

Computer-aided design (CAD) tools for bioproduction and biosensing pathway engineering

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université d'Évry-Val-d'Essonne

École doctorale n°577 Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la Vie et de la Santé

Thèse présentée et soutenue à Jouy-en-Josas, le 7 décembre 2017, par

Baudoin Delépine

Composition du Jury :

Janet Thornton	
Professeure, European Bioinformatics Institutes (European Molecular Biology Laboratory)	Présidente
Stéphanie Heux	
Directrice de recherche, INRA (LISBP, Université de Toulouse, CNRS)	Rapporteuse
Olivier Taboureau	
Professeur, Université Paris Diderot (Inserm UMR-S 973)	Rapporteur
David Vallenet	
Chercheur, CEA (Laboratory of Bioinformatics Analyses for Genomics and Metabolism)	Examineur
Jean-Loup Faulon	
Directeur de recherche, INRA (Micalis UMR 1319, AgroParisTech, Université Paris-Saclay)	Directeur de thèse
Pablo Carbonell	
Chargé de recherche, The University of Manchester (BBSRC/EPSC SynBioChem Centre, MIB)	Co-directeur de thèse

Abstract

Advances in systems and synthetic biology are fueling our ability to develop successful metabolic engineering applications for the sustainable production of bio-based chemicals. We can envision a future in which *designer cells* could be engineered to transform any carbon source into any target compound. This daunting task will be achieved by leveraging methods that proved themselves in other engineering disciplines. Among those, the use of Computer Aided Design (CAD) softwares is expected to reduce the amount of time and expert knowledge needed to design *de novo* metabolic pathways. The first part of this thesis is dedicated to our pathway prediction algorithm and its CAD implementations. Most notably, we will present RetroPath2.0, a versatile reaction network prediction framework focused on retrosynthesis that is built to be easily extensible by the community. In the second part, we will highlight the interest of intracellular biosensors for metabolic engineering and introduce SensiPath, a web application that uses a reaction prediction engine to design biosensing circuits for compounds for which no direct biosensors are known. Altogether, this thesis proposes that bioCAD tools should focus on empowering users' creativity and encourage them to explore original applications.

Résumé

Les récentes avancées en biologie des systèmes et en biologie synthétique contribuent déjà au fleurissement d'applications en ingénierie métabolique visant une bioproduction renouvelable de composés chimiques. Nous pouvons entrevoir un futur où des microbes seraient conçus à la carte afin de valoriser n'importe quelle source de carbone en n'importe quel composé d'intérêt. Si la route est longue avant l'accomplissement d'un tel objectif, son parcours devrait en être grandement facilité par l'exploitation de méthodes d'ingénierie déjà éprouvées dans d'autres disciplines. On s'attend entre autre à ce que l'utilisation de logiciels de Conception Assistée par Ordinateur (CAO) diminue le temps et l'expertise nécessaires à la construction de voies métaboliques n'existant pas dans la nature. La première partie de cette thèse est dédiée à notre méthode de prédiction de voies métaboliques et à ses implémentations. Nous décrivons tout particulièrement RetroPath2.0, un outil de prédiction de réseaux de réactions mettant l'accent sur les applications de rétrosynthèse, et qui est construit pour être facilement extensible par la communauté. Dans la seconde partie, nous détaillons l'intérêt des biosenseurs intracellulaires pour l'ingénierie métabolique et introduisons SensiPath; une application web qui exploite un outil de prédiction de réactions pour concevoir des circuits métaboliques permettant la biodétection de composés pour lesquels aucun biosenseur direct n'est connu. Dans l'ensemble, cette thèse propose que les outils de bioCAO devraient permettre de révéler la créativité de leurs utilisateurs et encourager l'exploration de nouvelles applications.

Remerciements

Je suis particulièrement reconnaissant au Jury d'avoir trouvé le temps de lire ma thèse et d'avoir accepté de prendre part à ma soutenance.

Jean-Loup Faulon et Pablo Carbonell, merci de m'avoir donné l'opportunité de travailler sur vos thématiques de recherche en acceptant de m'encadrer. Vous m'avez donné les codes pour comprendre et apprécier les rouages du monde scientifique. C'est grâce à vous si je fais aujourd'hui un peu partie de cette communauté.

Merci aux membres de l'équipe BioRetroSynth ainsi que tous ceux que j'ai pu rencontrer au sein de l'iSSB et de Micalis, et qui ont partagé leur enthousiasme, leur expérience, et finalement contribué durant ces trois ans à un contexte professionnel particulièrement agréable et enrichissant.

Vincent Libis et Thomas Duigou, sans vous cette thèse serait bien peu de choses ! Ce fût un véritable plaisir de collaborer ensemble et j'espère que vous en garderez d'aussi bons souvenirs que moi. Vous avez été une source de motivation constante et avez largement contribué à développer mes compétences scientifiques et techniques. Merci !

Ioana, Cyrille, Valérie, Heykel, André, Philippe, Pierre, Bedis, Anne-Laure, merci pour vos conseils et ces riches discussions qui m'ont permis de me familiariser avec les subtilités de la "réalité de terrain" qu'on peut observer au laboratoire. Merci à vous pour votre patience et votre bienveillance.

Julie, Aline et Mathilde; merci pour votre indéfectible soutien moral ! Merci aussi à tous les autres doctorants que je ne saurais nommer ici et que j'ai eu le plaisir de rencontrer à Évry et à Jouy.

Sylvie et Bernadette, merci pour votre bonne humeur et votre aide précieuse dans mes déboires administratifs !

Enfin, j'aimerais remercier mes grand-parents, mon oncle et mon beau-père pour m'avoir donné le goût des sciences et l'envie d'y contribuer.

Contents

Abstract	i
Résumé	iii
Remerciements	v
1 Introduction	1
1.1 Metabolic engineering for a sustainable chemistry	1
1.1.1 Concerns about the sustainability of synthetic organic chemistry	1
1.1.2 Bioproduction and renewable carbon-source	2
1.1.3 Development of microbial cell factories	4
1.2 Computer-aided pathway design	6
1.2.1 Earliest attempts and motivation to predict enzymatic promiscuity	6
1.2.2 Structure-based generalist reaction rules	7
1.2.3 From metabolic space generation to <i>de novo</i> pathways design	9
1.2.4 Retrosynthesis for an efficient navigation into the metabolic space	10
1.3 Intracellular biosensors	12
1.3.1 Microbial systems for the detection of small chemicals	12
1.3.2 Other biosensors	14
1.3.3 The need for biosensors in metabolic engineering	15
1.4 Thesis structure	18
I Retrosynthesis for metabolic engineering	19
2 Extended Metabolic Space Modelling	21
2.1 Abstract	21
2.2 Introduction	22
2.3 Materials	23
2.4 Methods	23
2.4.1 Computation of molecular signatures	23

2.4.2	Computation of reaction signature	27
2.4.3	Products enumeration	28
2.4.4	Chassis modeling in the extended metabolic space	29
2.4.5	Computing the scope	30
2.4.6	Enumerating pathways	31
2.4.7	Design in the extended metabolic space	32
2.5	Notes	34
3	RetroPath2.0: a retrosynthesis workflow for metabolic engineers	37
3.1	Abstract	38
3.2	Introduction	39
3.3	Theoretical background	40
3.3.1	Encoding reaction as reaction rules	41
3.3.2	Building (retrosynthesis) reaction network	46
3.4	Methods	48
3.4.1	Reaction rules	49
3.4.2	Building (retrosynthesis) reaction networks between two pools of compounds using the RetroPath2.0 workflow	50
3.4.3	Score rules by enzyme sequence consistency	51
3.4.4	Enumerating pathways between two pools of compounds	54
3.5	Results	55
3.5.1	Rules validation	55
3.5.2	Workflow validation and applications	59
3.6	Discussion	61
3.7	Supporting information	68
3.7.1	Reaction rule diameter and promiscuity recovery	68
3.7.2	Details about multimolecular rules	70
3.7.3	Detection of biomarkers through metabolic circuits	71
3.7.4	Description of supplementary data	73
	Conclusion & perspectives	75
II	Computer-aided design of sensing-enabling metabolic path- ways	79
4	Sensing new chemicals with bacterial transcription factors	81
4.1	Highlights	81
4.2	Abstract	82
4.3	Introduction	82
4.4	Mining allosteric transcription factors in nature	83
4.5	Engineering the sensing scope	84
4.6	Regulation via indirect sensing	87

4.7	Conclusion	90
5	SensiPath: computer-aided design of Sensing-enabling metabolic Pathways	93
5.1	Abstract	94
5.2	Introduction	94
5.3	Material and Methods	95
5.3.1	Source databases	96
5.3.2	Reaction and compound encoding	97
5.3.3	Matching algorithm	98
5.3.4	Metabolic graph	98
5.3.5	Similarity search	99
5.3.6	Web server implementation	99
5.4	Input and Output	99
5.4.1	Input	99
5.4.2	Output	99
5.5	Case studies	100
5.5.1	Cocaine detection	100
5.5.2	Parathion detection	102
5.6	Discussion	102
6	Expanding biosensing abilities through computer-aided design of metabolic pathways	105
6.1	Abstract	106
6.2	Introduction	106
6.3	Results	108
6.3.1	The scope of detectable molecules is significantly expanded <i>in silico</i> via enzymatic pathways	108
6.3.2	Implementation of SEMP <i>s in vivo</i> expands <i>E. coli</i> sensing abilities	111
6.4	Discussion	115
6.5	Methods	117
6.5.1	Data source	117
6.5.2	Molecular and reaction signatures	117
6.5.3	Compounds and reactions preprocessing	118
6.5.4	SEMP prediction	118
6.5.5	Chemicals and reagents	119
6.5.6	Genetic constructs	119
6.5.7	Biosensor dose-response characterization	120
6.5.8	Characterization of the transcriptional response to nitro- glycerin of different <i>E. coli</i> strains	121
6.6	Supporting information	122

Conclusion & perspectives	133
Bibliography	139

List of Figures

1.1	Concept of microbial cell factory.	3
1.2	Classification of compounds produced by metabolic engineering. . .	5
1.3	Interface of the first computer-aided design software dedicated to retrosynthesis.	10
1.4	Concept of whole-cell biosensor.	15
1.5	Metabolic valve to balance between growth and production objectives.	15
1.6	Selection of superproducers among a library of variants with biosensors.	16
2.1	Steps involved in the construction of the extended metabolic space.	24
2.2	Reaction signature of an aspartate transaminase (EC 2.6.1.1, $d=4$). . .	28
2.3	Example of pathway enumeration in the extended metabolic space.	33
3.1	Graphical abstract for RetroPath2.0.	37
3.2	Example of reaction rules.	42
3.3	RetroPath2.0 rules and corresponding SMARTS for reaction 2.6.1.1 at various diameters.	43
3.4	Reactions rules for EC class 4.5.1.-.	45
3.5	RetroPath2.0 KNIME workflow.	48
3.6	Receiving operating characteristic curves (ROC) curves for the rules of RetroPath2.0 of diameter $d = 8$	58
3.7	Box plot comparing the distribution of reaction scores for specialist and generalist enzymes in <i>E. coli</i>	58
3.8	Enumerated pathways for the production of styrene.	62
3.9	Enumerated pathways for the production of the non-natural compound terephthalic acid (TPA, compound A) from <i>E. coli</i>	63
3.10	Enumerated pathways for the production of ethylene glycol.	64
3.11	Example of a scope obtained for sarcosine.	66
3.12	Specificity of rules vs. promiscuity recovery in function of diameter.	70
4.1	Schema of the dual selection employed while screening allosteric transcription factor (aTF) variants.	84
4.2	Recent examples and chemical structures of new effectors yielded by various aTF redesign strategies.	85

4.3	Examples of synthetic and natural sensing-enabling metabolic pathways (SEMP).	88
5.1	Concept of sensing-enabling metabolic pathway.	93
5.2	SensiPath flowchart.	96
5.3	Results pages for querying "CHEBI:60056" at one step.	100
6.1	Graphical abstract depicting the experimental validation of Sensing Enabling Metabolic Pathways.	105
6.2	General concept of a cell equipped with a sensing-enabling metabolic pathway (SEMP) allowing for the detection of a new chemical. . . .	106
6.3	Sensing-enabling metabolic pathway (SEMP) design workflow. . . .	107
6.4	<i>In silico</i> prediction of detectable compounds among three data sets.	110
6.5	<i>In vivo</i> characterization of sensing-enabling metabolic pathways (SEMP).	112
6.6	Extended metabolic space around the DrugBank dataset.	122
6.7	Monitoring of the response time to hippurate (a), cocaine (b) and benzoate of strains equipped with SEMP.	123
6.8	Summary of the SEMP successfully implemented <i>in vivo</i> over the course of this work.	124
6.9	Architecture of metabolic and sensing modules.	124
6.10	Dose response of control strains lacking metabolic module to the different compounds.	125
6.11	Summary of SEMP-associated specificity modes of failure.	126
6.12	Flow cytometry analysis of strains equipped with SEMP in presence of toxic concentrations of inducer.	127
6.13	Growth curves of the different strains in presence of the associated compounds.	129
6.14	Swing and RFP/OD of all strains with every compounds.	131

List of Tables

2.1	A selection of software tools for modeling in the extended metabolic space.	25
3.1	Retrosynthesis networks generation tools.	46
3.2	Assignment of sequences to rules at different diameters and using decreasing EC class level of specificity.	53
3.3	Examples of metabolic pathways enabling the detection of prostate cancer biomarkers.	65
4.1	Examples of indirect sensing via metabolism.	87
6.1	Example of Sensing-Enabling Metabolic Pathway (SEMP)	109

Introduction

1.1 Metabolic engineering for a sustainable chemistry

1.1.1 Concerns about the sustainability of synthetic organic chemistry

Organic chemistry rose at the dusk of the 20th century from a descriptive science to an engineering discipline that quickly contributed to the wealth of our society. Advances in synthetic chemistry allowed the development of polymers and bioactive compounds for which we have a direct or indirect usage every day of our life. Fertilizers and pesticides secure our food production, synthetic drugs protect us from misery, while plastics shape our world. However, concerns are rising about the sustainability of this model, and the chemical industry, once cherished as the promise of a better world, is now seen as a hazard for our society.

The attractiveness of synthetic chemistry was built on the development of methods to efficiently assemble small chemicals into more complex ones through standardized sequences of transformations. The catalysts and solvents needed for those transformations can be toxic [1], and even reactions reputed and appreciated for the mild conditions they require use nonrenewable materials. Their disposal and recycling is a challenge [2]. For instance, the widely used Sonogashira reaction that forms carbon-carbon bonds at room temperature uses palladium, a catalyst that is already considered rare and expensive [3, 4].

At the root of organic chemistry's success, commodity chemicals are building blocks used as standard parts in numerous synthesis routes. Most of them are produced from nonrenewable fossil carbon sources by the petrochemical industry [5] at low-price and high-volume. With the rarefaction of oil reserves that can be exploited at low-price, it is the very basis of the organic chemical industry model that is threatened. However, not all commodity chemicals come from the petrochemical industry and some of them, such as ethanol, are already isolated from renewable sources [6].

1.1.2 Bioproduction and renewable carbon-source

There is much hope in the development of biorefineries, and ultimately their contribution to the replacement of fossil carbon sources. Biorefineries transform biomass coming from the agriculture industry into valuable compounds using chemical, biological or a combination of technologies coming from both worlds [5, 7]. State-of-the-art biorefineries mostly use edible biomass to produce a small palette of commodity chemicals from microbes [8, 9]. The diversion of arable lands and feedstock toward the chemical industry is raising concerns about its putative impact on the prices for human (and animal) consumption, and ultimately about the sustainability of this model. Moreover, the relative high price of edible biomass compared to oil compromises the competitiveness of bioproduction for most commodity chemicals since they would require titers, rates of production and yields that are still unreachable [10].

Consequently, wastes from the agriculture industry are regarded as a promising alternative carbon-source. In that spirit, we assisted in the recent years to considerable efforts to increase the conversion efficiency of lignocellulosic biomass and other low-value carbon sources by microbes [7, 11] in order to unleash their fantastic catalytic potential.

Enzymes are already used in the industry by synthetic chemists for their unmatched catalytic capabilities [12, 13]. Indeed, enzymes benefit from a superior stereoselectivity permitted by the conformation of their binding pocket [12, 14]. On the other hand, they are much more fragile and difficult to produce (extract and isolate) than their synthetic catalysts counterparts. In that context, an attractive solution is to use living organisms to perform the synthesis, in which enzymes and cofactors are naturally recycled, and then extract the desired product.

Semisynthesis is a common strategy that consists in initiating the synthesis from a compound extracted and isolated from a biological source and complete the synthesis with usual organic chemistry methods. Semisynthesis is particularly relevant when the structure of a target compound is too complex for a competitive total synthesis from commodity chemicals, and that a structurally similar compound can be bioproduced by an enzymatic pathway (and extracted) at a lower cost. This makes it a strategy of choice for the production of fine chemicals that have a high market demand for their special properties or structure. For instance, the worldwide production of morphine is made in crops (opium poppy, *Papaver somniferum*) despite the fact that more than 30 total chemical synthesis routes are known for this compound [15]. Once isolated, morphine

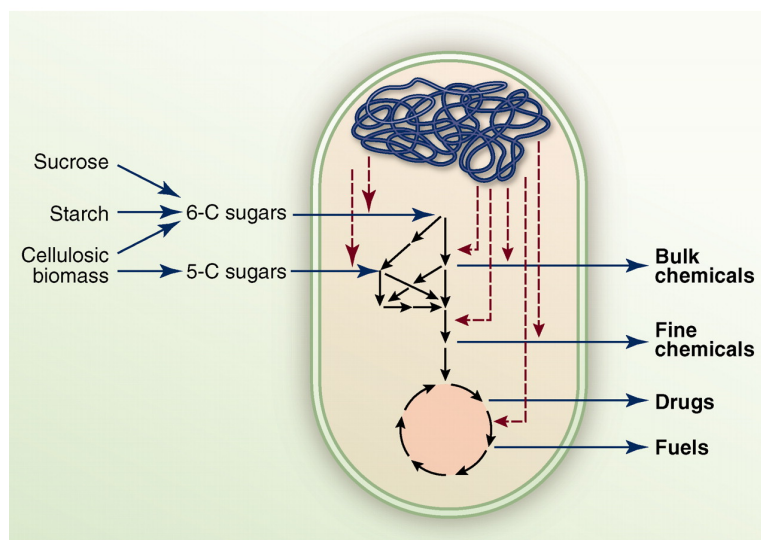


Fig. 1.1: Concept of microbial cell factory. The hope carried out by metabolic engineering is that, by the right set of genetic modifications, we would transform bacteria so that they could consume any suitable carbon source and transform it into any desirable chemical for our industry. This would be the guarantee of a sustainable chemical production model. Extracted from [18].

can later be used as a platform and transformed into less addictive drugs such as codeine [16].

The exploitation of crops is not exempt of limitations, be it to use them as a carbon source or as a catalytic vessel. Like any industrial installation, biorefineries must be fed continuously along the year with a sufficient amount of material respecting quality standards to meet profitability. This is hardly the case for biomass, whose availability obviously spikes during harvest seasons and is vulnerable to climatic events and, more broadly, to climate change [5]. The same goes with semisynthesis, with the additional risks arising from an increase susceptibility to political instabilities and pest catastrophes if the bioproduction is made by specific plants with precise climate needs.

For those reasons, the development of synthetic processes able to tap into flexible (renewable) carbon sources with a catalytic power equivalent to synthetic organic chemistry is highly desirable as it would secure our supply of chemicals [17]. Bacteria and yeasts are ideal candidates for that since they have already proved themselves in biotechnological processes, and they can easily be modified by current genetic engineering tools (chassis organisms) to reach specific production objectives.

1.1.3 Development of microbial cell factories

Metabolic engineering [19] is a research field that vowed to transform cells into factories by redirecting their resources toward the production of chemicals of interest (Figure 1.1). Metabolic engineering significantly grew with the advances in systems and synthetic biology that respectively provided the map to understand metabolism and the tools to modify it. The hope carried out by the development of metabolic engineering is that, by performing the right genetic modifications, we could unchain cell's productivity up to a profitable level by increasing their titer, rate of production and yield [10] while being able to consume any carbon source [20].

There is no definitive method to increase the production of a compound, for the reason that organisms are complex systems and their enhancement needs to compromise between multiple objectives. Indeed, cells are already constrained by their evolutionary history and inserting new objectives without breaking cells' viability is a daunting task [10]. Importantly, it is not a matter of overproducing every enzymatic step since that would end up in a misallocation of resources (cofactors, amino acids, ribosomes) that could wreck metabolism by starving other processes, or reach toxicity levels for intermediates or final compounds. Modern metabolic engineering (sometimes called *systems* metabolic engineering) focuses more on balancing the whole metabolism toward the production objective at the maximum of what the cell can take [21]. This includes but is not limited to optimizing the carbon uptake, deleting competitive pathways, improving product tolerance, designing regulation mechanisms, and of course finding a proper producing pathway [22].

The fast paced advances of synthetic organic chemistry during the last century was hallmarked by the total synthesis of emblematic chemicals, and this should not come as a surprise that we assist nowadays to the same pattern in metabolic engineering [22, 24, 25]. Jeong Wook Lee *et al.* [23] proposed to classify those chemicals into four categories on the basis of whether they are known to be naturally produced by unmodified organisms (natural vs. non-natural), and whether the pathways proposed for the production were naturals (inherent), reconstructed from natural enzymes (noninherent) or needed new enzymatic activities (created). See also Figure 1.2. This classification sheds some light on the degree of engineering that is required for the production of a compound, and the advances metabolic engineering is making. The bioproduction of natural compounds by inherent pathways is well-known and was exploited for centuries (fermentations). The current challenge is to streamline the bioproduction through noninherent pathways, both for natural and non-natural compounds, at a time

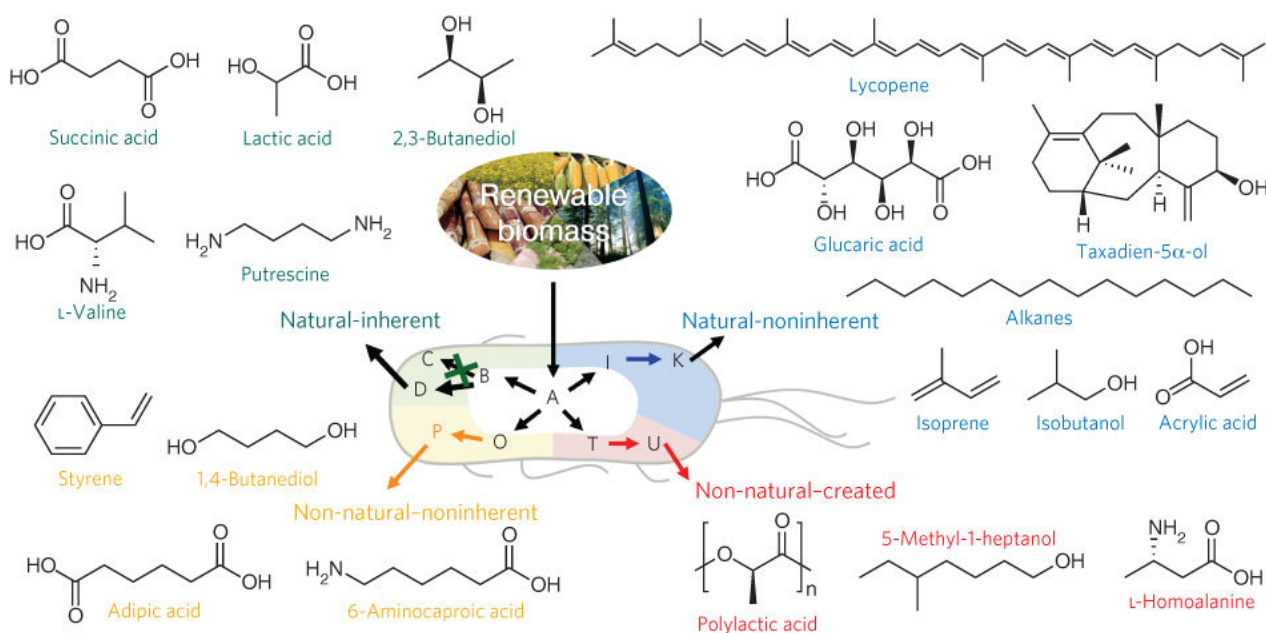


Fig. 1.2: Classification of compounds produced by metabolic engineering. This simple classification in four categories was proposed by Sang Yup Lee based on the degree of novelty that required the synthesis of a compound. *Natural* compounds are those already produced by unmodified organisms, whereas *non-natural* compounds are typically man-made. A producing pathway is said *inherent* if it did not required the addition of new enzymes. *A contrario*, if the production required the addition of heterologous enzymes (from other organisms) then it is said *non-inherent*. Pathways that required enzyme engineering are classified as *created* since they can neither be truly *inherent* or *non-inherent*. Importantly, enzymatic substrate promiscuity can be exploited to yield *non-natural noninherent* compounds. Extracted from [23].

when the first created activities are starting to get reported [26, 27] and protein engineering can be used to alleviate bottlenecks. For instance, despite being a natural opioid, the bioproduction of thebaine (a precursor of morphine) from the central metabolism of *Saccharomyces cerevisiae* required the impressive design of a pathway with 21 heterologous genes from plants and bacteria [28] (natural non-inherent), and even required some enzyme engineering to prevent unfortunate glycosylations. Another example is the bioproduction of 1,4-butanediol, a non-natural commodity chemical, that required the insertion of only 5 heterologous genes [29] in *E. coli* for its bioproduction (non-natural noninherent); the pathway was subsequently optimized by enzyme engineering and reached commercial viability.

Numerous proofs-of-concept exist in the literature but only few of them lead to industrial applications due to the tremendous research efforts that are needed to scale-up a promising pathway [30]. In this regard, the synthesis of artemisinic acid in yeast is a remarkable scale-up success story. Artemisinic acid is a precursor of a medicine to treat malaria that was previously extracted from an herb (*Artemisia annua*) at a prohibitive price for those who needed it the most. Its proof-of-concept synthesis was kick-started by a private grant of \$42 million, and it took 8 more years before the first competitive titers (from 100mg/L to

25g/L) were reported [31, 32]. Artemisinic acid is now produced at industrial scale by this very process [33]. Similarly, despite being close to a titer of 20g/L, the proof-of-concept synthesis of 1,4-butanediol (mentioned above) was still too low to compete with petrochemical processes [29]. The Genomatica company increased further their process up to 110g/L by a round of modifications including combinatorial mutagenesis on key enzymes to alleviate "enzyme-level bottlenecks" [34]. This was the first example of a non-natural commodity chemical produced industrially in a cell factory able to compete with petrochemical industry's processes. This project relied heavily on *in silico* predictive tools since its beginning, notably to predict promising pathways [29], which allowed the authors to invest time and financial resources on the scale-up that took 5 years of intense efforts [34].

It appears that, if the development of new cell factories is appealing to support the sustainability of our chemical industry, the road ahead is still long before we master the design of cells to fit our objectives. The development of new metabolic engineering projects is expected to fasten by embracing methods that have proven themselves in other engineering disciplines [10, 35]. Among those methods, the use of computer aided design tools (CAD) should continue to rise. In the next section, we will introduce some background information on *de novo* pathway design, focusing on rational methods supported by CAD softwares.

1.2 Computer-aided pathway design

1.2.1 Earliest attempts and motivation to predict enzymatic promiscuity

Early attempts to programmatically predict biological pathways relied on logic programming and inference engines to generate new routes between source and target metabolites [36–38]. Logic programming is indeed particularly well-suited to develop artificial intelligence solutions to problems that can be framed in list of constraints (*rules*, e.g. " $A + B \rightarrow C + D$ ", " A and B are needed to produce C and D ") and statements (*facts*, " A is endogenous"). This view of the metabolism (that ignores chemical structures and focus on stoichiometry information) gave birth around the same time to the first metabolic networks analysis methods based on linear optimization theory [39]. Constraints-based methods are still widely popular today, such as Flux Balance Analysis and prediction of isotope labeling experiments. Nonetheless, those methods generally assume that relationships between reactions and metabolites are fully known, which is hardly the case due to enzymatic promiscuity.

Enzymatic substrate promiscuity is the ability that have enzymes to perform the same catalytic activity on a variety of alternative substrates. Those chemicals usually share some degree of structural similarity; either by having the same overall aspect, sharing some groups, or some bonds [40]. The patchwork model of enzyme evolution [41] proposes that the latent catalytic plasticity of enzymes allowed organisms to gain versatile functions that eventually contributed to their adaptability and that were ultimately selected and optimized at a later stage [42, 43]. Enzymatic promiscuity is banal [40, 44] and already exploited to find alternative catalysts for metabolic engineering applications [45], and to fill gaps in metabolic models (underground metabolism) [46]. Importantly, promiscuous enzymes can reach high catalytic efficiency despite their low substrate specificity [47, 48]. Moreover, both the activity and the specificity can be optimized during directed-evolution experiments [49]. Nonetheless, it is difficult to get information about alternative substrates of an enzyme without going back to publications of enzymatic assays [50, 51]. Indeed, most enzyme databases record only the main reaction that has the most efficient catalytic activity among those that are known. Thus, it is necessary to model enzymatic promiscuity to assess the full potential of enzymatic reactions for pathways predictions.

Synthetic organic chemists grew an interest for biochemical pathways to predict the outcome of synthetic compounds once metabolized [52]; especially for medicines in biological tissue [53, 54], or pollutants in the environment [55]. The former has obvious applications to predict drugs' toxicity, but also in drug delivery in order to design drugs specifically so that they could be activated by enzymatic transformations (prodrugs) that occur only in some tissues [56]. Interestingly, since the chemicals used for those analyses are typically xenobiotics, they were absent from maps of metabolic networks and it was thus essential to model enzymatic promiscuity. The idea that a catalyst can perform the same transformation over a wide range of similar substrates is trivial for organic chemists. As a matter of facts, reactions are more often than not depicted with so-called Markush structures that allow side chains that do not take part in the reaction to be replaced by "R groups" that are used as a wild card. This led to the popularization in biology of generalist reaction rules that were already used to encode reactions in organic chemistry.

1.2.2 Structure-based generalist reaction rules

Reaction rules are designed to abstract (bio)chemical reactions to predict (1) if a catalyst would act on a given substrate(s), and if so, (2) what would be the resulting product(s). Reaction rules were originally defined by hard-coded heuristics in reaction prediction softwares [52, 57]. It made it difficult to update the rules to

take into account most recent knowledge about new reactions and exceptions to those that were already known [58]. Consequently, a great deal of efforts was made to create information systems that could easily be updated by human curators.

Those efforts eventually culminated with the adoption of graph-based reaction rules, most notably with the diversion of the SMARTS language that was originally developed to provide an in-line notation for substructures [59]. SMiles ARbitrary Target Specification language (SMARTS) is itself based on the Simplified Molecular Input Line Entry System (SMILES) that uses methods from graph theory to depict a molecule in a simple ASCII string [60]. The main advantage of encoding reaction rules in SMARTS is that it is a compact text data format that can easily be read and used both by humans and any cheminformatic toolkit. Reaction rules can be defined as a graph transformation between two subgraphs (SMARTS) encoding the changes occurring at the reaction center [61]: one SMARTS covering the substructure common to all known substrates, and another one covering the products. The former is used to filter only compounds that are believed to work with the catalyst and thus is responsible for the modeling of enzymatic promiscuity; while the latter provides the result of the transformation operated by the reaction. Importantly, the Atom-Atom Mapping (AAM) between the two SMARTS is mandatory to encode the actual transformation between the graph of the substrate(s) and the graph of the product(s). Reaction SMARTS (also called SMIRKS) were eventually used by softwares to predict metabolic reactions [55, 62–64] (see also Chapter 3). Others preferred to stay away from this language while adopting the exact same idea of encoding a transformation by two mapped subgraphs with in-house data format based on Bond Electron Matrix (BEM) [65], Reaction-center Difference Match (RDM) patterns [66, 67], or other [68].

Both finding the Atom-Atom Mapping (AAM) and applying a transformation on a graph are related to the difficult Maximum Common Substructure (MCS) problem, itself related to the graph isomorphism problem which is NP-hard and thus computationally intensive [69]. Moreover, AAM is very sensitive to the quality of input data that vary greatly between databases in practice (unequilibrated reactions, missing chemical structure, redundancies). Those limitations motivated the development of another class of structure-based reaction rules based on vectors of chemical features that do not require any AAM or graph handling. This approach was pioneered in biology by the work of Faulon and Carbonell on "reaction signatures" (see Chapter 2) [70, 71] and followed by others with subtle variations [72, 73]. Those methods have in common to encode compounds in a vectors where each bit refers to the occurrence of overlapping substructures in the compound (either paths or subgraphs, sometimes called atom environments or fragments), much like the Extended Connectivity FingerPrint (ECFP) [74] that

are widely popular to perform compounds similarity calculation. From that definition we can see a reaction as a simple difference between the products and the substrates vectors. The resulting vector summarizes the net changes occurring during the reaction in term of chemical fragments; it is the reaction rule. A simple addition on the vector associated to a putative substrate will give the corresponding putative product that can then be found in database or regenerated from the fragments. We will see that in more details in Chapter 2. Ultimately, those vectors were used in yet another class of reaction prediction softwares that use machine learning in place of rules [75].

All reaction rules do not encode enzymatic promiscuity to the same extent, some being more generalist than others. For instance, BNICE models its rules after the Enzyme Commission (EC) classification [76] that sorts enzymatic reactions by the chemistry involved (e.g. oxidoreductases, transferase, etc.) and are typically very generalists [65]. RetroPath adopts a more data-driven approach by automatically generating rules from reaction databases and brings the additional feature to enable users to choose the degree of promiscuity that should be allowed [70, 77]. This has a critical importance during the generation of metabolic spaces since more generalist rules will produce more (putative) products that would in-turn need to be evaluated in an iterative process to generate the metabolic space. One can clearly see that this leads to a combinatorial explosion that needs to be controlled by search algorithms or dealt with by throwing more computational power at it.

1.2.3 From metabolic space generation to *de novo* pathways design

Reaction rules can be used iteratively on known endogenous metabolites to reconstruct the metabolic network of specific organisms, or even the combined metabolism of all organisms. As we have seen above, reaction rules can model enzymatic promiscuity to palliate the lack of information we have on enzymes' alternative substrates and unveil metabolism latent capabilities. This has applications to identify compounds from metabolome data (mass spectrometry) [78, 79], fill gaps in metabolic models [46] and of course for the design of *de novo* metabolic pathways [80]. Nonetheless, pathway design is not limited to the ability to predict reactions in a metabolic space.

We must keep in mind that a chassis is not a blank state but already an inherently complex system that will get perturbed by the implementation of new reactions. Since this will eventually impact the pathway efficiency it should be taken into account during the pathway design, and computed-aided design tools should

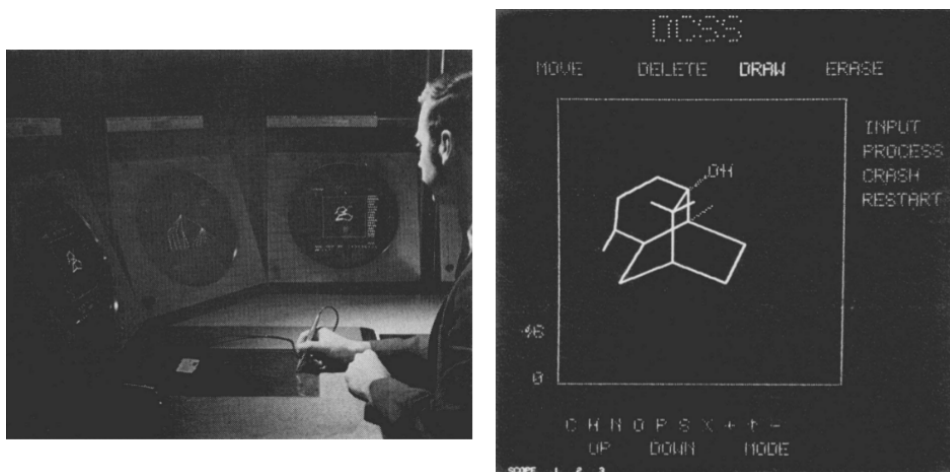


Fig. 1.3: Interface of the first computer-aided design software dedicated to retrosynthesis. This software was developed back in 1969 for synthetic organic chemists by Corey and Wipke. Natural reactions catalyzed by enzymes were outside its scope. Left: a user draws the compound he wishes to synthesize. Right: input's graphical user interface. Extracted from [57].

provide features in that direction. For instance, numerous CAD tools use the net Gibbs free energy of putative pathways to assess their spontaneity and thus their driving force [64, 65, 81]. Systems-level methods to predict maximum theoretical yield are also quite often integrated into pathway prediction softwares [62, 82–84]. Others tools take into account important parameters at the metabolite-level influencing pathway and that are more difficult to estimate such as the toxicity of intermediate metabolites [81]. We will not enter into the details of those features but simply note that they are exploited to score predicted pathways and sort out the most promising before the eyes of the user.

This highlights an important aspect of pathways CAD softwares that is that their purpose is first and foremost to empower users to make pertinent design choices, contrary to the notion that CAD software should provide definitive plug-and-play answers. In that context, it appears that tools proposing to design novel pathways between one source and one target compound may very well hinder users' creativity by enforcing too soon strong constraints (e.g. carbon source). This becomes clearly a liability considering the fact that reaction databases are incomplete and that missing a reaction can be the cause for missing a whole pathway toward a target metabolite.

1.2.4 Retrosynthesis for an efficient navigation into the metabolic space

Retrosynthesis is a rational method to find non-intuitive production routes toward chemicals of interest that was first developed by organic and synthetic

chemists. The Father of retrosynthesis analysis certainly is E. J. Corey as he devoted his life's work to rationalize organic synthetic chemistry and received in 1990 the Nobel Prize in Chemistry for the "development of the theory and methodology of organic synthesis" [85]. In his Nobel lecture [86], Corey defines retrosynthesis analysis as: "a problem-solving technique for transforming the structure of a synthetic target molecule to a sequence of progressively simpler structures along a pathway which ultimately leads to simple or commercially available starting materials for chemical synthesis". Thus, retrosynthesis is an iterative backward process that starts from the target molecule and then explores the chemical space until a suitable route toward a starting material is found. At each step, reversed reactions determine the allowed next moves in the chemical space, i.e. how to transform one compound into another. The chemical space exploration is thus fueled by our understanding of chemical reactions.

The beauty of retrosynthesis is that this rather simple idea requires a massive amount of expert knowledge to be successfully executed. This makes it a difficult task even nowadays with the help of computers, not because the computational power would be lacking, but because the underlying phenomena are difficult to predict and many data sources need to be aggregated together to summarize the extent of current knowledge [58, 86, 87].

Naturally, many attempts have been made to automate retrosynthesis analysis with computer-aided design (CAD) tools. The purpose of such tools, as with any other CAD software, is to lower the amount of expert knowledge needed by the user and ultimately to accelerate the development of new applications [35, 88]. As one might expect, the first attempts to leverage computational power for retrosynthesis (in synthetic organic chemistry) were pioneered by Corey himself in the late sixties [57] (see Figure 1.3). Interestingly, the challenges associated to the development of a retrosynthesis CAD software were already the same as the ones modern solutions are still facing [58, 89]; namely, the design of an intuitive graphical user interface, the need for interconvertible chemical data format, a way to deal reasonably with combinatorial complexity, and of course the modeling of the chemical reactions themselves.

The concept of retrosynthesis analysis can be applied to biology and metabolic engineering to find synthesis routes to be implemented in living organisms as heterologous enzymatic pathways. The term of bioretrosynthesis (or retrobiosynthesis) is used to emphasize that the retrosynthesis analysis takes place in a metabolic engineering context and thus inherits from its advantages and limitations. Obviously, available reactions are not the same for designing pathways in prokaryotes, eukaryotes or *in vitro*; starting compounds are different too, since compounds supplemented to the media would be subject to membrane perme-

ability, etc. As of today, a great number of tools have been developed both by research groups and companies to perform retrosynthesis analysis, and a large portion of them are specialized in bioretrosynthesis [62, 68, 70, 81, 90–92].

We have seen that predicting enzymatic pathways is a complex task that requires a massive amount of data and efficient algorithms to navigate through the metabolic space. There is no reason to believe that metabolic and reaction databases will ever be complete for information as difficult to collect as enzymes' alternative substrates. This highlights the importance of reaction prediction and in particular the importance of modeling enzymatic promiscuity that is at the core of *de novo* pathway design. However, computer-aided pathway design tools are not limited to the search of pathways for bioproduction applications, and encouraging the development of new applications for reaction prediction frameworks might be another path to contribute significantly to the advances of metabolic engineering. We will exemplify this point in the second part of this thesis with the description of a strategy to design metabolic circuits for the biodetection of small chemicals. In the next section we will quickly introduce the interest of intracellular biosensors focusing our efforts on metabolic engineering applications.

1.3 Intracellular biosensors

1.3.1 Microbial systems for the detection of small chemicals

All living systems need to monitor the fluctuations of their environment to better fit the associated constraints by adapting their behavior and/or metabolism. For instance, it is well-known that *E. coli* has a strong taste for D-glucose over L-lactose, itself being preferred to D-xylose and other sugars as a carbon source [93]. *E. coli* follows this utilization hierarchy to consume all of preferred sugar before switching to second-choice in an "all-or-none" response. This mechanism is orchestrated by a tight regulation of sugar assimilation pathways that was fine-tuned by the evolution; some sugars are indeed easier to assimilate thus providing a competitive advantage. Consequently, microbes have the ability to sense their carbon source and react by adapting their metabolism. The same could be said for the detection of toxic compounds that triggers microbial countermeasures such as the expression of transporters, or the detection of the level of critical compounds such as cofactors to maintain their homeostasis within the cells. Microbes have developed a variety of molecular mechanisms dedicated to

the detection of small chemicals for the purpose of triggering a response at the genetic level.

Allosteric transcription factors (aTF) are proteins that have the property to change their affinity to DNA upon the transitional binding of an effector (detectable compound) through an allosteric conformational shift. This shift can either promote or decrease DNA-affinity on a specific motif (operator domain) that is part of the promoter of the regulated operon. aTFs act on gene transcription either by blocking the access of the RNA polymerase (repressors), or on the contrary by stabilizing its binding (activators). For instance, AraC and XylR are both (de)repressors aTFs that bind to their respective DNA operator in absence of their effector, effectively preventing gene expression. In presence of their effector (arabinose and xylose), AraC and XylR adopt a DNA-free conformation that allows gene expression. Transcription factors are the most recurrent regulation system for bacteria, but other systems exist.

Riboswitches are RNA sequences in *cis* of an mRNA's CDS able to regulate the associated gene expression either at the transcription or translation level by a variety of mechanisms, most notably by premature transcription termination or by preventing translation initiation [94]. Riboswitches are considered by some to be the first elements dedicated to regulation that appeared, presumably during the RNA World [95]. Simple riboswitches have one expression platform able to undergo a conformational switch triggering a gene expression regulation mechanism upon the binding of a chemical on a dedicated aptamer domain. This simple architecture and the fact that we can efficiently find aptamer domain *in vitro* using high-throughput techniques nourished tremendous research efforts to give them new detection capabilities [96]. However, results for *in vivo* applications are still limited so far due to their detection limit that often falls in a toxic range for living cells [97].

Two-component signal transduction systems have a membrane-bound histidine kinase able to induce the autophosphorylation of a cytoplasmic response regulator upon the extracellular detection of a chemical. The regulator can then bind DNA and achieve its purpose. While the particular feature of extracellular detection can be interesting for compounds that cannot cross the membrane, two-component systems are subject to a risk of cross-talk that limit their range of applications [98].

We will see in more details how to find natural biosensors in Chapter 4. For now, we will quickly acknowledge the existence of other kind of biosensors coming from other disciplines.

1.3.2 Other biosensors

The term of "biosensor" is widely used in several communities to describe different concepts, which can make it confusing. In biology, it usually refers to a genetically encoded system able to detect chemicals, as we described in the previous section. Importantly, biosensors are not limited to the detection of small organic compounds and can also detect physical and biological states (e.g. oxidative stress, light, DNA damage, etc.) [99]. The term is also used for applications that aims at the detection of pathogenic bacteria by PCR directed against specific DNA loci [100].

In the nanotech world, the term of biosensors is used for any sensor able to detect biological material. Those sensors typically use an aptamer (DNA or RNA) bound onto a surface (microchips, quartz, nanoparticles) and an apparatus able to monitor mass loading by a change of physical properties (colorimetry, fluorescence, conductivity, resonance, etc.). For instance, a biosensors for cocaine has been reported using an anti-cocaine DNA aptamer bound onto a quartz surface monitored by an electromagnetic piezoelectric acoustic sensor [101].

Coming from physical chemistry, Förster Resonance Energy Transfer (FRET) phenomenon states that two chromophores that are physically close enough and with an overlapping absorption and emission spectra will transfer the absorbed energy of one chromophore (donor) to another (acceptor), so that the other would yield the energy in its own emission spectra. This can be exploited to make biosensors by functionalizing the chromophores on an aptamer so that their distance would be dependent of the binding of a chemical, and so would be the FRET emission [102]. Interestingly, they can be genetically encoded when they use fluorescent proteins [103, 104].

Finally, let us note that not all biosensors have a switch-like behavior and that some are specifically designed to have a one-time action. This is the case of so-called hammerhead ribozyme, a riboswitch that has the property to autocleave in presence of its effector [105].

In this thesis, we will use the term of biosensor exclusively to refer to genetically-encoded biological systems able to detect small organic compounds, like those that we described in the previous subsection, with a special emphasis on allosteric transcription factors (aTF).

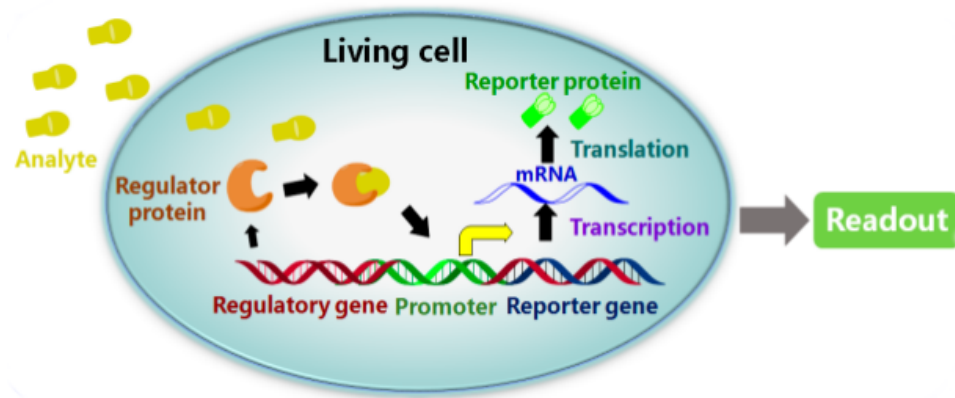


Fig. 1.4: Concept of whole-cell biosensor. The whole cell act as a detection device to capt the signal (presence of the analyte, i.e. a compound) and transform it into a measurable output, such as fluorescent proteins. Extracted from [106].

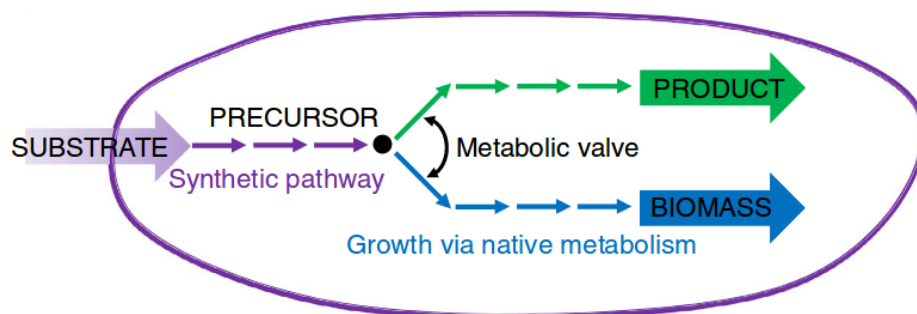


Fig. 1.5: Metabolic valve to balance between growth and production objectives. Biosensors have been proposed to act on key points of the metabolism to redirect cells resources from growth to the production objective. Extracted from [107].

1.3.3 The need for biosensors in metabolic engineering

The usage of biosensors in synthetic biology and metabolic engineering applications is more and more prevalent and concentrated around three main applications [108–111]:

- the design of whole-cell biosensors,
- dynamic regulation of heterologous pathways,
- high-throughput screening, eventually to achieve biosensor-driven adaptive evolution.

Whole-cell biosensors are cells, usually microbes, dedicated to the detection and report of stimuli thanks to a molecular biosensing mechanism [112] and a

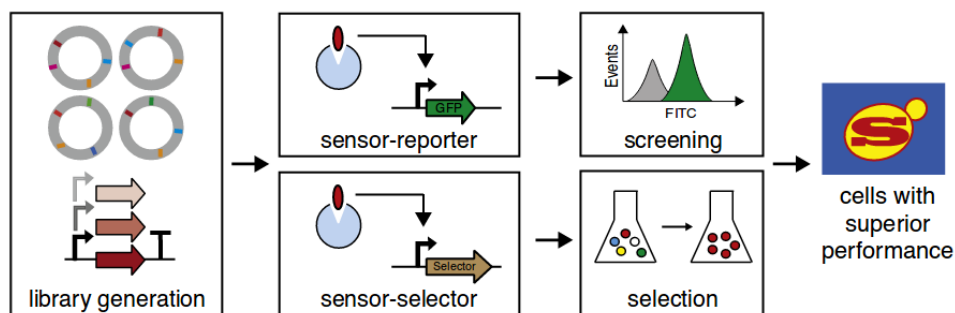


Fig. 1.6: Selection of superproducers among a library of variants with biosensors. Genetically encoded biosensors (such as allosteric transcription factors) can be used with a reporter system to screen (top, fluorescent proteins) or select (bottom, toxin/antitoxin systems) the right variants among a population of cells. Extracted from [108].

reporter system [113] (see Figure 1.4). They are often proposed to report environmental pollutant and biomarkers in medical analysis applications [106, 114, 115]. Their main advantage against other detection methods is that they directly evaluate the bioavailability of tested compounds (if the biosensing mechanism is intracellular); they are also renewable, cheap to produce, auto-replicable, and can be functionalized to make portable devices for real-time detection of chemicals. Interestingly, whole-cell biosensors can easily be used as a platform for the detection of multiple compounds of interest, either by using together strains dedicated to the detection of one compound (in consortia or as independent tests) [116], by implementing several responsive elements in the same strain, or using a promiscuous responsive element [117]. However, whole-cell biosensors still lack reproducibility and stability during long-term storage and transport [106], which limits their usage for real-life applications. Their use on the field is also constrained by the legislation related to genetically modified organisms.

Biosensors have been successfully exploited to regulate the flux of heterologous pathways so that they would not consume cofactors too fast and impair cells viability in the process. For instance, Zhang *et al.* reported a pathway toward biodiesels in which the expression of key enzymes is controlled by acyl-CoA availability (through fatty acids availability) in a negative-feedback fashion [118]. They reported a 3 fold increase of yield at 28% of the theoretical maximum, a steadier production, a diminution of the accumulation of byproducts (acetic acid), and a better plasmid integrity for the regulated system. The same idea was tested with manonyl-CoA for the production of flavonoïds [119] and fatty acids [120, 121]. Dahl *et al.* adopted a similar but more systemic approach to minimize the impact of toxic intermediates in an amorphadiene producing pathway by using the promoter of host's native stress response system to regulate pathway's flux [122].

Another strategy to improve yields with dynamic regulation is to use biosensors to balance the utilization of the carbon source toward biomass or the production of interest with so-called "metabolic valves" [123] (see Figure 1.5). Indeed, fast population growth and fast production are both desirable traits for cells in a bioreactor, but those objectives compete for the use of nutriment. An elegant solution is to decouple growth and production and trigger the latter only once the optimal cell density is reached [107]. Tan *et al.* implemented this idea by putting key enzymes for glucose assimilation toward biomass under the control of an inducible promoter. Once cells are supplemented with the relevant effector, it blocks their growth and redirect carbon flux toward the bioproduction [124]. Others prefer to use quorum sensing signals to automatically start the bioproduction of interest once a satisfactory biomass is reached in the bioreactor [125].

Biosensors also have a central place in the development and optimization of new bioproducing strains [126]. Indeed, not all enzymatic pathways have a colored product or precursor that allows the utilization of colorimetric assays to assess their efficiency and select the best producers, like it is the case for instance for isoprenoid and carotenoids [127]. Classical detection methods (HPLC, mass spectrometry) can always be used, but they do not work at the cell-level which is a desirable trait for the high-throughput screening of libraries of variants. To tackle that issue, genetically encoded intracellular biosensors able to detect a chemical of interest have successfully been used to report the production efficiency by linking it to cells' growth [128] or the expression of a fluorescent reporter [129] (see Figure 1.6). Modern fluorescence-activated cell-sorting (FACS) methods are capable of ultra-high-throughput screening (HTS) in the order of 10^8 cells by day. That makes them particularly attractive to sort large libraries of variants from gene mining approaches [130], or directed evolution experiments [131]. For instance, Binder *et al.* used an aTF for the HTS of a mutant library to pick best lysine producers and gain insights on the mutations that were needed at the chromosomal level to improve the production [129]. HTS used conjointly with a relevant biosensor can also be used to improve specific enzymes [132] and pathways [133].

We have seen that a multitude of biosensors exists, both natural and synthetic, to detect chemicals using a large palette of molecular mechanisms. The need for biosensors is palpable in synthetic biology and metabolic engineering applications, and it is not always easy to find a ready-made sensor for one's application as we will see in Chapter 4. Hopefully, this quick introduction on biosensors and what we already wrote about CAD reaction prediction softwares and metabolic engineering advances paves the way to a clear understanding of the results presented in this thesis.

1.4 Thesis structure

Part I – Retrosynthesis for metabolic engineering

The first part of this thesis is dedicated to the retrosynthesis tools developed in the team. Their main function is the prediction of metabolic pathways for the bioproduction of chemicals in microbes.

With Chapter 2, we will see in details the first version of the retrosynthesis algorithm that was developed prior to my arrival in the team (RetroPath). Its algorithm was used as a basis for the SensiPath project described in Chapter 5.

With Chapter 3, we will see our latest implementation of RetroPath in a version that has the originality to be embedded in a third-party workflow management software (KNIME) to allow for an easier access to retrosynthesis and other related tasks for the metabolic engineering community.

Part II – Computer-aided design of sensing-enabling metabolic pathways

The second part of this thesis exemplifies the use of reaction prediction frameworks for other purpose than predicting bioproduction pathways. In our case, it is the design of sensing-enabling metabolic pathways (SEMPs); short metabolic circuits able to detect and report through enzymatic reactions and allosteric transcription factors the availability of intracellular compounds. SEMPs are particularly appealing when no direct biosensors are known for the compounds of interest.

With Chapter 4, we will review the strategies that can be leveraged to detect specific chemicals within bacteria's cytosol and trigger a genetic response.

With Chapter 5, we will present SensiPath, a webserver for the *in silico* computer-aided design of SEMPs. SensiPath is powered by a brand new implementation of the algorithm presented in Chapter 2 and backed up with a database compounds known to be detectable by allosteric transcription factors.

Finally, with Chapter 6, we will see the validation of SensiPath's algorithm with the *in vivo* implementation of several new whole-cell biosensors.

Part I

Retrosynthesis for metabolic
engineering

Extended Metabolic Space Modelling

This work was published in Springer Protocols, Methods in Molecular Biology for a volume on Synthetic Metabolic Pathways by Pablo Carbonell, Baudoin Delépine, and Jean-Loup Faulon.

Only minor modifications have been introduced in the chapter below.

Full reference

Carbonell, P., Delépine, B., Faulon, J.-L. (2017). "Extended metabolic space modeling". In: Michael Krogh Jensen, Jay D. Keasling, Synthetic metabolic pathways. Methods and protocols. *Methods in Molecular Biology*, 1671. Springer.

Contribution

The methods discussed in this publication were first described in [71] and [81]. P.C. lead the project and B.D. assisted him in the preparation of the manuscript.

2.1 Abstract

Determining the fraction of the chemical space that can be processed *in vivo* by using natural and synthetic biology devices is crucial for the development of advanced synthetic biology devices. The extended metabolic space is a coding system based on molecular signatures that enables the derivation of reaction rules for metabolic reactions and the enumeration of all possible substrates and products corresponding to the rules. The extended metabolic space expands capabilities for controlling the production, processing, sensing and the release of specific molecules in chassis organisms.

2.2 Introduction

The set of chemical compounds that organisms can process and synthesize is finite. Such finite set, however, is not fully known yet. Based on a model that accounts for versatility of enzymatic reactions, we describe here a computational protocol to estimate the extent of such full metabolic space. The extended metabolic space can be screened to list any possible biological circuit that can be conceived, such as the ones that are used to produce, detect and process chemicals.

To fully exploit the metabolic space, an essential requirement is having a thorough knowledge of the metabolome associated with any given organism. However, experimental evidences from metabolomics analyses often show that with currently known metabolites one cannot cover the ranges of masses found in actual samples, and consequently there is an impelling need of completing the metabolomes and reactomes of interest for metabolic design [134, 135]. Furthermore, the metabolic phenotype of an organism may vary upon different conditions such as during different growth states leading to variations in the metabolite profile [136]. Besides such sources of uncertainty in samples, many unassigned peaks should be due to promiscuous activities of enzymes not yet characterized because of the lack of an appropriate description of the mechanisms of enzyme promiscuity.

Our group has addressed the issue of complexity by proposing a tradeoff solution based on molecular signatures [137]. Our molecular signature codes for changes in atom bonding environments where the reaction is taking place. The advantage of the signature method is that the reaction rules describe the changes in the environments of the atoms belonging to the catalytic center of the reactions, and the size of the environment (named diameter) can be tuned to control the combinatorial explosion of possible compounds. Moreover, reaction signatures are robust to unbalanced reactions and can be created automatically without the need of any atom-atom mapping. The signature representation has shown itself to be specially well-suited for modeling the mechanisms of enzyme promiscuity [138], paving by these means the way towards engineering innovation in metabolic networks. Either through directed evolution [139] or random selection [140], latent capabilities present in enzymes as modeled by the extended metabolic space can be potentially enhanced in order to optimize the desired activity and eventually implemented as a biological part containing a metabolic circuit.

Here, we describe the necessary steps in order to generate an extended metabolic space and how to compute all viable routes within the extended space that

determine a viable pathway connecting a desired target to the chassis organism (Fig. 2.1).

2.3 Materials

Materials for the described computational protocols consist basically of datasets obtained from public databases and processing software.

- A metabolic database of reference covering chemical structures and reactions. Metanetx [141] is a consensus database that reconciliates multiple databases.
- Models of metabolism for chassis organisms. Biomodels [142] and BiGG [135], among others, are databases containing genome-scale models for most commonly used organisms.
- Software to compute molecular signatures, which are specialized type of topological chemical descriptors. Molsig [137], among others, is an open-source package that provides such capabilities.
- Matrix manipulation software such as octave, matlab, scipy, R, etc.
- Computation of elementary modes. Efntool [143] provides both a Java and matlab-based efficient implementations.
- Software for chemical manipulation. Some of the most popular implementations are RDKit, Marvin, CDK, KNIME (Table 2.1).

2.4 Methods

2.4.1 Computation of molecular signatures

The first step to generate an extended metabolic space is to encode all compounds of a metabolic database in a format that will allow the subsequent encoding of enzymatic reactions. We propose here to showcase the important steps that should be kept in mind through the use of one of the available encoding methods, the molecular signature [137] (see Note 1).

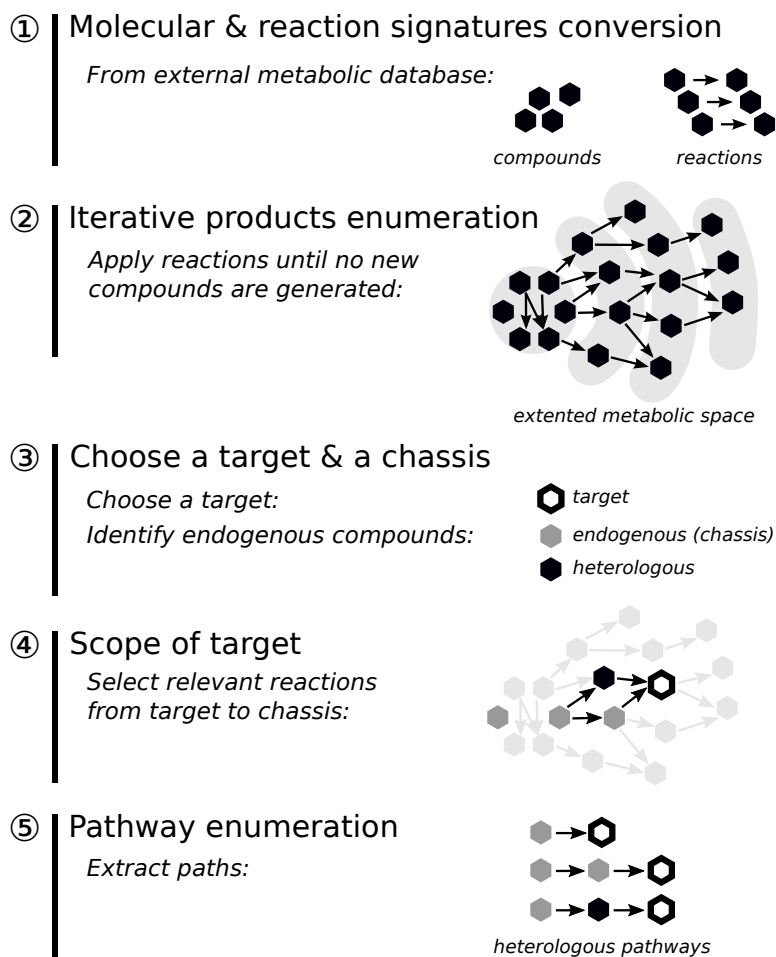


Fig. 2.1: Steps involved in the construction of the extended metabolic space. The 1st step consists of converting compounds and reactions into molecular signatures. The 2nd step enumerates new products by an iterative algorithm applied to the reaction signatures. The 3rd step consists on choosing a target, i.e., a reaction or a compound, and a chassis organism. The 4th step determines the metabolic scope linking the chassis to the target. Finally, the 5th step enumerates all viable pathways connecting the chassis to the target.

Tab. 2.1: A selection of software tools for modeling in the extended metabolic space.

Name	Keyword	Comment
Stand-alone softwares		
Cytoscape	Graph visualization	Cytoscape can be used to manually explore and visualize the EMS [144]
efmtool	Elementary flux	Computation of elementary flux modes [143]
KNIME	Workflow	Knime propose to create automatic processes (“workflow”) through a drag-n-drop interface of small tasks (“node”). It is useful for reproducibility of data analysis [145]
MarvinSketch	Chemical editor	ChemAxon’s chemical editor. Useful to visualize compounds and reactions, especially to manually inspect difficult cases. URL: http://www.chemaxon.com
MolSig	Molecular signatures	Compute molecular signatures from MDL MolFile. URL: http://molsig.sourceforge.net/
Python libraries		
COBRApy	Constraints-based models	A constraint-based steady-state simulation analysis for genome-scale models [146]
NetworkX	Graph exploration	NetworkX has an intuitive interface and an extensive documentation. It is a good solution to handle the conversion of the EMS into standard graphs format, or to programmatically explore the EMS.
RDKit	Chemoinformatic toolbox	RDKit makes it very easy to handle chemical structures, especially to standardize compounds.

1. Initially gather compounds from a metabolic database. This database must have structural data for compounds and reactions, and ideally be linked to a whole-cell model (see Note 2).
2. Check compounds for incomplete structural data. Some compounds can be defined with incomplete Markush structure or wildcard atoms. Those compounds typically stand to define classes of compounds (e.g. "an alcohol") and should be removed since they cannot be interpreted through the molecular signatures algorithm used in this protocol.
3. Standardize compounds. Molecular signatures encode directly molecular graphs from a MDL MolFile input. Users must ensure that compounds (resp., chemical groups) that should be considered identical have the same molecular graph (resp., sub-graph) (see Note 3).
 - a) Neutralize or remove charges. As much as possible, chemical groups should be represented with the same protonation state to prevent different tautomeric forms. One can either use heuristics to add or remove hydrogen when necessary or simply remove all charges from the compound dataset.
 - b) Choose one conjugated form by compound. This is particularly important for aromatic compounds, which could appear under different kekulé forms in the database. A good solution is to explicitly use aromatic bonds in the molecular graph description.
 - c) Use a consistent hydrogen representation, either implicit or explicit.
4. To compute the signature of a chemical compound, we need initially to consider its molecular graph. Let $G(V, E)$ be the molecular graph associated with some chemical compound C and let $a \in V$ ($b \in B$) be an atom (bond) of G . The atomic signature of atom a of diameter d , $^d\sigma(a)$, is a canonical representation of the sub-graph of G spanned by its vertices at a maximum distance of $d/2$ from a . From a chemical point of view, this corresponds to a circular fragment of the compound centered on d .
5. The molecular signature of a molecular graph G of diameter d associated to C , $^d\sigma(a)$, is defined as the list of all atomic signatures of diameter d (one by atom). Therefore, a molecular signature is a list of overlapping molecular fragments.

6. Depending on the diameter d , a molecular fingerprint can show degeneracy, i.e. a same molecular signature can represent more than one molecular graph G , much like a chemical formula can correspond to several compounds.
7. Based on previous definitions, the computation of the molecular signature involves two steps:
 - a) Choose a diameter to encode enzymatic promiscuity. To some extent, enzymes have the ability to process additional reactants that are structurally similar to the known ones. In a context where it is important to maximize the number of reactions in order to get more leads, modeling promiscuity can reveal itself to be a critical feature (see Note 4). We recommend starting with a diameter of 12 and to go lower (down to 4) if no satisfying solution can be found.
 - b) Compute molecular signatures. The MolSig software [137] computes molecular signatures starting with compounds in MDL MolFiles format, which can be easily retrieved from metabolic and chemical databases or converted from other equivalent formats (see Note 5).

2.4.2 Computation of reaction signature

The step following the encoding of compounds is the encoding of reactions into reaction signatures. Reaction signatures should be understood as an exchange of fragments. Unlike other reaction models, reaction signatures do not need any atom-atom mapping to be computed, nor do they need reaction to be balanced (see Note 6).

1. Let R be a reaction for which all substrates $\{S_i, i \in [1, n]\}$ and products $\{P_j, j \in [1, m]\}$ are encoded in molecular signatures, respectively $\{^d\sigma(S_i), i \in [1, n]\}$ and $\{^d\sigma(P_j), j \in [1, m]\}$. The reaction signature is defined as follows (see Note 7):

$$^d\sigma(R) = \bigcup_{j=1}^m ^d\sigma(P_j) - \bigcup_{i=1}^n ^d\sigma(S_i)$$

2. Thus, $^d\sigma(R)$ is the difference in term of atomic signatures (i.e. molecular fragments) occurring during a reaction; created (resp. consumed or needed) fragments being positives (resp. negatives). In this context, the diameter d corresponds to the reacting moieties and their neighboring

- a) if the signatures of candidate substrates include all fragments consumed by R :

$$\{x \in {}^d\sigma(R), x < 0\} \subseteq \bigcup_{i=1}^n {}^d\sigma(S'_i)$$

- b) and the predicted product(s) signature(s) ${}^d\sigma(P')$ correspond to some previously known compound(s) in DB, with (see Note 8):

$${}^d\sigma(P') = \bigcup_{i=1}^n {}^d\sigma(S'_i) + {}^d\sigma(R)$$

4. Being able to model enzymatic promiscuity assumes that reaction signatures can be used with other substrates than the ones in the native reaction. In turn, alternative substrates produce new products. Those compounds may be absent from the metabolic space, i.e., the set of known metabolites. Therefore reaction signatures extend the metabolic space by linking potentially new compounds to the metabolism (see Note 9).

2.4.4 Chassis modeling in the extended metabolic space

In the previous sections we have described the protocol that allows extending the metabolic space. When the extension is applied to a metabolic network consisting of all known metabolic reactions, we arrive at the full description of all available metabolic capabilities. Some of these capabilities are going to be common to several groups of organisms, such as reactions in the central metabolism, while others like secondary metabolism will be specific to some groups. In applications such a biotechnology, the organism that is engineered is known as the chassis organism and often the objective will be to expand the natural capabilities of the chassis by introducing heterologous enzymes. In this section we will describe how to model the chassis organism as a subset of the extended metabolic space.

1. The extended metabolic space of diameter d , denoted by M_d , represents all the possible compounds C and allowed transformations (reactions) R between compound as spanned by the enumerated reactions computed by following the method described.
2. A chassis is a subset of the extended metabolic space $O_d \subset M_d$ that corresponds to the extended metabolic network of an organism at signature

diameter d . A chassis is defined by the set of nominal reactions annotated for the enzymes present in the organism.

3. The list of nominal metabolic reactions for a given organism can be compiled from databases such as KEGG [147], MetaCyc [148], BiGG [135], BRENDA [51], etc. The choice of one database over the others depend on several factors:
 - a) The degree of curation of the model;
 - b) The free and open availability of the model;
 - c) The way the model is going to be analyzed, i.e., network analysis, steady-state simulation or simply as a reference list of metabolites and reactions (see Note 10).
4. *In silico* organism models showing a good degree of accuracy and reproducibility are currently available for many industrial strains, including *Escherichia coli*, *Saccharomyces cerevisiae* or *Bacillus subtilis*. They can be generally downloaded in SBML format [149].
5. In order to determine O_d , each reaction in the reference model is augmented with the set of enumerated reactions of the chassis in the extended metabolic space, resulting in an extended model (see Note 11).

2.4.5 Computing the scope

The next step in modeling in the extended metabolic space is to have an understanding of the design space for a given target metabolic activity. In other words, we want to compute the metabolic scope connecting some target reaction to the chassis. To that end, we provide in this section some relevant definitions and a two-step procedure that allows the determination of the metabolic scope.

1. A minimal pathway is defined as any set of reactions connecting the chassis to the target that are minimal:
 - a) They form a viable production pathway in terms of precursors availability;
 - b) All reactions are essential, i.e., the removal of any reaction renders non-viable the pathway (see Note 12).

2. Based on that definition, the metabolic scope is defined as follows: given an initial set of source metabolites S (the chassis) and a final set of target metabolites T , the scope is the set of enzymes that are at least involved in one minimal pathway connecting elements of T to the source S , i.e., the scope should contain only enzymes that are at least essential for establishing one of the metabolic pathways. In order to compute the scope for a given compound, a two-step procedure can be applied, as described in the following.
3. Reduction of the extended metabolic space to the reachable space of reactions. It consists of the following steps:
 - a) A compound is defined as reachable if there exists a reachable reaction that can produce it, i.e., a reaction for which all substrates are available.
 - b) Start from the set of initial compounds S and iteratively find newly reachable compounds.
 - c) The process stops when no new reachable compounds are found.
 - d) Build a graph to keep track of which reactions produced each compound.
4. Backward determination of the scope. It consists of the following steps:
 - a) Start from the target compound(s) T . For each reaction that can produce the target compound(s), add it to the scope.
 - b) Recursively apply the same procedure on each substrate of the reaction.
 - c) The recursion stops when initial compounds S are reached.

2.4.6 Enumerating pathways

Once the extended metabolic scope has been determined, we should be interested in enumerating all viable metabolic pathways connecting the source to the target. This turns out to be a computationally complex problem that can be solved through several approaches [150]. We describe here a solution based on the computation of elementary flux modes [151] (see Note 13). EFMs are the set

of minimal pathways that are non-trivial solutions to the steady-state equation whose combination can describe any possible path in the network (see Note 14).

1. Define the augmented metabolic space formed by the union of the reactions in the chassis and in the scope (Fig. 2.3a).
2. Construct a stoichiometric matrix where each row corresponds to a compound and each column to a reaction of the previous augmented metabolic spaces and the value of each cell is the stoichiometric coefficient (Fig. 2.3b).
3. Remove all rows representing initial compounds (see Note 15).
4. Remove all rows representing compounds that are produced by a reaction but never used in any other.
5. Merge identical columns by deleting redundant columns and renaming the remaining column with the names of all reactions (see Note 16).
6. Add an additional column to create a flux out for the target compound.
7. Several toolboxes exist that allow efficiently computing the elementary modes (Fig. 2.3c). For instance `efmtool` [143] provides an efficient implementation that can either run in matlab or in Java.
8. Expand resulting elementary modes into the pathway solutions by enumerating all combinations of merged reactions in each elementary mode (Fig. 2.3d).

2.4.7 Design in the extended metabolic space

We have described in previous sections step-by-step methods that generate extended metabolic spaces for a) global metabolic capabilities; b) chassis organisms; c) organisms augmented with desired target activities. From here, resulting extended models can be used in multiple engineering biology applications, from production of chemicals to their sensing and regulation. Some of the main applications developed to date in extended metabolic spaces include the following:

1. Engineering of heterologous pathways for the production of a desired chemical in a chassis organism. In order to select enzyme sequences for each enzymatic step in the pathway for the most promising routes in the

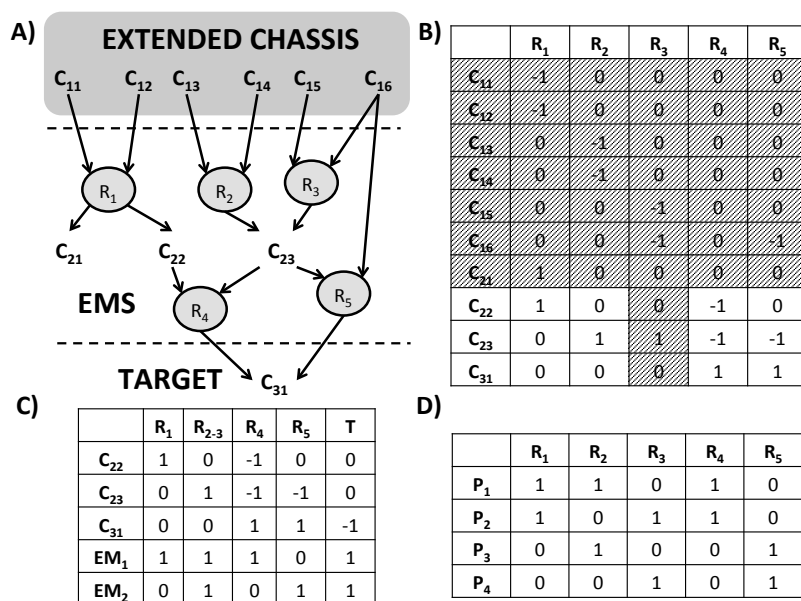


Fig. 2.3: Example of pathway enumeration in the extended metabolic space. Panel (a) shows the scope graph connecting compounds in the extended chassis (C₁₁, C₁₂, C₁₃, C₁₄, C₁₅, C₁₆) to target compound C₃₁ through reactions R₁, R₂, R₃, R₄ and intermediate compounds C₂₁, C₂₂, C₂₃ in the extended metabolic space (EMS). Panel (b) displays the equivalent stoichiometric matrix. Grayed columns and rows are discarded in the enumeration, as described in the enumeration protocol. Panel (c) shows the reduced matrix used for enumeration, containing an additional reaction T for the selected target compound. The enumeration algorithm found two elementary modes EM₁ and EM₂. Panel (d) shows the resulting 4 pathways solution P₁ to P₄ after expansion of topological equivalent reactions. Pathways P₁ and P₂ involve 3 reactions, while pathways P₃ and P₄ involve 2 reactions.

extended metabolic space, a pathway ranking function needs to be defined. The approach is described in detail in the retrosynthetic RetroPath protocol [152] and a demonstration of the application of such protocol is shown in the XTMS web service [81].

2. Development of novel biosensors based of metabolic pathways. Metabolic pathways that transform a target compound into a detectable compound allow the expansion of the observable extended metabolic space [71]. Such application has been demonstrated through the SensiPath web service [153].

2.5 Notes

1. Molecular signatures are an efficient and intuitive way to model metabolites. They are similar to the well-known Extended Connectivity FingerPrint (ECFP) topological fingerprint, which summarizes compounds in lists of circular molecular fragments.
2. Chemical structures and reactions can be found in multiple formats. Reactions are often defined in a database-specific flat-file where reactants are referenced by their compound identifier. Most of the time, you will find a file in MDL SDF or MOL format binding the compound identifiers to their respective structures. Other interchangeable formats are usually available such as SMILES and InChI. Inter-conversion between formats using standard software such as Open Babel [154] yields to equivalent representations of the compound. A sanity check can help to ensure that they all refer to the same compound. This will eventually filter out wrong annotations.
3. Before being converted into molecular signatures, molecular graphs do not need to represent chemically valid compounds in term of valence, charges, etc. The important point is that compounds (moieties) that should be considered identical according to the final application share the same molecular graph (subgraph). Of course, those simplifications introduced at the compound encoding step must be kept in mind while interpreting the results.
4. Putative enzymes promiscuity can be modeled through molecular signatures given an appropriate diameter. Obviously, as we lower the diameter, the stronger is the promiscuity hypothesis and the riskier are the predictions.

5. Molecular signatures can take into account stereo-chemistry, which is particularly appealing when working with enzymes. Nonetheless, if stereo information is considered, it is important to ensure that it is available (and valid) for most of the compounds, otherwise compounds with and without stereo information will be perceived differently through signatures.
6. Metabolic databases contain generally a substantial portion of reactions that are not stoichiometrically-balanced. Reactions signatures can be computed for reactions that are not need strictly balanced input reaction. Nonetheless, working with balanced reactions is always recommended and is a sign of a well-curated database.
7. This mathematical expression simply states that the reaction signature is the set formed by the difference between product signatures and reactions signatures. Intuitively, can be understood as the chemical groups that are transferred or transformed through the reaction.
8. Multi-substrate reactions are difficult to handle with the proposed equation. Indeed, testing all compounds with a reaction would take N^m tries, where N is the total number of compounds in the database and m the number of substrates anticipated for that reaction. A more practical option is to allow promiscuity for only one substrate at a time, therefore limiting the number of trials to $N * m$. A complementary approach is to allow promiscuity only for non-cofactors compounds.
9. This feature is particularly desirable to untap enzymes full potential in metabolic engineering applications since it can find unexpected synthesis route.
10. There is a basic difference between the information that is required in the model in order to design heterologous metabolic pathways and to estimate steady-state fluxes. In the former case, the most essential information is the knowledge about the metabolites that are endogenous to the organism and therefore can be used as precursors in the heterologous pathway. In the latter case, the accuracy of the stoichiometric relationship between those reactions that directly influence the pathway is required, while partial knowledge about upstream reactions with low influence into the pathway can be tolerated.
11. The extended metabolic space of the model of an organism provides useful information in order to discover previously unidentified routes and to fill gaps in present models.

12. Pathway minimality is a heuristic condition based on reducing metabolic burden in the cell (a pathway with less number of enzymes should be more tolerated by the cell because it potentially imposes less stress).
13. Metabolic networks are formally modeled as hypergraphs for pathway enumeration. Basically, the availability of each substrate is required in the reaction in order to produce the product. That creates some level of complexity higher than in classical graph pathway enumeration algorithm. Moreover, standard graph approaches do not consider stoichiometry. The stoichiometric approach, in turn, based on linear algebraic decomposition provides an easier analytic approach.
14. Pathway enumeration based on elementary flux modes can become computationally intractable for highly connected networks such as central metabolism. However, in cases where we want to produce some heterologous compound in a chassis organism, pathways are generally almost linear and the elementary flux mode enumeration remains tractable. The enumeration of elementary flux modes can be also expressed as a dual problem using minimal cut sets.
15. We remove all initial compounds in the chassis, as we already know that they are available. Products of reactions in the scope consuming the initial compounds will be kept for the enumeration.
16. Identical columns represent routes that are topologically equivalent. In order to make the enumeration algorithm more efficient, we remove duplicated columns. However, for the final enumeration we should list each topologically equivalent reaction as an alternative pathway.

RetroPath2.0: a retrosynthesis workflow for metabolic engineers

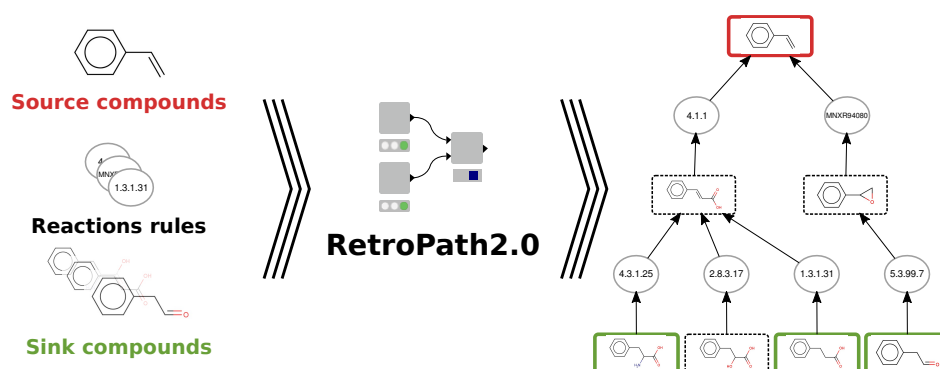


Fig. 3.1: Graphical abstract for RetroPath2.0. RetroPath2.0 computes the reaction networks between *source* and *sink* compounds thanks to reactions *rules* that encode enzymatic reactions.

This work was published in *Metabolic Engineering* by Baudoin Delépine, Thomas Duigou, Pablo Carbonell and Jean-Loup Faulon.

In this version, supplementary figures and tables were inserted into the main text for a better integration with its flow. Supplementary figures and tables are annotated as such in their captions.

Last version of the workflow can be downloaded at myexperiment.org. The supplementary data are hosted by Metabolic Engineering.

Full reference

Baudoin Delépine, Thomas Duigou, Pablo Carbonell, and Jean-Loup Faulon. “RetroPath2.0: A retrosynthesis workflow for metabolic engineers”. In: *Metabolic Engineering* 45 (Jan. 2018), pp. 158–170. DOI: 10.1016/j.ymben.2017.12.002

Contribution

B.D. and T.D. contributed equally to the work. Project was initiated by J-L.F. All authors participated to the preparation of the manuscript.

Highlights

- State-of-the-art Computer-Aided Design retrosynthesis solutions lack open source tools and easiness of use
- We propose RetroPath2.0 a modular and open-source workflow to perform retrosynthesis
- RetroPath2.0 computes reaction network between Source and Sink sets of compounds
- RetroPath2.0 is distributed as a KNIME workflow for desktop computers
- RetroPath2.0 is ready-for-use and distributed with reaction rules

3.1 Abstract

Synthetic biology applied to industrial biotechnology is transforming the way we produce chemicals. However, despite advances in the scale and scope of metabolic engineering, the research and development process still remains costly. In order to expand the chemical repertoire for the production of next generation compounds, a major engineering biology effort is required in the development of novel design tools that target chemical diversity through rapid and predictable protocols. Addressing that goal involves retrosynthesis approaches that explore the chemical biosynthetic space. However, the complexity associated with the large combinatorial retrosynthesis design space has often been recognized as the main challenge hindering the approach. Here, we provide RetroPath2.0, an automated open source workflow for retrosynthesis based on generalized reaction rules that perform the retrosynthesis search from chassis to target through an efficient and well-controlled protocol. Its easiness of use and the versatility of its applications make this tool a valuable addition to the biological engineer's desk. We show through several examples the application of the workflow to biotechnological relevant problems, including the identification of alternative biosynthetic routes through enzyme promiscuity or the development of biosen-

sors. We demonstrate in that way the ability of the workflow to streamline retrosynthesis pathway design and its major role in reshaping the design, build, test and learn pipeline by driving the process toward the objective of optimizing bioproduction. The RetroPath2.0 workflow is built using tools developed by the bioinformatics and cheminformatics community, because it is open source we anticipate community contributions will likely expand further the features of the workflow.

3.2 Introduction

Despite the increasing number of small molecules that are bioproduced, the research and development process (R&D) is still costly and rather slow. For instance, the metabolic engineering of artemisinic acid is claimed to have taken more than 130 person-years and about 10 years to complete [32, 156]. Among the challenges that industrial biotechnology is facing to deliver sustainable solutions are 1) the reduction of R&D costs and 2) the bioproduction of a wider palette of compounds. To address these challenges, computational/experimental strategies where alternative metabolic pathways are first designed and assessed before being built and tested have been proposed (see reviews [22, 35, 80, 157]). While some computationally-driven strategies make use of known metabolic reactions albeit not necessarily in the same species [158, 159] others allow to design pathways that encompass novel reactions not stored in metabolic databases, these latter tools make use of retrosynthesis algorithms [29, 62, 92, 159–162].

Retrosynthesis algorithms take as input a set of metabolites, for instance the metabolites in a growth medium or the metabolites of a chassis strain model, and the set of target compounds to bioproduce. Ideally the target compounds could be any molecule in the chemical space. The algorithms generate retrosynthesis networks linking the target compound(s) (the source) to the metabolites of the chassis strain (the sink) through reactions.

Such retrosynthesis networks should be further processed to map or extract information relevant for the biological application. For instance, some algorithms can be applied to enumerate pathways [150] and rank them based on several criteria including enzyme availability and performance, product and intermediate compound toxicities [163] or the theoretical yield of the desired compound [62, 81, 90, 92]. Interestingly, retrosynthesis networks exploitation is not strictly limited to retrosynthesis. Applications have been proposed to predict biodegradation routes [66, 164, 165] in order to identify unknown compounds from the underground metabolism [79], to predict the transitions of labelled atoms in metabolic

networks [166, 167], and to design biosensing circuits for compounds for which no direct biosensors are known [153]. The main difference of the aforementioned applications lies in the definition of source and sink compounds sets; the current paper focuses on retrosynthesis but our solutions still stand for other applications requiring network generation.

One issue users of retrosynthesis-based solutions are facing is that algorithms and underlying data have not been fully documented and released. In most cases, authors provided fine-tuned webserver [62, 79, 81, 162] often filled with pre-generated data that focuses on some exemplar cases. Based on this information, it is difficult for users to grasp methods' limitations, to improve them, or to exploit them for different uses. At a time when open-data principles gain more and more traction [168–170] we believe this lack of flexibility should be overcome.

In this spirit, we developed the RetroPath2.0 workflow on the KNIME analytics platform [145] to answer the need for a modular and easy-to-use tool to predict reaction networks. Workflows have several advantages over scripting languages. A graphical user interface allows for rapid test and prototyping, even for users with little to no knowledge in programming. For instance, parallelization of tasks is inferred from workflow topology and does not need any special library or technical knowledge from the user. Once configured, workflows are readily deployable on all platforms where KNIME can be installed. KNIME workflows are popular in cheminformatics to prepare and analyse data, as shown by the number of extensions maintained by users in this field [145, 171]. Thus, metabolic engineers benefit from a large panel of tools to analyse the chemical diversity and features of their data. As a matter of fact, RetroPath2.0 was developed using only community tools. We foresee it will make the workflow easier to modify and at the very least a good proof of concept of what can be done with workflows.

The current paper provides for the first time a simple workflow encompassing the main steps of the retrosynthesis process. We hereby review the main steps of retrosynthesis algorithms in order to demystify their use and shed light on the shortcomings of current tools [29, 62, 92, 159–162]. We then outline our proposed solution through several applications in metabolic engineering and biosensor engineering. RetroPath2.0 is available in supplementary along a set of reaction rules and some classic metabolic engineering examples to test RetroPath2.0 features.

3.3 Theoretical background

3.3.1 Encoding reaction as reaction rules

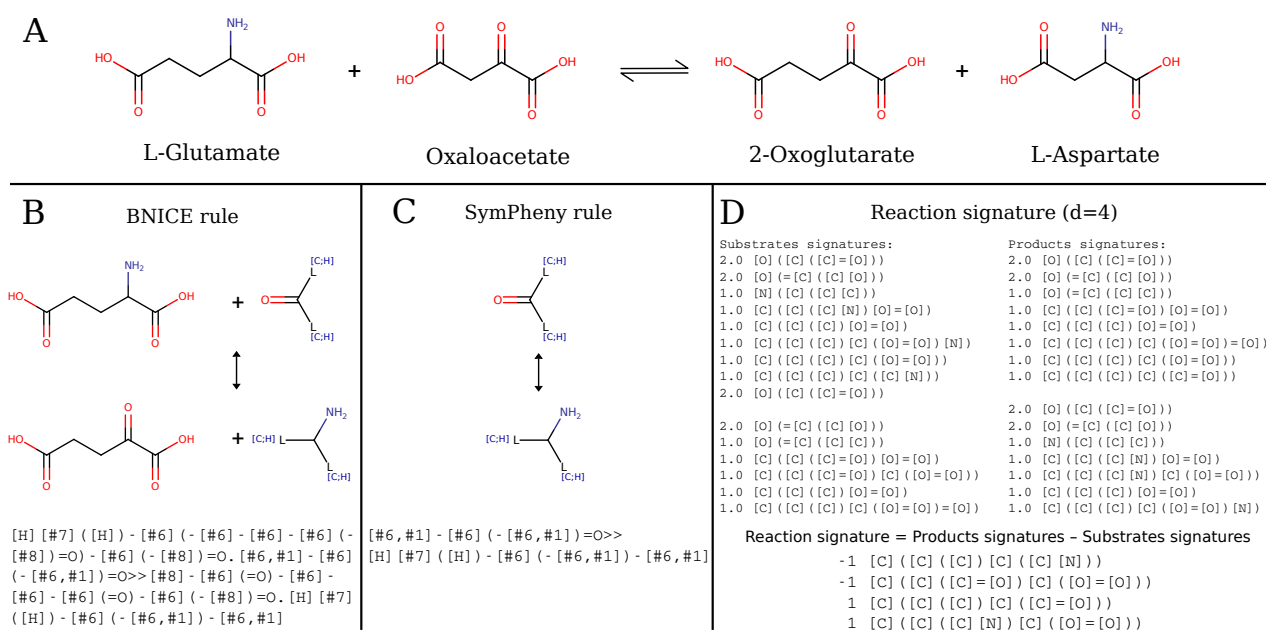
The first challenge that retrosynthesis algorithms have to address is linked to the way reactions are encoded. Most retrosynthesis algorithms are based on reaction rules, but other strategies exist to encode reactions [172, 173]. A reaction rule generally depicts the change in bonding patterns when transforming a set of substrates (reactants) into a set of products. For retrosynthesis applications, rules are reversed such that one computes the substrates from the products.

Several solutions have been proposed to code for reaction rules, namely Bond-Electron (BE) matrices [174], reaction SMARTS [59], RDM patterns [66], and reaction signatures [137]. Examples of coding systems are illustrated in Figure 3.2. We highlight below some key concepts to understanding reaction rule encoding in a retrosynthesis context.

Enzymatic promiscuity

Reactions for retrosynthesis applications should be modelled with a controlled degree of generalization for their substrates and products. Indeed, reaction rules containing a full description of substrates and products chemical structures cannot be applied on new compounds. This is the case for classic metabolic models and database and their lack of generalization prohibits the generation of novel pathways. The use of generalized chemical transformations is required in order to be able to predict new metabolic transformations. Such predictions are necessary since reaction databases are not complete [50, 51] and side enzymatic activities are often underestimated.

This lack of knowledge on alternative enzymatic activities is currently a critical limiting factor for metabolic engineering since it has been estimated that 37% of *E. coli* K12 enzymes have a promiscuous activity for other substrates structurally similar to their main known substrate [44]. In order to be able to generate new metabolic transformations (and new compounds) one thus needs to use generalized reactions to model enzymatic promiscuity, i.e. rules that can be applied to different substrates, and eventually on compounds absent from the databases. For instance, BNICE [79, 161, 162] and SimPheny [29] use a collection of reaction rules that, as depicted in Figure 3.2, can be applied to any ketones (including oxaloacetate) since their encoding is focused on the reaction centre.



Identification of the reaction centre

Reaction rules used in retrosynthesis generally require a solved Atom-Atom Mapping (AAM, see panel A in Figure 3.3) between the atoms of the substrates and those of the products to identify the reaction centre of the reaction [55, 66, 90, 92, 175]. The AAM problem is equivalent to the Maximum Common Substructure, or the subgraph isomorphism problem which turns out to be NP-hard [69]. Avoiding the use of AAM to generate rules is nevertheless possible in some cases, as it was originally shown by a previous version of the RetroPath algorithm based on fingerprint subtraction [71] (see Figure 3.2).

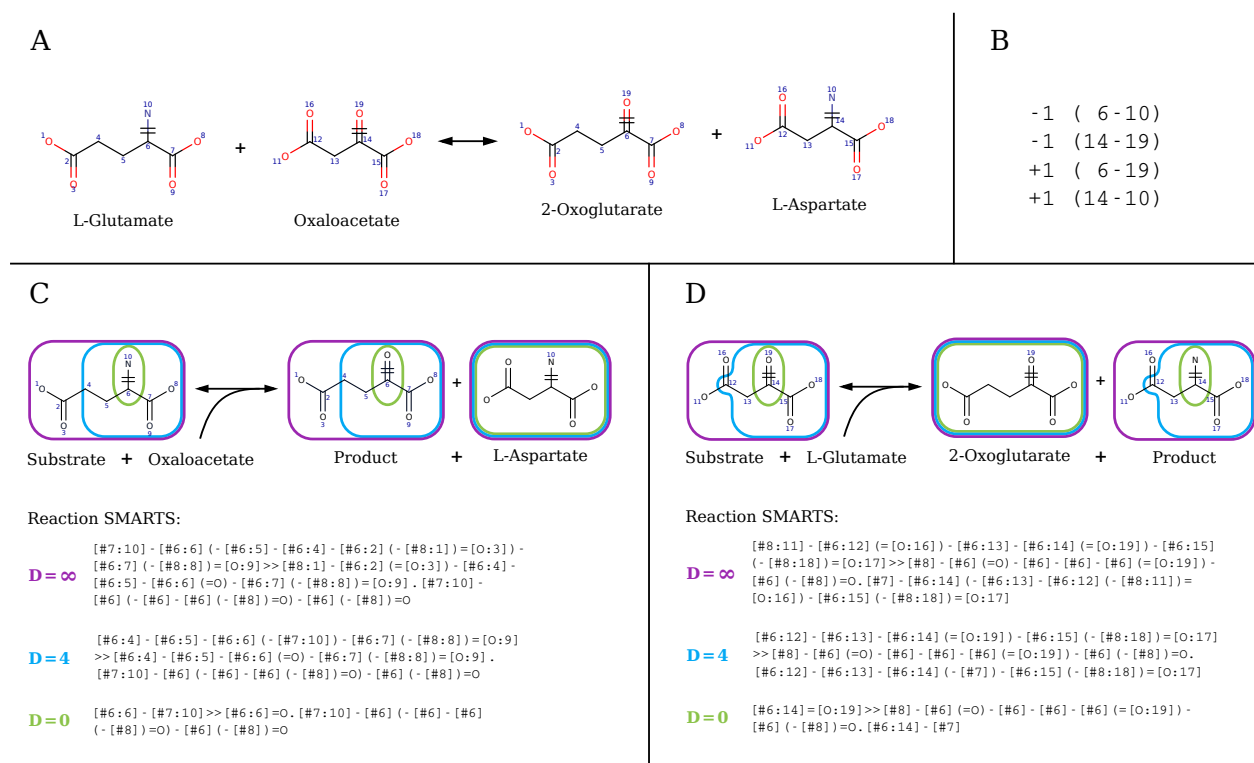


Fig. 3.3: RetroPath2.0 rules and corresponding SMARTS for reaction 2.6.1.1 at various diameters. A. Full reaction 2.6.1.1 with atom mapping. **B.** The list of broken bonds (-1) and bonds formed (+1) is given by their atom numbers. **C.** The corresponding SMARTS for the component modelling promiscuity on L-glutamate: Substrate + Oxaloacetate = Product + 2-Oxoglutarate. **D.** The corresponding SMARTS for the component modelling promiscuity on oxaloacetate: Substrate + L-Glutamate = L-Aspartate + Product. C and D. Rules are encoded as reaction SMARTS and characterized by their diameter (∞ purple, 4 blue or 0 green), that is the number of bonds around the reaction centre (atoms 6, 10 and 14, 19) defining the atoms kept in the rule. This allow for a controlled and flexible modelling of enzymatic promiscuity. Note that for the case of 2.6.1.1 the co-product is always the same (C: L-aspartate; D: 2-oxoglutarate) but that is not always the case, depending on the connectivity of the atoms belonging to the reaction centre.

problem, the definition of the reacting centre is extended to neighbour atoms, either systematically at a predefined bond-distance (diameter, panel C and D Figure 3.3) or based on expert-knowledge.

Systematic rule generation

Reaction rules can be computed in principle by processing the set of reactions stored in metabolic databases. However there are some difficulties associated with this task. Exhaustive rules generation is certainly another major challenge for retrosynthesis. We can distinguish two main philosophies to systematically encode enzymatic reactions.

The first approach consists in encoding a small set of generalist rules guaranteed by a model to cover all possible classes of reactions. According to the Enzyme Commission (EC) nomenclature all reactions that belong to the same third level EC number should follow the same chemistry, while the fourth and last level is for disambiguation [76]. Both SimPheny and BNICE use the third EC number level to guide their reaction encoding effort. SimPheny [29] has 50 manually curated reaction rules, and the number of rules of BNICE systems are of the same order; 86 for [161], 198 for [79], 722 for [162]. This approach is well-suited for manual curation, since even if the number of reactions to annotate is rather small, it is supposed to be exhaustive in terms of the involved chemistry. Nonetheless, relying on EC numbers often requires adding exceptions since some reactions at the third level of EC numbers do not share any common substructure and thus cannot be expressed by the same rule. For instance, the carbon-halide lyases class (EC 4.5.1.*) is composed of five fourth level reactions which all remove a chlorine atom, but some reactions also remove a primary amine from a substrate and replace it either by a double bonded carbon, a hydrogen, an oxygen atom or a more complex functional group (Figure 3.4). Their number of substrates and products also varies. Clearly, these reactions cannot be encoded using a single BE matrix, a reaction signature, or an intelligible reaction SMARTS. Another need for exceptions arises from the fact that many reactions have no EC number assigned by the Commission [76].

The second approach, which is more data-driven, is to automatically compute rules for all available metabolic reactions by selecting only the atoms belonging to a sphere of fixed diameter around the reaction centre. This is the approach adopted by the workflow proposed in this paper, RetroPath [70, 81], and others [64, 69, 176]. Ideally, the diameter used should directly be linked to known promiscuity of an enzyme's sequence. In our experience, a diameter of 6-8 (see 3.7.1 for a detailed discussion on diameter selection and promiscuity) is generally

	<p>Rules</p> <p>R1</p> <p>[#6]-[Cl]>>[#6]=Cl</p>
	<p>R2</p> <p>[#7]-[Cl].[N]>>[#7]=N.Cl.N</p>
	<p>R3</p> <p>Cl[Cl]>>Cl=Cl.Cl.Cl</p>
	<p>R4</p> <p>[#6]-[Cl].[N]>>[#6]=N.Cl.N</p>
	<p>R5</p> <p>[#6]Cl.[S]>>[#6]=S.Cl.S</p>
<p>4.5.1.1, 4.5.1.3, 4.5.1.5</p>	<p>R6</p> <p>Cl*[A]>>[A]=Cl.Cl</p>

Fig. 3.4: Reactions rules for EC class 4.5.1.-. Hydrogen are omitted for simplicity thus O stands for H₂O, Cl for HCl and N for NH₃. The last rule (R6) does not apply 4.5.1.2 and 4.5.1.4 since in addition to the removal of chlorine these reactions also involved removal of nitrogen. This figure was originally in Supplementary.

a good trade-off to cover known reactions' specificity with a reasonable amount of promiscuity predictions (see Section 3.5.1 for an evaluation of rules performance for promiscuity classification and [70, 177]). Using the procedure outlined in the caption of Figure 3.3, when applied to the MetaNetX database [141] the number of rules returned is between 6,900 and 19,000 depending on the parameters used to model enzymatic promiscuity (diameter) for the 31,527 reactions stored in MetaNetX (MNXR identifiers, v.2.0). Interestingly, not only multiple generated rules can belong to the same EC class, but also a same rule can correspond to several EC classes. For instance, at diameter 4, three EC numbers (2.6.1.1, 2.6.1.17, 2.6.1.67) from three distinct reactions (resp. MNXR32641, MNXR32641, MNXR31792) are associated to the same rule depicted in Figure 3.2D (promiscuity on oxaloacetate, MNXM42).

Cosubstrates, cofactors and coproducts

Another challenge for retrosynthesis algorithms is the need to handle reactions processing multiple substrates and/or multiple products. Dealing with multi-substrate reactions requires more computational resources in order to model enzymatic promiscuity for each combination of promiscuous substrates (Figure 3.3).

For these purposes, cosubstrates and coproducts that are currency cofactors (such as water, CO₂, ATP, NADP, etc.) can be ignored from the rules under the

Tab. 3.1: Retrosynthesis networks generation tools.

	Reaction rules calculation	Rules coverage	Number of rules	Reaction rule specificity	Multiple product & substrates	Enzyme sequence search	Combinatorial complexity	Availability
SimPheny (BioPathway Predictor) [29]	Computed from 3rd EC level followed by manual curation	All metabolic reactions	50	Fixed	No	No	Controlled by network and molecules size	No
BNICE [162]	Automated from KEGG followed by manual curation	All metabolic reactions	722	Fixed	Yes	No	Controlled by network size	Web server
PathPred [159]	Automated from the KEGG RPAIR database	Xenobiotic degradation and biosynthesis of secondary metabolites	853 (degradation) 1126 (biosynthesis)	Fixed	No	No	Controlled by similarity	Web server
GEM-Path [62]	Computed from 3rd EC level	All metabolic reactions	443	Fixed	Yes	No	Controlled by similarity and thermodynamics	No
METEOR [160]	Knowledge-based expert system (Lhasa Ltd.)	All metabolic reactions?	357	Fixed	No?	No	Controlled by “reasoning” rules	Commercial
Reverse Pathway Engineering (THERESA) [92]	Automated from the BioPath database (Molecular Networks GmbH)	All metabolic reactions	3,516 reference reactions	Fixed	Yes	Comparative genomics	Manual compound selection	Commercial
RetroPath [71]	Automated computation on MetaCyc	All metabolic reactions	between 3,000 (d=4) and 5,000 (d=14)	Variable, controlled by diameter	Yes	Post-process using machine learning	Controlled by diameter	Web server
RetroPath2.0 (this study)	Automated computation on MetaNetX	All metabolic reactions	between 6,900 (d=2) and 19,000 (d=16)	Variable, controlled by diameter	Yes	Embedded using sequence clustering	Controlled by diameter and enzyme score	Open source

assumptions that they are available in the cell and that there is no gain for retrosynthesis analysis in modelling promiscuity on them. However, information about cofactors participating in reactions should not be discarded since they could be used at a later stage to sort pathways by their efficiency in terms of cofactor exchange and the burden they impose on central metabolism.

Nonetheless, even if we ignore currency metabolites in the rules, around a third of metabolic reactions still remains multimoleculars (see 3.7.2). Our practical solution is to model enzymatic promiscuity for only one substrate at a time, meaning that for any multi-substrate reaction “ $A + B \rightarrow C + D$ ”, alternatives substrates A' and B' are never tested together to limit the combinatorial complexity. RetroPath2.0 follows this solution as we encode one rule per reference substrate (for components “ $A \rightarrow C + D$ ” and “ $B \rightarrow C + D$ ”) as shown in Figure 3.3. Others embrace combinatorial complexity (Figure 1B and D) or simply ignore all cosubstrates (Figure 1C).

3.3.2 Building (retrosynthesis) reaction network

In all algorithms listed in Table 3.1, retrosynthesis maps are constructed by applying reaction rules in an iterative fashion starting from a source set of compounds until the molecules in a sink set of compounds are found in the map. In the context of metabolic engineering, if the rules are applied in a forward manner, the source set is composed of the native metabolites of the chassis strain and the

sink set are the molecule we wish to produce. If the rules are applied in a reverse manner then the source set are the molecules to be produced and the sink set are the metabolites of the chassis. One bottleneck that all algorithms face is computation complexity due to the combinatorial explosion of the number of reactions predicted from the rules. This is true regardless of whether the reactions are applied in a forward or reverse manner. As an example, let us assume we wish to perform retrosynthesis for some FDA approved drugs in *E. coli*. In the reaction list we have at our disposal there is one for reversed hydro-lyases (i.e. reversed 4.2.1). According to [161] the rule for that reversed reaction is $R1C(=O)C(R2)=C(R3)R4 + O-R5 \rightarrow R1C(=O)C(R2)-C(R3)(R4)OR5$, where $R5$ can be C, H, O, and S and all other R s can be any atoms. Assuming $R1C(=O)C(R2)=C(R3)R4$ is the main substrate (our drug target) and $O-R5$ the cosubstrate, 68 FDA approved drugs from DrugBank contain the first substructure. If we restrict the cosubstrate to be in the *E. coli* model iJO1366 then 653 metabolites out of 810 compounds in the model contain the second substructure, while 50,810 compounds from MetaNetX will pass the substructure test. Taking Vitamin C as an example of a DrugBank compound that passes the substructure filter, one finds 1,883 unique products when applying the reversed rule 4.2.1 to Vitamin C and *E. coli* metabolites and 343,177 products when the cosubstrate is in MetaNetX. There are more products than substrates because for some substrates the reversed rule 4.2.1 applies to more than one location.

As already mentioned, for a given retrosynthesis target one needs to apply all rules to the target, all rules to the products obtained by application of the reversed reactions to the target, and so on until a predefined stop condition occurs (often the number of iterations). Clearly, if reaction rules generate more than 1,000 products even with 50 rules the problem starts to be challenging -if not impossible- to manage computationally after 2 or 3 iteration steps.

Strategies are needed in order to cope with that complexity. RetroPath proposes a solution where reactions are scored according to their ability to retrieve enzyme sequences catalysing substrate to product transformations. Reactions below a predefined score are removed from the retrosynthesis map. For any given reaction the sequence scores are computed by machine learning using a technique that we developed earlier. The model is trained on all known pair "enzyme sequence" x "(substrate, product)" using Support Vector Machines [177] or Gaussian Processes [178]. GEM-Path [62] proposes another strategy where for each reaction the substrates are accepted if they are similar enough to the substrates of the reference reactions.

We detail in the next sections a new implementation of RetroPath to predict reaction networks and perform retrosynthesis among other applications. RetroPath2.0

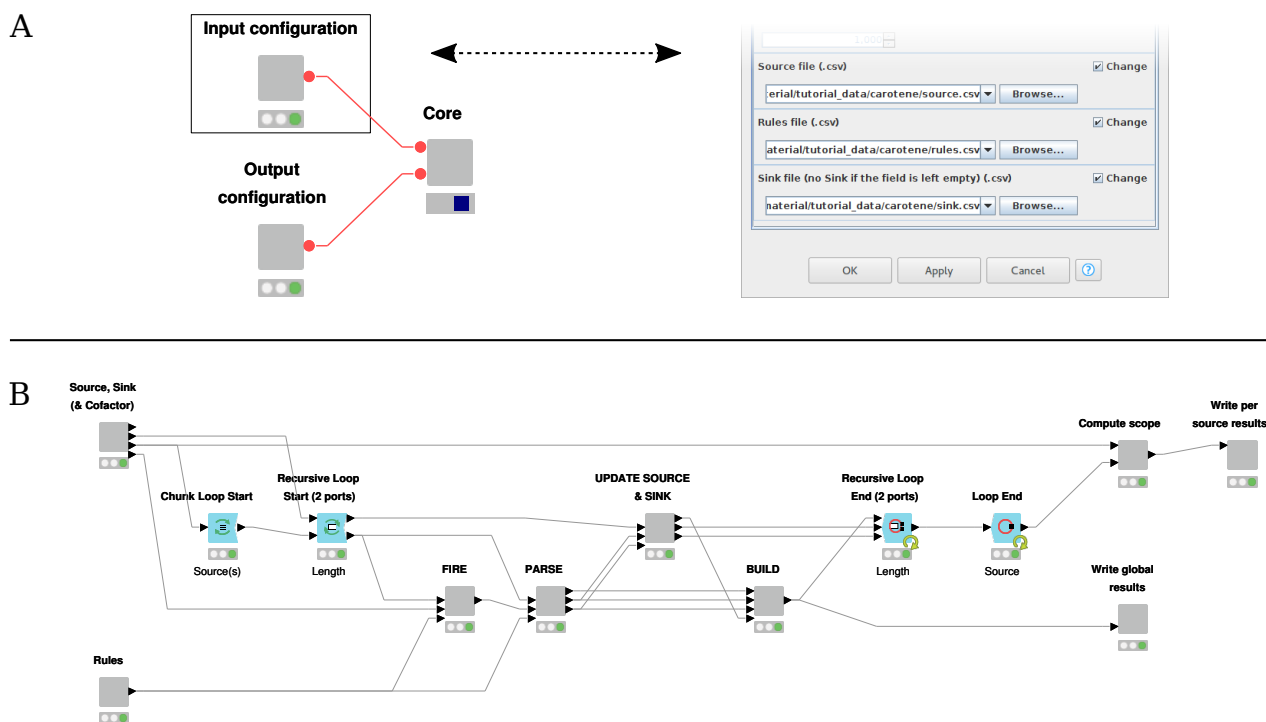


Fig. 3.5: RetroPath2.0 KNIME workflow. **A.** Main panel view (left) and input configuration window (right) that allow the user to set up parameters. **B.** Inner view of the "Core" node where the computation takes place. The "Source, Sink..." and "Rules" nodes parse the source, sink and rules input files provided by the user and sanitize data so that it can be processed by downstream nodes. The outer loop ("Source" loop) iterates over each source compounds, while the inner loop ("Length" loop) allows to iterate the process up to a maximum number of steps predefined by the user. The nodes (i) "FIRE", (ii) "PARSE", (iii) "UPDATE SOURCE..." and (iv) "BUILD" are sequentially executed at each inner iteration. Respectively, they (i) apply all the rules on source compounds, (ii) parse and sanitize new products, (iii) update the lists of source and sink compounds for the next iteration and (iv) merge results that will be written by the node "Write global results". Once the maximum number of steps is reached (or no new product is found), the "Compute scope" node identify the scope linking each source to the sink compounds, then these results are written by the node "Write per source results". Only the main nodes involved in the process are shown.

addresses the challenges listed above with a special attention to remaining easy to use and modifiable by end users, unlike tools developed so far. In that sense, both the encoding of reactions into generalized rules and the actual use of those rules to predict new reactions depend strictly on resources developed by the community.

3.4 Methods

The workflow proposed in this section can a priori be used to all systems presented in Table 3.1 to construct retrosynthesis maps as long as reaction rules can be coded by reaction SMARTS. As examples, we provide such set of rules extracted from the SimPheny, BNICE and RetroPath systems.

Our computational methods make use of in-house algorithms [71], RDKit routines [41] and KNIME nodes [145]. They have been implemented in the form of a KNIME workflow -called RetroPath2.0- (Figure 3.5) that we provide in the supplementary materials, in addition of sets of rules, examples, and useful data files.

3.4.1 Reaction rules

RetroPath2.0 uses reaction SMARTS to encode reactions. It is a SMIRK-like reaction rules [59] format defined by RDKit and is mostly compatible with other tools (see Figure 3.2).

Collected reaction rules

Rules for SimPheny and BNICE were extracted from [29] and [161] respectively, and manually entered by using Chemaxon Marvin Sketch software products (15.5.18, 2015, <http://www.chemaxon.com>). For each rule atom mapping was calculated by Marvin Sketch and the resulting rules were stored in SMARTS format as in Figure 3.3.

Generated reaction rules

We used MetaNetX version 2.0 [141] as the reference database for metabolic reactions that we encoded in rules. MetaNetX is a meta-database that compiles into a single reference namespace both reactions and metabolites extracted from main metabolic databases such as KEGG, Metacyc, Rhea or Reactome. Reactions can contain many substrates and many products. We performed an Atom-Atom Mapping (AAM) using the tool developed by [179] on all MetaNetX reactions (Figure 3.3A). We filtered out transports reactions and those involving compounds with incomplete structures (class of compounds, R-groups, etc.). Stereochemistry was removed.

Multiple substrates reactions were decomposed into components (panel C and D in Figure 3.3). There are as many components as there are substrates and each component gives the transformation between one substrate and the products. Each product must contain at least one atom from the substrate according to the AAM. This strategy enforces that only one substrate can differ at a time from the substrates of the reference reaction when applying the rule (see section 2.1.1 enzymatic promiscuity modelling).

Next step consisted in computing reactions rules as reaction SMARTS for each component. We did it for diameters 2 to 16 around the reaction centre (panels C and D in Figure 3.3) by removing from the reaction components all atoms that were not in the spheres around the reaction centre atoms.

We extracted more than 24,000 reaction components from MetaNetX reactions, each one of those leading to a rule at each diameter (from 2 to 16).

We provide in Supplementary a subset of 14,300 rules for *E. coli* metabolism, both in direct and reversed direction. The rules were selected based on the MetaNetX binding to external databases and the iJO1366 whole-cell *E. coli* metabolic model [180].

3.4.2 Building (retrosynthesis) reaction networks between two pools of compounds using the RetroPath2.0 workflow

The RetroPath2.0 workflow essentially follows an algorithm proposed by some of us [43]. After removing all source compounds already in the sink set, the workflow applies the rules to each of the compounds of the source set. For each compound, the products are computed using the RDKit KNIME nodes [41]. Products are standardised and duplicates are merged. All pairs substrate-product are added to the growing network along with the reaction rules linking them.

In the next iteration, the set of products becomes the new source set. However, before iterating, the workflow removes from the new source set all compounds that belong to the sink (as these are already solutions and there is no need to iterate) and the workflow adds the product set to the sink in order to avoid applying reactions on the same products during subsequent iterations. Consequently, the workflow computes only the minimal routes between source and sink, i.e. routes in which all reactions are essential for their viability, and thus minimizes the number of enzymes to be added to a chassis strain when implementing the pathways. This feature can be ignored by not specifying a sink for the first iteration.

The RetroPath2.0 workflow iterates until a predefined number of iterations is reached or until the source set is empty. The final produced graph is composed of the list of links between substrates and products annotated with their corresponding reaction rule. Products belonging to the sink are annotated as such.

Note that the iterative process can reveal itself to be quite computationally demanding. To tackle this issue, RetroPath2.0 has a feature to bias the reaction space exploration toward compounds generated by trusted rules, using a rule-wise penalty score. If too many compounds are generated and need to be handled at once, only a predefined number of compounds with the lowest penalties according to their generating rules are kept in the new-source of the following iteration. Of course, both the definition of the penalty and maximum number of compounds to keep are critical and fall within the responsibility of the user. As described next, the rules we provide are scored to optimize *in vivo* pathway feasibility by penalizing rules associated to enzymatic reactions with inconsistent sequence annotation.

3.4.3 Score rules by enzyme sequence consistency

Predicted reactions in the final graph generated by the RetroPath2.0 retrosynthesis workflow need to be associated with enzyme sequences in the final engineering of the pathways. The selection of such sequences should look for a trade-off between the specificity of the reaction rule and the information available in enzyme databases for the reaction through the EC classification. Whereas the EC classification has traditionally provided a hierarchical numerical classification of enzyme-catalysed reactions to progressively describe reactions in finer detail, RetroPath2.0 introduces a similar hierarchical classification that is controlled by the diameter used in rule generation. In some cases the diameter of the reaction rule found by the RetroPath2.0 workflow might be high, i.e. highly specific to that reaction. However, it often occurs that there is no annotated enzyme sequence for the rule. In order to find some candidate sequences, we look into reactions that are close according to the EC hierarchy for each EC class containing at least one instance of the rule at given diameter. Traversing both rules diameter hierarchy and the underlying EC classes allows the selection of plausible sequence candidates for each reaction rule.

We compiled the set of Uniprot sequence identifiers annotated for reactions by looking at the cross-link annotations in MetaNetX for Rhea and MetaCyc databases. In total 208,980 sequences from 5,388 organisms were associated to 7,793 reactions. At a given diameter of the rule, we iteratively assigned sequences to rules. First, reactions with annotated sequences were collected for each generated rule. Since a rule can represent one or more reactions at a given diameter, sequences coming from different reactions sharing the same rule were aggregated into a single set for that rule. These direct annotations only provided a partial coverage for the total rules in the database. For instance, at diameter $d = 8$, there were 7,898 orphan rules, i.e. rules that were generated from reactions lacking

sequence annotation (Table 3.2). Similarly, there were 6,280 orphan reactions at diameter $d = 8$. In order to increase the coverage, we considered the EC class of the reaction when such information was available. Sequences associated with reactions sharing strictly the same EC class were combined together. Adding together such annotations for the same EC class fixed issues related to partial annotations for the less common reactions. In that way, the number of orphan rules was significantly reduced to 1,719, which is approximately a 13% of the total rules. Similar ratios were observed for reactions.

For the orphan rules having no sequence annotation after considering the EC class of the reactions, we followed the strategy of reducing the specificity of the EC class by reducing the number of digits. In other words, if a rule had no annotation based on the EC class at 4 digits, we looked at reactions that shared same EC class at 3 digits with one reaction associated with the rule and so on until we found sequence annotations. Notably, a sharp decrease on the number of orphan rules already occurred at the level of three digits of the EC class. The remaining orphan rules, less than 1%, was eventually annotated once we reduced the specificity from 3 digits down to 1 digit in the EC class.

We should emphasize that in the procedure described below, sequence annotations that merged multiple EC classes sharing same initial digits were only used for those cases where no sequence information was available at higher EC class levels. This annotation from higher to lower specificity in the set of sequences associated with the rules depending on known sequences allowed us to score the rules. A rule that has associated sequences with low diversity should in general correspond to cases where the sequence information is highly specific to that rule. As the diversity of sequences increases the specificity of those sequences to their associated rules becomes lower. We evaluated such degree of specificity by considering the degree of clustering of the sequences associated with the rules. Clustering of the sequences was performed by using Cd-hit [181]. According with this algorithm, our database of 208,980 amino-acid sequences was clustered into 22,221 clusters for a similarity threshold of 0.5. We used a penalty score for the rules based on the number of sequence clusters n_{rule} contained in the sequences selected for a given rule:

$$score(rule) = \log_{10}(n_{rule}) \quad (3.1)$$

where the logarithm is applied for regularization. A penalty score of 0 implies high specificity, as this means that all sequences belong to a single cluster, while high penalty scores imply multiple clusters and therefore low specificity in the sequence annotation.

Tab. 3.2: Assignment of sequences to rules at different diameters and using decreasing EC class level of specificity. Orphan SMARTS and reactions at a given EC level are the only ones that are considered for sequence assignment at a lower level. This table was originally in Supplementary.

Diameter	EC level	SMARTS	Orphan SMARTS	Reactions	Orphan reactions
2			3680		3259
	4		655		823
	3	8210	47	13782	86
	2		5		12
	1		0		0
4			5141		4377
	4		1058		1280
	3	10501	61	13782	90
	2		15		25
	1		0		0
6			6806		5593
	4		1466		1661
	3	12573	72	13782	96
	2		16		25
	1		0		0
8			7898		6280
	4		1719		1918
	3	13855	73	13782	97
	2		16		25
	1		0		0
10			8603		6611
	4		1836		1863
	3	14772	76	13782	99
	2		16		25
	1		0		0
12			9134		6874
	4		1936		1846
	3	15460	76	13782	94
	2		16		25
	1		0		0
14			9433		7007
	4		1975		1856
	3	15867	77	13782	98
	2		16		25
	1		0		0
16			9741		7159
	4		2025		1811
	3	16227	79	13782	72
	2		16		24
	1		0		0

3.4.4 Enumerating pathways between two pools of compounds

The lists of pathways linking (i) a pool of source compounds to (ii) a pool of sink compounds are computed running an algorithm we developed earlier [71]. This algorithm consists of the following steps for a given source compound:

1. Compute the scope, a subset of predicted reactions between the sink compounds and the set of source compounds. The scope represents the set of compounds and reactions that are involved in at least one pathway. It is computed in a two steps search. First the forward step starting from source compounds finds all reachable compounds that are producible through reactions. Secondly the backward step starting from the sink compound adds to the scope all reactions that can be involved in at least one producible-pathway.
2. Build the stoichiometric matrix. The stoichiometric matrix describes the directed subnetwork involving the set of compounds and reactions identified at the scope step, starting from the source compounds.
3. Enumerate elementary flux modes. An elementary mode corresponds to a minimal unique set of reactions that
 - a) verified the stoichiometric constraints of the network and
 - b) is able to carry non zero-fluxes at the system's steady-state [182]. In order to efficiently compute elementary modes, stoichiometric matrix dimension is generally reduced through lossless compression. Only enumerated flux modes linking source compounds to the sink compound are kept in order to form the final list of pathways. These three steps are performed iteratively for each source compound.

RetroPath2.0 computes the scope for each queried compound. It can be visualized and explored to retrieve the pathways thanks to ScopeViewer, a humble web-application that we provide in Supplementary. Note that the provided workflow does not explicitly extract the pathways and does not rank them. Yet, we provide at <https://github.com/brsynth/rp2paths> a separate utility program “RP2paths” allowing one to enumerate pathways from the results generated by RetroPath2.0.

3.5 Results

We validated our set of rules with RetroPath2.0 by checking that they were able to reproduce the known metabolic space, and that they could be used to perform reaction classification. The capability of RetroPath2.0 to perform retrosynthesis was confronted to in vivo experiments by counting the number of bioproduction pathways found for targets extracted from a database of metabolic engineering successes. We also emphasized the versatile usage of RetroPath2.0 by an original application to design biosensors (see 3.7.3).

3.5.1 Rules validation

The quality of the output of the workflow depends largely on feeding it with the proper set of reaction rules. Some authors [29, 161] have published sets of rules that already constitute an initial test bed. We collected those in addition of a set of SMARTS rules that we compiled for all reactions of the last *E. coli* whole-cell model [180] based on MetaNetX cross-references. Those rules are available in Supplementary. All rules were checked to ensure they could be used with the workflow and yield at least one product.

Coverage of known metabolic space

In order to check the potency of the rules, i.e. that they could indeed be used to predict reactions, we tried to retrieve all reference reactions of MetaNetX from the rules. We compared three dataset of monosubstrate rules according to their origin: SimPheny [29], BNICE [161] and RetroPath2.0. To make a fair comparison we selected from all MetaNetX reactions a subset of 13,000 reactions having an associated EC number and a structure for all its compounds (SimPheny and BNICE rules are based on EC numbers). We extracted from those 6,000 substrates and 7,000 products (MetaNetX identifiers) excluding cofactors. For each rule dataset, all rules were applied on the set of substrates using the workflow with default parameters. We counted the number of products that could be regenerated and the number of generated compounds that were referenced in MetaNetX.

Remarkably given the number of rules considered, 34% of MetaNetX products were recovered by SimPheny rules (50), and 41% by BNICE rules (86). They respectively generated 75,400 and 59,000 compounds, among which 5% and 7% could be found in MetaNetX and are thus connected to a biological database. Since RetroPath2.0 rules were generated from MetaNetX data we expected a

better coverage over the products. This was indeed the case with 96% recovered products from MetaNetX's reactions. The few missed products originated from reactions that could not be encoded in rules due to atom-atom mapping issues. Additionally, 63% of the 17,500 compounds generated by RetroPath2.0 are new to MetaNetX, which highlights the capability of our rule dataset to generate a reasonable amount of new compounds.

The fact that RetroPath2.0 rules generates less compounds than the other tested sets of rules is explained by the differences in term of diameter used. RetroPath2.0 uses a flexible diameter, which by default ranges from 16 to 2, decreasing if no rule can be used on a substrate at higher diameters. This has for effect to prioritize more conservative results (higher diameter) while ensuring that broader promiscuity are tested as a last resort (lower diameter). The few missed products originated from reactions that could not be encoded in rules due to atom-atom mapping issues. Overall, product coverage shows us that RetroPath2.0 rules are able to reproduce most of MetaNetX products, hence most of what is known of the metabolic space.

RetroPath2.0 rules for reaction classification

We evaluated the ability of our rules to perform automated reaction classification. To that end, reactions in the database that contained EC class annotations were grouped into their corresponding EC class at the third level. We then computed the similarity between reactions based on the signature content of their rules. For a given diameter d , each rule was decomposed into its elementary signatures [71] and similarity between two given reactions R_1 and R_2 was computed by means of the Jaccard similarity coefficient $T^d(R_1, R_2)$ applied to the two reaction rules:

$$T^d(R_1, R_2) = \frac{|\sigma^d(R_1) \cap \sigma^d(R_2)|}{|\sigma^d(R_1) \cup \sigma^d(R_2)|} \quad (3.2)$$

The previous expression ranges between 0 (minimum similarity) and 1 (maximum similarity) and has been often applied to compute similarity between compounds or even reaction that are described by binary fingerprints (EC-BLAST [62, 159, 176]). The advantage and main difference of using rules with a selectable diameter is that we can compute the Jaccard similarity coefficient in function of the diameter d . That generates a sequence of monotonically decreasing similarities starting from 0 up to the maximum diameter of the reactants. Similarity of two reactions at diameter 0 contains the basic information about common patterns of bonds that were broken or formed in the two reactions. As we ex-

tend similarity to higher diameters, information becomes more specific to the substrates and products involved in each reaction.

In order to capture efficiently this feature of diameter dependence for Jaccard similarities between rules, we defined a global similarity parameter between reactions $S(R_1, R_2)$ extended to a diameter range $[0, d]$ as an exponentially increasing weighted sum of the Jaccard similarity coefficients:

$$S(R_1, R_2) = \frac{\sum_{k=1}^{n+1} J^{k-1}(R_1, R_2) k^a}{\sum_{k=1}^{n+1} k^a} \quad (3.3)$$

where a is a regularization parameter.

For each reaction in the database, we computed its corresponding rule and similarities based on a diameter range from 0 to 8. In total, rules were computed for 13,782 reactions contained in the database. We used $a = 2$ as regularization parameter.

We then tested the discriminant ability of using such reaction global similarity measure for reaction classification. Our tests were performed using the R package ROCR. We created a positive and negative set for each EC class. The positive set was formed by the set of reactions annotated for this EC class. A balanced training set was then built by randomly selecting from the negative set. For each EC class containing at least 10 data points, as well as for the total set of balanced training set we computed the area under the ROC curve (AUC), resulting in an overall AUC of 0.884 for diameter $d = 8$ (Fig. 3.6). Such performance values are slightly higher than the ones obtained by EC-BLAST [176] by using fingerprint-based similarities, showing the ability of the rules as reaction classifiers.

Score vs. specificity

The ability of substrate generalization of SMARTS rules can potentially be used to assess enzyme specificity. Enzyme specificity is an important factor that needs to be considered for metabolic pathway engineering. Moreover, several studies have shown that enzymes that can catalyse multiple reactions or can process multiple substrates have more evolvability capabilities than specific enzymes [44, 46, 183, 184]. Such property can be approached through our rules as they provide a means for representing chemical transformations for generalized substrates. The level of generalization of reactions and ultimately of their associated enzyme sequences could be therefore quantified using our rules. As described in Methods, one can define a specificity score by assessing the level of generalization of both the reactions and sequences having such reactions at a given rule diameter.

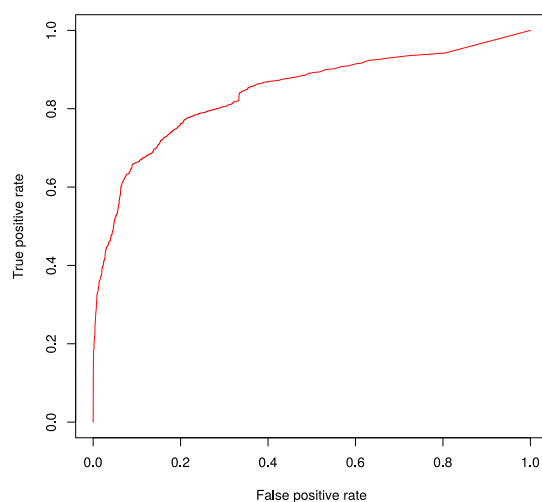


Fig. 3.6: Receiving operating characteristic curves (ROC) curves for the rules of RetroPath2.0 of diameter $d = 8$. This figure was originally in Supplementary.

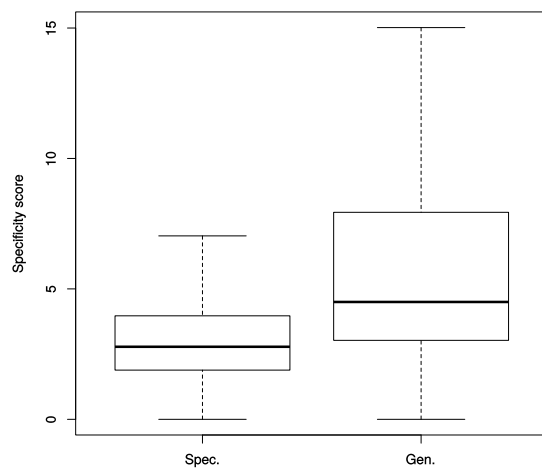


Fig. 3.7: Box plot comparing the distribution of reaction scores for specialist and generalist enzymes in *E. coli*. This figure was originally in Supplementary.

The algorithm traverses both the reaction and sequence space in order to score reaction specificity and more specific rules get lower scores.

To evaluate the ability of the score to represent enzyme specificity, we have analysed a reference set of enzymes in *E. coli* that have been classified as either specific or generalist, i.e. if they can catalyse one or multiple reactions [44]. For each gene, we took their associated reactions in the EcoCyc database [185]. Each reaction was mapped into their associated rule at several diameters d . The resulting scores for each gene were then aggregated. We mapped in total 787 *E. coli* genes, with 602 specific vs. 185 generalist enzymes, respectively.

Notably, the scores computed in that way, as shown in Figure 3.7, displayed the ability to differentiate between these two groups of enzymes, ($t = -6.5144$, p-value of $2.3e-10$ for a Welch's two sample t-test), with specific enzymes receiving lower ranking. We should note that the classification between specific vs. non-specific enzymes depends on the actual knowledge and degree of detail in the description of the reactions in the reference organism and therefore the list of generalist enzymes should be updated as long as new activities are discovered [46]. For instance, we observed a clear outlier in the set of specific enzymes that received a high score based on rules and therefore we should expect wider specificity. This was the case of gene *phoA*, *b0383*, alkaline phosphatase EC 3.1.3.1. It turned out that this enzyme has been reported to have wide specificity [186] in agreement with the high score.

3.5.2 Workflow validation and applications

We tested the reaction network prediction features of RetroPath2.0 workflow with two applications. The typical prediction of bioproduction pathways (see below), and the prediction of small biosensing metabolic circuits for biomarkers (see 3.7.3).

Coverage of bioproduction pathways

The Learning Assisted Strain EngineerRing (LASER) database is a repository for metabolic engineering strain designs [187]. It stores more than 600 successful metabolic engineering designs [188] that have been manually curated from the literature. Those examples are particularly appealing for testing retrosynthesis features since they include an ideal dataset of authentic positive examples of bioproduction pathways, sometimes involving heterologous enzymes. We extracted all compounds targeted for production described in the LASER database (release f6ce080a8993) and used them to assess the ability of RetroPath2.0 to find

retrosynthesis pathways for real-life applications when used with all the rules from MetaNetX.

The structures of the target compounds were inferred from their name by querying PubChem and ChemSpider. 160 compounds targeted for bioproduction were extracted from LASER. To complete further this dataset, we extracted 68 compounds (MBE dataset) published in Metabolic Engineering in 2016 (volumes 33 to 38), a period not covered by LASER. These two datasets contained 203 distinct compounds once merged together based on their structure (standard InChI). Furthermore, we removed *E. coli* endogenous compounds that were used as our “sink”. Finally, 146 distinct compounds were collected to serve as “source” compounds.

Compounds from *E. coli* were extracted from iJO1366 whole-cell model [180] and MetaNetX cross-references. We ignored compounds that belong to so-called “blocked pathways” which are by definition impossible to produce or consume at steady-state in a metabolic model. Such compounds do not constitute a proper source (or sink) for retrosynthesis applications because reactions explaining compound availability in the chassis could be missing. We performed a flux variability analysis to identify them. Overall, we collected 962 MetaNetX identifiers of compounds belonging to *E. coli* that we provide in Supplementary along with their structure (InChI).

All results were generated with a maximum of five retrosynthesis iterations and a timeout of three hours per target on a recent desktop computer. This puts us in realistic operational conditions for users that might have access to modest computational resources. Given those constraints, we successfully found at least one pathway for 81% of the targets (119/146), i.e. a set of reactions allowing the production of the target compound exclusively from *E. coli* endogenous metabolites. Interestingly, we found more than one pathway in most of the cases (104/119).

One of such compounds for which several pathways was found is styrene. Styrene is a building block used in the fabrication of plastics [189]. LASER references one pathway for the bioproduction of styrene from phenylalanine with heterologous enzymes in *E. coli* [190, 191] and in *S. cerevisiae* [192]. RetroPath2.0 found this pathway (Figure 3.8, in red) along with five alternative one from *E. coli* endogenous compounds: 3-phenylpropionic acid, phenylacetaldehyde, and phenylpyruvic acid (Figure 3.8, resp. F, G, and H).

Another non-natural example for which several pathways were found is terephthalic acid (TPA). TPA is a non-natural commodity chemical widely-used for its

ability to form synthetic fibres, and ultimately in the fabrication of polyesters such as PET. TPA is traditionally produced from p-xylene by synthetic chemistry processes [193]. The p-xylene can eventually come from lignocellulosic biomass, making the TPA a bio-based compound in such cases [189]. Interestingly, two enzymatic bioproduction pathways have been reported for TPA, and they follow the same chemical transformations as the ones from synthetic chemistry [193]; one from p-xylene [194] in *Burkholderia* genus, and another from p-toluic acid in *Comamonas testosteroni* [195]. RetroPath2.0 retrieved those routes and proposed alternative shorter paths from endogenous *E. coli* compounds such as phenylalanine, phenylpyruvic acid, and 3-phenylpropionic acid (Fig. 3.9, resp. K, P, and M). To the best of our knowledge, those pathways have never been implemented *in vivo*.

Those results highlight the interest of RetroPath2.0 for retrosynthesis applications. As an additional example, see also the pathways predicted toward ethylene glycol (Figure 3.10). RetroPath2.0 is able to reproduce validated pathways and to propose new ones, both for natural and non-natural compounds. All results are provided in Supplementary.

3.6 Discussion

The RetroPath2.0 workflow is a versatile reaction network tool, built to be modular enough to answer most metabolic engineering needs. RetroPath2.0 takes as input a first set of compounds (the source), a second set of compounds (the sink) and a set of reaction rules (see Figure 3.5). The workflow produces a network linking the source set to the sink set, where each link in the network corresponds to a reaction rule. The RetroPath2.0 workflow runs under the KNIME analytics platform and is available in Supplementary Material.

The choice of source, sink and rule sets depends on the application. For instance, if one wishes to find all possible synthesis routes that can be engineered for a target compound, then the source set will be the target, the sink will be the set of metabolites of the chassis strain, and the rules will be the reversed form of all known metabolic reactions (cf. 3.5.2). If one is interested in finding pathways to be engineered to degrade a given xenobiotic, the source set will be the xenobiotic, the sink set can be composed of those metabolites in the central metabolism of a chassis strain and the rule set could comprise all known catabolic reactions. In the same vein, one can find sensing-enabling pathways with the set of known detectable compounds as sink, the set of target compounds to detect as source, and by using the forward rules (cf. 3.7.3 for the detection of biomarkers). Finally if

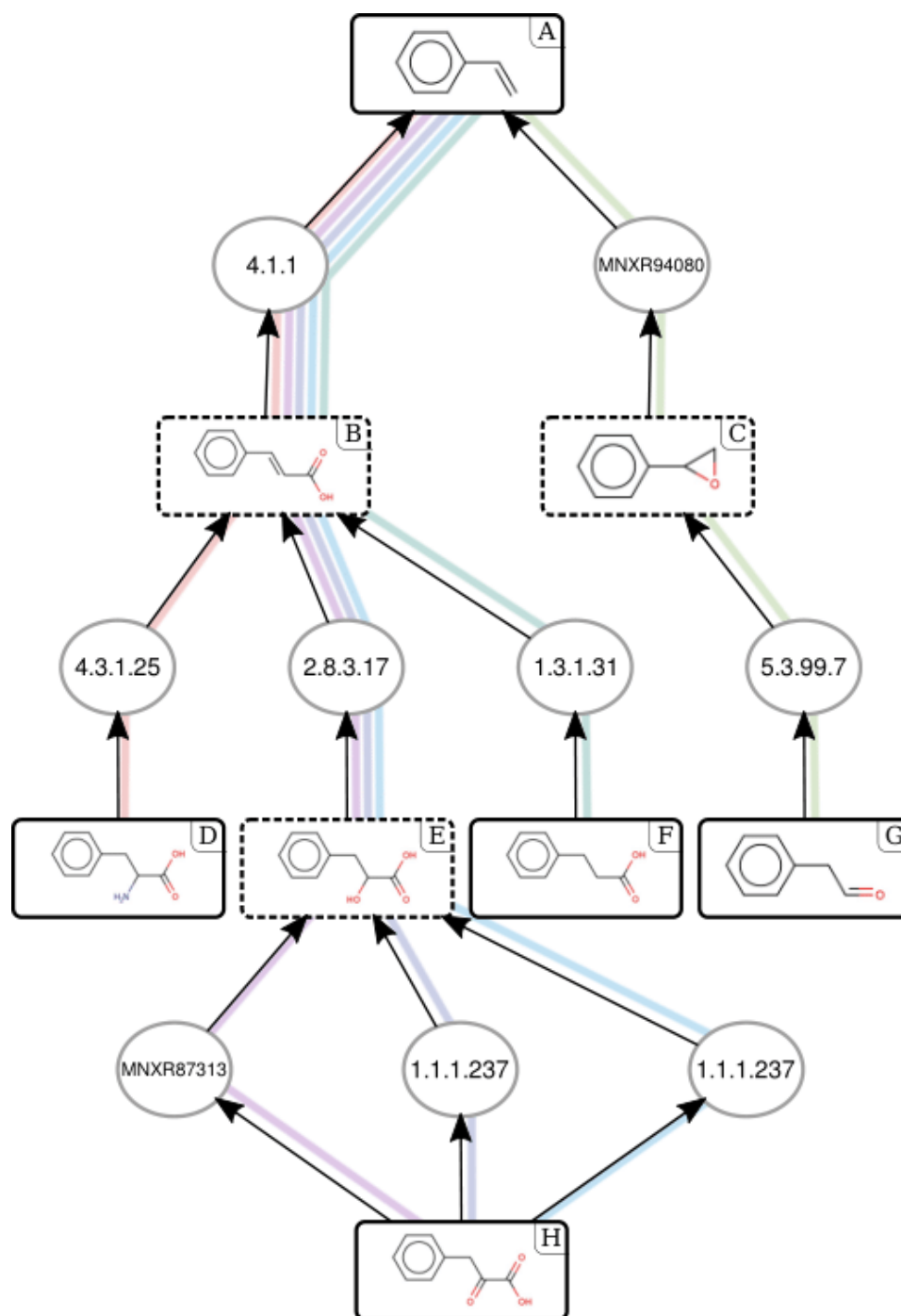


Fig. 3.8: Enumerated pathways for the production of styrene. Each pathway is depicted by a distinct colour. Pathway referenced in [190] is in red (D-B-A). Compounds are represented by their structures, and reactions by their EC numbers. Styrene and sink compounds are surrounded by a solid line, intermediates by a dashed line. A: styrene; B: phenylacrylic acid; C: styrene oxide; D: phenylalanine; E: 3-phenyllactic acid; F: 3-phenylpropionic acid; G: phenylacetaldehyde; H: phenylpyruvic acid. Cofactors have been removed for clarity; the whole scope is available in Supplementary.

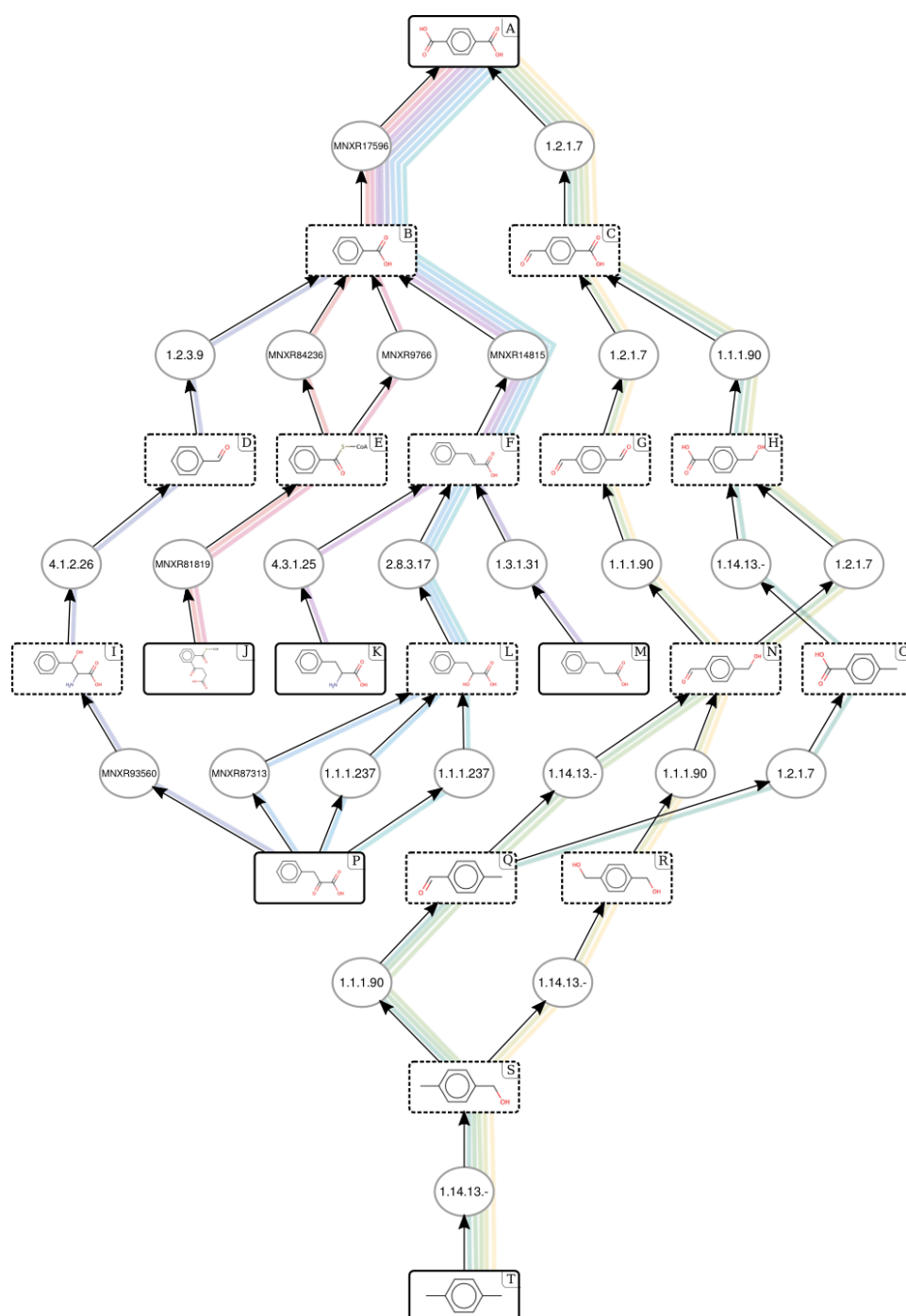


Fig. 3.9: Enumerated pathways for the production of the non-natural compound terephthalic acid (TPA, compound A) from *E. coli*. Each pathway is depicted by a distinct colour. Pathway referenced in [194] is in teal blue (T-S-Q-O-H-C-A). Compounds are represented by their structures, reactions by their EC numbers. TPA and sink compounds are surrounded by a solid line, intermediates by a dashed line. Reactions with unknown EC number according to MetaNetX are referenced by their MetaNetX ID. A: terephthalic acid; B: benzoic acid; C: 4-formylbenzoic acid; D: benzaldehyde; E: benzoyl-CoA; F: phenylacrylic acid; G: terephthalaldehyde; H: p-hydroxymethyl benzoic acid; I: 3-phenylserine; J: 2-succinylbenzoyl-CoA; K: phenylalanine; L: 3-phenyllactic acid; M: 3-phenylpropionic acid; N: 4-(hydroxymethyl)benzaldehyde; O: p-toluic acid; P: phenylpyruvic acid; Q: p-tolualdehyde; R: 1,4-benzenedimethanol; S: 4-methylbenzyl alcohol; T: p-xylene. Cofactors have been removed for clarity; the whole scope is available in Supplementary.

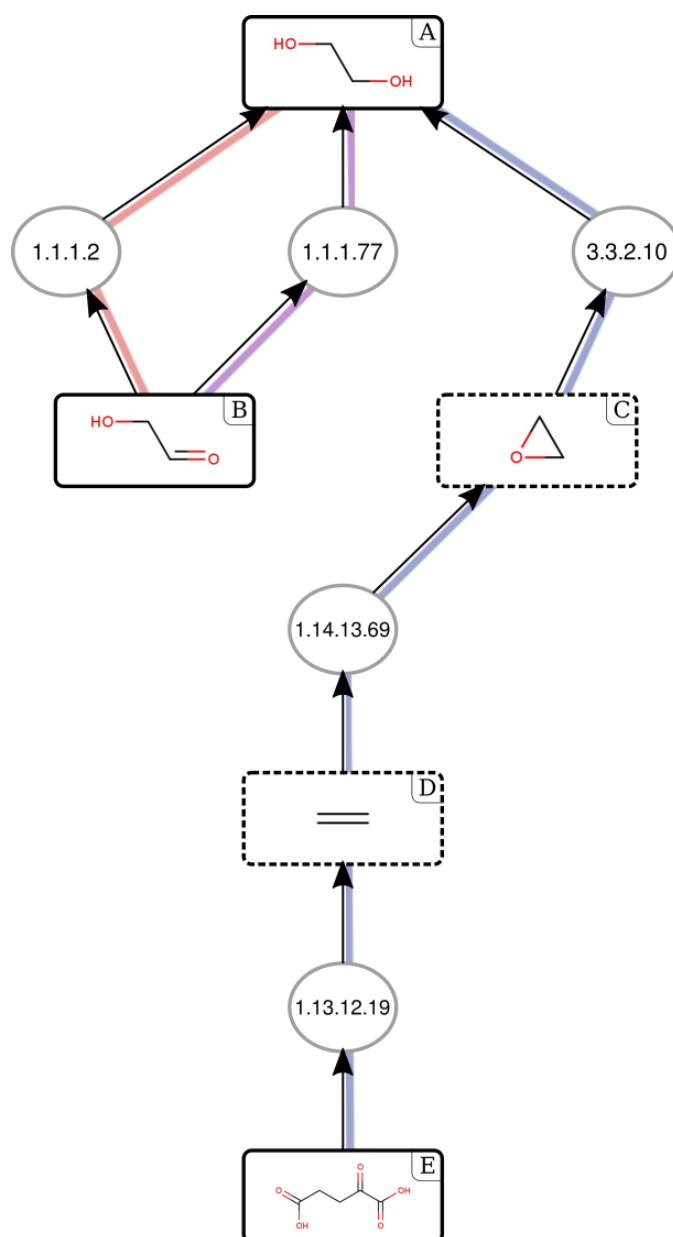
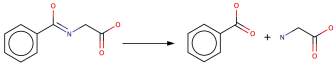
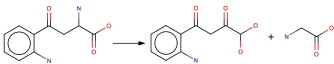
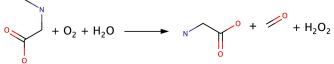

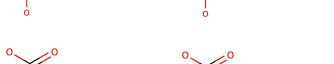
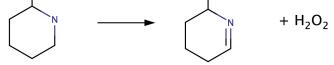
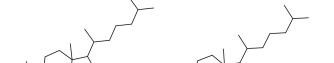
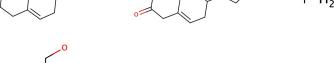


Fig. 3.10: Enumerated pathways for the production of ethylene glycol. Each pathway is depicted by a distinct color. The final step of the pathway engineered in [92] is in red (B, A). Enzymatic step 1.13.12.19 is also known as RXN-12538 (MetaCyc) and R09784 (Kegg). Compounds are represented by their structures, and reactions by their EC numbers. Ethylene glycol and sink compounds are surrounded by a solid line, others by a dashed line. Cofactors and currency metabolites (such as NADPH, NADH, water, proton, dioxygen) have been removed for clarity. This figure was originally in Supplementary. Involved compounds: ethylene glycol (A), glyceraldehyde (B), ethylene oxide (C), ethylene (D) and oxoglutarate (E).

Tab. 3.3: Examples of metabolic pathways enabling the detection of prostate cancer biomarkers. The fourth column indicates the sample type (Serum, Urine, Metastatic Tissue) and if the biomarker has been found to be up or down regulated compared to a controlled sample of the same type. References for enzymes were taken from MetaNetX and references for transcription factor were taken from the SensiPath server [153]. Figure 3.11 illustrates the scope of sarcosine. This table was originally in Supplementary.

Biomarker	Metabolic Reaction	Effector(s)	References: Sample (S), Enzyme (E), Transcription-Factor (TF)
Hippuric acid		Benzoic acid, Glycine	S: Urine down-regulated (Struck-Lewicka <i>et al.</i> , 2015) E: hippurate hydrolase (3.5.1.32) TF: benzoate BenR [196], glycine GcvR [197]
Kynurenine		Glycine	S: Serum, Urine, Tissue up-regulated [198] E: kynurenine-glyoxylate aminotransferase (2.6.1.63) TF: GcvR [197]
Sarcosine		Glycine, H2O2	S: Serum, Urine, Tissue up-regulated [198] E: sarcosine oxydase (1.5.3.1) TF: glycine GcvR [198], H2O2: OxyR [199]
N-acetyl-aspartate		H2O2	S: Serum, Urine, Tissue up-regulated [198] E: aspartate oxydase (1.4.3.16) TF: OxyR [199]
Pipecolate		H2O2	S: Serum, Urine, Tissue up-regulated [198] E: pipecolate oxydase (1.5.3.7) TF: OxyR [199]
Cholesterol		H2O2	S: Serum, Urine, Tissue down-regulated [198] E: cholesterol oxydase (1.1.3.6) TF: OxyR [199]
L-Sorbose		H2O2	S: Urine down-regulated [200] E: sorbose oxydase (1.1.3.13) TF: OxyR [199]
Creatinine		Urea, H2O2 and glycine via sarcosine	S: Serum, Urine, Tissue down-regulated [198] E: creatininase (1.1.3.13) followed by creatine amidinohydrolase (3.5.3.3) TF: urea NtcA [201], glycine GcvR [197], H2O2: OxyR [199]

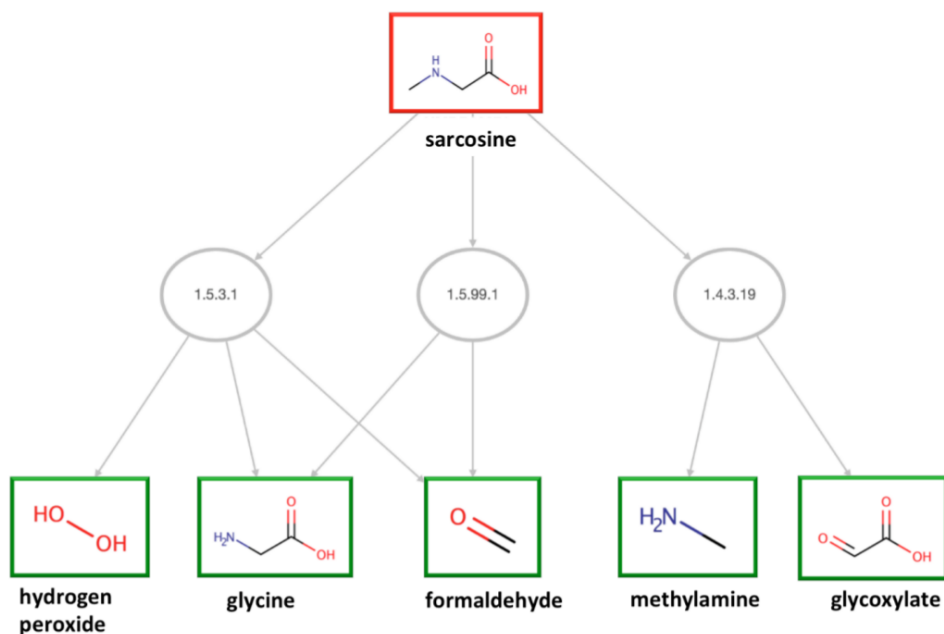


Fig. 3.11: Example of a scope obtained for sarcosine. All the products of the pathways are transcription factor effectors (i.e. in the sink file, green squares). Reactions are given by their EC numbers. This figure was originally in Supplementary.

one wishes to know all possible compounds that can be produced with a chassis strain when adding heterologous enzymes, the source set is composed of the metabolites of the chassis strain, the sink set can be either empty or a set of compounds in a vendor catalogue, and the rule set should cover all reactions that could occur in the chassis strain, including heterologous enzymatic ones. Moreover, any other applications where the problem can be framed into source, sink and rule sets can be processed by the workflow including problems where compounds are not metabolites and reactions are not metabolic reactions.

The most critical feature of a reaction network prediction system is certainly how the reactions are encoded and from where this knowledge was extracted. In our case, we choose to adopt a reaction encoding based on SMARTS, a widely accepted compound query language [59] that was already used successfully in such context [80]. Unlike most rule-based reaction prediction systems, RetroPath2.0 rules are not built around the Enzyme Commission nomenclature, but rather from an automatic translation of enzymatic reactions extracted from databases, which we believe offers a refined view of enzyme's capabilities.

We showed that our rules were able to classify reactions and that our set of rules extracted from MetaNetX had a good coverage over the known reactome. A good part of the reactions that were not covered were actually reactions involving compound classes (e.g. "an alcohol"), which were removed during the rule

generation steps. This type of generalized reactions were, in turn, represented in our set through our unique way of encoding reactions as generalized rules. One substantial improvement could probably be met by constraining the atom-atom mapping and reaction centre identification steps based on the exploitation of additional knowledge on the reaction and the associated enzyme. For instance by using the known alternative substrates associated to a single enzyme sequence or the EC number assignation.

Evaluating the coverage of a reaction database is interesting in order to assert the coverage of the known reactome by a given set of rules, but it cannot be used to assert the efficiency of a retrosynthesis tool. Indeed, the coverage of a reaction database depends mainly on the database from which the rules were inferred and how exhaustive the cross-links are between those two. Ideally, we would desire a set of rules being able to recover all known biochemical reactions. It means that anything less than 100% coverage shows that the set of rules is incomplete and that more data could have been aggregated. Note that in this work we focused our efforts on MetaNetX for the sake of simplicity but it is clear that more data can be imported from other databases such as BRENDA [51].

To the author's opinion, a better indicator of retrosynthesis tools efficiency should be found in the coverage of known pathways realized in a metabolic engineering context. This is precisely what we did using the LASER database as a reference for examples of successfully engineered metabolic pathways. In that way we provided a comprehensive overview of the capabilities of our tool in order to identify metabolic engineering solutions to bioproduction for well-studied cases. The main source of misprediction that we observed in our analysis came from cases in which additional compounds absent from *E. coli* metabolism were needed to perform the synthesis. Indeed, we performed all computations within five iterations from *E. coli*, with target compounds that were not necessarily produced in this chassis, nor at five enzymatic steps; moreover, some substrates could be supplemented in the media of the chassis organism. For instance, the synthesis of morphine is described for *Saccharomyces cerevisiae* in [202] by two pathways at three and four steps from thebaine. Thebaine is not naturally present in *E. coli* metabolism thus absent from the sink we used. Consequently, this example has no scope at five steps and was counted as mispredicted. Once thebaine is supplemented in the sink, RetroPath2.0 can generate a scope with both pathways. Note that thebaine was already predicted before being added to the sink, and that doing so only allowed RetroPath2.0 to use this compound as a valid starting point for synthesis instead of continuing further the retrosynthesis.

Importantly, not all predicted pathways can be readily implemented in *E. coli*. Indeed, translation of *in silico* models into *in vivo* experiments require much more

constraints to be satisfied, some of those being hardly predictable. To name but a few, enzyme sequence availability, chassis ability to fold the enzymes, kinetics, intermediate compounds toxicity, and overall pathway induced stress on the cell should all be checked before going any further. In this context, RetroPath2.0 can be seen as a base on which everyone is invited to build new features in order to further improve its metabolic space exploration abilities.

Exploiting chemical diversity in order to gain access to the large catalogue of natural and non-natural chemical resources is arguably one of the most important goals for biotechnology applications. By extending metabolic capabilities of enzymes, applications in metabolic engineering, biosensors and synthetic circuits can be greatly enlarged and diversified. To that end, RetroPath2.0 brings to the community a flexible and scalable open source platform with unique metabolic design capabilities. For the first time, we allow the systematic application of a full set of validated and standardized reaction rules that can be expressed with a selectable level of specificity. Such representation, which parallels the versatility of enzyme promiscuity, allows an in-depth exploration of latent abilities of natural enzymes.

The excellent coverage of the workflow along with its proved ability for recovering both known pathways and putative alternative candidate pathways show its power as an engineering tool. For that reason, we have no doubt that the tool will be received as a valuable addition to the toolbox for engineering biology. Moreover, community contributions to the workflow will likely expand further the features of the tool, even beyond metabolic design. In summary, we believe that the ability of RetroPath2.0 to rationalize and standardize design steps of biological engineering that have been traditionally performed manually by trial and error, constitutes a major contribution towards the development of automated workflows across the whole design, build, test and learn cycle.

3.7 Supporting information

3.7.1 Reaction rule diameter and promiscuity recovery

Reaction diameter is a parameter related to enzyme promiscuity. In this note, we investigate the ability of the reaction rules to recover promiscuity in function of the diameter. We have performed several tests on the reference list of promiscuous enzymes in *E. coli* [44] (see Section 3.5.1 of the manuscript).

Test 1. Does using a reaction diameter recover all the known examples of promiscuity for any enzymes?

We looked at the coverage of the rule for the total number of annotated reactions in the promiscuous enzymes. As shown in the Figure 3.12 (red line), at low diameter, the coverage is quite high (above 90%), i.e. most of the promiscuous reactions annotated for one gene are recovered by the rule. As the diameter increases the coverage decreases to 85%.

Test 2. What reaction diameter is best?

In order to answer this question, we should consider the specificity of the rules, i.e. how specific is a rule to the enzyme? Figure 3.12 shows the percentage of rules that appear annotated for a single enzyme depending on the diameter. Rules at low diameter are sometimes shared by more than one promiscuous enzyme (around 20% of the cases), whereas rules become more specific for one single enzyme at higher diameters ($d > 6$). Taking into account this result and the previous one about reaction recovery, we think that a good trade-off between coverage and specificity is for diameters between 6 and 8.

Test 3. If the reaction diameter is too small (too general) are false positives predicted?

Enumeration of reaction rules can generate a large list of pairs of substrate-products, especially at low diameters. Depending on each enzyme sequence and on the experimental conditions, the expressed enzyme might display more or less affinity toward the substrates and more or less level of efficiency for some of the reactions. Therefore some predicted reactions might be false positives depending on the selected enzyme sequence. For instance, phenylalanine ammonia-lyases (PAL) EC 4.3.1.24 often also show tyrosine ammonia-lyase (TAL) EC 4.3.1.23 activity. Both of them can be encoded through the same reaction rule at low diameters. Whether the selected enzyme will show PAL, TAL or both activities will depend on the chosen enzyme sequence. This is a problem that needs to be addressed through enzyme design rather than retrosynthesis.

Concerning general trends, our previous answer showed that at low diameters there are some rules that are shared by more than one enzyme. More precisely, at diameter $d = 4$, 62% of enzymes had reaction rules with no false positives, 74% of enzymes had at most one false positive (a reaction in the dataset predicted by the rule that is not annotated for the enzyme). At diameter $d = 8$, these percentages

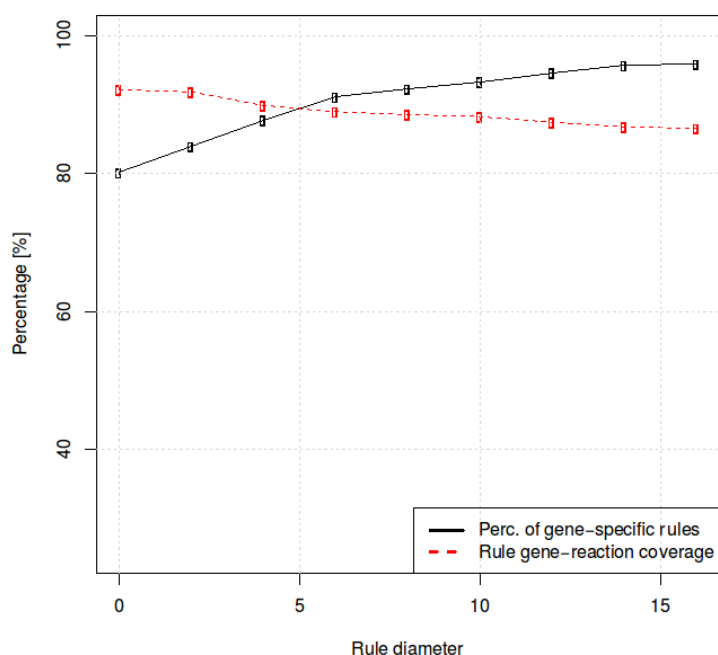


Fig. 3.12: Specificity of rules vs. promiscuity recovery in function of diameter. Rules specificity is calculated as percentage of gene-specific rules (solid black line), which represents the percentage of the total rules at a given diameter that are associated with a single gene in *Escherichia coli*. Promiscuity recovery is calculated as rule gene-reaction coverage (red dotted line), which represents the average maximum percentage of reactions associated with a gene that are covered by a rule at the given diameter. This figure was originally in Supplementary.

were of 78% and 88%, respectively. These results suggest again that a good choice for recovering promiscuity is a diameter around $d = 8$.

3.7.2 Details about multimolecular rules

Nonetheless, there are still many metabolic reactions that are multimolecular, even after removing currency cofactors. MetaNetX version 2.0 (110,000 compounds after canonicalization and 31,527 reactions) comprise 42% of reactions that remains multimolecular after removing currency cofactors. Metabolic databases such as MetaCyc or KEGG do identify main substrates and products, albeit not in all cases, and 29% of reactions in MetaCyc have multiple main substrates and 27% have multiple main product and 15% have both. A good example of such reactions are transaminases (EC class 2.6.1). There are 178 reactions in that class in MetaCyc, most of them involving two substrates and two products. Cofactors in 2.6.1 reactions are glutamate and oxoglutarate in 51% of the cases, but other cofactors are found such as oxaloacetate, 2-oxobutanoate, oxoglutarate, oxooctonal, glyoxylate, pyruvate, glutamine, oxosuccinamate, or butamine.

Clearly reactions of class 2.6.1 admit multiple substrates and products and they all vary from one reaction to another, thus these reactions cannot all be coded as monomolecular transformations ($R1C(=O)R2 \rightarrow R1C(NH_2)R2$) in the way that is done in [29] nor they can be coded in the form Glutamate + $R1C(=O)R2 \rightarrow$ Oxoglutarate + $R1C(NH_2)R2$, (where $R1$ and $R2$ can be C or H) as done in [161].

Another issue that retrosynthesis algorithms need to overcome is to handle multiple products when the reactions are reversed (as they should be in any retrosynthesis process). Indeed when a reaction proceeds forward one assumes that the substrates are readily available and this is generally the case when moving down through a metabolic pathway, where the substrates of any given reactions are the products of upstream reactions. In retrosynthesis, the products become the substrates of the reversed reaction, and these substrates are not necessarily known. To illustrate this issue let us consider the reversed reaction $T + P \rightarrow S$, where T is our retrosynthesis target. T is known but not P , in principle we should apply the rule to any compound P of the chemical universe (since P is not necessarily a known metabolite). This solution is of course not practical. To palliate this issue, RetroPath do not reverse multiproduct reactions but construct an extended metabolic space using reaction rules fired on the metabolites chassis strains. The other retrosynthesis algorithms do not explicitly address this issue, albeit mono product reactions can certainly be reversed, and in principle, all SimPheny rules can be used for retrosynthesis purposes.

As summarized in Table 3.1, SimPheny [29] does not deal with multiple substrates and products as all rules are monomolecular, BNICE in [161] handles partially the problem, as 70 out of 86 reactions are multimolecular but only 4 reactions have multiple substrates different than cofactors (compared to 35% in MetaNetX). GEM-Path and RetroPath work with rules handling multiple substrates and multiple products, thus reflecting better the complexity of metabolic reactions. Nonetheless, RetroPath allows only one substrate at a time to undergo promiscuity modelling so that reaction prediction remains tractable.

3.7.3 Detection of biomarkers through metabolic circuits

Asides from metabolic engineering, reaction network prediction algorithms can also be used to develop whole-cell biosensors. Typical synthetic biosensors [203] currently being developed comprise systems capable of sensing a small molecule generally through allosteric interactions with RNA aptamers (e.g. riboswitches) or transcription factors [114] that upon sensing will express a reporter gene. In the context of medical diagnostics based on biomarkers detection the main

advantages of synthetic cell-based technologies over abiotic detection based on purified antibodies, nucleic acid hybridization, or metabolomics analysis are lower cost, improved stability, and the possibility to be ultimately used as a personal home healthcare device.

However, as of today, typical whole-cell biosensors are triggered by no more than half a dozen input signals. To palliate this shortcomings, we have recently proposed a method to expand the range of biologically detectable biomarkers by systematically engineering sensing enabling metabolic pathways (SEMP) [153, 196], i.e., metabolic pathways that transform non-detectable chemicals into molecules for which sensors already exist. The SEMP method has been successfully benchmarked to engineer biosensors that detect pollutants, drugs and biomarkers such as benzoic acid and hippuric acid [196].

Here we investigate the use of RetroPath2.0 to search for all prostate cancer biomarkers that could potentially be detected using *E. coli* as a sensing device.

Prostate cancer biomarkers were retrieved from the Human Metabolome Database (HMDB) and scanned literature to select biomarkers in various physiological fluids: serum [198, 200, 204, 205], urine [198, 200, 206–208], and tissue [198, 200, 209, 210]. The above references gave a final list of about 800 small molecule biomarkers. Because engineered in *E. coli* we removed all *E. coli* native metabolites, we also removed duplicates and biomarkers that could not be found in HMDB because of ambiguous names. The resulting sanitized set was composed of 421 biomarkers (provided in supplementary materials).

RetroPath2.0 was run taking as source all (non-*E. coli*) prostate cancer biomarkers, and as sink a list of 500 effector molecules known to either activate or inhibit transcription factors (extracted from [153]). SEMP were generated by enumerating pathways linking source to sink in a single iteration by firing rules computed from MetaNetX (provided in supplementary materials).

Among the 421 biomarkers, we found 27 biomarkers directly detectable by transcription factors, and 415 pathways enabling the transformations of 164 different biomarkers into 76 different effectors. Some of these results are presented in Table 3.3.

Notable amongst the biosensor listed in Table 3.3 are H₂O₂ and glycine that are detectable by the native *E. coli* transcription factors OxyR and GcvR, respectively [197, 211], and benzoate for which biosensors have already been built in *E. coli* [196] (detailed results are provided in the supplementary materials). Interestingly several biomarkers could be transformed into the same effector, thus enabling

the integration of multiple biomarker signals into a unique detectable biosensor. Those results highlight the versatile use that a generic retrosynthesis and reaction network prediction algorithm such as RetroPath2.0 can have.

3.7.4 Description of supplementary data

Last version of the workflow can be downloaded at <https://www.myexperiment.org/workflows/4987.html>. The supplementary data are hosted by Metabolic Engineering.

RetroPath 2.0 workflow

RetroPath2.0 is provided as a KNIME workflow (*RetroPath2.0.knwf* file). KNIME and RetroPath2.0 installation procedures are described in *RetroPath2.0_installation.pdf*.

Tutorial & examples

The *RetroPath2.0_tutorial.pdf* file can be used as a starting point for beginning with RetroPath2.0. It is a step-by-step tutorial describing how to use the tool through basic retrosynthetic examples. Data for examples are in the *tutorial_data* folder.

Scope Viewer

The Scope Viewer is a modest tool dedicated to the visualization of scope files outputted by RetroPath2.0. It is available in the *scope_viewer* folder.

Set of SMARTS rules

Several sets of SMARTS rules are provided in the *data/rules* subfolder. Files *knime-ready-rules_mnx-all-forward_ECOLI-iJO1366.csv* and *knime-ready-rules_mnx-all-reverse_ECOLI-iJO1366.csv* are subsets of SMARTS rules we generated from reactions available in the MetaNetX database. Both subsets correspond to the *E. coli* metabolism for diameters 2 to 16 (see main text). The first set contains rules in the direct direction (*[...]-forward-[...].csv* file) while the second contains rules in the reverse direction (for retrosynthesis, *[...]-reverse-[...].csv* file). Extracted rules for BNICE (mono- and bi-substrate rules) and Sympheny are also provided. All provided sets of rules are ready to be used with RetroPath2.0.

Set of compounds from *E. coli* metabolism

The *data/ecoli-ijO1366-mnx-compounds.csv* file contains the list of compounds that we extracted from the *E. coli* iJO1366 whole-cell model and MetaNetX cross-references. See main text for details.

Coverage of bioproduction pathways

Generated results are provided in the *results/bioproduction_pathways* folder.

The *bioproduction_pathways/LASER_MBE* subfolder contains results regarding the compounds extracted from the LASER database and Metabolic Engineering papers published in 2016 (see main text for details). The *laser.csv* and *mbe.csv* files provide details on each compound (*input* subfolder). Scope (**_scope.json* file), enumerated pathways (**.png* and *pathways.csv* files), and structures of involved compounds (*chemical_structures.csv* file) are listed in the *pathways* subfolder for each tested compound that leads to at least one pathway.

The *bioproduction_pathways* folder does also contain the whole scope computed for styrene (*styrene_scope* subfolder) and terephthalic acid (*TPA* subfolder) that are respectively shown by figure 3.8 and 3.9 of the paper. These scopes can be generated again using the source (*source.csv*), sink (*sink.csv*) and rule (*rules.csv*) sets that belong to each folder.

Detection of biomarkers through metabolic circuits

Data are provided in the *results/detectable_biomarkers* folder. The source file is composed of prostate cancer metabolites. The sink file comprises effectors (small molecules) activating or inhibiting 5 transcription factors. Transcription factors are not provided but can easily be retrieved by entering the InChIs of the effectors in the SensiPath web server [153]. Table 3.3 in the main text was generated from results extracted in the file *result.csv* (*res* folder). The *.csv* and *.json* files attached to each biomarkers comprise a lower number of pathways than in the *result.csv* file, the reason is that in the former case pathways are enumerated only when all the products of the pathways are in the sink.

Part I: conclusion & perspectives

Part I summarizes the efforts that have been undertaken during this thesis to give biological engineers a clear understanding of the retrosynthesis tools they could use to design metabolic pathways.

In Chapter 2, we detailed a protocol to create an Extended Metabolic Space (EMS), i.e. a network of all the metabolites that can be synthesized or consumed by enzymatic reactions under a certain degree of enzymatic promiscuity. This approach has several original features that deserved to be highlighted.

First, we use a vector-based description to encode compounds (molecular signature) and reactions, and that has several interesting consequences. Reaction rules do not require the use of an Atom-Atom Mapping (AAM) step and are thus more robust to poor quality data (unequilibrated reactions, missing structures, etc.). Moreover, searching for a valid alternative substrate or applying a reaction are simple additions of vectors that are faster to execute than graph-editing operations [72].

Second, our protocol is essentially a two-steps process to create and *then* explore the metabolic space: (i) the *forward* enumeration gives the EMS as a graph linking chemicals (nodes) by reactions (edges); it can also be seen as an exhaustive database of predicted reactions, much like what has been done in MINEs [79] and ATLAS [162]. (ii) the EMS is transformed into a stoichiometry matrix representation that benefit from efficient search algorithms to perform the actual retrosynthesis analysis (*backward* search). However, this two-step process loses most of its interest if users are interested only in a bunch of pathways and not by the whole EMS. Indeed, a lot of computational power is lost during step (i) by wandering in pathways that may have no interest for the user and that may even be based on false promiscuity assumptions. Moreover, if for any reason a target compound is absent from the EMS (or disconnected) at step (ii), then finding a pathway would require the reconstruction of a whole new EMS under

relaxed conditions (such as a stronger enzymatic promiscuity hypothesis). This realization motivated us to develop a retrosynthesis tool that would be easier to adapt and that could be dedicated to end-users that are interested mostly in the study of a few pathways.

In Chapter 3, we presented a new version of RetroPath that is embedded in a third-party workflow management software (KNIME). As we discussed, the main interest of workflows is that they allow end-users to easily build on existing features, which ultimately extend the range of their applications. We have seen through several examples how RetroPath2 could be used to perform retrosynthesis analysis and more original applications such as the design of biosensing circuits, a subject that we will discuss in more details in Part II.

RetroPath2 has several important conceptual differences from its first version and the method we described in Chapter 2. The most remarkable change is without any doubt the utilization of SMARTS to encode reactions instead of molecular signatures. Molecular signatures have many advantages but are more difficult to handle than SMARTS for non-expert users [137]. Moreover, SMARTS are becoming a standard to encode reaction rules since they have been used successfully in the recent years to make metabolic pathway prediction by several independent groups [62, 64, 212]. However, SMARTS can suffer from some ambiguity while encoding resonance structures, stereochemistry, and organometallic bonds (in heme). Consequently, alternative ways to encode reactions (and compounds) are still under active development and are expected to be more flexible to describe chemical reactions [213].

Let us note here that several groups attempted to use machine learning to predict pathways. The main trend is to use machine learning conjointly with a set of generalist rules in order to learn *which* rule to apply on a chemical to find a suitable pathway. This has been shown in organic chemistry using Support Vector Machine classifier [75], Kernel Density Estimation [67], and neural network [214, 215]. In those cases, the use of machine learning is used to reduce the combinatorial complexity of the pathway search and not to predict reaction mechanism. On the contrary, Baldi's group tried to predict reaction mechanism using an orbital electron model to describe chemicals with a set of rules describing the flux of electrons that would occur during a reaction [63, 172]; again the machine learning is used to choose which rule to apply, but those rules are at a much finer grain than what we have discussed so far. It is still unclear what added-value machine learning has to offer in comparison to graph exploration heuristics for metabolic pathway prediction. Graph search algorithms have the merit to have their underlying choice mechanism explicitly written (and thus debatable) unlike machine learning models where the choices depend on the intricacies of the

learning dataset. Being able to backtrack and correct prediction errors is arguably an essential feature for expert systems, especially at a time when user trust comes at a high price.

Strangely enough, it seems that the use of computer aided pathway design is still rather limited, aside from collaborations with groups expert in this field. Computer-aided pathway design tools are more often cited by other CAD tools and reviews than actual users. This could indicate that user-demand is not met yet and that there is still place for improvements and research in retrosynthesis. The same phenomenon was observed in synthetic organic chemistry and was attributed to a lack of trust from end-user into both the software's results and its added-value capabilities compared to the work of a skillful chemist [58, 89, 216].

Obviously, the trust in CAD predictions should grow progressively with the accumulation of successful *in vivo* validations reported in the literature and by the industry. However, most pathway prediction softwares do not assess their predictions with new *in vivo* experiments and often propose the re-discovery of few examples of pathways that are naturals or that were reported by others. If those examples are necessary to show-case softwares capabilities they constitute by no means a serious statistical evaluation and could contribute to the lack of user trust. Following that idea, the interest of modeling enzymatic promiscuity to discover new enzymatic activities is rarely put to trial in *in vivo* experiments [29, 217, 218], and the same goes with the prediction of pathways efficiency.

This trust problem is not new in bioinformatics and has been tackled more than once, most notably in structural bioinformatics. For instance, the CASP (Critical Assessment of protein Structure Prediction) is a friendly competition that proposes every few years a series of challenges for structural bioinformatics groups interested into the modeling of proteins structures [219]. Recently discovered 3D structures are kept secrets while the primary sequences are sent to the competitors. To participate, they must provide a predicted 3D model that gets evaluated at the end of the competition with the release of experimental data. The same kind of "scientific Olympics" exist for the prediction of Protein Interactions (CAPRI) [220], gene Function Annotation (CAFA) [221], biological text-mining (BioCreative), sequence assembly (Assemblathon), and many more [222]. Instead of punctual assessments, others proposed a platform for a Continuous Automated Model EvaluatiOn (CAMEO) of methods related to the 3D structure, model quality, contact residue and ligand binding site using daily newly deposited data in the Protein Data Bank [223]. We could transpose the exact same concept of "community computational challenge" [224] to models in systems biology and metabolic engineering [225, 226] to assess the quality of pathway and enzymatic

promiscuity predictions tools. As a matter of fact, Salis's team that is responsible for the RBS calculator, a predictive model of translation initiation rate, recently embraced this idea of continuous tests by implementing the first automated test system for gene expression models [227]. We believe a critical assessment of reaction network prediction softwares would greatly benefit the field by providing a fair comparison of such expert systems, which would ultimately encourage users' trust in the results provided by those softwares.

Another trust issue comes from the fact that biological engineers are not only experts in their own field but also at searching information in the scientific literature (and on the Internet); tasks that are not so trivial for a software. Failing at reporting available information is obviously a problem for the efficiency of CAD programs but maybe more importantly it is also a breach in the trust of its users. Gathering information is a difficult task not only because databases are incomplete and not always up to date with the literature, but also because we assist nowadays to the multiplication of primary data sources. For instance, since 2012 four new databases have emerged to store xenobiotic degradation pathways (MetaPath [228], enviPath [229], XMetDB [230], MetaTox [231]), a task that was pioneered by UM-BBD in 2003 [232] and was last updated in 2010 [233]. Each of those databases may contain reactions that are not reported somewhere else and that might be essential for some application.

It is inspiring to see so many efforts to gather data about reactions, but on the other hand it is saddening that the community did not succeed yet to systematically deposit such data in a centralized repository under a recognized authority. A great example to follow would be the one of PubChem (NCBI), a database of chemicals that propose other databases to deposit their data as "Substances" and then proceed to canonicalize and cross-link them under a "Compound" identifier [234]. The result is that virtually all known chemicals are on PubChem and are easy to query both for humans and softwares, which makes the latter much more able to shine by their features instead of their capacity to import data from multiple sources. We could imagine a similar strategy for metabolic reactions or reaction rules using already well-established infrastructures such as Rhea, the reaction database from the European Bioinformatics Institute [235]. In the end, the creation of a centralized deposition service for reactions would reduce the cost of record curation efforts and favor the completeness of reaction data for CAD softwares. This would ultimately increase the quality of their predictions so that nothing a user could do by himself would be out of the capabilities of an automated software, hence contributing to build user trust.

Part II

Computer-aided design of
sensing-enabling metabolic pathways

Sensing new chemicals with bacterial transcription factors

This work was originally published in *Current Opinion in Microbiology* by Baudoin Delépine, Vincent Libis and Jean-Loup Faulon.

Only minor modifications have been introduced in the chapter presented below.

Full reference

Vincent Libis, Baudoin Delépine, and Jean-Loup Faulon. “Sensing new chemicals with bacterial transcription factors”. In: *Current Opinion in Microbiology* 33 (Oct. 2016), pp. 105–112. DOI: 10.1016/j.mib.2016.07.006

Contribution

V.L. and B.D. contributed equally to the preparation of the manuscript.

4.1 Highlights

- Transcription factors represent a starting material of choice to build new biosensors
- Evolution and rational design are making advances toward tailor-made transcriptions factors
- Metabolism indirectly expands the scope of chemicals that can be detected

4.2 Abstract

Bacteria rely on allosteric transcription factors (aTFs) to sense a wide range of chemicals. The variety of effectors has contributed in making aTFs the most used input system in synthetic biological circuits. Considering their enabling role in biotechnology, an important question concerns the size of the chemical space that can potentially be detected by these biosensors. From digging into the ever changing repertoire of natural regulatory circuits, to advances in aTF engineering, we review here different strategies that are pushing the boundaries of this chemical space. We also review natural and synthetic cases of indirect sensing, where aTFs work in combination with metabolism to enable detection of new molecules.

4.3 Introduction

Allosteric transcription factors (aTFs) responding to chemicals have an intimate link with the history of molecular biology since the unraveling of the *lac* operon. The chemical diversity that they can sense is now considered critical for the development of a wide range of biotech applications. In addition to being central in the standing efforts to develop biosensors monitoring environmental pollutants [112], aTFs are now enabling synthetic biology applications in the medical field as well [237, 238]. They are also becoming critical in metabolic engineering where they enable high-throughput screening of strains producing added-value chemicals [128, 239] and optimization of metabolic pathways through dynamic regulation of the expression of key enzymes [120]. Advances in these fields should accelerate as synthetic circuits are becoming more and more predictable, modular and scalable [240] but the limited diversity of chemical inputs is a bottleneck in the development of applications. The list of well-characterized aTFs is regularly increasing to tackle this issue [241] but the needs are not met yet. While synthetic biologists dream about tailor-made aTFs, such ability will require a deep understanding of how they work and evolve in nature.

Here we showcase recent approaches aiming at broadening the scope of chemicals that can induce aTFs. Several other reviews are available for more details concerning the applications of aTFs in biotechnology [109] or comparison of aTFs to other types of sensors [110].

4.4 Mining allosteric transcription factors in nature

In nature, aTFs evolved to recognize a multitude of chemicals such as antibiotics, primary or secondary metabolites, quorum sensing molecules, or man-made pollutants. Decades of effortful individual characterization regularly expanded the known frontiers of the chemical space within reach of these proteins. This knowledge is now organized and regularly updated in several databases where couples of aTF-effectors can be found [242–245]. BioNemo for instance compiles almost a hundred xenobiotic-responding aTFs that are the fruits of a very recent evolution as most of these molecules have appeared in the environment since the industrial chemistry era [244]. Many lessons about aTF families, mechanisms, and evolutionary potential have been learned on the way, thoroughly reviewed by De Lorenzo *et al.* [246] in the case of xenobiotic sensing.

Most of the natural sensing abilities remain probably uncharted as effectors are unknown for the majority of sequenced aTFs. A promising strategy consists in using comparative genomics to identify the binding sites of an aTF on the genome and thus predict its target genes which allow reconstruction of the regulon [247]. By assessing the function of the regulated metabolic pathway, known metabolites can be proposed as potential effectors for the aTF. This kind of approach has been demonstrated both with members of the LacI family [247] and of the ROK family [248].

With some aTFs, for instance those belonging to the large TetR-family, genetic context alone can be sufficient to make educated guesses about the effectors. Most aTFs from this family have their genes divergently oriented and less than 200 base pair away from the genes that they regulate [249]. Recently, predictions drawn from this feature have been coupled with phylogenomics to propose candidate effectors that were successfully validated experimentally. After building a relational tree organizing thousands of sequences from TetR-family members in subfamilies, Cuthbertson *et al.* [250] incorporated all known effectors on the tree and observed that similar effectors would cluster with the same subfamilies. Given the important diversity of effectors recognized by TetR- family members, the predictive aspect of such relational tree could be a powerful tool to identify new couples of aTF-effectors from the hundreds of thousands of TetR sequences available in databases.

A more direct way to dig into natural diversity is to use a screening method for substrate-induced gene expression (SIGEX). SIGEX begins with the systematic

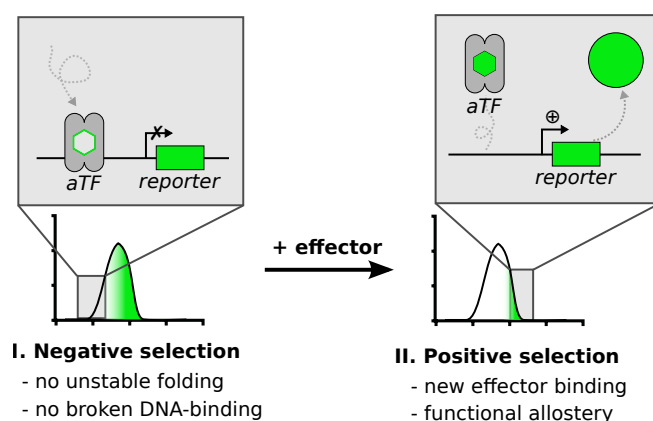


Fig. 4.1: Schema of the dual selection employed while screening allosteric transcription factor (aTF) variants. Top panels represent targeted molecular state in presence or absence of the investigated effector. The graphs represent the distribution of fluorescence intensity among a population of variants with the gates that must be used to select the cells exhibiting the appropriate response. The first step of selection (left) aims at eliminating aTF that lost the ability to repress their cognate promoter. Instead of fluorescence-based cell sorting, this step can also be replaced by a selection on the basis of the ability to repress expression of a toxic protein [252]. The second step (right) aims at selecting aTF induced in presence of the effector.

cloning of fragments from metagenomic libraries into an operon-trap vector containing a promoter-less fluorescent protein. Fragments containing an aTF and associated promoter can be isolated by fluorescence-based screening in presence of potential effectors. This approach regularly yields new characterized aTFs by screening environmental DNA extracted from contaminated areas [251]. Because the chemical space recognized by nature is rapidly evolving at the contact of the human civilization, these approaches tapping directly into this rich diversity will probably continue to be a first line solution to the lack of biosensors.

4.5 Engineering the sensing scope

The regulatory function of aTFs makes them suitable for high-throughput screening protocols aiming to isolate variants induced by new chemicals. Importantly, aTF redesign experiments necessitate a dual screening strategy to avoid selecting inactivated or non-inducible mutants (Figure 4.1). When structures of the aTF are available (ideally co-crystallized with effectors) to guide the process, such screenings coupled with mutagenesis methods can be sufficient to isolate aTF able to recognize new chemicals [253–255].

While aTFs are composed of modules such as ligand-binding domain (LBD), DNA-binding domain (DBD) and linker domain, an important difficulty to engineer them resides in their interdependency to ensure allostery of the protein [256].


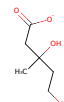

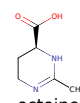
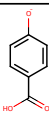
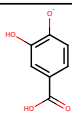
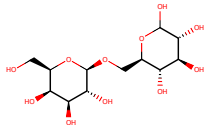
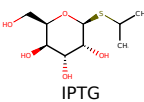
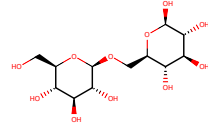
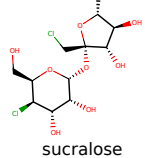
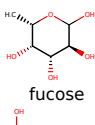

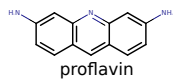
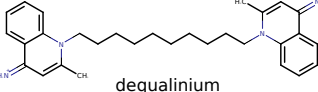
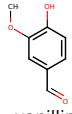
aTF	Previously known effectors	Adopted strategy	Target effector	Reference
AraC	 L-arabinose	<ul style="list-style-type: none"> Site-directed saturation mutagenesis 	 mevalonate	251
			 triacetic acid lactone	252
			 ectoine	253
PobR (IclR family)	 4-dihydroxybenzoate	<ul style="list-style-type: none"> Computational design (Rosetta), then "natural drift". 	 3-4-dihydroxybenzoate	258
LacI	 allolactose  IPTG	Mixed: <ul style="list-style-type: none"> computational design (Rosetta), protein-wide single-amino-acid saturation mutagenesis, error-prone PCR, and activity maturation. 	 gentiobiose  sucralose  fucose  lactitol	250
QacR (TetR-family)	 proflavin  dequalinium	<ul style="list-style-type: none"> <i>In Silico</i> sequence generation, then cell-free prototyping. 	 vanillin	261

Fig. 4.2: Recent examples and chemical structures of new effectors yielded by various aTF redesign strategies.

Changes in conformation induced by an effector must propagate successfully in the protein to change its affinity to DNA and this propagation can easily be disrupted by mutations. Counterintuitive lessons have been learned, for instance through the observation that mutations outside the LBD could lead to important changes in effector specificity. Nonetheless, allostery can be preserved through careful engineering guided by expert knowledge. For example, shuffling LBD and DBD from members of the same family have been demonstrated with the LacI/GalR family [257, 258]. This approach could yield more functional chimeras since a new database containing structural and mutagenesis data is now available to help predicting allosteric positions in this family [259]. Hopefully, the knowledge gained about maintenance of allostery in such chimeras will help engineering LBD recognizing new chemicals in a modular way.

However, in most cases, the wiring diagram of allostery is not known and an aTF must be considered as a fragile whole. Consequently, mutagenesis efforts should not be targeted at subdomains but ideally cover the whole sequence. This put a major limit on the chances of fruitful explorations since mutagenesis campaigns are bound by the transformation efficiency of bacteria (10^8). In this context, computational design tools can be used to reduce the design space. By helping the identification of amino acids playing roles in binding, specificity or allostery, these tools can define priority targets on which mutagenesis efforts should be focused. Several recent successes suggest that these approaches outperform random mutagenesis at altering effector specificity.

For instance, despite lacking an experimentally determined structure, Jha *et al.* [260] used structural bioinformatics tools to model the aTF PobR, and performed docking experiments that identified 16 important residues in the predicted LBD. The authors then used a "natural drift" heuristic in which they restricted themselves to one mutation by identified codon, and used the BLOSUM62 substitution matrix to keep only the most naturally observed substitutions. After flow cytometry screening, they found several hits with gain of function for induction by their target chemical. These variants had around ten mutations compared to the native sequence, covering most of the predicted positions.

Taylor *et al.* [252] proposed an end-to-end workflow for redesigning aTF specificity and compared several strategy to explore variants space, including computational design, random and targeted mutagenesis. They notably took advantage of a full-length single amino-acid substitution scanning of the LacI aTF to discover positions away from the LBD involved in allostery and specificity. Protein design approaches leveraging such mutational scanning data are considered to have an important potential [256, 261, 262]. The authors challenged their computational design abilities by developing biosensors for four chemicals with

Tab. 4.1: Examples of indirect sensing via metabolism.

Host	Sensed compound	Intermediates	Regulator (organism)	Ref.
<i>E. coli</i> (natural)	Lactose	Allolactose	LacI	[264]
<i>Thauera aromatica</i> (natural)	Toluene	Benzylsuccinate	TutBC	[265]
<i>Mycobacterium tuberculosis</i> (natural)	Cholesterol	Cholest-4-en-3-one; 3-oxocholest-4-en-26-oyl-CoA	KstR	[266]
<i>Mycobacterium smegmatis</i> (natural)	Cholesterol	Cholest-4-en-3-one; 3-oxo-4-cholestenoic acid	KstR	[267]
<i>Paracoccus</i> sp. (natural)	L-Gluconate	L-5-Ketogluconate; D-Idonate	LgnR	[268]
<i>Azoarcus</i> sp. (natural)	3-Methylbenzoate	3-Methylbenzoyl-CoA	MbdR	[269]
<i>Sphingobium</i> sp. (natural)	Ferulate	Feruloyl-CoA	FerR	[270]
<i>Rhodopseudomonas palustris</i>	p-Coumarate	p-Coumaroyl-CoA	CouR	[271]
<i>Comamonas testosteroni</i> (natural)	Benzoic acid	Benzoyl-CoA	GenR	[272]
<i>Thermus thermophilus</i> (natural)	Phenylacetic acid	Phenylacetyl-CoA	PaaR	[273]
<i>E. coli</i> (synthetic)	Salicylaldehyde	Salicylate	NahR (<i>Pseudomonas</i>)	[274]
<i>E. coli</i> (synthetic)	Lindane	1,2,4-Trichlorobenzene	XylR (<i>Pseudomonas putida</i>)	[130]
<i>E. coli</i> (synthetic)	2-Chloro-4-nitrophenol	Chloro-1,4-benzoquinone; chlorohydroquinone	LinR (<i>Sphingomonas paucimobilis</i>)	[196]
<i>E. coli</i> (synthetic)	Cocaine	Benzoic acid	BenR (<i>Pseudomonas putida</i>)	[196]
<i>E. coli</i> (synthetic)	Parathion	4-Nitrophenol	DmpR (<i>Pseudomonas</i> sp.)	[196]
<i>E. coli</i> (synthetic)	Hippuric acid	Benzoic acid	BenR (<i>Pseudomonas putida</i>)	[196]
<i>E. coli</i> (synthetic)	3-Hydroxypropionate(3HP)	3HP-CoA; acrylyl-CoA; propionyl-CoA; 2-methylcitrate	PrpR (<i>E. coli</i>)	[275]
<i>E. coli</i> (synthetic)	3-Hydroxypropionate(3HP)	3HP-CoA; acrylyl-CoA; acrylate	AcuR (<i>Rhodobacter sphaeoides</i>)	[275]

increasingly dissimilar structures from the natural LacI effectors. Interestingly, they stressed the importance of a last step of ‘activity maturation’ consisting of making chimeras of good variants in terms of induction with good variants in terms of narrow specificity. This strategy allowed them to lose the natural induction by IPTG while maintaining the newly developed induction by their target chemicals.

An ambitious approach consists in performing part of the screening *in silico*. This is a daunting task and expert knowledge of tools initially developed in other areas such as drug design is necessary. One difficulty is that straightforward docking of effectors on *in silico* variants to predict sequences with increased affinity is usually futile. Indeed, affinity does not necessarily correlate with induction abilities, and lower affinity can translate in higher induction abilities. Despite these difficulties, De Los Santos *et al.* [263] recently used several co-crystallized structures of the well characterized aTF QacR as starting point and performed an *in silico* screening to modify QacR specificity to recognize vanillin. They tested the best predictions with a cell-free screening protocol and obtained a few positive hits. Remarkably, the structure of vanillin is completely different from the known effectors of QacR (Figure 4.2). Although the cell-free *in vitro* screening selected aTFs variants that turned out to be toxic *in vivo*, this framework could become an important source of new biosensors in the coming years if it proves adaptable to other aTFs and chemicals.

4.6 Regulation via indirect sensing

Both synthetic biologists and natural organisms can take advantage of indirect sensing through metabolism to expand the chemical space detectable by aTF. Chemicals that can be transformed in situ by enzymes into aTF effectors are able to induce a response at the genetic level (Figure 4.3). Therefore, filling a need for a biosensor toward a particular chemical can sometimes be done

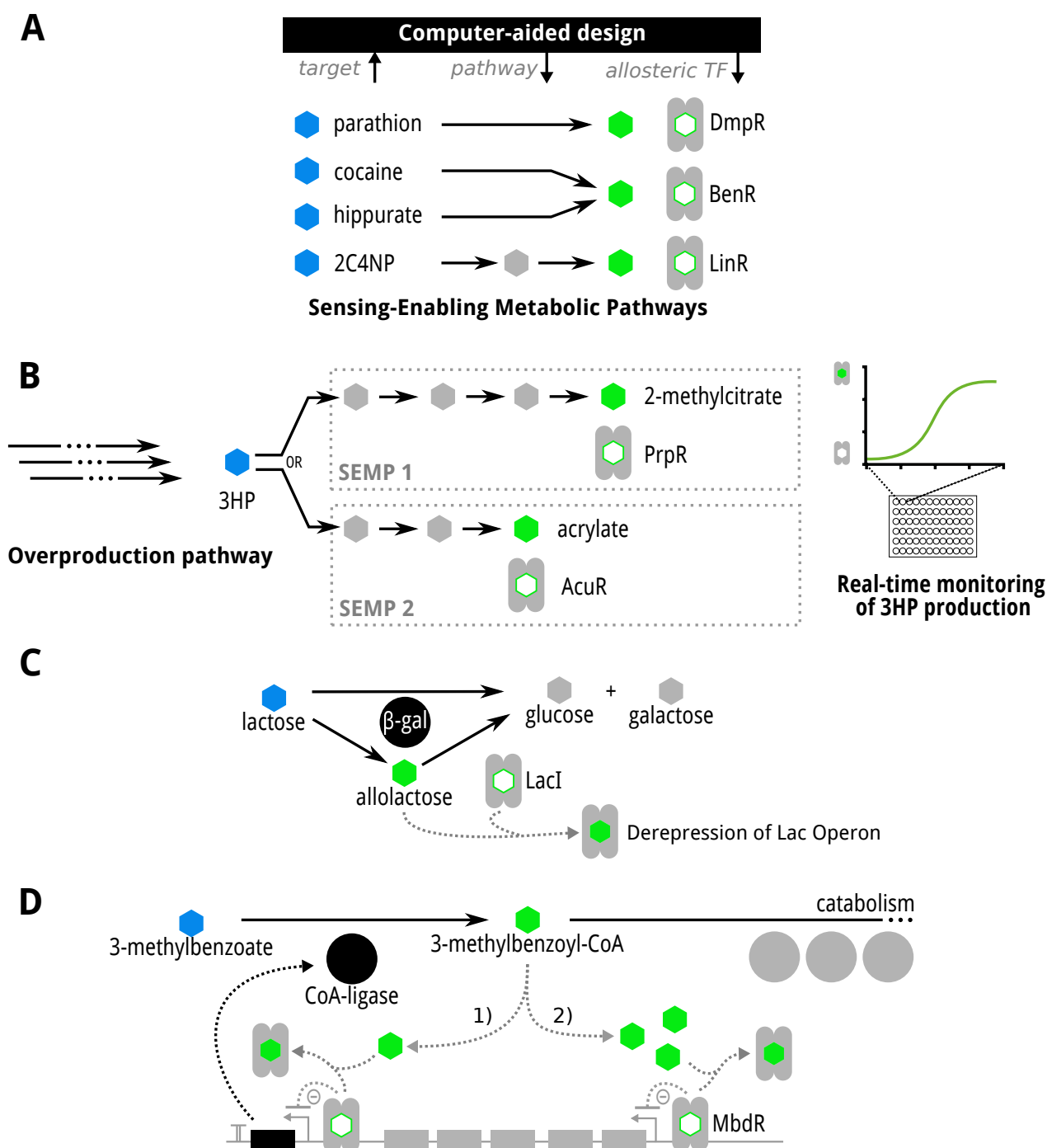


Fig. 4.3: Examples of synthetic and natural sensing-enabling metabolic pathways (SEMP). **A.** Computer-aided design can leverage biochemical reaction databases to construct synthetic routes that transform a chemical of choice into a known allosteric transcription factor (aTF) effector. **B.** In addition to the enzymes required for production of 3-hydroxypropionate (3HP), enzymes able to transform 3HP into detectable molecules were implemented in *E. coli*. These two alternative SEMP allowed monitoring of 3HP via two different aTF. SEMP 2 exhibited the best response to 3HP and was therefore used to help the high-throughput optimization of 3HP bioproduction process. **C.** The *lac* operon necessitates the secondary activity of β -galactosidase to transform lactose into the natural effector of LacI, allolactose. **D.** The specific regulation of the anaerobic degradation pathway of 3-methylbenzoate in *Azoarcus* sp. (1) The induction of the operon starts with derepression of the nested promoter controlling specifically the CoA-ligase which in turn increase the pool of the effector 3-methylbenzoyl-CoA. (2) When the effector reach higher concentrations, the rest of the operon is transcribed and the catabolism is initiated.

more simply by chemically modifying the target molecule than by engineering the specificity of aTF. Thanks to the immense reservoir of known biochemical reactions, many sensing-enabling metabolic pathways (SEMP) can be designed to transform nondetectable molecules into detectable ones. Engineered bacteria harboring such proper combination of enzymes and transcription factors have seen their sensing abilities extended to previously non-detectable molecules such as lindane [130], salicylaldehyde [274] or cocaine [196] to name a few. In the context of metabolic engineering, this strategy has recently enabled the monitoring of 3-hydroxypropionate (3HP) to speed up the development of a biobased production process. Because of the lack of aTF able to sense 3HP, Rogers and Church [275] developed two alternative SEMPs connected to PrpR or AcuR that allowed combinatorial testing of fermentation conditions (Figure 4.3 b). Systematic design of SEMP has been achieved by leveraging metabolic pathway databases, databases of known aTF effectors and retrosynthetic pathway design tools, leading to the development of 5 new whole cell biosensors [196]. A webserver is now available to the community to predict SEMP for chemicals on demand [153].

In nature, a growing number of SEMP-like motifs have been recently reported (Table 4.1). The reasons why indirect sensing is selected by evolution in some cases are not understood yet. The lac operon for instance is induced indirectly in presence of lactose only after its transformation into allolactose by a secondary activity of β -galactosidase (Figure 4.3 c). This secondary activity and the LacI aTF have been shown to be strictly co-selected [276] thus raising the question of the advantage of this SEMP motif over an evolution of LacI toward direct sensing of lactose. Another example of regulation motif with occurrence among several bacteria able to detect aromatics consists in using aromatic CoA-thioesters as effectors. In the case of 3-methylbenzoate sensing by *Azoarcus* sp. this SEMP has been well characterized and shine some lights on the type of effects these motifs can have on the genetic response profile. Juárez *et al.* [269] observed that the CoA-ligase producing the bona fide effector molecule is under the control of two promoters, one controlling the whole multi-enzyme degradation operon and a second one which is nested in the operon to add a second layer of control specifically on this enzyme. Both promoters are equally essential for growth in 3-methylbenzoate and they are both regulated by the same aTF, MbdR, but their induction kinetics and dynamic range appear different. When put in presence of 3-methylbenzoate, the expression of the CoA-ligase is first induced moderately without the rest of the operon. This leads to an increase of the 3-methylbenzoyl-CoA pool in a positive feedback fashion, ultimately triggering the expression of the whole degradation operon through induction of its main promoter (Figure 4.3 d). Positive feedbacks are usually associated with interesting effects in response to signal and such exchanges between the metabolic and the genetic layers sup-

port the growing view of metabolism taking part in the computations performed by cells [277]. How much SEMP motifs actually play a role in computing signals is unclear. The reflexion could be fueled by studies demonstrating that rich biological behaviors such as linearization of dose-response can emerge at the metabolic level without requiring genetic circuits [278]. Understanding the effects of SEMP motifs could even shed light on regulatory circuits essential in pathogenesis of bacteria such as *Mycobacterium tuberculosis* where the sensing of cholesterol is also mediated by several enzymatic steps [266, 267].

4.7 Conclusion

The methods described in this review represent promising progress toward "on demand" aTFs, however, they still require important expertise and efforts. Because the range of applications extends over many disciplines, one issue is that laboratories with such expertise in aTFs are often not the same as laboratories facing the needs to detect new chemicals and they do not necessarily know on which relevant target molecules to focus their efforts. An era of synthetic biology was therefore dominated by applications inspired by the availability of bioparts. A new paradigm is emerging where applications are instead inspired directly by problems but this will necessitate minimizing the needs for expert knowledge through standardized methods of biosensor development.

While different methods may lead to functional biosensors for the same chemical, the overall cost and efforts involved can vary considerably. It is therefore recommended to wisely choose the most appropriate method considering each new target chemical. If an aTF detecting a structurally similar molecule can be found in databases, evolution protocols are probably the fastest approach. In other cases, the most appropriate approach can be determined by considering the nature of the chemical, itself being usually related to the field of application. For example, environmental monitoring applications typically need biosensors for xenobiotics that are already in contact with microbial communities in the environment. Mining exposed metagenomes seems therefore more straightforward than engineering approaches. Because aTFs may evolve subsequently to enzymes, indirect sensing through metabolism is an alternative when environmental exposure is too recent for aTFs to be found. For instance, Product-Induced Gene EXpression (PIGEX) approaches using aTFs detecting likely metabolites from the target chemical can yield new enzymes from metagenomes and create SEMP biosensors at the same time [279–281].

A substantial part of metabolic engineering efforts is focused on bioproduction of natural products of secondary metabolisms. Biosensors enabling high-throughput screening of strains have chances to be immediately identified by mining microbiomes in contact with these products [282]. On the contrary metabolic engineering for xenobiotic production might require development of biosensor *de novo* through more sophisticated engineering approaches. Luckily the field of metabolic engineering is motivated by the promise of short term industrial applications, and is therefore becoming the biggest contributor of new biosensors. It is likely that approaches to develop tailor-made biosensors will be further improved and standardized through this driving force and will in turn benefit to other fields.

Some fundamental knowledge may also be gained on the way, as in the case of indirect sensing via SEMP. Whether the role of SEMP is simply limited to sensing in nature, or whether metabolism enables some level of computation will require further investigation but synthetic biologists will certainly follow such discoveries if additional design principles can be unraveled.

SensiPath: computer-aided design of Sensing-enabling metabolic Pathways

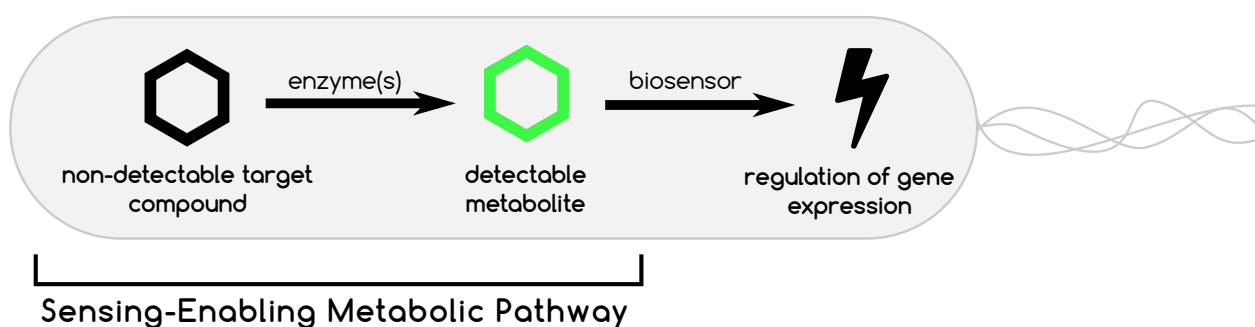


Fig. 5.1: Concept of sensing-enabling metabolic pathway.

This work was originally published in *Nucleic Acids Research* web server issue and authored by Baudoin Delépine, Vincent Libis, Pablo Carbonell and Jean-Loup Faulon.

Only minor modifications have been introduced in the chapter below.

Full reference

Baudoin Delépine, Vincent Libis, Pablo Carbonell, and Jean-Loup Faulon. “SensiPath: computer-aided design of sensing-enabling metabolic pathways”. In: *Nucleic Acids Research* 44 (W1 July 8, 2016), W226–231. DOI: 10.1093/nar/gkw305

Contribution

B.D. lead the project and was assisted by V.L. for selecting the use-cases presented in the manuscript and providing essentials advises as an alpha tester. The core of SensiPath is a rewrite by B.D. of the algorithm presented in [71] which was motivated by the desire to overcome technical limitations that were hindering the implementation of new features (on-the-fly computation, use of stereochem-

istry, single-compound query). All authors contributed to the preparation of the manuscript.

5.1 Abstract

Genetically-encoded biosensors offer a wide range of opportunities to develop advanced synthetic biology applications. Circuits with the ability of detecting and quantifying intracellular amounts of a compound of interest are central to whole-cell biosensors design for medical and environmental applications, and they also constitute essential parts for the selection and regulation of high-producer strains in metabolic engineering. However, the number of compounds that can be detected through natural mechanisms, like allosteric transcription factors, is limited; expanding the set of detectable compounds is therefore highly desirable. Here, we present the SensiPath web server, accessible at <http://sensipath.micalis.fr>. SensiPath implements a strategy to enlarge the set of detectable compounds by screening for multi-step enzymatic transformations converting non-detectable compounds into detectable ones. The SensiPath approach is based on the encoding of reactions through signature descriptors to explore sensing-enabling metabolic pathways, which are putative biochemical transformations of the target compound leading to known effectors of transcription factors. In that way, SensiPath enlarges the design space by broadening the potential use of biosensors in synthetic biology applications.

5.2 Introduction

Synthetic biology and metabolic engineering applications often require as part of their design a way to assess the presence or to quantify the amount of a compound of interest. Genetically-encoded biosensors such as riboswitches and allosteric transcription factors offer the possibility to control the expression of a gene of choice. This feature makes them valuable for many applications [109, 110] such as pollutant monitoring or high-throughput screening of optimized strains and enzymes [283–285], as expression of reporter genes like fluorescent proteins can be linked to the concentration of the compound of interest. Moreover, the ability of these biosensors to provide input at the genetic level opens the way to more complex downstream signal processing and actuation [286]. Examples of applications of such circuits range from threshold activation in presence of pathological concentration levels of biomarkers [287] to the creation of a feedback control motif leading to yield improvement for a chemical producing strain [120].

There is thus a critical need for biosensors, but it appears that current strategies for finding new biosensors may not be sufficient to answer all the needs. Although remarkable progress has been made in the field of genetically encoded biosensor design [252, 260, 288] and genome mining [289], the number of chemicals that can be detected is still limited and thus constitute a bottleneck in the development of synthetic biology applications.

New strategies of biosensing can be considered to tackle this issue. One of them relies on indirect sensing by transforming the molecule of interest into a detectable one. Such strategy has been successfully used with the help of enzymes to transform a key metabolite such as L-tyrosine [290] or L-DOPA [291] into pigments and thus allowing high-throughput screening of overproducers. The same strategy can also be employed to transform the molecule of interest into a molecule for which a genetically-encoded biosensor is available [130, 275]. We recently demonstrated that this approach could be attempted in a systematic fashion by combining information on the available biosensors and automatic design of enzymatic networks. This led to the development of five new whole-cell biosensors for pollutants (parathion, 2C4NP), biomarker (hippuric acid) and drugs (cocain, nitroglycerin) [196].

In order to open this untapped source of biosensors for synthetic biologists, we hereby present SensiPath (<http://sensipath.micalis.fr>), a web-based tool assisting the design of sensing-enabling metabolic pathways (SEMPs). SensiPath will serve users wishing to perform cell-mediated detection of a compound when no direct-sensing solution is feasible. The primary objective of SensiPath, thus, is to enlarge the number of detectable compounds for synthetic biology applications. The algorithms we implemented to simulate biochemical reactions are derived from the well- tested RetroPath [71]. It notably allows to take advantage of enzymatic promiscuity, i.e. the ability that enzymes have to process structurally similar substrates, thus yielding more results. SensiPath is built from a comprehensive list of more than 100 000 compounds and 87 000 reactions from four metabolic databases, covering most of the known metabolism. We also collected a large dataset of more than 500 detectable compounds for which intracellular biosensors exist from several gene expression regulation databases, focusing our search on allosteric transcription factors.

5.3 Material and Methods

Figure 5.2 shows an overview of how SensiPath works, the details are exposed in the following subsections. SensiPath is based on a comprehensive internal

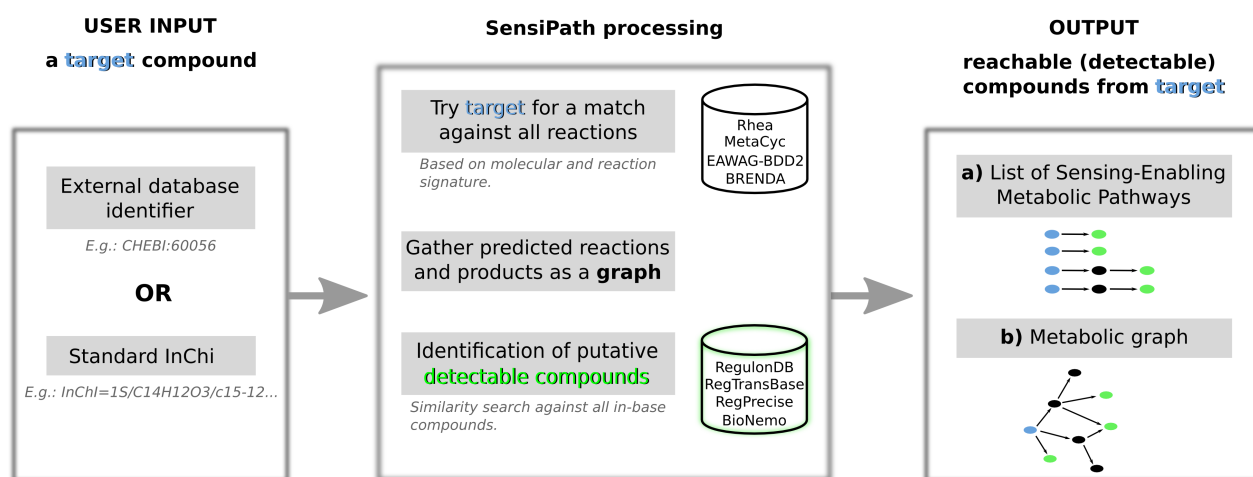


Fig. 5.2: SensiPath flowchart. Users query a target compound (blue), either as an external database identifier or as a standard InChI. Target is processed to gather reachable compounds through enzymatic reactions; detectable compounds are annotated (green). Results are displayed both as (a) the set of pathways leading to recognized detectable compounds; and (b) the whole computed graph around the target.

database of biochemical reactions and compounds encoded as chemical signatures. Once a compound query is submitted, it performs a search in order to find a match against all the enzymatic reactions that we have collected in our database. The search is carried out in order to predict reachable compounds from the target. This search generates a metabolic graph at up to two enzymatic steps away from the target, in which nodes are compounds and edges are reactions. Detectable compounds are identified and annotated by a score of similarity based on searching against the list of known detectable compounds in the database. For later reference, all SensiPath sources in its current online version are available on FigShare (<https://dx.doi.org/10.6084/m9.figshare.3144616.v1>) in addition of our list of detectable compounds (<https://dx.doi.org/10.6084/m9.figshare.3144715.v1>).

5.3.1 Source databases

SensiPath predictions are based on imported data from metabolic and gene expression regulation databases. We gathered data from multiple sources to cover most of available knowledge in current databases.

Reactions

Known biochemical reactions were extracted from main common reaction databases (Rhea (v66, <http://www.rhea-db.org>) [292], MetaCyc (v19.1, <http://metacyc.org>) [293], BRENDA (v15.2, <http://www.brenda-enzymes.info>) [51] as well

as from a more specialized database, the Biocatalysis/Biodegradation Database (<http://eawag-bbd.ethz.ch>, accessed in December 2015) [233]. We considered only reactions for which structures of all reactants were available, fully defined and valid. Overall, we collected more than 100 000 compounds and 87 000 reactions with references to external databases.

Detectable compounds

We gathered a list of 504 putative detectable compounds focusing our search on effectors of allosteric transcription factors from prokaryotes. Data were collected from several gene expression regulation databases: RegulonDB (v9.0, <http://regulondb.ccg.unam.mx>) [294], RegPrecise (v4, <http://regprecise.lbl.gov>) [243], RegTransBase (v7, <http://regtransbase.lbl.gov>) [242] and BioNemo (v6.0, <http://bionemo.bioinfo.cnio.es>) [244].

5.3.2 Reaction and compound encoding

In order to encode the reactions we first normalized the compounds, next computed molecular signatures and finally computed reaction signatures.

Compound normalization

The representation of compounds must be normalized in order to improve the performance of the encoding method. In particular, compounds were represented under their aromatic form while charges and hydrogens were removed; stereochemistry was kept.

Molecular signature

All compounds were encoded internally through their molecular signature [137]. The molecular signature of a compound is a list of overlapping molecular fragments, each of them centred on a distinct atom. Thus, fragments represent atom neighbourhood (also called atomic signature or atomic environment) in terms of atom and bond type. Basically, a molecular signature is similar to the extended connectivity circular fingerprint (ECFP) [74]. We used fragments (atomic signatures) with an environment diameter of 12 bonds.

Reaction signature

All biochemical reactions were represented internally by reaction signatures [177]. The reaction signature $\sigma(R)$ is defined in a vector space as the sum of molecular signatures of products less the sum of molecular signatures of substrates:

$${}^d\sigma(R_n) = \sum_i {}^d\sigma(P_i) - \sum_j {}^d\sigma(S_j)$$

where ${}^d\sigma(P_i)$ and ${}^d\sigma(S_j)$ are the molecular signatures of substrate S_j and product P_i at diameter d .

This approach allows us to encode biochemical reactions by looking at the changes occurring at the reaction center. Note that the specificity of a reaction signature is determined by the diameter of the molecular signature, as lower diameters encode multiple compounds while higher diameters are specific. Therefore, reaction signatures have been shown as a handy way to model enzymatic substrate promiscuity [71, 81, 177], i.e. the ability that enzymes have to process structurally similar substrates. Our chosen diameter of 12 assumes a relatively low degree of enzymatic promiscuity for the encoded reactions.

5.3.3 Matching algorithm

After integrating reaction signatures in our database, we can predict on-the-fly if a compound can act as substrate of a reaction by using a new implementation of the RetroPath forward algorithm [152]. If a compound C has a list of fragments (atomic signatures and their respective occurrence) embedding the substrate fragments contained in a reaction signature R (i.e. the negative part of reaction signature), then the compound is said to match the reaction. The sum of the signatures of compound C and those of the reaction R generates a new list of (positive) fragments P , representing the putative products generated by the reaction signature acting upon compound C . If we can retrieve a set of known compounds from those fragments, then the reaction is accepted and C is considered a valid substrate for R to produce P .

5.3.4 Metabolic graph

Pathways are handled as a graph (where nodes are compounds and edges reactions) with NetworkX python library [295].

5.3.5 Similarity search

In order to annotate compounds structurally similar to detectable compounds in predicted metabolic graphs, we precomputed the similarities between all compounds and detectable ones. Indeed, promiscuous detection of structurally similar compounds may not be reported in databases and should be checked in the literature if no suitable detectable compound is found by SensiPath.

Similarity was evaluated with RDKit python library (<http://www.rdkit.org/>), representing compounds with RDKit's ECFP4 fingerprint implementation and a Jaccard- Tanimoto index [296]. A Tanimoto of one is a perfect match.

5.3.6 Web server implementation

SensiPath web server is a Docker application running the following standard software packages: Nginx, gUnicorn, Django and Postgres. Data and matching functions are stored in the database.

5.4 Input and Output

5.4.1 Input

Users query SensiPath with the compound they wish to detect (figure 5.2, left panel), either as an identifier from an external database (e.g. ChEBI available at <https://www.ebi.ac.uk/chebi/>) or as a standard InChI (<http://www.inchi-trust.org/>). InChI is a IUPAC string representation of compounds and can be easily obtained from compound databases. Users can choose to search for detectable compounds that are at one or two enzymatic steps away from their target.

5.4.2 Output

SensiPath displays its results in two views; (i) pathway view: the set of pathways leading to recognized detectable compounds (figure 5.3 A); and (ii) graph view: the whole computed graph around the target (figure 5.3 B), also available for download as a standard Graph Markup Language file.

not allow the signal to be transferred to the genetic layer of a living organism, a requirement for further *in situ* signal processing.

Here, we show how SensiPath was used in order to design a SEMP that detects cocaine *in vivo*. To that end, SensiPath web server is queried using a chemical identifier of cocaine, either through CHEBI:60056 or InChI=1S/C17H21NO4/c1-18-12-8-9-13(18)15(17(20)21- 2)14(10-12)22-16(19)11-6-4-3-5-7-11/h3-7,12-15H,8-10H2,1-2H3/t12-,13+,14-,15+/m0/s1. SensiPath finds a candidate SEMP allowing detection that is one enzymatic step away from the target. On the Graph view (figure 5.3 B), the five different products obtained through known enzymatic activities on cocaine are displayed. Clicking on an edge of the graph provides a link to databases providing information on each reaction. One of these compounds has a green border indicating that a biosensor is known to interact with an identical or highly similar chemical structure. This suggests that the information of the presence of cocaine in the medium can be transferred to the genetic layer and thus constitutes a putative SEMP. All found SEMPs are summarized on the Pathway view (figure 5.3 A). In the present case study, cocaine can be hydrolysed and forms the detectable molecule benzoate. Clicking on the arrow that represents the enzymatic transformation will display cross reference links to external databases of enzymatic transformations. It is strongly recommended to carefully check the bibliography that motivated the annotation of the reaction in the database, since important results might be omitted or misrepresented due to an incorrect curation process. In the case of cocaine hydrolysis, several publications confirm the benzoate conversion and databases such as Rhea and MetaCyc provide a direct link to Uniprot or GenBank where the sequence coding for the enzyme can be found (GenBank AF173165.1). This sequence can be synthesized and cloned into an expression vector of choice to constitute the metabolic module part of the SEMP.

In parallel, a query on BioNemo or RegTransBase for benzoate (or benzoic acid), the compound reported by SensiPath as having a biosensor, leads to several potential transcription factors that are known to interact with this compound (BenM, BenR, CbdS, PcaR, TcbR, CatR, BadR and XylS). In our experimental implementation, we chose the couple composed of BenR and its responsive promoter pBen from *Pseudomonas putida* KT2440 after a quick assessment of the available literature.

The sequence of pBen can then be synthesized and cloned in front of a reporter gene of choice (e.g. a fluorescent protein) in addition to the transcription factor coding sequence in order to form the sensing module part of the SEMP. To maximize the chances of proper expression of the heterologous proteins, we recommend to perform a step of codonoptimization on all the coding sequences,

to place them under control of inducible promoters and to use strain such as BL21(DE3) due to its efficient protein expression capabilities.

5.5.2 Parathion detection

Synthetic biology application of biosensors in the field of environmental protection could take the form of microorganisms programmed with a "seek and destroy" behaviour toward pollutants [298]. However, the task of engineering tailor-made biosensors for pollutants has been difficult to date [299].

Parathion is listed as one of the twelve worst offenders persistent organic pollutants according to the United Nations Environment Program and could benefit from such synthetic biology applications provided that a biosensor is available.

A request on Sensipath for parathion, with identifier CHEBI:27928, leads to the identification of a 1-step SEMP that depends on a phosphotriesterase (PTE) allowing transformation of parathion into 4-nitrophenol. As in the previous cocaine example, the proposed transformation could be verified in the literature [300]. We have experimentally validated this SEMP with a metabolic module based on the PTE coding sequence coupled with the sensing module made up of the transcription factor DmpR and its responsive promoter Pu from *Pseudomonas* sp. CF600. However, both PTE enzyme and DmpR promoter are known to be promiscuous, and other pollutants harbouring phenolic structures could activate DmpR. As this could impair applications requiring a high specificity, alternative SEMP were also explored.

Interestingly, with a 2-steps query, SensiPath's Pathway view shows that 4-nitrophenol can be an intermediate compound to another SEMP based on nitrite detection. Indeed a second enzymatic step mediated by a monooxygenase [301] is able to further transform 4-nitrophenol into nitrite, which is known to interact with regulators such as NarL from *E. coli*. This alternative offers the possibility of developing a more specific biosensor, effectively discarding any risk of cross-activation by phenolic compounds, as long as they do not have a nitro group. Going further with this idea, high specificity target detection could be guaranteed by building up combinations of alternative SEMP in one or several strains.

5.6 Discussion

The development of novel biosensors is presently needed in order to enlarge the set of detectable and observable metabolites that are available for synthetic

biology applications such as in health, environment or fine chemical production. In that direction, the SensiPath web server provides synthetic biologists with new solutions to build circuits having the ability of triggering a genetic response when a compound of interest is present. Our biosensor design solution is based on the strategy, not fully explored previously, of performing an *in silico* screening for enzymatic pathways linking the target to known detectable compounds. The originality of the approach lies in the systematic search through a full enumeration that SensiPath carries out, allowing discovery of novel sensing pathway candidates in the metabolic space. Resulting SEMP's are appealing for synthetic biologists because they can be easily built using conventional DNA assembling techniques and tested *in vivo*. SensiPath thus provides an easy way to explore right out of the box multiple biosensor constructs.

Depending on the application, the reliability of the candidate SEMP's identified by our method may vary. Limitations of the SEMP method include the need for the target compound to be able to co-localize with the enzyme (i.e. to enter the cells or to be internally produced in the cell), and the need for enzymatic products of the sensing pathway to be not too toxic to the cell. Such issues need to be addressed in a case-by-case manner, since they greatly depend on the application and on physico-chemical properties that are not always known for the compound. Other potential limitations of the method hold with regards to the choice of the biosensor. Although some information about the degree of promiscuity of transcription factors may be available from databases and literature, this aspect should be carefully considered in each application, especially if the final application requires a high level of specificity. The choice of the biosensor should also take into account dose response parameters such as the dynamic range and linear range of detection. SEMP's properties will depend on the actual properties of the biosensor, an information that therefore should be considered and retrieved from the available literature. In addition, promoter sequences responding to transcription factors may not be always found in databases, often requiring an investigation of associated references. This information nevertheless is progressively becoming more available through repositories like the Registry of Standard Biological Parts (<http://parts.igem.org/MainPage>).

In conclusion, we believe that the SEMP detection method is an interesting alternative worth considering with respect to tailored solutions such as rational design [252, 260] or genome mining [289]. To the authors acknowledgement, this is the first time a web-based tool is proposed to design biosensors based on the SEMP's approach. Other tools (such as M-path [73] or BioSynther [91] to name a few) proposed finding pathways from one compound to another, but they did not include any detectability concept in the way it was considered here. In that sense, SensiPath and SEMP's will surely contribute to the design of new

synthetic biology applications. Moreover, we should expect in the next years to see the broadness of applicability of SEMP to increase in parallel with progress in reaction and gene expression regulation knowledge sources.

Expanding biosensing abilities through computer-aided design of metabolic pathways

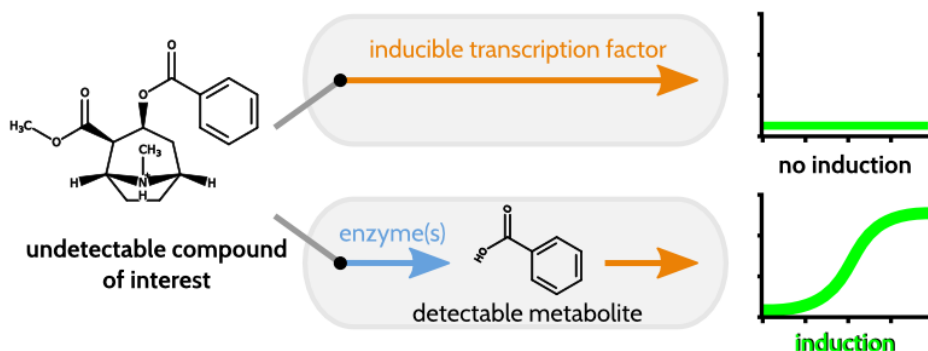


Fig. 6.1: Graphical abstract depicting the experimental validation of Sensing Enabling Metabolic Pathways. Cells with a *sensing module* (orange) but without *enzymatic module* (blue) do not show any induction in presence of the target compound (top). Once the *enzymatic module* is inserted, the signal is transmitted through a *detectable compound* that triggers the induction, revealing the presence of the target compound (down).

This work was originally published in ACS Synthetic Biology by Vincent Libis, Baudoin Delépine, and Jean-Loup Faulon.

Only minor modifications have been introduced in the chapter presented below.

Full reference

Vincent Libis, Baudoin Delépine, and Jean-Loup Faulon. "Expanding Biosensing Abilities through Computer-Aided Design of Metabolic Pathways". In: *ACS synthetic biology* 5.10 (2016), pp. 1076–1085. DOI: 10.1021/acssynbio.5b00225

Contribution

V.L., B.D., and J.-L.F. designed the study. V.L. designed, built, and characterized the biosensors. B.D. adapted the algorithms and conducted simulations and

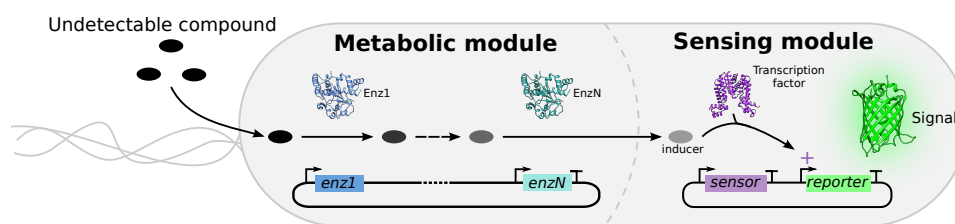


Fig. 6.2: General concept of a cell equipped with a sensing-enabling metabolic pathway (SEMP) allowing for the detection of a new chemical. A naturally undetectable molecule is transformed by a metabolic module into an inducer molecule triggering a genetic response upon binding to a transcription factor. While the figure illustrates the case of transcription factor working as a positive activator, the same concept can be applied with a repressor or a riboswitch.

analysis. All authors participated in the interpretation of the results and in the preparation of the manuscript.

6.1 Abstract

Detection of chemical signals is critical for cells in nature as well as in synthetic biology, where they serve as inputs for designer circuits. Important progress has been made in the design of signal processing circuits triggering complex biological behaviors, but the range of small molecules recognized by sensors as inputs is limited. The ability to detect new molecules will increase the number of synthetic biology applications, but direct engineering of tailor-made sensors takes time. Here we describe a way to immediately expand the range of biologically detectable molecules by systematically designing metabolic pathways that transform nondetectable molecules into molecules for which sensors already exist. We leveraged computer-aided design to predict such sensing-enabling metabolic pathways, and we built several new whole-cell biosensors for molecules such as cocaine, parathion, hippuric acid, and nitroglycerin.

6.2 Introduction

Engineering of circuits in cells has made fast progress since the dawn of synthetic biology. New modular tools and strategies regularly expand the toolbox [286]. Just considering the progress made in the last two years, signal processing in biological systems can now rely on elements such as load drivers [302], memory systems [303], amplifiers [304], coupling systems [305], or bow-tie architectures [306], to name a few. Such tools enable the use of synthetic circuits in real life applications where the complexity of the signals encountered in the environment was until now problematic [287]. Despite these intense efforts allowing

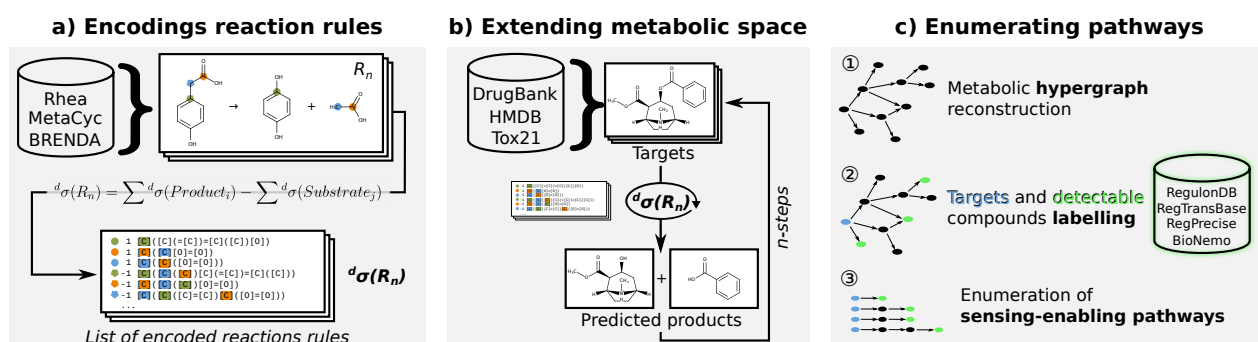


Fig. 6.3: Sensing-enabling metabolic pathway (SEMP) design workflow. (a) Reactions are collected from biochemical databases and encoded as reaction signatures. A reaction signature describes a biochemical transformation rule as a list of fragments of the molecules that are modified during the reaction (See Methods). (b) The effect of each reaction signature is simulated on each target. If a product is predicted to form (i.e., if a target's fragments match the ones of a natural substrate of the reaction), the generated structure is injected back into the target list, in case it can be further transformed. The metabolic space is extended by this iterative process until no more products are formed or if a limit is set on the number of iteration ($n = 3$ in this work). (c) Finally, the extended metabolic space is represented as an oriented graph where the nodes are compounds and the edges represent enzymatic reactions. This graph is colored by labeling the compounds that are found in the transcription factor effectors databases. Pathways linking a target to a gene inducer can then be automatically enumerated.

precise control of circuit behavior, the development of applications is slowed by the limited number of inputs available. Usually, inputs to the genetic layer of circuits are mediated by sensors such as transcription factors, riboswitches, or two-component signaling pathways. Unfortunately, the number of organic molecules detectable by well-characterized natural sensors is relatively small [241]. Rational engineering of sensors through protein engineering or riboswitch engineering has been accomplished [252, 307], but the time and effort necessary to deploy such approaches still limits the number of available tailor-made sensors.

Alternative strategies of biosensing could play a role in tackling this lack of inputs. In nature, information about a chemical signal can be indirectly conveyed through enzymatic transformations. A classic example can be observed in the Lac operon where information about the quantity of lactose in the medium is not acquired by direct interaction with a transcription factor. Instead, a fraction of the available lactose is transformed by β -galactosidase into allolactose, which is the molecule detected by the transcription factor LacI. The use of metabolic transformation to convey information to the genetic layer has also been demonstrated in synthetic biology to detect an aromatic and to obtain cell to cell communication [274, 308, 309].

Here we explore the full potential of metabolism to enable detection of new molecules and thus expand the scope of chemicals that can serve as input in synthetic biology applications. We systematically search for enzymatic ways to

transform undetectable molecules of interest into molecules detectable by existing biosensors (Figure 6.2). This requires the design of tailor-made pathways out of thousands of individual enzymatic modules available in the pool of known biochemical reactions, and necessitates the development of computer-aided design (CAD) tools. As synthetic biologists developed CAD tools to guide the engineering of genetic circuits [310, 311], metabolic engineers created powerful computational methods in the context of small molecule production in microorganisms [35, 81]. Among them, mathematical abstractions have been developed to represent and simulate biochemical reactions *in silico*. When fed with entire databases of known biochemical reactions, such a tool can extract a set of biochemical reaction rules that can then be applied to any given substrate to generate potential products (i.e., predicted metabolites). Retrosynthesis tools can iteratively apply these rules to reconstruct natural or synthetic metabolic pathways. We leveraged this expertise to build a CAD tool exploiting 9,319 biochemical reaction rules to search for sensing-enabling metabolic pathways connecting molecules of interest to the genetic layer. We first evaluated the potential of this approach *in silico* by predicting sensing-enabling metabolic pathways (SEMP) of molecules such as drugs, biomarkers, and toxics. We then tested several predictions in *E. coli* by assembling circuits made of heterologous enzymes and transcription factors. We observed the successful fluorescent response of *E. coli* to several molecules, including cocaine, parathion, 2-chloro-4-nitrophenol, hippuric acid, and nitroglycerin. In synergy with ongoing efforts of biosensor development, this strategy offers an immediate expansion of the scope of inputs for synthetic circuits and has the potential to open the way to new synthetic biology applications in fields such as medicine and environment.

6.3 Results

6.3.1 The scope of detectable molecules is significantly expanded *in silico* via enzymatic pathways

We evaluated the potential of this approach by predicting sensing pathways of target molecules such as drugs, biomarkers of human diseases and molecule with risk of toxicity for health and the environment. We gathered data sets representative of these types of molecules from three public databases: DrugBank [312], HMDB [134], Tox21 [313]. Systematic design of tailor-made SEMP for these targets requires biochemical retrosynthesis and subsequent identification of inducers in the generated products. The workflow we used consists of four steps: (i) gather the whole trans-species Reactome (ts-Reactome) as a list of encoded biochemical reaction rules; (ii) submit each target to the ts-Reactome rules to

Tab. 6.1: Example of Sensing-Enabling Metabolic Pathway (SEMP) Targets marked by a star ("*") have their SEMP implemented and tested *in vivo* over the course of this work (see Supporting Information Figure 6.8).

Target	Enzymes and metabolites	Sensor and effectors
Cocaine* illicit drug	Cocaine esterase (<i>Rhodococcus</i> sp.) 3.1.1.84: Benzoate	BenR (<i>Pseudomonas putida</i>) Benzoate
Heroin illicit drug	Heroin esterase (<i>Rhodococcus</i> sp.) 3.1.-.-: Acetate (2×)	NR I (<i>Escherichia coli</i>) Acetate
Aspirin drug	Acetylsalicylate deacetylase (<i>Rattus norvegicus</i>) 3.1.1.55: Salicylate	NahR (<i>Pseudomonas putida</i>) Salicylate
Caffeine drug	Methylxanthine N1-demethylase (<i>Pseudomonas putida</i>) 1.14.13.178: Formaldehyde	FrmR (<i>Escherichia coli</i>) Formaldehyde
Paracetamol drug	Aryl acylamidase (<i>Rhodococcus erythropolis</i>) 3.5.1.13: Acetate	NR I (<i>Escherichia coli</i>) Acetate
Barbituric acid drug	Bar (<i>Rhodococcus erythropolis</i>) 3.5.2.1: Ureidomalonate N-malonylurea hydrolase (<i>Rhodococcus erythropolis</i>) 3.5.1.95: Malonate	FapR (<i>Bacillus subtilis</i>) Malonate
Nitroglycerin* drug/explosive	NemA (<i>Escherichia coli</i>) 4.99.1.-: Nitrite (2x)	NarL (<i>Escherichia coli</i>) Nitrite
Chlorpropham* pollutant	AmpA (<i>Paracoccus</i> sp.) 3.5.1.-: 3-Chloroaniline	TadR (<i>Delftia tsuruhatensis</i>) 3-Chloroaniline
2-chloro-4-nitrophenol* pollutant	PnpA (<i>Burkholderia</i> sp.) 1.14.13.-: Chloro-1,4-benzoquinone PnpB (<i>Burkholderia</i> sp.) 1.6.5.-: Chlorohydroquinone	LinR (<i>Sphingomonas paucimobilis</i>) Chlorohydroquinone
Propanil pollutant	AmpA (<i>Paracoccus</i> sp.) 3.5.1.-: Propionate	PrpR (<i>Corynebacterium glutamicum</i>) Propionate
Parathion* pollutant	PTE (<i>Pseudomonas diminuta</i>) 3.1.8.1: 4-nitrophenol	DmpR (<i>Pseudomonas</i> sp.) 4-nitrophenol
Hydrogen cyanide chemical warfare agent	Cyanide hydratase (<i>Gloeocercospora sorghi</i>) 4.2.1.66: Formamide Formamidase (<i>Paracoccidioides brasiliensis</i>) 3.5.1.49: Formate	FdsR (<i>Ralstonia eutropha</i>) Formate
Cyclosarin chemical warfare agent	PTE (<i>Pseudomonas diminuta</i>) 3.1.8.-: Cyclohexanol ChnA (<i>Acinetobacter</i> sp.) 3.5.1.49: 1.1.1.245: Cyclohexanone	ChnR (<i>Acinetobacter</i> sp.) Cyclohexanone
Hippurate* biomarker	HipO (<i>Campylobacter jejuni</i>) 3.5.1.32: Benzoate	BenR (<i>Pseudomonas putida</i>) Benzoate

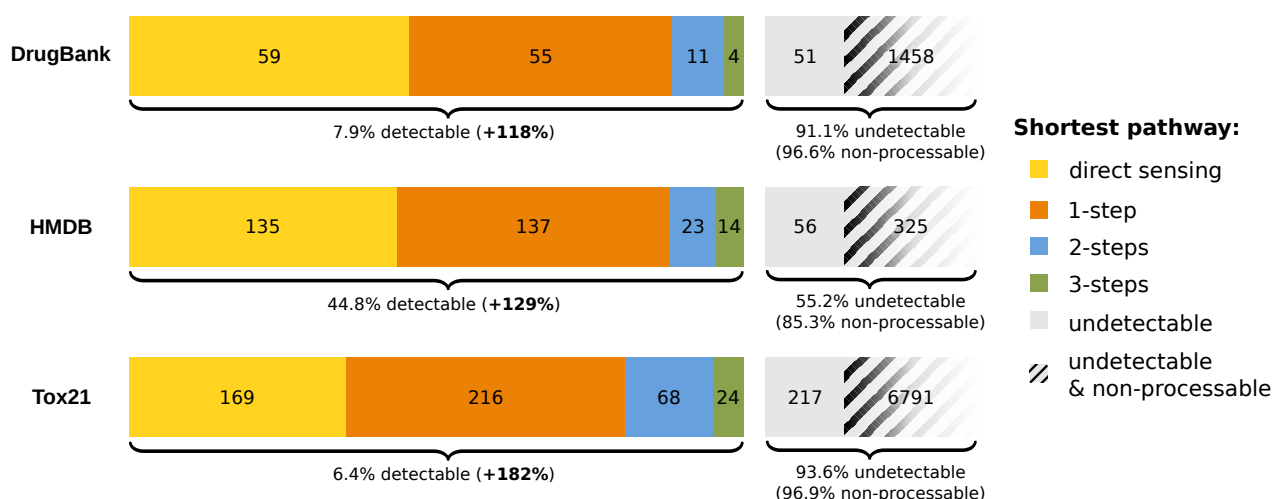


Fig. 6.4: In silico prediction of detectable compounds among three data sets. Compounds labeled in the "direct sensing" category (yellow) are already present in the data set of natural transcription factor's effectors. Compounds in other colors (orange, blue, green) are reached by sensing-enabling metabolic pathways (SEMP). Nonprocessable compounds (hashed) are compounds that do not participate in known enzymatic reactions (as a product or a substrate). Data sets correspond to the DrugBank sections "approved drugs" and "illicit drugs"; HMDB's biomarker compounds are associated with a human disease and the entire Tox21 data set of putative toxic compounds.

generate products; (iii) iterate on the products and generate a hypergraph around targets; and (iv) screen the hypergraph for inducers and output putative SEMPs. The ts-Reactome was based on 3 major biochemical databases (BRENDA [51], Metacyc [314], and Rhea [292]) that were merged and encoded in reaction signatures [177], a mathematical representation of reactions that we developed previously. Reaction signatures have been proven useful for metabolic pathway design in the context of microbial production of value-added compounds [217]. The ts-Reactome reached a total of 9,319 unique reaction rules. When iteratively applying these reaction rules to the targets and their generated products, combinatorial explosions can be computationally demanding and we therefore limited to 3 steps the maximum length of the pathways. Once the hypergraph was generated it was colored with known inducers sourced from 4 databases of transcription factor effectors: BioNemo [244], RegTransBase [242], RegulonDB [294], and RegPrecise [243]. An overview of the workflow is represented in Figure 6.3, and the labeled graph output from the DrugBank data set can be seen in Supporting Information Figure 6.6. The pathways in the graph linking a target molecule to a natural transcription factor effector were automatically enumerated to allow statistical analysis and selection of proof of concept examples for *in vivo* implementation.

In each of the considered target data sets a number of compounds were found to be naturally detectable by existing biosensors, respectively 59, 135, and 169 for DrugBank, HMDB, and Tox21. Through metabolism, the number of detectable

compounds grows to 123 (Drugbank), 280 (HMDB), and 477 (Tox21). Therefore, this approach at least doubled the number of detectable molecules that can be considered for sensor development in each of the studied data sets (Figure 6.4).

Most compounds in the data sets that remain undetectable are products of organic chemistry that are not processed by any known enzyme and therefore no SEMP could be predicted by this method. Remarkably, if a compound is processed by an enzyme the chances are high (>66%) that at least one SEMP will be found in 3-steps or less. Moreover, the connection with an existing biosensor is usually possible in a small number of enzymatic steps as the number of compounds that necessitate long SEMP (3 steps) is significantly lower than the ones connected through 1 and 2 step pathways. A short selection of interesting candidates for biosensor development and associated SEMPs is displayed in Table 6.1.

6.3.2 Implementation of SEMPs *in vivo* expands *E. coli* sensing abilities

In order to validate the concept *in vivo*, we implemented in *E. coli* representative SEMPs from the predictions related to each of the targets data sets (1 drug, 2 pollutants, 1 biomarker). For each SEMP, the genes coding for the required enzymes were cloned into a metabolic module plasmid allowing control of the enzymes' expression level by IPTG. In parallel the genes coding for the required heterologous transcription factors were cloned into a sensing module plasmid under the control of an arabinose-inducible promoter and a red fluorescent protein (RFP) was placed under the control of their associated promoters. The two modules are compatible for cotransformation in the same strain. Prior to full characterization of SEMPs we identified optimal expression levels of the heterologous transcription factors by varying the arabinose concentration in the presence of the natural effectors (see Methods).

Whole-cell biosensors for illicit compound detection can provide a cheap way to determine the presence of a molecule in an unknown mixture. Among the predictions we obtained by processing the DrugBank database's section "Illicit Drugs" we selected a SEMP providing *E. coli* with the ability to detect cocaine. As shown in Table 6.1, the transformation of cocaine by an esterase (CocE from *Rhodococcus* sp.) produces benzoate which can be detected by the transcription factor BenR from *Pseudomonas putida*. *E. coli* BL21(DE3) was first transformed only with the sensing module bearing the transcription factor BenR and a RFP under the control of pBEN promoter. This strain emits a strong fluorescent signal in the presence of benzoate but not in the presence of cocaine (Supporting

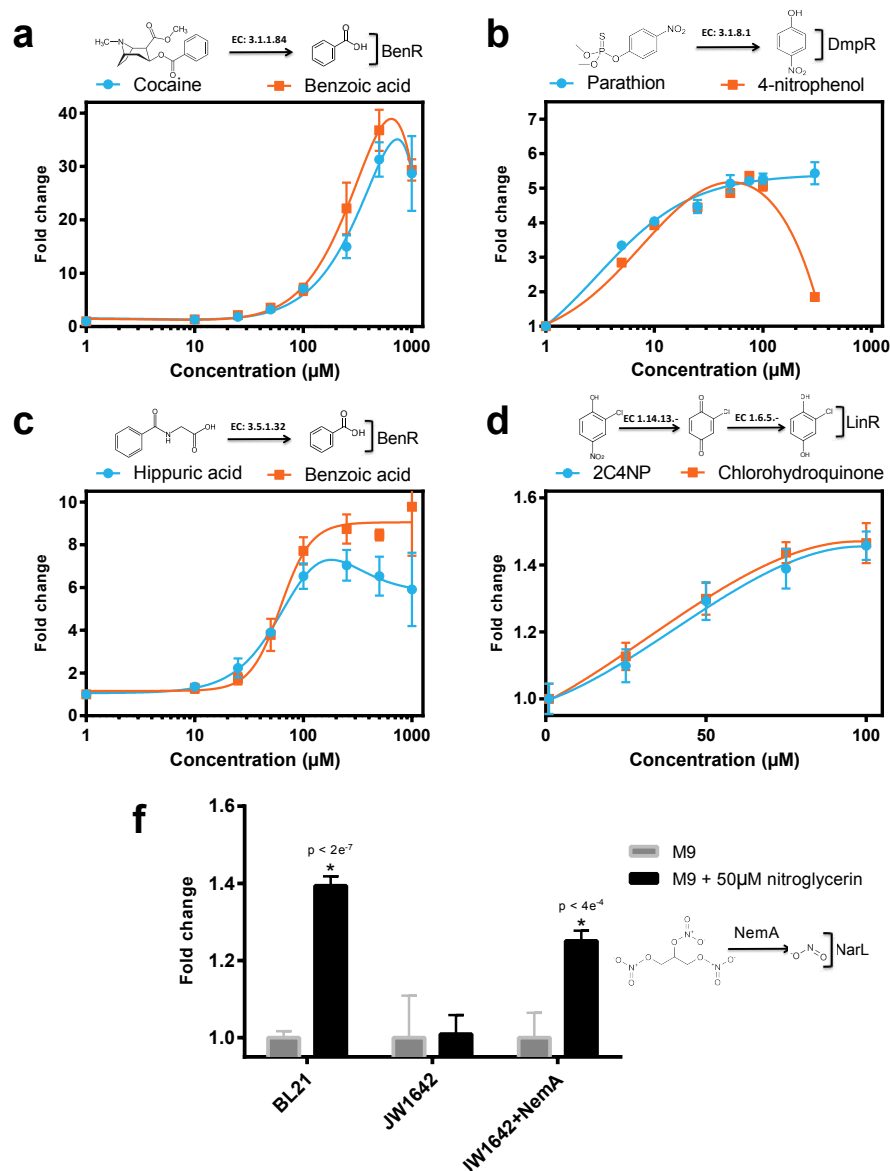


Fig. 6.5: *In vivo* characterization of sensing-enabling metabolic pathways (SEMP).

Cotransformation of *E. coli* with both the metabolic module and sensing module confers sensing abilities toward new molecules. The dose-response relationship of engineered *E. coli* strains to the natural effectors and to the target molecules cocaine (a), parathion (b), hippurate (c), and 2C4NP (d). (e) shows the fluorescent response of three strains of *E. coli* harboring a RFP fused to pYearR promoter (controlled by NarL) in M9 medium or in M9 supplemented with 50 μM nitroglycerin. Laboratory strain BL21 responds to nitroglycerin, but strain JW1642, which is knocked out for the NemaA enzyme, does not. However, response can be restored in JW1642 by providing a functional copy of NemaA on a plasmid. Fold change represents the ratio between the fluorescence observed at a given concentration and the fluorescence of uninduced controls of the same strain; a value of 1 represents no induction. Each data point is the mean of at least 3 replicates, and error bars represent standard deviations. Smooth curves represent dose-response fitting of the scattered data points. Statistical significance was determined using Student's t-test with a P-value cutoff of 0.005.

Information Figure 6.10a). Upon cotransformation with a metabolic module containing CocE, a fluorescent response also occurs in the presence of cocaine, indicating the expected extension of the sensing scope of the bacteria (Figure 6.5a). Comparison of the dose-response curves of the strain toward cocaine and the natural effector benzoate suggest that the enzymatic step have no or little impact on the biosensor performance features. The linear range of detection and dynamic range for cocaine stays in the same order of magnitude as benzoate.

Another interesting application is the development of whole-cell biosensors for monitoring chemicals in the environment. Parathion is a major environmental threat and counts among the "dirty dozen", the 12 worst offenders persistent organic pollutants according to the United Nations Environment Programme. The transformation of parathion by a phosphotriesterase (PTE from *Pseudomonas diminuta*) produces 4-nitrophenol, which can be detected by the transcription factor DmpR from *Pseudomonas sp.* (see Table 6.1). An *E. coli* strain harboring a sensing module based on DmpR and the associated DmpK promoter is able to detect 4-nitrophenol in a dose-dependent fashion but not parathion (Supporting Information Figure 6.10b). Upon addition of the metabolic module containing PTE to this strain, the extension of the sensing scope takes place and a strong fluorescent response occurs in the presence of parathion (Figure 6.5b). To our surprise the fluorescent response of this strain to parathion is higher than for the native inducer above 100 μM . The drop in signal in response to 4-nitrophenol at 300 μM is probably due to the associated toxicity that diminishes growth (see Supporting Information Figure 6.13). In addition to this effect on growth, flow cytometry measurements indicate a decreased fluorescence in individual cells at higher 4-nitrophenol concentration (See single cell data at 1 mM in Supporting Information Figure 6.12).

In contrast, parathion is less toxic for these cells and in fact is not toxic at these concentrations if the PTE enzyme is absent (Supporting Information Figure 6.13). Moreover, despite the decrease in growth rate associated with the intracellular formation of 4-nitrophenol, the individual fluorescence of these cells in response to parathion keeps increasing gradually up to 1 mM (Supporting Information Figure 6.12). Partial or delayed transformation of parathion at high concentration may allow the cells to stay healthy longer and produce more signal.

Medical applications of synthetic biology often rely on biosensors for biomarkers of human diseases [315]. We chose to implement a biosensor for an interesting biomarker, hippuric acid, that is found at high concentration in the urine of a person intoxicated with toluene. As predicted by the algorithm, transformation of *E. coli* with a metabolic module harboring hippurase HipO from *Campylobacter jejuni* allows degradation of hippuric acid into benzoic acid, which is detected

in a dose-dependent fashion by the BenR-based sensing module (Figure 6.5c). We monitored the activation of this hippurate sensor and the cocaine sensor in order to investigate if SEMP's have kinetics of detection different from a regular benzoate sensor. Comparison of the response time to target compounds or to the natural inducer benzoate shows that neither the HipO- or CocE-mediated transformation step delays the fluorescent response (see Supporting Information Figure 6.7).

The engineering of transfer of matter through metabolism has made important progress during the last 20 years. Synthetic pathways as large as 23 enzymes long have been successfully implemented to divert natural carbon flux toward valuable compound production [28]. We wondered if transfer of information could as well rely on multienzymatic pathways and we selected a target compound that necessitates two consecutive transformations to allow detection. 2-Chloro-4-nitrophenol (2C4NP) belongs to a family of molecule with high toxicity for humans used in the chemical industry. Upon monooxygenation and subsequent reduction performed by two enzymes from a strain of *Burkholderia sp.*, 2C4NP is transformed into chlorohydroquinone, which can be detected by LinR, a transcription factor found in *Sphingomonas paucimobilis* (see Table 6.1). All three genes were cloned into the modules and conferred *E. coli* the ability to detect 2C4NP (Figure 6.5d), while it was not possible with the LinR-module alone (Supporting Information Figure 6.10d). With our setup, the sensing module based on LinR transcription factor and a fragment of the LinE promoter from *Sphingomonas paucimobilis* had a small dynamic range in *E. coli*. This could probably be overcome by directed evolution as promoter activity can be linked to a selectable output. 2C4NP is highly toxic for *E. coli*, thus limiting our measurements to a 100 μM upper limit. Remarkably, only the strain with the metabolic module could survive at a concentration of 2C4NP higher than 75 μM (Supporting Information Figure 6.13), this is probably due to the detoxifying effect of the two enzymes from *Burkholderia sp.*. This situation is the opposite to that observed in the case of the biosensing of parathion where the intermediate metabolite was more toxic to the cells than the initial target molecule.

Finally, an interesting case emerged from the predictions as a sensing-enabling pathway to the vasodilator drug and explosive nitroglycerin was identified with both metabolic module and sensing module component already present in wild type *E. coli*. Indeed, the promiscuous NemaA enzyme from *E. coli* is known to allow degradation of nitroglycerin into nitrites that are naturally monitored in *E. coli* by the NarL regulator. This suggests that wild type *E. coli* exhibits a fortuitous transcriptional response to nitroglycerin through a SEMP-like circuit. In order to test this hypothesis we transformed *E. coli* with a plasmid harboring a RFP fused to the native promoter pYeaR, which is controlled by NarL. This strain

successfully produces a fluorescent response when cultivated in the presence of nitroglycerin. To confirm the SEMP dependency of this response we transformed the same reporter plasmid into JW1642, an *E. coli* strain with a knocked-out NemaA enzyme. This strain is not anymore able to respond to nitroglycerin. Finally, we introduced a functional copy of NemaA on a second plasmid into JW1642 and observed a restoration of the fluorescent response to nitroglycerin (Figure 6.5e). This confirms that NemaA is enabling the transcriptional response to nitroglycerin of *E. coli* and represents an interesting example of a computationally elucidated naturally occurring SEMP. A summary of all the SEMP validated experimentally in this work is represented in Supporting Information Figure 6.8.

6.4 Discussion

Biological sensors are central for synthetic biology to solve real-world problems. Numbers of promising systems involving biosensors have been developed for medical, environmental, and industrial applications. The approach described here at least doubles the number of molecules that can be considered for sensor development in each of the studied compound classes (toxics, biomarkers, and drugs). SEMPs can be created without the need for time-consuming protein and riboswitch engineering or discovery of natural sensors. Moreover, the number of chemical candidates for such a sensing strategy should automatically grow in time, as the number of characterized sensors and biochemical reactions will continuously increase in databases. For these reasons, we envision SEMPs as an important new source of biosensors that will fruitfully be integrated within the modular synthetic biology toolbox.

Very recently a SEMP approach was used for the monitoring of a product of interest for industrial bioproduction, 3-hydroxypropionate [275]. By increasing the number of value added chemicals that can be monitored through fluorescent readout, SEMPs could alleviate the bottleneck that metabolic engineers currently face with the low throughputs of conventional measurement methods.

As another example of potential application, the hippuric acid sensor described here could be a starting point for a cheap way of controlling urine from workers in the paint industry, especially in the developing world, where a lot of toluene intoxications occur due to lack of regular testing. However, this application requires the cells to emit a signal upon reaching a clinically relevant threshold in complex medium. Encouragingly, such ability was achieved recently with a modular signal processing strategy (digitalization and amplification) allowing

glucose detection at selected thresholds in clinical urine samples from diabetic patients [287].

In addition to the proof-of-principle examples described in this work, it is interesting to note that the detection of parathion was achieved with the phosphotriesterase variant PTE-S5 [300], which is known to have an important substrate promiscuity against several organophosphorous compounds [316]. Among them are chemical warfare agents such as cyclosarin that could also be detected through the same strategy. Transformation of cyclosarin by PTE-S5 produces cyclohexanol, a compound being subsequently transformed by *Acinetobacter sp.*'s ChnA enzyme into cyclohexanone, for which a ChnR-based biosensor has already been characterized for metabolic engineering applications [317]. While cyclosarin counts among the most toxic substances ever created, its toxicity comes from the inhibition of the enzyme acetylcholinesterase in the brain; therefore, organisms such as *E. coli* can detect it without suffering the associated toxicity.

Intuitive limitations of SEMP are shared with other intracellular biosensors, such as the need for sufficient membrane permeability to extracellular targets and a limited toxicity of the targets for the chassis. As a potential option to minimize these issues, we expect SEMP to be easily transplanted to artificial cells that have been shown to serve as viable chassis for both biosensors and metabolic pathways [318, 319]. Specific limitations of SEMP reside in the need for non-toxic metabolic intermediates and the risk of specificity issues. These biosensors are unable to discriminate between the presence of the target molecule in the medium and any intermediates of the SEMP. Careful consideration should thus be given to the risk of crosstalk if one of the intermediate is a possible contaminant in the envisioned biosensing application. Additionally, the potential ligand promiscuity of enzymes and transcription factors must be kept in mind if high specificity is required for a particular application. Potential cases of specificity issues linked to promiscuity are depicted in Supporting Information Figure 6.11. While detection of multiple targets could be advantageous in environmental applications, this would usually be problematic in medical applications. Directed evolution of a SEMP guided by its fluorescent output is an option to overcome false-positive activation. Lastly, changes of physicochemical properties between the target molecule and the intermediates of the pathway might result in unexpected properties. For instance, a decrease in the permeability coefficient of metabolites resulting from the transformations could lead to their accumulation inside cells. This could lower the detection limit of the sensor.

Aside from the synthetic biology tool aspect, one may wonder if SEMP represent a motif frequently used in nature and why. To our knowledge, this has not been in-

vestigated yet. Nevertheless, recent reviews suggest that the role of the metabolic layer in integrating information about the medium or the internal state of the cell is underestimated compared to the role of the genetic layer [277, 320]. While the natural occurrence of a response to nitroglycerin seems fortuitous, recent evidence concerning the Lac operon shows that evolution strongly coselected the LacI transcription factor and the side-reaction site of β -galactosidase that leads to production of allolactose [276]. The fact that LacI never evolved to directly detect lactose suggests that this conserved SEMP motif is advantageous to the homeostasis although the mechanism is not elucidated yet.

6.5 Methods

6.5.1 Data source

We retrieved 9,319 distinct encoded reactions (reaction signature diameter 12, see below) and their associated compounds from BRENDA 2015.1 [51], MetaCyc 18.5 [314], and Rhea v61 [292]. Targets were gathered from DrugBank 4.1 [312] (all approved drugs and illicit drugs data sets), the Human Metabolomic DataBase [134] (all compounds linked to a human disease) and the Distributed Structure-Searchable Toxicity (DSSTox) Database Network [313] (all compounds). Finally, we retrieved 505 distinct transcription factor effectors from BioNemo [244], RegTransBase [242], RegulonDB [294], and RegPrecise [243].

6.5.2 Molecular and reaction signatures

Molecular signatures [77] (MS) are graph-based descriptors that encodes the "neighborhood" of each atom of a molecule, similarly to Morgan's or ECFP fingerprints. Each kind of "neighborhood", or atom environment, is a feature of MS. A reaction's signature (RS) is computed by subtracting the MS of the substrates to the MS of the products of the reaction [177], and takes the general form:

$$^d\sigma(R_n) = \sum_i ^d\sigma(P_i) - \sum_j ^d\sigma(S_j)$$

where $^d\sigma(P_i)$ and $^d\sigma(S_j)$ are the molecular signatures of substrate S_j and product P_i at diameter d . The variable diameter (d) of a molecular signature determines the size of the scope of the neighborhood that will be described around each atom. When this diameter is set to a small value, structurally similar molecules

will share a lot of atoms with identical neighborhoods. This will allow a reaction signature to be applied to substrates that are close but different to the ones originally described in biochemical reaction databases. While small diameters allows to simulate promiscuous activities of enzymes and predict more products, it arbitrarily assumes a certain level of promiscuity from enzymes and thus lead to the generation of a higher number of incorrect predictions. In this work we used molecular and reaction signature at a large diameter ($d = 12$) to encode compounds and biochemical reactions, we thus consider enzymes largely non-promiscuous (unless several distinct reactions are reported in the databases for the same enzyme). See Carbonell *et al.* [81] for detailed informations about metabolic pathway design at lower diameter ($d < 12$).

6.5.3 Compounds and reactions preprocessing

The reactions were filtered in order to gather only biochemical reactions with a structure available for all involved compounds. The compounds were filtered using ChemAxon's Checker and Standardizer tools (JChem v.15.4.27, 2015). We let aside compounds with R-groups and the associated reactions. We performed the necessary treatments to standardize the compounds such that their molecular signature would be comparable. This involved an aromatization step and the removal of explicit hydrogens. The resulting compounds and reactions were processed to generate molecular (MS) and reaction signatures (RS). In the end, we gathered 9,319 unique RS involving more than 18000 unique MS. Target compounds were pretreated the same way and were encoded as MS.

6.5.4 SEMP prediction

Our previous work focused on the development of a synthetic pathway retrosynthesis algorithm named Retropath [81]. We built further on this basis by developing a Python pipeline adapted to predict SEMP. In our implementation, each target compound is successively used in-place of each substrate of each reaction R . If the resulting putative reaction R' has the same reaction signature as R , then we accept R' as a pathway step. In order to extend the pathway, the products of R' are then considered themselves as targets. We generated pathways with up to three steps. The result is a graph where the nodes represent compounds, and edges represent reactions. Compounds are then matched to the list of transcription factor effectors compounds (Jaccard-Tanimoto coefficient [296] over 0.99). Finally, the sensing pathways are extracted from the graph with NetworkX graph library v.1.11 [295]. Frequently a target can be sensed through several SEMPs leading to different sensible compounds, with different pathway lengths. For

the analysis of the predictions, detectable targets were counted only once even if several SEMP were predicted.

Request for predictions on custom list of compounds can be addressed to Jean-Loup Faulon.

6.5.5 Chemicals and reagents

Benzoic acid, cocaine hydrochloride, hippuric acid, 2-chloro-4-nitrophenol, chlorohydroquinone, parathion-ethyl and 4-nitrophenol were purchased from Sigma (St. Louis, MO, USA). Permission to purchase cocaine hydrochloride was given by the French drug regulatory agency (Agence Nationale de Sécurité du Médicament et des Produits de Santé) to allow development of a new biosensor. Nitroglycerin was purchased from AccuStandard Europe (Niederbipp, Switzerland). Enzymes for cloning procedures (BsaI and DNA polymerase Q5) were purchased from New England BioLabs (Evry, France) and primers were purchased from Eurofins Genomics (Ebersberg, Germany).

6.5.6 Genetic constructs

Two custom plasmids were assembled to serve as vector for each module. Metabolic module vector is based on BioBrick standard vector pSB4T5 with pSC101 origin of replication and Tetracycline resistance marker, modified to harbor (i) type IIs restriction sites BsaI flanking cloning site and (ii) LacI transcription factor. Sensing module vector is based on BioBrick standard vector pSB1K3 with pMB1 origin of replication and Kanamycin resistance marker, modified to harbor (i) AraC transcription factor, (ii) pBAD promoter, (iii) type IIs restriction sites BsaI and (vi) a promoterless RFP. Coding sequences of genes BenR, CocE, HipO, were codon-optimized for *E. coli* with Jcat 44 and natural BsaI sites were removed. These genes were synthesized by Genscript (Piscataway, NJ, USA). Coding sequence of PTE-S5, LinR, pnpA, pnpB and DmpR were extracted by PCR from plasmids pMalc2x-PTE-S5 [300], pMEU2R [321], pET-pnpA, pET-pnpB [322], and *bba_k1413001* from the registry of standard biological parts (<http://parts.igem.org>), kind gifts from Prof. Dan TAWFIK, Prof. Yuji NAGATA, Prof. Ning-Yi ZHOU, and iGEM team EVERY 2014. Concerning sensing modules, LinR, BenR and DmpR sequences were followed by transcription terminator *BBa₀015* from the registry of standard biological parts, followed respectively with sequence of the LinE gene promoter (i.e., the 153 base pairs in front of LinE CDS on the *Sphingomonas paucimobilis* UT26 chromosome), pBEN promoter (i.e., the 150 base pairs in front of BenA CDS on *Pseudomonas putida* KT2440 chromosome) or sequence of the Pu promoter

(i.e., the 189 base pairs in front of DmpK CDS on *Pseudomonas* sp. CF600 plasmid pVI150). Concerning metabolic modules, T7 promoter with LacO operator and a ribosome binding site were placed in front of CocE, HipO, PTE-S5 and pnpA-pnpB CDS. All genes were amplified by PCR to add appropriate BsaI restriction sites and they were inserted in either metabolic or sensing module vectors using one-step GoldenGate assembly method [323]. In this setup the one-step assembly places the heterologous transcription factor under control of pBAD promoter and the promoterless RFP is placed under control of the heterologous promoter (see Supporting Information Figure 6.9 for a detailed map of metabolic and sensing modules). Annotated sequences for all constructs were deposited on GenBank (accession numbers KU746628, KU746629, KU746630, KU746631, KU746632, KU746633, KU746634, KU746635, and KU746636). Cloning of individual modules was done in DH5alpha and dose-response characterizations were carried out in BL21(DE3) after transformation with either sensing module only or both modules.

6.5.7 Biosensor dose-response characterization

For each biosensor strain, an isolated colony of BL21(DE3) harboring the appropriate plasmid(s) was inoculated in 2 mL of selective LB and grown overnight at 37°C. The overnight culture was diluted 1:100 in fresh selective LB and grown for 90 min under agitation at 37°C. Cells were then induced with IPTG 1 mM and arabinose and further grown until OD600 reached 0.1 for parathion, 4-nitrophenol, benzoic acid, hippuric acid and cocaine sensors. For 2C4NP and chlorohydroquinone, sensor cells were grown until OD600 reached 0.4 to minimize artifacts on the signal due to the high toxicity of 2C4NP. In our setup, optimal induction levels of transcription factors were found to be obtained with arabinose levels of 0.001% (BenR) 0.1% (LinR) and none (DmpR). Candidate chemicals for biosensing were dissolved in ethanol and 2 µL of different concentrations were mixed with samples of 200 µL of cell culture. 2 µL of pure ethanol was mixed with negative control cell samples. Cocaine sensor cells and associated control cells lacking metabolic module were grown in Eppendorf 1.5 mL microtubes with vigorous agitation for 18 h hours at 30 °C, thus following optimal conditions described in the literature for CocE activity. All other sensor cells were grown for 18 h with agitation at 37°C in microplate reader TECAN Infinite 500. Absorbance at OD600 and fluorescence (Exc: 580 nm/Em: 610 nm) was measured with microplate reader TECAN Infinite 500. All experiments were repeated at least 3 times on different days with similar results.

6.5.8 Characterization of the transcriptional response to nitroglycerin of different *E. coli* strains

pYeaR promoter (i.e., 146 base pairs in front of YeaR CDS on *E. coli* MG1655 chromosome) was amplified by PCR from purified chromosome and fused through GoldenGate assembly with a promoterless RFP into a pACYC plasmid backbone. The resulting plasmid pYeaR-pACYC was transformed into BL21- (DE3) or JW1642 from the Keio collection [324]. The coding sequence of NemaA was extracted by PCR from a BL21(DE3) strain and cloned under the control of constitutive promoter J23100 into a pCDF plasmid backbone. A strain of JW1642 was co-transformed with plasmids pYeaR-pACYC and NemaA-pCDF to investigate if it would restore the response to nitroglycerin. Cells were grown at 37°C in selective M9 minimal medium with 0.5% glucose. Overnight cultures of each strain were diluted 1:100 and grown until OD600 reached 0.1. 2 µL of nitroglycerin diluted in ethanol was added to 200 µL of cell samples to reach a final concentration of 50 µM. 2 µL of pure ethanol was added to control samples. Cells were grown for 18 h with agitation at 37°C and absorbance at OD600 and fluorescence (Exc: 580 nm/Em: 610 nm) was measured in microplate reader TECAN Infinite 500.

6.6 Supporting information

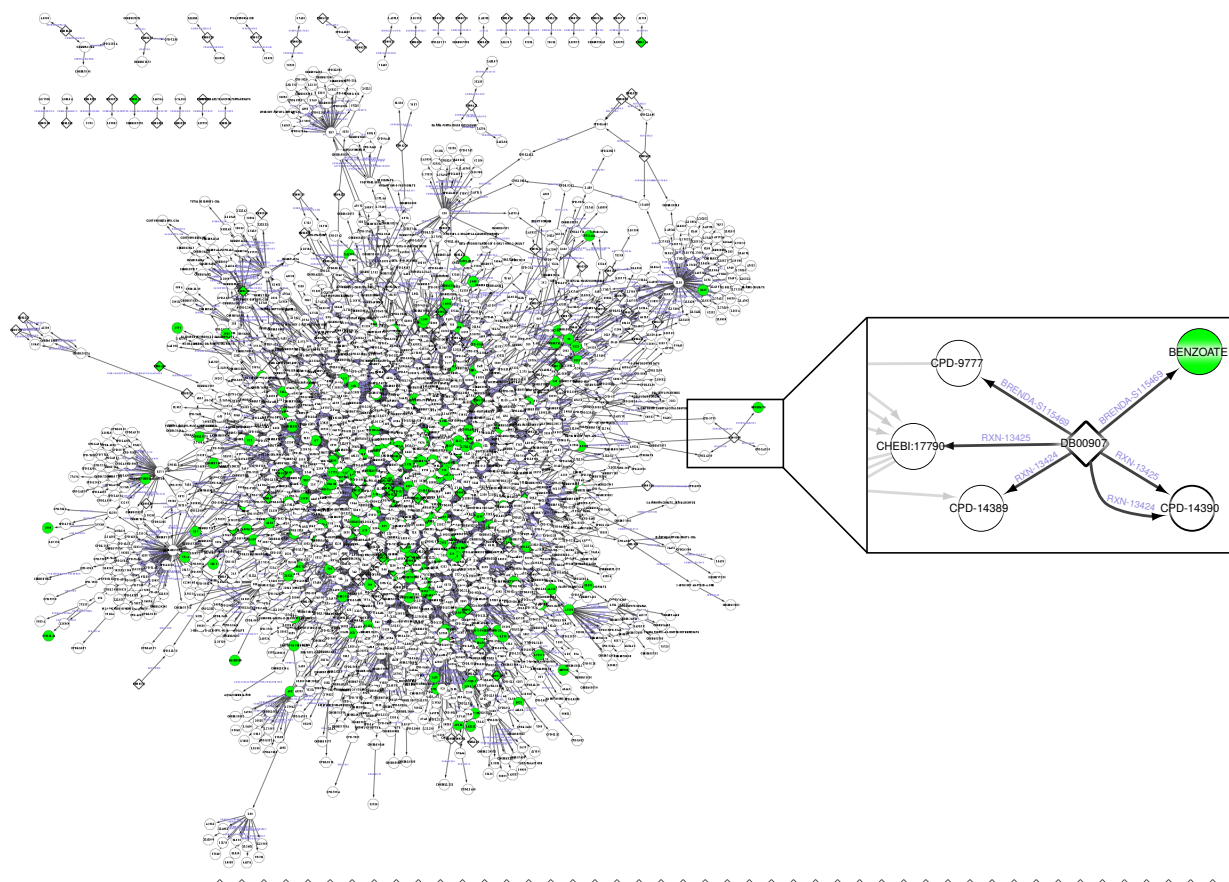


Fig. 6.6: Extended metabolic space around the DrugBank dataset. This is an oriented hypergraph generated by the bioinformatic pipeline described in Fig. 6.3, here visualized with Cytoscape. Similar graphs were generated and (computationally) analyzed for the HMDB and Tox21 datasets. Nodes represent chemical compounds and edges represent biochemical reactions. Each compound is associated strictly to one node. Enzymes can be associated to different pairs of compounds, and thus can be associated to several edges. Diamond shaped nodes represent target compounds from DrugBank (illicit or accepted drugs) and circular shaped nodes represent compounds that are generated through up to three metabolic steps from the initial target compounds. Nodes colored in green are naturally detectable compounds found in the transcription factor effector databases. A zoom on the area around target compound DB00907 (cocaine) shows the predicted products of different enzymatic reactions including the natural effector benzoate. Compounds not involved with any reaction ("non-processed compounds") are not represented.

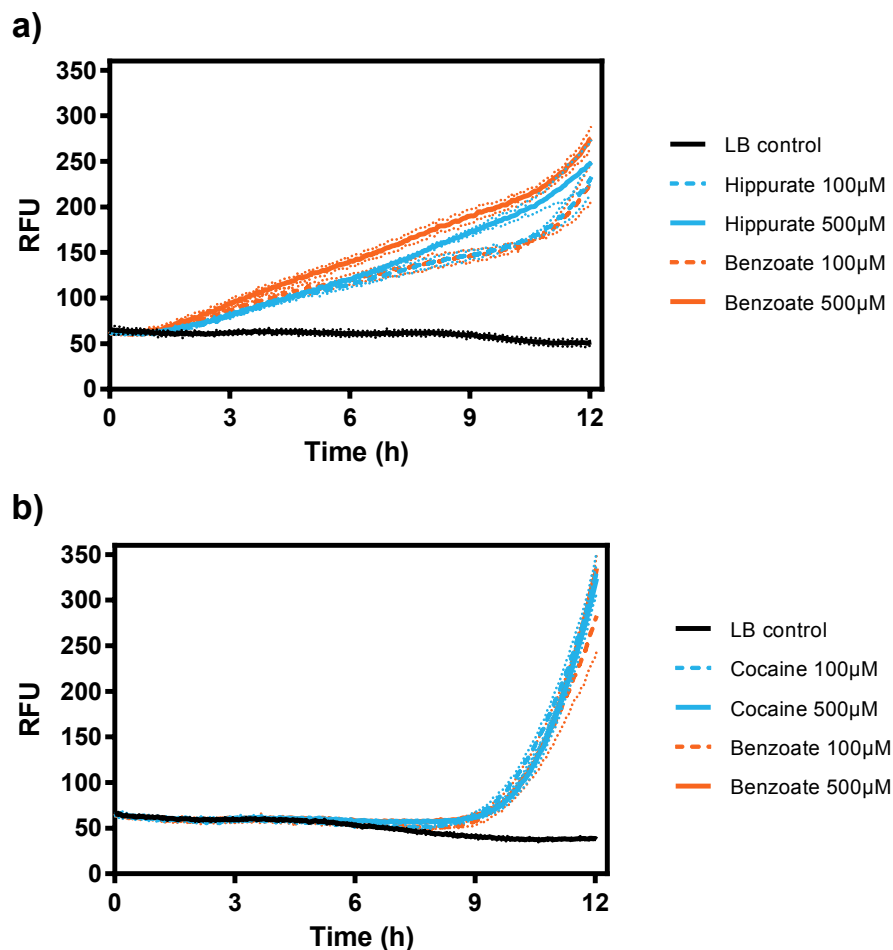


Fig. 6.7: Monitoring of the response time to hippurate (a), cocaine (b) and benzoate of strains equipped with SEMP. No delay is observed between response time to the natural inducer benzoate and to the SEMP targets hippurate and cocaine whether in presence of 100μM or 500μM of inducers. Cells were prepared in the same way as for dose-response characterization experiments (see Methods). The longer response time for the cocaine sensor might be explained by the difference of temperature at which the experiment was performed (30°C instead of 37°C for hippurate sensor). RFU stands for relative fluorescence units and consists in the raw fluorescence value measured by the platereader. Dashed lines represent standard deviation of the three replicates.

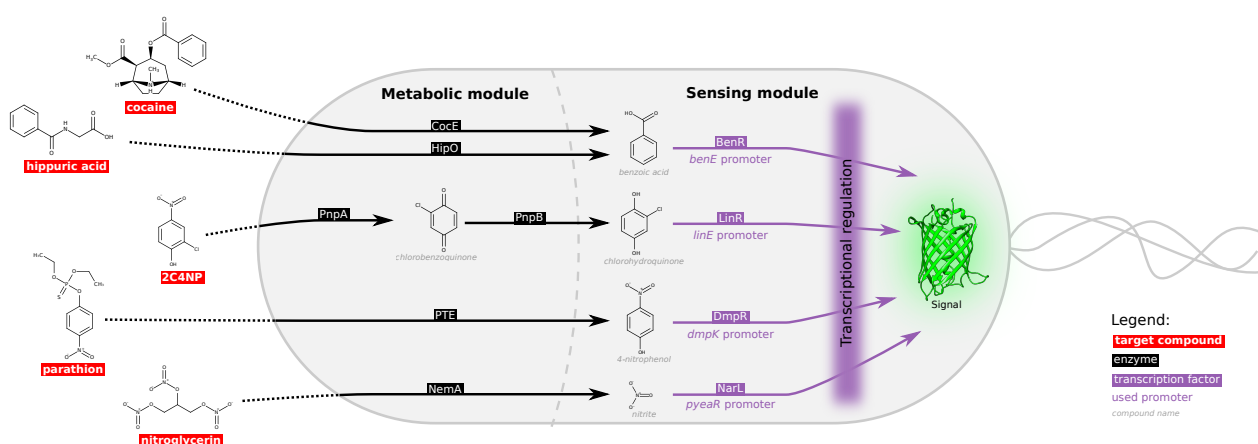


Fig. 6.8: Summary of the SEMP successfully implemented *in vivo* over the course of this work. Only one pair of the represented metabolic and sensing modules was transformed in each strain. Implementation of a sixth SEMP allowing detection of the pollutant chlorpropham (see Table 6.1) was not successful because we were unable to obtain a response to 3-chloroaniline in *E. coli* based on the TadR transcription factor from *Delftia tsuruhatensis*. This failure may be due to a gene induction mechanism that necessitates compatibility with the host transcription machinery or that requires a co-inducing protein. Intuitively, the choice of SEMP elements coming from organisms closely related to the host in terms of evolution is expected to maximize chances of successful implementation.

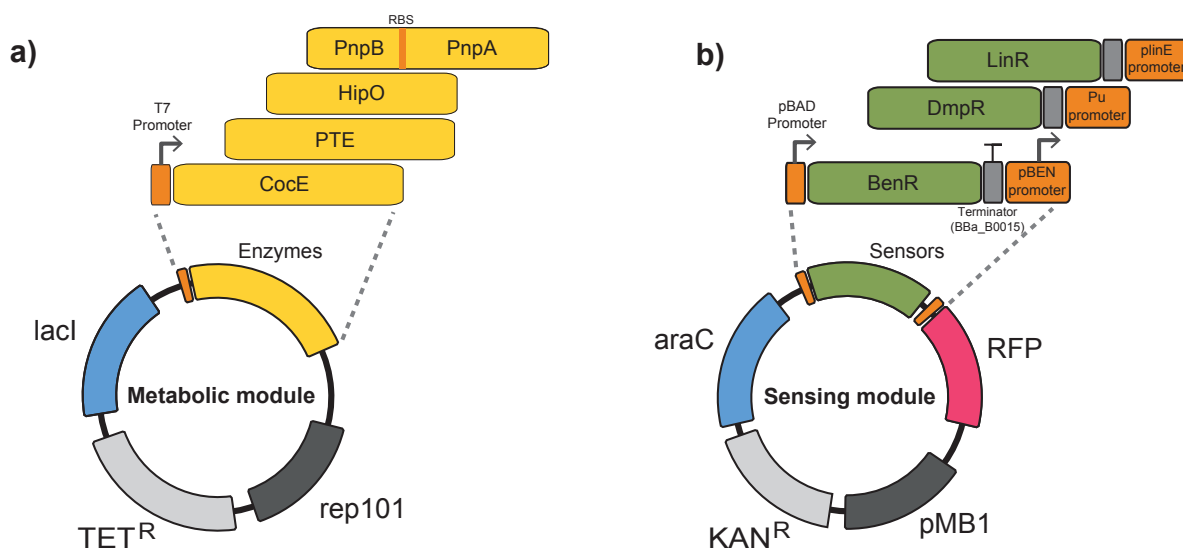


Fig. 6.9: Architecture of metabolic and sensing modules. (a) Map of the metabolic module harboring either CocE, PTE-S5, pnpB-pnpA or HipO coding sequences. (b) Map of the sensing module harboring either BenR-pBen, DmpR-Pu or LinR-plinE sequences to control RFP. Transcriptional isolation is ensured by BBa_B0015, a double terminator from the registry of standard biological parts.

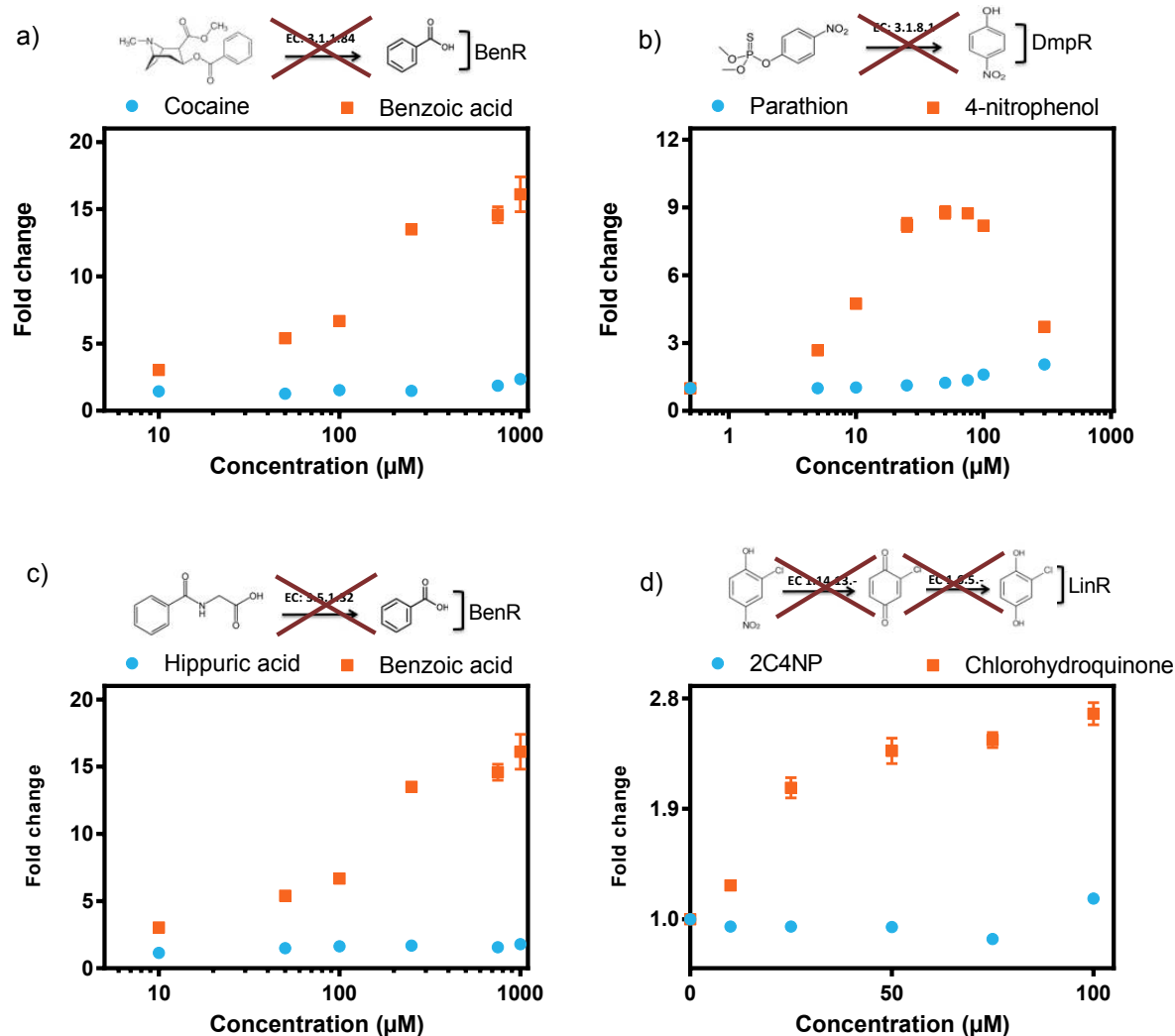


Fig. 6.10: Dose response of control strains lacking metabolic module to the different compounds. (a)(c) Benzoate sensing module alone; (b) 4-nitrophenol sensing module alone; (d) chlorohydroquinone sensing module alone. Dose-dependent response is limited to the natural effector of transcription factors BenR, DmpR, LinR (benzoic acid, 4-nitrophenol, chlorohydroquinone) when metabolic module is not present to enable sensing scope extension to target molecules. At high concentration of target molecules, traces of signal appeared even without metabolic module, this may be due to partial spontaneous hydrolysis, impurities from synthesis or promiscuity of the transcription factor.

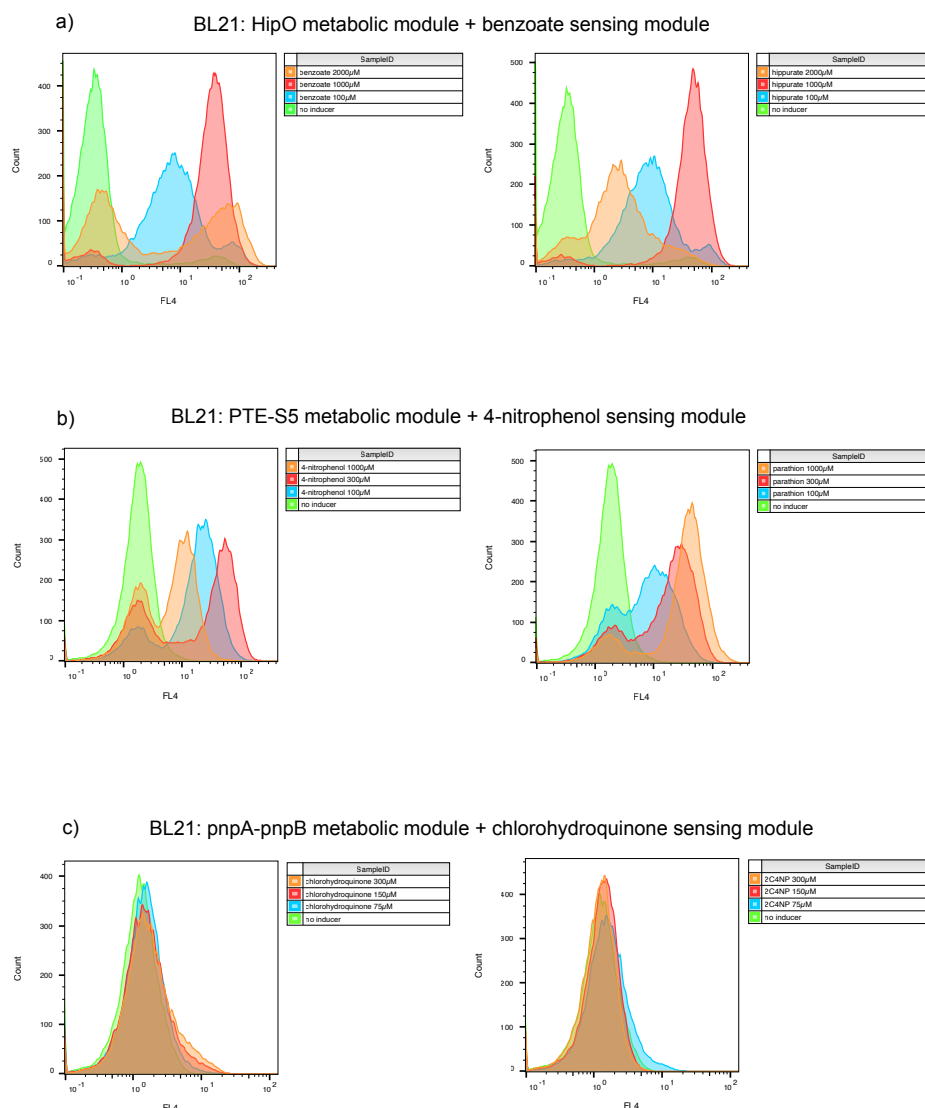
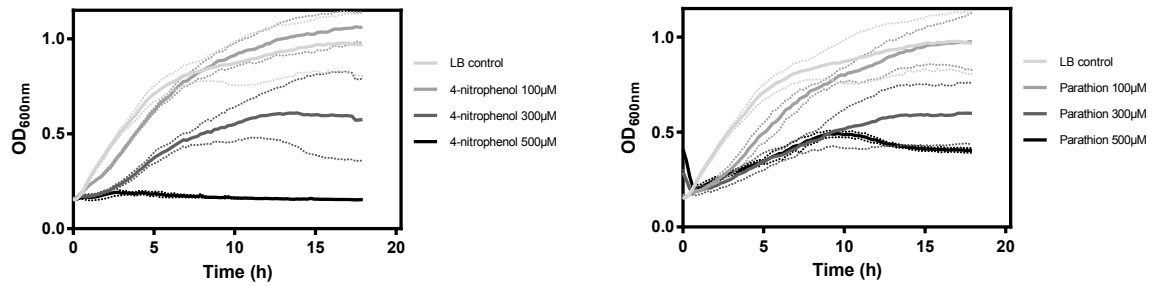
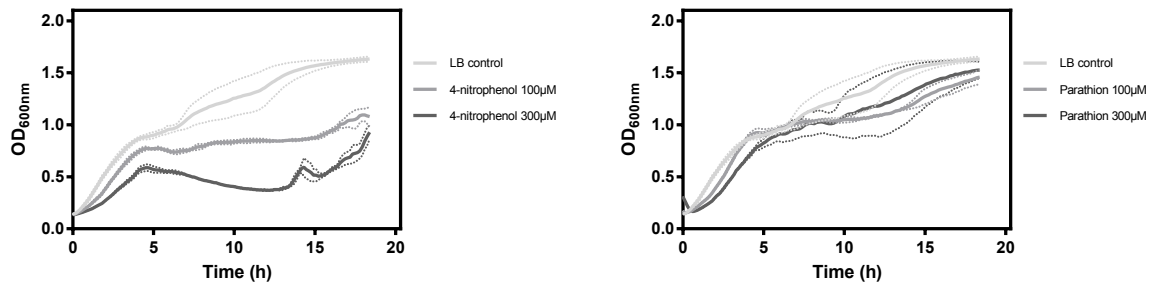


Fig. 6.12: Flow cytometry analysis of strains equipped with SEMP in presence of toxic concentrations of inducer. Cells were cultivated in the same conditions as for dose-response characterization (see Methods) and red fluorescence was analysed on a CyFlow Space cytometer (PARTEC) counting 20,000 events for each sample. (a) Toxicity of benzoic acid leads to bimodal distribution of cell fluorescence at 2mM and half of the population emits more fluorescence than with lower concentrations. Interestingly, hippuric acid toxicity has a different effect on cell response as fluorescence at 2mM decreases below the level observed with 100μM. (b) Toxicity of 4-nitrophenol leads to lower level of fluorescence at 2mM than at 100μM, similarly to what is observed with hippuric acid. On the contrary, proportionality of the response to parathion is maintained even at toxic levels. As shown in suppl. Fig 8, parathion by itself is not toxic unless PTE-S5 enzyme is present to transform it into 4-nitrophenol. This is probably the cause of the robust graded fluorescent response at high concentration although the exact mechanism is unknown. (c) Despite the low dynamic range of LinR, a small but graded response to chlorohydroquinone is observed while response to 2C4NP only occurs at low concentration (75μM) and is completely abolished upon reaching toxic levels.

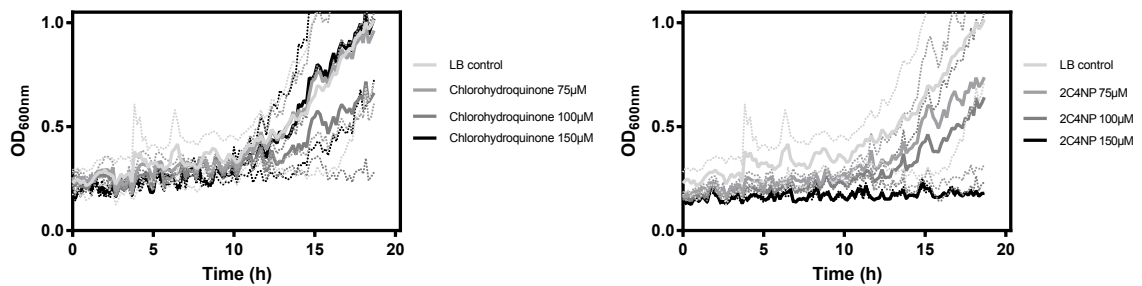
a) BL21: PTE-S5 metabolic module + 4-nitrophenol sensing module



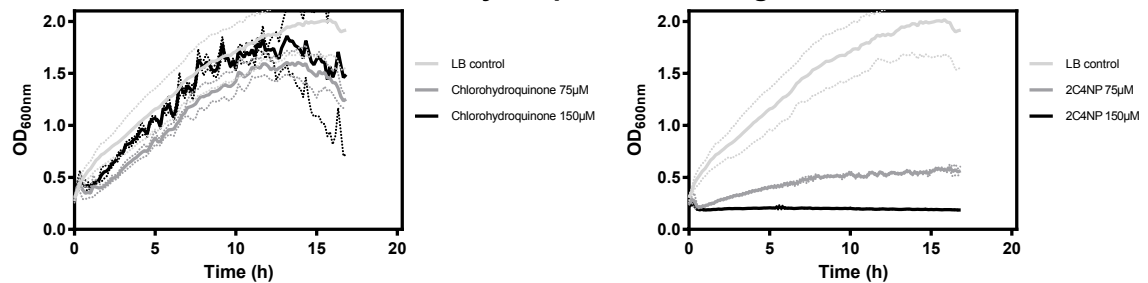
b) BL21: 4-nitrophenol sensing module



c) BL21: pnpB-pnpA metabolic module + chlorohydroquinone sensing module



d) BL21: chlorohydroquinone sensing module



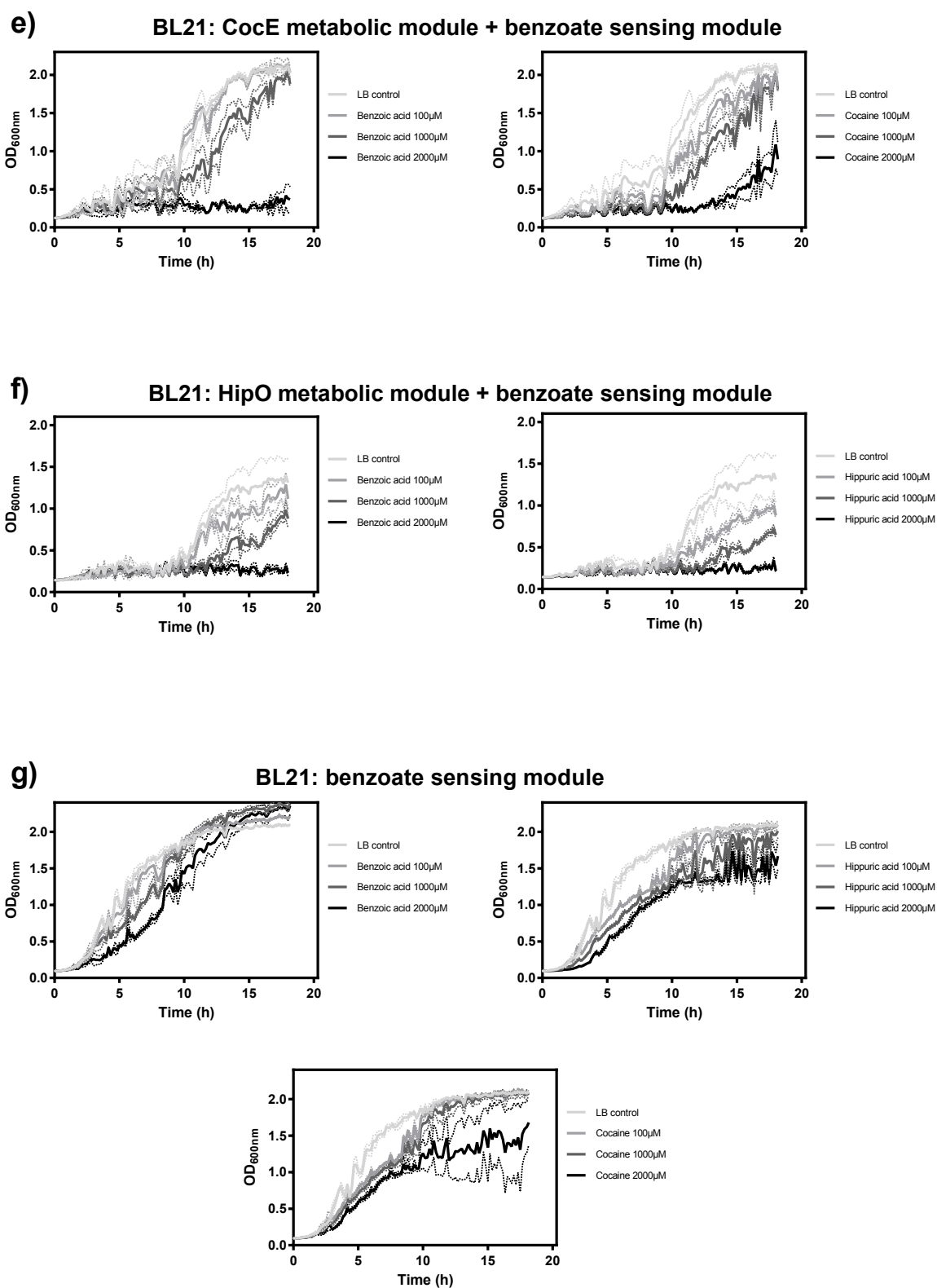
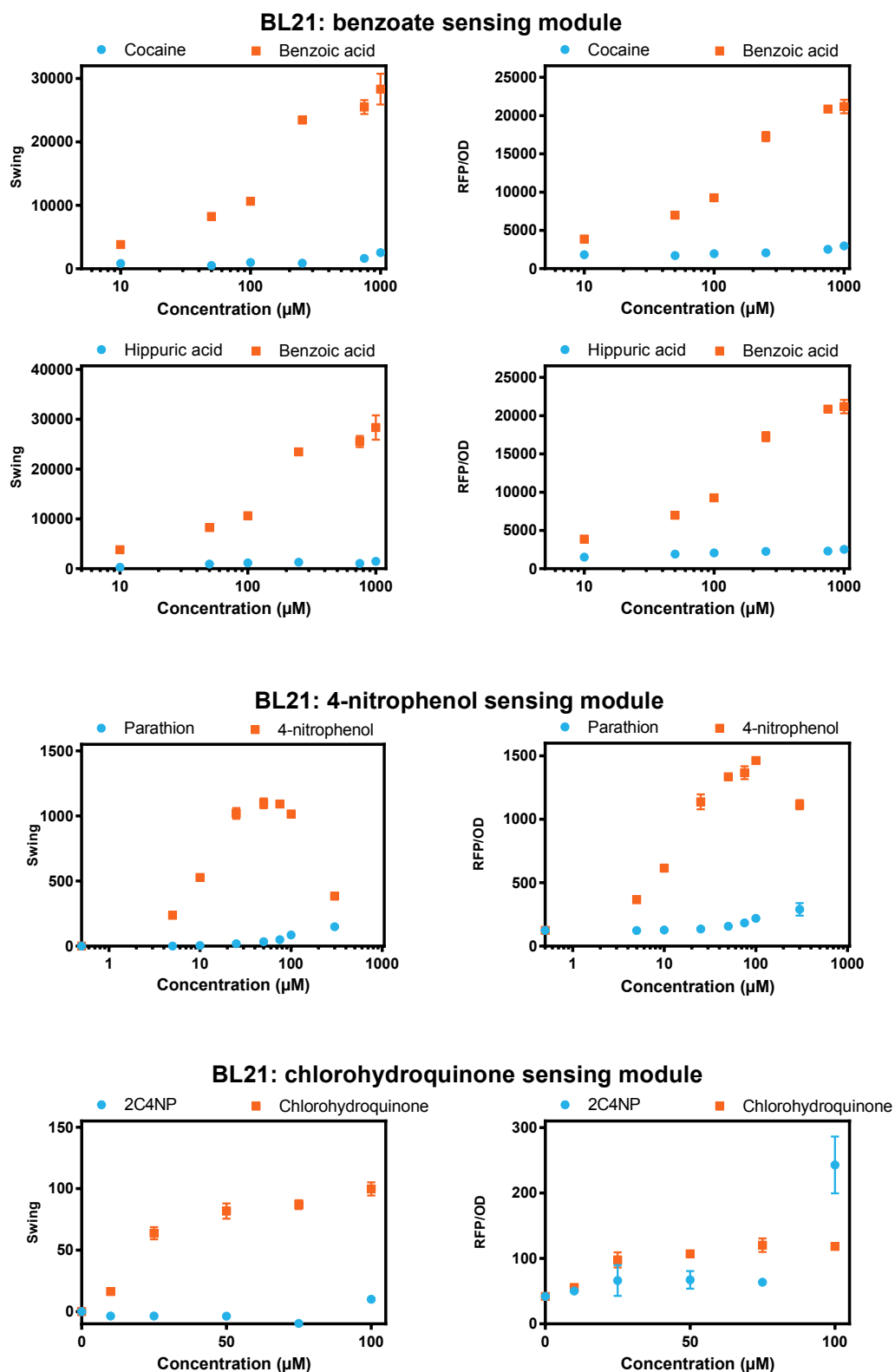


Fig. 6.13: Growth curves of the different strains in presence of the associated compounds.



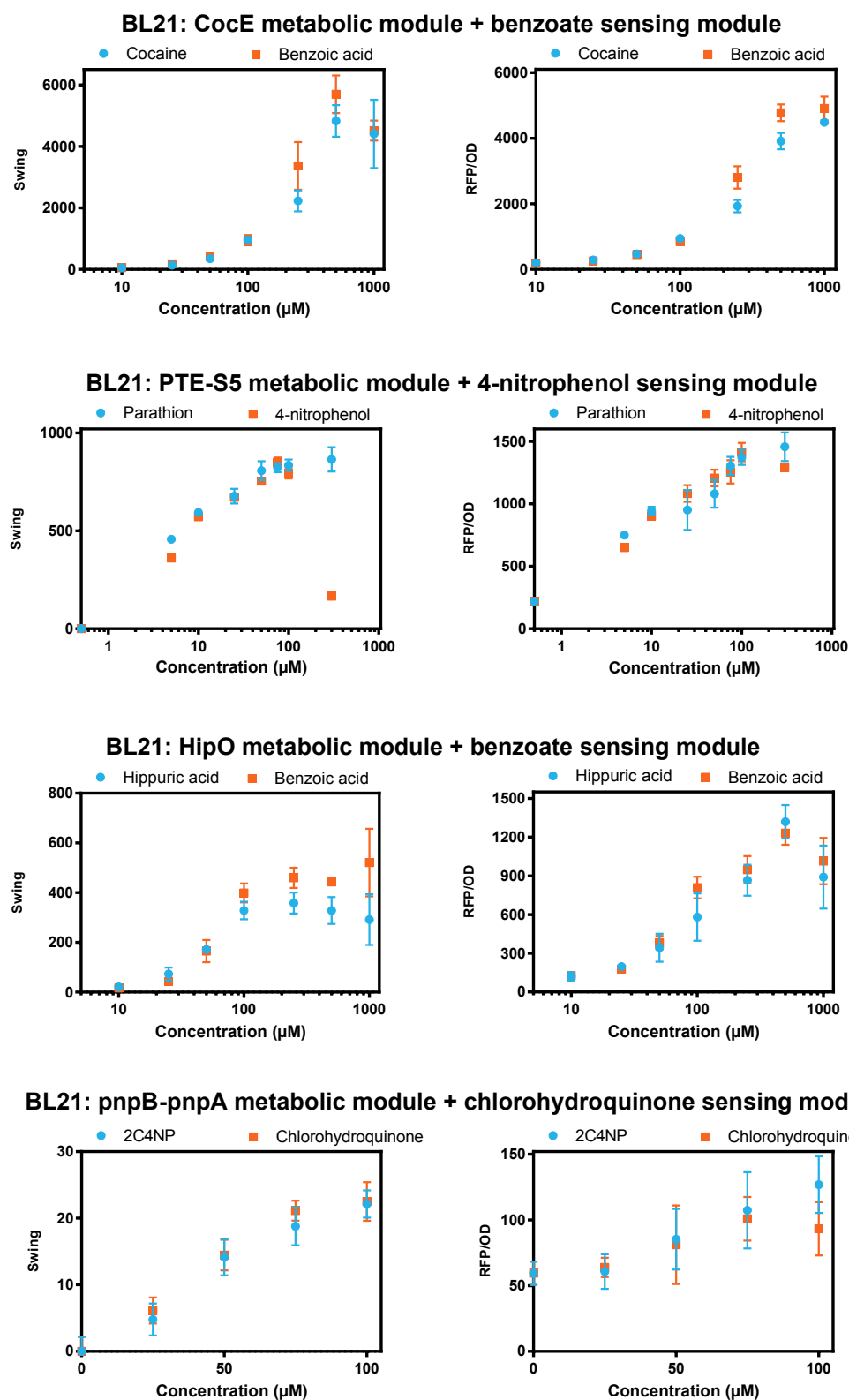


Fig. 6.14: Swing and RFP/OD of all strains with every compounds. Swing is defined as the difference in raw fluorescence observed between induced cells and non-induced control population.

Part II: conclusion & perspectives

Part II summarizes our work related to biosensor design, in particular the use of computer-aided pathway design softwares to find Sensing-Enabling Metabolic Pathways (SEMPs).

In Chapter 4 we reviewed the interest of allosteric transcription factors (aTF) as biosensors for metabolic engineering applications and different strategies that can be deployed to discover new biosensors for specific compounds of interest. Importantly, if the usage of biosensors is widespread in synthetic biology with the utilization of inducible promoters, the variety of detectable compounds is quite low in synthetic biology applications compared to what cells are capable of [241]. This is explained by the desire in that community to use well-characterized standard parts and should not be regarded as the limit of what can actually be used in term of biosensors. Indeed, it is becoming quite common to find [325, 326] and optimize [274, 275, 327] biosensors, even as a preliminary step of a bigger metabolic engineering study [328–331]. This trend should not go away anytime soon with the ever growing addition of microbial genomes in databases that will continue to provide new regulatory elements [332].

In Chapter 5 we presented SensiPath, a web application for the computer-aided design of sensing-enabling metabolic pathways (SEMPs). We collected a dataset of compounds detectable by allosteric transcription factors and used a new implementation of a pathway design algorithm to predict short metabolic circuits able to transform undetectable compounds into detectable ones. SensiPath has several interesting features that makes it more than a random computer aided pathway design software with a mapping of detectable compounds.

First, SensiPath is aware of compounds' stereochemistry thanks to an update of the molecular signatures [137]. This feature is still lacking from many reaction prediction softwares despite being one of the major advantages of enzymatic synthesis [12]. Stereochemical information is indeed difficult to gather from

metabolic databases since it may be missing, incomplete or incoherent both for compounds and reactions. It may also results in an additional combinatorial complexity since chemicals are stored with all their stereoisomers. That is why the easiest path is simply to ignore stereochemical data. Let us note that the way SensiPath deals with stereochemistry is still quite naive (compounds are stored both with and without stereo) and subject to improvements.

We can foresee that the exploitation of chemical ontologies like the one proposed by ChEBI [333] will greatly help to store and interpret stereochemical relationship between compounds in a meaningful way. It will also have a salvative impact on the interpretation of compounds (and reactions) with partial structures (Markush structures, R-groups) that are typically filtered out while creating reaction rules, at the risk of losing important reaction mechanisms. In both cases, processing concepts rather than actual structures could give insights to limit computational burden by applying rules directly on class of compounds (e.g. "an alcohol") instead of each of its members (e.g. ethanol, methanol, etc.). This idea would be particularly well-adapted to a scenario in which users would interactively decide how to consider a compound, i.e. with or without stereo information, as a characteristic instance of a class of compounds, etc.

Second, SensiPath computes pathways on-demand and stores the results in a database to serve future identical queries. This is quite different from other reaction prediction framework such as XTMS [81] that use a precomputed Extended Metabolic Space (EMS) to find pathways, and thus cannot work with input compounds absent from the EMS. SensiPath is closer to an interactive use, like what can be seen with MetaRoute [334], MRSD [335], and BioSynther [91].

In truth, both stereochemical perception and real-time computation features were technical prototypes to prepare a new implementation of XTMS with additional data. RetroPath2, that we presented in Chapter 3, superseded those plans by giving users the flexibility to compute any kind of metabolic circuit [71], including SEMP. The dataset of 500 detectable compounds that is the main novelty of SensiPath from an application standpoint is open-source and can readily be used by other reaction prediction frameworks.

In Chapter 6 we presented the *in vivo* validation of the SEMP concept with the design of four new whole-cell biosensors, in addition of some statistics related to the number of compounds that could theoretically benefit from such an approach. Our main finding is that sensing through the metabolism significantly increases the number of readily detectable compounds, as we doubled this number based solely on our list of 500 effectors of allosteric transcription factors.

Interestingly, the main limitation of the prediction of SEMP was due to the fact that many target compounds would not be transformed even once by any of our reaction rules. Since this observation stands for compounds from the Human Metabolite DataBase, the natural explanation would be that we were missing reaction mechanisms. This highlights once again the importance of an exhaustive reaction rules dataset. We could have used a more permissive enzymatic promiscuity hypothesis while designing the rules (lower diameter, see Chapter 2), but this would have generated more false positives and ultimately slowed down our *in vivo* experimental validation effort. Thus, we can expect much more compounds to be identified as detectable with a broader dataset of enzymatic reactions. A thorough dataset of detectable compounds would also greatly benefit our predictions.

Some of our predictions lead to cofactors or other commodity metabolites that are endogenous to the chassis, but this is not detrimental to their practicality as detectable compounds. Instinctively, we could think that this would lead to an overwhelming specificity problem, in addition to a sensitivity challenge since cells would react with their own regulation mechanism to a non-orthogonal SEMP. However, we showed with the nitrite biosensor (used for nitroglycerin detection) that the detection of an endogenous compound is possible and could be used in SEMP design.

The development of generalist biosensors is highly desirable since they can be used for many applications. As a matter of fact, the intracellular detection and fluorescent report of NAD(P)H/NAD(P) ratios was the subject of intense research. In 2013, Siedler *et al.* used SoxR to sense NADPH levels in *E. coli* and showed in a proof-of-concept that it was suitable to screen alcohol dehydrogenase variants by the impact of their activity on NADPH level (that is a cofactor) [336]. They successfully increased the activity for an alternative substrate by 38% in a single screen and proposed that this method could be used for the guided evolution of all NADPH-dependent enzymes. Similarly, in 2016 Keasling's group reported a NADPH/NADP⁺ yeast biosensor that could be used for the dynamic regulation of the redox state of cell factories [337]. Therefore, it appears that the detection of cofactors that are widely abundant in cells is not only feasible but also desirable for metabolic engineering applications.

The SEMP approach was tested by numerous *in vivo* applications. De Lorenzo's lab devised a similar strategy as soon as 2006 for the discovery of specific novel enzymatic activities with a "genetic trap" that would report the transformation of a substrate (supplemented) into a known detectable product. They proposed a proof-of-concept with the search of dehydrochlorination enzymes that would process lindane (a pesticide) into 1,2,4-trichlorobenzene (effector of an evolved

XylR) [130]. In 2014, De Lorenzo used a SEMP-like concept to engineer a communication channel between two *Pseudomonas Putida* strains using benzoic acid as a messenger for the detection of aromatic compounds [308]. Xue *et al.* detected salicylaldehyde by a one-step transformation into salicylate that is detectable by NahR [274] and proposed to use it in association with other aTF to make whole-cell biosensors for the detection of a variety of aromatic pollutants. Chong *et al.* developed a whole-cell biosensor with a pigment output for organophosphate pesticides detection using enzymatic promiscuity to transform parathion, methylparathion, and fenitrothion into 4-nitrophenol that is readily detectable by DmpR [338]. As we already mentioned, Church's lab used the SEMP approach to screen for variants producing 3-hydroxypropionate [275]. All those results comfort the interest of SEMP's methodology.

In the end, the perfect biosensor depends heavily on the final application and the resources that are at disposition for the project [110]. SEMP have the merit to require few research effort if enzymes and aTF are available and are well-expressed in the chassis organism. However, direct sensing should be preferred most of the time since it allows a better handle on biosensors' parameters [327]. Interestingly, we should note that systems that were heavily optimized by evolutive processes are quite different from what we could think of when keeping engineering principles in mind (abstract hierarchical thinking, modularization and decoupling, part standardization). One of the best examples might be found in enzymes that, in addition of their catalytic function, also harbor an allosteric mechanism that is controlled by metabolites of their pathway to regulate their efficiency. This is indeed an ideal answer to the dynamic regulation problem since it is much faster than anything necessitating a round of gene expression. If strategies exist to engineer such allosteric mechanisms [339, 340], they require case-by-case protein engineering that does not seem amenable yet to automated design in metabolic engineering.

Overall, we argue in this thesis that the design of *de novo* metabolic pathways greatly benefits from the integration of exotic data since they can bring innovative solutions before the eyes of a human expert. We discussed in Part I how CAD softwares could build users' trust and encourage data exploration. In Part II, we showed that the simple addition of a detectable compounds dataset revealed a new utilization for our reaction prediction framework. We envision that the integration of other datasets answering questions that are somehow related to pathway design ("which compounds can be bought?", "which are toxics?", "which pass the membrane?", etc.) will contribute too to the emergence of new applications and to the perceived added-value of CAD softwares by users. In that context, we need to go toward a better integration of predictive models, notably Quantitative Structure Activity Relationship (QSAR), in order to predict relevant

properties. Importantly, to maintain their trust, users must be kept in control of the degree of confidence in data used by CAD softwares. This will be amenable by more interactivity between users and bioCAD softwares. In that spirit, we should take a leaf out of other CAD softwares' book, notably those for architecture and product design that adopted along the years plenty of features to empower their users' creativity and to enhance their collaborativity.

We expect that the advances in CAD will continue to lower the amount of necessary prerequisite expert knowledge for biological engineers to express their creativity, and we believe this will eventually contribute to the flourishing of novel and smart metabolic engineering applications to solve the grand challenges our society is facing. In this regard, let us quote a word of wisdom from Kasparov, a chess-player that was considered the best of all time when he was defeated in a ground-breaking game by the brute force power of Deep Blue, IBM's artificial intelligence (AI). Later, Kasparov conducted a rich reflection on AI's potential for humankind and remains to this day one of its best advocates. Interestingly, he designed a new kind of chess (freestyle chess) in which human players are each paired with an AI to free their creativity and came to some exciting conclusions about human-computer interaction:

"I reached the formulation that a weak human player plus machine plus a better process is superior, not only to a very powerful machine, but most remarkably, to a strong human player plus machine plus an inferior process.

At the end of the day, it's about interface. Creating an interface that will help us to coach machine towards more useful intelligence will be the right step forward." ([@2])

Bibliography

- [1] Ksenia S. Egorova and Valentine P. Ananikov. “Which Metals are Green for Catalysis? Comparison of the Toxicities of Ni, Cu, Fe, Pd, Pt, Rh, and Au Salts”. In: *Angewandte Chemie International Edition* 55.40 (Sept. 26, 2016), pp. 12150–12162. DOI: 10.1002/anie.201603777 (cit. on p. 1).
- [2] C.-J. Li and B. M. Trost. “Green chemistry for chemical synthesis”. In: *Proceedings of the National Academy of Sciences* 105.36 (Sept. 9, 2008), pp. 13197–13202. DOI: 10.1073/pnas.0804348105 (cit. on p. 1).
- [3] Amanda J. Hickman and Melanie S. Sanford. “High-valent organometallic copper and palladium in catalysis”. In: *Nature* 484.7393 (Apr. 11, 2012), pp. 177–185. DOI: 10.1038/nature11008 (cit. on p. 1).
- [4] Rafael Chinchilla and Carmen Nájera. “The Sonogashira Reaction: A Booming Methodology in Synthetic Organic Chemistry[†]”. In: *Chemical Reviews* 107.3 (Mar. 2007), pp. 874–922. DOI: 10.1021/cr050992x (cit. on p. 1).
- [5] Herbert Danner and Rudolf Braun. “Biotechnology for the production of commodity chemicals from biomass”. In: *Chemical Society Reviews* 28.6 (1999), pp. 395–405. DOI: 10.1039/a806968i (cit. on pp. 1–3).
- [6] Lindsay Soh and Matthew J. Eckelman. “Green Solvents in Biomass Processing”. In: *ACS Sustainable Chemistry & Engineering* 4.11 (Nov. 7, 2016), pp. 5821–5837. DOI: 10.1021/acssuschemeng.6b01635 (cit. on p. 1).
- [7] M.G. Adsul, M.S. Singhvi, S.A. Gaikawari, and D.V. Gokhale. “Development of biocatalysts for production of commodity chemicals from lignocellulosic biomass”. In: *Bioresource Technology* 102.6 (Mar. 2011), pp. 4304–4312. DOI: 10.1016/j.biortech.2011.01.002 (cit. on p. 2).
- [8] Mustafa Vohra, Jagdish Manwar, Rahul Manmode, Satish Padgilwar, and Sanjay Patil. “Bioethanol production: Feedstock and current technologies”. In: *Journal of Environmental Chemical Engineering* 2.1 (Mar. 2014), pp. 573–584. DOI: 10.1016/j.jece.2013.10.013 (cit. on p. 2).
- [9] Guozhan Jiang, David Hill, Marek Kowalczyk, et al. “Carbon Sources for Polyhydroxyalkanoates and an Integrated Biorefinery”. In: *International Journal of Molecular Sciences* 17.7 (July 19, 2016), p. 1157. DOI: 10.3390/ijms17071157 (cit. on p. 2).
- [10] Jens Nielsen and Jay D. Keasling. “Engineering Cellular Metabolism”. In: *Cell* 164.6 (Mar. 10, 2016), pp. 1185–1197. DOI: 10.1016/j.cell.2016.02.004 (cit. on pp. 2, 4, 6).

- [11] Vinuselvi Parisutham, Sathesh-Prabu Chandran, Aindrila Mukhopadhyay, Sung Kuk Lee, and Jay D. Keasling. "Intracellular cellobiose metabolism and its applications in lignocellulose-based biorefineries". In: *Bioresource Technology* 239 (Sept. 2017), pp. 496–506. DOI: 10.1016/j.biortech.2017.05.001 (cit. on p. 2).
- [12] H. E. Schoemaker. "Dispelling the Myths—Biocatalysis in Industrial Synthesis". In: *Science* 299.5613 (Mar. 14, 2003), pp. 1694–1697. DOI: 10.1126/science.1079237 (cit. on pp. 2, 133).
- [13] Adrie J. J. Straathof, Sven Panke, and Andreas Schmid. "The production of fine chemicals by biotransformations". In: *Current Opinion in Biotechnology* 13.6 (Dec. 2002), pp. 548–556 (cit. on p. 2).
- [14] Fernando Lopez-Gallego and Claudia Schmidt-Dannert. "Multi-enzymatic synthesis". In: *Current Opinion in Chemical Biology* 14.2 (Apr. 2010), pp. 174–183. DOI: 10.1016/j.cbpa.2009.11.023 (cit. on p. 2).
- [15] Josephine W. Reed and Tomas Hudlicky. "The Quest for a Practical Synthesis of Morphine Alkaloids and Their Derivatives by Chemoenzymatic Methods". In: *Accounts of Chemical Research* 48.3 (Mar. 17, 2015), pp. 674–687. DOI: 10.1021/ar500427k (cit. on p. 2).
- [16] Paul R. Blakemore and James D. White. "Morphine, the Proteus of organic molecules". In: *Chemical Communications* 11 (May 17, 2002), pp. 1159–1168. DOI: 10.1039/b111551k (cit. on p. 3).
- [17] B. Kamm and M. Kamm. "Principles of biorefineries". In: *Applied Microbiology and Biotechnology* 64.2 (Apr. 2004), pp. 137–145. DOI: 10.1007/s00253-003-1537-7 (cit. on p. 3).
- [18] Jay D. Keasling. "Manufacturing Molecules Through Metabolic Engineering". In: *Science* 330.6009 (Dec. 3, 2010), pp. 1355–1358. DOI: 10.1126/science.1193990 (cit. on p. 3).
- [19] J. Bailey. "Toward a science of metabolic engineering". In: *Science* 252.5013 (June 21, 1991), pp. 1668–1675. DOI: 10.1126/science.2047876 (cit. on p. 4).
- [20] Volker F. Wendisch, Luciana Fernandes Brito, Marina Gil Lopez, et al. "The flexible feedstock concept in Industrial Biotechnology: Metabolic engineering of *Escherichia coli*, *Corynebacterium glutamicum*, *Pseudomonas*, *Bacillus* and yeast strains for access to alternative carbon sources". In: *Journal of Biotechnology* 234 (Sept. 2016), pp. 139–157. DOI: 10.1016/j.jbiotec.2016.07.022 (cit. on p. 4).
- [21] Jeong Wook Lee, Tae Yong Kim, Yu-Sin Jang, Sol Choi, and Sang Yup Lee. "Systems metabolic engineering for chemicals and materials". In: *Trends in Biotechnology* 29.8 (Aug. 2011), pp. 370–378. DOI: 10.1016/j.tibtech.2011.04.001 (cit. on p. 4).
- [22] Sang Yup Lee and Hyun Uk Kim. "Systems strategies for developing industrial microbial strains". In: *Nature Biotechnology* 33.10 (Oct. 2015), pp. 1061–1072. DOI: 10.1038/nbt.3365 (cit. on pp. 4, 39).

- [23] Jeong Wook Lee, Dokyun Na, Jong Myoung Park, et al. “Systems metabolic engineering of microorganisms for natural and non-natural chemicals”. In: *Nature Chemical Biology* 8.6 (May 17, 2012), pp. 536–546. DOI: 10.1038/nchembio.970 (cit. on pp. 4, 5).
- [24] Yunzi Luo, Bing-Zhi Li, Duo Liu, et al. “Engineered Biosynthesis of Natural Products in Heterologous Hosts”. In: *Chemical Society reviews* 44.15 (Aug. 7, 2015), pp. 5265–5290. DOI: 10.1039/c5cs00025d (cit. on p. 4).
- [25] Judith Becker and Christoph Wittmann. “Advanced Biotechnology: Metabolically Engineered Cells for the Bio-Based Production of Chemicals and Fuels, Materials, and Health-Care Products”. In: *Angewandte Chemie International Edition* 54.11 (Mar. 9, 2015), pp. 3328–3350. DOI: 10.1002/anie.201409033 (cit. on p. 4).
- [26] Stephan C. Hammer, Anders M. Knight, and Frances H. Arnold. “Design and evolution of enzymes for non-natural chemistry”. In: *Current Opinion in Green and Sustainable Chemistry*. 7 New Synthetic Methods 2017 7 (Oct. 1, 2017), pp. 23–30. DOI: 10.1016/j.cogsc.2017.06.002 (cit. on p. 5).
- [27] Justin B. Siegel, Amanda Lee Smith, Sean Poust, et al. “Computational protein design enables a novel one-carbon assimilation pathway”. In: *Proceedings of the National Academy of Sciences* (Mar. 9, 2015), p. 201500545. DOI: 10.1073/pnas.1500545112 (cit. on p. 5).
- [28] Stephanie Galanie, Kate Thodey, Isis J. Trenchard, Maria Filsinger Interrante, and Christina D. Smolke. “Complete biosynthesis of opioids in yeast”. In: *Science (New York, N.Y.)* 349.6252 (Sept. 4, 2015), pp. 1095–1100. DOI: 10.1126/science.aac9373 (cit. on pp. 5, 114).
- [29] Harry Yim, Robert Haselbeck, Wei Niu, et al. “Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol”. In: *Nature Chemical Biology* 7.7 (May 22, 2011), pp. 445–452. DOI: 10.1038/nchembio.580 (cit. on pp. 5, 6, 39–42, 44, 46, 49, 55, 71, 77).
- [30] Kevin W. George, Jorge Alonso-Gutierrez, Jay D. Keasling, and Taek Soon Lee. “Isoprenoid Drugs, Biofuels, and Chemicals—Artemisinin, Farnesene, and Beyond”. In: *Biotechnology of Isoprenoids*. Ed. by Jens Schrader and Jörg Bohlmann. Vol. 148. DOI: 10.1007/10_2014_288. Cham: Springer International Publishing, 2015, pp. 355–389 (cit. on p. 5).
- [31] Dae-Kyun Ro, Eric M. Paradise, Mario Ouellet, et al. “Production of the antimalarial drug precursor artemisinic acid in engineered yeast”. In: *Nature* 440.7086 (Apr. 13, 2006), pp. 940–943. DOI: 10.1038/nature04640 (cit. on p. 6).
- [32] C. J. Paddon, P. J. Westfall, D. J. Pitera, et al. “High-level semi-synthetic production of the potent antimalarial artemisinin”. In: *Nature* 496.7446 (Apr. 10, 2013), pp. 528–532. DOI: 10.1038/nature12051 (cit. on pp. 6, 39).
- [33] Chris J. Paddon and Jay D. Keasling. “Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development”. In: *Nature Reviews Microbiology* 12.5 (May 2014), pp. 355–367. DOI: 10.1038/nrmicro3240 (cit. on p. 6).

- [34] Anthony Burgard, Mark J. Burk, Robin Osterhout, Stephen Van Dien, and Harry Yim. “Development of a commercial scale process for production of 1,4-butanediol from sugar”. In: *Current Opinion in Biotechnology* 42 (Dec. 2016), pp. 118–125. DOI: 10.1016/j.copbio.2016.04.016 (cit. on p. 6).
- [35] Marnix H. Medema, Renske van Raaphorst, Eriko Takano, and Rainer Breitling. “Computational tools for the synthetic design of biochemical pathways”. In: *Nature Reviews Microbiology* 10.3 (Jan. 23, 2012), pp. 191–202. DOI: 10.1038/nrmicro2717 (cit. on pp. 6, 11, 39, 108).
- [36] A. Seressiotis and J. E. Bailey. “MPS: An artificially intelligent software system for the analysis and synthesis of metabolic pathways”. In: *Biotechnology and Bioengineering* 31.6 (Apr. 20, 1988), pp. 587–602. DOI: 10.1002/bit.260310611 (cit. on p. 6).
- [37] Ferenc Darvas. “Predicting metabolic pathways by logic programming”. In: *Journal of Molecular Graphics* 6.2 (June 1, 1988), pp. 80–86. DOI: 10.1016/0263-7855(88)85004-5 (cit. on p. 6).
- [38] Michael L. Mavrovouniotis, George Stephanopoulos, and Gregory Stephanopoulos. “Computer-aided synthesis of biochemical pathways”. In: *Biotechnology and Bioengineering* 36.11 (Dec. 20, 1990), pp. 1119–1132. DOI: 10.1002/bit.260361107 (cit. on p. 6).
- [39] J. M. Savinell and B. O. Palsson. “Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism”. In: *Journal of Theoretical Biology* 154.4 (Feb. 21, 1992), pp. 421–454 (cit. on p. 6).
- [40] Jonathan D Tyzack, Nicholas Furnham, Ian Sillitoe, Christine M Orengo, and Janet M Thornton. “Understanding enzyme function evolution from a computational perspective”. In: *Current Opinion in Structural Biology* 47 (Dec. 2017), pp. 131–139. DOI: 10.1016/j.sbi.2017.08.003 (cit. on p. 7).
- [41] R. A. Jensen. “Enzyme recruitment in evolution of new function”. In: *Annual Review of Microbiology* 30 (1976), pp. 409–425. DOI: 10.1146/annurev.mi.30.100176.002205 (cit. on p. 7).
- [42] Takuji Yamada and Peer Bork. “Evolution of biomolecular networks — lessons from metabolic and protein interactions”. In: *Nature Reviews Molecular Cell Biology* 10.11 (Nov. 2009), pp. 791–803. DOI: 10.1038/nrm2787 (cit. on p. 7).
- [43] P. Carbonell, G. Lecointre, and J.-L. Faulon. “Origins of Specificity and Promiscuity in Metabolic Networks”. In: *Journal of Biological Chemistry* 286.51 (Dec. 23, 2011), pp. 43994–44004. DOI: 10.1074/jbc.M111.274050 (cit. on pp. 7, 50).
- [44] Hojung Nam, Nathan E. Lewis, Joshua A. Lerman, et al. “Network Context and Selection in the Evolution to Enzyme Specificity”. In: *Science* 337.6098 (Aug. 31, 2012), pp. 1101–1104. DOI: 10.1126/science.1216861 (cit. on pp. 7, 41, 57, 59, 68).
- [45] Masahito Hosokawa, Yuri Hoshino, Yohei Nishikawa, et al. “Droplet-based microfluidics for high-throughput screening of a metagenomic library for isolation of microbial enzymes”. In: *Biosensors & Bioelectronics* 67 (May 15, 2015), pp. 379–385. DOI: 10.1016/j.bios.2014.08.059 (cit. on p. 7).

- [46] Gabriela I. Guzmán, José Utrilla, Sergey Nurk, et al. “Model-driven discovery of underground metabolic functions in *Escherichia coli*”. In: *Proceedings of the National Academy of Sciences of the United States of America* 112.3 (Jan. 20, 2015), pp. 929–934. DOI: 10.1073/pnas.1414218112 (cit. on pp. 7, 9, 57, 59).
- [47] B. van Loo, S. Jonas, A. C. Babbie, et al. “An efficient, multiply promiscuous hydrolase in the alkaline phosphatase superfamily”. In: *Proceedings of the National Academy of Sciences* 107.7 (Feb. 16, 2010), pp. 2740–2745. DOI: 10.1073/pnas.0903951107 (cit. on p. 7).
- [48] Ann Babbie, Nobuhiko Tokuriki, and Florian Hollfelder. “What makes an enzyme promiscuous?” In: *Current Opinion in Chemical Biology* 14.2 (Apr. 2010), pp. 200–207. DOI: 10.1016/j.cbpa.2009.11.028 (cit. on p. 7).
- [49] Sandeep Chakraborty, Renu Minda, Lipika Salaye, et al. “Promiscuity-based enzyme selection for rational directed evolution experiments”. In: *Methods in Molecular Biology (Clifton, N.J.)* 978 (2013), pp. 205–216. DOI: 10.1007/978-1-62703-293-3_15 (cit. on p. 7).
- [50] Tomer Altman, Michael Travers, Anamika Kothari, Ron Caspi, and Peter D Karp. “A systematic comparison of the MetaCyc and KEGG pathway databases”. In: *BMC Bioinformatics* 14.1 (2013), p. 112. DOI: 10.1186/1471-2105-14-112 (cit. on pp. 7, 41).
- [51] Antje Chang, Ida Schomburg, Sandra Placzek, et al. “BRENDA in 2015: exciting developments in its 25th year of existence”. In: *Nucleic Acids Research* 43 (Database issue Jan. 28, 2015), pp. D439–D446. DOI: 10.1093/nar/gku1068 (cit. on pp. 7, 30, 41, 67, 96, 110, 117).
- [52] John F. Tinker and Herbert Gelernter. “Computer Simulation of Metabolic Transformation”. In: *Journal of Computational Chemistry* 7.5 (Oct. 1986), pp. 657–665. DOI: 10.1002/jcc.540070508 (cit. on p. 7).
- [53] Gilles Klopman, Mario Dimayuga, and Joseph Talafous. “META. 1. A Program for the Evaluation of Metabolic Transformation of Chemicals”. In: *Journal of Chemical Information and Computer Sciences* 34.6 (Nov. 1, 1994), pp. 1320–1325. DOI: 10.1021/ci00022a014 (cit. on p. 7).
- [54] N. Greene, P. N. Judson, J. J. Langowski, and C. A. Marchant. “Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR”. In: *SAR and QSAR in environmental research* 10.2 (1999), pp. 299–314. DOI: 10.1080/10629369908039182 (cit. on p. 7).
- [55] Bo Kyeng Hou, Lawrence P. Wackett, and Lynda B. M. Ellis. “Microbial pathway prediction: a functional group approach”. In: *Journal of Chemical Information and Computer Sciences* 43.3 (June 2003), pp. 1051–1057. DOI: 10.1021/ci034018f (cit. on pp. 7, 8, 42).
- [56] Raoul Walther, Jarkko Rautio, and Alexander N. Zelikin. “Prodrugs in medicinal chemistry and enzyme prodrug therapies”. In: *Advanced Drug Delivery Reviews* (July 2017). DOI: 10.1016/j.addr.2017.06.013 (cit. on p. 7).
- [57] E. J. Corey and W. T. Wipke. “Computer-Assisted Design of Complex Organic Syntheses”. In: *Science* 166.3902 (Oct. 10, 1969), pp. 178–192. DOI: 10.1126/science.166.3902.178 (cit. on pp. 7, 10, 11).

- [58] Sara Szymkuć, Ewa P. Gajewska, Tomasz Klucznik, et al. "Computer-Assisted Synthetic Planning: The End of the Beginning". In: *Angewandte Chemie International Edition* 55.20 (May 10, 2016), pp. 5904–5937. DOI: 10.1002/anie.201506101 (cit. on pp. 8, 11, 77).
- [59] Daylight Chemical Information Systems, Inc. *Daylight Theory Manual* (cit. on pp. 8, 41, 49, 66).
- [60] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of Chemical Information and Modeling* 28.1 (Feb. 1, 1988), pp. 31–36. DOI: 10.1021/ci00057a005 (cit. on p. 8).
- [61] Eric Gifford, Mark Johnson, and Chun-che Tsai. "A graph-theoretic approach to modeling metabolic pathways". In: *Journal of Computer-Aided Molecular Design* 5.4 (Aug. 1991), pp. 303–322. DOI: 10.1007/BF00126665 (cit. on p. 8).
- [62] Miguel A. Campodonico, Barbara A. Andrews, Juan A. Asenjo, Bernhard O. Palsson, and Adam M. Feist. "Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path". In: *Metabolic Engineering* 25 (Sept. 2014), pp. 140–158. DOI: 10.1016/j.ymben.2014.07.009 (cit. on pp. 8, 10, 12, 39, 40, 46, 47, 56, 76).
- [63] Matthew A. Kayala and Pierre Baldi. "ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning". In: *Journal of Chemical Information and Modeling* 52.10 (Oct. 22, 2012), pp. 2526–2540. DOI: 10.1021/ci3003039 (cit. on pp. 8, 76).
- [64] Tadi Venkata Sivakumar, Varun Giri, Jin Hwan Park, Tae Yong Kim, and Anirban Bhaduri. "ReactPRED: a tool to predict and analyze biochemical reactions". In: *Bioinformatics* 32.22 (Nov. 15, 2016), pp. 3522–3524. DOI: 10.1093/bioinformatics/btw491 (cit. on pp. 8, 10, 44, 76).
- [65] Chunhui Li, Christopher S. Henry, Matthew D. Jankowski, et al. "Computational discovery of biochemical routes to specialty chemicals". In: *Chemical Engineering Science*. ISCRE18 59.22 (Nov. 2004), pp. 5051–5060. DOI: 10.1016/j.ces.2004.09.021 (cit. on pp. 8–10).
- [66] Mina Oh, Takuji Yamada, Masahiro Hattori, Susumu Goto, and Minoru Kanehisa. "Systematic Analysis of Enzyme-Catalyzed Reaction Patterns and Prediction of Microbial Biodegradation Pathways". In: *Journal of Chemical Information and Modeling* 47.4 (July 1, 2007), pp. 1702–1712. DOI: 10.1021/ci700006f (cit. on pp. 8, 39, 41, 42).
- [67] M. J. L. de Groot, R. J. P. van Berlo, W. A. van Winden, et al. "Metabolite and reaction inference based on enzyme specificities". In: *Bioinformatics (Oxford, England)* 25.22 (Nov. 15, 2009), pp. 2975–2982. DOI: 10.1093/bioinformatics/btp507 (cit. on pp. 8, 76).
- [68] Aravind Sankar, Sayan Ranu, and Karthik Raman. "Predicting Novel Metabolic Pathways through Subgraph Mining". In: *Bioinformatics* (July 27, 2017). DOI: 10.1093/bioinformatics/btx481 (cit. on pp. 8, 12).

- [69] William Lingran Chen, David Z. Chen, and Keith T. Taylor. “Automatic reaction mapping and reaction center detection”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3.6 (Nov. 1, 2013), pp. 560–593. DOI: 10.1002/wcms.1140 (cit. on pp. 8, 42, 44).
- [70] Pablo Carbonell, Anne-Gaëlle Planson, Davide Fichera, and Jean-Loup Faulon. “A retrosynthetic biology approach to metabolic pathway design for therapeutic production”. In: *BMC systems biology* 5 (2011), p. 122. DOI: 10.1186/1752-0509-5-122 (cit. on pp. 8, 9, 12, 44, 45).
- [71] Pablo Carbonell, Pierre Parutto, Claire Baudier, Christophe Junot, and Jean-Loup Faulon. “Retropath: Automated Pipeline for Embedded Metabolic Circuits”. In: *ACS Synthetic Biology* (Oct. 16, 2013), p. 131016105628002. DOI: 10.1021/sb4001273 (cit. on pp. 8, 21, 34, 42, 46, 49, 54, 56, 93, 95, 98, 134).
- [72] Masaomi Nakamura, Tsuyoshi Hachiya, Yutaka Saito, Kengo Sato, and Yasubumi Sakakibara. “An efficient algorithm for de novo predictions of biochemical pathways between chemical compounds”. In: *BMC bioinformatics* 13 Suppl 17 (2012), S8. DOI: 10.1186/1471-2105-13-S17-S8 (cit. on pp. 8, 75).
- [73] Michihiro Araki, Robert Sidney Cox, Hiroki Makiguchi, et al. “M-path: A Compass for Navigating Potential Metabolic Pathways”. In: *Bioinformatics (Oxford, England)* (Nov. 13, 2014). DOI: 10.1093/bioinformatics/btu750 (cit. on pp. 8, 103).
- [74] David Rogers and Mathew Hahn. “Extended-connectivity fingerprints”. In: *Journal of chemical information and modeling* 50.5 (May 24, 2010), pp. 742–754. DOI: 10.1021/ci100050t (cit. on pp. 8, 97).
- [75] Masaaki Kotera, Yasuo Tabei, Yoshihiro Yamanishi, Toshiaki Tokimatsu, and Susumu Goto. “Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets”. In: *Bioinformatics* 29.13 (July 1, 2013), pp. i135–i144. DOI: 10.1093/bioinformatics/btt244 (cit. on pp. 9, 76).
- [76] International Union of Biochemistry {and} Molecular Biology Nomenclature Committee and Edwin Clifford Webb. *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Google-Books-ID: C3lqAAAAMAAJ. Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992. 888 pp. (cit. on pp. 9, 44).
- [77] Jean-Loup Faulon, Donald P. Visco, and Ramdas S. Pophale. “The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies”. In: *Journal of chemical information and computer sciences* 43.3 (2003), pp. 707–720 (cit. on pp. 9, 117).
- [78] M. Reza Anari and Thomas A. Baillie. “Bridging cheminformatic metabolite prediction and tandem mass spectrometry”. In: *Drug Discovery Today* 10.10 (May 2005), pp. 711–717. DOI: 10.1016/S1359-6446(05)03445-8 (cit. on p. 9).
- [79] James G. Jeffryes, Ricardo L. Colastani, Mona Elbadawi-Sidhu, et al. “MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics”. In: *Journal of Cheminformatics* 7.1 (Aug. 28, 2015), p. 44. DOI: 10.1186/s13321-015-0087-1 (cit. on pp. 9, 39–41, 44, 75).

- [80] Noushin Hadadi and Vassily Hatzimanikatis. "Design of computational retro-biosynthesis tools for the design of de novo synthetic pathways". In: *Current Opinion in Chemical Biology* 28 (Oct. 2015), pp. 99–104. DOI: 10.1016/j.cbpa.2015.06.025 (cit. on pp. 9, 39, 66).
- [81] Pablo Carbonell, Pierre Parutto, Joan Herisson, Shashi Bhushan Pandit, and Jean-Loup Faulon. "XTMS: pathway design in an eXTended metabolic space". In: *Nucleic Acids Research* 42 (Web Server issue July 2014), W389–394. DOI: 10.1093/nar/gku362 (cit. on pp. 10, 12, 21, 34, 39, 40, 44, 98, 108, 118, 134).
- [82] Di Wu, Qin Wang, Rajeev S. Assary, Linda J. Broadbelt, and Goran Krilov. "A computational approach to design and evaluate enzymatic reaction pathways: application to 1-butanol production from pyruvate". In: *Journal of Chemical Information and Modeling* 51.7 (July 25, 2011), pp. 1634–1647. DOI: 10.1021/ci2000659 (cit. on p. 10).
- [83] Meric Ataman and Vassily Hatzimanikatis. "Heading in the right direction: thermodynamics-based network analysis and pathway engineering". In: *Current Opinion in Biotechnology*. Pathway engineering 36 (Dec. 2015), pp. 176–182. DOI: 10.1016/j.copbio.2015.08.021 (cit. on p. 10).
- [84] Matthew Moura, Justin Finkle, Sarah Stainbrook, et al. "Evaluating enzymatic synthesis of small molecule drugs". In: *Metabolic Engineering* 33 (Jan. 2016), pp. 138–147. DOI: 10.1016/j.ymben.2015.11.006 (cit. on p. 10).
- [85] R. Pool. "Chemistry "Grand Master" Garners a Nobel Prize: E. J. Corey developed a logical methods for synthesizing molecules by working backward from the desired product". In: *Science* 250.4980 (Oct. 26, 1990), pp. 510–511. DOI: 10.1126/science.250.4980.510 (cit. on p. 11).
- [86] Elias James Corey. "The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules(Nobel Lecture)". In: *Angewandte Chemie International Edition in English* 30.5 (May 1991), pp. 455–465. DOI: 10.1002/anie.199104553 (cit. on p. 11).
- [87] Alexis D. C. Parenty, William G. Button, and Martin A. Ott. "An Expert System To Predict the Forced Degradation of Organic Molecules". In: *Molecular Pharmaceutics* 10.8 (Aug. 5, 2013), pp. 2962–2974. DOI: 10.1021/mp400083h (cit. on p. 11).
- [88] Pablo Carbonell, Andrew Currin, Adrian J. Jervis, et al. "Bioinformatics for the synthetic biology of natural products: integrating across the Design-Build-Test cycle". In: *Natural Product Reports* 33.8 (Aug. 27, 2016), pp. 925–932. DOI: 10.1039/c6np00018e (cit. on p. 11).
- [89] Anthony Cook, A. Peter Johnson, James Law, et al. "Computer-aided synthesis design: 40 years on". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 2.1 (Jan. 2012), pp. 79–107. DOI: 10.1002/wcms.61 (cit. on pp. 11, 77).
- [90] Ayoun Cho, Hongseok Yun, Jin Hwan Park, Sang Yup Lee, and Sunwon Park. "Prediction of novel synthetic pathways for the production of desired chemicals". In: *BMC Systems Biology* 4 (2010), p. 35. DOI: 10.1186/1752-0509-4-35 (cit. on pp. 12, 39, 42).

- [91] Weizhong Tu, Haoran Zhang, Juan Liu, and Qian-Nan Hu. “BioSynther: a customized biosynthetic potential explorer”. In: *Bioinformatics (Oxford, England)* 32.3 (Feb. 1, 2016), pp. 472–473. DOI: 10.1093/bioinformatics/btv599 (cit. on pp. 12, 103, 134).
- [92] Mengjin Liu, Bruno Bienfait, Oliver Sacher, et al. “Combining chemoinformatics with bioinformatics: in silico prediction of bacterial flavor-forming pathways by a chemical systems biology approach “reverse pathway engineering””. In: *PloS One* 9.1 (2014), e84769. DOI: 10.1371/journal.pone.0084769 (cit. on pp. 12, 39, 40, 42, 46, 64).
- [93] Guy Aidelberg, Benjamin D Towbin, Daphna Rothschild, et al. “Hierarchy of non-glucose sugars in Escherichia coli”. In: *BMC Systems Biology* 8.1 (Dec. 2014). DOI: 10.1186/s12918-014-0133-z (cit. on p. 12).
- [94] Ronald R. Breaker. “Prospects for Riboswitch Discovery and Analysis”. In: *Molecular Cell* 43.6 (Sept. 2011), pp. 867–879. DOI: 10.1016/j.molcel.2011.08.024 (cit. on p. 13).
- [95] Ronald R Breaker. “Riboswitches and the RNA world”. In: *Cold Spring Harbor perspectives in biology* 4.2 (Feb. 2012). DOI: 10.1101/cshperspect.a003566 (cit. on p. 13).
- [96] Brent Townshend, Andrew B. Kennedy, Joy S. Xiang, and Christina D. Smolke. “High-throughput cellular RNA device engineering”. In: *Nature Methods* (Aug. 10, 2015). DOI: 10.1038/nmeth.3486 (cit. on p. 13).
- [97] Florian Groher and Beatrix Suess. “Synthetic riboswitches—A tool comes of age”. In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* (2014) (cit. on p. 13).
- [98] Anna I. Podgornaia and Michael T. Laub. “Determinants of specificity in two-component signal transduction”. In: *Current Opinion in Microbiology* 16.2 (Apr. 2013), pp. 156–162. DOI: 10.1016/j.mib.2013.01.004 (cit. on p. 13).
- [99] Fuzhong Zhang and Jay Keasling. “Biosensors and their applications in microbial metabolic engineering”. In: *Trends in Microbiology* 19.7 (July 2011), pp. 323–329. DOI: 10.1016/j.tim.2011.05.003 (cit. on p. 14).
- [100] Md Eshrat E. Alahi and Subhas Chandra Mukhopadhyay. “Detection Methodologies for Pathogen and Toxins: A Review”. In: *Sensors* 17.8 (Aug. 16, 2017), p. 1885. DOI: 10.3390/s17081885 (cit. on p. 14).
- [101] Miguel A. D. Neves, Christophe Blaszykowski, Sumra Bokhari, and Michael Thompson. “Ultra-high frequency piezoelectric aptasensor for the label-free detection of cocaine”. In: *Biosensors & Bioelectronics* 72 (Oct. 15, 2015), pp. 383–392. DOI: 10.1016/j.bios.2015.05.038 (cit. on pp. 14, 100).
- [102] Sohila Zadran, Steve Standley, Kaylee Wong, et al. “Fluorescence resonance energy transfer (FRET)-based biosensors: visualizing cellular dynamics and bioenergetics”. In: *Applied Microbiology and Biotechnology* 96.4 (Nov. 2012), pp. 895–902. DOI: 10.1007/s00253-012-4449-6 (cit. on p. 14).

- [103] Chang Zhang, Zi-Han Wei, and Bang-Ce Ye. “Imaging and tracing of intracellular metabolites utilizing genetically encoded fluorescent biosensors”. In: *Biotechnology Journal* 8.11 (Nov. 2013), pp. 1280–1291. DOI: 10.1002/biot.201300001 (cit. on p. 14).
- [104] Bernhard Hochreiter, Alan Pardo Garcia, and Johannes A. Schmid. “Fluorescent Proteins as Genetically Encoded FRET Biosensors in Life Sciences”. In: *Sensors* 15.10 (Oct. 16, 2015), pp. 26281–26314. DOI: 10.3390/s151026281 (cit. on p. 14).
- [105] Yoko Nomura, Deepak Kumar, and Yohei Yokobayashi. “Synthetic mammalian riboswitches based on guanine aptazyme”. In: *Chemical communications (Cambridge, England)* 48.57 (July 21, 2012), pp. 7215–7217. DOI: 10.1039/c2cc33140c (cit. on p. 14).
- [106] Qingyuan Gui, Tom Lawson, Suyan Shan, Lu Yan, and Yong Liu. “The Application of Whole Cell-Based Biosensors for Use in Environmental Analysis and in Medical Diagnostics”. In: *Sensors* 17.7 (July 13, 2017), p. 1623. DOI: 10.3390/s17071623 (cit. on pp. 15, 16).
- [107] Aditya Vikram Pandit, Shyam Srinivasan, and Radhakrishnan Mahadevan. “Re-designing metabolism based on orthogonality principles”. In: *Nature Communications* 8 (May 30, 2017), ncomms15188. DOI: 10.1038/ncomms15188 (cit. on pp. 15, 17).
- [108] Jie Zhang, Michael K. Jensen, and Jay D. Keasling. “Development of biosensors and their application in metabolic engineering”. In: *Current Opinion in Chemical Biology* 28 (June 4, 2015), pp. 1–8. DOI: 10.1016/j.cbpa.2015.05.013 (cit. on pp. 15, 16).
- [109] Regina Mahr and Julia Frunzke. “Transcription factor-based biosensors in biotechnology: current state and future prospects”. In: *Applied Microbiology and Biotechnology* (Oct. 31, 2015). DOI: 10.1007/s00253-015-7090-3 (cit. on pp. 15, 82, 94).
- [110] Di Liu, Trent Evans, and Fuzhong Zhang. “Applications and advances of metabolite biosensors for metabolic engineering”. In: *Metabolic Engineering* (July 1, 2015). DOI: 10.1016/j.ymben.2015.06.008 (cit. on pp. 15, 82, 94, 136).
- [111] Byung Eun Min, Hyun Gyu Hwang, Hyun Gyu Lim, and Gyoo Yeol Jung. “Optimization of industrial microorganisms: recent advances in synthetic dynamic regulators”. In: *Journal of Industrial Microbiology & Biotechnology* (Nov. 10, 2016). DOI: 10.1007/s10295-016-1867-y (cit. on p. 15).
- [112] Raul Fernandez-López, Raul Ruiz, Fernando de la Cruz, and Gabriel Moncalián. “Transcription factor-based biosensors enlightened by the analyte”. In: *Frontiers in Microbiology* 6 (2015), p. 648. DOI: 10.3389/fmicb.2015.00648 (cit. on pp. 15, 82).
- [113] Sylvia Daunert, Gary Barrett, Jessika S. Feliciano, et al. “Genetically Engineered Whole-Cell Sensing Systems: Coupling Biological Recognition with Reporter Genes”. In: *Chemical Reviews* 100.7 (July 1, 2000), pp. 2705–2738. DOI: 10.1021/cr990115p (cit. on p. 16).

- [114] Jan Roelof van der Meer and Shimshon Belkin. “Where microbiology meets microengineering: design and applications of reporter bacteria”. In: *Nature Reviews. Microbiology* 8.7 (July 2010), pp. 511–522. DOI: 10.1038/nrmicro2392 (cit. on pp. 16, 71).
- [115] Miso Park, Shen-Long Tsai, and Wilfred Chen. “Microbial biosensors: engineered microorganisms as the sensing machinery”. In: *Sensors (Basel, Switzerland)* 13.5 (2013), pp. 5777–5795. DOI: 10.3390/s130505777 (cit. on p. 16).
- [116] Angela Ivask, Taisia Rõlova, and Anne Kahru. “A suite of recombinant luminescent bacterial strains for the quantification of bioavailable heavy metals and toxicity testing”. In: *BMC Biotechnology* 9.1 (2009), p. 41. DOI: 10.1186/1472-6750-9-41 (cit. on p. 16).
- [117] S. Tauriainen, M. Karp, W. Chang, and M. Virta. “Recombinant luminescent bacteria for measuring bioavailable arsenite and antimonite”. In: *Applied and Environmental Microbiology* 63.11 (Nov. 1997), pp. 4456–4461 (cit. on p. 16).
- [118] Fuzhong Zhang, James M. Carothers, and Jay D. Keasling. “Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids”. In: *Nature Biotechnology* 30.4 (Apr. 2012), pp. 354–359. DOI: 10.1038/nbt.2149 (cit. on p. 16).
- [119] Tamás Fehér, Vincent Libis, Pablo Carbonell, and Jean-Loup Faulon. “A Sense of Balance: Experimental Investigation and Modeling of a Malonyl-CoA Sensor in *Escherichia coli*”. In: *Frontiers in Bioengineering and Biotechnology* 3 (2015), p. 46. DOI: 10.3389/fbioe.2015.00046 (cit. on p. 16).
- [120] Peng Xu, Lingyun Li, Fuming Zhang, Gregory Stephanopoulos, and Mattheos Koffas. “Improving fatty acids production by engineering dynamic pathway regulation and metabolic control”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.31 (Aug. 5, 2014), pp. 11299–11304. DOI: 10.1073/pnas.1406401111 (cit. on pp. 16, 82, 94).
- [121] Di Liu, Yi Xiao, Bradley S. Evans, and Fuzhong Zhang. “Negative feedback regulation of fatty acid production based on a malonyl-CoA sensor-actuator”. In: *ACS synthetic biology* 4.2 (Feb. 20, 2015), pp. 132–140. DOI: 10.1021/sb400158w (cit. on p. 16).
- [122] Robert H. Dahl, Fuzhong Zhang, Jorge Alonso-Gutierrez, et al. “Engineering dynamic pathway regulation using stress-response promoters”. In: *Nature Biotechnology* 31.11 (Nov. 2013), pp. 1039–1046. DOI: 10.1038/nbt.2689 (cit. on p. 16).
- [123] Kevin V. Solomon, Tarielle M. Sanders, and Kristala L.J. Prather. “A dynamic metabolite valve for the control of central carbon metabolism”. In: *Metabolic Engineering* 14.6 (Nov. 2012), pp. 661–671. DOI: 10.1016/j.ymben.2012.08.006 (cit. on p. 17).
- [124] Sue Zanne Tan, Shawn Manchester, and Kristala L. J. Prather. “Controlling Central Carbon Metabolism for Improved Pathway Yields in *Saccharomyces cerevisiae*”. In: *ACS synthetic biology* 5.2 (Feb. 19, 2016), pp. 116–124. DOI: 10.1021/acssynbio.5b00164 (cit. on p. 17).

- [125] Apoorv Gupta, Irene M Brockman Reizman, Christopher R Reisch, and Kristala L J Prather. “Dynamic regulation of metabolic flux in engineered bacteria using a pathway-independent quorum-sensing circuit”. In: *Nature Biotechnology* 35.3 (Feb. 13, 2017), pp. 273–279. DOI: 10.1038/nbt.3796 (cit. on p. 17).
- [126] Jameson K Rogers, Noah D Taylor, and George M Church. “Biosensor-based engineering of biosynthetic pathways”. In: *Current Opinion in Biotechnology. Pharmaceutical biotechnology • Chemical biotechnology* 42 (Dec. 2016), pp. 84–91. DOI: 10.1016/j.copbio.2016.03.005 (cit. on p. 17).
- [127] Daniel Klein-Marcuschamer, Parayil Kumaran Ajikumar, and Gregory Stephanopoulos. “Engineering microbial cell factories for biosynthesis of isoprenoid molecules: beyond lycopene”. In: *Trends in Biotechnology* 25.9 (Sept. 2007), pp. 417–424. DOI: 10.1016/j.tibtech.2007.07.006 (cit. on p. 17).
- [128] Jeffrey A. Dietrich, David L. Shis, Azadeh Alikhani, and Jay D. Keasling. “Transcription factor-based screens and synthetic selections for microbial small-molecule biosynthesis”. In: *ACS synthetic biology* 2.1 (Jan. 18, 2013), pp. 47–58. DOI: 10.1021/sb300091d (cit. on pp. 17, 82).
- [129] Stephan Binder, Georg Schendzielorz, Norma Stäbler, et al. “A high-throughput approach to identify genomic variants of bacterial metabolite producers at the single-cell level”. In: *Genome Biology* 13.5 (2012), R40. DOI: 10.1186/gb-2012-13-5-r40 (cit. on p. 17).
- [130] William W. Mohn, Junkal Garmendia, Teca C. Galvao, and Víctor de Lorenzo. “Surveying biotransformations with à la carte genetic traps: translating dehydrochlorination of lindane (gamma-hexachlorocyclohexane) into lacZ-based phenotypes”. In: *Environmental Microbiology* 8.3 (Mar. 2006), pp. 546–555. DOI: 10.1111/j.1462-2920.2006.00983.x (cit. on pp. 17, 87, 89, 95, 136).
- [131] Michael Bott. “Need for speed – finding productive mutations using transcription factor-based biosensors, fluorescenceactivated cell sorting and recombineering”. In: *Microbial Biotechnology* 8.1 (Jan. 2015), pp. 8–10. DOI: 10.1111/1751-7915.12248 (cit. on p. 17).
- [132] Guangyu Yang and Stephen G. Withers. “Ultrahigh-throughput FACS-based screening for directed enzyme evolution”. In: *Chembiochem: A European Journal of Chemical Biology* 10.17 (Nov. 23, 2009), pp. 2704–2715. DOI: 10.1002/cbic.200900384 (cit. on p. 17).
- [133] Marcus Schallmey, Julia Frunzke, Lothar Eggeling, and Jan Marienhagen. “Looking for the pick of the bunch: high-throughput screening of producing microorganisms with biosensors”. In: *Current Opinion in Biotechnology* 26 (Apr. 2014), pp. 148–154. DOI: 10.1016/j.copbio.2014.01.005 (cit. on p. 17).
- [134] David S. Wishart, Timothy Jewison, An Chi Guo, et al. “HMDB 3.0–The Human Metabolome Database in 2013”. In: *Nucleic Acids Research* 41 (Database issue Jan. 2013), pp. D801–807. DOI: 10.1093/nar/gks1065 (cit. on pp. 22, 108, 117).
- [135] Jan Schellenberger, Junyoung O. Park, Tom M. Conrad, and Bernhard Ø Pals-son. “BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions”. In: *BMC Bioinformatics* 11 (2010), p. 213. DOI: 10.1186/1471-2105-11-213 (cit. on pp. 22, 23, 30).

- [136] B. P. Tu, R. E. Mohler, J. C. Liu, et al. “Cyclic changes in metabolic state during the life of a yeast cell”. In: *Proceedings of the National Academy of Sciences* 104.43 (Oct. 23, 2007), pp. 16886–16891. DOI: 10.1073/pnas.0708365104 (cit. on p. 22).
- [137] Pablo Carbonell, Lars Carlsson, and Jean-Loup Faulon. “Stereo Signature Molecular Descriptor”. In: *Journal of Chemical Information and Modeling* 53.4 (Apr. 22, 2013), pp. 887–897. DOI: 10.1021/ci300584r (cit. on pp. 22, 23, 27, 41, 42, 76, 97, 133).
- [138] Pablo Carbonell and Jean-Loup Faulon. “Molecular signatures-based prediction of enzyme promiscuity”. In: *Bioinformatics* 26.16 (Aug. 15, 2010), pp. 2012–2019. DOI: 10.1093/bioinformatics/btq317 (cit. on p. 22).
- [139] Michael J Dougherty and Frances H Arnold. “Directed evolution: new parts and optimized function”. In: *Current Opinion in Biotechnology* 20.4 (Aug. 2009), pp. 486–491. DOI: 10.1016/j.copbio.2009.08.005 (cit. on p. 22).
- [140] Juhan Kim, Jamie P Kershner, Yehor Novikov, Richard K Shoemaker, and Shelley D Copley. “Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5'-phosphate synthesis”. In: *Molecular Systems Biology* 6 (Nov. 30, 2010). DOI: 10.1038/msb.2010.88 (cit. on p. 22).
- [141] Sébastien Moretti, Olivier Martin, T. Van Du Tran, et al. “MetaNetX/MNXref—reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D523–526. DOI: 10.1093/nar/gkv1117 (cit. on pp. 23, 45, 49).
- [142] V. Chelliah, N. Juty, I. Ajmera, et al. “BioModels: ten-year anniversary”. In: *Nucleic Acids Research* 43 (D1 Jan. 28, 2015), pp. D542–D548. DOI: 10.1093/nar/gku1181 (cit. on p. 23).
- [143] Marco Terzer and Jörg Stelling. “Large-scale computation of elementary flux modes with bit pattern trees”. In: *Bioinformatics* 24.19 (Oct. 1, 2008), pp. 2229–2235. DOI: 10.1093/bioinformatics/btn401 (cit. on pp. 23, 25, 32).
- [144] P. Shannon. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks”. In: *Genome Research* 13.11 (Nov. 1, 2003), pp. 2498–2504. DOI: 10.1101/gr.1239303 (cit. on p. 25).
- [145] Michael R. Berthold, Nicolas Cebon, Fabian Dill, et al. “KNIME: The Konstanz Information Miner”. In: *Data Analysis, Machine Learning and Applications*. DOI: 10.1007/978-3-540-78246-9_38. Springer, Berlin, Heidelberg, 2008, pp. 319–326 (cit. on pp. 25, 40, 49).
- [146] Ali Ebrahim, Joshua A Lerman, Bernhard O Palsson, and Daniel R Hyduke. “COBRApy: CONstraints-Based Reconstruction and Analysis for Python”. In: *BMC Systems Biology* 7.1 (2013), p. 74. DOI: 10.1186/1752-0509-7-74 (cit. on p. 25).
- [147] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe. “KEGG for integration and interpretation of large-scale molecular data sets”. In: *Nucleic Acids Research* 40 (D1 Jan. 1, 2012), pp. D109–D114. DOI: 10.1093/nar/gkr988 (cit. on p. 30).

- [148] R. Caspi, T. Altman, K. Dreher, et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases”. In: *Nucleic Acids Research* 40 (D1 Jan. 1, 2012), pp. D742–D753. DOI: 10 . 1093/nar/gkr1014 (cit. on p. 30).
- [149] M. Hucka, A. Finney, H. M. Sauro, et al. “The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models”. In: *Bioinformatics* 19.4 (Mar. 1, 2003), pp. 524–531. DOI: 10 . 1093/bioinformatics/btg015 (cit. on p. 30).
- [150] Pablo Carbonell, Davide Fichera, Shashi B. Pandit, and Jean-Loup Faulon. “Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms”. In: *BMC systems biology* 6.1 (2012), p. 10 (cit. on pp. 31, 39).
- [151] Jürgen Zanghellini, David E. Ruckerbauer, Michael Hanscho, and Christian Jungreuthmayer. “Elementary flux modes in a nutshell: Properties, calculation and applications”. In: *Biotechnology Journal* 8.9 (Sept. 2013), pp. 1009–1016. DOI: 10 . 1002/biot.201200269 (cit. on p. 31).
- [152] Pablo Carbonell, Anne-Gaëlle Planson, and Jean-Loup Faulon. “Retrosynthetic design of heterologous pathways”. In: *Methods in Molecular Biology (Clifton, N.J.)* 985 (2013), pp. 149–173. DOI: 10 . 1007/978-1-62703-299-5_9 (cit. on pp. 34, 98).
- [153] Baudoin Delépine, Vincent Libis, Pablo Carbonell, and Jean-Loup Faulon. “SensiPath: computer-aided design of sensing-enabling metabolic pathways”. In: *Nucleic Acids Research* 44 (W1 July 8, 2016), W226–231. DOI: 10 . 1093/nar/gkw305 (cit. on pp. 34, 40, 65, 72, 74, 89, 93).
- [154] Noel M O’Boyle, Michael Banck, Craig A James, et al. “Open Babel: An open chemical toolbox”. In: *Journal of cheminformatics* 3 (2011), p. 33. DOI: 10 . 1186/1758-2946-3-33 (cit. on p. 34).
- [155] Baudoin Delépine, Thomas Duigou, Pablo Carbonell, and Jean-Loup Faulon. “RetroPath2.0: A retrosynthesis workflow for metabolic engineers”. In: *Metabolic Engineering* 45 (Jan. 2018), pp. 158–170. DOI: 10 . 1016/j.ymben.2017.12.002 (cit. on p. 37).
- [156] Jay D. Keasling. “Hearing on Policies to Spur Innovative Medical Breakthroughs from Laboratories to Patients”. U.S. House of Representatives, July 17, 2014 (cit. on p. 39).
- [157] Wilbert B. Copeland, Bryan A. Bartley, Deepak Chandran, et al. “Computational Tools for Metabolic Engineering”. In: *Metabolic Engineering* 14.3 (May 2012), pp. 270–280 (cit. on p. 39).
- [158] Guillermo Rodrigo, Javier Carrera, Kristala Jones Prather, and Alfonso Jaramillo. “DESHARKY: automatic design of metabolic pathways for optimal cell growth”. In: *Bioinformatics (Oxford, England)* 24.21 (Nov. 1, 2008), pp. 2554–2556. DOI: 10 . 1093/bioinformatics/btn471 (cit. on p. 39).
- [159] Yuki Moriya, Daichi Shigemizu, Masahiro Hattori, et al. “PathPred: an enzyme-catalyzed metabolic pathway prediction server”. In: *Nucleic Acids Research* 38 (Web Server issue July 2010), W138–143. DOI: 10 . 1093/nar/gkq318 (cit. on pp. 39, 40, 46, 56).

- [160] Carol A. Marchant, Katharine A. Briggs, and Anthony Long. “In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic”. In: *Toxicology Mechanisms and Methods* 18.2 (2008), pp. 177–187. DOI: 10.1080/15376510701857320 (cit. on pp. 39, 40, 46).
- [161] Christopher S. Henry, Linda J. Broadbelt, and Vassily Hatzimanikatis. “Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate”. In: *Biotechnology and Bioengineering* 106.3 (June 15, 2010), pp. 462–473. DOI: 10.1002/bit.22673 (cit. on pp. 39–42, 44, 47, 49, 55, 71).
- [162] Noushin Hadadi, Jasmin Hafner, Adrian Shajkofci, Katerina Zisaki, and Vassily Hatzimanikatis. “ATLAS of Biochemistry A repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies”. In: *ACS synthetic biology* (July 12, 2016). DOI: 10.1021/acssynbio.6b00054 (cit. on pp. 39–41, 44, 46, 75).
- [163] Anne-Gaëlle Planson, Pablo Carbonell, Elodie Paillard, Nicolas Pollet, and Jean-Loup Faulon. “Compound toxicity screening and structure-activity relationship modeling in *Escherichia coli*”. In: *Biotechnology and Bioengineering* 109.3 (Mar. 2012), pp. 846–850. DOI: 10.1002/bit.24356 (cit. on p. 39).
- [164] Bo Kyeng Hou, Lynda B. M. Ellis, and Lawrence P. Wackett. “Encoding microbial metabolic logic: predicting biodegradation”. In: *Journal of Industrial Microbiology and Biotechnology* 31.6 (July 1, 2004), pp. 261–272. DOI: 10.1007/s10295-004-0144-7 (cit. on p. 39).
- [165] Stacey D. Finley, Linda J. Broadbelt, and Vassily Hatzimanikatis. “Computational framework for predictive biodegradation”. In: *Biotechnology and Bioengineering* 104.6 (Dec. 15, 2009), pp. 1086–1097. DOI: 10.1002/bit.22489 (cit. on p. 39).
- [166] Masanori Arita. “In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism”. In: *Genome Research* 13.11 (Nov. 2003), pp. 2455–2466. DOI: 10.1101/gr.1212003 (cit. on p. 40).
- [167] Noushin Hadadi, Jasmin Hafner, Keng Cher Soh, and Vassily Hatzimanikatis. “Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites”. In: *Biotechnology Journal* (Nov. 1, 2016). DOI: 10.1002/biot.201600464 (cit. on p. 40).
- [168] Paul N. Schofield, Tania Bubela, Thomas Weaver, et al. “Post-publication sharing of data and tools”. In: *Nature* 461.7261 (Sept. 10, 2009), pp. 171–173. DOI: 10.1038/461171a (cit. on p. 40).
- [169] Marcia McNutt, Kerstin Lehnert, Brooks Hanson, et al. “RESEARCH INTEGRITY. Liberating field science samples and data”. In: *Science (New York, N.Y.)* 351.6277 (Mar. 4, 2016), pp. 1024–1026. DOI: 10.1126/science.aad7048 (cit. on p. 40).
- [170] Kenneth Haug, Reza M. Salek, and Christoph Steinbeck. “Global open data management in metabolomics”. In: *Current Opinion in Chemical Biology* 36 (Feb. 2017), pp. 58–63. DOI: 10.1016/j.cbpa.2016.12.024 (cit. on p. 40).
- [171] Wendy A. Warr. “Scientific workflow systems: Pipeline Pilot and KNIME”. In: *Journal of Computer-Aided Molecular Design* 26.7 (July 2012), pp. 801–804. DOI: 10.1007/s10822-012-9577-7 (cit. on p. 40).

- [172] Matthew A. Kayala, Chloé-Agathe Azencott, Jonathan H. Chen, and Pierre Baldi. “Learning to Predict Chemical Reactions”. In: *Journal of Chemical Information and Modeling* 51.9 (Sept. 26, 2011), pp. 2209–2222. DOI: 10.1021/ci200207y (cit. on pp. 41, 76).
- [173] Diogo A. R. S. Latino and João Aires-de-Sousa. “Classification of chemical reactions and chemoinformatic processing of enzymatic transformations”. In: *Methods in Molecular Biology (Clifton, N.J.)* 672 (2011), pp. 325–340. DOI: 10.1007/978-1-60761-839-3_13 (cit. on p. 41).
- [174] James Dugundji and Ivar Ugi. “An algebraic model of constitutional chemistry as a basis for chemical computer programs”. In: *Computers in Chemistry*. DOI: 10.1007/BFb0051317. Springer, Berlin, Heidelberg, 1973, pp. 19–64 (cit. on p. 41).
- [175] Vassily Hatzimanikatis, Chunhui Li, Justin A. Ionita, et al. “Exploring the diversity of complex metabolic networks”. In: *Bioinformatics (Oxford, England)* 21.8 (Apr. 15, 2005), pp. 1603–1609. DOI: 10.1093/bioinformatics/bti213 (cit. on p. 42).
- [176] Syed Asad Rahman, Sergio Martinez Cuesta, Nicholas Furnham, Gemma L. Holliday, and Janet M. Thornton. “EC-BLAST: a tool to automatically search and compare enzyme reactions”. In: *Nature Methods* 11.2 (Feb. 2014), pp. 171–174. DOI: 10.1038/nmeth.2803 (cit. on pp. 44, 56, 57).
- [177] Jean-Loup Faulon, Milind Misra, Shawn Martin, Ken Sale, and Rajat Sapra. “Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor”. In: *Bioinformatics* 24.2 (Jan. 15, 2008), pp. 225–233. DOI: 10.1093/bioinformatics/btm580 (cit. on pp. 45, 47, 98, 110, 117).
- [178] Joseph Mellor, Ioana Grigoras, Pablo Carbonell, and Jean-Loup Faulon. “Semisupervised Gaussian Process for Automated Enzyme Search”. In: *ACS synthetic biology* 5.6 (2016), pp. 518–528. DOI: 10.1021/acssynbio.5b00294 (cit. on p. 47).
- [179] Syed Asad Rahman, Gilliean Torrance, Lorenzo Baldacci, et al. “Reaction Decoder Tool (RDT): extracting features from chemical reactions”. In: *Bioinformatics (Oxford, England)* 32.13 (July 1, 2016), pp. 2065–2066. DOI: 10.1093/bioinformatics/btw096 (cit. on p. 49).
- [180] Jeffrey D. Orth, Tom M. Conrad, Jessica Na, et al. “A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011”. In: *Molecular Systems Biology* 7 (Oct. 11, 2011), p. 535. DOI: 10.1038/msb.2011.65 (cit. on pp. 50, 55, 60).
- [181] Weizhong Li and Adam Godzik. “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences”. In: *Bioinformatics (Oxford, England)* 22.13 (July 1, 2006), pp. 1658–1659. DOI: 10.1093/bioinformatics/bt1158 (cit. on p. 52).
- [182] S. Schuster, D. A. Fell, and T. Dandekar. “A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks”. In: *Nature Biotechnology* 18.3 (Mar. 2000), pp. 326–332. DOI: 10.1038/73786 (cit. on p. 54).

- [183] Olga Khersonsky {and} Dan S. Tawfik. "Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective". In: *Annual Review of Biochemistry* 79.1 (June 7, 2010), pp. 471–505. DOI: 10.1146/annurev-biochem-030409-143718 (cit. on p. 57).
- [184] Jeffrey D. Orth and Bernhard Pálsson. "Gap-filling analysis of the iJO1366 Escherichia coli metabolic network reconstruction for discovery of metabolic functions". In: *BMC systems biology* 6 (May 1, 2012), p. 30. DOI: 10.1186/1752-0509-6-30 (cit. on p. 57).
- [185] Ingrid M. Keseler, Amanda Mackie, Martin Peralta-Gil, et al. "EcoCyc: fusing model organism databases with systems biology". In: *Nucleic Acids Research* 41 (Database issue Jan. 2013), pp. D605–612. DOI: 10.1093/nar/gks1027 (cit. on p. 59).
- [186] Kechao Yang and William W. Metcalf. "A new activity for an old enzyme: Escherichia coli bacterial alkaline phosphatase is a phosphite-dependent hydrogenase". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.21 (May 25, 2004), pp. 7919–7924. DOI: 10.1073/pnas.0400664101 (cit. on p. 59).
- [187] James D. Winkler, Andrea L. Halweg-Edwards, and Ryan T. Gill. "The LASER database: Formalizing design rules for metabolic engineering". In: *Metabolic Engineering Communications* 2 (Dec. 2015), pp. 30–38. DOI: 10.1016/j.meten.2015.06.003 (cit. on p. 59).
- [188] James D. Winkler, Andrea L. Halweg-Edwards, and Ryan T. Gill. "Quantifying complexity in metabolic engineering using the LASER database". In: *Metabolic Engineering Communications* 3 (Dec. 2016), pp. 227–233. DOI: 10.1016/j.meten.2016.07.002 (cit. on p. 59).
- [189] Furkan H. Isikgor and C. Remzi Becer. "Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers". In: *Polymer Chemistry* 6.25 (June 16, 2015), pp. 4497–4559. DOI: 10.1039/C5PY00263J (cit. on pp. 60, 61).
- [190] Rebekah McKenna and David R. Nielsen. "Styrene biosynthesis from glucose by engineered E. coli". In: *Metabolic Engineering* 13.5 (Sept. 2011), pp. 544–554. DOI: 10.1016/j.ymben.2011.06.005 (cit. on pp. 60, 62).
- [191] Rebekah McKenna, Luis Moya, Matthew McDaniel, and David R. Nielsen. "Comparing in situ removal strategies for improving styrene bioproduction". In: *Bio-process and Biosystems Engineering* 38.1 (Jan. 2015), pp. 165–174. DOI: 10.1007/s00449-014-1255-9 (cit. on p. 60).
- [192] Rebekah McKenna, Brian Thompson, Shawn Pugh, and David R. Nielsen. "Rational and combinatorial approaches to engineering styrene production by *Saccharomyces cerevisiae*". In: *Microbial Cell Factories* 13 (Aug. 21, 2014), p. 123. DOI: 10.1186/s12934-014-0123-2 (cit. on p. 60).
- [193] Richard J. Sheehan. "Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid". In: *Ullmann's Encyclopedia of Industrial Chemistry*. DOI: 10.1002/14356007.a26_193.pub2. Wiley-VCH Verlag GmbH & Co. KGaA, 2000 (cit. on p. 61).

- [194] “Microbial production of terephthalic acid and isophthalic acid”. US6187569 B1. Michael G. Bramucci, Carol M. McCutchen, Vasantha Nagarajan, and Stuart M. Thomas. U.S. Classification 435/136, 435/41, 435/146, 435/142, 435/170, 435/156, 435/155; International Classification C12P7/46; Cooperative Classification C12P7/46; European Classification C12P7/46. Feb. 13, 2001 (cit. on pp. 61, 63).
- [195] Jing Wang, Jianhe Xu, and Jing Tian. “A Method for Producing Terephthalic Acid by Comamonas testosteroni DSM6577”. In: *Chinese Journal of Catalysis* 27.4 (2006), p. 297 (cit. on p. 61).
- [196] Vincent Libis, Baudoin Delépine, and Jean-Loup Faulon. “Expanding Biosensing Abilities through Computer-Aided Design of Metabolic Pathways”. In: *ACS synthetic biology* 5.10 (2016), pp. 1076–1085. DOI: 10.1021/acssynbio.5b00225 (cit. on pp. 65, 72, 87, 89, 95, 100, 105).
- [197] L. T. Stauffer and G. V. Stauffer. “Characterization of the gcv control region from Escherichia coli”. In: *Journal of Bacteriology* 176.20 (Oct. 1994), pp. 6159–6164 (cit. on pp. 65, 72).
- [198] Arun Sreekumar, Laila M. Poisson, Thekkelnaycke M. Rajendiran, et al. “Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression”. In: *Nature* 457.7231 (Feb. 12, 2009), pp. 910–914. DOI: 10.1038/nature07762 (cit. on pp. 65, 72).
- [199] Jacob R. Rubens, Gianluca Selvaggio, and Timothy K. Lu. “Synthetic mixed-signal computation in living cells”. In: *Nature Communications* 7 (June 3, 2016), p. 11658. DOI: 10.1038/ncomms11658 (cit. on p. 65).
- [200] Ana Rita Lima, Maria de Lourdes Bastos, Márcia Carvalho, and Paula Guedes de Pinho. “Biomarker Discovery in Human Prostate Cancer: an Update in Metabolomics Studies”. In: *Translational Oncology* 9.4 (Aug. 2016), pp. 357–370. DOI: 10.1016/j.tranon.2016.05.004 (cit. on pp. 65, 72).
- [201] S. E. D’Orazio, V. Thomas, and C. M. Collins. “Activation of transcription at divergent urea-dependent promoters by the urease gene regulator UreR”. In: *Molecular Microbiology* 21.3 (Aug. 1996), pp. 643–655 (cit. on p. 65).
- [202] Kate Thodey, Stephanie Galanie, and Christina D. Smolke. “A microbial biomanufacturing platform for natural and semisynthetic opioids”. In: *Nature Chemical Biology* 10.10 (Oct. 2014), pp. 837–844. DOI: 10.1038/nchembio.1613 (cit. on p. 67).
- [203] Ahmad S. Khalil and James J. Collins. “Synthetic biology: applications come of age”. In: *Nature Reviews. Genetics* 11.5 (May 2010), pp. 367–379. DOI: 10.1038/nrg2775 (cit. on p. 71).
- [204] Xiaoling Zang, Christina M. Jones, Tran Q. Long, et al. “Feasibility of detecting prostate cancer by ultraperformance liquid chromatography-mass spectrometry serum metabolomics”. In: *Journal of Proteome Research* 13.7 (July 3, 2014), pp. 3444–3454. DOI: 10.1021/pr500409q (cit. on p. 72).
- [205] Yuanfeng Li, Shi Qiu, and Ai Hua Zhang. “High-throughput metabolomics to identify metabolites to serve as diagnostic biomarkers of prostate cancer”. In: *Analytical Methods* 8.16 (Apr. 21, 2016), pp. 3284–3290. DOI: 10.1039/C6AY00127K (cit. on p. 72).

- [206] Tong Zhang, David G. Watson, Lijie Wang, et al. “Application of Holistic Liquid Chromatography-High Resolution Mass Spectrometry Based Urinary Metabolomics for Prostate Cancer Detection and Biomarker Discovery”. In: *PLoS One* 8.6 (2013), e65880. DOI: 10.1371/journal.pone.0065880 (cit. on p. 72).
- [207] Wiktoria Struck-Lewicka, Marta Kordalewska, Renata Bujak, et al. “Urine metabolic fingerprinting using LC-MS and GC-MS reveals metabolite changes in prostate cancer: A pilot study”. In: *Journal of Pharmaceutical and Biomedical Analysis* 111 (2015), pp. 351–361. DOI: 10.1016/j.jpba.2014.12.026 (cit. on p. 72).
- [208] M. A. Fernández-Peralbo, E. Gómez-Gómez, M. Calderón-Santiago, et al. “Prostate Cancer Patients-Negative Biopsy Controls Discrimination by Untargeted Metabolomics Analysis of Urine by LC-QTOF: Upstream Information on Other Omics”. In: *Scientific Reports* 6 (Dec. 2, 2016), p. 38243. DOI: 10.1038/srep38243 (cit. on p. 72).
- [209] Jonathan E. McDunn, Zhen Li, Klaus-Peter Adam, et al. “Metabolomic signatures of aggressive prostate cancer”. In: *The Prostate* 73.14 (Oct. 2013), pp. 1547–1560. DOI: 10.1002/pros.22704 (cit. on p. 72).
- [210] Tao Huan, Dean A. Troyer, and Liang Li. “Metabolite Analysis and Histology on the Exact Same Tissue: Comprehensive Metabolomic Profiling and Metabolic Classification of Prostate Cancer”. In: *Scientific Reports* 6 (Aug. 31, 2016), p. 32272. DOI: 10.1038/srep32272 (cit. on p. 72).
- [211] L. A. Tartaglia, G. Storz, and B. N. Ames. “Identification and molecular analysis of oxyR-regulated promoters important for the bacterial adaptation to oxidative stress”. In: *Journal of Molecular Biology* 210.4 (Dec. 20, 1989), pp. 709–719 (cit. on p. 72).
- [212] Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green, and Klavs F. Jensen. “Prediction of Organic Reaction Outcomes Using Machine Learning”. In: *ACS Central Science* 3.5 (May 24, 2017), pp. 434–443. DOI: 10.1021/acscentsci.7b00064 (cit. on p. 76).
- [213] Chihai Yang, Aleksey Tarkhov, Jörg Marusczyk, et al. “New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling”. In: *Journal of Chemical Information and Modeling* 55.3 (Mar. 23, 2015), pp. 510–528. DOI: 10.1021/ci500667v (cit. on p. 76).
- [214] Jennifer N. Wei, David Duvenaud, and Alán Aspuru-Guzik. “Neural Networks for the Prediction of Organic Chemistry Reactions”. In: *ACS Central Science* 2.10 (Oct. 26, 2016), pp. 725–732. DOI: 10.1021/acscentsci.6b00219 (cit. on p. 76).
- [215] Marwin H. S. Segler and Mark P. Waller. “Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction”. In: *Chemistry - A European Journal* 23.25 (May 2, 2017), pp. 5966–5971. DOI: 10.1002/chem.201605499 (cit. on p. 76).
- [216] Asher Mullard. “The drug-maker’s guide to the galaxy”. In: *Nature* 549.7673 (Sept. 26, 2017), pp. 445–447. DOI: 10.1038/549445a (cit. on p. 77).

- [217] Tamas Fehér, Anne-Gaëlle Planson, Pablo Carbonell, et al. “Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering”. In: *Biotechnology Journal* 9.11 (Nov. 2014), pp. 1446–1457. DOI: 10.1002/biot.201400055 (cit. on pp. 77, 110).
- [218] Andrew Stine, Miaomin Zhang, Soo Ro, et al. “Exploring De Novo metabolic pathways from pyruvate to propionic acid”. In: *Biotechnology Progress* 32.2 (Mar. 1, 2016), pp. 303–311. DOI: 10.1002/btpr.2233 (cit. on p. 77).
- [219] John Moult, Krzysztof Fidelis, Andriy Kryshchuk, Torsten Schwede, and Anna Tramontano. “Critical assessment of methods of protein structure prediction (CASP) - round x: Critical Assessment of Structure Prediction”. In: *Proteins: Structure, Function, and Bioinformatics* 82 (Feb. 2014), pp. 1–6. DOI: 10.1002/prot.24452 (cit. on p. 77).
- [220] Shoshana J. Wodak and Joel Janin. “Modeling protein assemblies: Critical Assessment of Predicted Interactions (CAPRI) 15 years hence.: 6TH CAPRI evaluation meeting April 17-19 Tel-Aviv, Israel”. In: *Proteins: Structure, Function, and Bioinformatics* 85.3 (Mar. 2017), pp. 357–358. DOI: 10.1002/prot.25233 (cit. on p. 77).
- [221] Iddo Friedberg and Predrag Radivojac. “Community-Wide Evaluation of Computational Function Prediction”. In: *The Gene Ontology Handbook*. Ed. by Christophe Dessimoz and Nives Škunca. Vol. 1446. DOI: 10.1007/978-1-4939-3743-1_10. New York, NY: Springer New York, 2017, pp. 133–146 (cit. on p. 77).
- [222] J C Costello and G Stolovitzky. “Seeking the Wisdom of Crowds Through Challenge-Based Competitions in Biomedical Research”. In: *Clinical Pharmacology & Therapeutics* 93.5 (May 2013), pp. 396–398. DOI: 10.1038/clpt.2013.36 (cit. on p. 77).
- [223] J. Haas, S. Roth, K. Arnold, et al. “The Protein Model Portal—a comprehensive resource for protein structure and model information”. In: *Database* 2013 (Apr. 26, 2013), bat031–bat031. DOI: 10.1093/database/bat031 (cit. on p. 77).
- [224] Iddo Friedberg, Mark N. Wass, Sean D. Mooney, and Predrag Radivojac. “Ten Simple Rules for a Community Computational Challenge”. In: *PLOS Computational Biology* 11.4 (Apr. 23, 2015), e1004150. DOI: 10.1371/journal.pcbi.1004150 (cit. on p. 77).
- [225] G. Stolovitzky, D. Monroe, and A. Califano. “Dialogue on Reverse-Engineering Assessment and Methods: The DREAM of High-Throughput Pathway Inference”. In: *Annals of the New York Academy of Sciences* 1115.1 (Oct. 9, 2007), pp. 1–22. DOI: 10.1196/annals.1407.021 (cit. on p. 77).
- [226] Pablo Meyer, Leonidas G Alexopoulos, Thomas Bonk, et al. “Verification of systems biology research in the age of collaborative competition”. In: *Nature Biotechnology* 29.9 (Sept. 8, 2011), pp. 811–815. DOI: 10.1038/nbt.1968 (cit. on p. 77).
- [227] Sean M. Halper, Iman Farasat, and Howard M. Salis. “Automated Parameterization of Predictive Kinetic Metabolic Models from Sparse Datasets for Efficient Optimization of Many-Enzyme Heterologous Pathways”. In: *bioRxiv* (July 9, 2017), p. 161372. DOI: 10.1101/161372 (cit. on p. 78).

- [228] Richard C. Kolanczyk, Patricia Schmieder, William J. Jones, et al. “MetaPath: an electronic knowledge base for collating, exchanging and analyzing case studies of xenobiotic metabolism”. In: *Regulatory toxicology and pharmacology: RTP* 63.1 (June 2012), pp. 84–96. DOI: 10.1016/j.yrtph.2012.02.013 (cit. on p. 78).
- [229] Jörg Wicker, Tim Lorschach, Martin Gütlein, et al. “enviPath – The environmental contaminant biotransformation pathway resource”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D502–D508. DOI: 10.1093/nar/gkv1229 (cit. on p. 78).
- [230] Ola Spjuth, Patrik Rydberg, Egon L. Willighagen, Chris T. Evelo, and Nina Jeli-azkova. “XMetDB: an open access database for xenobiotic metabolism”. In: *Journal of Cheminformatics* 8.1 (Dec. 2016). DOI: 10.1186/s13321-016-0161-3 (cit. on p. 78).
- [231] Anastasia V. Rudik, Vladislav M. Bezhentsev, Alexander V. Dmitriev, et al. “Meta-Tox: Web Application for Predicting Structure and Toxicity of Xenobiotics’ Metabolites”. In: *Journal of Chemical Information and Modeling* 57.4 (Apr. 24, 2017), pp. 638–642. DOI: 10.1021/acs.jcim.6b00662 (cit. on p. 78).
- [232] Lynda B. M. Ellis, Bo Kyeng Hou, Wenjun Kang, and Lawrence P. Wackett. “The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining”. In: *Nucleic Acids Research* 31.1 (Jan. 1, 2003), pp. 262–265 (cit. on p. 78).
- [233] Junfeng Gao, Lynda B. M. Ellis, and Lawrence P. Wackett. “The University of Minnesota Biocatalysis/Biodegradation Database: improving public access”. In: *Nucleic Acids Research* 38 (Database issue Jan. 2010), pp. D488–491. DOI: 10.1093/nar/gkp771 (cit. on pp. 78, 97).
- [234] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, et al. “PubChem Substance and Compound databases”. In: *Nucleic Acids Research* (Sept. 22, 2015), gkv951. DOI: 10.1093/nar/gkv951 (cit. on p. 78).
- [235] Anne Morgat, Thierry Lombardot, Kristian B. Axelsen, et al. “Updates in Rhea – an expert curated resource of biochemical reactions”. In: *Nucleic Acids Research* 45 (D1 Jan. 4, 2017), pp. D415–D418. DOI: 10.1093/nar/gkw990 (cit. on p. 78).
- [236] Vincent Libis, Baudoin Delépine, and Jean-Loup Faulon. “Sensing new chemicals with bacterial transcription factors”. In: *Current Opinion in Microbiology* 33 (Oct. 2016), pp. 105–112. DOI: 10.1016/j.mib.2016.07.006 (cit. on p. 81).
- [237] Katrin Rössger, Ghislaine Charpin-El-Hamri, and Martin Fussenegger. “A closed-loop synthetic gene circuit for the treatment of diet-induced obesity in mice”. In: *Nature Communications* 4 (2013), p. 2825. DOI: 10.1038/ncomms3825 (cit. on p. 82).
- [238] Christian Kemmer, Marc Gitzinger, Marie Daoud-El Baba, et al. “Self-sufficient control of urate homeostasis in mice by a synthetic circuit”. In: *Nature Biotechnology* 28.4 (Apr. 2010), pp. 355–360. DOI: 10.1038/nbt.1617 (cit. on p. 82).
- [239] Srivatsan Raman, Jameson K. Rogers, Noah D. Taylor, and George M. Church. “Evolution-guided optimization of biosynthetic pathways”. In: *Proceedings of the National Academy of Sciences* (2014), p. 201409523 (cit. on p. 82).

- [240] Alec A. K. Nielsen, Bryan S. Der, Jonghyeon Shin, et al. “Genetic circuit design automation”. In: *Science (New York, N.Y.)* 352.6281 (Apr. 1, 2016), aac7341. DOI: 10.1126/science.aac7341 (cit. on p. 82).
- [241] Jameson K. Rogers, Christopher D. Guzman, Noah D. Taylor, et al. “Synthetic biosensors for precise gene control and real-time monitoring of metabolites”. In: *Nucleic Acids Research* 43.15 (Sept. 3, 2015), pp. 7648–7660. DOI: 10.1093/nar/gkv616 (cit. on pp. 82, 107, 133).
- [242] Michael J Cipriano, Pavel N Novichkov, Alexey E Kazakov, et al. “RegTransBase—a database of regulatory sequences and interactions based on literature: a resource for investigating transcriptional regulation in prokaryotes”. In: *BMC genomics* 14 (2013), p. 213. DOI: 10.1186/1471-2164-14-213 (cit. on pp. 83, 97, 110, 117).
- [243] Pavel S. Novichkov, Alexey E. Kazakov, Dmitry A. Ravcheev, et al. “RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria”. In: *BMC genomics* 14 (2013), p. 745. DOI: 10.1186/1471-2164-14-745 (cit. on pp. 83, 97, 110, 117).
- [244] Guillermo Carbajosa, Almudena Trigo, Alfonso Valencia, and Ildefonso Cases. “Bionemo: molecular information on biodegradation metabolism”. In: *Nucleic Acids Research* 37 (suppl 1 Jan. 1, 2009), pp. D598–D602. DOI: 10.1093/nar/gkn864 (cit. on pp. 83, 97, 110, 117).
- [245] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, et al. “RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond”. In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D133–D143. DOI: 10.1093/nar/gkv1156 (cit. on p. 83).
- [246] Víctor de Lorenzo, Rafael Silva-Rocha, Guillermo Carbajosa, Teca C. Galvão, and Ildefonso Cases. “Sensing Xenobiotic Compounds: Lessons from Bacteria that Face Pollutants in the Environment”. In: *Sensory Mechanisms in Bacteria: Molecular Aspects of Signal Recognition*. Caister Academic Press, U.K., Sept. 2010, pp. x + 268 (cit. on p. 83).
- [247] Dmitry A. Ravcheev, Matvei S. Khoroshkin, Olga N. Laikova, et al. “Comparative genomics and evolution of regulons of the LacI-family transcription factors”. In: *Frontiers in Microbiology* 5 (2014), p. 294. DOI: 10.3389/fmicb.2014.00294 (cit. on p. 83).
- [248] Marat D. Kazanov, Xiaoqing Li, Mikhail S. Gelfand, Andrei L. Osterman, and Dmitry A. Rodionov. “Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the Thermotogae phylum”. In: *Nucleic Acids Research* 41.2 (Jan. 1, 2013), pp. 790–803. DOI: 10.1093/nar/gks1184 (cit. on p. 83).
- [249] Sang Kyun Ahn, Leslie Cuthbertson, and Justin R. Nodwell. “Genome Context as a Predictive Tool for Identifying Regulatory Targets of the TetR Family Transcriptional Regulators”. In: *PLOS ONE* 7.11 (Nov. 30, 2012), e50562. DOI: 10.1371/journal.pone.0050562 (cit. on p. 83).
- [250] Leslie Cuthbertson, Sang Kyun Ahn, and Justin R. Nodwell. “Deglycosylation as a Mechanism of Inducible Antibiotic Resistance Revealed Using a Global Relational Tree for One-Component Regulators”. In: *Chemistry & Biology* 20.2 (Feb. 21, 2013), pp. 232–240. DOI: 10.1016/j.chembio.2012.11.011 (cit. on p. 83).

- [251] Taku Uchiyama and Kentaro Miyazaki. “Metagenomic Screening for Aromatic Compound-Responsive Transcriptional Regulators”. In: *PLOS ONE* 8.9 (Sept. 30, 2013), e75795. DOI: 10.1371/journal.pone.0075795 (cit. on p. 84).
- [252] Noah D. Taylor, Alexander S. Garruss, Rocco Moretti, et al. “Engineering an allosteric transcription factor to respond to new ligands”. In: *Nature Methods* advance online publication (Dec. 21, 2015). DOI: 10.1038/nmeth.3696 (cit. on pp. 84, 86, 95, 103, 107).
- [253] Shuang-Yan Tang and Patrick C. Cirino. “Design and application of a mevalonate-responsive regulatory protein”. In: *Angewandte Chemie (International Ed. in English)* 50.5 (Feb. 1, 2011), pp. 1084–1086. DOI: 10.1002/anie.201006083 (cit. on p. 84).
- [254] Shuang-Yan Tang, Shuai Qian, Olubolaji Akinterinwa, et al. “Screening for enhanced triacetic acid lactone production by recombinant *Escherichia coli* expressing a designed triacetic acid lactone reporter”. In: *Journal of the American Chemical Society* 135.27 (July 10, 2013), pp. 10099–10103. DOI: 10.1021/ja402654z (cit. on p. 84).
- [255] Wei Chen, Shan Zhang, Peixia Jiang, et al. “Design of an ectoine-responsive AraC mutant and its application in metabolic engineering of ectoine biosynthesis”. In: *Metabolic Engineering* 30 (July 2015), pp. 149–155. DOI: 10.1016/j.ymben.2015.05.004 (cit. on p. 84).
- [256] Srivatsan Raman, Noah Taylor, Naomi Genuth, Stanley Fields, and George M. Church. “Engineering allostery”. In: *Trends in genetics: TIG* 30.12 (Dec. 2014), pp. 521–528. DOI: 10.1016/j.tig.2014.09.004 (cit. on pp. 84, 86).
- [257] David L. Shis, Faiza Hussain, Sarah Meinhardt, Liskin Swint-Kruse, and Matthew R. Bennett. “Modular, multi-input transcriptional logic gating with orthogonal LacI/GalR family chimeras”. In: *ACS synthetic biology* 3.9 (Sept. 19, 2014), pp. 645–651. DOI: 10.1021/sb500262f (cit. on p. 86).
- [258] Sarah Meinhardt, Michael W. Manley, Nicole A. Becker, et al. “Novel insights from hybrid LacI/GalR proteins: family-wide functional attributes and biologically significant variation in transcription repression”. In: *Nucleic Acids Research* 40.21 (Nov. 2012), pp. 11139–11154. DOI: 10.1093/nar/gks806 (cit. on p. 86).
- [259] Filipa L. Sousa, Daniel J. Parente, David L. Shis, et al. “AlloRep: A Repository of Sequence, Structural and Mutagenesis Data for the LacI/GalR Transcription Regulators”. In: *Journal of Molecular Biology* (Sept. 25, 2015). DOI: 10.1016/j.jmb.2015.09.015 (cit. on p. 86).
- [260] Ramesh K. Jha, Subhendu Chakraborti, Theresa L. Kern, David T. Fox, and Charlie E. M. Strauss. “Rosetta comparative modeling for library design: Engineering alternative inducer specificity in a transcription factor”. In: *Proteins: Structure, Function, and Bioinformatics* (May 1, 2015), n/a–n/a. DOI: 10.1002/prot.24828 (cit. on pp. 86, 95, 103).
- [261] HyeonSeok Shin and Byung-Kwan Cho. “Rational Protein Engineering Guided by Deep Mutational Scanning”. In: *International Journal of Molecular Sciences* 16.9 (2015), pp. 23094–23110. DOI: 10.3390/ijms160923094 (cit. on p. 86).

- [262] Douglas M. Fowler and Stanley Fields. “Deep mutational scanning: a new style of protein science”. In: *Nature Methods* 11.8 (Aug. 2014), pp. 801–807. DOI: 10.1038/nmeth.3027 (cit. on p. 86).
- [263] Emmanuel L. C. de los Santos, Joseph T. Meyerowitz, Stephen L. Mayo, and Richard M. Murray. “Engineering Transcriptional Regulator Effector Specificity Using Computational Design and In Vitro Rapid Prototyping: Developing a Vanillin Sensor”. In: *ACS synthetic biology* 5.4 (Apr. 15, 2016), pp. 287–295. DOI: 10.1021/acssynbio.5b00090 (cit. on p. 87).
- [264] Alan Jobe and Suzanne Bourgeois. “lac repressor-operator interaction”. In: *Journal of Molecular Biology* 69.3 (Aug. 28, 1972), pp. 397–408. DOI: 10.1016/0022-2836(72)90253-7 (cit. on p. 87).
- [265] Peter W. Coschigano and Bradley J. Bishop. “Role of benzylsuccinate in the induction of the tutE tutFDGH gene complex of T. aromatica strain T1”. In: *FEMS Microbiology Letters* 231.2 (Feb. 1, 2004), pp. 261–266. DOI: 10.1016/S0378-1097(04)00005-9 (cit. on p. 87).
- [266] Ngoc Anh Thu Ho, Stephanie S. Dawes, Adam M. Crowe, et al. “The Structure of the Transcriptional Repressor KstR in Complex with CoA Thioester Cholesterol Metabolites Sheds Light on the Regulation of Cholesterol Catabolism in Mycobacterium tuberculosis”. In: *Journal of Biological Chemistry* 291.14 (Apr. 1, 2016), pp. 7256–7266. DOI: 10.1074/jbc.M115.707760 (cit. on pp. 87, 90).
- [267] Esther García-Fernández, Francisco Javier Medrano, Beatriz Galán, and José Luis García. “Deciphering the Transcriptional Regulation of Cholesterol Catabolic Pathway in Mycobacteria IDENTIFICATION OF THE INDUCER OF KstR REPRESSOR”. In: *Journal of Biological Chemistry* 289.25 (June 20, 2014), pp. 17576–17588. DOI: 10.1074/jbc.M113.545715 (cit. on pp. 87, 90).
- [268] Tetsu Shimizu and Akira Nakamura. “Characterization of LgnR, an IclR family transcriptional regulator involved in the regulation of l-gluconate catabolic genes in Paracoccus sp. 43P”. In: *Microbiology* 160.3 (2014), pp. 623–634. DOI: 10.1099/mic.0.074286-0 (cit. on p. 87).
- [269] Javier F. Juárez, Huixiang Liu, María T. Zamarro, et al. “Unraveling the Specific Regulation of the Central Pathway for Anaerobic Degradation of 3-Methylbenzoate”. In: *Journal of Biological Chemistry* 290.19 (May 8, 2015), pp. 12165–12183. DOI: 10.1074/jbc.M115.637074 (cit. on pp. 87, 89).
- [270] Daisuke Kasai, Naofumi Kamimura, Kenta Tani, et al. “Characterization of FerC, a MarR-type transcriptional regulator, involved in transcriptional regulation of the ferulate catabolic operon in Sphingobium sp. strain SYK-6”. In: *FEMS Microbiology Letters* 332.1 (July 1, 2012), pp. 68–75. DOI: 10.1111/j.1574-6968.2012.02576.x (cit. on p. 87).
- [271] Hidetada Hirakawa, Amy L. Schaefer, E. Peter Greenberg, and Caroline S. Harwood. “Anaerobic p-Coumarate Degradation by Rhodopseudomonas palustris and Identification of CouR, a MarR Repressor Protein That Binds p-Coumaroyl Coenzyme A”. In: *Journal of Bacteriology* 194.8 (Apr. 15, 2012), pp. 1960–1967. DOI: 10.1128/JB.06817-11 (cit. on p. 87).

- [272] Dong-Wei Chen, Yun Zhang, Cheng-Ying Jiang, and Shuang-Jiang Liu. “Benzoate Metabolism Intermediate Benzoyl Coenzyme A Affects Gentisate Pathway Regulation in *Comamonas testosteroni*”. In: *Applied and Environmental Microbiology* 80.13 (July 1, 2014), pp. 4051–4062. DOI: 10.1128/AEM.01146-14 (cit. on p. 87).
- [273] Keiko Sakamoto, Yoshihiro Agari, Seiki Kuramitsu, and Akeo Shinkai. “Phenylacetyl Coenzyme A Is an Effector Molecule of the TetR Family Transcriptional Repressor PaaR from *Thermus thermophilus* HB8”. In: *Journal of Bacteriology* 193.17 (Sept. 1, 2011), pp. 4388–4395. DOI: 10.1128/JB.05203-11 (cit. on p. 87).
- [274] Haoran Xue, Hailing Shi, Zhou Yu, et al. “Design, construction, and characterization of a set of biosensors for aromatic compounds”. In: *ACS synthetic biology* 3.12 (Dec. 19, 2014), pp. 1011–1014. DOI: 10.1021/sb500023f (cit. on pp. 87, 89, 107, 133, 136).
- [275] Jameson K. Rogers and George M. Church. “Genetically encoded sensors enable real-time observation of metabolite production”. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.9 (Mar. 1, 2016), pp. 2388–2393. DOI: 10.1073/pnas.1600375113 (cit. on pp. 87, 89, 95, 115, 133, 136).
- [276] Robert W. Wheatley, Summie Lo, Larisa J. Jancewicz, Megan L. Dugdale, and Reuben E. Huber. “Structural explanation for allolactose (lac operon inducer) synthesis by lacZ β -galactosidase and the evolutionary relationship between allolactose synthesis and the lac repressor”. In: *The Journal of Biological Chemistry* 288.18 (May 3, 2013), pp. 12993–13005. DOI: 10.1074/jbc.M113.455436 (cit. on pp. 89, 117).
- [277] Katharina Pflüger-Grau and Víctor de Lorenzo. “From the phosphoenolpyruvate phosphotransferase system to selfish metabolism: a story retraced in *Pseudomonas putida*”. In: *FEMS microbiology letters* 356.2 (July 2014), pp. 144–153. DOI: 10.1111/1574-6968.12459 (cit. on pp. 90, 117).
- [278] Yonatan Savir, Benjamin P. Tu, and Michael Springer. “Competitive Inhibition Can Linearize Dose-Response and Generate a Linear Rectifier”. In: *Cell Systems* 1.3 (Sept. 23, 2015), pp. 238–245. DOI: 10.1016/j.cels.2015.09.001 (cit. on p. 90).
- [279] Su-Lim Choi, Eugene Rha, Sang Jun Lee, et al. “Toward a Generalized and High-throughput Enzyme Screening System Based on Artificial Genetic Circuits”. In: *ACS Synthetic Biology* 3.3 (Mar. 21, 2014), pp. 163–171. DOI: 10.1021/sb400112u (cit. on p. 90).
- [280] Young-Su Jeong, Su-Lim Choi, Hyun-Ho Kyeong, et al. “High-throughput screening system based on phenolics-responsive transcription activator for directed evolution of organophosphate-degrading enzymes”. In: *Protein engineering, design & selection: PEDS* 25.11 (Nov. 2012), pp. 725–731. DOI: 10.1093/protein/gzs071 (cit. on p. 90).
- [281] Taku Uchiyama and Kentaro Miyazaki. “Product-Induced Gene Expression, a Product-Responsive Reporter Assay Used To Screen Metagenomic Libraries for Enzyme-Encoding Genes”. In: *Applied and Environmental Microbiology* 76.21 (Nov. 1, 2010), pp. 7029–7035. DOI: 10.1128/AEM.00464-10 (cit. on p. 90).

- [282] Solvej Siedler, Steen G Stahlhut, Sailesh Malla, Jérôme Maury, and Ana Rute Neves. “Novel biosensors based on flavonoid-responsive transcriptional regulators introduced into *Escherichia coli*”. In: *Metabolic engineering* 21 (Jan. 2014), pp. 2–8. DOI: 10.1016/j.ymben.2013.10.011 (cit. on p. 91).
- [283] Regina Mahr, Cornelia Gätgens, Jochem Gätgens, et al. “Biosensor-driven adaptive laboratory evolution of L-valine production in *Corynebacterium glutamicum*”. In: *Metabolic Engineering* 32 (Nov. 2015), pp. 184–194. DOI: 10.1016/j.ymben.2015.09.017 (cit. on p. 94).
- [284] Jason T Boock, Apoorv Gupta, and Kristala LJ Prather. “Screening and modular design for metabolic pathway optimization”. In: *Current Opinion in Biotechnology. Pathway engineering* 36 (Dec. 2015), pp. 189–198. DOI: 10.1016/j.copbio.2015.08.013 (cit. on p. 94).
- [285] Aitor de las Heras and Víctor de Lorenzo. “In situ detection of aromatic compounds with biosensor *Pseudomonas putida* cells preserved and delivered to soil in water-soluble gelatin capsules”. In: *Analytical and Bioanalytical Chemistry* 400.4 (May 2011), pp. 1093–1104. DOI: 10.1007/s00216-010-4558-y (cit. on p. 94).
- [286] Jennifer A. N. Brophy and Christopher A. Voigt. “Principles of genetic circuit design”. In: *Nature Methods* 11.5 (May 2014), pp. 508–520. DOI: 10.1038/nmeth.2926 (cit. on pp. 94, 106).
- [287] Alexis Courbet, Drew Endy, Eric Renard, Franck Molina, and Jérôme Bonnet. “Detection of pathological biomarkers in human clinical samples via amplifying genetic switches and logic gates”. In: *Science Translational Medicine* 7.289 (May 27, 2015), 289ra83. DOI: 10.1126/scitranslmed.aaa3601 (cit. on pp. 94, 106, 116).
- [288] Shuang-Yan Tang, Hossein Fazelinia, and Patrick C. Cirino. “AraC regulatory protein mutants with altered effector specificity”. In: *Journal of the American Chemical Society* 130.15 (Apr. 16, 2008), pp. 5267–5271. DOI: 10.1021/ja7109053 (cit. on p. 95).
- [289] Taku Uchiyama and Kazuya Watanabe. “Substrate-induced gene expression (SIGEX) screening of metagenome libraries”. In: *Nature Protocols* 3.7 (2008), pp. 1202–1212. DOI: 10.1038/nprot.2008.96 (cit. on pp. 95, 103).
- [290] Christine Nicole S. Santos and Gregory Stephanopoulos. “Melanin-based high-throughput screen for L-tyrosine production in *Escherichia coli*”. In: *Applied and Environmental Microbiology* 74.4 (Feb. 2008), pp. 1190–1197. DOI: 10.1128/AEM.02448-07 (cit. on p. 95).
- [291] William C. DeLoache, Zachary N. Russ, Lauren Narcross, et al. “An enzyme-coupled biosensor enables (S)-reticuline production in yeast from glucose”. In: *Nature Chemical Biology* 11.7 (July 2015), pp. 465–471. DOI: 10.1038/nchembio.1816 (cit. on p. 95).
- [292] Anne Morgat, Kristian B. Axelsen, Thierry Lombardot, et al. “Updates in Rhea—a manually curated resource of biochemical reactions”. In: *Nucleic Acids Research* 43 (Database issue Jan. 2015), pp. D459–464. DOI: 10.1093/nar/gku961 (cit. on pp. 96, 110, 117).

- [293] Ron Caspi, Richard Billington, Luciana Ferrer, et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases". In: *Nucleic Acids Research* 44 (D1 Jan. 4, 2016), pp. D471–480. DOI: 10.1093/nar/gkv1164 (cit. on p. 96).
- [294] Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, et al. "RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more". In: *Nucleic acids research* 41 (Database issue Jan. 2013), pp. D203–213. DOI: 10.1093/nar/gks1201 (cit. on pp. 97, 110, 117).
- [295] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11–15 (cit. on pp. 98, 118).
- [296] P. Jaccard. "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines". In: *Bulletin del la Société Vaudoise des Sciences Naturelles* 37 (1901), pp. 241–272 (cit. on pp. 99, 118).
- [297] Ahmad Sarreshtehdar Emrani, Noor Mohammad Danesh, Mohammad Ramezani, Seyed Mohammad Taghdisi, and Khalil Abnous. "A novel fluorescent aptasensor based on hairpin structure of complementary strand of aptamer and nanoparticles as a signal amplification approach for ultrasensitive detection of cocaine". In: *Biosensors & Bioelectronics* 79 (May 15, 2016), pp. 288–293. DOI: 10.1016/j.bios.2015.12.025 (cit. on p. 100).
- [298] Joy Sinha, Samuel J. Reyes, and Justin P. Gallivan. "Reprogramming bacteria to seek and destroy an herbicide". In: *Nature Chemical Biology* 6.6 (June 2010), pp. 464–470. DOI: 10.1038/nchembio.369 (cit. on p. 102).
- [299] Joy Sinha, Samuel Reyes, and Justin P. Gallivan. "Retraction: Reprogramming bacteria to seek and destroy an herbicide". In: *Nature Chemical Biology* 10.3 (Mar. 2014), pp. 239–239. DOI: 10.1038/nchembio0314-239a (cit. on p. 102).
- [300] C. Roodveldt and D. S. Tawfik. "Directed evolution of phosphotriesterase from *Pseudomonas diminuta* for heterologous expression in *Escherichia coli* results in stabilization of the metal-free state". In: *Protein engineering, design & selection: PEDS* 18.1 (Jan. 2005), pp. 51–58. DOI: 10.1093/protein/gzi005 (cit. on pp. 102, 116, 119).
- [301] Jun-Jie Zhang, Hong Liu, Yi Xiao, Xian-En Zhang, and Ning-Yi Zhou. "Identification and characterization of catabolic para-nitrophenol 4-monooxygenase and para-benzoquinone reductase from *Pseudomonas* sp. strain WBC-3". In: *Journal of Bacteriology* 191.8 (Apr. 2009), pp. 2703–2710. DOI: 10.1128/JB.01566-08 (cit. on p. 102).
- [302] Deepak Mishra, Phillip M. Rivera, Allen Lin, Domitilla Del Vecchio, and Ron Weiss. "A load driver device for engineering modularity in biological networks". In: *Nature Biotechnology* 32.12 (Dec. 2014), pp. 1268–1275. DOI: 10.1038/nbt.3044 (cit. on p. 106).
- [303] Fahim Farzadfard and Timothy K. Lu. "Genomically encoded analog memory with precise in vivo DNA writing in living cell populations". In: *Science* 346.6211 (Nov. 14, 2014), p. 1256272. DOI: 10.1126/science.1256272 (cit. on p. 106).

- [304] Jerome Bonnet, Peter Yin, Monica E. Ortiz, Pakpoom Subsoontorn, and Drew Endy. “Amplifying Genetic Logic Gates”. In: *Science* 340.6132 (May 3, 2013), pp. 599–603. DOI: 10.1126/science.1232758 (cit. on p. 106).
- [305] Arthur Prindle, Jangir Selimkhanov, Howard Li, et al. “Rapid and tunable post-translational coupling of genetic circuits”. In: *Nature* 508.7496 (Apr. 17, 2014), pp. 387–391. DOI: 10.1038/nature13238 (cit. on p. 106).
- [306] Laura Prochazka, Bartolomeo Angelici, Benjamin Haefliger, and Yaakov Benenson. “Highly modular bow-tie gene circuits with programmable dynamic behaviour”. In: *Nature Communications* 5 (Oct. 14, 2014), p. 4729. DOI: 10.1038/ncomms5729 (cit. on p. 106).
- [307] Travis S. Bayer and Christina D. Smolke. “Programmable ligand-controlled riboregulators of eukaryotic gene expression”. In: *Nature Biotechnology* 23.3 (Mar. 2005), pp. 337–343. DOI: 10.1038/nbt1069 (cit. on p. 107).
- [308] Rafael Silva-Rocha and Victor de Lorenzo. “Engineering multicellular logic in bacteria with metabolic wires”. In: *ACS synthetic biology* 3.4 (Apr. 18, 2014), pp. 204–209. DOI: 10.1021/sb400064y (cit. on pp. 107, 136).
- [309] Thomas Bulter, Sun-Gu Lee, Wilson WaiChun Wong, et al. “Design of artificial cell-cell communication using gene and metabolic networks”. In: *Proceedings of the National Academy of Sciences of the United States of America* 101.8 (Feb. 24, 2004), pp. 2299–2304 (cit. on p. 107).
- [310] Jacob Beal, Ting Lu, and Ron Weiss. “Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks”. In: *PloS One* 6.8 (2011), e22490. DOI: 10.1371/journal.pone.0022490 (cit. on p. 108).
- [311] Kevin Clancy and Christopher A. Voigt. “Programming cells: towards an automated ‘Genetic Compiler’”. In: *Current Opinion in Biotechnology* 21.4 (Aug. 2010), pp. 572–581. DOI: 10.1016/j.copbio.2010.07.005 (cit. on p. 108).
- [312] Vivian Law, Craig Knox, Yannick Djoumbou, et al. “DrugBank 4.0: shedding new light on drug metabolism”. In: *Nucleic acids research* 42 (Database issue Jan. 2014), pp. D1091–1097. DOI: 10.1093/nar/gkt1068 (cit. on pp. 108, 117).
- [313] Daniel Krewski, Melvin E. Andersen, Ellen Mantus, and Lauren Zeise. “Toxicity testing in the 21st century: implications for human health risk assessment”. In: *Risk Analysis: An Official Publication of the Society for Risk Analysis* 29.4 (Apr. 2009), pp. 474–479. DOI: 10.1111/j.1539-6924.2008.01150.x (cit. on pp. 108, 117).
- [314] Ron Caspi, Tomer Altman, Richard Billington, et al. “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases”. In: *Nucleic Acids Research* 42 (D1 Jan. 1, 2014), pp. D459–D471. DOI: 10.1093/nar/gkt1103 (cit. on pp. 110, 117).
- [315] Ryosuke Kojima, Dominique Aubel, and Martin Fussenegger. “Novel theranostic agents for next-generation personalized medicine: small molecules, nanoparticles, and engineered mammalian cells”. In: *Current Opinion in Chemical Biology* 28 (Oct. 2015), pp. 29–38. DOI: 10.1016/j.cbpa.2015.05.021 (cit. on p. 113).

- [316] Luis Briseño-Roa, Christopher M. Timperley, Andrew D. Griffiths, and Alan R. Fersht. “Phosphotriesterase variants with high methylphosphonate activity and strong negative trade-off against phosphotriesters”. In: *Protein engineering, design & selection: PEDS* 24.1 (Jan. 2011), pp. 151–159. DOI: 10.1093/protein/gzq076 (cit. on p. 116).
- [317] Magnus Steigedal and Svein Valla. “The *Acinetobacter* sp. *chnB* promoter together with its cognate positive regulator ChnR is an attractive new candidate for metabolic engineering applications in bacteria”. In: *Metabolic Engineering* 10.2 (Mar. 2008), pp. 121–129. DOI: 10.1016/j.ymben.2007.08.002 (cit. on p. 116).
- [318] Roberta Lentini, Silvia Perez Santero, Fabio Chizzolini, et al. “Integrating artificial with natural cells to translate chemical messages that direct *E. coli* behaviour”. In: *Nature Communications* 5 (May 30, 2014), p. 4012. DOI: 10.1038/ncomms5012 (cit. on p. 116).
- [319] Yuval Elani, Robert V. Law, and Oscar Ces. “Vesicle-based artificial cells as chemical microreactors with spatially segregated reaction pathways”. In: *Nature Communications* 5 (Oct. 29, 2014), p. 5305. DOI: 10.1038/ncomms6305 (cit. on p. 116).
- [320] Victor Chubukov, Luca Gerosa, Karl Kochanowski, and Uwe Sauer. “Coordination of microbial metabolism”. In: *Nature Reviews Microbiology* 12.5 (May 2014), pp. 327–340. DOI: 10.1038/nrmicro3238 (cit. on p. 117).
- [321] Keisuke Miyauchi, Haeng-Seog Lee, Masao Fukuda, Masamichi Takagi, and Yuji Nagata. “Cloning and characterization of *linR*, involved in regulation of the downstream pathway for gamma-hexachlorocyclohexane degradation in *Sphingomonas paucimobilis* UT26”. In: *Applied and Environmental Microbiology* 68.4 (Apr. 2002), pp. 1803–1807 (cit. on p. 119).
- [322] Jun Min, Jun-Jie Zhang, and Ning-Yi Zhou. “The gene cluster for para-nitrophenol catabolism is responsible for 2-chloro-4-nitrophenol degradation in *Burkholderia* sp. strain SJ98”. In: *Applied and Environmental Microbiology* 80.19 (Oct. 2014), pp. 6212–6222. DOI: 10.1128/AEM.02093-14 (cit. on p. 119).
- [323] Carola Engler, Romy Kandzia, and Sylvestre Marillonnet. “A one pot, one step, precision cloning method with high throughput capability”. In: *PloS One* 3.11 (2008), e3647. DOI: 10.1371/journal.pone.0003647 (cit. on p. 120).
- [324] Tomoya Baba, Takeshi Ara, Miki Hasegawa, et al. “Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection”. In: *Molecular Systems Biology* 2 (2006), p. 2006.0008. DOI: 10.1038/msb4100050 (cit. on p. 121).
- [325] Jingwei Zhang, Jesus F. Barajas, Mehmet Burdu, et al. “Development of a Transcription Factor Based Lactam Biosensor”. In: *ACS synthetic biology* (Dec. 20, 2016). DOI: 10.1021/acssynbio.6b00136 (cit. on p. 133).
- [326] Regina Mahr, Raphael Freiherr von Boeselager, Johanna Wiechert, and Julia Frunzke. “Screening of an *Escherichia coli* promoter library for a phenylalanine biosensor”. In: *Applied Microbiology and Biotechnology* (May 12, 2016). DOI: 10.1007/s00253-016-7575-8 (cit. on p. 133).

- [327] Brecht De Paepe, Gert Peters, Pieter Coussement, Jo Maertens, and Marjan De Mey. "Tailor-made transcriptional biosensors for optimizing microbial cell factories". In: *Journal of Industrial Microbiology & Biotechnology* (Nov. 11, 2016). DOI: 10.1007/s10295-016-1862-3 (cit. on pp. 133, 136).
- [328] Christopher S. Frei, Zhiqing Wang, Shuai Qian, et al. "Analysis of amino acid substitutions in AraC variants that respond to triacetic acid lactone". In: *Protein Science: A Publication of the Protein Society* 25.4 (Apr. 2016), pp. 804–814. DOI: 10.1002/pro.2873 (cit. on p. 133).
- [329] Jieyuan Wu, Peixia Jiang, Wei Chen, et al. "Design and application of a lactulose biosensor". In: *Scientific Reports* 7 (Apr. 7, 2017), p. 45994. DOI: 10.1038/srep45994 (cit. on p. 133).
- [330] Solvej Siedler, Narendar K. Khatri, Andrea Zsohár, et al. "Development of a bacterial biosensor for rapid screening of yeast p-coumaric acid production". In: *ACS synthetic biology* (May 23, 2017). DOI: 10.1021/acssynbio.7b00009 (cit. on p. 133).
- [331] Christian Kasey, Mounir Zerrad, Yiwei Li, Thomas Ashton Cropp, and Gavin J. Williams. "Development of transcription factor-based designer macrolide biosensors for metabolic engineering and synthetic biology". In: *ACS Synthetic Biology* (Sept. 26, 2017). DOI: 10.1021/acssynbio.7b00287 (cit. on p. 133).
- [332] Supratim Mukherjee, Rekha Seshadri, Neha J Varghese, et al. "1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life". In: *Nature Biotechnology* 35.7 (June 12, 2017), pp. 676–683. DOI: 10.1038/nbt.3886 (cit. on p. 133).
- [333] J. Hastings, P. de Matos, A. Dekker, et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013". In: *Nucleic Acids Research* 41 (D1 Jan. 1, 2013), pp. D456–D463. DOI: 10.1093/nar/gks1146 (cit. on p. 134).
- [334] Torsten Blum and Oliver Kohlbacher. "MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization". In: *Bioinformatics* 24.18 (Sept. 15, 2008), pp. 2108–2109. DOI: 10.1093/bioinformatics/btn360 (cit. on p. 134).
- [335] Deguo Xia, Haoran Zheng, Zhiqiang Liu, et al. "MRSD: a web server for Metabolic Route Search and Design". In: *Bioinformatics* 27.11 (June 1, 2011), pp. 1581–1582. DOI: 10.1093/bioinformatics/btr160 (cit. on p. 134).
- [336] Solvej Siedler, Georg Schendzielorz, Stephan Binder, et al. "SoxR as a Single-Cell Biosensor for NADPH-Consuming Enzymes in *Escherichia coli*". In: *ACS Synthetic Biology* 3.1 (Jan. 17, 2014), pp. 41–47. DOI: 10.1021/sb400110j (cit. on p. 135).
- [337] Jie Zhang, Nikolaus Sonnenschein, Thomas Peter Boye Pihl, et al. "Engineering an NADPH/NADP⁺ redox biosensor in yeast". In: *ACS synthetic biology* (July 15, 2016). DOI: 10.1021/acssynbio.6b00135 (cit. on p. 135).
- [338] Huiqing Chong and Chi Bun Ching. "Development of Colorimetric-Based Whole-Cell Biosensor for Organophosphorus Compounds by Engineering Transcription Regulator DmpR". In: *ACS Synthetic Biology* (July 5, 2016). DOI: 10.1021/acssynbio.6b00061 (cit. on p. 136).

- [339] Jacques Fastrez. “Engineering allosteric regulation into biological catalysts”. In: *Chembiochem: A European Journal of Chemical Biology* 10.18 (Dec. 14, 2009), pp. 2824–2835. DOI: 10.1002/cbic.200900590 (cit. on p. 136).
- [340] Nikolay V. Dokholyan. “Controlling Allosteric Networks in Proteins”. In: *Chemical Reviews* (Feb. 19, 2016). DOI: 10.1021/acs.chemrev.5b00544 (cit. on p. 136).

Websites

- [@1] Greg Landrum. *RDKit: Open-source cheminformatics*. URL: <http://www.rdkit.org/> (visited on Feb. 8, 2016) (cit. on pp. 49, 50).
- [@2] Mercatus Center. *Garry Kasparov on AI, Chess, and the Future of Creativity*. Medium. May 10, 2017. URL: <https://medium.com/conversations-with-tyler/garry-kasparov-tyler-cowen-chess-iq-ai-putin-3bf28baf4dba> (visited on Oct. 3, 2017) (cit. on p. 137).

Title : Computer-aided design (CAD) tools for bioproduction and biosensing pathway engineering

Keywords: computer-aided design, retrosynthesis, metabolic engineering, synthetic biology, biosensor

Abstract: Advances in systems and synthetic biology are fueling our ability to develop successful metabolic engineering applications for the sustainable production of bio-based chemicals. We can envision a future in which *designer cells* could be engineered to transform any carbon source into any target compound. This daunting task will be achieved by leveraging methods that proved themselves in other engineering disciplines. Among those, the use of Computer Aided Design (CAD) softwares is expected to reduce the amount of time and expert knowledge needed to design *de novo* metabolic pathways. The first part of this thesis is dedicated to our pathway prediction algorithm and its

CAD implementations. Most notably, we will present RetroPath2.0, a versatile reaction network prediction framework focused on retrosynthesis that is built to be easily extensible by the community. In the second part, we will highlight the interest of intracellular biosensors for metabolic engineering and introduce SensiPath, a web application that uses a reaction prediction engine to design biosensing circuits for compounds for which no direct biosensors are known. Altogether, this thesis proposes that bioCAD tools should focus on empowering users' creativity and encourage them to explore original applications.

Titre: Outils de conception assistée par ordinateur pour l'ingénierie de voies métaboliques de bioproduction et de biodétection

Mots clefs : conception assistée par ordinateur, rétrosynthèse, ingénierie métabolique, biologie synthétique, biosenseur

Résumé : Les récentes avancées en biologie des systèmes et en biologie synthétique contribuent déjà au fleurissement d'applications en ingénierie métabolique visant une bioproduction renouvelable de composés chimiques. Nous pouvons entrevoir un futur où des microbes seraient conçus à la carte afin de valoriser n'importe quelle source de carbone en n'importe quel composé d'intérêt. Si la route est longue avant l'accomplissement d'un tel objectif, son parcours devrait en être grandement facilité par l'exploitation de méthodes d'ingénierie déjà éprouvées dans d'autres disciplines. On s'attend entre autre à ce que l'utilisation de logiciels de Conception Assistée par Ordinateur (CAO) diminue le temps et l'expertise nécessaires à la construction de voies métaboliques n'existant pas dans la nature. La première partie de

cette thèse est dédiée à notre méthode de prédiction de voies métaboliques et à ses implémentations. Nous décrivons tout particulièrement RetroPath2.0, un outil de prédiction de réseaux de réactions mettant l'accent sur les applications de rétrosynthèse, et qui est construit pour être facilement extensible par la communauté. Dans la seconde partie, nous détaillons l'intérêt des biosenseurs intracellulaires pour l'ingénierie métabolique et introduisons SensiPath; une application web qui exploite un outil de prédiction de réactions pour concevoir des circuits métaboliques permettant la biodétection de composés pour lesquels aucun biosenseur direct n'est connu. Dans l'ensemble, cette thèse propose que les outils de bioCAO devraient permettre de révéler la créativité de leurs utilisateurs et encourager l'exploration de nouvelles applications.