliques. La modulation du métal de transition ainsi que ses ligands permet d'accéder à une grande gamme de réactions tels que l'isomérisation, l'addition, l'élimination et autres.[22] Puisque les molécules issues de la chimie fine deviennent de plus en plus complexes, une chimio-, régio- et stéréosélectivité est nécessaire. Une panoplie de catalyseurs métalliques à base de platine, ruthénium, rhodium, iridium et plusieurs autres peuvent accomplir ces réactions en conjonction avec les principes de la chimie verte.[23] Par contre, leur toxicité pose un problème en industrie pharmaceutique, car ils doivent être entièrement retirés lors du développement de médicaments.[24] De plus, ces métaux sont classés dans la catégorie des métaux précieux à haut risque, car ils ont des propriétés uniques, mais sont d'abondance limitée. Ces ressources ne sont donc pas durables et leur prix varie en fonction du marché et de leur rareté.[25] Finalement, leur utilisation peut nécessiter de hautes températures, de longs temps de réaction et un excès de réactifs.[26]

Pour pallier à ces problèmes, une multitude d'alternatives ont été développées telles que l'utilisation de métaux de transition de base (première rangée du tableau périodique), la combinaison de métaux de transition avec des liquides ioniques [27] ainsi que l'utilisation d'acides/bases solides.[28] Malgré les efforts déployés pour l'environnement en synthèse organique, les exigences de la chimie moderne deviennent toujours plus complexes et il sera toujours nécessaire de développer les voies de synthèse en accord avec la chimie verte. Par conséquent, la chimie verte ne peut que tendre vers une poursuite continue des procédés idéaux. Dans cette optique, une alternative en plein essor fera l'objet de ce mémoire : la biocatalyse.

1.2 La biocatalyse

1.2.1 Les avantages et limitations des enzymes

La biocatalyse consiste à utiliser les enzymes pour accélérer les réactions chimiques. Certains organismes contenant des enzymes sont utilisés depuis plus de 6000 ans pour la fermentation de la bière et du fromage [29-30]; les processus de fermentation sont toujours exploités à grande échelle, et sont définis par l'utilisation de microorganismes entiers. Cependant, l'incorporation d'enzymes isolées dans l'industrie de la chimie fine et pharmaceutique n'existe que depuis près

de 50 ans malgré son existence remontant jusqu'à 1894. Cela a eu une énorme influence sur la façon d'aborder la chimie de synthèse.[31] L'utilisation d'enzymes pour remplacer des étapes de synthèse organique est une approche de plus en plus convoitée non seulement pour leur attrait environnemental mais aussi pour leur efficacité. De plus, la diversité naturelle d'enzymes ayant évolué au sein des organismes de tous les règnes du vivant définit cette grande variété de réactions.

Les enzymes peuvent être classées dans 6 grandes classes : oxydoréductases, transférases, hydrolases, lyases, isomérases et ligases.[32] Ayant évolué pendant des millions d'années, la plupart des enzymes sont devenues spécialisées, c'est-à-dire très efficaces pour convertir en une seule étape une molécule qui peut être complexe et posséder de nombreux groupements ayant une réactivité similaire. Les enzymes peuvent donc présenter un avantage relativement aux synthèses chimiques multi-étapes, incluant souvent des étapes de protection/déprotection qui réduisent généralement l'efficacité de la synthèse. En plus, dans le cas des médicaments, il est important de contrôler l'excès énantiomérique (ee), car les impuretés énantiomériques peuvent être toxiques, réduire l'efficacité du médicament, ainsi qu'engendrer des coûts importants en lien avec les pertes et les étapes d'isolation de l'isomère désiré.[33] Par exemple, dans les années 50, la thalidomide était prescrite auprès des femmes enceintes pour traiter les nausées matinales. Ce médicament avait été jugé sécuritaire, mais il ne lui a fallu que quelques années avant de devenir l'un des plus grands désastres médicaux de l'histoire. En effet, son utilisation fut responsable de plus de 10 000 nouveaux nés ayant des malformations congénitales ainsi qu'une augmentation du taux d'avortements spontanés dont la moitié n'ont pas survécu. Ces effets étaient dus à l'isomère (*S*)- de la thalidomide qui, contrairement à l'isomère (*R*)- possédant les effets bénéfiques, est tératogène.[34] Il est à noter que la synthèse du composé énantiopur n'est pas un objectif raisonnable, car la thalidomide se racémise aisément dans l'eau. Néanmoins, ceci illustre l'importance qui doit être accordée à l'énantiopureté des composés pharmaceutiques.

Les enzymes peuvent accomplir des transformations chimio-, régio-, et stéréospécifiques avec un haut rendement. Certaines atteignent un taux de conversion exceptionnel telle l'anhydrase carbonique de nos poumons qui peut convertir 10^6 molécules de CO_2 en HCO_3^- par seconde.[35] De plus, les enzymes fonctionnent souvent sous des conditions douces et dans le solvant universel : l'eau. Par contre, cette incroyable efficacité est généralement jumelée à une forte

spécificité qui fait en sorte qu'une enzyme ne peut pas accommoder n'importe quel substrat. En effet, la spécificité d'une enzyme envers un substrat dépend de la structure du site actif ainsi que son dynamisme.[36] La modification de la spécificité est généralement accomplie par l'ingénierie de l'enzyme, où une évolution accélérée, en laboratoire, peut permettre à l'enzyme de réagir avec les substrats désirés. L'ingénierie d'enzymes permet également d'améliorer les problèmes de repliement, d'expression et de stabilité sous des conditions d'intérêt industriel. Enfin, des enjeux particuliers aux biotransformations se présentent lorsqu'il s'agit de les porter à grande échelle.[32] Les difficultés susmentionnées ont ralenti l'adoption industrielle des procédés biocatalysés. Toutefois, une panoplie d'avancées technologiques fait que nous vivons actuellement une période d'effervescence où des procédés biocatalytiques fiables, rapides et économiquement avantageux se font rapidement adopter par les multinationales de chimie.[37]

1.2.2 Applications de la biocatalyse en industrie

La biocatalyse a évolué en 3 vagues distinctes à travers le temps pour se rendre accessible au marché industriel. La première vague a débuté il y a plus de 100 ans lorsque l'on s'était rendu compte que le vivant pouvait servir à faire des réactions chimiques. Rosenthaler a effectué la synthèse de la (*R*)-mandelonitrile à partir de benzaldéhyde et d'hydrure de cyanure en utilisant un extrait de plante.[38] Ensuite, à partir de 1980-1990, la deuxième vague consistait à explorer la diversité des enzymes pour découvrir de nouvelles réactions inconnues à partir de l'étude des structures et du criblage d'organismes. Un exemple notable est l'hydratation de l'acrylonitrile pour synthétiser l'acrylamide à l'aide d'une hydratase de nitrile.[39] Cette amélioration synthétique a grandement contribué à l'expansion de ce polymère qui devrait atteindre un marché de 6,83 milliards USD en 2025.[40] Cette vague a été de courte durée, car les scientifiques se sont rapidement rendu compte de la portée limitée des enzymes naturelles.[41] C'est pourquoi dès la fin des années 90, Frances Arnold et Willem Stemmer ont révolutionné la biocatalyse en instaurant la troisième vague avec l'évolution dirigée qui sera décrite en plus de détail ci-dessous.[42] C'est cette dernière vague qui a permis la croissance fulgurante de la biocatalyse en industrie au cours des 10 dernières années. Par exemple, dans le domaine pharmaceutique, les 2/3 des 30 000 publications/brevets sur la biocatalyse furent publiés entre 2004 à 2014, démontrant bien la tendance de l'industrie (Figure 1-2).[43]

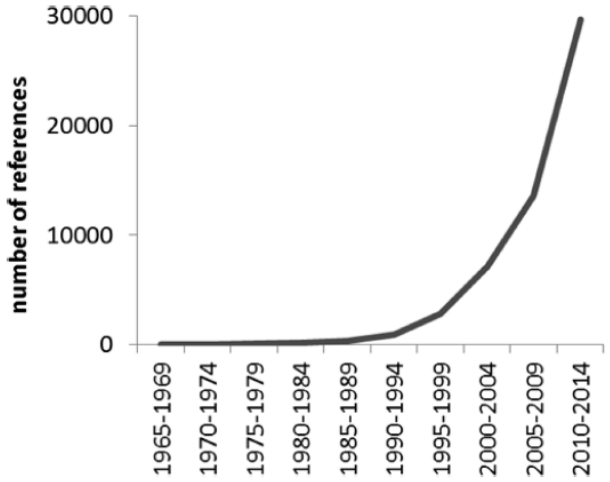


Figure 1-2. Évolution du nombre de publications et références sur la biocatalyse en industrie pharmaceutique.

Figure reproduite de [43].

En plus, les pressions publiques et politiques ont poussé les industries pharmaceutiques et les gouvernements à se pencher sur la question. En 2012, selon une étude réalisée auprès de 24 compagnies pharmaceutiques, une majorité d'entre elles planifiait d'augmenter l'utilisation d'enzymes dans leurs procédés.[44] La même année, le plus grand consortium européen publique-privé : « *Chemical Manufacturing Methods for the 21st Century (CHEM21)* » a été créé pour rendre la synthèse de médicaments plus écologique. En effet, 26.4 millions d'euros ont été investis dans ce projet regroupant six compagnies pharmaceutiques, cinq petites-moyennes entreprises et 13 universités au travers de l'Europe.[45]

Tous ces efforts ont engendré le développement d'une panoplie de biotransformations au travers du monde.[14, 33, 46-48] Un exemple important fut rapporté par Merck qui, en 2006, a mis sur le marché le médicament Sitagliptin (Januvia) pour le traitement du diabète de type II représentant des revenues de 3.86 milliards de dollars US en 2015.[49] La synthèse de ce médicament comprenait la réduction asymétrique d'une énamine, catalysée par le rhodium sous haute pression, allant à l'encontre des principes la chimie verte. Merck a fait appel à Codexis, un des leadeurs mondiaux en ingénierie de biocatalyseurs industriels, pour développer et intégrer une transaminase optimisée ((R)-ATA) (Figure 1-3). Le nouveau procédé élimine l'utilisation de rhodium, réduit les coûts, l'énergie utilisée, les déchets totaux de 19% et le nombre d'étapes de synthèse, tout en augmentant les rendements de 10-13% et la productivité

de 53%.[14, 50] Cette innovation leur a valu le « *Presidential Green Chemistry Award* » en 2010.

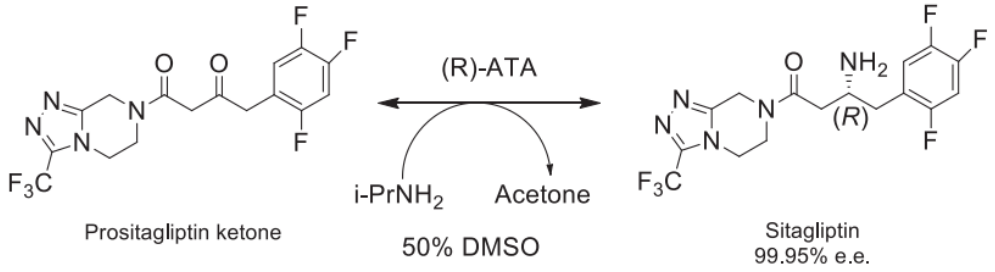


Figure 1-3. Synthèse de la sitagliptin par une transaminase modifiée par ingénierie ((R)-ATA).

Figure reproduite de [47].

Les réactions d'oxydation suscitent également un grand intérêt dans l'industrie pharmaceutique. Un exemple dans ce domaine a pour objet la diminution du taux de cholestérol par l'atorvastatin. Ce médicament a été le meilleur vendeur de tous les temps en générant plus de 125 milliards de dollars US.[51] Une réaction en deux étapes est effectuée pour synthétiser un intermédiaire de l'atorvastatin en utilisant 3 enzymes modifiées : une déhalogénase d'halogénohydrine (HHDH), une déshydrogénase de glucose (GDH) et une réductase de cétone (KRED). Les centres chiraux rend la synthèse organique beaucoup plus difficile justifiant l'utilisation d'enzymes (Figure 1-4).

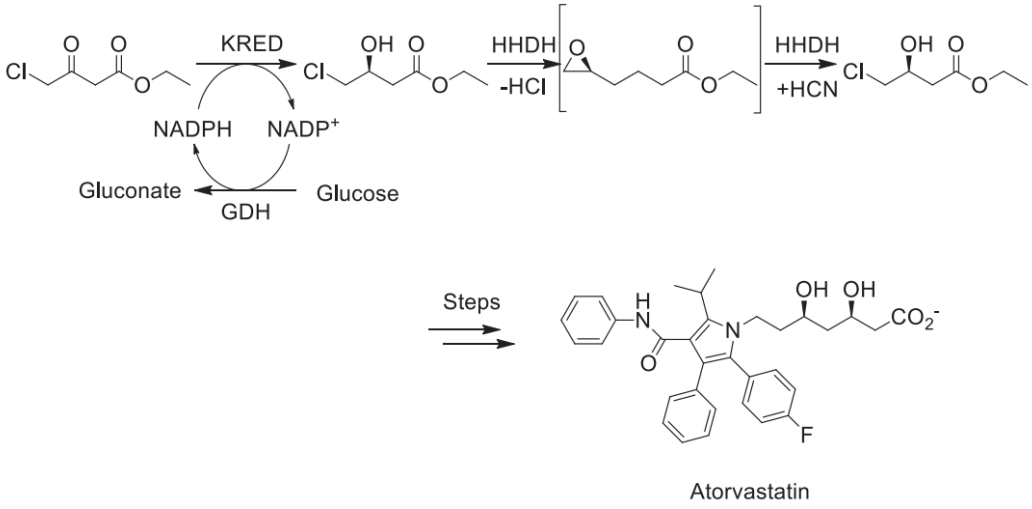


Figure 1-4. Synthèse de l'intermédiaire de l'atorvastatin.

Synthèse effectuée à l'aide de trois enzymes modifiée par ingénierie : une déhalogénase d'halogénohydrine (HHDH), une déshydrogénase de glucose (GDH) et une réductase de cétone (KRED). Figure reproduite de [47].

Le succès des réactions biocatalysées dans les synthèses industrielles en augmente la demande.[52] Toutefois, il demeure laborieux d'améliorer une enzyme pour une application donnée. Ce constat décrit le contexte dans lequel s'inscrit le sujet de ce mémoire : l'accélération de la découverte de nouvelles transformations enzymatiques pour l'industrie, illustrée à l'aide d'une oxygénase de type cytochrome P450.

1.2.3 Les enzymes cytochrome P450 : applications et mécanisme

La première enzyme cytochrome P450 (P450) a été détectée dans les années 1950 dans le foie de lapin.[53-54] Elle lie fortement le monoxyde de carbone (CO) traduit par un pic caractéristique d'absorption à 450 nm, permettant sa quantification directe.[55-58] Depuis, les P450 ont été étudiées de façon intensive car elles offrent une chimie unique de l'oxygène. Ces mono-oxygénases oxydent des carbones non-activés avec grande sélectivité et ce, en une seule étape. Elles catalysent une multitude de réactions d'oxydo-réduction incluant l'hydroxylation, l'oxygénation, la désalkylation, l'époxydation et plusieurs autres.[59-62] La famille des P450 est retrouvée dans tous les règnes ayant divergé à partir du même ancêtre commun.[63] Se retrouvant principalement dans le foie chez les mammifères, elles jouent un rôle dans la synthèse des hormones stéroïdiennes, le contrôle du métabolisme de vitamines et la transformation des acides gras insaturés.[64] De plus, elles contribuent à la détoxification de l'organisme en oxydant les médicaments et autres composés xénobiotiques pour mieux les excréter.

Les métabolites des médicaments peuvent faire preuve de toxicité, un facteur principal à considérer avant d'en amorcer le développement clinique.[65] La caractérisation des métabolites produits lors de l'exposition d'un médicament-candidat à des P450 purifiées permet de sonder le potentiel de toxicité rapidement et à faible coût.[66] À cet effet, plusieurs P450 optimisées pour métaboliser un large spectre de molécules sont commercialement disponibles. Par exemple, l'acétaminophène, est un des médicaments les plus consommés au monde dû à son efficacité analgésique et à son accessibilité en vente libre. Toutefois, il est la cause de plus de

30 000 hospitalisations et 300 décès par année aux États-Unis, des suites d'une hépatoxicité dûe à sa conversion par les P450 en métabolite *N*-acétyl-*p*-benzoquinone imine (NAPQI).[67-68]

Il est à noter que le criblage *in vitro* permet également la validation de pro-médicaments pouvant être activés par la réaction avec une P450.[69-70] Les P450 sont également directement utilisées dans la synthèse de composés pharmaceutiques, telles les statines utilisées pour la régularisation du taux de lipides et cholestérol élevés. C'est le cas de l'hydroxylation de la compactine par la $P450_{sca\text{-}2}$ chez *Streptomyces carbophilus* pour donner le médicament Pravastatin, utilisé depuis les années 1990 (Figure 1-5A).[71-72] En chimie fine, les P450 sont appliquées à la synthèse de molécules naturelles odorantes et gustatives dont l'abondance est faible, engendrant des coûts d'extraction élevés. C'est le cas, entre autres, du (+)-nootkatone, une saveur commerciale de grande valeur. Son extraction des pamplemousses, où elle est présente en faibles concentrations, est avantageusement remplacée par l'oxydation de la (+)-valencène, abondante chez les oranges, par la P450 CYP109B1 de *Bacillus subtilis* (Figure 1-5B).[73]

En plus des avantages pour l'environnement engendrés par ces conversions industrielles enzymatiques, les P450 peuvent également être appliquées directement à la détoxification l'environnementale (biorémédiation). L'utilisation massive de surfactants dans les produits chimiques et les produits de consommation courante est un réel problème puisqu'ils se retrouvent dans les eaux et sols. La dégradation de surfactants tels les alkylphénols éthoxylés génère principalement des nonylphénols étant des modulateurs du système endocrinien (Figure 1-5C). Une multitude de P450 provenant du champignon *Phanerochaete chrysosporium* dégradent cette classe de surfactants, réduisant leur toxicité.[74-75]

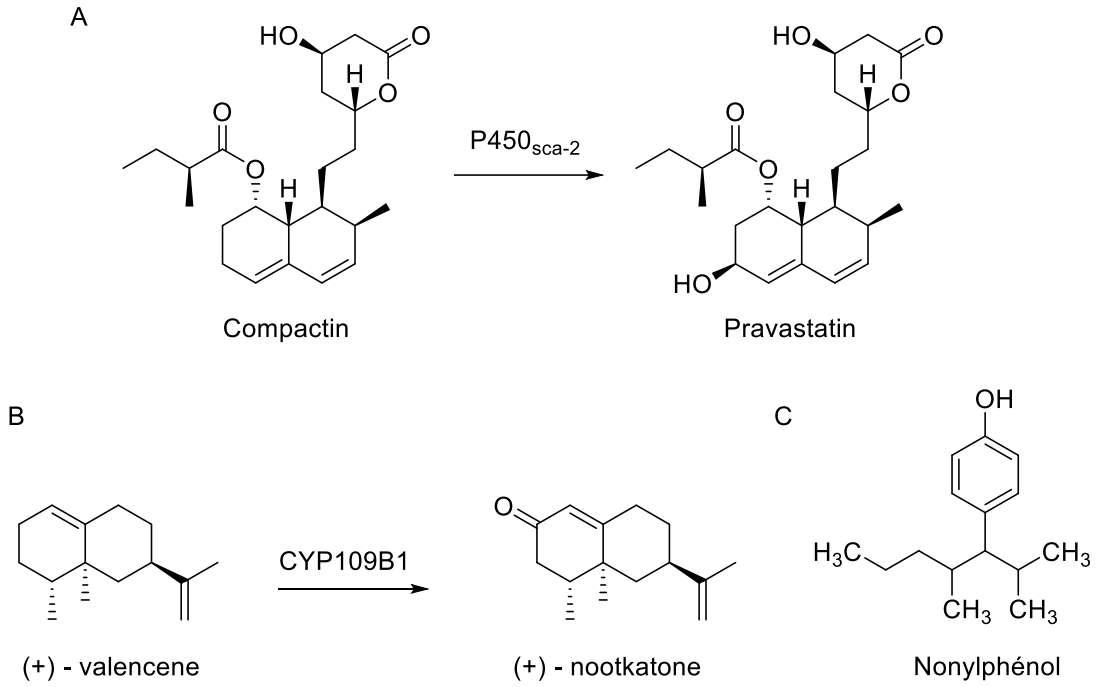


Figure 1-5. Exemples d'applications industrielles des P450.

A) Hydroxylation de la compactine en Pravastatin par la $P450_{sca-2}$. B) Oxydation de la (+)-valencène en (+)-nootkatone par la P450 CYP109B1. C) Le nonylphénol pouvant être dégradé par les P450 de P. chrysosporium.

Le vaste éventail d'applications des P450 découle du fait qu'elles ont évolué pour accepter une grande variété de réactifs. Malgré les différences entre P450, elles conservent néanmoins le même cœur catalytique: un hème ferrique lié par une cystéine qui catalyse les réactions d'oxydo-réduction.[76] Les P450 se différencient par le partenaire catalytique qui assure le transfert d'électrons, processus essentiel au cycle catalytique (Figure 1-6). Certaines assurent le transfert d'électrons par leur association non-covalente à une enzyme réductase de type fer-sulfure (ferredoxine) tandis que d'autres incluent un domaine FMN et un domaine FAD dépendant du cofacteur NAD(P)H.[77]

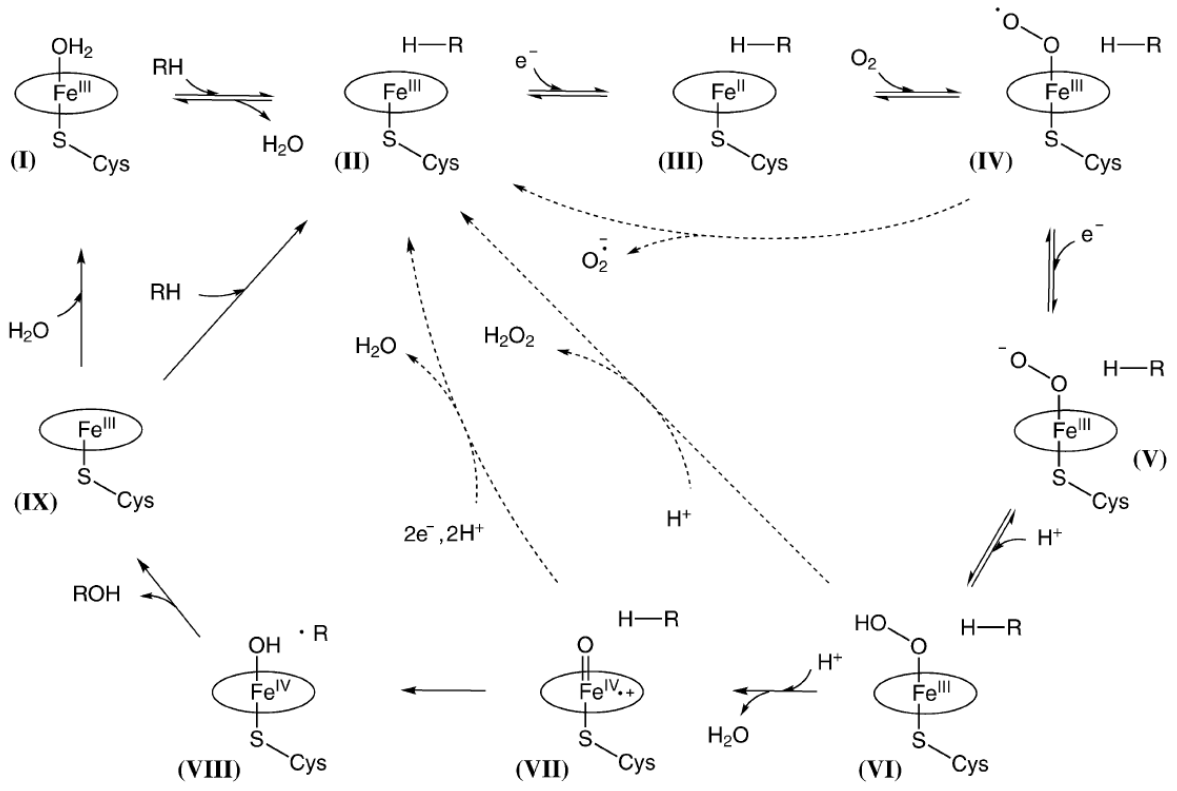


Figure 1-6. Cycle catalytique du cytochrome P450.
Figure reproduite de [76].

Le mécanisme catalytique des P450 a été l'objet de nombreux débats avant d'en arriver au consensus actuel (Figure 1-6). Brièvement, la liaison du substrat réduit (RH) déplace une molécule d'eau liée au Fe^{III} (étape II) et permet au fer de passer d'un état « low spin » à « high spin ». Un premier électron réduit le Fe^{III} en Fe^{II} (étape III). La liaison de l'oxygène peut donc avoir lieu (étape IV) et un second électron peut réduire l'oxygène (étape V). L'évacuation d'une molécule d'eau forme une espèce très réactive (étape VII) et permet d'abstraire un radical hydrogène au substrat (étape VIII). Le substrat est oxydé puis relâché (étape IX) et réduisant le fer à son état d'oxydation initial pour recommencer le cycle. Plusieurs phénomènes caractérisés par des évènements de découplage entre les différents états de l'hème peuvent entraver l'oxydation du substrat. Il peut donc y avoir retour non-productif à l'état initial par le relâchement d'un anion superoxyde, de peroxyde d'hydrogène ou d'eau (des étapes IV, VI et VII à l'étape II).

Les cytochromes P450 bactériennes sont généralement solubles tandis que celles provenant d'eucaryotes sont plus souvent membranaires. La solubilité des P450 bactériennes simplifie grandement leur surexpression, purification et cristallisation. Cela en facilite l'analyse structurale et mécanistique, ainsi que leur application en biotechnologie.[78] À cet égard, le cytochrome P450 CYP102A1 de *Bacillus megaterium* (P450 BM3) est un choix de prédilection : soluble, elle fait preuve d'une forte efficacité catalytique car son domaine hème est fusionné à un domaine réductase de type FMN/FAD, ce qui augmente l'efficacité du transfert d'électrons.[79]

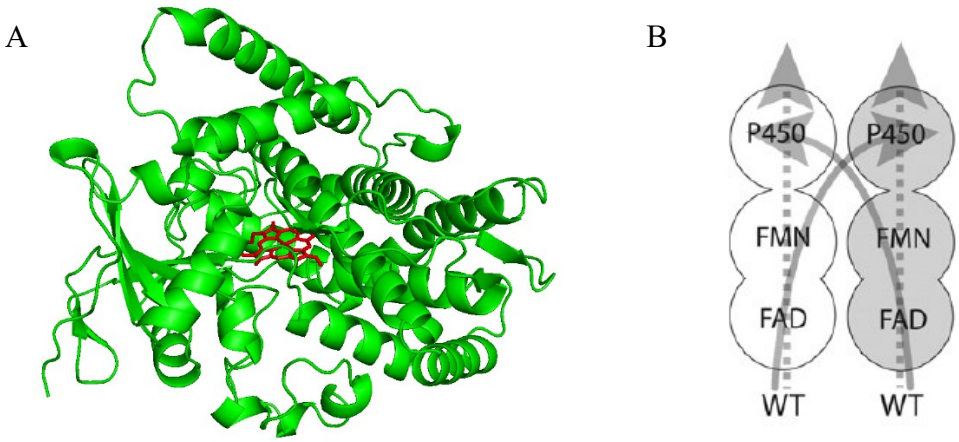


Figure 1-7. Le cytochrome P450 BM3

A) Représentation de la structure tridimensionnelle du domaine hème du cytochrome P450 BM3. Code PDB : 1BU7, protéine : ruban vert, hème : rouge. B) Démonstration de la dépendance de la dimérisation de P450 BM3 pour le transfert d'électrons se faisant d'un monomère à l'autre (*trans*).

P450 BM3 est un homodimère constitué de deux unités de 119 kDa (1049 acides aminés par monomère) (Figure 1-7A). Elle est active uniquement lorsque dimérisée car le transfert d'électrons s'effectue d'un monomère à l'autre (Figure 1-7B).[80] P450 BM3 possède l'une des plus grandes activités catalytiques de sa famille, avec un taux de roulement (TTN) de 17000 min^{-1} pour l'arachidonate.[81] Le rôle physiologique primaire de P450 BM3 n'est toujours pas élucidé, mais ses substrats naturels sont des acides gras possédant des chaînes alkyles de C_{12}-C_{18}. P450 BM3 jouerait tout de même un rôle de détoxification, puisque tout comme le reste de sa famille, elle catalyse une panoplie de réactions.[76] Cette propriété, se nommant la

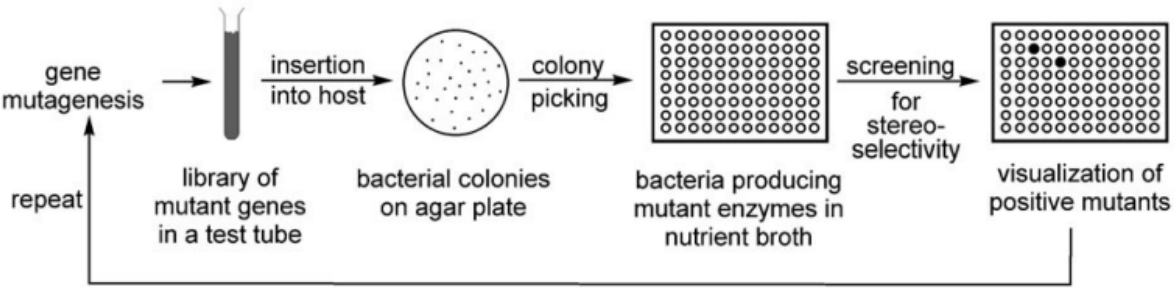
promiscuité, n'est pas commune chez la plupart des enzymes. Tel que mentionné précédemment, les enzymes évoluent généralement pour transformer un seul substrat de façon spécifique. La promiscuité de P450 BM3 peut s'expliquer, entre autre, par la particularité de son site actif semblable à une cavité se situant au bout d'un long tunnel. Cette disposition permet l'ancrage de la tête polaire des acides gras à l'entrée du tunnel et la pénétration de la chaine hydrophobe jusqu'au site actif afin d'y être oxydé de façon spécifique des positions ω-1, ω-2 et ω-3.[82] Ensuite, les P450 possèdent, en général, une cavité relativement grande à la suite de ce tunnel.[83-84] Cela augmente la mobilité du substrat, car il n'est pas contraint dans une seule orientation lors de la liaison. La spécificité de l'enzyme est aussi diminué, car il y a moins de résidus faisant contact avec le substrat dans le site actif.[76] Ces distinctions structurelles et mécanistiques démontrent la tolérance envers différents substrats possédant différentes formes et grandeurs en plus de pouvoir se lier dans multiples orientations. Pour ces raisons, P450 BM3 a été étudiée de façon intensive depuis plus de 40 ans.[85] La combinaison de son efficacité, de sa grande promiscuité, de sa facilité de modification ainsi que sa capacité à oxyder les carbones non-activés suscite énormément d'intérêt pour la biocatalyse industrielle et fait d'elle une cible de choix pour l'ingénierie d'enzymes. Nous croyons que la P450 BM3 peut remplacer de multiples procédés chimiques tel que démontré dans la prochaine section.

1.2.4 L'ingénierie du cytochrome P450 BM3

L'arrivée de la troisième vague en ingénierie des enzymes, soit le développement des méthodes d'évolution dirigée, a révolutionné la biocatalyse industrielle (Figure 1-8A). Il est actuellement possible d'incorporer des dizaines de mutations au sein d'une enzyme pour totalement changer sa sélectivité. Par exemple, la transaminase modifiée par Merck et Codexis pour la synthèse de la sitagliptine, présentée en section 1.2.2, comporte 27 mutations.[50] L'évolution dirigée consiste à reproduire l'évolution naturelle d'une protéine de manière accélérée, et en dirigeant l'évolution vers un but précis, tel que l'augmentation de stabilité ou l'activité pour un substrat donné. Cela se fait en naviguant le plus efficacement possible dans l'espace chimique de l'enzyme par itération de mutagénèse et de criblage pour un phénotype (propriété) désiré. L'espace chimique correspond à l'ensemble des possibilités de mutations au sein d'une enzyme; cet espace atteint rapidement une quasi-infinité de possibilités. Tel qu'illustré (Figure 1-8B), l'incorporation de mutations peut mener de façon directe (a) ou progressive (d) vers un

maximum global de la propriété désirée. Cependant, il arrive d'être pris dans un maximum local (b) ou bien diverger longuement avant d'obtenir une amélioration de la propriété d'intérêt (c).

A



B

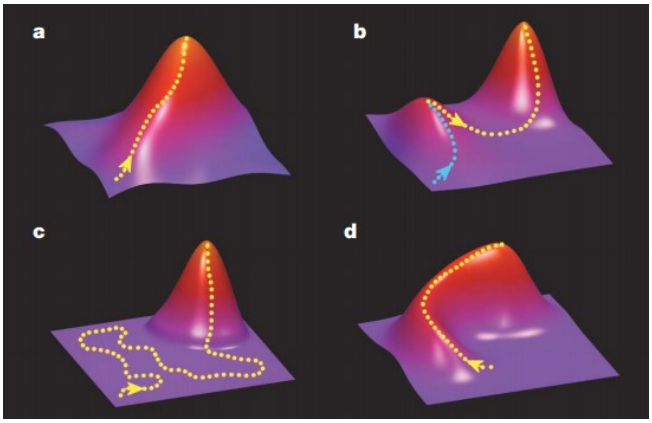


Figure 1-8. Évolution dirigée.

A) Cycle utilisé dans l'ingénierie de protéines pour améliorer une fonction. Le gène de l'enzyme désirée est muté de façon rationnelle et criblé dans un format haut débit. Le meilleur variant est ensuite utilisé pour le prochain cycle. Figure reproduite de [86]. B) Représentation des différentes possibilités lors de l'exploration de l'espace chimique d'une enzyme. Chaque point dans l'espace représente l'aptitude d'un variant à accomplir une fonction précise. Plus c'est élevé (rouge) plus elle est apte. Le chemin utilisé (pointillé jaune) définit la facilité avec laquelle une enzyme peut être modifiée par ingénierie. a. Chemin direct pour atteindre le maximum global. b. Atteinte d'un maximum local avant d'atteindre le maximum global. c. Évolution latente avant l'atteinte du maximum globale. d. Évolution ralentie, mais progressive vers le maximum global. Figure reproduite de [87].

Diverses approches existent pour explorer l'espace chimique d'une protéine. Les plus fréquentes sont la mutagenèse aléatoire et la recombinaison de fragments d'ADN (« *DNA shuffling* »). La mutagenèse aléatoire permet de créer d'énormes librairies sans avoir besoin de comprendre le mécanisme du système. Cependant cette méthode génère beaucoup de variants inactifs dû aux plus grands effets de la combinaison de mutations sur l'activité. Cela exige donc une grande capacité de criblage pour augmenter les chances de trouver des variants actifs. La recombinaison de fragments d'ADN, quant à elle, permet de recombiner, et ainsi augmenter, la diversité de séquences rapidement à partir des précédentes techniques.

Une approche différente, la conception rationnelle, consiste à modifier une protéine en se basant sur le mécanisme et la structure de l'enzyme. Bien que ciblée, elle permet aisément d'abroger une propriété mais difficilement d'en améliorer une. Néanmoins, la prédiction de mutations fructueuses devient actuellement plus efficace grâce aux outils bio-informatiques les plus performants. Ce sujet sera abordé dans la prochaine section. Finalement, ces méthodes peuvent être combinées selon le besoin pour accélérer la découverte de nouvelles fonctions et activités.

Un des principaux objectifs de ce projet de maitrise est de déterminer à quel point il est aisé – ou non – de modifier le site actif d'une enzyme connue pour réagir avec plusieurs substrats afin d'en améliorer la réactivité de réactions désirées. P450 BM3 sera le sujet de ces travaux et les effets d'une seule ronde de mutagénèse seront examinés. Les conditions requises pour cette adaptabilité ne sont pas encore tout à fait élucidées, mais la tolérance aux mutations est un atout pour l'acceptabilité de nouveaux substrats. Cela lui confère une variété de séquence tout en gardant son repliement et sa fonction catalytique.[88] Il a été découvert que le dynamisme de P450 BM3 influence grandement sa réactivité. En effet, P450 BM3 existe sous deux conformations majoritaires nommées sans substrat (SF : «*substrate-free*») et avec substrat (SB : «*substrate-bound*») (Figure 1-9A). Pour qu'un substrat se coordonne avec l'hème dans le site actif, il doit premièrement s'y rendre et ensuite déplacer une molécule d'eau. Le déplacement de cette molécule d'eau engendre le changement d'état d'oxydation de l'hème requis pour que la biocatalyse ait lieu. Cela cause aussi le bris de plusieurs ponts hydrogènes de cette molécule d'eau avec différents résidus de l'hélice I se trouvant au-dessus de l'hème (Figure 1-9B). Cette modification engendre la conversion de P450 BM3 à l'état SB correspondant à un abaissement des hélices G et F pour fermer le tunnel et exclure l'eau de la cavité. [76, 89-91] En plus du

tunnel du substrat, 9 autres tunnels ont été identifiés dans P450 BM3 ayant d'autres fonctions telles que le transport de l'oxygène et de l'eau (Figure 1-9C). Le mécanisme de P450 BM3 est donc complexe et démontre la difficulté avec laquelle il est dur de prédire l'effet d'une mutation dans cette dernière.

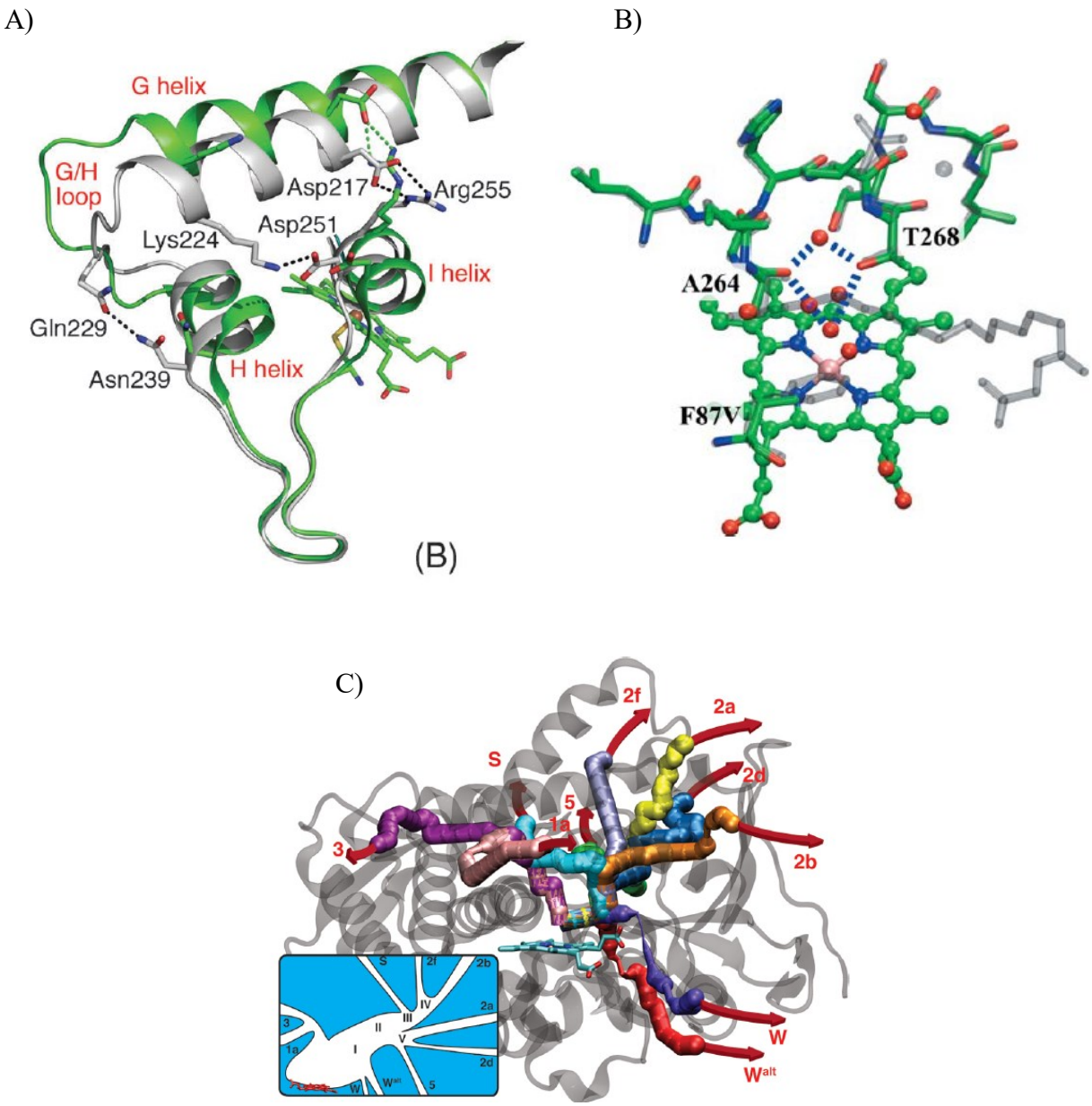


Figure 1-9. Illustrations des différentes caractéristiques structurelles de P450 BM3

(A) Comparaison de la conformation sans substrat (SF) et avec substrat (SB) de P450 BM3 démontrant certaines différences structurelles pour l'hélice G et l'hélice I. Figure reproduite de

[76]. B) Réseaux de ponts d'hydrogènes avec d'importantes molécules d'eau au-dessus de l'hème. Le déplacement de la molécule d'eau intéragissant avec A264 et T268 engendre le changement de conformation en SF et SB. Figure reproduite de [89]. C) Représentation des tunnels identifiés dans P450 BM3 ayant différents rôles tels que substrat, eau ou oxygène. Figure reproduite de [92].

Une revue exhaustive des mutations entreprises sur P450 BM3 jusqu'en 2012 répertorie 350 mutations individuelles à 167 positions dans le domaine hème, regroupant 4 substrats naturels et 32 substrats non naturels.[76] La majorité de ces mutations ont été entreprises dans le contexte de la modification de la réactivité de P450 BM3. Un des exemples les plus impressionnants concerne l'ingénierie de P450 BM3 afin qu'elle effectue la cyclopropanation de styrènes. Cette réaction est fort convoitée, entre autres, dans l'industrie pétrochimique et de polymères. Classiquement, elle dépend de métaux de transition dont les ressources mondiales s'épuisent et pourrait donc être avantageusement remplacée par une réaction enzymatique.[93] Des variants de P450 BM3 furent créés au laboratoire de Frances Arnold et criblés pour la cyclopropanation. Plusieurs variants furent identifiés, donnant des rendements jusqu'à 65% en plus d'offrir différentes diastéréosélectivités et des énantiosélectivités allant jusqu'à 95%. Ainsi, plutôt que de catalyser une réaction d'oxydation, les variants de P450 BM3 catalysent un transfert de carbène. Les variants pouvaient contenir jusqu'à 23 mutations à des positions proche ou loin du site actif.[94] Cet ouvrage démontre qu'il est possible de moduler et améliorer P450 BM3 afin d'exploiter son centre réactif d'une façon différente (sans O_2) pour donner lieu à une chimie jamais observée dans la nature.

Toutefois, dans le contexte de ce mémoire, nous examinerons des réactions d'oxydation semblables aux réactions natives de P450 BM3. Une application importante des variants de P450 BM3 accomplissant des réactions d'oxydation est la synthèse de métabolites de médicaments. Ces métabolites ont une grande valeur en synthèse afin d'améliorer les propriétés de têtes de série telles que l'activité, la solubilité, la toxicité, la stabilité et biodisponibilité. Cependant, ces métabolites sont nombreux et souvent de synthèse complexe car ils résultent justement de l'action des P450 hépatiques. Déjà en 2012, P450 BM3 avait été utilisée pour synthétiser les métabolites connus de plus de 24 différents médicaments comprenant des grands vendeurs tels que l'acétaminophène, le citalopram et le naproxen.[95] Depuis, la liste de médicaments ne fait que s'agrandir et permet de découvrir de nouvelles réactions.[96-97] Ces

médicaments ayant des structures diversifiées, témoignent de la promiscuité de P450 BM3. Notons que ces composés contiennent majoritairement des cycles aromatiques.

Ainsi, nous émettons une première hypothèse qu'il est possible d'identifier de multiples mutants étant capables de catalyser une même réaction. Dans ces travaux, nous explorons une réaction qui s'apparente au métabolisme des médicaments, soit l'hydroxylation de l'indole en indoxyle. Cette réaction permet le suivi aisé d'une oxydation aromatique puisque l'indoxyle produit se dimérise spontanément en indigo suite à son oxidation par l'air (Figure 1-10A). L'indigo est l'une des plus anciennes teintures. Extraite de multiples plantes telles que *Indigofera*, *Strobilanthes* et *Isatis*, son utilisation remonte à plus de 7800 ans.[98-100] Sa synthèse organique, entreprise en 1878 et commercialisée en 1890 par Adolf Baeyer, lui a valu le prix Nobel de chimie en 1905.[101] La demande croissante en fin des années 90, due entre autres à la popularité des jeans, a promu le développement d'une voie enzymatique pour remplacer les réactifs et déchets toxiques. Par biologie synthétique, *Escherichia coli* a été développée pour exprimer une tryptophanase et une dioxygénase de naphtalène afin de synthétiser l'indigo à partir du glucose, une molécule de départ accessible et peu couteuse.[102] En plus, une hydrolase d'isatine a été incorporée pour diminuer la production d'indirubine, un sous-produit de couleur rouge.[103]

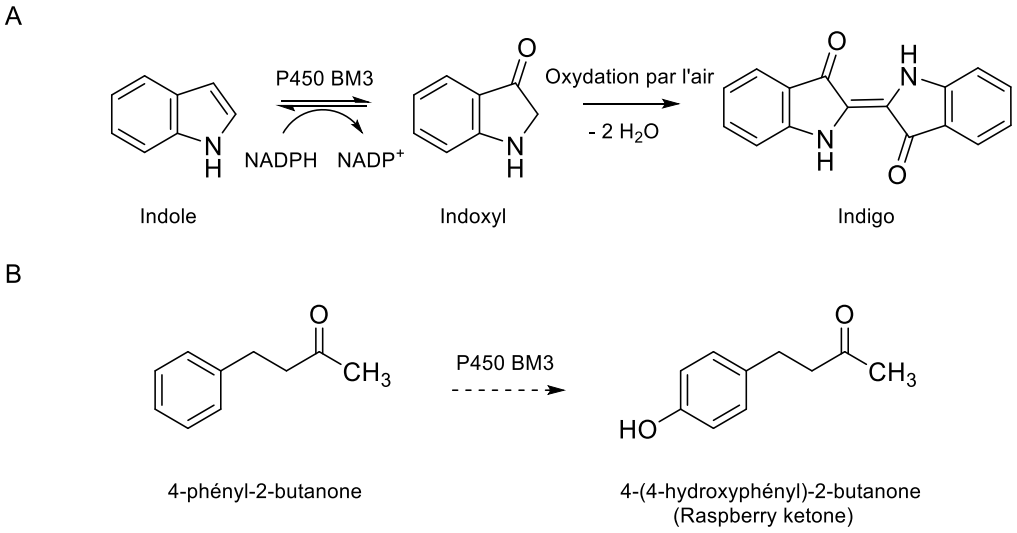


Figure 1-10. Schéma des biosynthèses explorées dans ce mémoire.

A) Biosynthèse de l'indigo catalysée par l'enzyme cytochrome P450 BM3. B) Biosynthèse hypothétique de la cétone de framboise catalysée par l'enzyme cytochrome P450 BM3.

À la même époque, il avait été observé que les cultures bactériennes coexprimant la P450 2E1 et sa réductase démontraient une formation de couleur bleue, suggérant que la P450 oxydait l'indole provenant de la voie métabolique du tryptophane.[104-105] Depuis, plusieurs se sont penchés sur la production de l'indigo par les P450, et particulièrement par P450 BM3. Par exemple, l'évolution dirigée de P450 BM3 a relevé que certaines substitutions aux positions F87, L188 et A74 produisaient de l'indigo et leur combinaison pour donner le triple mutant F87V/L188Q/A74G en augmente l'activité.[105-106] La structure cristalline du variant F87V, une position qui se situe directement au-dessus de l'hème dans le site actif, a révélé la création d'une cavité qui change la spécificité de l'enzyme.[96] Au contraire, les variants A82W et A82F, également fortement actifs pour la synthèse d'indigo, referment une cavité dans le site actif (Figure 1-11).[107] Ces exemples démontrent à quel point il est difficile de prédire l'effet spécifique d'une mutation sur une activité donnée, et donnent raison à l'approche de criblage de banques de variants. Notons qu'au départ de ces travaux, certaines sources proposaient que P450 BM3 natif ne produit pas d'indigo [105-106] alors que d'autres indiquaient une faible production d'indigo [107-109], observation qui sera vérifiée dans ce mémoire.

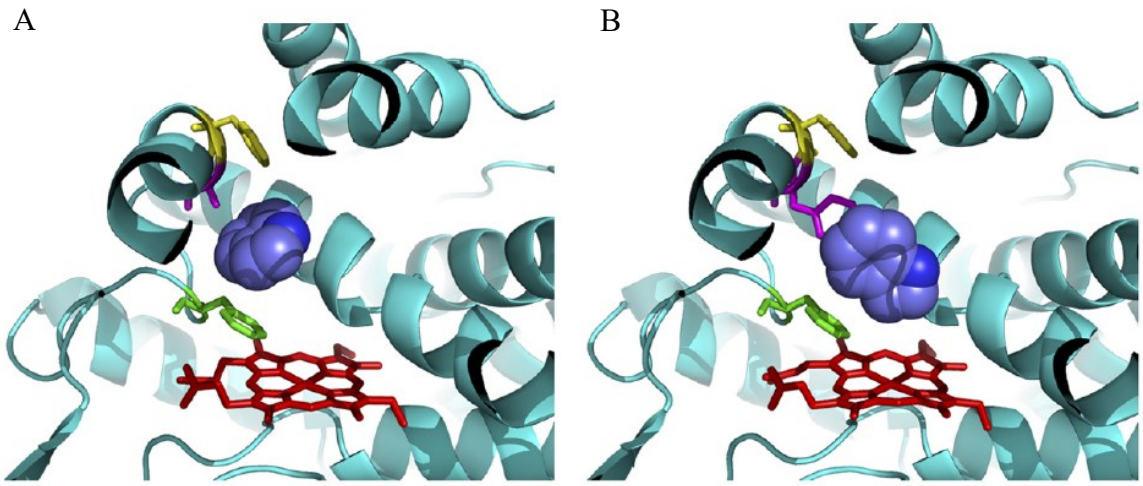
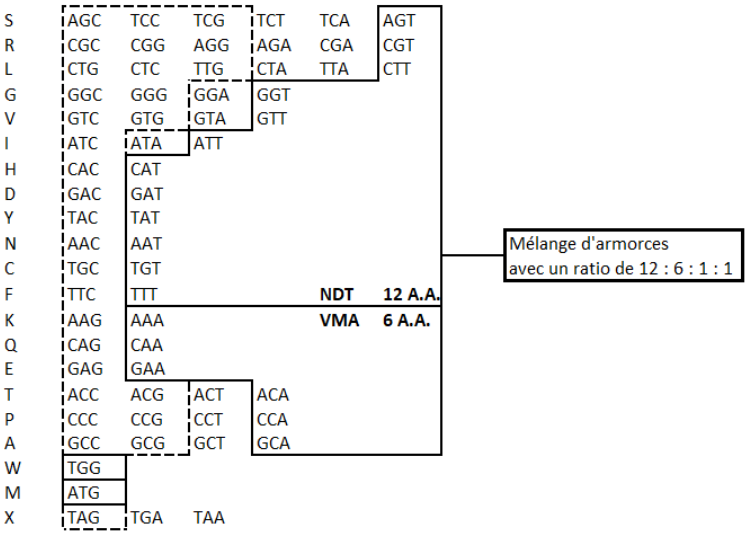


Figure 1-11. Représentation de l'arrimage moléculaire de l'indole dans le site actif de la P450 BM3.

A) Arrimage moléculaire de l'indole dans P450 BM3 natif (PDB code : 1JPZ). B) Arrimage moléculaire de l'indole dans le variant P450 BM3 A82F (PDB code : 2UWH). La chaine latérale de A82/F82 est représentée en rose, celle de F87 en vert, l'indole en sphère bleue et l'hème en rouge. Figure reproduite de [107].

L'espace chimique à explorer pour une protéine de la taille d'une P450 étant quasi-infini, il est impossible de créer l'ensemble des mutations et encore moins de toutes les cribler. En effet, la création de chaque mutation ponctuelle constituerait une banque de 20^{1049}, chiffre plus grand que le nombre d'atomes dans l'univers. Nous proposons de créer une librairie suffisamment grande pour y obtenir une représentation statistique valide des effets de diverses substitutions, en demeurant à l'intérieur des contraintes permettant la caractérisation exhaustive. Dans la première phase de ces travaux, nous avons donc ciblé cinq positions par saturation, c'est-à-dire de tenter d'obtenir des variants avec chacune des 19 substitutions ponctuelles à chacune des cinq positions, pour un maximum de 95 variants. Nous nous sommes penchés sur les positions A82, F87, V78 A264 et T268 du site actif, la probabilité d'avoir un impact positif sur l'activité étant plus élevée lorsqu'on cible le site actif.[110] Les raisons pour ces choix spécifiques seront élaborées au chapitre 2. Pour ce faire, nous avons utilisé un mélange précis d'amorces partiellement dégénérées (NDT, VMA, TGG et ATG) nous permettant théoriquement d'obtenir une distribution uniforme de chaque mutation à chaque position sélectionnée (Figure 1-12A).[111] Un mélange NDT, où N encode pour A, T, C, G et D encode pour A, T, G se traduit en 12 acides aminés. Pour sa part, VMA, où V encode pour A, C, G et M encode pour A, C se traduit en six autres acides aminés. Le tout est complété par ATG pour la méthionine et TGG pour le tryptophane. En comparaison avec d'autres méthodes, telle l'utilisation d'une seule amorce entièrement dégénérée (encodant A, C, G et T en mêmes proportions aux positions ciblées) dite 'NNN', celle-ci accélère l'obtention de la librairie complète désirée en réduisant la redondance de codons et en éliminant l'inclusion de codons stop. En effet, pour obtenir une couverture de 95%, il faut seulement 36 colonies pour viser 12 mutations par la méthode NDT, tandis qu'il en faut 192 pour viser l'ensemble des mutations par la méthode NNN. Plus on incorpore de mutations simultanément par ces méthodes, plus il est avantageux d'utiliser la méthode NDT (Figure 1-12B).[112]

A



B

Nombre de codons saturés simultanément	Type de codons dégénérés utilisés dans mutagénèse par saturation de site					
	NNN		NNK		NDT	
	Diversité	Clones requis pour une couverture de 95%	Diversité	Clones requis pour une couverture de 95%	Diversité	Clones requis pour une couverture de 95%
1	**64**	**192**	**32**	**96**	**12**	**36**
2	**4096**	**12288**	**1024**	**3072**	**144**	**432**
3	262,144	786,432	32,768	98,304	1728	5184
4	1.68×10^7	5.03×10^7	1.05×10^6	3.15×10^6	20,736	62,208
5	1.07×10^9	3.22×10^9	3.36×10^7	1.01×10^8	248,832	746,496

N = adénine [A], thymine [T], cytosine [C], ou guanine [G]
K = guanine [G] ou thymine [T]
D = adénine [A], guanine [G] ou thymine [T]

Figure 1-12. Utilisation de codons dégénérés pour la mutagénèse dirigée par saturation.

A) Chaque acide aminé peut être naturellement encodé un ou plusieurs codons. Un mélange de codons optimisés NDT, VMA, ATG et TGG permet d'avoir une distribution uniforme lors de la mutagénèse par site-dirigée. N représente les 4 nucléotides, D représente G, T, A, V représente A, C, G et M représente A et C. Figure adaptée de [111]. B) Comparaison des méthodes de saturation NNN, NNK et NDT pour démontrer l'effet de la redondance sur le nombre de colonies à cribler pour obtenir une couverture de 95%. Figure adaptée de [112].

L'exploration de la production d'indigo par l'ingénierie de la P450 est facilitée par la simplicité de l'essai colorimétrique direct pour caractériser les variants. Ce criblage peut se faire facilement par inspection visuelle des colonies bactériennes sur Pétri exprimant chacune un variant, puisque la colonie devient bleue s'il y a oxydation d'indole à indoxyle, puis formation spontanée d'indigo. Par contre, cette approche n'est que qualitative et ne nous permet pas de quantifier

l'indigo formé. À cet effet, nous avons optimisé un essai de criblage en liquide à haut débit, en plaque 96 puits à l'aide d'un robot Beckman Coulter Biomek NX^P (Figure 1-13). Dans le même ordre d'idées, l'expression des variants de P450 en bactéries et la lyse chimique ont été optimisées dans un format de plaque 96 puits afin d'automatiser au maximum le processus.

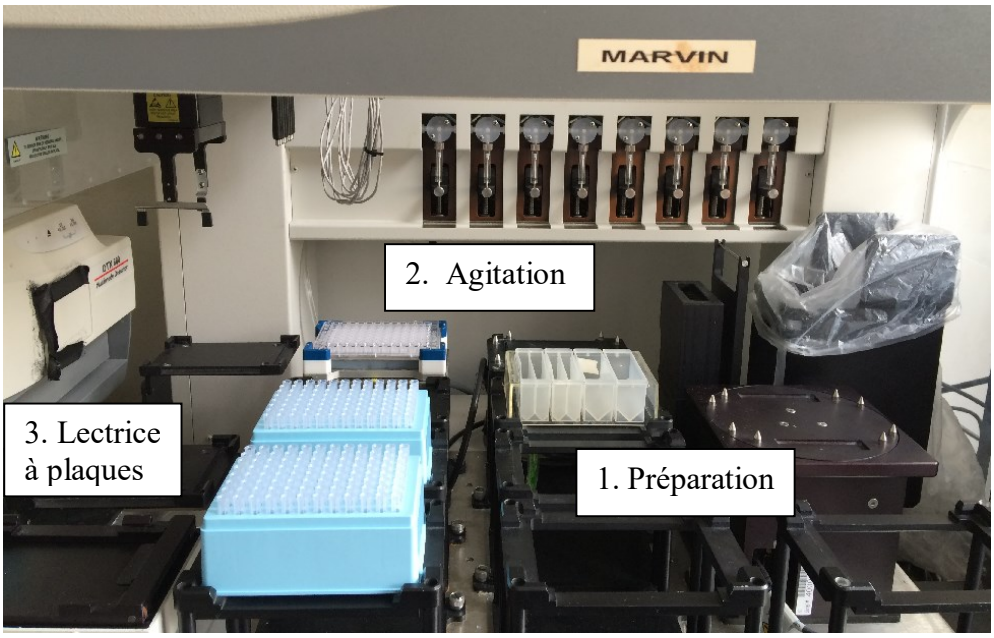


Figure 1-13. Disposition de l'expérience de criblage à haut débit sur le robot Beckman Coulter Biomek NX^P.

1. La réaction est préparée à l'aide du robot en ajoutant tous les réactifs et tampons dans une plaque 96 puits. 2. La réaction est agitée sur un module. 3. Une mesure analytique est prise à l'aide de la lectrice à plaques à chaque intervalle de temps.

En plus de donner lieu à un produit désirable, les variants de P450 producteurs d'indigo ont une utilité additionnelle : il a été observé qu'ils affichent en général une plus grande promiscuité pour d'autres substrats non naturels. Cette approche a permis d'identifier de nouveaux variants actifs pour la synthèse de dérivés d'éthylebenzène et de diphényleméthane dans le système $P450_{CAM}$.[113-114] Tandis que pour P450 BM3, de nouveaux variants ont été identifiés qui produisent des métabolites de coumarines indiquant une promiscuité s'apparentant aux P450 humaines.[115]

De ce fait, nous émettons une seconde hypothèse que l'indigo peut être utilisé en tant que prédicteur pour la découverte de variants actifs pour une seconde réaction d'intérêt, soit la synthèse de la cétone de framboise (Figure 1-10B). La cétone de framboise est utilisée massivement dans les secteurs industriels des saveurs et fragrances. L'extraction de la molécule naturelle étant très dispendieuse (20000$/kg), une alternative compatible avec les secteurs alimentaires et cosmétiques est de mise.[116] Les oxydations aromatiques étant accessibles avec les P450, nous désirions cribler les variants actifs en synthèse d'indigo envers cette seconde réaction. Cependant la réaction de la 4-phénylbutan-2-one à la cétone de framboise ne donne pas lieu à un produit coloré ou fluorescent. Nous avons donc optimisé une seconde méthode générale de criblage à haut débit utilisant le changement d'absorbance qui accompagne la consommation du cofacteur NADPH.[117] Les variants consommant le NADPH plus rapidement en présence de substrat ont été retenus pour l'identification de l'indigo par LC-MS et de la cétone de framboise par GC-MS, cette dernière donnant lieu à un patron de fragmentation inattendu en LC-MS. Une fois la masse et le temps de rétention confirmés pour le produit d'oxydation de la 4-phénylbutan-2-one, la position de l'hydroxylation fut confirmée par RMN.

1.3 L'application d'outils informatiques en biocatalyse

1.3.1 Les mégadonnées et le séquençage nouvelle génération

Indépendamment du domaine, une explosion de l'utilisation du mot mégadonnées («*Big data*») a débuté en 2010 grâce au développement des technologies informatiques.[118] La masse de données générées demande des ressources croissantes pour en faire l'analyse et l'entreposage. En biologie, le séquençage d'ADN de nouvelle génération (NGS) a permis de construire d'énormes séries de données pour le suivi d'études cliniques sur le cancer, le séquençage de génomes complets de nombreux organismes, le profilage du transcriptome, l'épigénétique et autres. Ces méthodes permettent actuellement de séquencer le génome humain pour 1000$, soit 50 000 fois moins que le coût du Projet du Génome Humain complété en 2000. Ces bases de données peuvent atteindre des dizaines de Pétaoctets (10^6 Gigaoctets), rendant le développement de nouvelles méthodes puissantes d'analyse impératif.[119-120]

Dans un contexte d'ingénierie de protéines, le séquençage 'classique' d'échantillons individuels par la méthode de Sanger est de mise lorsque la taille des librairies demeure petite. Cependant, l'augmentation de la capacité de mutagenèse et de criblage fait en sorte que cette méthode devient rapidement trop couteuse. Saisissant l'opportunité, le NGS peut être appliqué au séquençage de librairies massives de variants en ingénierie de protéines, permettant de relier le génotype au phénotype à plus grande échelle. Cette approche, dénommée « *deep mutational scanning* », a été utilisée pour la détection de mutations stabilisantes [121-122] et d'interactions protéine-protéine [123-125]. Cette approche avancée est actuellement appliquée en industrie à l'ingénierie de protéines appliquées à la biocatalyse mais n'a que très peu été rapportée dans la littérature.

Nous avons donc appliqué cette méthode à la découverte de variants de P450 BM3 actifs pour la biotransformation de l'indigo tel que décrit ci-dessus, mais à une plus grande échelle. Nous avons identifié et muté 49 positions dans un rayon de 20 Å au-dessus du site actif de P450 BM3, excluant les 5 positions mutées ci-dessus et incluant des mutations doubles et triples à des positions consécutives. La librairie résultante, qui comprend théoriquement plus de 7000 variants, comprend des mutations plus éloignées que la première sphère (*first shell*) entourant le site actif; ceci nous permettra d'observer à quel point les mutations plus éloignées peuvent avoir un impact significatif sur l'activité catalytique, tel que rapporté pour d'autres systèmes.[126-128] Pour un criblage de cette envergure, nous avons criblé chaque ensemble de variants (un ensemble = variants à une position, ou à un groupe de positions consécutives) en bactéries par l'observation de la couleur de l'indigo formé. Les colonies bactériennes furent récoltées en trois groupes: blanc (aucune activité), bleu pâle (activité faible) et bleu foncé (activité forte).

Le principal avantage du NGS est de pouvoir combiner tous les échantillons, puis utiliser des codes-barres pour déconvoluer l'origine de chaque séquence lors de l'analyse du résultat de séquençage. La préparation des librairies est une étape cruciale au succès d'une telle expérience qui doit être compatible avec la plateforme de NGS utilisée. Les technologies de NGS se classent en deux grandes catégories : les séquences courtes et longues (Figure 1-14). Nous avons considéré d'utiliser le séquençage PacBio générant de longues séquences, une méthode de troisième génération développée par Pacific BioSciences. Le principal avantage de cette

technique est qu'elle permet de séquencer des séquences allant jusqu'à 20 000 bases. Cependant, cette méthode génère beaucoup d'erreurs dues aux insertions et délétions de bases, et le nombre de total de séquences obtenues est relativement faible (50 000 séquences par expérience). Puisque notre librairie comporte théoriquement 7000 variants, que nous ne désirons séquencer qu'un seul gène d'environ 1 400 bases et que nous désirons obtenir plusieurs lectures par variant, cette approche fut écartée.[119, 129]

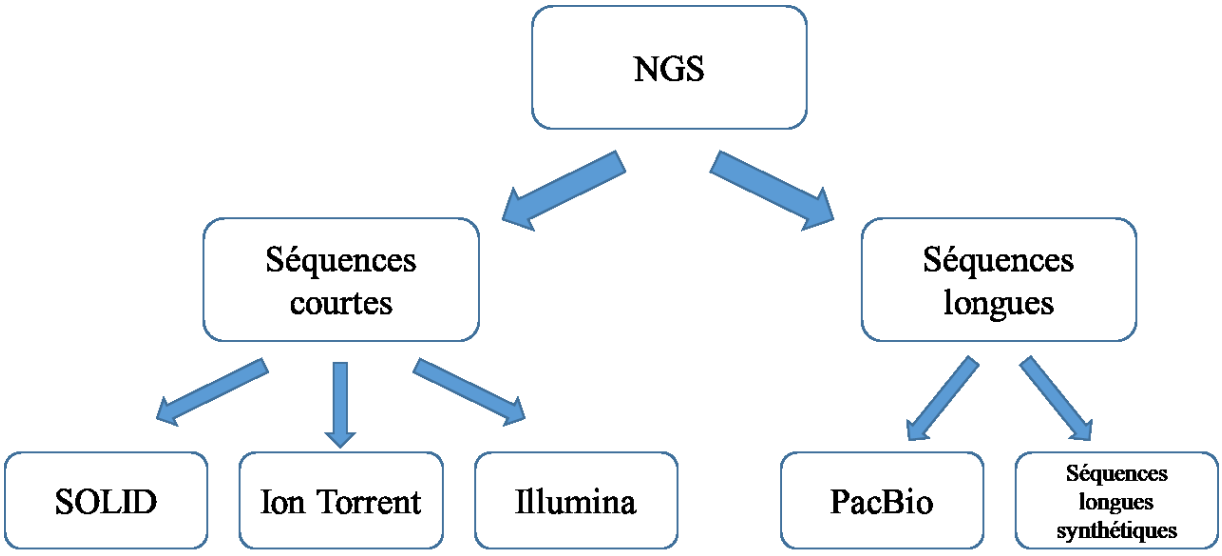


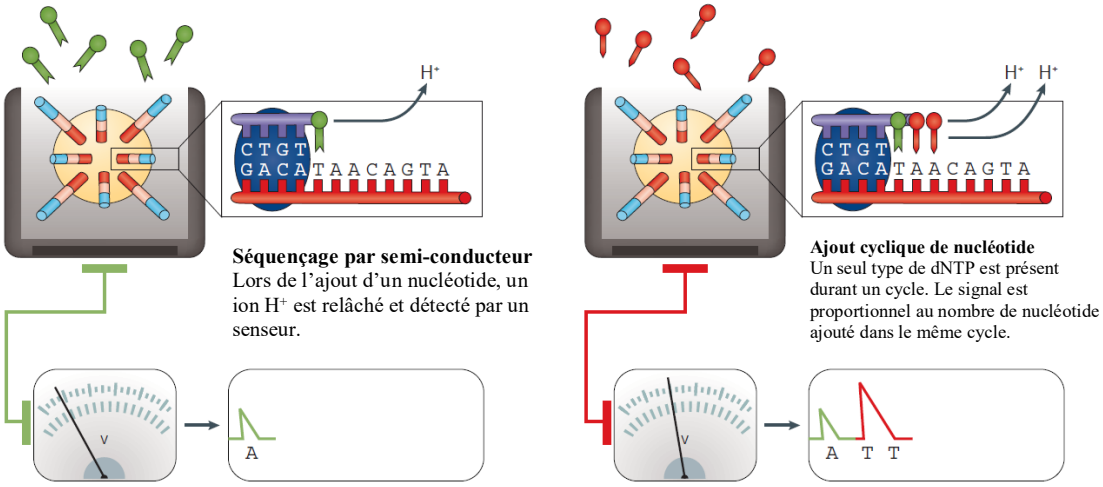
Figure 1-14. Hiérarchie des principales techniques utilisées en NGS.

Nous nous sommes tournés vers les méthodes de séquençage générant beaucoup plus de séquences courtes. Les trois plateformes de séquences courtes dominant présentement le marché sont : SOLID, Ion Torrent et Illumina (Figure 1-14).[119] La plateforme SOLID utilise une PCR par émulsion sur des billes pour amplifier le matériel génétique. Ensuite le séquençage se fait par des cycles d'hybridation et ligaturation de dinucléotides fluorescents engendrant quatre différents signaux représentant les quatre nucléotides. Cette méthode est très précise dans l'identification du nucléotide, mais en contrepartie, elle souffre d'une plus faible sensibilité et spécificité. En plus, les fragments ont une taille maximale de 75 bases, compliquant la préparation de la librairie. Nous avons donc écarté cette approche.

La plateforme Ion Torrent utilise la même méthode d'amplification que SOLID, mais exploite une méthode de séquençage par synthèse. Une polymérase amplifie les brins d'ADN sur les billes en ajoutant les nucléotides de façon itérative. Lors de l'ajout d'un nucléotide, le

relâchement de protons permet de suivre le changement de pH pour identifier le nucléotide ajouté (Figure 1-15A). La plateforme Illumina, quant à elle, est la plus utilisée sur le marché en égard à la maturité de sa technologie et sa compatibilité avec divers instruments. En Illumina, l'amplification se fait à l'aide d'adaptateurs se liant de chaque côté des brins d'ADN à séquencer. Suite à l'hybridation à la surface par ces adaptateurs, une polymérase amplifie chaque brin en formant des ponts, générant ainsi des zones où l'échantillon est amplifié à forte densité (Figure 1-15B). Le séquençage est ensuite similaire au séquençage Sanger, où l'ajout séquentiel de chacun des nucléotides est suivi par imagerie selon le relâchement de chacun de quatre fluorophores. (Figure 1-15C).

A

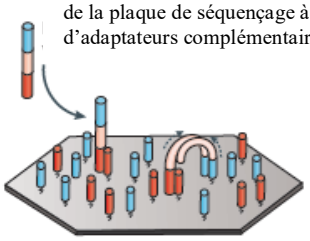


Séquençage par semi-conducteur
Lors de l'ajout d'un nucléotide, un ion H^+ est relâché et détecté par un senseur.

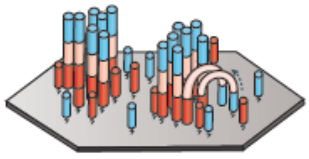
Ajout cyclique de nucléotide
Un seul type de dNTP est présent durant un cycle. Le signal est proportionnel au nombre de nucléotide ajouté dans le même cycle.

B

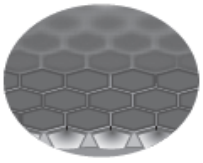
Liaison du gabarit
Le gabarit s'hybride sur la surface de la plaque de séquençage à l'aide d'adaptateurs complémentaires



Amplification par ponts
Chaque extrémité de brin d'ADN s'hybride et s'amplifie de chaque côté sur la plaque de façon à former un pont.



Génération de grappes d'ADN («*cluster*»)
Plusieurs cycles d'amplification forment de 100 à 200 millions de grappes contenant la même séquence d'ADN.



Motif des cellules sur la plaque de séquençage
Les cellules sont formées de puits dirigeant la formation et la densité des grappes.

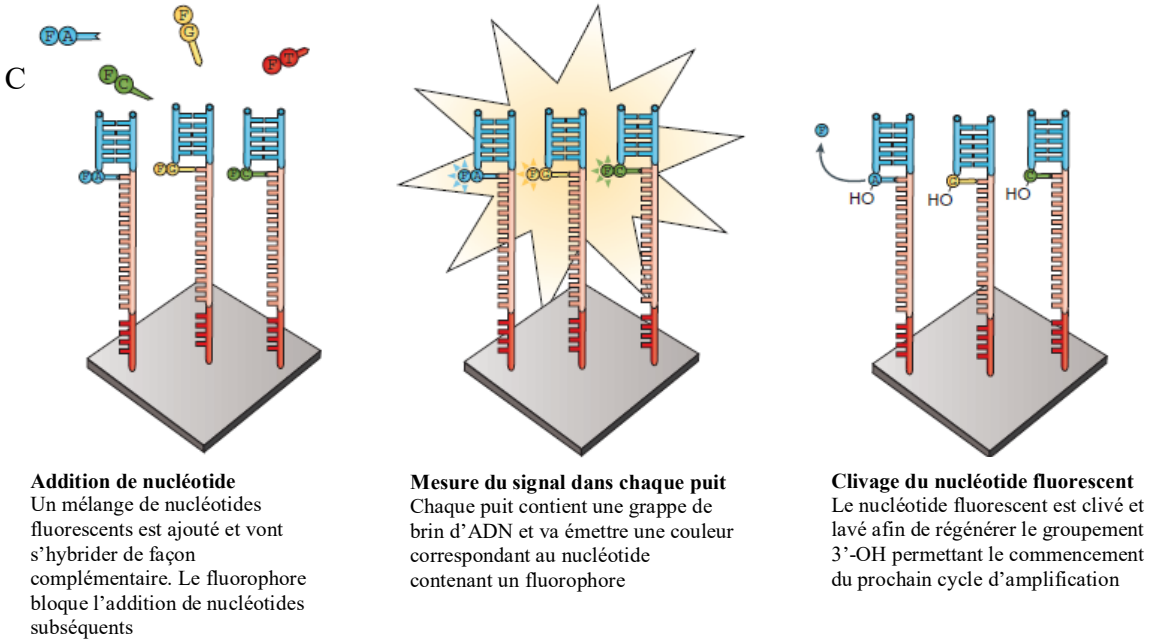


Figure 1-15. Représentation du fonctionnement de la plateforme Ion Torrent et Illumina en NGS.

A) Description de la méthode de séquençage Ion Torrent. B) Description de la méthode d'amplification de l'ADN par ponts. C) Description de la méthode de séquençage Illumina. Figure adaptée de [119].

Ces deux dernières techniques sont comparables en termes de puissance. Elles peuvent toutes deux générer des millions de séquences de façon rapide et abordable. Elles peuvent aussi traiter les homopolymères : l'ajout de deux nucléotides identiques consécutifs. Illumina peut sous-représenter des régions enrichies en AT ou CG, tandis que Ion Torrent possède un plus haut taux d'erreurs dues aux insertions et délétions de bases. Cependant, l'avantage unique d'Illumina est qu'elle peut générer des lectures de séquences appariées, c'est-à-dire une lecture en chaque direction (sens et antisens). La génération de séquences appariées réduit grandement les erreurs de séquençage, car elle fournit une confirmation de l'identité de chaque nucléotide. L'approche Illumina fut donc retenue.

La méthode de séquençage NGS Illumina nécessite de préparer les échantillons d'ADN en fragments d'une taille maximale de 300 paires de bases (pb). Sachant que les mutations que nous avons entreprises étaient ciblées dans des régions précises du gène de P450 BM3 et que la majorité de la séquence est native chez chaque variant, nous n'avons pas appliqué la méthode

de fragmentation aléatoire généralement utilisée.[130] Nous avons plutôt généré les fragments par PCR en ciblant cinq régions de 250 bp qui, globalement, incluent tous les sites mutés. Ainsi, l'information obtenue par l'utilisation de la plateforme MiSeq Illumina en séquences appariées offrira une excellente couverture de nos mutations. La fragmentation n'est pas compatible avec toutes les méthodes de mutagénèse. En effet, il n'est pas possible de retracer des combinaisons de mutations sur les fragments différents issus de mutagénèse aléatoire ou mutagénèse combinatoire. C'est pour cette raison que nos mutations doubles et triples sont à des positions consécutives.

La possibilité de séquencer des millions des séquences est d'autant plus intéressante, car il devient possible d'analyser les variants négatifs, chose qui est rarement rapportée par manque d'intérêt et de ressources. Cependant, ils renferment d'importantes informations. Au même titre que les variants positifs, les variants négatifs peuvent aussi aider à comprendre la structure, le mécanisme et les interactions importantes au sein de l'enzyme. Ces données peuvent aussi être utilisées dans la construction de séries de données pour l'entrainement et la validation d'algorithmes prédictifs. En plus, ces données permettent aux chercheurs de ne pas dupliquer des résultats négatifs qui n'auraient pas été publiés.

1.3.2 L'analyse du séquençage d'ADN

Le séquençage d'ADN par NGS a été développé à prime abord pour le séquençage de génomes complets d'humains, pour la détection de variations génétiques, et d'autres organismes pour comprendre l'évolution.[119, 131-133] Les fragments de génome sont séquencés puis alignés à l'aide d'algorithmes pour reconstruire ce dernier. Puisqu'un seul génome est séquencé à la fois, une différence de nucléotides à un site sera considérée comme une erreur de séquençage, décrivant le bruit de fond de la méthode. Une erreur typique de séquençage se situe habituellement dans moins de 1% des fragments de séquences générés.[119] Pour transposer l'utilisation du séquençage NGS dans le contexte de l'ingénierie d'enzymes, chacun des variants contenus dans un même échantillon (environ 2000, dans notre cas) devrait contenir entre une et trois mutations valides. Les algorithmes existants traduiront ces mutations, chacune représentant une fraction minime de l'échantillon global, comme étant des erreurs et elles ne seront donc pas répertoriées.

Afin de remédier à cette lacune, nous avons écrit un script pour extraire les mutations, en collaboration avec le professeur Sebastian Pechmann (Bio-informatique, U. de Montréal). Le script permet de filtrer afin d'éliminer les séquences de faible qualité, qui ne sont pas de pleine longueur ou qui ne sont pas appariées. Le fichier final liste toutes les mutations ainsi que leur fréquence dans chacun des trois groupes sélectionnés selon leur phénotype (Blanc, Bleu pâle, Bleu foncé).

Un second script était requis pour évaluer la validité de chacune des mutations identifiées, car une mutation peut être identifiée dans plus d'un groupe ou bien être non significative. Le test exact de Fisher permet d'évaluer, à l'aide d'un « *p-value* », si la comparaison d'une mutation au sein de deux groupes est significative, ainsi qu'un rapport des chances (« *odds ratio* ») pour déterminer l'appartenance de la mutation.

Pour reprendre l'image précédemment discutée, la quatrième vague de la biocatalyse correspondrait à l'utilisation d'outils informatiques pour analyser, modéliser, simuler et prédire à plus grande échelle la voie à prendre pour accélérer l'évolution dirigée d'une enzyme envers une réaction donnée.[37, 43] Malgré l'amélioration des technologies de mutagenèse et de criblage de variants, l'exploration complète de l'espace chimique reste inatteignable; il en découle que notre compréhension de systèmes enzymatiques demeure très limitée. Afin de visualiser la relation entre la structure, le dynamisme et l'activité des mutations de P450 BM3, des modèles ont été construits à l'aide de PyMOL et de courtes simulations de dynamique moléculaire (MD) ont été exécutées à l'aide du logiciel de simulations GROMACS. Ces étapes ont été effectuées par le post-doctorant de Sebastian Pechmann, Musa Ozboyaci. Les méthodes utilisées dans ces travaux sont expliquées brièvement ci-dessous.

Une structure tridimensionnelle est nécessaire pour effectuer des simulations. L'obtention de ces structures peut se faire de façon expérimentale par résonance magnétique nucléaire (RMN) ou par cristallographie aux rayons X. Ce sont de puissants outils ayant chacun leurs désavantages. La RMN contient beaucoup de données à analyser et est limitée par la taille de la protéine tandis qu'en cristallographie aux rayons X, le procédé pour obtenir des cristaux peut être difficile et l'analyse du dynamisme de la protéine est limitée.[134] Dans le contexte de

l'ingénierie d'enzymes, il n'est pas réaliste de tenter d'obtenir la structure de chacun des variants par ces méthodes. La modélisation permet de remédier à cette lacune.[135]

Lorsque la structure de l'enzyme native ou celle d'un homologue fiable est connue, il est possible de l'utiliser comme gabarit pour modéliser la structure tridimensionnelle des variants mutés. La fiabilité du modèle dépend du pourcentage d'identité entre les deux séquences.[135] Dans notre cas, puisque les variants d'intérêt ne contiennent qu'une seule mutation, nous avons utilisé le logiciel PyMOL pour générer leurs modèles. Bien que principalement utilisé pour la visualisation de protéines, PyMOL permet de modéliser les variants mutés. Cette approche est beaucoup moins intensive en ressources computationnelles que d'autres alternatives et a permis la génération des modèles selon un mode automatisé.

Les lois de la thermodynamique dictent que tout système tend à minimiser son énergie. Ainsi, les interactions des acides aminés entre eux et avec leur environnement causent le repliement de la protéine. Les différentes conformations qu'une enzyme peut adopter peuvent être exprimées par son profil énergétique (« *energy landscape* ») et l'énergie libre de Gibbs. Lors de la création de modèles de variants, la substitution d'un résidu peut faire en sorte que l'énergie du système est moins favorable. L'algorithme de PyMOL propose différents rotamères afin de réduire l'encombrement stérique dû à la substitution d'un acide aminé par un autre, et identifie le modèle le plus énergétiquement stable. Nous avons choisi de construire les modèles de variants produisant de l'indigo et de variants ne produisant pas d'indigo mais qui sont bien exprimés. Ces derniers servent de contrôles négatifs pour avoir une bonne représentation dans la série de données.

Ensuite nous avons soumis les modèles de chaque variant d'intérêt à de courtes simulations de dynamique moléculaire (MD) afin de visualiser le dynamisme de P450 BM3 et ses variants. Une dynamique moléculaire utilise un champ de force basé soit sur la mécanique classique (MM), de mécanique quantique (QM) ou une combinaison des deux (QM/MM).[136] Un champ de force est un ensemble d'équations et paramètres permettant de calculer l'énergie d'une protéine en fonction de ses coordonnées atomiques dans le temps. Plusieurs champs de force peuvent être utilisés : AMBER, CHARMM, GROMOS ou OPLS. La sélection du champ de force se fait en fonction du type molécule, de simulation et de l'échelle de temps à observer

puisqu'ils sont paramétrisés différemment.[137-139] Au final, les plus utilisés pour les protéines sont AMBER et CHARMM; nos collaborateurs ont choisi d'utiliser le champ de force AMBER pour les simulations.

Le champ de force en MM, basée sur les lois de Newton, est séparé en deux composantes : les paramètres intramoléculaires et intermoléculaires (Figure 1-16). Les paramètres intramoléculaires expriment la distance entre les liens atomiques, les angles, les torsions et autres tdandis que les paramètres intermoléculaires traitent de toutes les interactions possibles telles que les forces de Van der Waals, les interactions électrostatiques, l'effet hydrophobique et autres.[140] Ces simulations sont très intensives, alors elles ont été effectuées pour un sous-groupe de nos variants en cinq répliques indépendantes pour éviter tous biais.

$$E_{\text{total}} = \sum_{\text{bonds}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 +$$

$$\sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}} \right]$$

Figure 1-16. Exemple général du champ de force d'AMBER.

Les trois premiers termes représentent les interactions intramoléculaires tandis que le dernier représente les interactions intermoléculaires. Figure reproduite de [140]

Lors de ces simulations, il faut définir les paramètres du système pour qu'il soit le plus fidèle possible à la réalité. La loi des gaz parfaits est utilisée pour définir l'environnement de la simulation. Il est possible de choisir parmi différents ensembles tels que micro-canonique (nVE) : le nombre de particules, le volume et l'énergie sont conservés; canonique (nVT) : le nombre de particules, le volume et la température sont conservés; isotherme-isobare (nPT) : le nombre de particules, la pression et la température sont conservés.[141] Les ensembles nVT et nPT furent choisis pour différentes étapes de nos travaux.

Afin de diminuer le nombre de calculs et s'assurer de maintenir un environnement aqueux constant, les conditions périodiques aux limites (« *Periodic Boundary Conditions* » (PBC)) ont été appliquées. Lors d'une simulation, des contraintes sont appliquées pour maintenir la protéine d'intérêt dans une boite de dimensions fixes remplie de molécules d'eau. Le principe repose sur

la duplication de cette boite dans les trois dimensions de sorte que les molécules sortant de la boite entrent de l'autre côté avec la même énergie.[141]

Il y a deux méthodes pour définir la génération de cette boite d'eau, soit de façon implicite et l'explicite. La méthode implicite simplifie les calculs en considérant une approximation homogène d'eau pour alléger la simulation. Une approximation des interactions électrostatiques de longue portée réduit considérablement le nombre de molécules à considérer.[142] Il est ensuite possible d'inclure explicitement les molécules d'eau importantes. La méthode explicite signifie que chacune des molécules d'eau est simulée dans le système. Cela augmente la complexité de la simulation en augmentant le nombre de degrés de liberté, car le système contient beaucoup plus de molécules.[143] Nos collaborateurs ont choisi l'eau explicite, car les simulations sont plus fidèles à la réalité et il y a moins de limites computationnelles de nos jours.

Lorsque les paramètres ont été sélectionnés, il faut préparer le système pour la simulation. Il est important de minimiser l'énergie du système pour relaxer les contraintes imposées par les substitutions d'acides aminés. Les méthodes de minimisation d'énergie sont basées sur des algorithmes généraux d'optimisation. Les méthodes les plus utilisées sont le « *steepest descent* » (SD) et le « *conjugated gradient* » (CG). Leur explication détaillée dépasse le contexte de ce mémoire, mais de façon générale, SD converge plus rapidement vers le minimum (la conformation de protéine la plus stable), tandis que CG y arrive en moins d'itérations.[144] Nos collaborateurs ont minimisé le système par la méthode SD pendant 1 ns dans l'ensemble nVT suivi de 1 ns dans l'ensemble nPT avant d'effectuer des simulations de 25 ns dans l'ensemble nPT.

Plusieurs raisons pourraient expliquer la relation entre la structure, le dynamisme et l'activité des mutations qui seront discutées dans le chapitre 3. Nous tentons de les identifier par la méthode de tICA (« *time-structure independent analysis* ») suivi de classification par grappes («*clustering*»). Brièvement, tICA est semblable au PCA (« *Principal Component Analysis* »). PCA est une technique statistique qui consiste à réduire la dimensionnalité d'un système pour cerner les facteurs dominants décrivant son dynamisme.[145] La dimensionnalité est décrite par le nombre de composantes principales (PC). Cependant, cette méthode repose sur des propriétés statiques et linéaires faisant en sorte que la décomposition de ces PC serait moins valide pour le dynamisme des protéines.[146] À ces fins, nos collaborateurs utilisent le tICA, qui rend les PC

le plus indépendantes possible afin de mieux décrire et analyser le dynamisme du système. Des macro-états ont été identifiés durant les simulations pour chacun des variants analysés. La classification par grappes permet d'observer les similarités entre les macro-états de chacun des variants pour discerner un changement structural ou dynamique en commun qui permettrait de distinguer les variants actifs des variants inactifs pour la synthèse de l'indigo.

1.4 Objectifs de recherche

L'objectif principal de ce mémoire est de tenter d'approfondir les connaissances sur la biocatalyse en utilisant le cytochrome P450 BM3 comme système modèle. Nous explorons son espace chimique et tentons d'en déduire des corrélations parmi les différentes activités des variants expliquant la promiscuité de substrat ainsi que la raison pour laquelle une multitude de variants sont capables de synthétiser l'indigo.

Pour ce faire, nous avons vérifié l'hypothèse que les variants de P450 BM3 actifs pour indigo seraient plus portés à être actifs envers un second substrat, le précurseur de la cétone de framboise (chapitre 2). Dans un premier temps, nous avons mis sur pied un flux de travail allant de la fabrication de variants de P450 BM3 à leur criblage. L'objectif était d'optimiser chacune des étapes de ce flux de travail afin d'augmenter le nombre de variants actifs découverts en diminuant les coûts et le temps des manipulations. Par ces méthodes, nous avons accéléré la découverte de variants actifs pour la synthèse de l'indigo et les corrélons ensuite avec la production de cétone de framboise.

Suite à la confirmation de cette hypothèse, nous avons entrepris la cartographie à plus grande échelle de la relation entre la séquence et la fonction de P450 BM3 pour la synthèse de l'indigo grâce au NGS (chapitre 3). L'utilisation du NGS a permis de lier de façon massive le génotype au phénotype de milliers de variants simultanément. Puisque cette méthode est peu documentée pour la biocatalyse, nous avons établi les procédés pour la préparation des librairies ainsi que l'analyse des résultats par outils informatiques. Avec nos collaborateurs, nous avons développé des scripts pour l'analyse statistique des résultats de séquençage et l'analyse structurale des variants par combinaison de modélisation et dynamique moléculaire. Par ces moyens, nous avons identifié un changement de conformation de la P450 BM3 qui prédomine chez les mutants actifs pour la synthèse de l'indigo, nous permettant de proposer une hypothèse nouvelle quant à son rôle fonctionnel.

1.5 Références

[1] A. Zapf, M. Beller, *Top. Catal.* **2002**, *19*, 101-109.
[2] S. Sadasivaiah, Y. Tozan, J. G. Breman, *Am. J. Trop. Med. Hyg.* **2007**, *77*, 249-263.
[3] S. M. Snedeker, *Environ. Health Perspect.* **2001**, *109*, 35-47.
[4] T. J. Crowley, *Science* **2000**, *289*, 270.
[5] IPCC: Human and natural drivers of climate change, http://www.ipcc.ch/publications_and_data/ar4/wg1/en/spmsspm-human-and.html, **2007**, (accédé le 15 Juin 2018).
[6] IPCC report: A changing climate creates pervasive risks but opportunities exist for effective responses, https://web.archive.org/web/20140528144653/http://ipcc.ch/pdf/ar5/pr_wg2/140330_pr_wgII_spm_en.pdf, **2014**, (accédé le 15 Juin 2018).
[7] P. Harremoës, D. Gee, M. MacGarvin, A. Stirling, J. Keys, B. Wynne, S. Guedes Vaz, *EEA. Late Lessons from Early Warnings: The Precautionary Principle 1896–2000*, **2001**.
[8] S. A. Montzka, et al., *Nature* **2018**, *557*, 413-417.
[9] P. T. Anastas, M. M. Kirchhoff, *Acc. Chem. Res.* **2002**, *35*, 686-694.
[10] V. Zuin, L. Mammino, *Worldwide Trends in Green Chemistry Education*, Royal Society of Chemistry, **2015**.
[11] R. A. Sheldon, in *Methods and Reagents for Green Chemistry*, John Wiley & Sons, Inc., **2007**, pp. 189-199.
[12] R. A. Sheldon, I. W. C. E. Arends, U. Hanefeld, in *Green Chemistry and Catalysis*, Wiley-VCH Verlag GmbH & Co. KGaA, **2007**, pp. 1-47.
[13] C. Jiménez-González, A. D. Curzons, D. J. C. Constable, V. L. Cunningham, *The International Journal of Life Cycle Assessment* **2004**, *9*, 114-121.
[14] P. J. Dunn, *Chem. Soc. Rev.* **2012**, *41*, 1452-1461.
[15] R. K. Henderson, et al., *Green Chem.* **2011**, *13*, 854-862.
[16] K. Alfonsi, et al., *Green. Chem* **2008**, *10*, 31-36.
[17] I. M. N. Smallwood, *Solvent Recovery Handbook*, Blackwell Science, **2002**.
[18] B. M. Trost, *Science* **1991**, *254*, 1471.
[19] R. A. Sheldon, *Green Chem.* **2007**, *9*, 1273-1283.
[20] J. H. Clark, *Green Chem.* **2006**, *8*, 17-21.
[21] J. B. Manley, P. T. Anastas, B. W. Cue, *J. Clean. Prod.* **2008**, *16*, 743-750.
[22] Q.-L. Zhou, *Angew. Chem. Int. Ed.* **2015**, *55*, 5352-5353.
[23] B. M. Trost, *Angew. Chem. Int. Ed.* **1995**, *34*, 259-281.
[24] C. E. Garrett, K. Prasad, *Adv. Synth. Catal.* **2004**, *346*, 889-900.
[25] J. R. Ludwig, C. S. Schindler, *Chem* **2017**, *2*, 313-316.
[26] R. Ghosh, E. Lindstedt, N. Jalalian, B. Olofsson, *ChemistryOpen* **2014**, *3*, 54-57.
[27] P. Wasserscheid, in *Handbook of Green Chemistry*, Wiley-VCH Verlag GmbH & Co. KGaA, **2010**.
[28] R. A. Sheldon, I. Arends, U. Hanefeld, *Green chemistry and catalysis*, Wiley-VCH, **2007**.
[29] P. Grunwald, *Biocatalysis*, Imperial College Press, **2009**.
[30] M. Salque, P. I. Bogucki, J. Pyzel, I. Sobkowiak-Tabaka, R. Grygiel, M. Szmyt, R. P. Evershed, *Nature* **2012**, *493*, 522.
[31] G. W. Huisman, D. Gray, *Curr. Opin. Biotechnol.* **2002**, *13*, 352-358.

[32] K. Drauz, H. Gröger, O. May, *Enzyme Catalysis in Organic Synthesis*, John Wiley & Sons, **2012**.
[33] J. Albarrán-Velo, D. González-Martínez, V. Gotor-Fernández, *Biocatal. Biotransform.* **2018**, *36*, 102-130.
[34] N. Vargesson, *Birth Defects Res.* **2015**, *105*, 140-156.
[35] H. Steiner, B.-H. Jonsson, S. Lindskog, *Eur. J. Biochem.* **1975**, *59*, 253-259.
[36] E. Koshland Daniel, *Angew. Chem. Int. Ed.* **1995**, *33*, 2375-2378.
[37] U. T. Bornscheuer, *Philos. Trans. Royal Soc. A* **2018**, *376*.
[38] L. Rosenthaler, *Biochem. Z* **1908**, *14*, 238-253.
[39] T. Nagasawa, T. Nakamura, H. Yamada, *Appl. Microbiol. Biotechnol.* **1990**, *34*, 322-324.
[40] Grand View Research: Global Acrylamide Market, https://www.grandviewresearch.com/press-release/global-polyacrylamide-market, **2017**, (accédé le 7 Juillet 2018).
[41] M. Bilal, H. M. N. Iqbal, S. Guo, H. Hu, W. Wang, X. Zhang, *Int. J. Biol. Macromol.* **2018**, *108*, 893-901.
[42] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185.
[43] M. D. Truppo, *ACS Med. Chem. Lett.* **2017**, *8*, 476-480.
[44] W. J. W. Watson, *Green Chem.* **2012**, *14*, 251-259.
[45] S. Aldridge, *Nat. Biotechnol.* **2013**, *31*, 95.
[46] A. J. J. Straathof, S. Panke, A. Schmid, *Curr. Opin. Biotechnol.* **2002**, *13*, 548-556.
[47] J.-M. Choi, S.-S. Han, H.-S. Kim, *Biotechnol. Adv.* **2015**, *33*, 1443-1454.
[48] A. Schmid, J. S. Dordick, B. Hauer, A. Kiener, M. Wubbolts, B. Witholt, *Nature* **2001**, *409*, 258.
[49] Pmlive: Top 50 pharmaceutical products by global sales, http://www.pmlive.com/top_pharma_list/pharmaceutical_products/januvia, **2015**, (accédé le 23 Juillet 2018).
[50] C. K. Savile, et al., *Science* **2010**, *329*, 305.
[51] Statista, Worldwide revenue of Pfizer's Lipitor from 2003 to 2017, https://www.statista.com/statistics/254341/pfizers-worldwide-viagra-revenues-since-2003/, **2017**, (accédé le 16 Mai 2018).
[52] J. Chapman, A. Ismail, C. Dinu, *Catalysts* **2018**, *8*, 238.
[53] J. Axelrod, *J. Pharmacol. Exp. Ther.* **1955**, *114*, 430.
[54] B. B. Brodie, J. Axelrod, J. R. Cooper, L. Gaudette, B. N. La Du, C. Mitoma, S. Udenfriend, *Science* **1955**, *121*, 603.
[55] M. Klingenberg, *Arch. Biochem. Biophys.* **1958**, *75*, 376-386.
[56] D. Garfinkel, *Arch. Biochem. Biophys.* **1958**, *77*, 493-509.
[57] T. Omura, R. Sato, *J. Biol. Chem.* **1962**, *237*, 1375-1376.
[58] T. Omura, R. Sato, *J. Biol. Chem.* **1964**, *239*, 2370-2378.
[59] P. Manikandan, S. Nagini, *Curr. Drug Targets* **2018**, *19*, 38-54.
[60] T. Sakaki, *Biol. Pharm. Bull.* **2012**, *35*, 844-849.
[61] E. M. Isin, F. P. Guengerich, *Biochim. Biophys. Acta, Gen. Subj.* **2007**, *1770*, 314-329.
[62] F. P. Guengerich, A. W. Munro, *J. Biol. Chem.* **2013**, *288*, 17065-17073.
[63] M. Parvez, et al., *Scientific Reports* **2016**, *6*, 33099.
[64] R. W. Estabrook, *Drug Metab. Dispos.* **2003**, *31*, 1461.
[65] S. S. Singh, *Curr. Drug Metab.* **2006**, *7*, 165-182.

[66] K. P. Cusack, H. F. Koolman, U. E. W. Lange, H. M. Peltier, I. Piel, A. Vasudevan, *Bioorg. Med. Chem. Lett.* **2013**, *23*, 5471-5483.
[67] C. Bunchorntavakul, K. R. Reddy, *Clin. Liver Dis.* **2013**, *17*, 587-607, viii.
[68] M. Blieden, L. C. Paramore, D. Shah, R. Ben-Joseph, *Expert Rev. Clin. Pharmacol.* **2014**, *7*, 341-348.
[69] X. Qi, *Mol. Pharm.* **2006**, *3*, 187-195.
[70] P. R. Ortiz de Montellano, *Future Med. Chem.* **2013**, *5*, 213-228.
[71] I. Watanabe, F. Nara, N. Serizawa, *Gene* **1995**, *163*, 81-85.
[72] T. Matsuoka, S. Miyakoshi, K. Tanzawa, K. Nakahara, M. Hosobuchi, N. Serizawa, *Eur. J. Biochem.* **1989**, *184*, 707-713.
[73] M. Girhard, K. Machida, M. Itoh, R. D. Schmid, A. Arisawa, V. B. Urlacher, *Microb. Cell Fact.* **2009**, *8*, 36.
[74] V. Subramanian, J. S. Yadav, *Appl. Environ. Microbiol.* **2009**, *75*, 5570-5580.
[75] Z. Mao, X.-F. Zheng, Y.-Q. Zhang, X.-X. Tao, Y. Li, W. Wang, *Int J. Mol. Sci.* **2012**, *13*, 491-505.
[76] C. J. C. Whitehouse, S. G. Bell, L. L. Wong, *Chem. Soc. Rev.* **2012**, *41*, 1218-1260.
[77] S. Govindaraj, T. L. Poulos, *J. Biol. Chem.* **1997**, *272*, 7915-7921.
[78] W. Munro Andrew, J. G. Lindsay, *Mol. Microbiol.* **1996**, *20*, 1115-1125.
[79] A. W. Munro, et al., *Trends Biochem. Sci.* **2002**, *27*, 250-257.
[80] H. M. Girvan, et al., *Arch. Biochem. Biophys.* **2011**, *507*, 75-85.
[81] M. A. Noble, et al., *Biochem. J.* **1999**, *339*, 371-379.
[82] Y. Miura, A. J. Fulco, *Biochim. Biophys. Acta* **1975**, *388*, 305-317.
[83] E. M. Isin, F. P. Guengerich, *Anal. Bioanal. Chem.* **2008**, *392*, 1019-1030.
[84] S. C. Gay, A. G. Roberts, J. R. Halpert, *Future Med. Chem.* **2010**, *2*, 1451-1468.
[85] Y. Miura, A. J. Fulco, *J. Biol. Chem.* **1974**, *249*, 1880-1888.
[86] T. Reetz Manfred, *Angew. Chem. Int. Ed.* **2010**, *50*, 138-174.
[87] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, S. J. Tans, *Nature* **2007**, *445*, 383.
[88] S. T. Jung, R. Lauchli, F. H. Arnold, *Curr. Opin. Biotechnol.* **2011**, *22*, 809-817.
[89] L. Capoferri, et al., *Proteins* **2016**, *84*, 383-396.
[90] H. Li, T. L. Poulos, *Acta Crystallogr D Biol Crystallogr* **1995**, *51*, 21-32.
[91] M. G. Joyce, H. M. Girvan, A. W. Munro, D. Leys, *J. Biol. Chem.* **2004**, *279*, 23287-23293.
[92] M. C. C. J. C. Ebert, S. L. Dürr, A. A. Houle, G. Lamoureux, J. N. Pelletier, *ACS Catalysis* **2016**, *6*, 7426-7437.
[93] E. de Brito Sá, A. Rimola, L. Rodríguez-Santiago, M. Sodupe, X. Solans-Monfort, *The Journal of Physical Chemistry A* **2018**, *122*, 1702-1712.
[94] P. S. Coelho, E. M. Brustad, A. Kannan, F. H. Arnold, *Science* **2013**, *339*, 307-310.
[95] G. Di Nardo, G. Gilardi, *Int J. Mol. Sci.* **2012**, *13*, 15901-15924.
[96] C. F. Butler, C. Peet, A. E. Mason, M. W. Voice, D. Leys, A. W. Munro, *J. Biol. Chem.* **2013**, *288*, 25387-25399.
[97] X. Ren, A. Yorke Jake, E. Taylor, T. Zhang, W. Zhou, L. Wong Luet, *Chem. Eur. J.* **2015**, *21*, 15039-15047.
[98] B. D. Ensley, B. J. Ratzkin, T. D. Osslund, M. J. Simon, L. P. Wackett, D. T. Gibson, *Science* **1983**, *222*, 167-169.
[99] C. S. W. Koehler, *Today's Chemist at Work* **1999**, *8*, 85-91.
[100] J. C. Splitstoser, T. D. Dillehay, J. Wouters, A. Claro, *Sci. Adv.* **2016**, *2*, 1-4.

[101] E. M. J. Gillam, L. M. Notley, H. Cai, J. J. De Voss, F. P. Guengerich, *Biochemistry* **2000**, *39*, 13817-13824.
[102] D. Murdock, B. D. Ensley, C. Serdar, M. Thalen, *Bio/Technology* **1993**, *11*, 381.
[103] H. Bialy, *Nat. Biotechnol.* **1997**, *15*, 110.
[104] E. M. Gillam, et al., *Biochem. Biophys. Res. Commun.* **1999**, *265*, 469-472.
[105] Q. S. Li, U. Schwaneberg, P. Fischer, R. D. Schmid, *Chemistry* **2000**, *6*, 1531-1536.
[106] H. M. Li, L. H. Mei, V. B. Urlacher, R. D. Schmid, *Appl. Biochem. Biotechnol.* **2008**, *144*, 27-36.
[107] W. C. Huang, A. C. G. Westlake, J. D. M. M. Gordon Joyce, P. C. E. Moody, G. C. K. Roberts, *J. Mol. Biol.* **2007**, *373*, 633-651.
[108] Q.-S. Li, J. Ogawa, R. D. Schmid, S. Shimizu, *Biosci. Biotechnol. Biochem.* **2005**, *69*, 293-300.
[109] Y. Lu, L. Mei, *J. Ind. Microbiol. Biotechnol.* **2007**, *34*, 247-253.
[110] K. L. Morley, R. J. Kazlauskas, *Trends Biotechnol.* **2005**, *23*, 231-237.
[111] L. X. Tang, H. Gao, X. C. Zhu, X. Wang, M. Zhou, R. X. Jiang, *BioTechniques* **2012**, *52*, 149-+.
[112] R. Martínez, U. Schwaneberg, *Biol. Res.* **2013**, *46*, 395-405.
[113] P. P. Kelly, A. Eichler, S. Herter, D. C. Kranz, N. J. Turner, S. L. Flitsch, *Beilstein J. Org. Chem.* **2015**, *11*, 1713-1720.
[114] A. Celik, R. E. Speight, N. J. Turner, *Chem. Commun.* **2005**, 3652-3654.
[115] S. H. Park, et al., *Drug Metab. Dispos.* **2010**, *38*, 732-739.
[116] J. Beekwilder, I. M. van der Meer, O. Sibbesen, M. Broekgaarden, I. Qvist, J. D. Mikkelsen, R. D. Hall, *Biotechnol. J.* **2007**, *2*, 1270-1279.
[117] G. E. Tsotsou, A. E. G. Cass, G. Gilardi, *Biosens. Bioelectron.* **2002**, *17*, 119-131.
[118] A. Gandomi, M. Haider, *Int. J. Inf. Man.* **2015**, *35*, 137-144.
[119] S. Goodwin, J. D. McPherson, W. R. McCombie, *Nat. Rev. Genet.* **2016**, *17*, 333-351.
[120] C. S. Greene, J. I. E. Tan, M. Ung, J. H. Moore, C. Cheng, *J. Cell. Physiol.* **2014**, *229*, 1896-1900.
[121] D. M. Fowler, S. Fields, *Nat. Methods* **2014**, *11*, 801-807.
[122] P. A. Romero, T. M. Tran, A. R. Abate, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 7159-7164.
[123] T. A. Whitehead, et al., *Nat. Biotechnol.* **2012**, *30*, 543-548.
[124] D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, S. Fields, *Nat. Methods* **2010**, *7*, 741-746.
[125] J. D. Heredia, J. Park, R. J. Brubaker, S. K. Szymanski, K. S. Gill, E. Procko, *J. Immunol.* **2018**, *200*, 3825-3839.
[126] A. M. Freeman, B. M. Mole, R. E. Silversmith, R. B. Bourret, *J. Bacteriol.* **2011**, *193*, 4709-4718.
[127] L. M. F. Mendonça, S. R. Marana, *Biochim. Biophys. Acta, Proteins Proteomics* **2011**, *1814*, 1616-1623.
[128] J. Lee, N. M. Goodey, *Chem. Rev.* **2011**, *111*, 7595-7624.
[129] A. Rhoads, K. F. Au, *Genomics, Proteomics & Bioinformatics* **2015**, *13*, 278-289.
[130] S. R. Head, H. K. Komori, S. A. LaMere, T. Whisenant, F. Van Nieuwerburgh, D. R. Salomon, P. Ordoukhanian, *BioTechniques* **2014**, *56*, 61-passim.
[131] J. Wu, M. Wu, T. Chen, R. Jiang, *Quant. Biol.* **2016**, *4*, 115-128.

[132] J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, E. Trees, *Clin. Microbiol. Infect.* **2018**, *24*, 335-341.
[133] H. P. J. Buermans, J. T. den Dunnen, *Biochim. Biophys. Acta, Mol. Basis Dis.* **2014**, *1842*, 1932-1941.
[134] V. V. Krishnan, B. Rupp, *Macromolecular Structure Determination: Comparison of X-ray Crystallography and NMR Spectroscopy*, eLS John Wiley & Sons, **2012**.
[135] Z. Xiang, *Curr. Protein Pept. Sci.* **2006**, *7*, 217-227.
[136] H. M. Senn, W. Thiel, *Angew. Chem. Int. Ed.* **2009**, *48*, 1198-1229.
[137] F. Martín-García, E. Papaleo, P. Gomez-Puertas, W. Boomsma, K. Lindorff-Larsen, *PLoS One* **2015**, *10*, e0121114.
[138] E. A. Cino, W.-Y. Choy, M. Karttunen, *J. Chem. Theory Comput.* **2012**, *8*, 2725-2740.
[139] K. A. Beauchamp, Y.-S. Lin, R. Das, V. S. Pande, *J. Chem. Theory Comput.* **2012**, *8*, 1409-1414.
[140] W. D. Cornell, et al., *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
[141] H. C. Andersen, *J. Chem. Phys.* **1980**, *72*, 2384-2393.
[142] R. Anandakrishnan, A. Drozdetski, R. C. Walker, A. V. Onufriev, *Biophys. J.* **2015**, *108*, 1153-1164.
[143] C. J. Cramer, D. G. Truhlar, *Chem. Rev.* **1999**, *99*, 2161-2200.
[144] D. H. J. Mackay, A. J. Cross, A. T. Hagler, in *Prediction of Protein Structure and the Principles of Protein Conformation* (Ed.: G. D. Fasman), Springer US, Boston, MA, **1989**, pp. 317-358.
[145] C. C. David, D. J. Jacobs, *Methods Mol. Biol.* **2014**, *1084*, 193-226.
[146] Y. Naritomi, S. Fuchigami, *J. Chem. Phys.* **2011**, *134*, 065101.

Chapitre 2: Indigo synthesis is a robust predictor of raspberry ketone production by engineered cytochrome P450 BM3 variants

Olivier Rousseau[1,3,4], Maximilian C.C.J.C. Ebert[2,3,4], Daniela Quaglia[1,3,4], Saathanan Iyathurai[2,3,4], Joelle N. Pelletier[*,1,2,3,4]

[1] Department of Chemistry and [2] Department of Biochemistry, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada

[3] PROTEO, the Québec Network for Protein Function, Engineering and Applications, Québec, G1V 0A6, Canada

[4] CGCC, the Center of Green Chemistry and Catalysis, Montréal, Québec, H3A 0B8, Canada

Corresponding author:
Joelle N. Pelletier
Joelle.pelletier@umontreal.ca

Keywords: Biocatalysis, Enzyme engineering, High-throughput screening, Indigo, Raspberry ketone

2.1 Préface

Ce chapitre présente l'article « *Indigo synthesis is a robust predictor of raspberry ketone production by engineered cytochrome P450 BM3* variants ». Il a été soumis pour publication le 6 septembre 2018 au journal *ChemCatChem.* Le projet sur les P450 a débuté avec mon ancien superviseur de stage et étudiant au doctorat, Maximilian C.J.C.C Ebert et ma directrice de recherche Prof. Joelle N. Pelletier. Maximilian a instauré dans le laboratoire les premières procédures de mutagénèse, de quantification et de criblage de la P450 BM3, procédures utilisées dans le contexte de cet article qui traite de la découverte de nouveaux substrats de P450 BM3 par l'optimisation de son ingénierie.

En tant que premier auteur, j'ai optimisé les méthodes de criblage établies, développé de nouvelles méthodes de criblage et accompli l'entièreté des expériences pour ce manuscrit. J'ai créé les librairies de mutants, optimisé la méthode de quantification de la P450, la méthode de criblage à haut débit par absorbance de l'indigo et fluorescence du NADPH et la méthode de détection et d'identification de produits par LC-MS et GC-MS. Je les ai ensuite appliquées pour la synthèse de l'indigo et de la cétone de framboise. En plus, j'ai développé la méthode pour synthétiser et isoler un produit à plus grande échelle pour ensuite l'identifier par RMN. Finalement, j'ai développé un script pour accélérer l'analyse des résultats. En tant que deuxième auteur, Maximilian a grandement contribué à élaborer l'idée de cet article et me guider dans les expériences que j'ai effectuées. En tant que troisième auteure, Daniela Quaglia m'a guidé dans la planification de la rédaction et la correction de l'article. Saathanan Iyathurai a contribué à l'optimisation des étapes du criblage de haut débit. Finalement, Prof. Joelle N. Pelletier m'a guidé tout au long de ce projet et m'a dirigé lors de la rédaction de l'article.

2.2 Abstract

Natural raspberry ketone has a high value in the flavor, fragrance and pharmaceutical industries. Its extraction is costly, justifying the search for biosynthetic routes. We hypothesized that cytochrome P450 BM3 (P450 BM3) could be engineered to catalyze the hydroxylation of 4-phenylbutan-2-one, a naturally sourceable precursor, to raspberry ketone. The production of indigo by variants of P450 BM3 has previously served as a predictor for promiscuous oxidation reactions. To this end, we created a library of active-site variants of P450 BM3. Investigation of orthogonal high-throughput workflows identified the most streamlined route to all indigo-producing variants. Among the three known and 13 new indigo producers, eight hydroxylated 4-phenylbutan-2-one and raspberry ketone production was confirmed by NMR. In addition to validating indigo as a good predictor of this promiscuous activity, we propose a general workflow where NADPH consumption serves as a robust alternative predictor of promiscuous reactivity in P450 BM3.

2.3 Introduction

Substituted phenols are highly sought-after industrial intermediates and products.[1-2] Among these, 4-(4-hydroxyphenyl)-2-butanone, also known as rheosmin or raspberry ketone, is of high value for the flavor and fragrance industry and is under investigation for its potential health benefits.[3-6] The cost of extracting natural raspberry ketone from raspberries and other fruits (20 000$/kg) justifies the development of synthetic and biosynthetic routes to access this small molecule.[7-11] In particular, biocatalytic routes can be applied to transform natural precursors into products then considered to be 'natural', offering regulatory advantages for applications in the food and fragrance markets.[11]

Cytochrome P450 oxidases constitute a superfamily of NAD(P)H-dependent mono-oxygenases that catalyze reactions on a wide range of substrates, [12-13] including the degradation of xenobiotics and endogenous molecules.[14] They catalyze hydroxylation, epoxidation, deamination, dehalogenation, dealkylation and more, acting on non-activated carbon atoms in a single step. The breadth of their reactivity gives them great industrial importance.[15-17] Cytochrome P450 oxidase BM3 (P450 BM3), from *Bacillus megaterium*, is a self-sufficient cytochrome P450. Its reductase domain is fused to its heme domain, making it among the most efficient cytochrome P450 oxidases with respect to electron transfer. Although the primary physiological function of P450 BM3 remains unclear, fatty acids are its native substrates. The native P450 BM3 and, to a greater extent, its engineered variants, oxidize chemically diverse substrates, [15, 18-21] making P450 BM3 a versatile enzyme that offers good promiscuity for substrate discovery.

We hypothesized that P450 BM3 could be a potential candidate enzyme for the oxidative biotransformation of 4-phenyl-2-butanone (4-PB) to raspberry ketone (Figure 2-1A). The substrate, 4-PB, is a major fraction of the essential oils of the tropical Aquilaria trees and thus can be naturally sourced.[22-23] It has been demonstrated that screening cytochrome P450s for improved oxidative activity toward indigo formation can be a successful route to identifying variants that are promiscuously active toward other small aromatic substrates.[24-26]

Specifically, a study of $P450_{CAM}$ variants found a 96% correlation between the hydroxylation of indole and an unrelated substrate, diphenylmethane.[25]

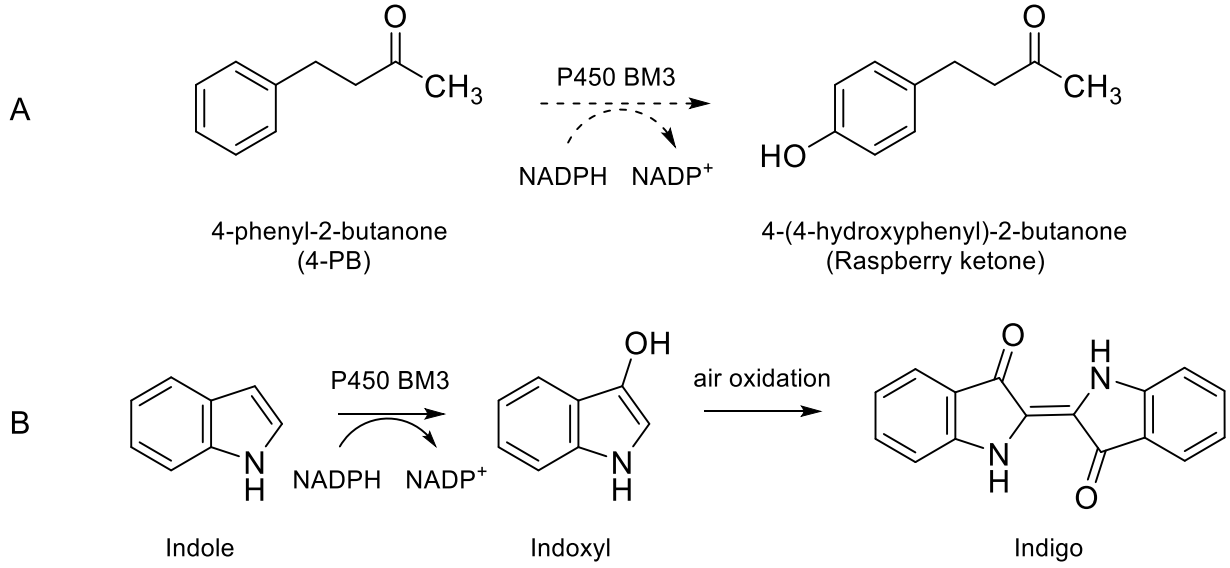


Figure 2-1. Role of cytochrome P450 BM3 in biotransformations.

A) The proposed biotransformation of 4-phenyl-2-butanone (4-PB) to raspberry ketone. B) The biotransformation of indole to indigo.

Screening assays to explore the versatility of cytochrome P450s have been developed, particularly assays that rely on chromogenic or derivatized substrates and products.[27-28] Monitoring cytochrome P450 oxidase variants for the oxidation of indole to indoxyl is particularly attractive because indoxyl undergoes spontaneous oxidative dimerization to the readily-detectable deep blue pigment indigo (Figure 2-1B). Chromogenic reactions address one of the main bottlenecks of high-throughput screening: reactivity can be detected in whole cells or in cell lysate, reducing time and cost of screening as only well-expressed, active and soluble variants are identified in a high-throughput format.

Here, we set out to verify the utility of indole oxidation in the microbial, self-sufficient P450 BM3 as a predictor of promiscuous activity for an unrelated substrate, 4-PB. Promiscuous activities are defined as the development of new functions that can be optimized through enzyme engineering. Both substrates are similar in the sense that they are small molecules and oxidation

occur on an aromatic ring. We considered important to determine whether, independently of indigo production, functional BM3 variants are equally good candidates for promiscuous activity. We compare experimental outputs to determine the most efficient workflow to identify novel P450 BM3 variants that hydroxylate 4-PB to raspberry ketone.

2.4 Results and Discussion

Indigo is among the oldest dyes and finds utility in diverse industries such as textiles, food, and pharmaceuticals.[29-31] Synthetic biology and biocatalytic solutions to indigo production have been sought in the past.[25, 32-33] Among these, engineered variants of P450 BM3 have been found to be active toward the oxidation of indole to indoxyl, and thence to indigo.[18, 20] We sought to exploit the correlation that has been observed between that oxidative transformation and the oxidation of other non-native substrates in some cytochrome P450s to identify variants active in the production of raspberry ketone.

The chemical space of P450 BM3 has been only lightly tapped into with respect to indigo-producing variants. With the aim of identifying new indigo-forming variants and broadening the search space for potential raspberry ketone production, we mutated and screened five active-site residues: V78, A82, F87, A264 and T268 (Figure 2-2). These positions were rationally selected based on distance from the heme-iron and side-chain orientation, and in some cases prior observation of indigo formation, as described below.

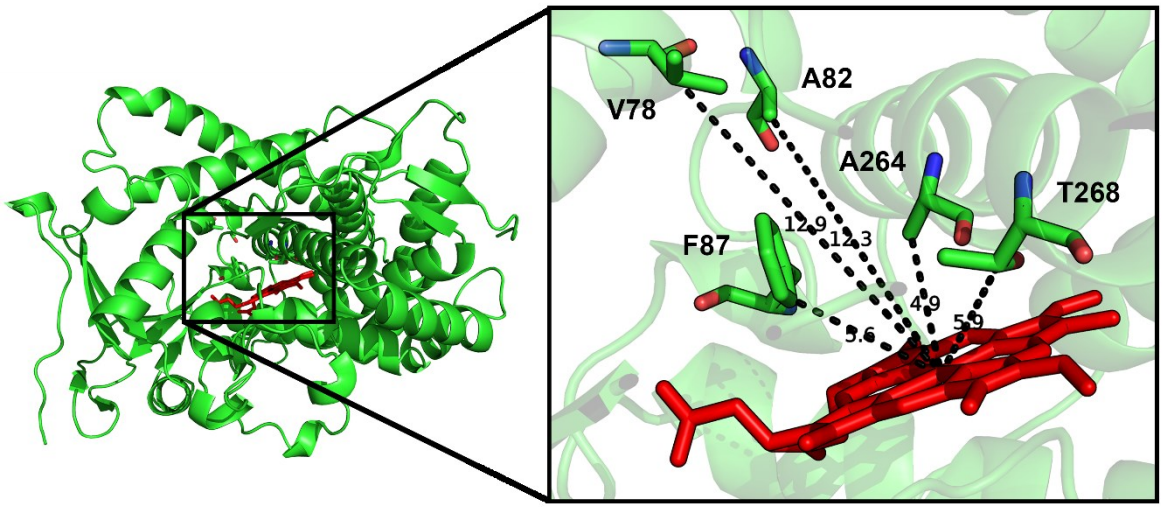


Figure 2-2. The five active-site positions selected for mutagenesis in P450 BM3.

Positions V78, A82, F87, A264 and T268 in a close-up of the active-site (PDB code 1BU7). P450 BM3 is represented in green cartoon form and the heme in red sticks. Distances from the heme iron are shown (Å).

Several variants at positions A82 and F87 have been well characterized in the oxidation of natural substrates, drugs and metabolites.[20, 34-35] Some are also known to catalyze indigo formation and to significantly modify the active site geometry; those two positions were further diversified.[18, 36-38] Some variants at position V78 oxidize small substrates, a feature that is compatible with our goal, justifying its inclusion.[39-40] Specific substitutions at position A264 displace the heme iron-coordinating water molecule, and others at the highly conserved T268 alter its role in the oxidation mechanism, demonstrating the impact of these core active-site positions on reactivity.[41-45] They were considered to be promising targets for modification. Site-saturation mutagenesis was performed at these five positions using a mix of degenerate primers (Table S2-2).[46] We obtained a coverage of 50-70% at each position, yielding a dataset of 53 variants (Table 2-1).

Table 2-1. P450 BM3 variants screened in this study.

Native residue	Variants screened	Coverage (/19)
V78	A, D, F, G, K, N, S, W, Y	9
A82	C, E, F, G, I, K, L, P, Q, R, T, V, W	13
F87	A, H, I, K, L, N, P, Q, S, V, W	11
A264	C, D, F, G, N, P, R, T, V	9
T268	A, D, E, G, K, M, N, P, S, W, Y	11

Direct visual identification of indigo-forming bacterial colonies can be done upon plating on indole-containing medium.[36] Though rapid, that method provides no quantitation. In our search for indigo-producing variants, we directly assayed clarified cell lysate for indigo

formation using a microtiter-well format to monitor absorption at 620 nm (Abs_{620}). Despite the poor water solubility of indigo, this method is sufficient to rank promising candidates. Direct absorbance in clarified lysate identified 15/53 variants producing indigo (Figure S2-5). These include three known variants, A82F, A82W and F87V, and we identified 12 unreported variants. Among these, substitutions at A82 and F87 displayed the highest activity; substitutions V78 and A264 also yielded positive hits. Only position T268 yielded no indigo-forming variants.

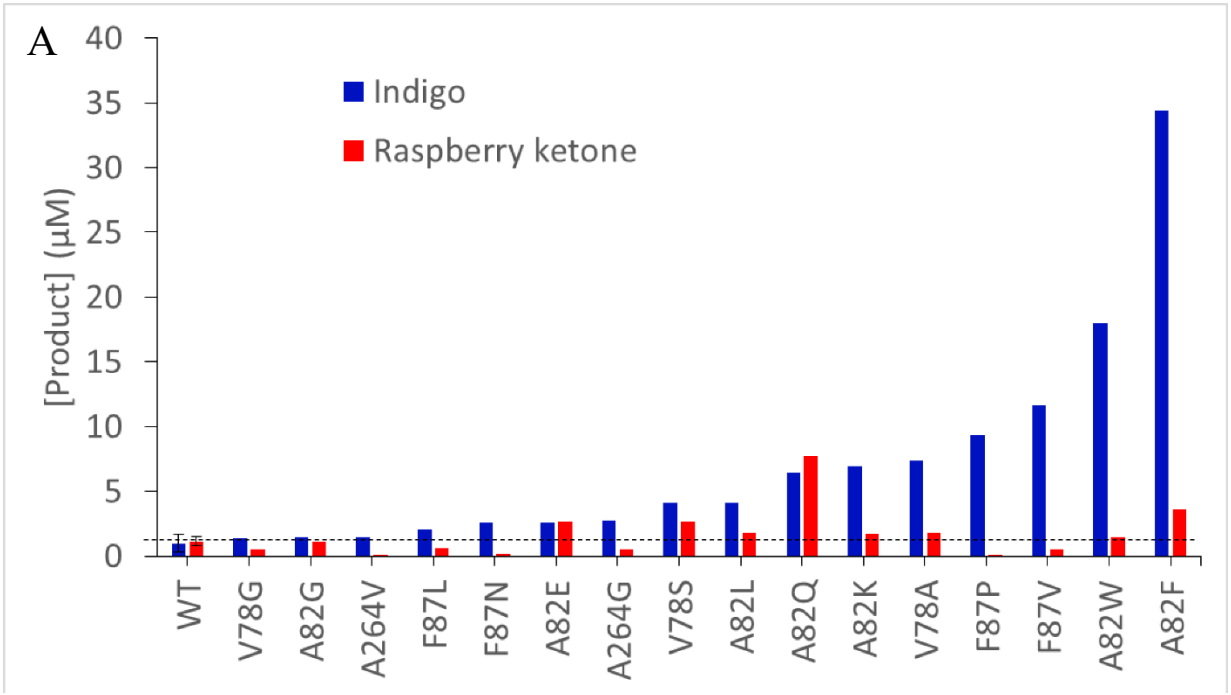
We considered that 15 indigo-forming variants out of 53 variants screened was quite successful, and asked how robust such an assay can be to predict activity for novel transformations. To this end, we established a high-throughput workflow to accelerate the discovery of highly active variants that are independent of the indigo predictor, with a focus on NADPH consumption. We quantified functional expression of the 53 variants and screened them for NADPH consumption, followed by product formation by mass spectrometry.

We identified the variants that are functional, including the heme iron, according to carbon monoxide (CO) difference spectroscopy in a microtiter plate format.[27] This rapid and reliable quantification of functional expression of each variant demonstrated that 31 out of the 53 variants were functionally expressed (Figure S2-6), including (as expected) all 15 indigo-positive variants. Under our conditions, we observed moderate variation of expression, as illustrated by the standard deviation on eight replicates of the wild-type P450 BM3. We note that SDS-PAGE analysis of protein expression in lysates identified all CO-test positive variants as being highly expressed; two further variants were highly expressed but negative in CO-testing (A82R and T268K), presumably because they did not successfully incorporate heme and/or iron (data not shown). Finally, this assay revealed one false positive, F87W, which was a weak producer in the Abs_{620} assay but was negative in the CO test.

We then optimized a NADPH-consumption fluorescence assay by adapting it to automation. This previously reported 96-well plate assay can apply to any substrate including non-colorigenic reactions.[47] Hits were defined as variants consuming NADPH faster in the presence of exogenous substrate. The high-throughput workflow was validated toward the conversion of indole. We originally performed the reaction with 0.375 mM NADPH to observe initial rates. Under those conditions, the difference in absence and in presence of the exogenous substrate did not allow the clear assignment of indole-reactive variants (data not shown). We

therefore designed an assay with 1.5 mM NADPH: initial rates are not directly observed due to saturation of the fluorescence detector, but clear rate discrimination is observed upon complete consumption of NADPH. Under these conditions, the NADPH fluorescence assay unveiled 29/31 CO-positive variants where NADPH was consumed faster in the presence than in the absence of indole (Table S2-3).

This result is substantially greater than the 15/53 indigo producers identified by Abs_{620}. We observed that indole activated NADPH consumption: all the well-expressed variants, except T268M and T268W which were poorly expressed, consumed faster in presence of this exogenous substrate. NADPH consumption in the bacterial lysates can result from the activity of other NADPH-dependent enzymes, from activity of P450 BM3 with endogenous substrates such as fatty acids or as a result of uncoupling of P450 BM3 to form H_2O_2, superoxide anion and water.[18, 48-51] Moreover, uncoupling can be substrate dependent: the T268A variant of P450 BM3 is known to exhibit uncoupling in the presence of C_{16} but not C_{14} fatty acid substrate.[52] The assay may therefore not be a direct reflection of product formation and requires validation of indole-activated NADPH-consuming variants by mass spectrometry.



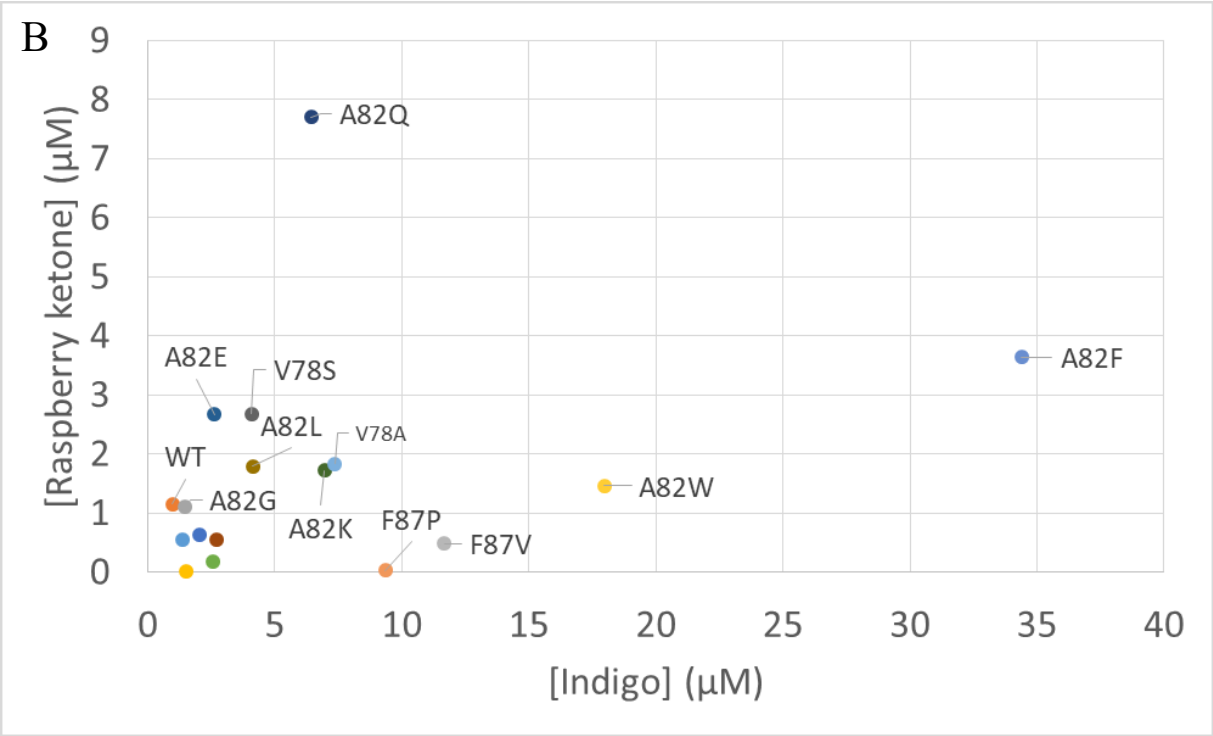


Figure 2-3. Quantification of indigo by LC-MS (blue) and raspberry ketone by GC-MS (red) for the 16 best indigo producing variants of P450 BM3.

A) Representation of the product concentration with a bar plot. The lowest indigo and raspberry ketone concentration in the linear portion of the calibration curve was 1.25 µM (black dashed), defining the lower limit of detection for each method. Product concentration in WT is the average eight replicates with standard deviation. B) Representation of the product concentration with a scatter plot.

Using LC-MS, we confirmed that 16 out of the 29 NADPH fluorescence hits were indigo-producing variants (Figure 2-3; Table S2-3; Figure S2-7). The higher sensitivity of LC-MS than Abs_{620} led to the detection of two additional variants, A264V and A82G, which were among the weakest indigo producers (Figure S2-8; Table S2-3). This demonstrates that the straightforward Abs_{620} assay rapidly identified nearly all the indigo producers, serving as a rapid and robust screen that is compatible with large libraries (Figure 2-4). Prolonging the Abs_{620} assay beyond 2 h did not reveal further indigo producers, but shortening the assay reduced the sensitivity of the method. LC-MS being more sensitive, it identified the weakly active variants. These variants can also be good starting point to evolve the enzyme toward new functions.

Remarkably, more than half of the NADPH-consuming variants produced indigo, demonstrating that active-site modification of P450 BM3 to oxidize indole is readily achievable.

The previously reported variants A82F, A82W and F87V were among the most efficient indigo-producing variants. Several of the best indigo producers are new variants at those two positions (A82 and F87), such as F87P. It has been shown that aggressive substitution in P450 BM3, including substitution to proline, can be accommodated and result in unexpected and desirable outcomes.[51] In addition, we report the first indigo-producing variants of positions V78 and A264. V78A, the fifth best indigo producer we identified, was one of several mutations in variants reported to hydroxylate small alkanes [40] or alter hydroxylation regiospecificity on C_{12} fatty acids.[53] Both A264 variants are weak indigo producers. Previously reported A264 substitutions to E, H, K, M, Q, or C all altered the heme iron ligand sets, inactivating the enzyme. [41-43, 54] However, within multiply-substituted variants, A264G contributes to the oxidation of polycyclic aromatic hydrocarbons [55] whereas A264V contributes to the oxidation of a terpene.[56] The two indigo-producing variants of A264 identified here harbour those same conservative substitutions (A264G/V) and constitute, to our knowledge, the first report of point-substituted reactive variants of A264.

Quantification of indigo production allowed us to address the long-standing debate of whether the wild-type P450 BM3 (WT) does [18, 57-58] or does not [36-37] hydroxylate indole. Absorption at 620 nm in bacterial lysates (n=8) was not consistently observed within the linear range of quantification, confirming that WT is, at best, a poor indole oxidizer. Nonetheless, LC-MS unequivocally confirmed production of indigo (Figure S2-8). Three out of eight replicates of the WT held indigo concentrations greater than the lowest indigo standard we verified (1.25 μM), establishing indigo formation by WT near the limit of quantification for this method.

Our identification of 3 known and 13 previously unreported indigo-producing variants demonstrates that there are numerous engineering solutions to achieve indigo production in this enzyme. Based on this, P450 BM3 may not need tremendous engineering efforts to evolve for oxidation reactions on a number of non-native substrates. We verified whether the previously reported correlation between oxidation of indole and of other substrates [24-26] also applies to the oxidation of 4-PB. WT has not been reported to synthesize raspberry ketone nor have any of its variants. GC-MS confirmed that four out of eight replicates of the WT held hydroxylated 4-PB concentrations greater than the lowest standard we verified (1.25 μM), establishing hydroxylated 4-PB formation by WT BM3 near the limit of quantification for this method.

We undertook the fluorescence-based NADPH assay with the exogenous substrate 4-PB. We identified 23/31 BM3 variants exhibiting higher NADPH consumption in the presence of 4-PB. Hydroxylation of 4-PB was confirmed by GC-MS for 8/23 variants (Figure 2-4; Table S2-3; Figure S2-9). As for indole oxidation, this demonstrates that substrate-specific NADPH consumption overestimates product formation. These samples have the same mass and retention time as the raspberry ketone standard, consistent with *p*-hydroxylation of the phenyl ring. We confirmed the production of raspberry ketone as the sole oxidation product by prep LC and ^{1}H-NMR with a scaled-up reaction of the most efficient variant, A82Q (Figure S2-10). Strikingly, A82Q and the seven further variants that hydroxylate 4-PB are all indigo producers. Our results support the initial hypothesis that indigo is a good predictor for raspberry ketone production. We further demonstrate that Abs_{620} is sufficient to identify promiscuous variants (Figure 2-4).

Although the fluorescence-based NADPH consumption assay overestimated the number of variants of interest, it held a second predictor of promiscuity. Three distinct NADPH consumption patterns were observed: A) No observable consumption; B) faster consumption in presence then in absence of exogenous substrate; C) very fast consumption both in presence and absence of exogenous substrate (Figure S2-11). The latter was observed for 9 variants (Table S2-3). Variants that rapidly consumed NADPH in absence of exogenous substrate could have been thought to exhibit high uncoupling. However, 8/9 fast NADPH consumers were active for indigo formation and 5/9 were active for raspberry ketone including the most active variant, A82Q (Figure 2-4; Table S2-4). Performing the assay at a lower concentration of NADPH (0.375 mM instead of 1.5 mM) also successfully identified 6 variants that consume NADPH rapidly in absence of exogenous substrate, where 6/6 were indigo producers and 5/6 were raspberry ketone producers (data not shown). This suggests that fast NADPH consumption is a good indicator of activity toward promiscuous substrates.

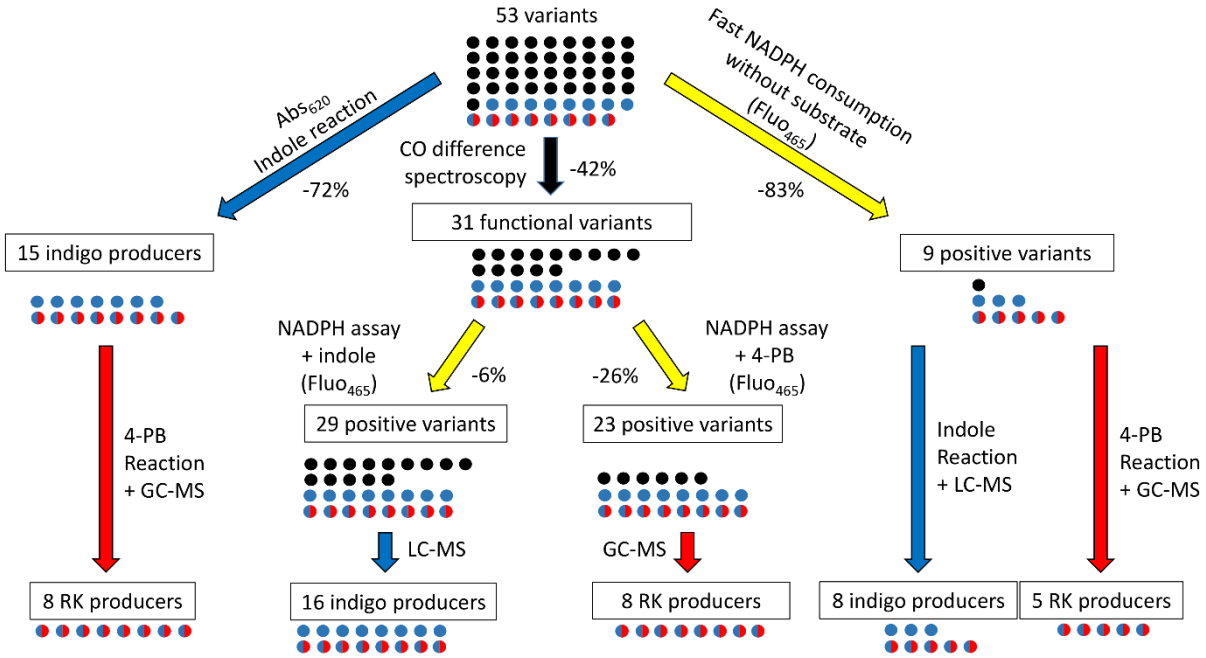


Figure 2-4. Comparison of workflows to enrich the pool of P450 BM3 variants with indigo or raspberry ketone (RK) producers.

Black, blue and blue/red circles represent negative variants, indigo-only producers and indigo/raspberry ketone producers, respectively. Blue, red and yellow arrows represent methods that identify indigo formation, raspberry ketone formation and NADPH consumption, respectively. The fraction of variants eliminated at each step is given beside each arrow as a negative percentage.

This work, along with previous reports, clearly demonstrate that indigo production is a predictor of several promiscuous reactions.[24-26] We have identified a number of new indigo and raspberry ketone-forming variants using readily accessible methods. Determination of Abs_{620} in bacterial lysate served as a rapid screening tool that is compatible with large libraries. It reduced by 72% the number of samples to analyze by MS for indigo and raspberry ketone production with only one false positive. We note that the three previously known indigo producers did not include the top raspberry ketone producer, demonstrating the importance of broadening the test-set of indigo-positive variants.

The breadth of new reactivities that can be predicted on the basis of indigo production is unknown. If indigo is not a good predictor for a desired reaction, the workflow consisting of the CO-difference assay, the fluorescence-based NADPH assay and MS quantitation is a robust

alternative. In our case, the workflow reduced by 45% and 57% the number of samples to analyze by MS for indigo and raspberry ketone production respectively. The fast NADPH-consuming variants reduced the need for MS analysis by 80% for both reactions; it failed to identify nearly 50% of positive hits but included several among the best. The workflow provides a universal platform for screening non-colorigenic reactions and is amenable to very high-throughput screening. These methods were successful for the discovery of new indigo producers as well as demonstrating, for the first time, that P450 BM3 can be engineered to produce raspberry ketone.

2.5 Experimental Section

Materials

The primers for mutagenesis were obtained from Sigma-Aldrich. Phusion Hot start II polymerase, purification kit, gel extraction kit and miniprep kit were obtained from New England Biolabs. FastDigest restriction enzymes and fast alkaline phosphatase were obtained from Thermo Fisher Scientific. TAKARA ligase was obtained from ClonTech. The cloning vector, pcWORI BM3 WT, was kindly provided by Prof. Frances Arnold (Caltech). Indole was obtained from Alfa Aesar. 4-phenyl-2-butanone, raspberry ketone and sodium hydrosulfite were obtained from Sigma-Aldrich. FastBreak lysis buffer was obtained from Promega.

PCR were performed with a T100 thermal cycler from Bio-Rad. Cultures were shaken using a Marshall Scientific Labline 4625 titer shaker. Screening was performed with a Beckman Coulter Biomek NX^P robot coupled to a DTX880 multimode detector plate reader. LC-MS analysis was performed on a Thermo Scientific Dionex UltiMate 3000 and masses were detected with a Finnigan TSQ Quantum Ultra mass detector. GC-MS analysis was performed with an Agilent 7890A and masses were detected with an Agilent 5975C with triple-axis detector. Preparative LC was performed on a Waters apparatus with a SQ Detector 2. Sequencing was operated by the Institute Research of Immunology and Cancerology.

Mutagenesis conditions

Individual mutagenesis reactions were carried out for each position. A mixture of four primers encoding: NDT, VMA, ATG or TGG at the position of interest was used to perform saturation site-directed mutagenesis at each of 5 position: V78, A82, F87, A264 and T268. Primer sequences can be found in the Supporting Information (Table S2-1). The megaprimer method of mutagenesis was applied.[59] PCR intermediates were purified before proceeding with further steps. Final PCR products were digested using BamHI and SacI according to the manufacturer's instructions. DpnI was used to eliminate parental DNA. The gel-purified inserts were ligated into the BamHI, SacI and Fast-AP treated pcWORI BM3 WT vector. Chemically competent *E.coli* BL21 (DE3) cells were transformed with each ligation and plated on Luria-Bertani (LB) agar with ampicillin (100 μg/mL).

Culture conditions and lysis

Unless otherwise stated, precultures were grown in 2 mL deep-well 96-well plates, in 1 mL LB with ampicillin (100 μg/mL) for 16 h at 37°C, shaking at intensity 5. Expression was carried in the same format, in 1 mL terrific broth (TB) supplemented with ampicillin (100 μg/mL), δ-aminolevulinic acid (12.5 μg/mL), thiamine (5 μg/mL), 0.125 μL trace metals (0.5g $MgCl_2$, 50g $FeCl_2{\cdot}6H_2O$, 1g $ZnCl_2{\cdot}4H_2O$, 0.2g $CoCl_2{\cdot}6H_2O$), 1g $Na_2MoO_4{\cdot}2H_2O$, 0.5g $CaCl_2{\cdot}2H_2O$, 1g $CuCl_2$, 0.2g H_2BO_5 in 1L HCl solution (90% v/v distilled water: concentrated HCl), and 20 μL of preculture. The cultures were propagated as above until O.D. reached 0.7-1.0, induced with 1 mM isopropyl-β-D-thiogalactoside (IPTG) and further propagated for 20 h at 22°C with shaking. Cultures were pelleted at 1643 g, 4°C for 20 min and pellets were frozen for at least 2 h at -80°C. Expression pellets were thawed on ice and resuspended in 400 μL lysis buffer (10mM $MgSO_4$, 1mM DDT, 0.5 mg/mL lysozyme (Sigma), DNase, 1.5 mM Benzamidine hydrochloride hydrate (Fisher), 0.25 mM PMSF (Sigma) in phosphate buffer 0.1M pH 8). Samples were shaken at room temperature for 2 h at intensity 8. Lysates were centrifuged at 1643 g, 4°C for 20 minutes and supernatants were used for further steps.

CO difference spectroscopy

A CO chamber was built by modifying a standard Pyrex desiccator to include a custom lid with an adjustable gas inlet and outlet. Quantification of the P450 BM3 variants was performed

based on a previously described protocol.[27] Lysate (160 µL) of each variant was reduced with 40 µL of 0.3 M sodium dithionite in 1 M phosphate buffer pH 8. They were quantified simultaneously in a 96-well plate. A reference measurement was taken at 485 nm. Then, the plate was placed in the CO chamber with a positive CO pressure for 15 seconds before releasing the pressure. This cycle was repeated 2 times. A last cycle was performed with CO held for 5 minutes. The CO difference spectroscopy measurement was immediately taken at 485 nm.

High-throughput absorbance and fluorescence screening assay

Assays were carried out in a 96-well plate format using 7.5 µL of bacterial lysate, 5 mM substrate (indole or 4-phenyl-2-butanone) in DMSO (2% final DMSO), 1.5 mM NADPH in a final volume of 140 µL with phosphate buffer 0.1 M pH 8. A reference plate was analyzed under the same conditions, where substrate was replaced with DMSO. Every 30 min, an absorbance measurement was taken at 620 nm for indigo formation and a fluorescence measurement at 465 nm (excitation 340 nm) for NADPH consumption was taken in parallel, for 16 hours.

LC-MS/GC-MS analysis

The indigo reactions were quenched with 300 µL acetonitrile and 15 µL of FastBreak lysis buffer. Raspberry ketone reactions were quenched with 300 µL ethyl acetate. All reactions were shaken for 20 min and centrifuged at 1643 g, room temperature, for 20 min to remove insoluble matter. Supernatant (100 µL) was transferred to a MS vial for analysis.

For LC-MS of indigo, a Phenomenex RP-polar 2 mm x 150 mm, 4 µm column was used at 50°C oven temperature. We injected 5 µL and applied a gradient of (A) H_2O 0.1% formic acid / (B) ACN 0.1% formic acid from 40-95% B over 5 min was applied and held for 3 min, then equilibrated at 60% A: 40% B for 2 min. Masses were detected under positive ionization and single-ion monitoring of indole (118 *m/z*), indigo (263 *m/z*) and full scan were taken.

For GC-MS of raspberry ketone, we used an Agilent J&W 122-7033 DB-wax column. We injected 1 µL with an inlet temperature of 270°C, pressure of 11.8 psi and a flow of 1.2 mL/min. A temperature gradient from 80°C to 240°C in 8.89 min was applied and held hold at 240°C for 16 min. Masses were detected under positive ionization and single-ion monitoring of 4-phenyl-2-butanone (121 *m/z*) and the fragment of raspberry ketone (107 *m/z*).

NMR identification of raspberry ketone

The 4-phenyl-2-butanone reaction was scaled-up in a 300 mL Erlenmeyer using 2.5 mL of lysate from variant A82Q, 1mL of 250 mM 4-phenyl-2-butanone, 5mL of 15 mM NADPH and 41.5 mL of 0.1 M potassium phosphate buffer pH 8 in a total volume of 50 mL. The reaction was performed for 16 hours at 25°C with shaking at 250 rpm. The product was extracted with 100 mL of ethyl acetate and evaporated at 40°C on a rotary evaporator and taken up in 1 mL of ethyl acetate for analysis.

Isolation of the hydroxylated product(s) was done by preparative LC with a Phenomenex RP-polar 19 mm x 150 mm, 4µm column. We injected 8 x 100 µL of the sample. An initial gradient of (A) H_2O 0.1% formic acid / (B) MeOH 0.1% formic acid from 85-33% B in 13.5 min was applied and held at 10% A: 90% B for 1 min, then equilibrated at 15% A: 85% B for 5 min. We monitored the mass of the raspberry ketone fragment as a sodium adduct (107.1 + 22 *m/z*). Fractions containing the correct mass and retention time were collected, combined and evaporated for NMR analysis.

^{1}H NMR (700 MHz, $CDCl_3$, 25°C, TMS): δ = 7.06 (m, 2H), δ = 6.76 (m, 2H), δ = 2.83 (t, *J* = 7.5 Hz, 2H), δ = 2.72 (t, *J* = 7.5, 2H), δ = 2.16 (s, 3H) ppm.

2.6 Acknowledgements

All authors belong to the Québec Network for Protein Function, Engineering and Applications (PROTEO) and the Center of Green Chemistry and Catalysis (CGCC). We thank Marie-Christine Tang, Louiza Mahrouche and Dr. Alexandra Furtos from the mass spectrometry laboratory of the Université de Montréal for invaluable help; the Université of Montréal Chemistry workshop for instrument design and technical assistance; Prof. Frances Arnold for providing the P450 BM3 gene; and Jacynthe L. Toulouse, Vanessa Kairouz, Cedric Malveau and Dr. Pedro Alguiar for helpful discussions and suggestions. This work was funded by Natural Sciences and Engineering Council (NSERC) grant 227853 and the Fonds Québecois de la Recherche sur la Nature et les Technologies (FQRNT) grant 181877. O.R. held a PROTEO scholarship, M.C.J.C.C. was the recipient of NSERC, PROTEO and UdeM scholarships and S.I. held a scholarship from CGCC.

2.7 Conflicts of interest

The authors declare no conflict of interest.

2.8 References

[1] D. Lin, et al., *Molecules* **2016**, *21*, 1374-1393.
[2] C. Caleja, A. Ribeiro, M. F. Barreiro, I. C. F. R. Ferreira, *Curr. Pharm. Des.* **2017**, *23*, 2787-2806.
[3] Y. C. Tsai, B. C. Yang, W. H. Peng, Y. M. Lee, M. H. Yen, P. Y. Cheng, *Phytomedicine* **2017**, *31*, 11-17.
[4] V. Khan, S. Sharma, U. Bhandari, S. M. Ali, S. E. Haque, *Life Sci.* **2018**, *194*, 205-212.
[5] K. S. Park, *Planta Med.* **2010**, *76*, 1654-1658.
[6] C. Ulbricht, et al., *Altern. Complement Ther.* **2013**, *19*, 98-100.
[7] J. Beekwilder, I. M. van der Meer, O. Sibbesen, M. Broekgaarden, I. Qvist, J. D. Mikkelsen, R. D. Hall, *Biotechnol. J.* **2007**, *2*, 1270-1279.
[8] L. R. Smith, *Chem. Educ.* **1996**, *1*, 1-18.
[9] D. Lee, N. D. Lloyd, I. S. Pretorius, A. R. Borneman, *Microb. Cell Fact.* **2016**, *15*, 49.
[10] W. F. Hoelderich, V. Ritzerfeld, *Appl. Catal., A* **2015**, *504*, 654-663.
[11] B. Kosjek, W. Stampfer, R. v. Deursen, K. Faber, W. Kroutil, *Tetrahedron* **2003**, *59*, 9517-9521.
[12] F. Hannemann, A. Bichet, K. M. Ewen, R. Bernhardt, *Biochim. Biophys. Acta, Gen. Subj.* **2007**, *1770*, 330-344.

[13] P. R. Ortiz de Montellano, *Cytochrome P450 : structure, mechanism, and biochemistry*, Springer, **2015**.
[14] P. Anzenbacher, E. Anzenbacherova, *Cell. Mol. Life Sci.* **2001**, *58*, 737-747.
[15] V. B. Urlacher, M. Girhard, *Trends Biotechnol.* **2012**, *30*, 26-36.
[16] S. Schulz, M. Girhard, V. B. Urlacher, *ChemCatChem* **2012**, *4*, 1889-1895.
[17] E. T. Farinas, M. Alcalde, F. H. Arnold, *Tetrahedron* **2004**, *60*, 525-528.
[18] W. C. Huang, A. C. G. Westlake, J. D. M. M. Gordon Joyce, P. C. E. Moody, G. C. K. Roberts, *J. Mol. Biol.* **2007**, *373*, 633-651.
[19] R. T. Ruettinger, L. P. Wen, A. J. Fulco, *J. Biol. Chem.* **1989**, *264*, 10987-10995.
[20] C. J. C. Whitehouse, S. G. Bell, L. L. Wong, *Chem. Soc. Rev.* **2012**, *41*, 1218-1260.
[21] C. J. C. Whitehouse, S. G. Bell, H. G. Tufton, R. J. P. Kenny, L. C. I. Ogilvie, L.-L. Wong, *Chem. Commun.* **2008**, 966-968.
[22] A. S. Norul, M. Nazira, S. N. Tajuddin, *Aust. J. Basic Appl. Sci.* **2015**, *9*, 155-159.
[23] A. Bahan, et al., *Malaysian J. Anal. Sci.* **2013**, *17*, 403-413.
[24] P. P. Kelly, A. Eichler, S. Herter, D. C. Kranz, N. J. Turner, S. L. Flitsch, *Beilstein J. Org. Chem.* **2015**, *11*, 1713-1720.
[25] A. Celik, R. E. Speight, N. J. Turner, *Chem. Commun.* **2005**, 3652-3654.
[26] S. H. Park, et al., *Drug Metab. Dispos.* **2010**, *38*, 732-739.
[27] F. H. Arnold, G. Georgiou, *Directed Enzyme Evolution : Screening and Selection Methods*, Humana Press, Totowa, **2003**.
[28] J. B. Behrendorff, W. Huang, E. M. Gillam, *Biochem. J.* **2015**, *467*, 1-15.
[29] J. C. Splitstoser, T. D. Dillehay, J. Wouters, A. Claro, *Sci. Adv.* **2016**, *2*, 1-4.
[30] B. D. Ensley, B. J. Ratzkin, T. D. Osslund, M. J. Simon, L. P. Wackett, D. T. Gibson, *Science* **1983**, *222*, 167-169.
[31] C. S. W. Koehler, *Today's Chemist at Work* **1999**, *8*, 85-91.
[32] L. Cheng, S. Yin, M. Chen, B. Sun, S. Hao, C. Wang, *Curr. Microbiol.* **2016**, *73*, 248-254.
[33] E. M. Gillam, et al., *Biochem. Biophys. Res. Commun.* **1999**, *265*, 469-472.
[34] A. M. Sawayama, M. M. Y. Chen, P. Kulanthaivel, M.-S. Kuo, H. Hemmerle, F. H. Arnold, *Chemistry* **2009**, *15*, 11723-11729.
[35] E. Vottero, V. Rea, J. Lastdrager, M. Honing, N. P. E. Vermeulen, J. N. M. Commandeur, *J. Biol. Inorg. Chem.* **2011**, *16*, 899-912.
[36] Q. S. Li, U. Schwaneberg, P. Fischer, R. D. Schmid, *Chemistry* **2000**, *6*, 1531-1536.
[37] H. M. Li, L. H. Mei, V. B. Urlacher, R. D. Schmid, *Appl. Biochem. Biotechnol.* **2008**, *144*, 27-36.
[38] Z. Pengpai, H. Sheng, M. Lehe, L. Yinlin, J. Zhihua, H. Guixiang, *Appl. Biochem. Biotechnol.* **2013**, *171*, 93-103.
[39] M. M. Y. Chen, C. D. Snow, C. L. Vizcarra, S. L. Mayo, F. H. Arnold, *Protein Eng., Des. Sel.* **2012**, *25*, 171-178.
[40] M. W. Peters, P. Meinhold, A. Glieder, F. H. Arnold, *J. Am. Chem. Soc.* **2003**, *125*, 13442-13450.
[41] H. M. Girvan, H. E. Seward, H. S. Toogood, M. R. Cheesman, D. Leys, A. W. Munro, *J. Biol. Chem.* **2007**, *282*, 564-572.
[42] H. M. Girvan, et al., *J. Biol. Chem.* **2004**, *279*, 23274-23286.
[43] H. M. Girvan, et al., *Biochem. J.* **2009**, *417*, 65-76.
[44] P. S. Coelho, E. M. Brustad, A. Kannan, F. H. Arnold, *Science* **2013**, *339*, 307-310.
[45] K. D. Dubey, B. Wang, M. Vajpai, S. Shaik, *Chem. Sci.* **2017**, *8*, 5335-5344.

[46] L. X. Tang, H. Gao, X. C. Zhu, X. Wang, M. Zhou, R. X. Jiang, *BioTechniques* **2012**, *52*, 149-+.
[47] G. E. Tsotsou, A. E. G. Cass, G. Gilardi, *Biosens. Bioelectron.* **2002**, *17*, 119-131.
[48] P. J. Loida, S. G. Sligar, *Biochemistry* **1993**, *32*, 11530-11538.
[49] D. Holtmann, F. Hollmann, *ChemBioChem* **2016**, *17*, 1391-1398.
[50] S. Kadkhodayan, E. D. Coulter, D. M. Maryniak, T. A. Bryson, J. H. Dawson, *J. Biol. Chem.* **1995**, *270*, 28042-28048.
[51] C. J. Whitehouse, et al., *ChemBioChem* **2010**, *11*, 2549-2556.
[52] M. J. Cryle, J. J. D. Voss, *ChemBioChem* **2008**, *9*, 261-266.
[53] M. Dietrich, T. A. Do, R. D. Schmid, J. Pleiss, V. B. Urlacher, *J. Biotechnol.* **2009**, *139*, 115-117.
[54] M. G. Joyce, H. M. Girvan, A. W. Munro, D. Leys, *J. Biol. Chem.* **2004**, *279*, 23287-23293.
[55] A. B. Carmichael, L. L. Wong, *Eur. J. Biochem.* **2001**, *268*, 3117-3125.
[56] A. Seifert, M. Antonovici, B. Hauer, J. Pleiss, *ChemBioChem* **2011**, *12*, 1346-1351.
[57] Q.-S. Li, J. Ogawa, R. D. Schmid, S. Shimizu, *Biosci. Biotechnol. Biochem.* **2005**, *69*, 293-300.
[58] Y. Lu, L. Mei, *J. Ind. Microbiol. Biotechnol.* **2007**, *34*, 247-253.
[59] J. Sanchis, et al., *Appl. Microbiol. Biotechnol.* **2008**, *81*, 387-397.

2.9 Supporting Information

Supplemental Tables

Table S2-2. Design of the degenerate primers for positions V78, A82, F87, A264 and T268.

V78(AHN): 5’ – CCTGCAAAATCACG**AHN**AAATTTAAGCGC – 3’
V78(TKB): 5’ – CCTGCAAAATCACG**TKB**AAATTTAAGCGC – 3’
V78(CAT): 5’ – CCTGCAAAATCACG**CAT**AAATTTAAGCGC – 3’
V78(CCA): 5’ – CCTGCAAAATCACG**CCA**AAATTTAAGCGC – 3’
A82(AHN): 5’ – CCCGTCTCC**AHN**AAAATCACGTAC – 3’
A82(TKB): 5’ – CCCGTCTCC**TKB**AAAATCACGTAC – 3’
A82(CAT): 5’ – CCCGTCTCC**CAT**AAAATCACGTAC – 3’
A82(CCA): 5’ – CCCGTCTCC**CCA**AAAATCACGTAC – 3’
F87(AHN): 5’ – GCGTCCAGCTTGT**AHN**TAACCCGTCTCC – 3’
F87(TKB): 5’ – GCGTCCAGCTTGT**TKB**TAACCCGTCTCC – 3’
F87(CAT): 5’ – GCGTCCAGCTTGT**CAT**TAACCCGTCTCC – 3’
F87(CCA): 5’ – GCGTCCAGCTTGT**CCA**TAACCCGTCTCC – 3’
A264(AHN): 5’ – GTTTCGTGTCC**AHN**AATTAAGAATGTAATAATTTG – 3’
A264(TKB): 5’ – GTTTCGTGTCC**TKB**AATTAAGAATGTAATAATTTG – 3’
A264(CAT): 5’ – GTTTCGTGTCC**CAT**AATTAAGAATGTAATAATTTG – 3’
A264(CCA): 5’ – GTTTCGTGTCC**CCA**AATTAAGAATGTAATAATTTG – 3’
T268(AHN): 5’ – GACCACTTGT**AHN**TTCGTGTCCCGC – 3’
T268(TKB): 5’ – GACCACTTGT**TKB**TTCGTGTCCCGC – 3’
T268(CAT): 5’ – GACCACTTGT**CAT**TTCGTGTCCCGC – 3’
T268(CCA): 5’ – GACCACTTGT**CCA**TTCGTGTCCCGC – 3’

For each selected position, a combination of 4 optimized primers was used to theoretically obtain a uniform distribution of mutations. AHN (NDT) encodes: SER, ARG, LEU, GLY, VAL, ILE, HIS, ASP, TYR, ASN, CYS, PHE. TKB (VMA) encodes: LYS, GLN, GLU, THR, PRO, ALA. CAT (ATG): MET. CCA (TGG): TRP. N represents the four nucleotides; H represents C, T, A; K represents T, G; and B represents T, G, C.

Table S2-3. Quantification of the WT and variants for P450 concentration as determined by the CO test, indigo production by LC-MS and raspberry ketone (RK) production by GC-MS.

Variant [a]	[P450] (μM)	Fluorescence-based NADPH assay [b]	
		[Indigo] (μM) [c, d]	[RK] (μM) [c, d]
V78A	2,44	7,37	1,83
V78D	0,03		
V78F	2,91	0,28	0,43
V78G	2,43	1,37	0,55
V78K	0,13		
V78N	0,01		
V78S	2,80	4,08	2,68
V78W	2,74	0,76	
V78Y	2,21	0,10	
A82C	2,91	0,40	0,65
A82E	1,97	2,61	2,67
A82F	2,82	34,4	3,63
A82G	2,45	1,46	1,10
A82I	0		
A82K	1,98	6,94	1,72
A82L	2,60	4,12	1,79
A82P	0		
A82Q	2,51	6,42	7,70
A82R	0		
A82T	0		
A82V	0,90	0,34	
A82W	2,41	18,0	1,46
F87A	0		
F87H	0		
F87I	0		
F87K	0		
F87L	2,47	2,05	0,62
F87N	2,85	2,57	0,18

Variant	[P450] (μM)	Fluorescence-based NADPH assay	
		[Indigo] (μM)	[RK] (μM)
F87P	1,63	9,35	0,03
F87Q	0,01		
F87S	2,65	1,09	0,10
F87V	2,63	11,7	0,48
F87W	0		
A264C	0		
A264D	0,93	0,39	
A264F	1,66	0	
A264G	2,59	2,72	0,55
A264N	0		
A264P	0		
A264R	0		
A264T	1,87	0,39	
A264V	2,12	1,49	0,00
T268A	2,01	0,27	0,06
T268D	0		
T268E	2,28	0,11	0,00
T268G	1,31	0,14	0,13
T268K	0,15		
T268M	0,58		
T268N	0		
T268P	0		
T268S	2,20	0,69	0,60
T268W	0,50		
T268Y	0		
WT [d]	1,60	0,98	1,15
Puc18 [e]	0,00	0,00	0,00

[a] Yellow: The fast NADPH-consuming variants without substrate.
[b] Grey: Variants that consumes NADPH faster in presence of substrate were carried forward for LC-MS determination of indigo concentration and for GC-MS determination of raspberry ketone concentration.
[c] The lowest indigo and raspberry ketone concentration in the linear portion of the calibration curve was 1.25 μM, defining the lower limit of detection for each method.
[d] The WT value is the average of eight replicates, with standard deviation.
[e] Puc18 is the negative control lacking a P450 BM3 gene.

Table S2-4. Comparison of the phenotype determination methods for the 16 best indigo producing variants[a]

		Abs_{620} [b]	LC-MS (indigo)[c]	GC-MS (raspberry ketone)[d]	Fastest NADPH, (indigo) [e]	Fastest NADPH, (raspberry ketone) [f]
1	A82F					
2	A82W					
3	F87V					
4	F87P					
5	V78A					
6	A82K					
7	A82Q*					
8	A82L					
9	V78S					
10	A264G					
12	A82E					
13	F87N					
14	F87L					
15	A264V					
16	A82G					
17	V78G					

[a] The variants are ranked according to the concentration of indigo determined by LC-MS. The top raspberry ketone producer, A82Q, is marked with an asterisk.

[b] Blue: indigo-producing variants according to Abs_{620}.

[c] Blue: indigo-producing variants according to LC-MS analysis.

[d] Red: raspberry ketone-producing variants according to GC-MS analysis.

[e, f] Variants consuming NADPH very fast in absence of exogenous substrate, colored according to indigo or raspberry ketone production.

Supplemental Figures

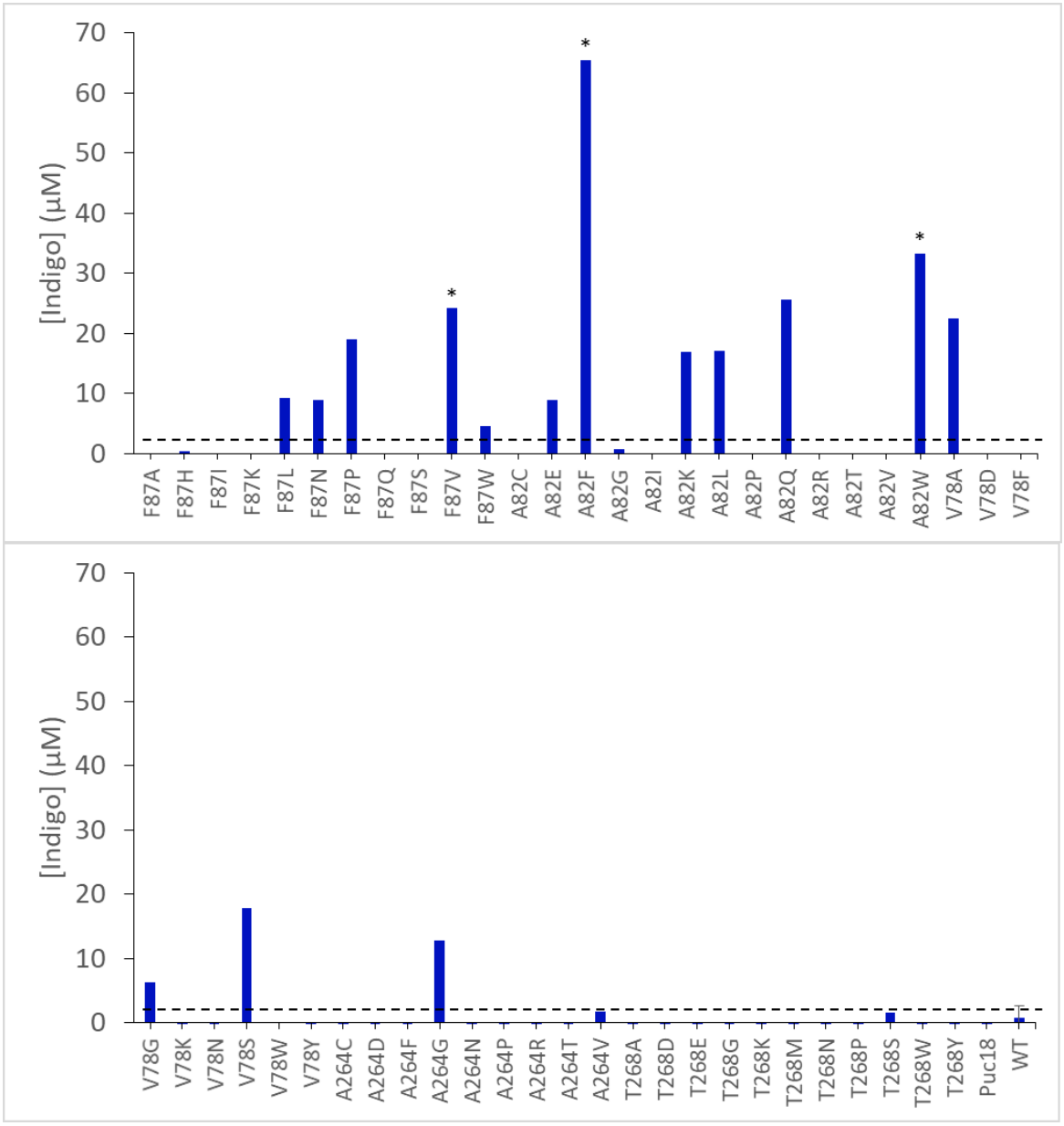


Figure S2-5. Direct absorbance at 620 nm for indigo synthesis in clarified lysate of 53 variants.

Asterisks identify the three previously reported indigo producing variants: A82F, A82W and F87V. The lowest indigo concentration in the linear portion of the calibration curve was 2.34 μM for Abs_{620} (black dashed line), defining the lower limit of detection. Indigo concentration in WT shown with standard deviation, was averaged from eight replicates.

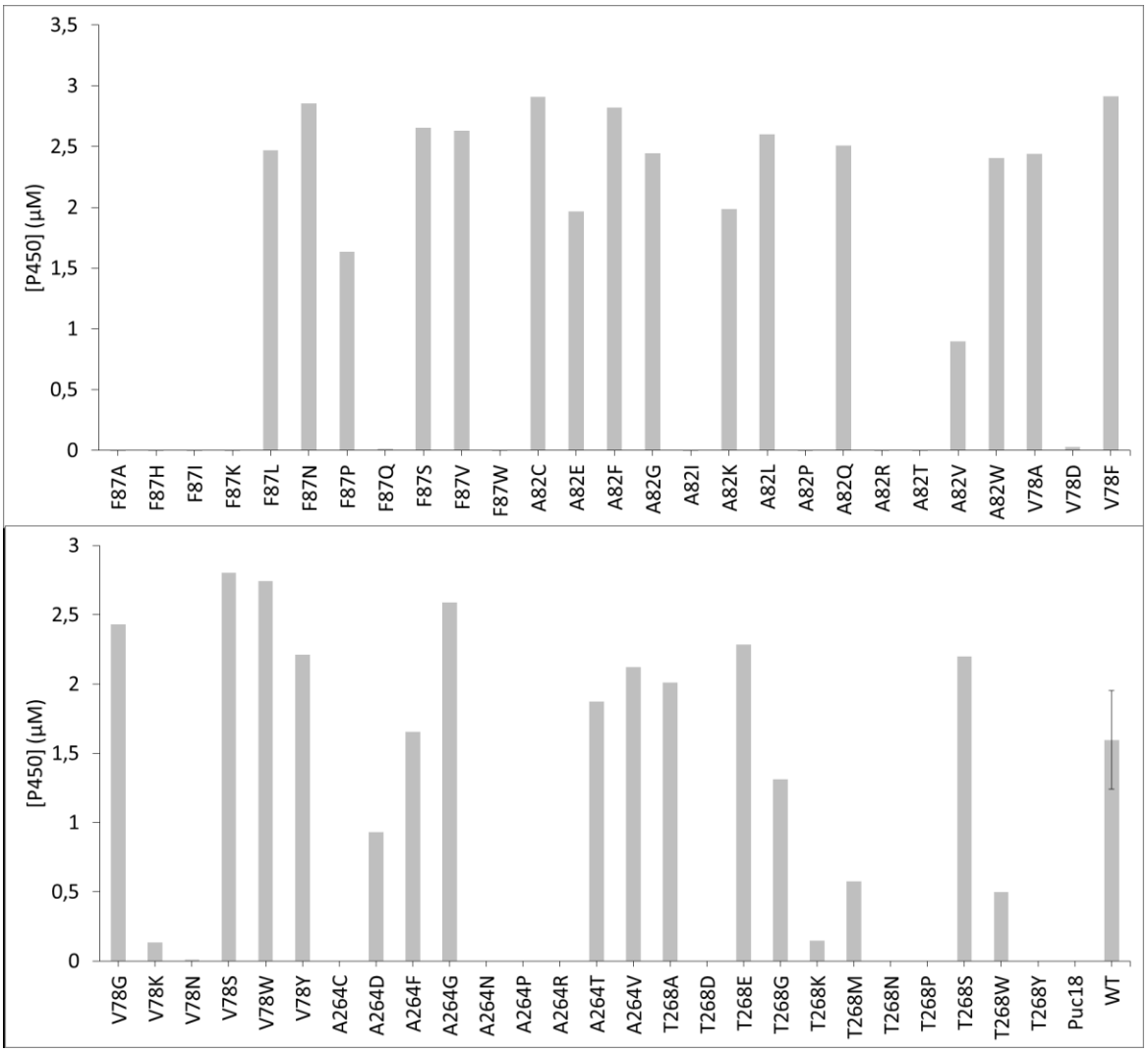


Figure S2-6. P450 BM3 concentration in clarified lysates from 1 mL expression of 53 variants.

The concentration were determined using CO difference spectroscopy. P450 concentration in WT shown with standard deviation, was averaged from eight replicates. Puc18 is the negative control lacking a P450 BM3 gene.

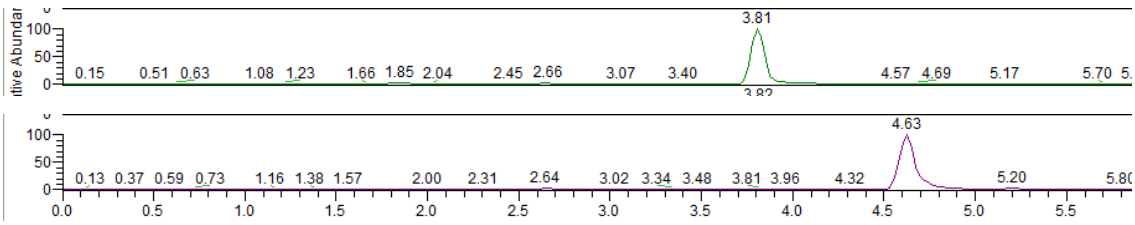


Figure S2-7. LC-MS example of the best variant for the indigo reaction, A82F.

The retention time of indole (117 + 1 H m/z) was at 3.81 min whereas indigo (262 + 1 H m/z) was at 4.63 min in a 10 min method. Masses were detected in positive ionization mode.

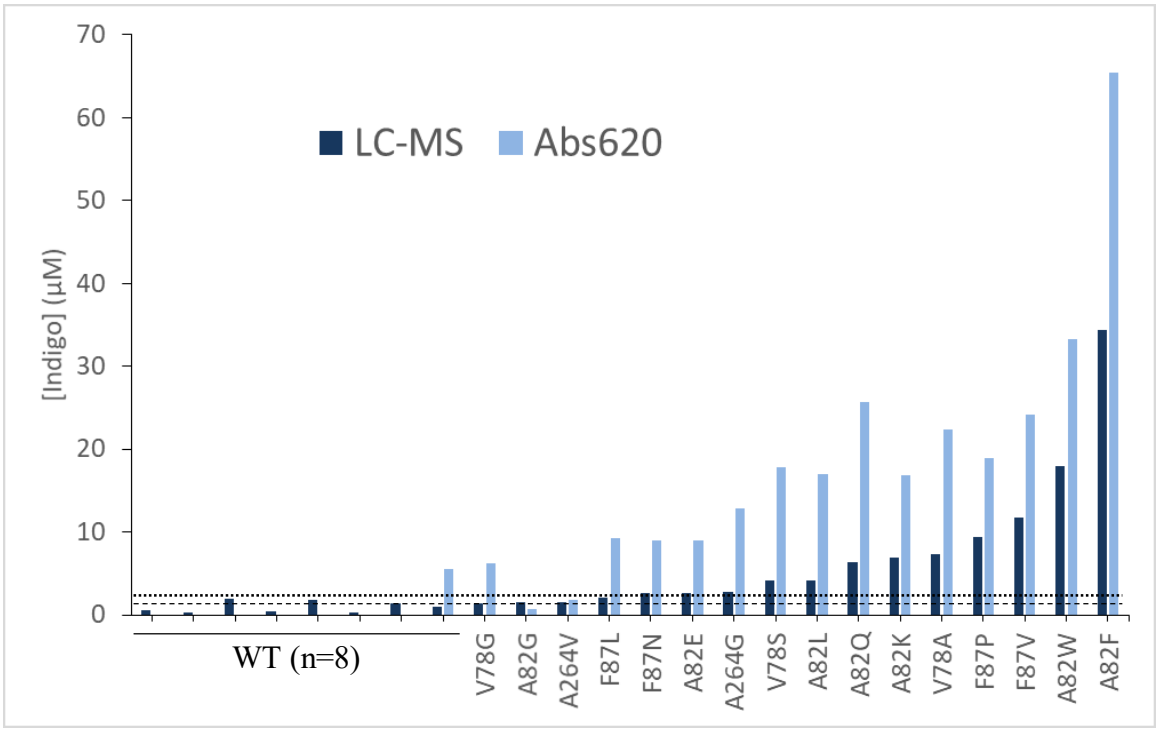


Figure S2-8. Indigo concentration produced by the 16 best variants and the WT determined by LC-MS (dark blue) and Abs620 nm (pale blue).

A264V and A82G were considered inactive according to the Abs_{620} assay. The lowest indigo concentration in the linear portion of the calibration curve was 1.25 µM for LC-MS (black dashed line) and 2.34 µM for Abs_{620} (black dotted line), defining the lower limit of detection for each method.

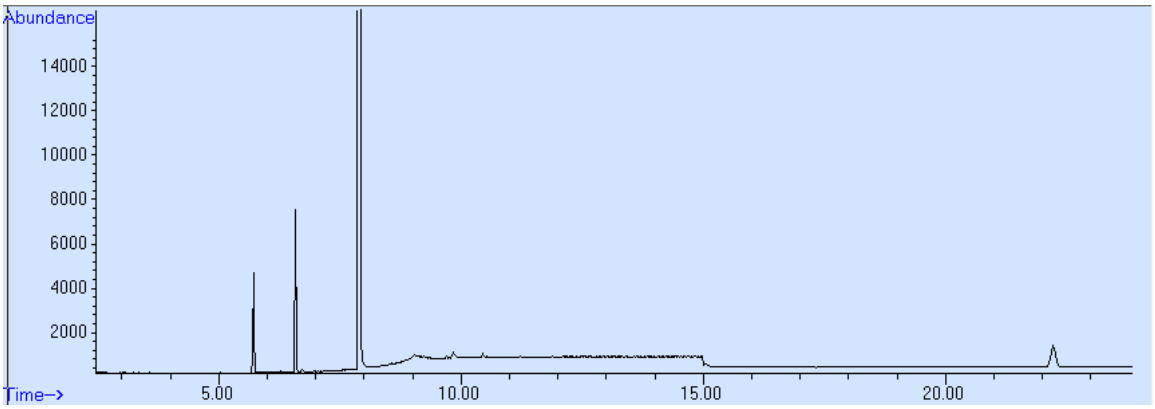


Figure S2-9. GC-MS example of the best variant for the raspberry ketone reaction, A82Q.

The retention time of 4-phenyl-2-butanone (148 m/z) was at 7.9 min whereas the raspberry ketone (fragment at 107 m/z) was at 22 min.

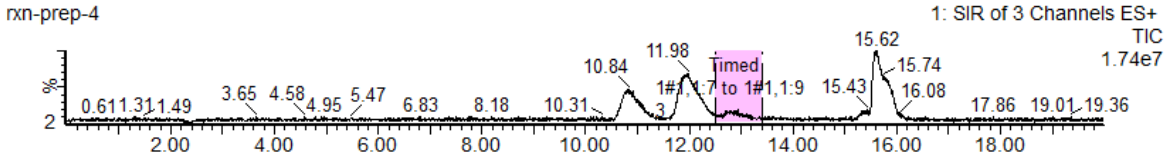


Figure S2-10. Characteristic chromatogram of one of eight injections for preparative LC of the scaled-up reaction of variant A82Q with 4-PB.

Single ion monitoring was conducted at 107.1 m/z (principal fragment of raspberry ketone), 171.2 m/z (4-PB with sodium adduct) and 187.2 m/z (raspberry ketone with sodium adduct). The hydroxylated product having the mass and retention time of the raspberry ketone standard is highlighted in pink.

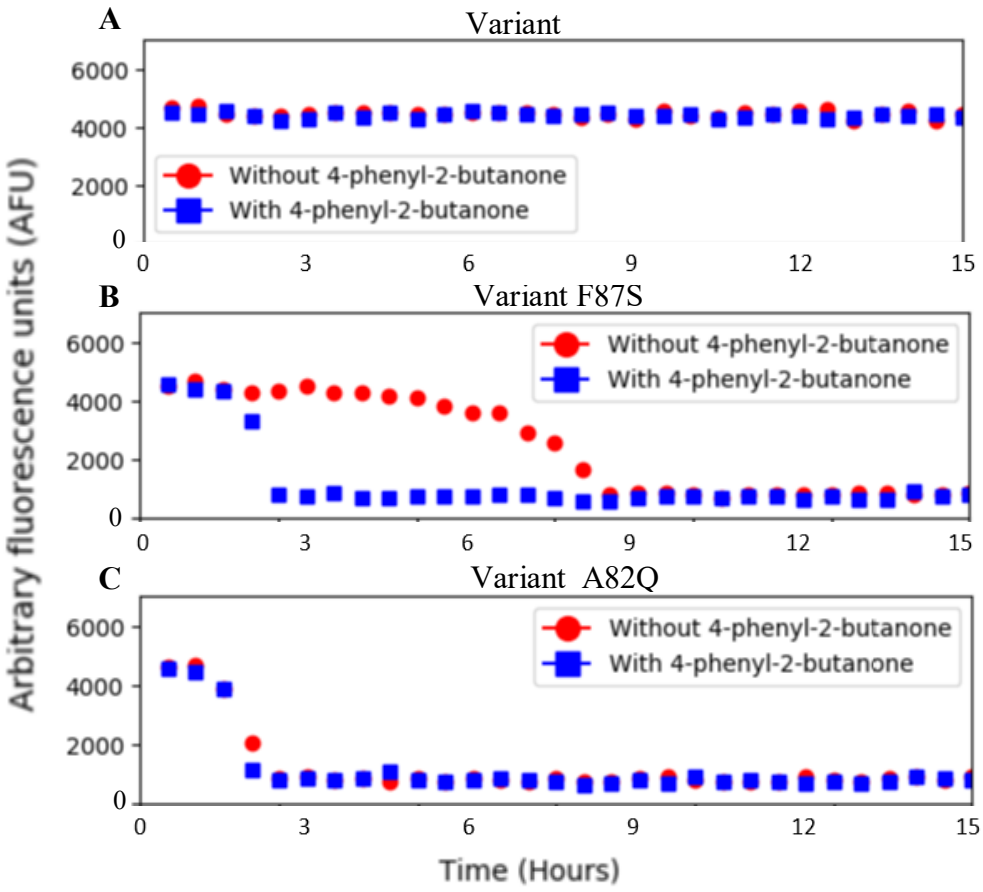


Figure S2-11. Three trends observed for NADPH consumption in the high-throughput fluorescence assay.

A) No NADPH consumption observed either with or without substrate (variant A264C); any slow consumption is masked by saturation of the fluorescence signal. B) Positive hit (variant F87S) where presence of substrate accelerates NADPH consumption. C) Both reactions (with or without substrate) consume NADPH (variant A82Q). The apparent lag in consumption (B, C) is due to the initial NADPH concentration (1.5 mM) saturating the fluorescence signal.

Chapitre 3: Next-generation sequencing applied to biocatalysis: Massive characterization of the cytochrome P450 BM3 for indigo formation

Olivier Rousseau[1,3,4], Musa Ozboyaci[2], Maximilian C.C.J.C. Ebert[2,3,4], Daniela Quaglia[3,4], Sebastian Pechmann [*,2], Joelle N. Pelletier[*,1,2,3,4]

[1] Department of Chemistry and [2] Department of Biochemistry, Université de Montréal, 2900 Boulevard Édouard-Montpetit, Montréal, Québec, H3T 1J4, Canada

[3] PROTEO, the Québec Network for Protein Function, Engineering and Applications, Québec, G1V 0A6, Canada

[4] CGCC, the Center of Green Chemistry and Catalysis, Montréal, Québec, H3A 0B8, Canada

Corresponding author:

Joelle N. Pelletier

Joelle.pelletier@umontreal.ca

Keywords: Biocatalysis, Enzyme engineering, Indigo, Molecular dynamics, Next-generation sequencing, tICA

3.1 Préface

Ce chapitre présente l'article « *Next-generation sequencing applied to biocatalysis : Massive characterization of the cytochrome P450 BM3 for indigo formation* » qui est prêt être soumis. Il traite de l'incorporation du séquençage nouvelle génération dans le processus de l'ingénierie du cytochrome P450 BM3. Il démontre les avantages d'utiliser une telle méthode pour accélérer la découverte de variants actifs de P450 BM3 en reliant la structure, le dynamisme et l'activité à grande échelle.

En tant que premier auteur, j'ai accompli toutes les expériences en laboratoire. J'ai créé les variants de P450 BM3 par la méthode de mutagénèse décrite au chapitre 2. J'ai optimisé l'essai de criblage sur milieu de croissance solide afin de cribler visuellement les librairies de variants de P450 BM3. J'ai fait la préparation des échantillons pour le séquençage de nouvelle génération et écrit un script pour en faire l'analyse. Prof. Sebastian Pechmann a développé un script pour faire l'extraction des mutations des résultats de séquençage et m'a guidé dans l'écriture du script d'analyse. Il a dirigé son stagiaire post-doctorant, Musa Ozboyaci, le deuxième auteur, pour faire les analyses bio-informatiques concernant les modèles par homologie, la dynamique moléculaire et l'analyse tICA. En tant que troisième auteur, Maximilian Ebert a élaboré l'idée de cet article et m'a guidé dans les expériences que j'ai effectuées. En tant que quatrième auteure, Daniela Quaglia m'a guidé dans la planification de la rédaction et la correction de l'article. Finalement, ma directrice de recherche, Prof. Joelle N. Pelletier, ma guidé tout au long de ce projet et a dirigé la rédaction de l'article.

3.2 Abstract

Despite key advances in enzyme engineering, our capacity to predict the effects of mutations on function remains nebulous. It has been determined that using indigo production as a primary screen is a robust predictor for promiscuous activity in variants of cytochrome P450 BM3 (P450 BM3). Expanding the dataset of indigo-producing variants will greatly accelerate the engineering of P450 BM3 for transformation of novel substrates. To this end, we apply bioinformatics tools to interpret next-generation DNA sequencing results of 46 libraries of P450 BM3 variants, examining the production of indigo. We identified 152 point substitutions, 125 double substitutions and 6 triple substitutions that produce indigo, distributed among 29 previously unreported positions. Our results immediately demonstrate the breadth of the fitness landscape that provides solutions for this synthesis. MD simulations revealed that, despite indigo-producing substitutions being in different regions, all gave rise to similar structural and dynamic perturbations. Macrostates of active variants were identified by tICA and were clustered to reveal that the main conformation associated with indigo production was a kinked I-helix, characteristic of the substrate-free state of P450 BM3. We propose that an altered equilibrium between the substrate-free and substrate-bound conformations plays an essential role in the numerous solutions we have identified to indigo synthesis. We look ahead to the potential for large experimental datasets to train smarter design algorithms for enzyme engineering.

3.3 Introduction

The *third wave of biocatalysis* has greatly accelerated enzyme engineering by introducing directed evolution and bioinformatics tools.[1] A plethora of methodologies have been developed to increase the throughput in the creation, screening and identification of libraries while reducing inactive variants, errors and biases.[2] Nonetheless, the choice of which amino acids to mutate to improve selected features of a target enzyme is still not trivial.[3] Mutagenesis is generally carried out either with a focused approach based on in-depth structural knowledge, or by application of random mutagenesis. If randomization is selected to generate diversity, the resulting library size is often a hurdle, even when a robust high-throughput screening is available. Indeed, the explorable fitness landscape of a protein is so vast that exhaustive experimental coverage is not feasible.

Progress in the *fourth wave of biocatalysis* [4] will depend on our capacity to establish strong links between massive datasets for phenotypic and genotypic characterization of very large libraries of variants. Phenotypic characterization of enzyme libraries on the order of 10^4 - 10^6 variants requires rapid and accurate high-throughput assays, potentially relying on microfluidics [5] or phage/yeast display combined with FACS.[6-9] Achieving such a throughput is challenging because colorigenic or fluorogenic read-outs are typically required, but the majority of relevant biocatalytic reactions do not natively produce such signals; indirect assays may be developed. Genotype determination for large libraries increasingly calls upon next-generation sequencing (NGS) because it allows simultaneous analysis of millions of DNA fragments.

To this day, NGS has been primarily used for public health applications. In particular, whole genome sequencing (WGS) is applied to map whole genomes according to the consensus obtained from multiple sequence alignments of sequenced genome fragments.[10-13] The identification of mutations within a library of a single gene, as commonly practised in enzyme engineering, is not straightforward with NGS because that method relies on determining the consensus sequence among a population of DNA fragments; in effect, standard software for NGS analysis such as Galaxy [14] or Breseq [15] is designed to interpret mutations as sequencing errors and discount them.

Nonetheless, the affordability and dramatic increase in sequencing throughput offered by NGS are attractive. In the context of enzyme engineering, the DNA encoding variants having a desired function determined by screening can be pooled and DNA-barcoded. Barcoded pools representing several phenotypes can be simultaneously processed by NGS. The resulting millions of short DNA sequences (reads) are tracked by the barcodes to trace their pool and thus link phenotype to genotype. This process requires rigorous statistical analysis because true mutations must be distinguished from the background of sequencing errors and biases. For this reason, NGS has not been extensively applied to biocatalysis. Here we develop computational approaches to make NGS readily available for application to enzyme engineering.[3]

High industrial interest exists for the superfamily of cytochrome P450s. These NADPH-dependent mono-oxygenases catalyze challenging regio-, chemo-, stereo-specific oxidation of non-activated carbon atoms in a single step.[16] Cytochrome P450 BM3 (P450 BM3), from *Bacillus megaterium*, presents particularly interesting features with respect to biocatalytic applications. P450 BM3 is a self-sufficient cytochrome P450: its reductase domain is fused to its heme domain, making it among the most efficient P450s.[17] Its natural substrate are fatty acids yet it has been reported to oxidize a wide range of substrates including alkanes, drugs and many more, following mutational modification.[18-22]

Indigo, one of the oldest known dyes, continues to draw interest with multiple applications in the textile, food and pharmaceutical industries.[23-24] P450 BM3 has previously been engineered to improve the oxidation of indole to indigo, on a limited scale (Figure 3-1).[25-28] In addition to direct interest in indigo itself, we and others have demonstrated that indole-reactive, indigo-producing variants of P450 BM3 have a high likelihood of displaying promiscuity towards a second unrelated substrate.[29-32] In that context, we recently demonstrated that P450 BM3 can be readily evolved to oxidize indole to indoxyl; specifically, screening only 53 point-substituents of five active-site positions yielded 16 indigo-producing variants, 13 of which had not been previously reported.[32] Eight among those variants, or 50%, catalyzed oxidation of the promiscuous substrate, 4-phenyl-2-butanone, to produce the industrially relevant raspberry ketone.[32] Based on the high predictive capacity of indigo production in the identification of new oxidation reactions, we set out to expand the database of indigo-producing variants of P450 BM3.

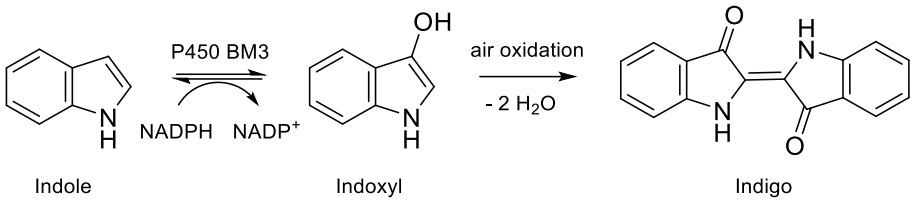


Figure 3-1. Biotransformation of indole to indigo.

The oxidation of indole to indoxyl by cytochrome P450 BM3 is followed by the spontaneous dimerization of indoxyl to indigo.

Herein, we present a combination of multiplexing, automation and bioinformatics to greatly accelerate mapping the fitness landscape of P450 BM3 for indigo formation (Figure 3-2). First, using site-directed saturation mutagenesis with a mix of degenerate primers, we obtained a diversified library of variants covering 46 positions.[33] The libraries were screened with a colony picker according to formation of blue colonies of *Escherichia coli* on solid growth media, allowing simultaneous processing of all variants. Colonies were pooled and barcoded according to their phenotype: high, low or no visible production of indigo. The pools were identified by NGS using a newly-developed script to extract and parse the mutations according to their frequency of occurrence and quality. A subset of 84 point-substituents thus identified were characterized according to models and short molecular dynamic simulations. Analysis with time-structure Independent Principal Component (tICA) allowed the identification of structural correlations explaining the propensity of the variants of P450 BM3 to convert indole to indigo.

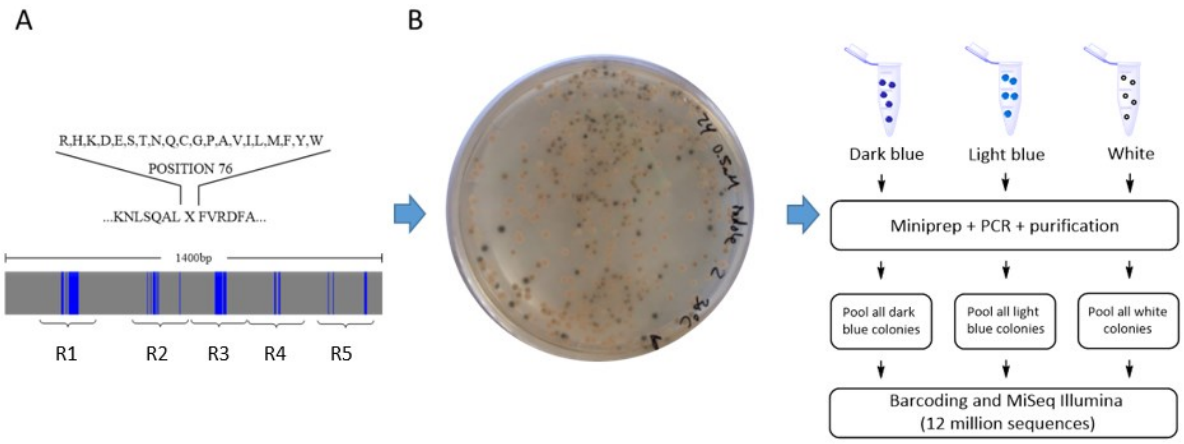


Figure 3-2. Workflow for creation and identification of sequence diversity.

A) Site-directed saturation mutagenesis was performed at each of 46 selected positions of P450 BM3. These positions are loosely clustered into five regions (R1 to R5), where each region is distinct in tertiary structure and role in the oxidation mechanism. B) Screening for indigo synthesis on solid media containing the substrate, indole. Colonies expressing an indigo-producing variant turn blue. C) Sample preparation for NGS.

3.4 Results

Creation of sequence diversity. Indigo production by variants of P450 BM3 is a promising predictor of novel substrate discovery. To date, few positions of P450 BM3 have been explored with respect to substitutions that give rise to oxidation of indole, limiting the potential to apply indigo production in discovery of new reactions.

In our previous report, we modified five positions of P450 BM3: V78, A82, F87, A264 and T268, and identified 16 variants that convert indole to indigo.[32] In this work, site-directed saturation mutagenesis was performed at 46 further positions that were selected for having at least one atom within a radius of 20 Å of the heme-iron, near or above the plane of the heme (Figure 3-3A and Table S3-1). In addition, these positions were selected for belonging to structural elements in one of five regions known to play different roles in the oxidation mechanism. Region 1 (R1) encompasses the B' helix, known to play a role in substrate recognition.[34] Several mutations in this helix have been reported to improve indigo synthesis.[19, 26, 28, 32] Region 2 (R2) encompasses the F/G-helices and is known to undergo

a large conformational change upon substrate binding, affecting the substrate access channel.[34-36] Region 3 (R3) includes the I-helix. It is the longest helix of P450 BM3 and it crosses the active-site cavity immediately above the heme. It contains residues that are critical to the mechanism and to the conformational changes observed in the F/G-helices.[34-36] Finally, regions 4 and 5 (R4 and R5) include loops in the vicinity of the heme that have the potential to modulate indigo synthesis.

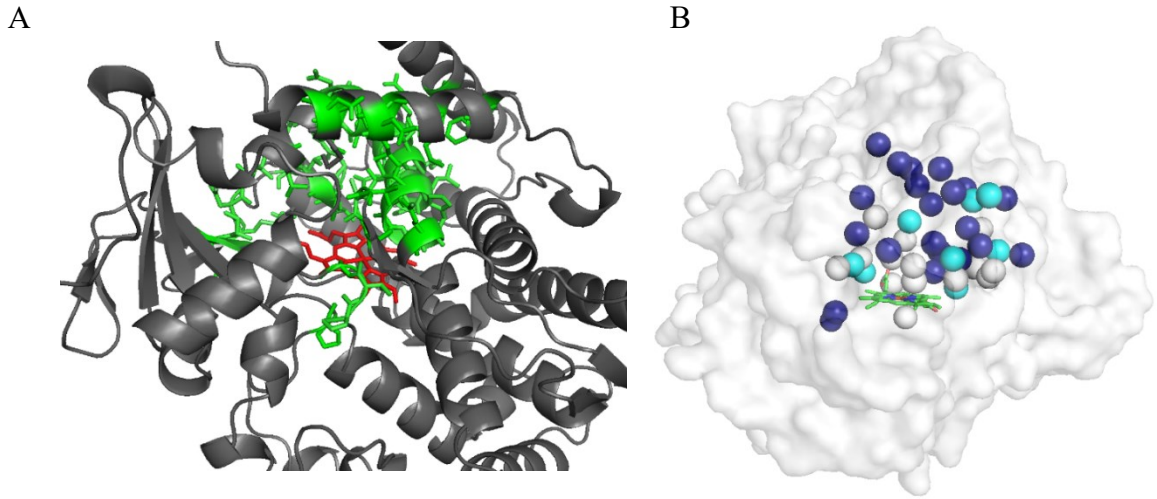


Figure 3-3. Modified active-site positions in P450 BM3 (PDB code 1BU7).
A) Positions selected for mutagenesis (green sticks) are generally above the heme plane (red sticks). B) Overview of results of the indigo screening assay. The surface of P450 BM3 is shown in white with heme in CPK sticks. Balls identify the phenotype associated with substitution at each selected position: dark blue, at least one substitution displayed high activity; cyan, at least one substitution displayed weak activity; white, no active substitution was identified.

We used a defined mix of primers optimized to eliminate codon redundancy and stop codons, theoretically encoding a uniform distribution of 20 amino acids at each position.[33] In total, 49 libraries were created: 42 point-mutated libraries, 4 double-mutant libraries and 3 triple-mutant libraries for a total of 46 mutated positions (some positions being mutated singly and in combination). The double and triple mutations were at consecutive positions, due to constraints imposed by NGS (details are provided under Methods).

Screening for production of indigo. Each of the 49 libraries was screened on agar plates containing indole (Figure 3-2B). Upon oxidation of indole to indoxyl and formation of indigo, colonies expressing active variants turn blue. Visual inspection immediately revealed the

potential of each position (or group of positions, in the case of double- and triple-mutant libraries) to give rise to indigo-producing variants. In addition, a range of color intensity was observed. This allowed for categorization of colonies as white, pale blue or dark blue, corresponding to inactive, weakly active and strongly active variants, respectively. Thirty of the 49 libraries screened held pale blue colonies and 15 among them also held dark blue colonies (Figure 3-3B and Table S3-1). Only one of the 30 positions active for indigo production was previously reported, such that we report 29 new positions where at least one substitution yields an indigo-producing variant of P450 BM3.

Colonies were picked to identify the substitutions allowing indigo production by NGS. A total of 10648 white, 664 pale blue and 123 dark blue colonies were picked (Table S3-2). The Illumina MiSeq NGS method allows sequencing of 250 bp-long DNA segments. The residues that were mutated in each of the five regions (R1 to R5) all cluster within 250 bp; we thus pooled colonies from each region. In addition, three distinct pools were created for each region, according to the phenotype (dark blue colonies, pale blue colonies and white colonies). This resulted in a maximum of 15 pools; two contained no colonies, yielding 13 pools (Figure 3-2C).

Statistical analysis of NGS results. Statistical analysis of the Illumina MiSeq NGS results was performed according to an in-house script that accounted for the inherent rate of sequencing error and experimental biases, as detailed under Methods. We identified 70 variants from the dark blue pools, 213 from the pale blue pools and 2006 from the white pools (Figure S3-7). We note that the white colonies can encode well-folded and functional indigo-negative variants as well as poorly folded and/or non-functional variants.[32] The amino acid coverage of the point-substituted libraries varied between 20% and 90% (Figure 3-4; Figure S3-8). A coverage of 3 to 45% was obtained for double and triple-position variants; lower coverage was expected because the number of variants in each of those libraries is greater (400 per doubly-substituted and 8000 for triply substituted libraries). We happened to identify ten point-substituted variants in the double/triple mutant libraries. For this reason, positions 78, 82 and 87 are included in Figure 3-4 although they had not been individually mutated.

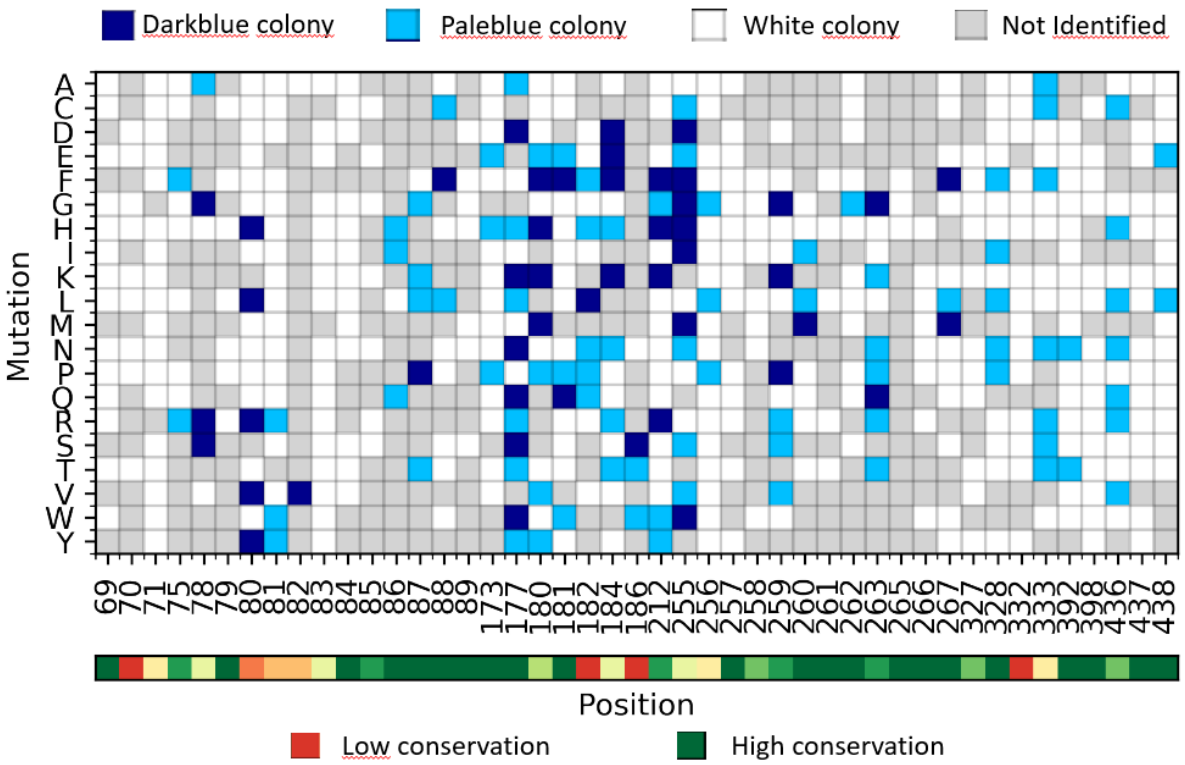


Figure 3-4. Map of the phenotype of all point-substituted variants identified by NGS.

Each square represents a single mutation (row) for a position (column). White, pale blue and dark blue identify inactive, weakly active and strongly active variants, respectively. Grey boxes indicate the substitutions that were not identified by NGS analysis. Below the map, the degree of protein sequence conservation, per position, is indicated.

We mapped the genotype/phenotype relation for all point-substituents that were identified to analyze the effect of individual amino acid substitutions (Figure 3-4). The positions that were selected for mutations in region 2 (residues 173 to 212) and region 3 (residues 213 to 267) are more prone to catalyze the conversion of indole to indigo. In particular, positions 177, 180, 184 and 212 of region 2 each held seven to twelve variants that produced indigo; together, they account for a high proportion of the highly active variants we identified. Positions 255, 259 and 263 in region 3 each held six to twelve indigo-active variants, with position 255 displaying the greatest number of highly active variants (seven) of all positions investigated. Several positions in regions 3 showed below-average mutational coverage, which may have led to an underestimation of that region's indigo-producing potential. Alternatively, the lower coverage could result from mutations in region 3 tending to be deleterious to bacterial propagation and

therefore not being identified by NGS. In contrast, regions 1 (residues 69 to 89), 4 (residues 327 to 333) and 5 (residues 392 to 438) included few highly active variants despite the numerous positions investigated in region 1 and the excellent sequence coverage obtained in regions 4 and 5; positions 78 and 80 are the exceptions, each displaying several highly active variants.

Comparison of the propensity for each position to yield indigo-producing variants with the sequence conservation rate at each position yielded no clear correlation. Among the positions that yielded many indigo-producing variants, positions 177, 212, 259 and 263 are highly conserved while positions 80, 180, 184, 255 and 436 are evolutionarily more tolerant to substitution. Positions 69, 71, 83, 84, 257, 327, 332, and 398 held no indigo-active variants despite each position being well covered. Among them, positions 69, 84, 257, 327 and 398 are highly conserved throughout evolution suggesting their modification affects the function or stability of P450 BM3. In fact, R398 is known to be highly conserved because it stabilizes the heme propionate by hydrogen bonding, influencing the heme orientation.[37-38]

Molecular dynamics analysis.

Our results clearly demonstrate that P450 BM3 offers multiple solutions to indigo synthesis. It is not known whether indole oxidation is activated by one or a few common mechanisms, or if each substitution contributes in a unique manner. To investigate this, we developed an analysis pipeline that does not make any prior assumptions on the modes of perturbation. Instead, it aimed to extract the most pronounced characteristic information from each variant and subsequently search for commonalities across the spectrum of active variants (Figure 3-5A).

Structural models were generated for 70 indigo-producing variants from our dataset and subjected to 25 ns of MD simulations. They were benchmarked against previously characterized variants A82W, A82F, F87V and F87P as indigo-positive controls and 10 inactive, but functional variants as negative controls (Table S3-4).[32] A feature-selection step was followed by the identification of characteristic representative structural conformations for individual variants and clustering across the full set of variants.

To maximize signal for downstream analysis, only positions whose average pairwise distances to neighbouring residues across the MD trajectory differed between mutant and wild-type simulations were selected. tICA analysis were applied to reduce dimensionality. This

allowed for identification of characteristic *microstates* from the MD trajectories, which were subsequently clustered into representative *macrostates* (Figure 3-5B).

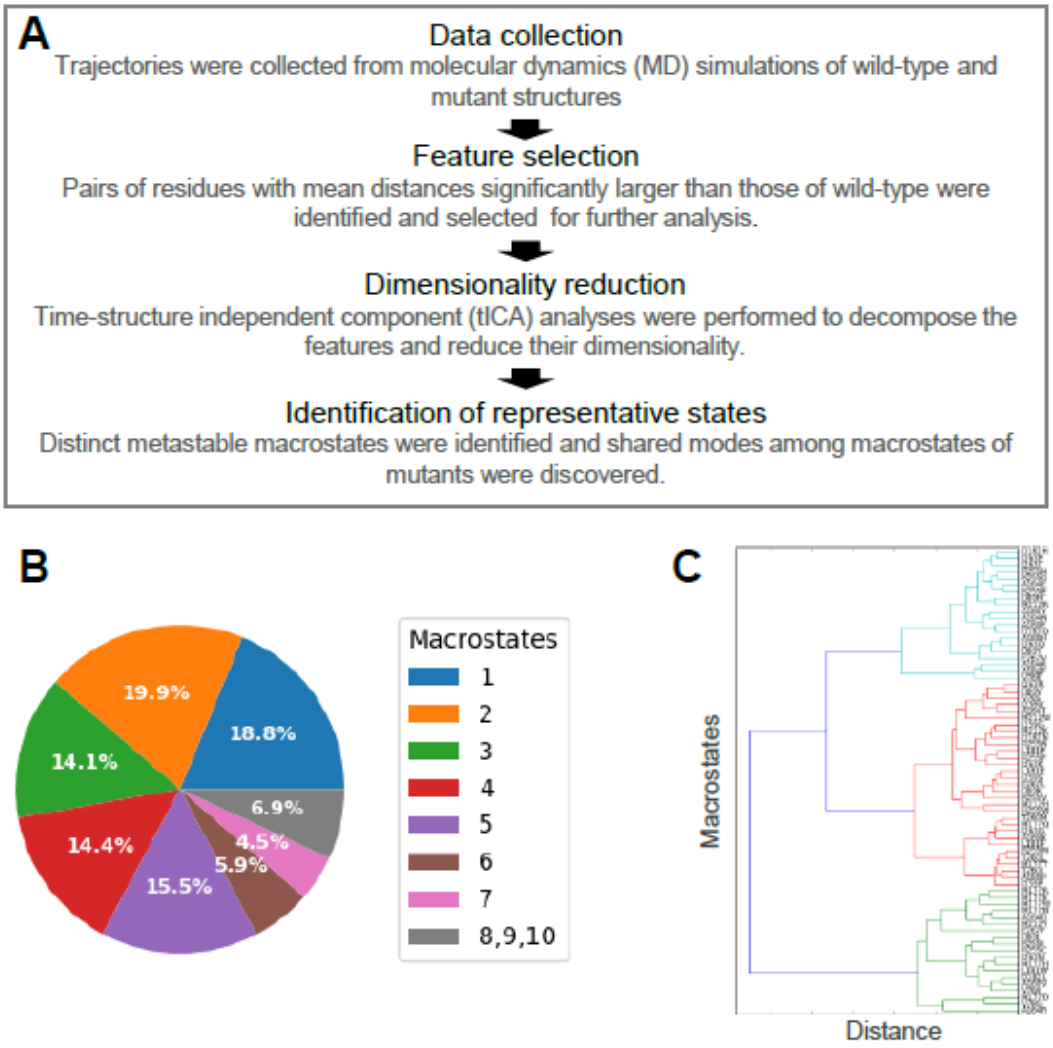


Figure 3-5. Computational analysis of the common perturbations caused by different point mutations.

(A) Workflow that describes the methodology used for this analysis. (B) Macrostate populations from the molecular dynamics simulations of the mutant structure M177K. Macrostates of different mutants have different population profiles for their distinct states. (C) Dendrogram that illustrates the arrangement of the clusters obtained from the clustering of macrostates identified from simulations of different mutations. Clustering helps group similar structural changes together and therefore reveals structurally distinct shared modes systematically.

We next quantified how often the representative macrostates were visited during a single trajectory. This was done by clustering the frames of the MD trajectory into the clusters of the

previously obtained macrostates for that trajectory. We clustered the macrostates of each variant across the full spectrum of variants to identify systematic structural perturbations that were distinct from any identified in the WT and negative controls (Figure 3-5C).

Strikingly, we observed a pronounced similarity between well-populated macrostates of different variants. The main cluster of macrostates from all indigo-positive variants defines the appearance of a kink in the I-helix. This is illustrated by one of the major conformational changes observed in indigo-positive variant M177K compared to the WT and to indigo-negative variant R255F (Figure 3-6). This kink pushes the F/G-helices away from the heme plane, opening the substrate channels. This makes sense as it is used in the gating mechanism to exclude water from the active site when the substrate binds the heme (This phenomenon is discussed below and is highlighted by a red circle in the figure 3-6). These observations are consistent with known P450 BM3 substrate-free (SF) and substrate-bound (SB) conformations.[36, 39] These observations and additional significant systematic conformational perturbations that are linked to enzyme activity will be further explored in the future.

3.5 Discussion

We have uncovered a detailed structural perturbation by means of a high-throughput enzyme evolution experiment and large-scale MD simulations. The structural perturbation we have observed was previously determined by means of labor-intensive, high-resolution structural biology methods. Our results strongly underline the premise, and promise, of this work: high-throughput methods have been applied to gain, for the first time, a more systematic and rational understanding of the sequence-structure-function landscape of a cytochrome P450 enzyme.

Substitutions that are active towards a given reaction cannot be systematically predicted by rational design. For example, multiple mutations in P450 BM3 at position A184 were found to catalyze epoxidation reactions, but the modulation of enantioselectivity was hard to justify.[40] Many of the indigo-positive variant identified here are outside of the first shell of the active site, rendering any predictions even more precarious. We demonstrate that a multiplexing approach greatly accelerates mapping the fitness landscape of P450 BM3 and the discovery of new active variants for indigo transformation.

Active substitutions could give rise to indigo production via numerous mechanisms. The fact that substitutions at several positions produce indigo suggests that, in addition to substrate binding, distinct factors inducing indole oxidation are in cause. P450 BM3 has been observed in two major conformations: substrate-free (SF) and substrate-bound (SB).[36, 39] Upon substrate binding, a shift between the SF and SB states displaces an important water molecule, thereby converting the six-coordinate heme (low-spin) to a five-coordinate heme (high-spin).[39, 41] This disrupts a hydrogen bonding network between the heme, water molecules and catalytic residues that include A264 and T268, [35] allowing oxygen binding, and permitting oxidation to proceed.

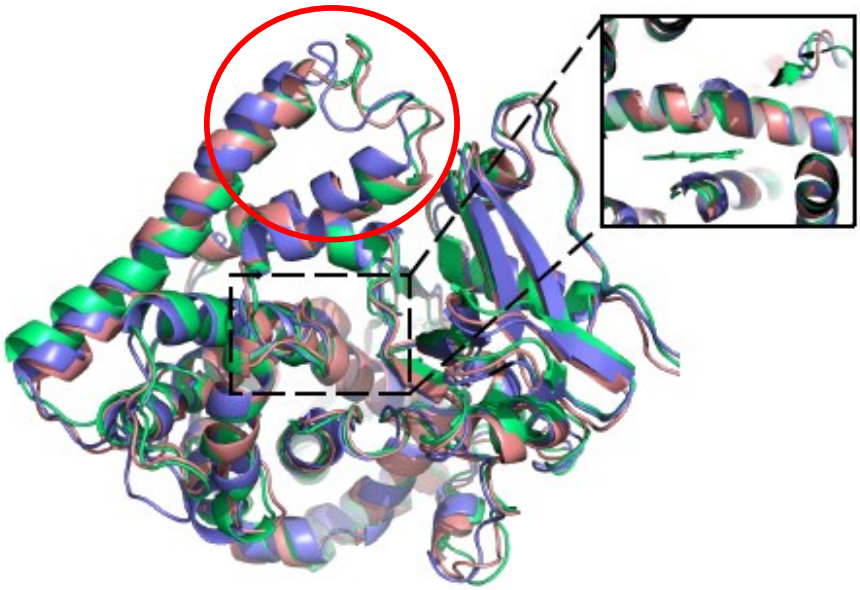


Figure 3-6. Models of the most populated macrostate representatives.

Models of WT P450 BM3, the indigo-positive variant M177K and the functional but indigo-negative variant R255F shown respectively in green, purple and pink. The M177K substitution produced in a kink involving residues 263-266 (right inset). The kink is in the active site, and it is shared by other indigo-positive structures mutated at different positions (green branches in dendrogram, Figure 5C). Macrostates of WT P450 BM3 and the exemplary indigo-negative variant R255F, on the other hand, do not depict any perturbation near that location.

The shift from the SF to the SB state induces a major conformational change in P450 BM3.[34, 36, 42] The I-helix, that contains a 13° kink in the SF state, becomes more linear (5°) in the SB state.[41] This kink is conserved in the SF states of a majority of cytochrome

P450s.[43] The SF-state kink pushes the F/G helices and the loop that links them, defined as the "lid domain", modulating the opening of the substrate access channel. Bulk solvent fills the active site in the SF state, but is excluded in the SB state.[41] This mechanism acts as a gating system to prevent unproductive use of electrons.[44-45] As a result, substitutions that alter the equilibrium between the SF and SB conformations have the potential to modulate indole uptake and oxidation.

Based on crystal structures, the SB I-helix conformation is observed in absence of substrate for several variants of P450 BM3.[22, 26, 39] Closing of the substrate access channel has been also observed in CYP1A2 for a distal mutation (F186L) affecting the D/E/F helices.[46] It has been further argued that the shift can occur independently of substrate binding, since mutations at position 264 adopt a SB conformation in absence of substrate by replacing water as the sixth ligand.[47-50] Thus, substrate binding is not the only cause for the SF to SB conformational shift. Here, MD simulations model the SF state as being the most highly represented for the active variants, as opposed to the WT and the inactive variants where the SB state predominates. This suggests that variants can bypass the gating system and shift the equilibrium between representative states independently of the presence of substrate.

In conclusion, we have identified nearly 300 new indigo-producing variants of P450 BM3. They involve 29 among the 46 positions (> 60%) we mutated in the core of the enzyme, none of which had been previously associated with indigo production. The high success rate suggests that more solutions to indigo-production exist: our imperfect coverage of the investigated positions and the possibility that distant sites also affect indole oxidation provide opportunities for further discovery. This massive influx of new information was made possible by the application of NGS and scripting to identify the mutated sites, which would otherwise be interpreted as sequencing errors by standard NGS analysis algorithms. High-throughput MD simulations revealed commonalities in structural perturbations, including a marked adoption of the kinked I-helix in indigo-positive variants only. Characteristic of the SF state of cytochrome P450s, its conformation and that of the SB state have been well documented; identification of altered equilibrium between these states is coherent with new substrate promiscuity and suggests that the simulations are robust. The identification of active mutations can help in a subsequent round of evolution by guiding enzyme engineering in the selection of positions to mutate and combination of active mutations to do. As indigo production has been demonstrated to be a

strong predictor of substrate promiscuity in P450 BM3 [29-32], we speculate that adoption of the SF state may be conducive to transformation of further promiscuous substrates.

3.6 Materials and Methods

Materials and instruments. Mutagenic primers were purchased from Sigma-Aldrich. The Phusion Hot Start II polymerase, purification kit, DNA miniprep kit, DNA gel extraction kit, and electrocompetent *E. coli* DH5alpha were purchased from New England Biolabs. FastDigest restriction enzymes BamHI, EcoRI, FastAP and DpnI were purchased from Thermo Fisher Scientific. Ligations were performed with the TAKARA ligase from ClonTech. The pcWORI P450 BM3 WT vector was kindly provided by the laboratory of Prof. Frances Arnold (Caltech, USA) and used as a cloning vector. Primers were designed using SnapGene software and purchased from Sigma. Indole was bought from Alfa Aesar. PCR reactions were performed with a T100 thermal cycler from Bio-Rad. Colonies were picked with a Pick-in Master PM-2s from Microtec co., LTD.

Library generation. Fourty-six active-site residues having at least one atom within 20 Å above the heme-iron were selected for mutagenesis. They were grouped into 42 single-mutant, four double-mutant and three triple-mutant libraries. Site-directed saturation mutagenesis was performed using a mix of four optimized primers for each mutated codon, where each primer encodes either NDT, VMA, ATG or TGG at the variable position.[33] Individual mutagenesis reactions were carried out for each library, using the megaprimer method of mutagenesis.[51] PCR intermediates were purified before proceeding with further steps. Final PCR products were digested using BamHI and EcoRI according to the manufacturer's instructions. DpnI was used to eliminate parental DNA. The gel-purified inserts were ligated into the BamHI, EcoRI and Fast-AP treated pcWORI P450 BM3 WT vector. Electrocompetent *E. coli* DH5alpha were transformed with each ligation mix and plated on LB agar with ampicillin (100 μg/mL). Following colony growth, each position-specific plate was washed with 2 x 750 μL of LB with 100 μg/mL ampicillin. The resuspended cells were incubated for 1h at 37°C with agitation at 230 rpm and glycerol stocks were conserved at -80°C.

Overnight cultures of each library were propagated in LB with 100 μg/mL ampicillin and plated (100 μL of a 10000 × dilution) on ZYP5052 agar medium containing 0.5 mM indole.

Following incubation at 37°C for 8h and at 30°C for at least 40 h, Petri dishes were stored at 4°C for at least 24 h to enhance the intensity of indigo.

Colonies were picked and pooled according to their color (dark blue, pale blue, white) and the region of the mutated residue (in nucleotides: R1= 79-385, R2 = 444-720, R3 = 667-944, R4 = 837-1108, R5 = 1120-1390). Regions size were approximately 250 bp in length due to restrictions inherent to the Illumina MiSeq NGS protocol. Blue colonies were hand-picked (787 colonies) and white colonies were picked using the colony picker (10648 colonies). Overnight cultures in LB with 100 mg/mL ampicillin at 37°C were prepared for each pool and miniprepped. PCR reactions were performed for each 250 bp region using specific primers containing CS1/CS2 sequences (Table S3-3). PCR products were pooled and submitted for barcoding and NGS using Illumina MiSeq 250bp technologies at the Génome Québec Innovation Center of McGill University.

NGS analysis. True mutations needed to be distinguished from sequencing errors and experimental biases. Illumina MiSeq sequencing delivered more than 12 million sequences, requiring an in-house script to extract the mutations from each read according to read quality and frequency of occurrence of that mutation relative to the background noise. Background was determined according to the frequency of incorrect bases at all positions that were not selected for mutation. We defined read frequency thresholds based on that information. We observed that the background noise was different for each pool (dark blue, pale blue, white) and region (R1-R5), illustrating the extent of experimental bias (Figure S3-9). We defined thresholds for each pool, for each number of mutation inserted (single, double, triple) and for each PCR region (R1-R5) with a tolerance of 0.5% of false positives. Mutations observed at a higher number of reads than their specific threshold were considered valid in their respective pool.

Some variants were identified in more than one pool; for example, they could be identified in both the pale blue and dark blue pools. We performed a Fisher exact test to determine the pool in which each mutant was most highly represented. For each mutation, a p-value expressed the probability of finding the mutation by chance in that pool by comparing its frequency in the different pools. We filtered the mutations with p-values lower than 0.01 to retain only mutations giving rise to valid comparisons. Then we applied the odds ratio to assign each variant to a pool, thus defining its most likely phenotype.

P450 BM3 MD simulations. The P450 BM3 crystal structure 1BU7 was the starting point for producing models of variants using PyMOL and performing MD simulations using GROMACS. Protonation states of histidines were predicted with PROPKA and pdb2pqr. The protein was embedded in a dodecahedral box using the TIP3P model for explicit water and periodic boundary conditions. The net charge of the protein was neutralized with 100 mM salt.

For energy minimization and MD simulations of each system, the AMBER99SB-ILDN force field was used. The structures were minimized with the steepest descent method for a maximum number of steps of 50000. Equilibration was performed for 1 ns with positional restraints applied on the heavy atoms of the protein and with an NVT ensemble at 300 K. Further 1 ns equilibration simulations were performed with constant NPT ensemble. Finally, for each variant, five simulations of 25 ns were performed at a constant pressure of 1 bar and a temperature of 300 K.

3.7 Acknowledgements

O. R., M. C. J. C. C. E., D. Q. and J. N. P. belong to the Québec Network for Protein Function, Engineering and Applications (PROTEO) and the Center of Green Chemistry and Catalysis (CGCC). We thank Prof. Frances Arnold for providing the P450 BM3 gene; Pierre Lepage at the McGill Génome Québec Innovation Center and Jacynthe L. Toulouse for helpful discussions and suggestions. This work was funded by Natural Sciences and Engineering Council (NSERC) grant 227853 and the Fonds Québecois de la Recherche sur la Nature et les Technologies (FQRNT) grant 181877. O.R. held a PROTEO scholarship and M.C.J.C.C.E. was the recipient of NSERC, PROTEO and UdeM scholarships.

3.8 Conflicts of interest

The authors declare no conflict of interest.

3.9 References

[1] U. T. Bornscheuer, G. W. Huisman, R. J. Kazlauskas, S. Lutz, J. C. Moore, K. Robins, *Nature* **2012**, *485*, 185.
[2] P. A. Tizei, E. Csibra, L. Torres, V. B. Pinheiro, *Biochem. Soc. Trans.* **2016**, *44*, 1165-1175.
[3] E. E. Wrenbeck, M. S. Faber, T. A. Whitehead, *Curr. Opin. Struct. Biol.* **2017**, *45*, 36-44.
[4] U. T. Bornscheuer, *Philos. Trans. Royal Soc. A* **2018**, *376*.
[5] P. A. Romero, T. M. Tran, A. R. Abate, *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 7159-7164.
[6] J. R. Klesmith, J.-P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, *Proceedings of the National Academy of Sciences* **2017**, *114*, 2265.
[7] T. A. Whitehead, et al., *Nat. Biotechnol.* **2012**, *30*, 543-548.
[8] L. M. Starita, et al., *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, E1263-1272.
[9] M. Nyyssonen, et al., *Front Microbiol* **2013**, *4*, 282.
[10] J. Wu, M. Wu, T. Chen, R. Jiang, *Quant. Biol.* **2016**, *4*, 115-128.
[11] J. Besser, H. A. Carleton, P. Gerner-Smidt, R. L. Lindsey, E. Trees, *Clin. Microbiol. Infect.* **2018**, *24*, 335-341.
[12] H. P. J. Buermans, J. T. den Dunnen, *Biochim. Biophys. Acta, Mol. Basis Dis.* **2014**, *1842*, 1932-1941.
[13] S. Goodwin, J. D. McPherson, W. R. McCombie, *Nat. Rev. Genet.* **2016**, *17*, 333-351.
[14] E. Afgan, et al., *Nucleic Acids Res.* **2018**, *46*, W537-w544.
[15] D. E. Deatherage, J. E. Barrick, *Methods Mol. Biol.* **2014**, *1151*, 165-188.
[16] T. Sakaki, *Biol. Pharm. Bull.* **2012**, *35*, 844-849.
[17] A. W. Munro, et al., *Trends Biochem. Sci.* **2002**, *27*, 250-257.
[18] R. T. Ruettinger, L. P. Wen, A. J. Fulco, *J. Biol. Chem.* **1989**, *264*, 10987-10995.
[19] C. J. C. Whitehouse, S. G. Bell, L. L. Wong, *Chem. Soc. Rev.* **2012**, *41*, 1218-1260.
[20] V. B. Urlacher, M. Girhard, *Trends Biotechnol.* **2012**, *30*, 26-36.
[21] C. J. C. Whitehouse, S. G. Bell, H. G. Tufton, R. J. P. Kenny, L. C. I. Ogilvie, L.-L. Wong, *Chem. Commun.* **2008**, 966-968.
[22] C. F. Butler, C. Peet, A. E. Mason, M. W. Voice, D. Leys, A. W. Munro, *J. Biol. Chem.* **2013**, *288*, 25387-25399.
[23] X. Han, W. Wang, X. Xiao, *Chinese Journal of Biotechnology* **2008**, *24*, 921-926.
[24] E. M. J. Gillam, L. M. Notley, H. Cai, J. J. De Voss, F. P. Guengerich, *Biochemistry* **2000**, *39*, 13817-13824.
[25] Q. S. Li, U. Schwaneberg, P. Fischer, R. D. Schmid, *Chemistry* **2000**, *6*, 1531-1536.
[26] W. C. Huang, A. C. Westlake, J. D. Marechal, M. G. Joyce, P. C. Moody, G. C. Roberts, *J. Mol. Biol.* **2007**, *373*, 633-651.
[27] Z. Pengpai, H. Sheng, M. Lehe, L. Yinlin, J. Zhihua, H. Guixiang, *Appl. Biochem. Biotechnol.* **2013**, *171*, 93-103.
[28] H. M. Li, L. H. Mei, V. B. Urlacher, R. D. Schmid, *Appl. Biochem. Biotechnol.* **2008**, *144*, 27-36.
[29] A. Celik, R. E. Speight, N. J. Turner, *Chem. Commun.* **2005**, 3652-3654.

[30] P. P. Kelly, A. Eichler, S. Herter, D. C. Kranz, N. J. Turner, S. L. Flitsch, *Beilstein J. Org. Chem.* **2015**, *11*, 1713-1720.
[31] S. H. Park, et al., *Drug Metab. Dispos.* **2010**, *38*, 732-739.
[32] O. Rousseau, M. C. J. C. C. Ebert, D. Quaglia, S. Iyathurai, J. N. Pelletier, *ChemCatChem* **2018**.
[33] L. X. Tang, H. Gao, X. C. Zhu, X. Wang, M. Zhou, R. X. Jiang, *BioTechniques* **2012**, *52*, 149-+.
[34] H. Li, T. L. Poulos, *Nat. Struct. Biol.* **1997**, *4*, 140.
[35] K. G. Ravichandran, S. S. Boddupalli, C. A. Hasemann, J. A. Peterson, J. Deisenhofer, *Science* **1993**, *261*, 731-736.
[36] H. Li, T. L. Poulos, *Acta Crystallogr D Biol Crystallogr* **1995**, *51*, 21-32.
[37] A. Rentmeister, et al., *ChemCatChem* **2011**, *3*, 1065-1071.
[38] E. Stjernschantz, et al., *Proteins: Struct., Funct., Bioinf.* **2008**, *71*, 336-352.
[39] L. Capoferri, et al., *Proteins* **2016**, *84*, 383-396.
[40] K. L. Tee, U. Schwaneberg, *Angew Chem Int Ed Engl* **2006**, *45*, 5380-5383.
[41] D. C. Haines, D. R. Tomchick, M. Machius, J. A. Peterson, *Biochemistry* **2001**, *40*, 13456-13465.
[42] P. Urban, T. Lautier, D. Pompon, G. Truan, *Int J. Mol. Sci.* **2018**, *19*.
[43] L. M. Podust, T. L. Poulos, M. R. Waterman, *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 3068-3073.
[44] S. N. Daff, S. K. Chapman, R. A. Holt, S. Govindaraj, T. L. Poulos, A. W. Munro, *Biochemistry* **1997**, *36*, 13816-13823.
[45] I. F. Sevrioukova, H. Li, H. Zhang, J. A. Peterson, T. L. Poulos, *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 1863-1868.
[46] T. Zhang, L. A. Liu, D. F. V. Lewis, D.-Q. Wei, *J. Chem. Inf. Model.* **2011**, *51*, 1336-1346.
[47] M. G. Joyce, H. M. Girvan, A. W. Munro, D. Leys, *J. Biol. Chem.* **2004**, *279*, 23287-23293.
[48] H. M. Girvan, H. E. Seward, H. S. Toogood, M. R. Cheesman, D. Leys, A. W. Munro, *J. Biol. Chem.* **2007**, *282*, 564-572.
[49] H. M. Girvan, et al., *Biochem. J.* **2009**, *417*, 65-76.
[50] H. M. Girvan, et al., *J. Biol. Chem.* **2004**, *279*, 23274-23286.
[51] J. Sanchis, et al., *Appl. Microbiol. Biotechnol.* **2008**, *81*, 387-397.

3.10 Supporting information

Supplemental Tables

Table S3-1. Libraries constructed in this study and number of colonies picked for each pool at each position.

Library	#White colonies	#Pale blue colonies	#Dark blue colonies
Lys69	96	-	-
Asn70	96	-	-
Leu71	96	-	-
Leu75[(a)]	96	22	-
Arg79	96	-	-
Asp80	96	-	-
Phe81	96	4	-
Gly83	96	-	-
Asp84	96	-	-
Gly85	96	-	-
Leu86	96	14	-
Thr88	96	-	-
Ser89	96	-	-
Phe81Asp82Gly83	1433	0+6	-
Val78Arg79Asp80[(b)]	1490	19+6	7
Phe87Thr88	375	55	3
Phe173	96	14	-
Met177	96	71	19
Ala180	96	30	5
Leu181	96	13	2
Asp182	96	20	-
Ala184	96	27	8
Asn186	96	5	-
Met212	96	28	10
Leu181Asp182	150	108	11

Arg255	96	43	20
Tyr256	96	8	1
Gln257	96	0+2	-
Ile258	96	-	-
Ile259	96	20	4
Thr260	96	5	2
Phe261	96	-	-
Leu262	96	-	-
Ile263	96	26	15
Gly265	96	-	-
His266	96	-	-
Glu267	96	10	12
Thr327	96	-	-
Ala328	96	18	-
Ser332	96	-	-
Leu333	96	11	-
His266Glu267Thr268	2068	3+6	-
Tyr256Gln257	600	6	-
Leu262Ile263	500	14	4
Pro392	96	10	-
Arg398	96	-	-
Thr436	96	36	-
Leu437	96	-	-
Thr438	96	4	-
Total number of	10648	664	123
Total positions	-	30/49	15/49
Potential active	-	531	122

[a] Pale blue rows are positions where pale blue colonies were picked.

[b] Dark blue rows are positions where pale blue and dark blue colonies were picked.

[c] Colonies were oversampled to account for redundancy and WT clones. The potential active variant row represents, for each pool, the maximum number of variants possible.

Table S3-2. Number of colonies picked according to the PCR regions.

	Residues	Nucleotides	Number of positions	Colonies picked			Total colonies
				Dark blue	Pale blue	White	
PCR1	34-121	100-363	16	10	126	4546	4682
PCR2	156-233	466-698	8	55	316	918	1289
PCR3	230-307	687-921	13	58	143	4320	4517
PCR4	288-362	860-1085	4	0	29	384	413
PCR5	377-456	1130-1368	5	0	50	480	530

Table S3-3. PCR primers for library preparation for MiSeq Illumina[(a)]

Oligo Name	Sequence 5' to 3'
MiSeq_F1(+)Lys69-Ser89	ACACTGACGACATGGTTCTACACGGATGAATTAGGAGAAATC
MiSeq_F1(-)Lys69-Ser89	TACGGTAGCAGAGACTTGGTCTGACCATCATCGCATG
MiSeq_F2(+)Phe173-Met212	ACACTGACGACATGGTTCTACACTTTGCGGCTTTAAC
MiSeq_F2(-)Phe173-Met212	TACGGTAGCAGAGACTTGGTCTTCATCGCTTTGTTCAC
MiSeq_F3(+)Arg255-Glu267	ACACTGACGACATGGTTCTACAAAGCGATGATTTATTAAC
MiSeq_F3(-)Arg255-Glu267	TACGGTAGCAGAGACTTGGTCTTTTGTAGCTTGGAAC
MiSeq_F4(+)Thr327-Leu333	ACACTGACGACATGGTTCTACATATTACAAAAAGCAGCAGAAG
MiSeq_F4(-)Thr327-Leu333	TACGGTAGCAGAGACTTGGTCTTGAAGCTGAGGAATCAG
MiSeq_F5(+)Pro392-Thr438	ACACTGACGACATGGTTCTACACAGAGCGTTTTGAAAATC
MiSeq_F5(-)Pro392-Thr438	TACGGTAGCAGAGACTTGGTCTAAGCGGAATTTTTTTCG

[(a)]Magenta represents CS1 sequences while cyan represents CS2 sequences.

Table S3-4. Variants selectied for homology modeling and MD simulations: 70 active variants identified by NGS, 4 previously reported indigo-producing positive controls and 10 active but indigo-negative controls.

Mutations in dark blue colonies	Mutations in pale blue colonies			Mutations in dark blue and pale blue colonies	
M177S	L75F	D182N	A328F	M177D	R255G
A180M	L75R	D182Q	A328L	M177H	R255H
L181F	F81R	D182Y	A328N	M177K	R255I
L181V	F81W	A184N	A328P	M177N	R255M
A184D	F81Y	A184R	A328V	M177Q	Y256P
A184E	L86I	A184T	L333A	M177W	I259G
A184K	L86Q	N186T	L333C	A180F	I259S
M212F	F173E	M212G	L333F	A180H	T260M
R255D	F173H	M212W	L333R	A180K	I263N
R255W	F173P	M212Y	L333S	A180V	I263Q
Y256C	M177A	R255C	L333T	A184F	I263R
Y256F	M177R	R255E	P392N	M212H	E267F
Y256H	M177T	R255N	P392T	M212K	E267L
I259K	M177Y	R255V	T436C	M212R	E267V
I259P	A180I	Y256G	T436H	R255F	
I263G	A180P	I259R	T436L		
E267H	A180Y	I259V	T436N		
	L181E	T260I	T436Q		
	L181P	T260L	T436R		
	L181W	I263K	T436V		
	D182H	I263P	T438E		
	D182K	I263T	T438L		
Positive control variants	A82W, A82F, F87V, F87P				
Negative control variants	V78F, V78W, V78Y, A82C, A82V, F87S, A264D, A264F, T268A, T268E				

Supplemental Figures

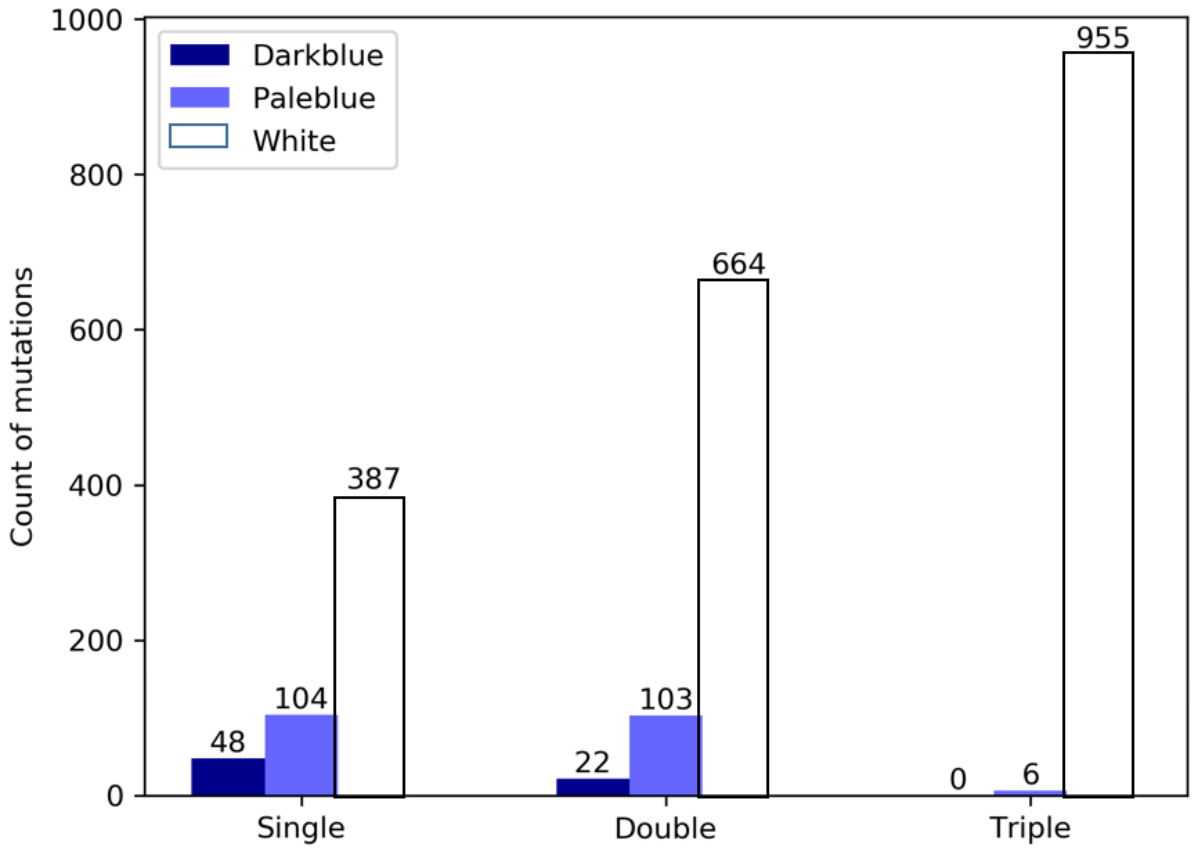


Figure S3-7. Number of individual single/double/triple variants identified in each pool.

Strong positive, weak positive and negative variants are represented in dark blue, pale blue and white respectively.

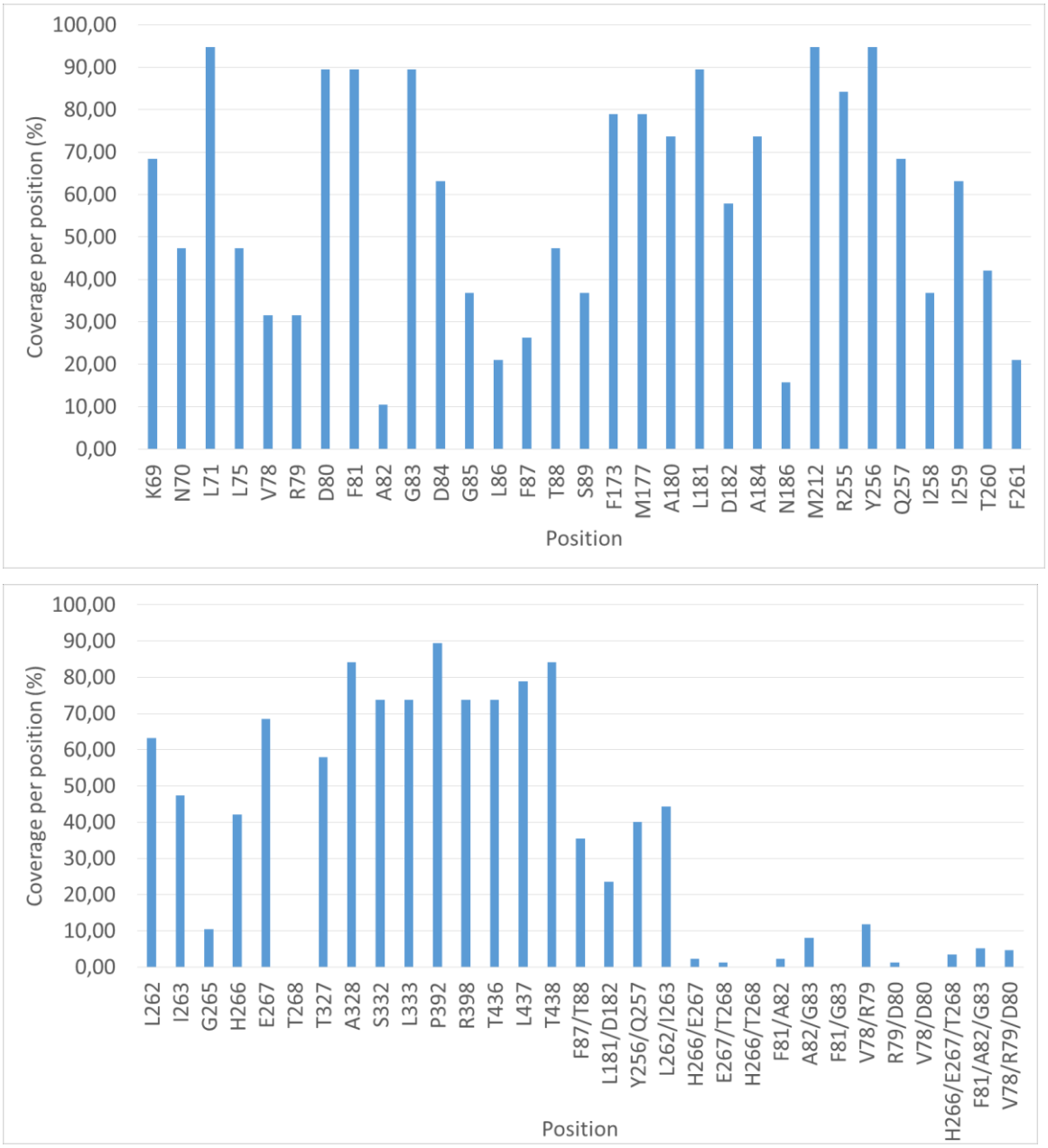


Figure S3-8. Coverage of mutations obtained by NGS for each position.

The coverage of double and triple mutations is lower because the number of mutations per library is greater (400 for doubly-substituted and 8000 for triple-substituted libraries). Single or double mutations can be observed in doubly or triply mutated libraries.

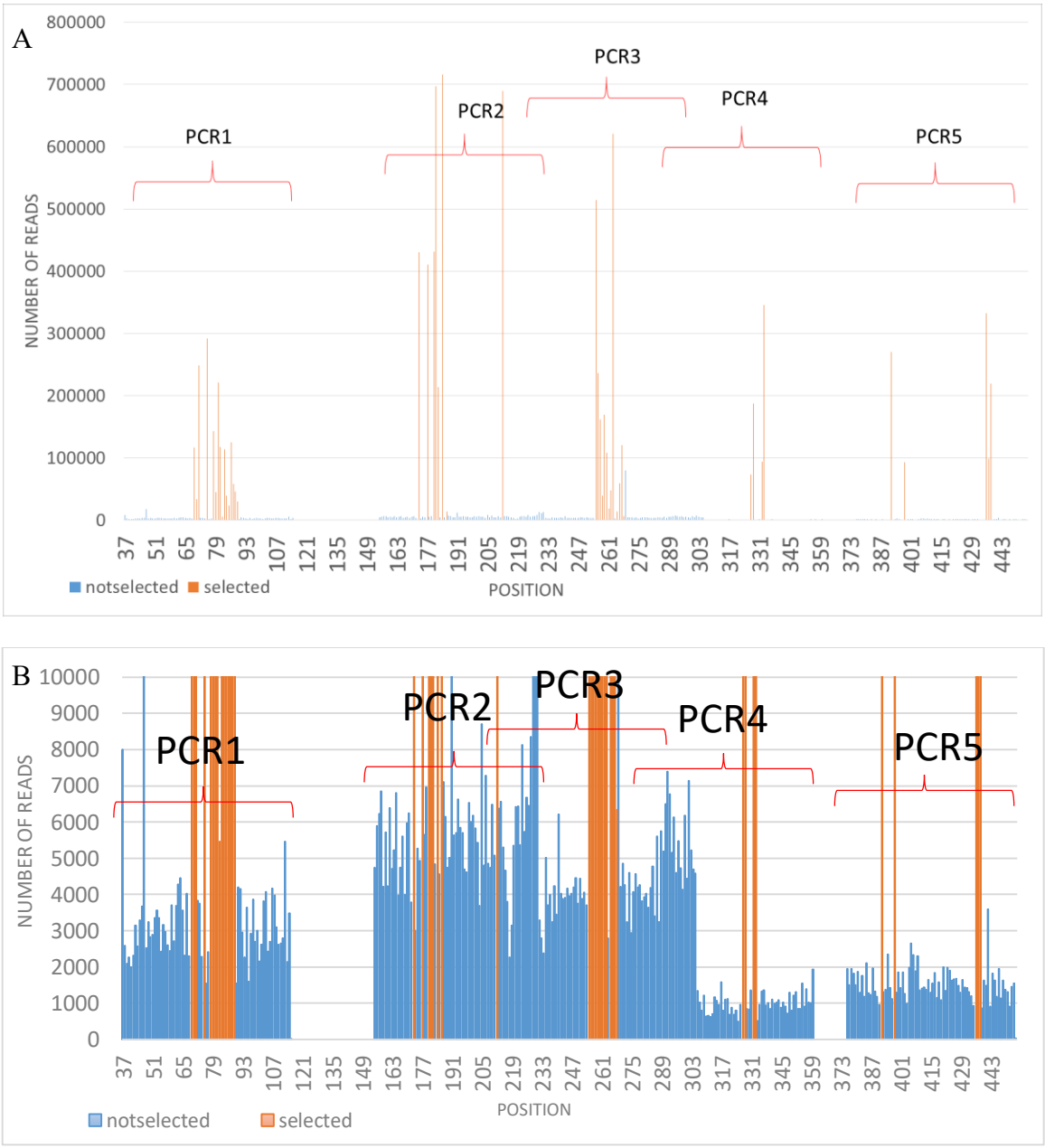


Figure S3-9. Comparison of number of reads between selected and non-selected positions.

Selected positions are in orange and non-selected positions are in blue, illustrating the background read noise for each position. A) Full view. B) Close-up of the position with low frequency reads to highlight the background noise.

Chapitre 4: Conclusions et perspectives

4.1 Conclusions

Le contenu de ce mémoire est une belle représentation de la complémentarité entre la chimie, la biochimie et l'informatique. Ne partant qu'avec une formation de chimiste, j'ai appris largement sur chacun de ces domaines et sur la façon dont ils s'entrecroisent. J'ai exploré la complexité d'un système enzymatique requérant des notions de biologie moléculaire pour la conception des variants, de biologie structurale pour comprendre le mécanisme et dynamisme d'une enzyme, de biochimie pour développer les méthodes de criblage, d'automatisation et d'informatique afin d'accélérer le développement des procédés à l'aide d'instruments sophistiqués et de scripts. Finalement, j'ai eu l'opportunité d'entreprendre un stage de quatre mois en industrie chez Codexis, un des leadeurs mondiaux en matière d'évolution d'enzymes, pour parfaire mes connaissances. Chacun de ces éléments était essentiel à la complétion de ce projet, ayant pour but d'accélérer la découverte de nouvelles réactions catalysées par des variants du cytochrome P450 BM3.

Le chapitre 2 de ce mémoire présente l'optimisation d'une plateforme de criblage à haut débit pour identifier de nouveaux variants actifs pour la synthèse d'indigo ainsi que d'une nouvelle cible : la cétone de framboise. Nous avons identifié 13 nouveaux variants qui produisent l'indigo ainsi que 8 variants produisant la cétone de framboise, approfondissant ainsi les connaissances sur P450 BM3. Nous avons confirmé, en second lieu, l'hypothèse que la production d'indigo à partir d'indole peut être utilisée comme prédicteur pour la transformation d'un substrat sensiblement différent de l'indole. Cela permettra d'accélérer grandement la découverte de nouvelles réactions possibles par P450 BM3 car notre méthodologie, maintenant établie, peut être directement appliquée à une panoplie de substrats industriellement intéressants.

Le chapitre 3 consistait à porter à grande échelle la découverte de variants actifs pour la production d'indigo par le biais du séquençage de nouvelle génération (NGS). Ces travaux ont démontré qu'il existe de nombreuses variations de P450 BM3 qui permettent d'augmenter la production d'indigo. En effet, nous avons rapporté plus de 2000 nouveaux résultats de variants de P450 BM3 rassemblant 29 nouvelles positions ayant la capacité d'oxyder l'indole en

indoxyle afin de donner lieu à l'indigo. L'application du NGS est une méthodologie quasi-inexistante en ingénierie de biocatalyseurs à ce jour. Par nos travaux, nous contribuons à l'implantation de la méthode dans ce domaine. Cela accélèrera grandement le développement de la biocatalyse puisque chacun pourra l'adapter à son système enzymatique.

Somme toute, ces deux approches ont permis d'avancer de façon significative la compréhension des propriétés de l'enzyme cytochrome P450 BM3 donnant lieu à une activité désirée. L'apport à la communauté sur le phénotype des variants pour une activité désirée, autant positif que négatif, contribue à cette explication. Il est à noter que le coût associé à l'identification de variants par NGS maintenant beaucoup plus abordable, permet à la communauté d'avoir accès aux résultats négatifs. Cela permet de bâtir une base de données afin de cartographier une enzyme et son activité pour chaque réaction tentée. En plus, ce genre de série de données permet de construire et perfectionner les algorithmes accélérant l'ingénierie d'enzymes.

4.2 Perspectives de recherche

Maintenant que les méthodologies ont été développées pour entreprendre l'ingénierie à haut débit de P450 BM3, il est possible d'emmener ce projet plus loin. Il serait possible d'améliorer le criblage de variants sur le robot de deux façons. Premièrement, l'essai de fluorescence pour le suivi de la consommation de NADPH pourrait être appliqué pour calculer des paramètres catalytiques des variants tels que l'activité spécifique (unités d'activité par mg d'enzyme), k_{cat} (le taux de conversion par molécule d'enzyme) et K_M (la constante d'affinité effective). Cela permettrait de rendre l'essai quantitatif ainsi que de classer rapidement les variants obtenus pour une réaction donnée. Deuxièmement, le criblage sur plateforme de pipetage automatisée se fait présentement par variants individuels (1 variant par puit). Il serait intéressant d'explorer la possibilité de cribler par position (19 variants par puit) pour éliminer plus rapidement les positions où aucun variant n'est actif. Cela devient pertinent lorsque les librairies à cribler atteignent de l'ordre de grandeur de celles présentées dans le chapitre 3, ou plus encore.

Ces méthodes peuvent ensuite être utilisées pour cribler nos librairies pour de nouveaux substrats industriellement intéressants. Par exemple, il existe plusieurs bases de données telles que ZINC [1] et CoCoCo [2] contenant des millions de molécules à des fins de criblages virtuels qu'il serait possible d'explorer pour découvrir de nouvelles réactions d'oxydation intéressantes.

En effet, ZINC peut identifier des molécules industriellement intéressantes par l'entremise d'analogues commercialement disponibles en plus de documenter leurs cibles ainsi que leurs applications. Par exemple, à partir du tryptophane, ZINC a attiré mon attention sur la possibilité de synthétiser le précurseur de la sérotonine par P450 BM3. Le précurseur, le 5-hydroxytryptophane (5-HTP), est synthétisé dans notre corps à partir du tryptophane et est actuellement vendu comme supplément afin traiter la dépression [3], l'anxiété [4], l'obésité [5] et autres troubles neurologiques (Figure 4-1).

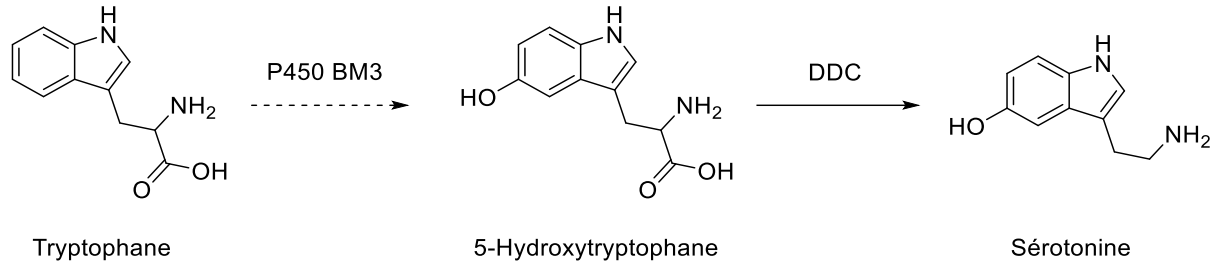


Figure 4-1. Biosynthèse hypothétique du précurseur de la sérotonine par le cytochrome P450 BM3 et ses variants.

Naturellement, le tryptophane est converti dans le corps en 5-hydroxytryptophane (5-HTP) par l'hydroxylase de tryptophane. P450 BM3 et ses variants pourraient remplacer cette étape pour produire le supplément alimentaire. Lorsque le précurseur traverse la barrière hémato-encéphalique, il est converti en sérotonine par la décarboxylase d'acide aminé (DDC). La sérotonine agit comme neurotransmetteur.

Au même titre que l'indigo, son extraction à partir d'une plante est beaucoup plus laborieuse que sa synthèse en laboratoire, alors il serait avantageux d'explorer cette avenue avec P450 BM3.[6] Cette réaction revient à hydroxyler le tryptophane à une position différente de l'indoxyle (le précurseur de l'indigo).

Une autre approche intéressante pourrait être de combiner différentes enzymes simultanément pour obtenir une voie de synthèse en un seul réacteur (*one-pot reaction*). Par exemple, il a été démontré qu'un agent pharmaceutique ayant des propriétés neuro-protectrices a été synthétisé selon une voie biocatalytique à l'aide d'une double oxydation par deux différents variants de P450 BM3.[7] Ces variants contiennent des mutations aux positions que nous avons explorées dans cet ouvrage, soit L75, V78, F87 et I263. P450 BM3 pourrait aussi être combiné avec des enzymes de différentes classes. Par exemple, P450 BM3 a été combiné dans un « one-pot

reaction » avec une lyase de phénol tyrosine pour synthétiser des précurseurs clés de médicaments anti-cancers.[8]

Lors de la synthèse de la cétone de framboise, il a fallu confirmer son identité par RMN car l'oxydation observée par spectrométrie de masse aurait pu avoir lieu à une autre position sur le cycle aromatique ou sur la chaîne aliphatique de la 4-phényl-2-butanone. Il a été laborieux d'obtenir suffisamment de produit réactionnel pour en effectuer une analyse sans équivoque. Il serait ainsi pertinent d'optimiser la biotransformation à plus grande échelle. Cela pourrait se faire en optimisant les conditions de la biotransformation ou par l'incorporation d'un système de régénération du NADPH : le facteur limitant lors de la réaction par P450 BM3 est la quantité du NADPH utilisé car il est dispendieux (1600$/g). Le couplage de P450 BM3 avec une déshydrogénase permettrait d'atteindre des quantités plus élevées de produit et ce, à moindre coût.[9]

Finalement, nos librairies de variants devenant de plus en plus grandes, les méthodes de mutagénèse et de criblage devraient être conçues avec l'objectif de maximiser l'apport de l'approche de NGS. Le nombre de mutations accumulées durant ces travaux peut être amplifié de façon exponentielle en appliquant les méthodes de mutagénèse aléatoire ou bien de recombinaison aléatoire (*DNA shuffling*) sur les librairies existantes. Elles pourraient ensuite être criblées et les variants actifs identifiés par NGS. L'application du NGS afin d'observer des cycles d'évolution lors de l'ingénierie de P450 BM3 est d'autant plus intéressant car chaque ronde d'évolution pourrait être associée à un code-barres pour observer l'augmentation de la fonction désirée. Cette méthode est normalement appliquée pour observer l'enrichissement de variants lors d'essais de sélection pour la survie lors de l'étude d'enzymes essentielles, mais pourrait être adaptée à l'évolution dirigée par criblage de biocatalyseurs tels P450 BM3.

4.3 Perspectives du domaine

Nous entrons dans la quatrième vague de la biocatalyse et avec son expansion constante, le futur du domaine est rayonnant. Il est prédit que le marché global d'enzymes industrielles croîtra de 5.01 milliards en 2016 à 6.32 milliards de dollars US en 2021.[10]

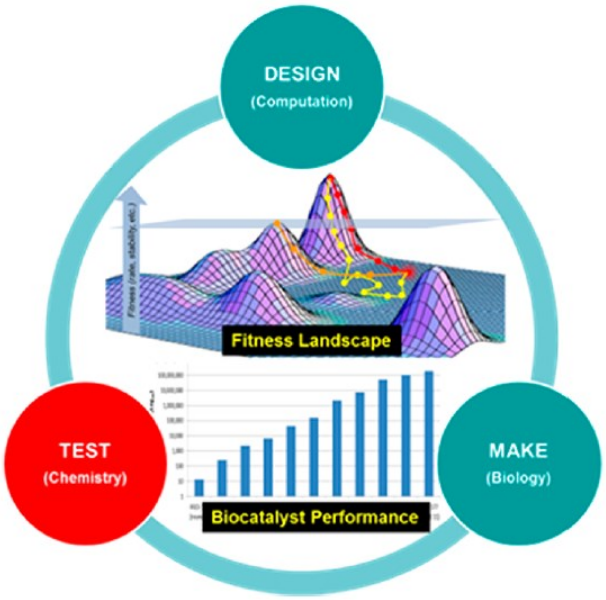


Figure 4-2. Le cycle « Design – Make – Test » de la quatrième vague de la biocatalyse.
Représentation du cycle effectué pour améliorer la fonction désirée d'un biocatalyseur. « Design » : Prédiction de mutations bénéfiques pour la fonction désirée à l'aide de logiciels bio-informatiques. « Make » : Conception des mutations prédites. « Test » : Criblage des mutations prédites. Figure reproduite de [11].

Cette quatrième vague propose un flux de travail pour guider le plus efficacement possible l'ingénierie d'enzymes en utilisant un cycle d'évolution de biocatalyseurs combinant étroitement les techniques de laboratoire et les outils informatiques (Figure 4-2). Chacune des sphères de ce cycle s'améliore continuellement faisant en sorte que la biocatalyse devient une partie intrinsèque de la chimie organique. Cela permet d'enrichir la boîte à outils de la synthèse organique.[12]

À cette fin, diverses avenues de recherche sont proposées en biocatalyse. Par exemple, de nouvelles classes d'enzymes sont développées par ingénierie d'enzymes telles réductases d'imines, les réductases d'acides carboxyliques, les Diels-Alderases et les méthyltransférases permettent d'avoir accès à de nouvelles réactivités. De nouvelles enzymes sont aussi découvertes grâce aux études métagénomiques démontrant des réactivités non explorées. En plus de ces nouvelles classes, de nouvelles réactions non-naturellement produites par les enzymes sont en émergence telles que la cyclopropanation sélective, l'arizidination ou bien la formation de lien Si-C.[13]

Ensuite, l'une des plus grandes contraintes lors du développement de biocatalyseurs est la stabilité des enzymes. Dû à l'utilisation de solvants organiques, les enzymes sont plus difficiles à incorporer dans les processus industriels à grande échelle. Beaucoup de recherches sont investies dans l'immobilisation d'enzymes leur permettant d'être actifs en solvants organiques et de les recycler pour leur réutilisation.[14-15] Par exemple, la synthèse d'un médicament anti-vomissement (EMEND), utilisé lors la chimiothérapie, est fortement améliorée par le remplacement d'une étape de synthèse organique par une réductase de cétone immobilisée. En plus d'obtenir un rendement de 98-99% avec >99% ee, le lot de la réductase de cétone a pu fonctionner dans des conditions de 90% de solvant organique en continu, ne perdant que 6% d'activité après sept jours.[14] De surcroît, cette approche simplifie les procédés par la réutilisation de l'enzyme et la facilitation de séparation de produits.[10]

Finalement, et non le moindre : l'apprentissage machine (« *Machine learning* »). Cette approche s'applique maintenant à une panoplie de domaines, l'ingénierie d'enzymes n'y faisant pas exception. L'apprentissage machine identifie les tendances manifestées au sein de séries de données contribuant à l'accélération de l'ingénierie d'enzymes. Il peut être utilisé pour prédire les propriétés de protéines [16], les interactions protéine-protéine [17], les structures secondaires [18] et plus encore [19]. En ingénierie d'enzymes, nous observons déjà des applications pour prédire des réactions enzymatiques promiscuitaires.[20] En plus, il est incorporé dans le développement d'algorithmes sophistiqués en industrie; chez Codexis par exemple, pour l'identification des meilleurs variants par l'analyse de données sur les séquences, l'activité et les structures.

Dans ce mémoire, nous avons exploré ces trois sphères pour le système du cytochrome P450 BM3, nous permettant d'être à la pointe de la technologie en biocatalyse. L'avancement des méthodes de criblage et l'informatique nous approche petit à petit vers la compréhension de ces systèmes biologiques complexes afin d'exploiter leur plein potentiel. La biocatalyse est vouée à être pleinement incorporée dans les procédés industriels au profit premier de l'efficacité, mais encore plus de l'environnement. C'est en utilisant la nature que nous serons capable de mieux la respecter.

« Jamais la nature ne nous trompe : c'est toujours nous qui nous trompons » (Jean-Jacques Rousseau)

4.4 Références

[1] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, R. G. Coleman, *J. Chem. Inf. Model.* **2012**, *52*, 1757-1768.

[2] A. Del Rio, A. J. M. Barbosa, F. Caporuscio, G. F. Mangiatordi, *Mol. BioSyst.* **2010**, *6*, 2122-2128.

[3] E. H. Turner, J. M. Loftis, A. D. Blackwell, *Pharmacol. Ther.* **2006**, *109*, 325-338.

[4] R. S. Kahn, H. G. Westenberg, W. M. Verhoeven, C. C. Gispen-de Wied, W. D. Kamerbeek, *Int. Clin. Psychopharmacol.* **1987**, *2*, 33-45.

[5] C. Cangiano, et al., *Am. J. Clin. Nutr.* **1992**, *56*, 863-867.

[6] P. A. Lemaire, R. K. Adosraku, *Phytochem. Anal.* **2002**, *13*, 333-337.

[7] P. Le-Huu, D. Petrović, B. Strodel, V. B. Urlacher, *ChemCatChem* **2016**, *8*, 3755-3761.

[8] A. Dennig, E. Busto, W. Kroutil, K. Faber, *ACS Catalysis* **2015**, *5*, 7503-7506.

[9] W. Hummel, H. Gröger, *J. Biotechnol.* **2014**, *191*, 22-31.

[10] J. Chapman, A. Ismail, C. Dinu, *Catalysts* **2018**, *8*, 238.

[11] M. D. Truppo, *ACS Med. Chem. Lett.* **2017**, *8*, 476-480.

[12] C. M. Clouthier, J. N. Pelletier, *Chem. Soc. Rev.* **2012**, *41*, 1585-1605.

[13] U. T. Bornscheuer, *Philos. Trans. Royal Soc. A* **2018**, *376*.

[14] H. Li, J. Moncecchi, M. D. Truppo, *Org. Process Res. Dev.* **2015**, *19*, 695-700.

[15] M. D. Truppo, H. Strotman, G. Hughes, *ChemCatChem* **2012**, *4*, 1071-1074.

[16] K. K. Yang, Z. Wu, C. N. Bedbrook, F. H. Arnold, *Bioinformatics* **2018**, *34*, 2642-2648.

[17] T. Sun, B. Zhou, L. Lai, J. Pei, *BMC Bioinformatics* **2017**, *18*, 277.

[18] S. Wang, J. Peng, J. Ma, J. Xu, *Scientific Reports* **2016**, *6*, 18962.

[19] J. Wang, H. Cao, J. Z. H. Zhang, Y. Qi, *Scientific Reports* **2018**, *8*, 6349.

[20] D. A. Pertusi, M. E. Moura, J. G. Jeffryes, S. Prabhu, B. Walters Biggs, K. E. J. Tyo, *Metab. Eng.* **2017**, *44*, 171-181.