Metabolic Engineering 45 (2018) 158–170

Contents lists available at ScienceDirect

Metabolic Engineering

journal homepage: www.elsevier.com/locate/meteng

RetroPath2.0: A retrosynthesis workflow for metabolic engineers

Baudoin Delépine[a,b,c,1], Thomas Duigou[c,1], Pablo Carbonell[d], Jean-Loup Faulon[a,b,c,d,*]

[a] CNRS-UMR8030/Laboratoire iSSB, Université Paris-Saclay, Évry 91000, France
[b] CEA, DRF, IG, Genoscope, Évry 91000, France
[c] Micalis Institute, INRA, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[d] SYNBIOCHEM Centre, Manchester Institute of Biotechnology, University of Manchester, Manchester, UK

ARTICLE INFO

Keywords:
Retrosynthesis
Workflow
CAD software
Pathways prediction
Metabolic space

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 bench 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 biosensors. 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.

1. 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 (Paddon et al., 2013; Keasling, 2014). 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 (Medema et al., 2012; Copeland et al., 2012; Hadadi and Hatzimanikatis, 2015; Lee and Kim, 2015)). While some computationally-driven strategies make use of known metabolic reactions albeit not necessarily in the same species (Rodrigo et al., 2008; Moriya et al., 2010) others allow to design pathways that encompass novel reactions not stored in metabolic databases. These latter tools make use of retrosynthesis algorithms (Marchant et al., 2008; Moriya et al., 2010; Henry et al., 2010; Carbonell et al., 2011b, 2014a; Yim et al., 2011; Liu et al., 2014; Campodonico et al., 2014; Hadadi et al., 2016a).

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 (Carbonell et al., 2012) and rank them based on several criteria including enzyme availability and performance, product and intermediate compound

* Correspondence to: Micalis, Institut National de la Recherche Agronomique, Domaine de Vilvert, 78352 Jouy-en-Josas, France.
E-mail address: jean-loup.faulon@inra.fr (J.-L. Faulon).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ymben.2017.12.002
Received 28 April 2017; Received in revised form 3 November 2017; Accepted 5 December 2017
Available online 09 December 2017
1096-7176/ © 2017 The Authors. Published by Elsevier Inc. on behalf of International Metabolic Engineering Society. This is an open access article under the CC BY license (http://creativecommons.org/licenses/BY/4.0/).

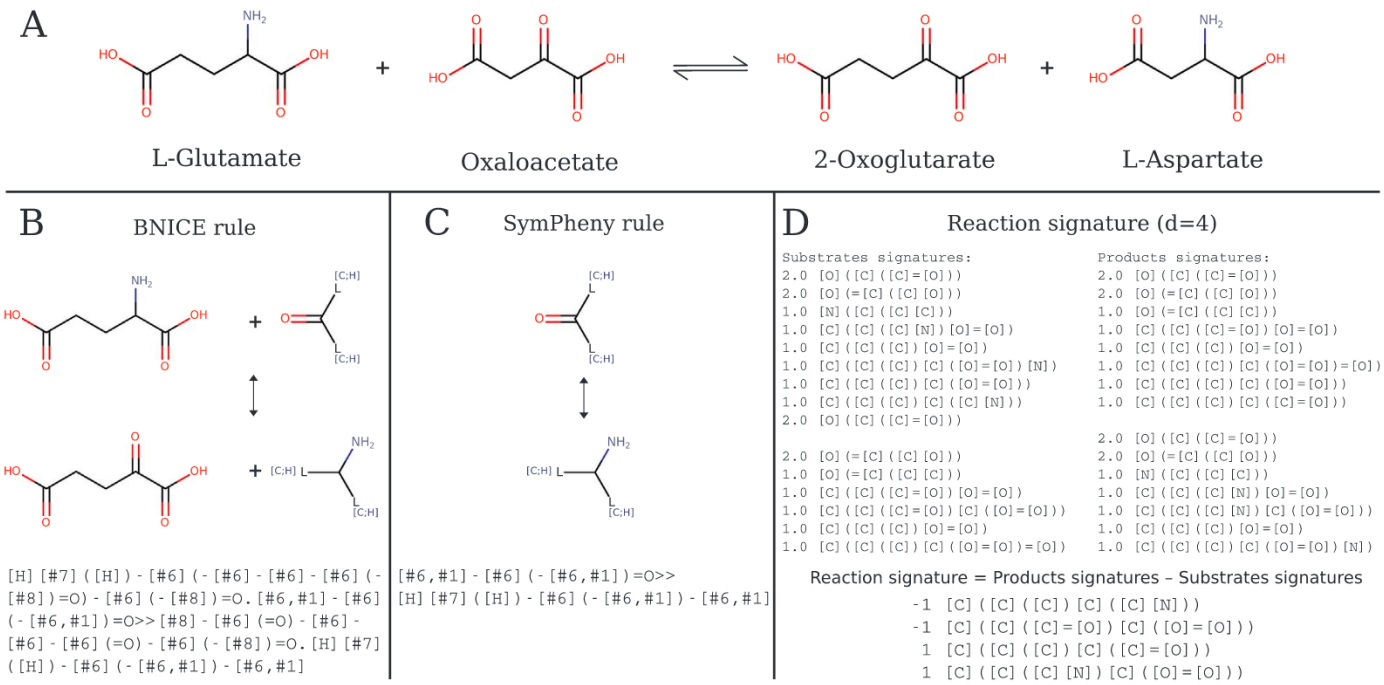


Fig. 1. Example of reaction rules. A. Generalized reaction rules for the transaminase 2.6.1.1. B and C. BNICE rules and SimPheny rules were extracted from Henry et al. (2010) and Yim et al., (2011). These are the only rules with EC number 2.6.1. In both cases the rules are represented by SMARTS strings. D. The reaction signature rule was computed using the MolSig package, *d* represents the signature diameter. See Carbonell et al. (2013) for definition and examples of signatures.

toxicities (Planson et al., 2012) or the theoretical yield of the desired compound (Campodonico et al., 2014; Carbonell et al., 2014b; Cho et al., 2010; Liu et al., 2014). Interestingly, retrosynthesis networks exploitation is not strictly limited to retrosynthesis. Applications have been proposed to predict biodegradation routes (Hou et al., 2004; Oh et al., 2007; Finley et al., 2009) in order to identify unknown compounds from the underground metabolism (Jeffryes et al., 2015), to predict the transitions of labelled atoms in metabolic networks (Arita, 2003; Hadadi et al., 2016b), and to design biosensing circuits for compounds for which no direct biosensors are known (Delépine et al., 2016). 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 reaction 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 webservers (Campodonico et al., 2014; Carbonell et al., 2014b; Jeffryes et al., 2015; Hadadi et al., 2016a) 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 (Schofield et al., 2009; McNutt et al., 2016; Haug et al., 2017) we believe this lack of flexibility should be overcome.

In this spirit, we developed the RetroPath2.0 workflow on the KNIME analytics platform (Berthold et al., 2008) 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 programing. 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 (Berthold et al., 2008; Warr, 2012). 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 (Marchant et al., 2008; Moriya et al., 2010; Henry et al., 2010; Yim et al., 2011; Liu et al., 2014; Carbonell et al., 2014a; Campodonico et al., 2014; Hadadi et al., 2016a). We then outline our proposed solution through several applications in metabolic engineering and biosensor engineering. RetroPath2.0 is available at myExperiment.org (https://www.myexperiment.org/workflows/4987.html) along a set of reaction rules and some classic metabolic engineering examples to test RetroPath2.0 features.

2. Theoretical background

2.1. Encoding reactions 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 (Kayala et al., 2011; Latino and Aires-de-Sousa, 2011). 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 (Dugundji and Ugi, 1973), reaction SMARTS (Daylight, 2017), RDM patterns (Oh et al., 2007), and reaction signatures (Carbonell et al., 2013). Examples of coding systems are illustrated in Fig. 1. We highlight below some key concepts to understanding reaction rule encoding in a retrosynthesis context.

2.1.1. 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 (Altman et al., 2013; Chang et al., 2015) 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 (Nam et al., 2012). 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 (Henry et al., 2010; Jeffryes et al., 2015; Hadadi et al., 2016a) and SimPheny (Yim et al., 2011) use a collection of reaction rules that, as depicted in Fig. 1, can be applied to any ketones (including oxaloacetate) since their encoding is focused on the reaction centre.

2.1.2. Identification of the reaction centre

The simplest way of controlling a degree of abstraction for reaction substrates is to encode the reactions around its centre. This requires compiling the list of atoms that belong to the reaction centre, i.e. atoms that change their configuration when the reaction is applied (panel B in Fig. 2). Atoms changing configuration are those attached to bonds that are broken, formed, or are changing order, as well atoms for which charge and stereochemistry is changing when the reaction is taking place.

Reaction rules used in retrosynthesis generally require a solved Atom-Atom Mapping (AAM, see panel A in Fig. 2) between the atoms of the substrates and those of the products to identify the reaction centre of the reaction (Hou et al., 2003; Hatzimanikatis et al., 2005; Oh et al., 2007; Cho et al., 2010; Liu et al., 2014). The AAM problem is equivalent to the Maximum Common Substructure, or the subgraph isomorphism problem which turns out to be NP-hard (Chen et al., 2013). 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 (Carbonell et al., 2014a) (see Fig. 1).

Importantly, if encoding the reacting centre is necessary, it may not be sufficient to properly define a reaction catalysed by an enzyme since other atoms far from the reacting centre could be involved in the ligand binding as well. To palliate this problem, the definition of the reacting centre is extended to neighbour atoms, either systematically at a predefined bond-distance (diameter, panel C and D Fig. 2) or based on expert-knowledge.

2.1.3. Systematic rule generation

Reaction rules can be computed in principle by processing the set of reactions stored in metabolic databases. However there are some

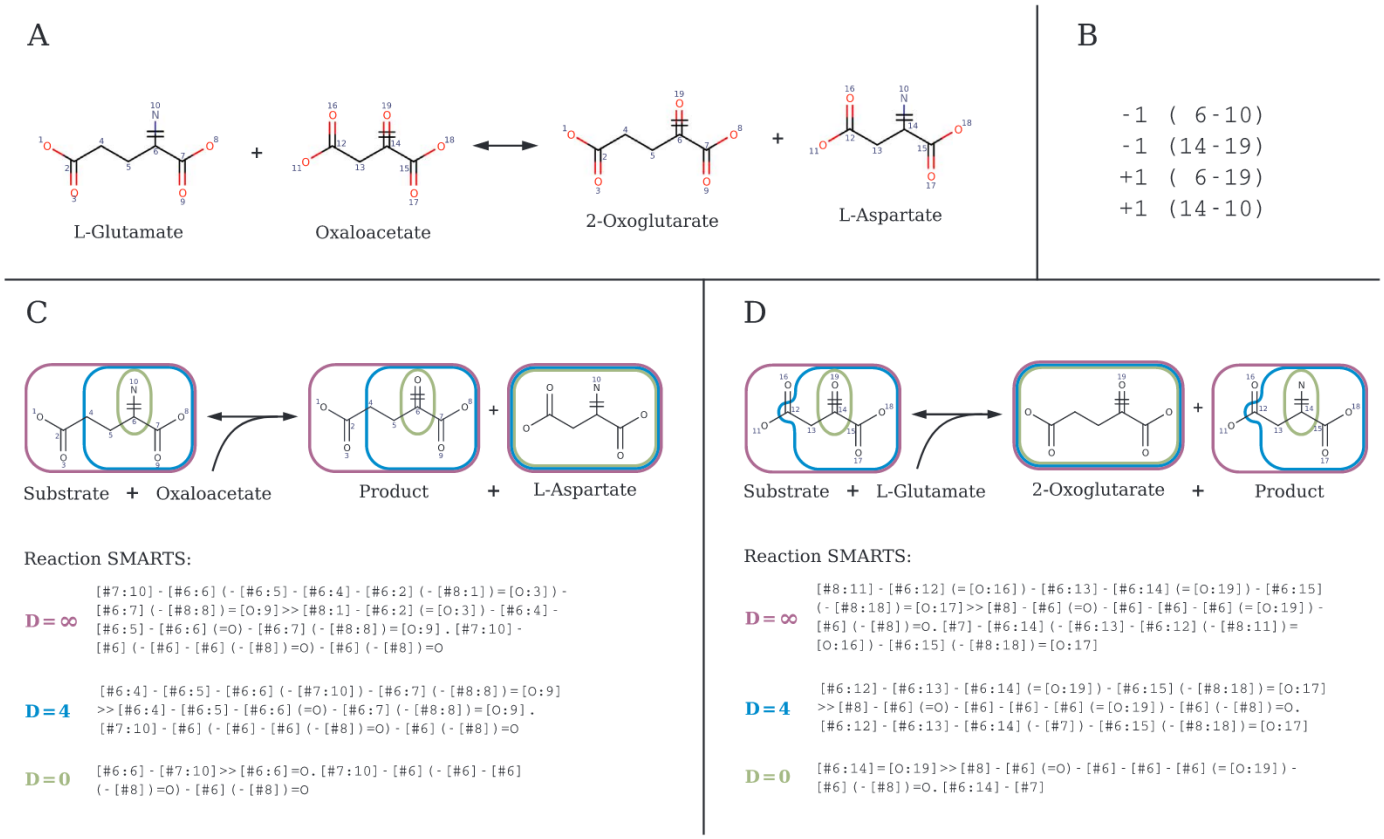


Fig. 2. RetroPath2.0 rules and corresponding SMARTS for reaction 2.6.1.1 at various diameters. A. Full reaction 2.6.1.1 with atom mapping. B. The list of broken bonds (−1) and bonds formed (+1) is given by their atom numbers. C. The corresponding SMARTS for the component modelling promiscuity on L-glutamate: 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 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) (Yim et al., 2011)	Computed from 3rd EC level followed by manual curation	All metabolic reactions	50	Fixed	No	No	Controlled by network and molecules size	No
BNICE (Hadadi et al., 2016a)	Automated from KEGG followed by manual curation	All metabolic reactions	722	Fixed	Yes	No	Controlled by network size	Web server (http://lcsb-databases.epfl.ch/atlas/)
PathPred (Moriya et al., 2010)	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 (http://www.genome.jp/tools/pathpred/)
GEM-Path (Campodonico et al., 2014)	Computed from 3rd EC level	All metabolic reactions	443	Fixed	Yes	No	Controlled by similarity and thermodynamics	No
METEOR (Marchant et al., 2008)	Knowledge-based expert system (Lhasa Ldt.)	All metabolic reactions?	357	Fixed	No?	No	Controlled by "reasoning" rules	Commercial (https://www.lhasalimited.org/products/meteor-nexus.htm)
Reverse Pathway Engineering (THERESA) (Liu et al., 2014)	Automated from the BioPath database (Molecular Networks GmbH)	All metabolic reactions	3516 reference reactions	Fixed	Yes	Comparative genomics	Manual compound selection	No
RetroPath (Carbonell et al., 2014b)	Automated computation on MetaCyc	All metabolic reactions	between 3000 (d = 4) and 5000 (d = 14)	Variable, controlled by diameter	Yes	Post-process using machine learning	Controlled by diameter	Web server (http://xtms.issb.genopole.fr/)
RetroPath2.0 (this study)	Automated computation on MetaNetX	All metabolic reactions	between 6900 (d = 2) and 19,000 (d = 16)	Variable, controlled by diameter	Yes	Embedded using sequence clustering	Controlled by diameter and enzyme score	Open source (https://www.myexperiment.org/workflows/4987.html)

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 (International Union of Biochemistry and Molecular Biology Nomenclature Committee and Webb, E.C, 1992). Both SimPheny and BNICE use the third EC number level to guide their reaction encoding effort. SimPheny (Yim et al., 2011) has 50 manually curated reaction rules, and the number of rules of BNICE systems are of the same order; 86 for (Henry et al., 2010), 198 for (Jeffryes et al., 2015), 722 for (Hadadi et al., 2016a). 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 (Supplementary Fig. S1). 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 (International Union of Biochemistry and Molecular Biology Nomenclature Committee and Webb, E.C, 1992).

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 (Carbonell et al., 2011b, 2014b), and others (Chen et al., 2013; Rahman et al., 2014; Sivakumar et al., 2016). 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 Supplementary Note 1 for a detailed discussion on diameter selection and promiscuity) is generally a good trade-off to cover known reactions' specificity with a reasonable amount of promiscuity predictions (see Section 4.1.3 for an evaluation of rules performance for promiscuity classification and (Carbonell et al., 2011a; Faulon et al., 2008)). Using the procedure outlined in the caption of Fig. 2, when applied to the MetaNetX database (Moretti et al., 2016) the number of rules returned is between 6900 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 Fig. 1D (promiscuity on oxaloacetate, MNXM42).

2.1.4. 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 (Fig. 2).

For these purposes, cosubstrates and coproducts that are currency cofactors (such as water, CO_2, ATP, NADP, etc.) can be ignored from the rules under the 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 Supplementary Note 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 → 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 → C + D" and "B → C + D") as shown in Fig. 2. Others embrace combinatorial complexity (Fig. 1B and D) or simply ignore all cosubstrates (Fig. 1C).

2.2. Building (retrosynthesis) reaction network

In all algorithms listed in Table 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 (Henry et al., 2010) the rule for that reversed reaction is R1C(=O)C(R2)=C(R3)R4 + O-R5 → R1C(=O)C(R2)-C(R3)(R4)OR5, where R5 can be C, H, O, and S and all other Rs 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 1883 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 1000 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 (Faulon et al., 2008) or Gaussian Processes (Mellor et al., 2016). GEM-Path (Campodonico et al., 2014) proposes another strategy where for each reaction the substrates are accepted if they are similar enough to the substrates of the reference reactions.

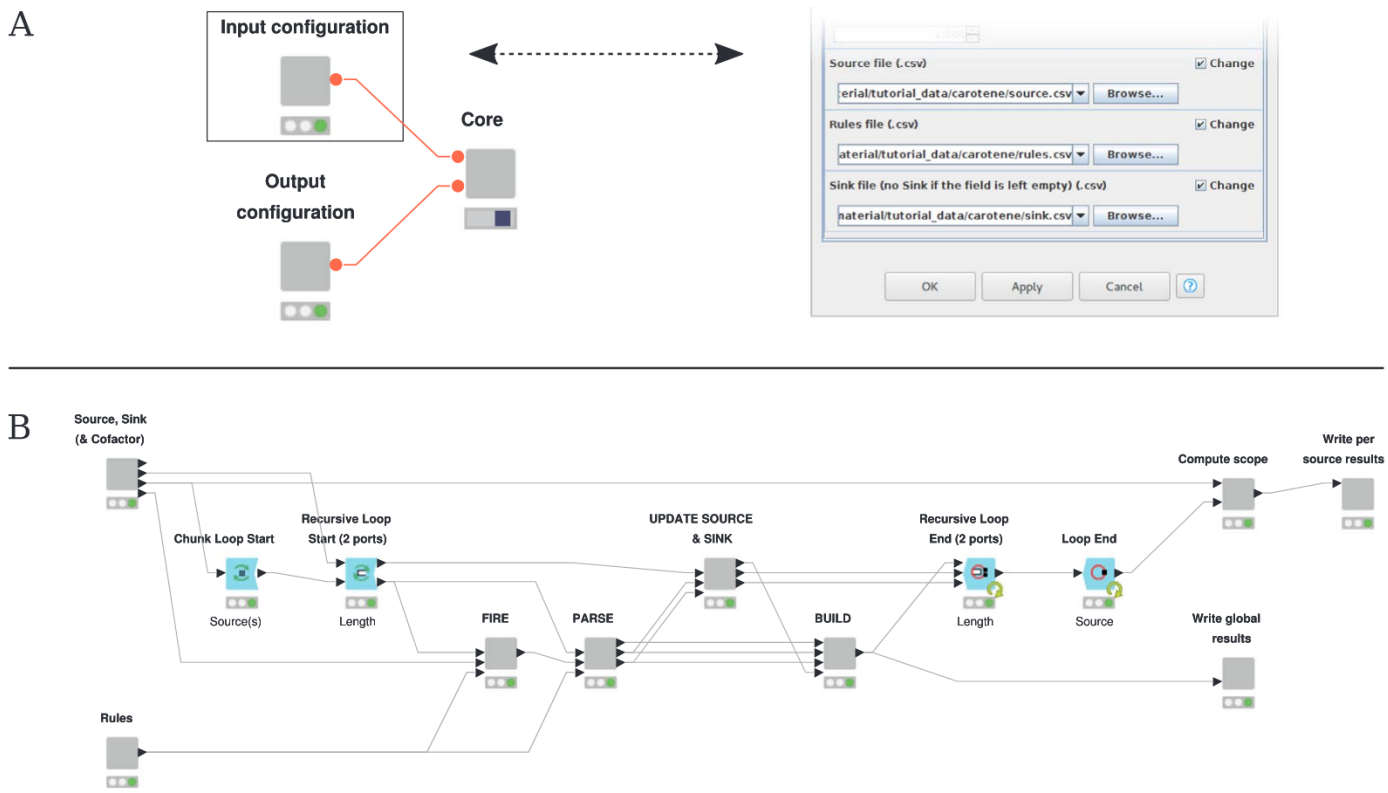


Fig. 3. RetroPath2.0 KNIME workflow. A. Main panel view (left) and input configuration window (right) that allow the user to set up parameters. B. Inner view of the "Core" node where the computation takes place. The "Source, Sink…" and "Rules" nodes parse the source, sink and rules input files provided 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.

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

The workflow proposed in this section can a priori be used to all systems presented in Table 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 (Carbonell et al., 2014a), RDKit routines (Landrum, 2016) and KNIME nodes (Berthold et al., 2008). They have been implemented in the form of a KNIME workflow -called RetroPath2.0- (Fig. 3) that we provide in the Supplementary materials, in addition of sets of rules, examples, and useful data files. Updates will be published on myExperiment.org (https://www.myexperiment.org/workflows/4987.html).

3.1. Reaction rules

RetroPath2.0 uses reaction SMARTS to encode reactions. It is a SMIRK-like reaction rules (Daylight, 2017) format defined by RDKit and is mostly compatible with other tools (see Fig. 1).

3.1.1. Collected reaction rules

Rules for SimPheny and BNICE were extracted from Yim et al. (2011) and Henry et al., (2010) 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 Fig. 2.

3.1.2. Generated reaction rules

We used MetaNetX version 2.0 (Moretti et al., 2016) 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 (Rahman et al., 2016) on all MetaNetX reactions (Fig. 2A). 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 Fig. 2). 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 Fig. 2) 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 (Orth et al., 2011).

3.2. 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 (Carbonell et al., 2011b). 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 (Landrum, 2016). 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.

3.3. 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 5388 organisms were associated to 7793 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 7898 orphan rules, i.e. rules that were generated from reactions lacking sequence annotation (Supplementary Table 1). Similarly, there were 6280 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 1719, 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 (Li and Godzik, 2006). 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:

$$score(rule) = log_{10}(n_{rule}) \quad (1)$$

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.

3.4. 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 (Carbonell et al., 2014a). 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 (i) verified the stoichiometric constraints of the network and (ii) is able to carry non zero-fluxes at the system's steady-state (Schuster et al., 2000). 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.

4. 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 (Supplementary Note 3).

4.1. Rules validation

The quality of the output of the workflow depends largely on feeding it with the proper set of reaction rules. Some authors (Henry et al., 2010; Yim et al., 2011) 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 (Orth et al., 2011) 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.

4.1.1. 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 mono-substrate rules according to their origin: SimPheny (Yim et al., 2011), BNICE (Henry et al., 2010) 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 6000 substrates and 7000 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). 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.

4.1.2. 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 (Carbonell et al., 2014a) 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:

$$T^d(R_1, R_2) = \frac{|\sigma^d(R_1) \cap \sigma^d(R_2)|}{|\sigma^d(R_1) \cup \sigma^d(R_2)|} \tag{2}$$

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 (Moriya et al., 2010; Campodonico et al., 2014; Rahman et al., 2014). 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 extend 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:

$$S(R_1, R_2) = \frac{\sum_{k=1}^{n+1} J^{k-1}(R_1, R_2)k^a}{\sum_{k=1}^{n+1} k^a} \tag{3}$$

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 $d = 8$ (Supplementary Fig. S4). Such performance values are slightly higher than the ones obtained by EC-BLAST (Rahman et al., 2014) by using fingerprint-based similarities, showing the ability of the rules as reaction classifiers.

4.1.3. 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 (Khersonsky and Tawfik, 2010; Nam et al., 2012; Orth and Palsson, 2012; Guzmán et al., 2015). 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. 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 (Nam et al., 2012). For each gene, we took their associated reactions in the EcoCyc database (Keseler et al., 2013). 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 Supplementary Fig. S5, 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 (Guzmán et al., 2015). 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 (Yang and Metcalf, 2004) in agreement with the high score.

4.2. 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 Supplementary Note 3).

4.2.1. Coverage of bioproduction pathways

The Learning Assisted Strain EngineeRing (LASER) database is a repository for metabolic engineering strain designs (Winkler et al., 2015). It stores more than 600 successful metabolic engineering designs (Winkler et al., 2016) 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 33–38), a period not covered by LASER. These two datasets contained 203 distinct compounds in total 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 (Orth et al., 2011) 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 (Isikgor and Becer, 2015). LASER references one pathway for the bioproduction of styrene from phenylalanine with heterologous enzymes in *E. coli* (McKenna and Nielsen, 2011; McKenna et al., 2015) and in *S.cerevisiae* (McKenna et al., 2014). RetroPath2.0 found this pathway (Fig. 4, in red) along with five alternative one from *E. coli* endogenous compounds: 3-phenylpropionic acid, phenylacetaldehyde, and phenylpyruvic acid (Fig. 4, 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 (Sheehan, 2000). The p-xylene can eventually come from lignocellulosic biomass, making the TPA a bio-based compound in such cases (Isikgor and Becer, 2015). Interestingly, two enzymatic bioproduction pathways have been reported for TPA, and they follow the same chemical transformations as the ones from synthetic chemistry (Sheehan, 2000); one from p-xylene (Bramucci et al., 2001) in *Burkholderia* genus, and another from p-toluic acid in *Comamonas testosterone* (Wang et al., 2006). 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. 5, 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 in Supplementary Fig. S3. 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.

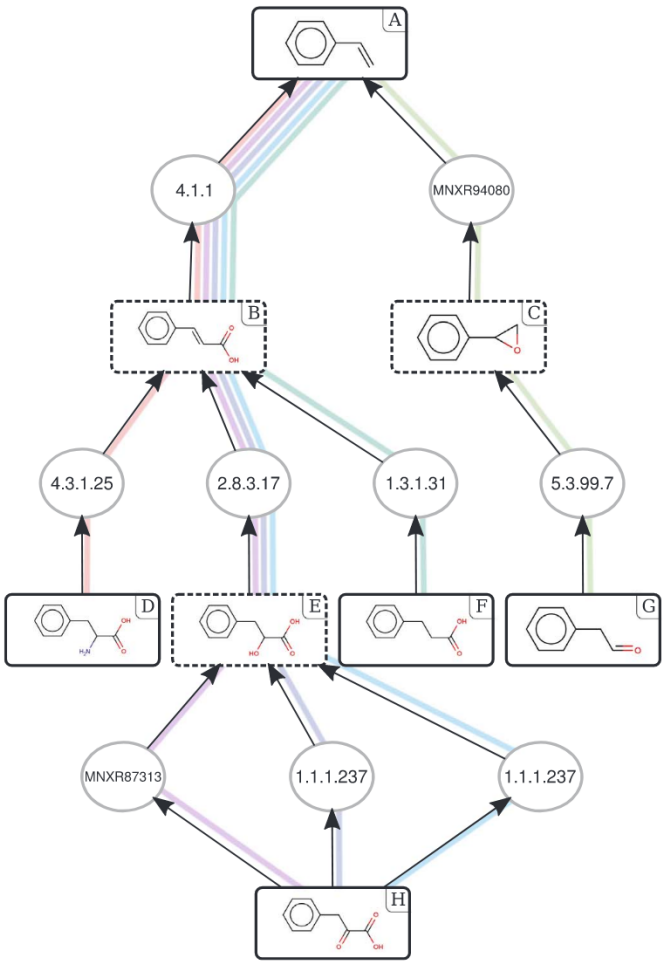


Fig. 4. Enumerated pathways for the production of styrene. Each pathway is depicted by a distinct colour. Pathway referenced in McKenna and Nielsen (2011) is in red (D-B-A). Compounds are represented by their structures, and reactions by their EC numbers. Styrene and sink compounds are surrounded by 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5. 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 Fig. 3). The workflow produces a network linking the source set to the sink set, where each link in the network correspond to a reaction rule. The RetroPath2.0 workflow runs under the KNIME analytics platform and is available in Supplementary material and at myExperiment.org

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. 4.2.1). 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 (see Supplementary Note 3 for the detection of biomarkers). Finally if 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 (Daylight, 2017) that was already used successfully in such context (Hadadi and Hatzimanikatis, 2015). 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 evidences 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 (Chang et al., 2015).

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 Thodey et al. (2014) 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

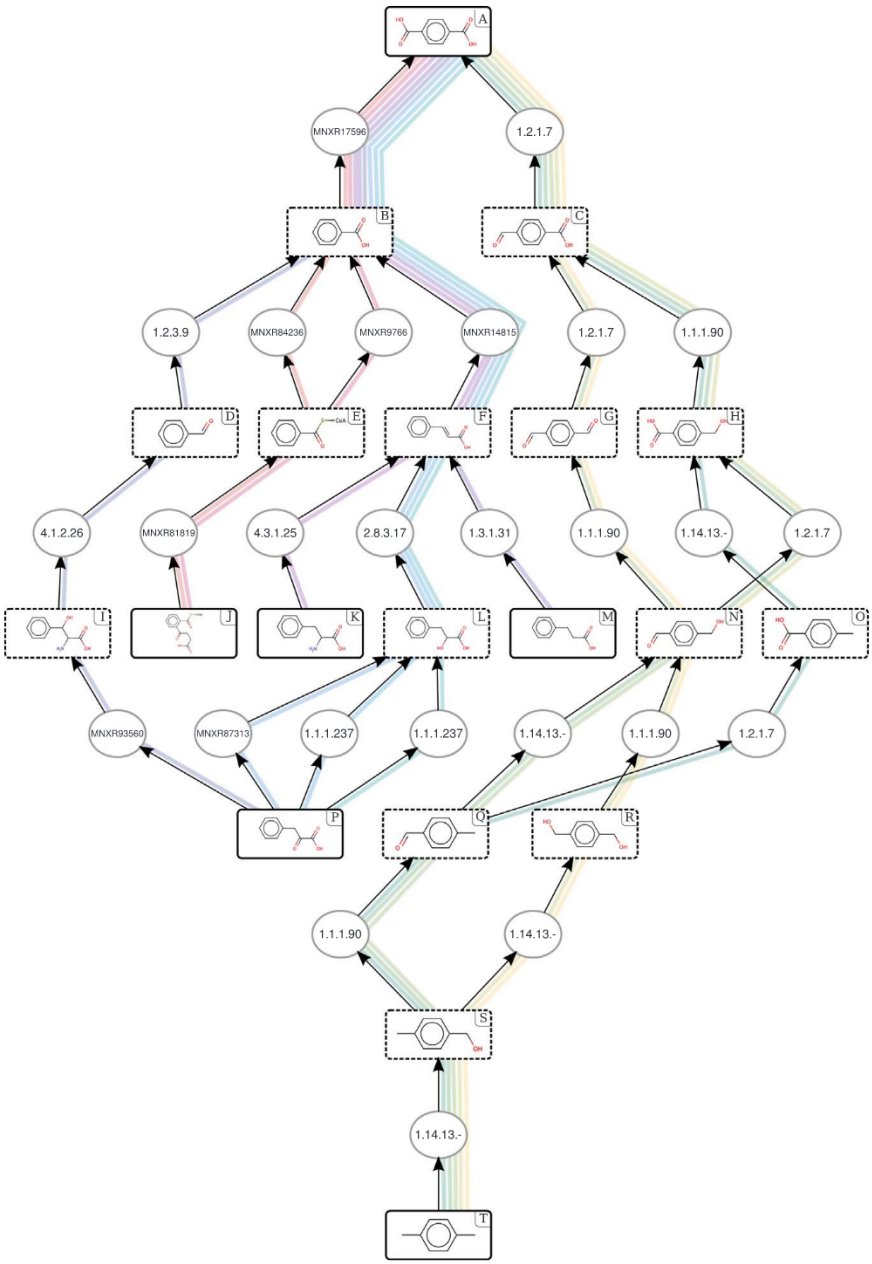


Fig. 5. Enumerated pathways for the production of the non-natural compound terephthalic acid (TPA, compound A) from *E. coli*. Each pathway is depicted by a distinct colour. Pathway referenced in Bramucci et al., (2001) is in teal blue (T-S-Q-O-H-C-A). Compounds are represented by their structures, reactions 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: terephthaldehyde; H: p-hydroxymethyl benzoic acid; I: 3-phenylserine; J: 2-succinylbenzoyl-CoA; K: phenylalanin; 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

Funding

This work was supported by the French National Research Agency

(ANR-15-CE1-0008), the Biotechnology and Biological Sciences Research Council, Centre for synthetic biology of fine and speciality chemicals (BB/M017702/1); Synthetic Biology Applications for Protective Materials (EP/N025504/1).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ymben.2017.12.002.

References

Altman, T., Travers, M., Kothari, A., Caspi, R., Karp, P.D., 2013. A systematic comparison of the MetaCyc and KEGG pathway databases. BMC Bioinform. 14, 112. http://dx.doi.org/10.1186/1471-2105-14-112.

Arita, M., 2003. In silico atomic tracing by substrate-product relationships in *Escherichia coli* intermediary metabolism. Genome Res. 13, 2455–2466. http://dx.doi.org/10.1101/gr.1212003.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2008. KNIME: The Konstanz Information Miner. In: Data Analysis, Machine Learning and Applications. Springer, Berlin, Heidelberg, pp. 319–326.

Bramucci, M.G., McCutchen, C.M., Nagarajan, V., Thomas, S.M., 2001. Microbial production of terephthalic acid and isophthalic acid. US6187569 B1.

Campodonico, M.A., Andrews, B.A., Asenjo, J.A., Palsson, B.O., Feist, A.M., 2014. Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path. Metab. Eng. 25, 140–158. http://dx.doi.org/10.1016/j.ymben.2014.07.009.

Carbonell, P., Lecointre, G., Faulon, J.-L., 2011a. Origins of Specificity and Promiscuity in Metabolic Networks. J. Biol. Chem. 286, 43994–44004. http://dx.doi.org/10.1074/jbc.M111.274050.

Carbonell, P., Planson, A.-G., Fichera, D., Faulon, J.-L., 2011b. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst. Biol. 5, 122. http://dx.doi.org/10.1186/1752-0509-5-122.

Carbonell, P., Fichera, D., Pandit, S.B., Faulon, J.-L., 2012. Enumerating metabolic pathways for the production of heterologous target chemicals in chassis organisms. BMC Syst. Biol. 6, 10. http://dx.doi.org/10.1186/1752-0509-6-10.

Carbonell, P., Carlsson, L., Faulon, J.-L., 2013. Stereo signature molecular descriptor. J. Chem. Inf. Model. 53, 887–897. http://dx.doi.org/10.1021/ci300584r.

Carbonell, P., Parutto, P., Baudier, C., Junot, C., Faulon, J.-L., 2014a. Retropath: automated pipeline for embedded metabolic circuits. ACS Synth. Biol. 3, 565–577. http://dx.doi.org/10.1021/sb4001273.

Carbonell, P., Parutto, P., Herisson, J., Pandit, S.B., Faulon, J.-L., 2014b. XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res. 42, W389–W394. http://dx.doi.org/10.1093/nar/gku362.

Chang, A., Schomburg, I., Placzek, S., Jeske, L., Ulbrich, M., Xiao, M., Sensen, C.W., Schomburg, D., 2015. BRENDA in 2015: exciting developments in its 25th year of existence. Nucleic Acids Res. 43, D439–D446. http://dx.doi.org/10.1093/nar/gku1068.

Chen, W.L., Chen, D.Z., Taylor, K.T., 2013. Automatic reaction mapping and reaction center detection. Wiley Interdiscip. Rev. Comput. Mol. Sci. 3, 560–593. http://dx.doi.org/10.1002/wcms.1140.

Cho, A., Yun, H., Park, J.H., Lee, S.Y., Park, S., 2010. Prediction of novel synthetic pathways for the production of desired chemicals. BMC Syst. Biol. 4, 35. http://dx.doi.org/10.1186/1752-0509-4-35.

Copeland, W.B., Bartley, B.A., Chandran, D., Galdzicki, M., Kim, K.H., Sleight, S.C., Maranas, C.D., Sauro, H.M., 2012. Computational tools for metabolic engineering. Metab. Eng. 14, 270–280. http://dx.doi.org/10.1016/j.ymben.2012.03.001.

Daylight, 2017. Daylight Theory Manual [WWW Document]. URL 〈http://www.daylight.com/dayhtml/doc/theory/〉 (Accessed 14 April 2017).

Delépine, B., Libis, V., Carbonell, P., Faulon, J.-L., 2016. SensiPath: computer-aided design of sensing-enabling metabolic pathways. Nucleic Acids Res. 44, W226–W231. http://dx.doi.org/10.1093/nar/gkw305.

Dugundji, J., Ugi, I., 1973. An algebraic model of constitutional chemistry as a basis for chemical computer programs. In: Computers in Chemistry. Springer, Berlin, Heidelberg, pp. 19–64.

Faulon, J.-L., Misra, M., Martin, S., Sale, K., Sapra, R., 2008. Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. Bioinformatics 24, 225–233. http://dx.doi.org/10.1093/bioinformatics/btm580.

Finley, S.D., Broadbelt, L.J., Hatzimanikatis, V., 2009. Computational framework for predictive biodegradation. Biotechnol. Bioeng. 104, 1086–1097. http://dx.doi.org/10.1002/bit.22489.

Guzmán, G.I., Utrilla, J., Nurk, S., Brunk, E., Monk, J.M., Ebrahim, A., Palsson, B.O., Feist, A.M., 2015. Model-driven discovery of underground metabolic functions in *Escherichia coli*. Proc. Natl. Acad. Sci. USA 112, 929–934. http://dx.doi.org/10.1073/pnas.1414218112.

Hadadi, N., Hatzimanikatis, V., 2015. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. Curr. Opin. Chem. Biol. 28, 99–104. http://dx.doi.org/10.1016/j.cbpa.2015.06.025.

Hadadi, N., Hafner, J., Shajkofci, A., Zisaki, A., Hatzimanikatis, V., 2016a. ATLAS of Biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. ACS Synth. Biol. 5, 1155–1166. http://dx.doi.org/10.1021/acssynbio.6b00054.

Hadadi, N., Hafner, J., Soh, K.C., Hatzimanikatis, V., 2016b. Reconstruction of biological pathways and metabolic networks from in silico labeled metabolites. Biotechnol. J. http://dx.doi.org/10.1002/biot.201600464.

Hatzimanikatis, V., Li, C., Ionita, J.A., Henry, C.S., Jankowski, M.D., Broadbelt, L.J., 2005. Exploring the diversity of complex metabolic networks. Bioinformatics 21, 1603–1609. http://dx.doi.org/10.1093/bioinformatics/bti213.

Haug, K., Salek, R.M., Steinbeck, C., 2017. Global open data management in metabolomics. Curr. Opin. Chem. Biol. 36, 58–63. http://dx.doi.org/10.1016/j.cbpa.2016.12.024.

Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2010. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. Biotechnol. Bioeng. 106, 462–473. http://dx.doi.org/10.1002/bit.22673.

Hou, B.K., Wackett, L.P., Ellis, L.B.M., 2003. Microbial pathway prediction: a functional group approach. J. Chem. Inf. Comput. Sci. 43, 1051–1057. http://dx.doi.org/10.1021/ci034018f.

Hou, B.K., Ellis, L.B.M., Wackett, L.P., 2004. Encoding microbial metabolic logic: predicting biodegradation. J. Ind. Microbiol. Biotechnol. 31, 261–272. http://dx.doi.org/10.1007/s10295-004-0144-7.

International Union of Biochemistry and Molecular Biology Nomenclature Committee, Webb, E.C, 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes. International Union of Biochemistry and Molecular Biology by Academic Press. ed. San Diego.

Isikgor, F.H., Becer, C.R., 2015. Lignocellulosic biomass: a sustainable platform for the production of bio-based chemicals and polymers. Polym. Chem. 6, 4497–4559. http://dx.doi.org/10.1039/C5PY00263J.

Jeffryes, J.G., Colastani, R.L., Elbadawi-Sidhu, M., Kind, T., Niehaus, T.D., Broadbelt, L.J., Hanson, A.D., Fiehn, O., Tyo, K.E.J., Henry, C.S., 2015. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J. Cheminform. 7, 44. http://dx.doi.org/10.1186/s13321-015-0087-1.

Kayala, M.A., Azencott, C.-A., Chen, J.H., Baldi, P., 2011. Learning to predict chemical reactions. J. Chem. Inf. Model. 51, 2209–2222. http://dx.doi.org/10.1021/ci200207y.

Keasling, 2014. Hearing on Policies to Spur Innovative Medical Breakthroughs from Laboratories to Patients.

Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A.G., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I., Karp, P.D., 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 41, D605–D612. http://dx.doi.org/10.1093/nar/gks1027.

Khersonsky, O., Tawfik, D.S., 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu. Rev. Biochem. 79, 471–505. http://dx.doi.org/10.1146/annurev-biochem-030409-143718.

Landrum, 2016. RDKit: Open-source Cheminformatics [WWW Document]. URL 〈http://www.rdkit.org/〉 (Accessed 2 August 2016).

Latino, D.A.R.S., Aires-de-Sousa, J., 2011. Classification of chemical reactions and chemoinformatic processing of enzymatic transformations. Methods Mol. Biol. 672, 325–340. http://dx.doi.org/10.1007/978-1-60761-839-3_13.

Lee, S.Y., Kim, H.U., 2015. Systems strategies for developing industrial microbial strains. Nat. Biotechnol. 33, 1061–1072. http://dx.doi.org/10.1038/nbt.3365.

Li, W., Godzik, A., 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658–1659. http://dx.doi.org/10.1093/bioinformatics/btl158.

Liu, M., Bienfait, B., Sacher, O., Gasteiger, J., Siezen, R.J., Nauta, A., Geurts, J.M.W., 2014. Combining chemoinformatics with bioinformatics: in silico prediction of bacterial flavor-forming pathways by a chemical systems biology approach "reverse pathway engineering". PLoS One 9, e84769. http://dx.doi.org/10.1371/journal.pone.0084769.

Marchant, C.A., Briggs, K.A., Long, A., 2008. In silico tools for sharing data and knowledge on toxicity and metabolism: derek for windows, meteor, and vitic. Toxicol. Mech. Methods 18, 177–187. http://dx.doi.org/10.1080/15376510701857320.

McKenna, R., Nielsen, D.R., 2011. Styrene biosynthesis from glucose by engineered *E. coli*. Metab. Eng. 13, 544–554. http://dx.doi.org/10.1016/j.ymben.2011.06.005.

McKenna, R., Thompson, B., Pugh, S., Nielsen, D.R., 2014. Rational and combinatorial approaches to engineering styrene production by *Saccharomyces cerevisiae*. Microb. Cell Factor. 13, 123. http://dx.doi.org/10.1186/s12934-014-0123-2.

McKenna, R., Moya, L., McDaniel, M., Nielsen, D.R., 2015. Comparing in situ removal strategies for improving styrene bioproduction. Bioprocess Biosyst. Eng. 38, 165–174. http://dx.doi.org/10.1007/s00449-014-1255-9.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B.A., Ellison, A.M., King, J.L., 2016. RESEARCH INTEGRITY. Liberating field science samples and data. Science 351, 1024–1026. http://dx.doi.org/10.1126/science.aad7048.

Medema, M.H., van Raaphorst, R., Takano, E., Breitling, R., 2012. Computational tools for the synthetic design of biochemical pathways. Nat. Rev. Microbiol. 10, 191–202. http://dx.doi.org/10.1038/nrmicro2717.

Mellor, J., Grigoras, I., Carbonell, P., Faulon, J.-L., 2016. Semisupervised Gaussian process for automated enzyme search. ACS Synth. Biol. 5, 518–528. http://dx.doi.org/10.1021/acssynbio.5b00294.

Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., Pagni, M., 2016. MetaNetX/MNXref–reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Res. 44, D523–D526. http://dx.doi.org/10.1093/nar/gkv1117.

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S., Kanehisa, M., 2010. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res. 38, W138–W143. http://dx.doi.org/10.1093/nar/gkq318.

Nam, H., Lewis, N.E., Lerman, J.A., Lee, D.-H., Chang, R.L., Kim, D., Palsson, B.O., 2012. Network context and selection in the evolution to enzyme specificity. Science 337, 1101–1104. http://dx.doi.org/10.1126/science.1216861.

Oh, M., Yamada, T., Hattori, M., Goto, S., Kanehisa, M., 2007. Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J. Chem. Inf. Model. 47, 1702–1712. http://dx.doi.org/10.1021/ci700006f.

Orth, J.D., Palsson, B., 2012. Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. BMC Syst. Biol. 6, 30. http://dx.doi.org/10.1186/1752-0509-6-30.

Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., Palsson, B.Ø., 2011. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. Mol. Syst. Biol. 7, 535. http://dx.doi.org/10.1038/msb.2011.65.

Paddon, C.J., Westfall, P.J., Pitera, D.J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M.D., Tai, A., Main, A., Eng, D., Polichuk, D.R., Teoh, K.H., Reed, D.W., Treynor, T., Lenihan, J., Fleck, M., Bajad, S., Dang, G., Dengrove, D., Diola, D., Dorin, G., Ellens, K.W., Fickes, S., Galazzo, J., Gaucher, S.P., Geistlinger, T., Henry, R., Hepp, M., Horning, T., Iqbal, T., Jiang, H., Kizer, L., Lieu, B., Melis, D., Moss, N., Regentin, R., Secrest, S., Tsuruta, H., Vazquez, R., Westblade, L.F., Xu, L., Yu, M., Zhang, Y., Zhao, L., Lievense, J., Covello, P.S., Keasling, J.D., Reiling, K.K., Renninger, N.S., Newman, J.D., 2013. High-level semi-synthetic production of the potent antimalarial artemisinin. Nature 496, 528–532. http://dx.doi.org/10.1038/nature12051.

Planson, A.-G., Carbonell, P., Paillard, E., Pollet, N., Faulon, J.-L., 2012. Compound toxicity screening and structure-activity relationship modeling in *Escherichia coli*. Biotechnol. Bioeng. 109, 846–850. http://dx.doi.org/10.1002/bit.24356.

Rahman, S.A., Cuesta, S.M., Furnham, N., Holliday, G.L., Thornton, J.M., 2014. EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat. Methods 11, 171–174. http://dx.doi.org/10.1038/nmeth.2803.

Rahman, S.A., Torrance, G., Baldacci, L., Martínez Cuesta, S., Fenninger, F., Gopal, N., Choudhary, S., May, J.W., Holliday, G.L., Steinbeck, C., Thornton, J.M., 2016. Reaction Decoder Tool (RDT): extracting features from chemical reactions. Bioinformatics 32, 2065–2066. http://dx.doi.org/10.1093/bioinformatics/btw096.

Rodrigo, G., Carrera, J., Prather, K.J., Jaramillo, A., 2008. DESHARKY: automatic design of metabolic pathways for optimal cell growth. Bioinformatics 24, 2554–2556. http://dx.doi.org/10.1093/bioinformatics/btn471.

Schofield, P.N., Bubela, T., Weaver, T., Portilla, L., Brown, S.D., Hancock, J.M., Einhorn, D., Tocchini-Valentini, G., Hrabe de Angelis, M., Rosenthal, N., CASIMIR Rome Meeting participants, 2009. Post-publication sharing of data and tools. Nature 461, 171–173. http://dx.doi.org/10.1038/461171a.

Schuster, S., Fell, D.A., Dandekar, T., 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. Nat. Biotechnol. 18, 326–332. http://dx.doi.org/10.1038/73786.

Sheehan, R.J., 2000. Terephthalic Acid, Dimethyl Terephthalate, and Isophthalic Acid. In: Ullmann's Encyclopedia of Industrial Chemistry. Wiley-VCH Verlag GmbH & Co. KGaA.

Sivakumar, T.V., Giri, V., Park, J.H., Kim, T.Y., Bhaduri, A., 2016. ReactPRED: a tool to predict and analyze biochemical reactions. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btw491.

Thodey, K., Galanie, S., Smolke, C.D., 2014. A microbial biomanufacturing platform for natural and semisynthetic opioids. Nat. Chem. Biol. 10, 837–844. http://dx.doi.org/10.1038/nchembio.1613.

Wang, J., Tian, J., Xu, J., 2006. A method for producing terephthalic acid by *Comamonas testosteroni* DSM6577. Chin. J. Catal. 27, 297.

Warr, W.A., 2012. Scientific workflow systems: Pipeline Pilot and KNIME. J. Comput. Aided Mol. Des. 26, 801–804. http://dx.doi.org/10.1007/s10822-012-9577-7.

Winkler, J.D., Halweg-Edwards, A.L., Gill, R.T., 2015. The LASER database: Formalizing design rules for metabolic engineering. Metab. Eng. Commun. 2, 30–38. http://dx.doi.org/10.1016/j.meteno.2015.06.003.

Winkler, J.D., Halweg-Edwards, A.L., Gill, R.T., 2016. Quantifying complexity in metabolic engineering using the LASER database. Metab. Eng. Commun. 3, 227–233. http://dx.doi.org/10.1016/j.meteno.2016.07.002.

Yang, K., Metcalf, W.W., 2004. A new activity for an old enzyme: *Escherichia coli* bacterial alkaline phosphatase is a phosphite-dependent hydrogenase. Proc. Natl. Acad. Sci. USA 101, 7919–7924. http://dx.doi.org/10.1073/pnas.0400664101.

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H.B., Andrae, S., Yang, T.H., Lee, S.Y., Burk, M.J., Van Dien, S., 2011. Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. Nat. Chem. Biol. 7, 445–452. http://dx.doi.org/10.1038/nchembio.580.

Supplementary Material

RetroPath2.0: A retrosynthesis workflow for metabolic engineers

Baudoin Delépine, Thomas Duigou, Pablo Carbonell, Jean-Loup Faulon

TABLE OF CONTENT

Supplementary Notes 2
Supplementary Note 1: Reaction rule diameter and promiscuity recovery 2
Supplementary Note 2: Details about multimolecular rules 4
Supplementary Note 3: Detection of biomarkers through metabolic circuits 6
Supplementary Figures 8
Supplementary Figure S1 8
Supplementary Figure S2 9
Supplementary Figure S3 10
Supplementary Figure S4 11
Supplementary Figure S5 12
Supplementary figure S6 13
Supplementary Tables 14
Supplementary Table S1 14
Supplementary Table S2 15
Supplementary Data 16
RetroPath2.0 workflow 16
Tutorial & examples 16
Scope Viewer 16
Set of SMARTS rules 16
Set of compounds from *E. coli* metabolism 16
Coverage of bioproduction pathways 16
Detection of biomarkers through metabolic circuits 17
RP2paths open source code for enumerating pathways 17
References 18

SUPPLEMENTARY NOTES

SUPPLEMENTARY NOTE 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* (Nam et al., 2012) (see Section 4.1.3 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 S6 (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 S6 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 were of 78% and 88%, respectively. These results suggest again that a good choice for recovering promiscuity is a diameter around d =8.

SUPPLEMENTARY NOTE 2: DETAILS ABOUT MULTIMOLECULAR RULES

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, oxoglutaramate, 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 → R1C(NH2)R2) in the way that is done in (Yim et al., 2011) nor they can be coded in the form Glutamate + R1C(=O)R2 → Oxoglutarate + R1C(NH2)R2, (where R1 and R2 can be C or H) as done in (Henry et al., 2010).

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 => 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 applied 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 1, SimPheny (Yim et al., 2011) does not deal with multiple substrates and products as all rules are monomolecular, BNICE in (Henry et al., 2010) 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 handing multiple substrates and multiple products, thus reflecting better the complexity of metabolic reactions. Nonetheless, RetroPath2.0 allows only one substrate at a time to undergo promiscuity modelling so that reaction prediction remains tractable.

SUPPLEMENTARY NOTE 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 (Khalil and Collins, 2010) currently being developed comprise systems capable of sensing a small molecule generally though allosteric interactions with RNA aptamers (e.g. riboswitches) or transcription factors (van der Meer and Belkin, 2010) 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) (Libis et al., 2016; Delépine et al., 2016), 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 (Libis et al., 2016).

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 (Sreekumar et al., 2009; Zang et al., 2014; Li et al., 2016; Lima et al., 2016), urine (Sreekumar et al., 2009; Zhang et al., 2013; Struck-Lewicka et al., 2015; Fernández-Peralbo et al., 2016; Lima et al., 2016), and tissue (Sreekumar et al., 2009; McDunn et al., 2013; Lima et al., 2016; Huan et al., 2016). The above references gave a final list of about 800 small molecule biomarkers. Because we are considering 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 (Delépine et al., 2016)). SEMPs 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 Supplementary Table S2.

Notable amongst the biosensor listed in Supplementary Table S2 are H_2O_2 and glycine that are detectable by the native *E. coli* transcription factors OxyR and GcvR, respectively (Tartaglia et al., 1989; Stauffer and Stauffer, 1994), and benzoate for which biosensors have already been built in *E. coli* (Libis et al., 2016) (detailed results are provided in the supplementary Table S2). Interestingly several biomarkers could be transformed into the same effector, thus enabling the integration of multiple biomarker signals into a unique detectable biosensor.

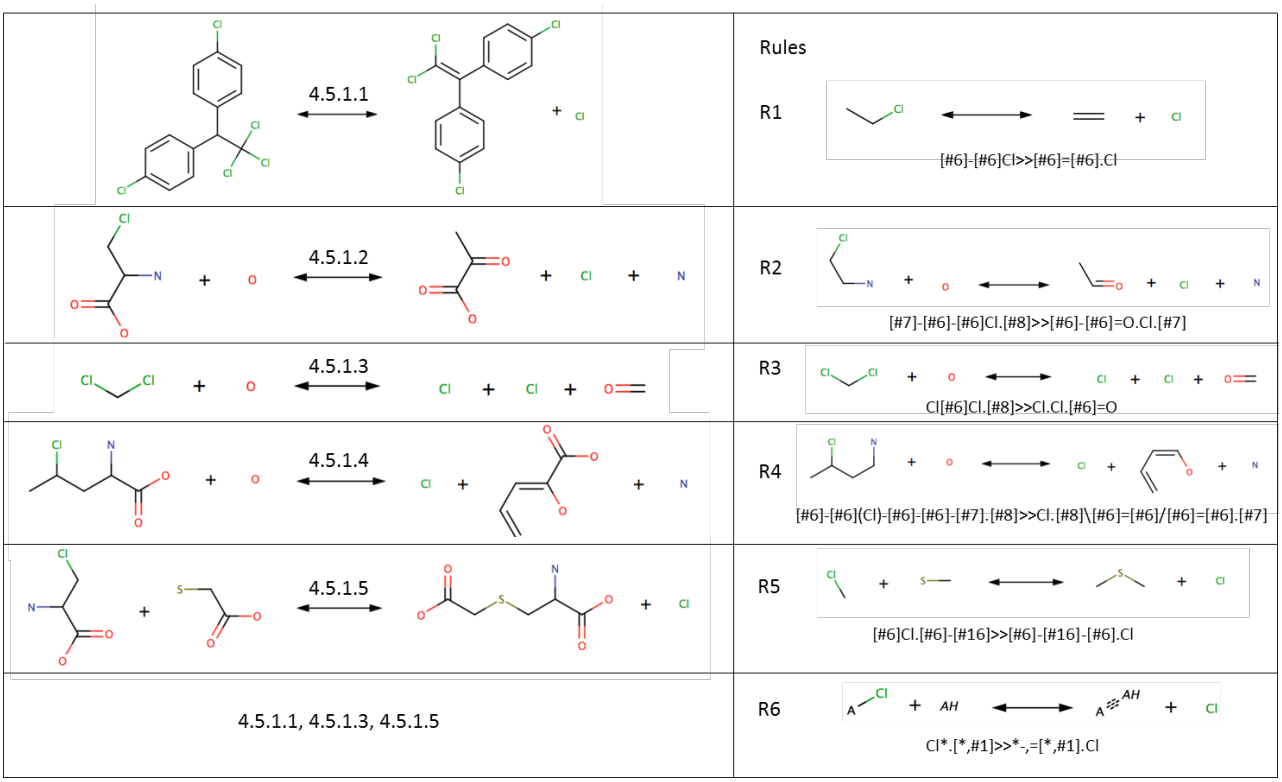
The choice of the enzymes or the transcription factors involved in SEMP biosensors is important when one wishes the biosensor to be specific to a given analyte. Taking the example of N-acetyl-aspartate from supplementary Table S2, some aspartate oxidase (1.4.3.16) transform also alternative substrates, such as L-aspartate, but other will have no activity on aspartate and transform only N-acetyl-aspartate (Tedeschi et al., 2010, p. 121). Another strategy to increase specificity is to engineer multiple SMEP for a given analyte, for instance in supplementary Table S2, hippurate, sarcosine and creatinine can each be detected by two transcription factors.

Examples of SEMP provided in Table S2 mostly involve only one enzyme, however RetroPath can of course produce longer pathways, for examples of such pathways the reader can consult our previous papers (Delépine et al., 2016; Libis et al., 2016).

The results presented in supplementary Table S2 highlight the versatile use that a generic retrosynthesis and reaction network prediction algorithm can have beyond metabolic engineering.

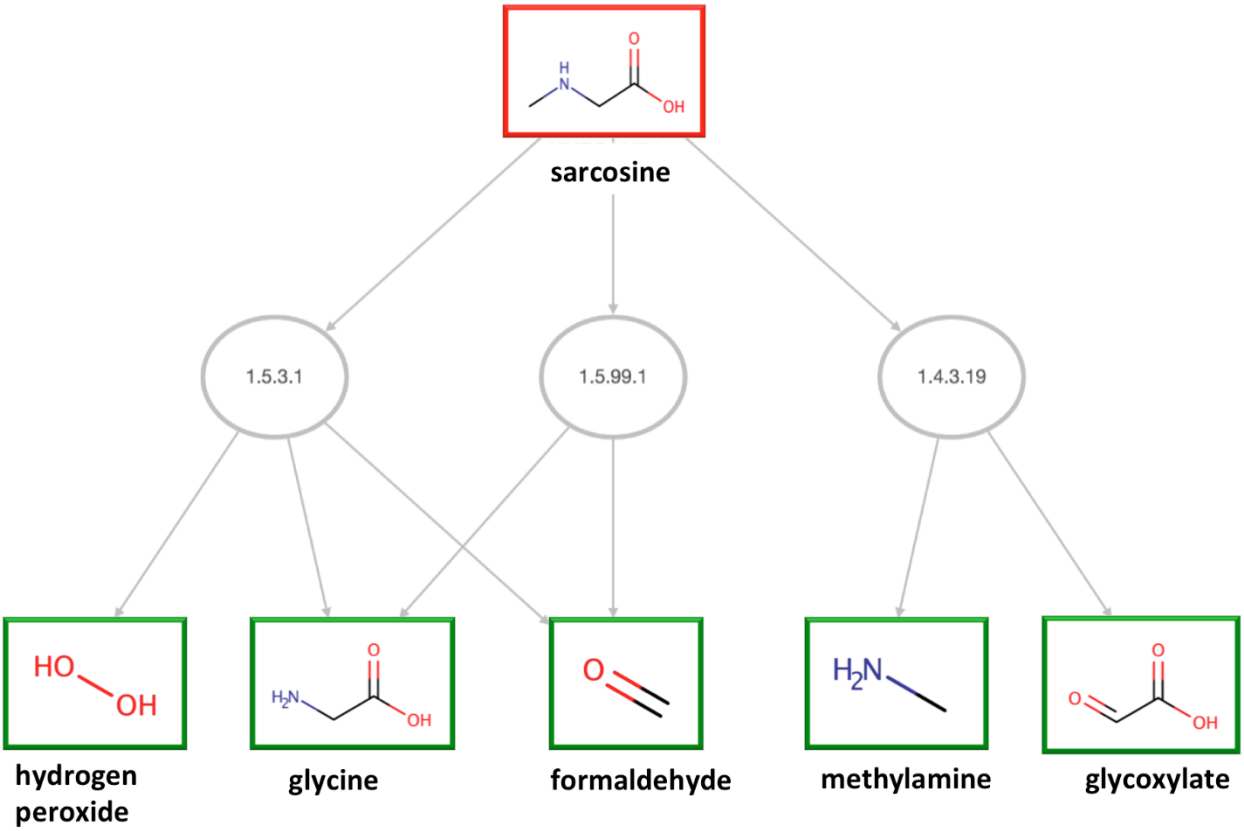
SUPPLEMENTARY FIGURES

SUPPLEMENTARY FIGURE S1



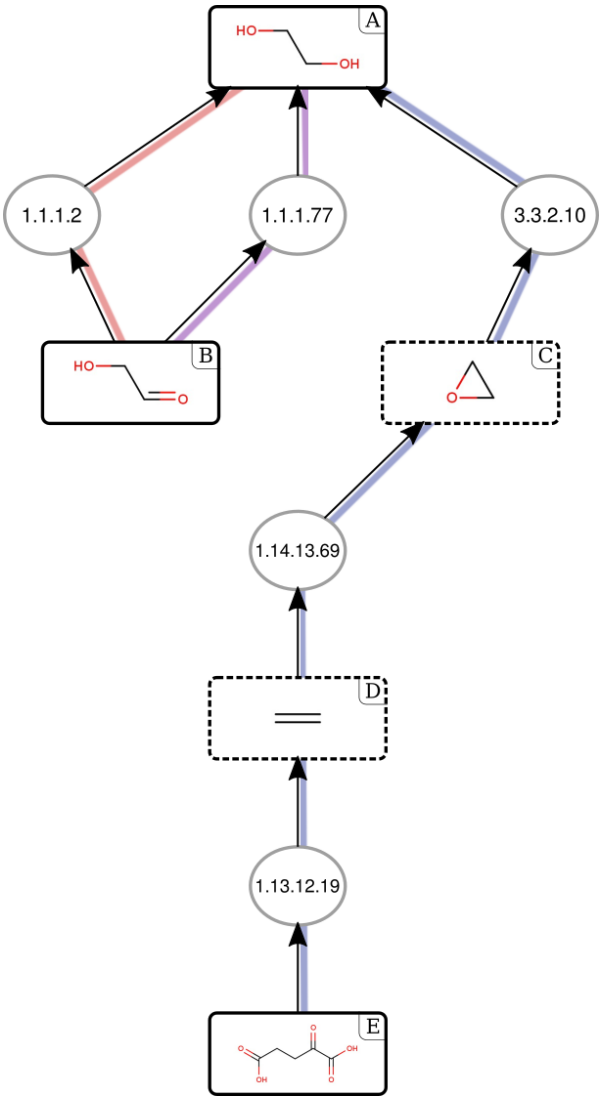
Supplementary Figure S1 – Reactions rules for EC class 4.5.1.-. Hydrogen are omitted for simplicity thus O stands for H_2O, Cl for HCl and N for NH_3. The last rule (R6) does not apply 4.5.1.2 and 4.5.1.4 since in addition to the removal of chlorine these reactions also involved removal of nitrogen.

SUPPLEMENTARY FIGURE S2



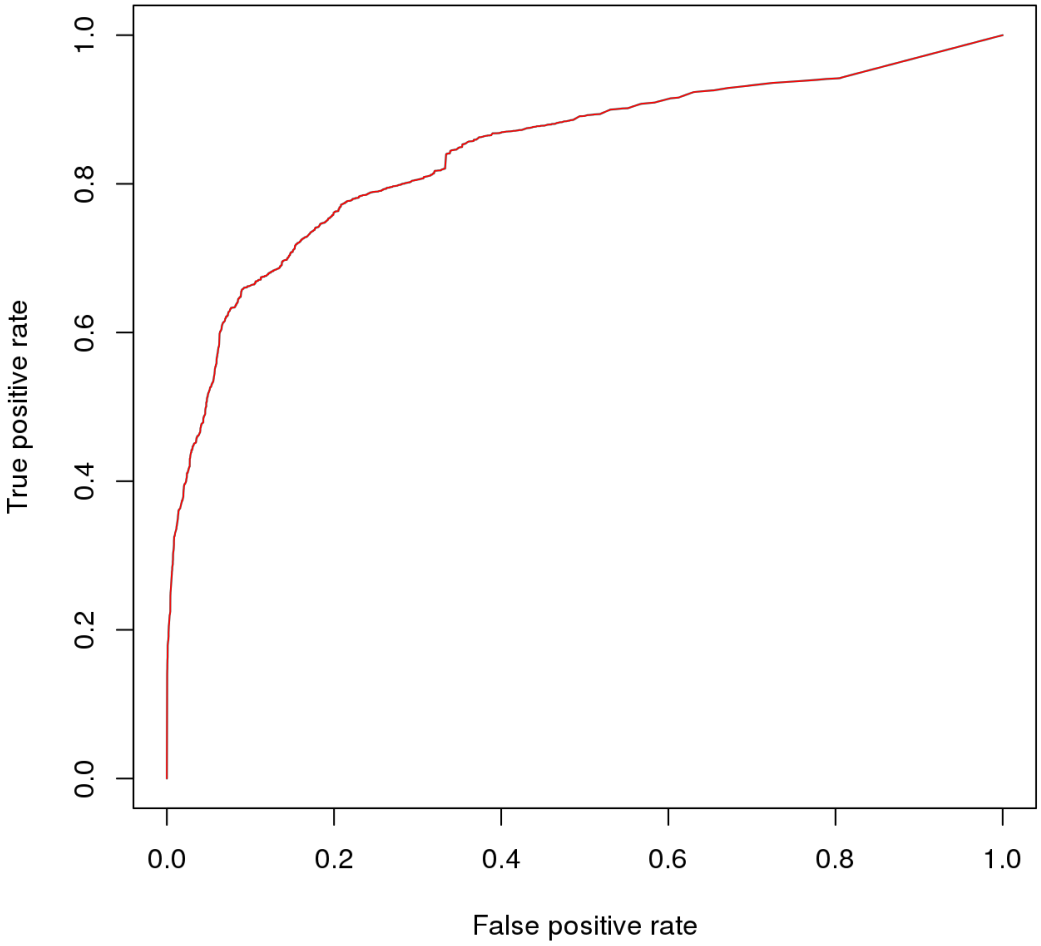
Supplementary Figure S2. Example of a scope obtained for sarcosine. All the products of the pathways are transcription factor effectors (i.e. in the sink file, green squares). Reactions are given by their EC numbers.

SUPPLEMENTARY FIGURE S3



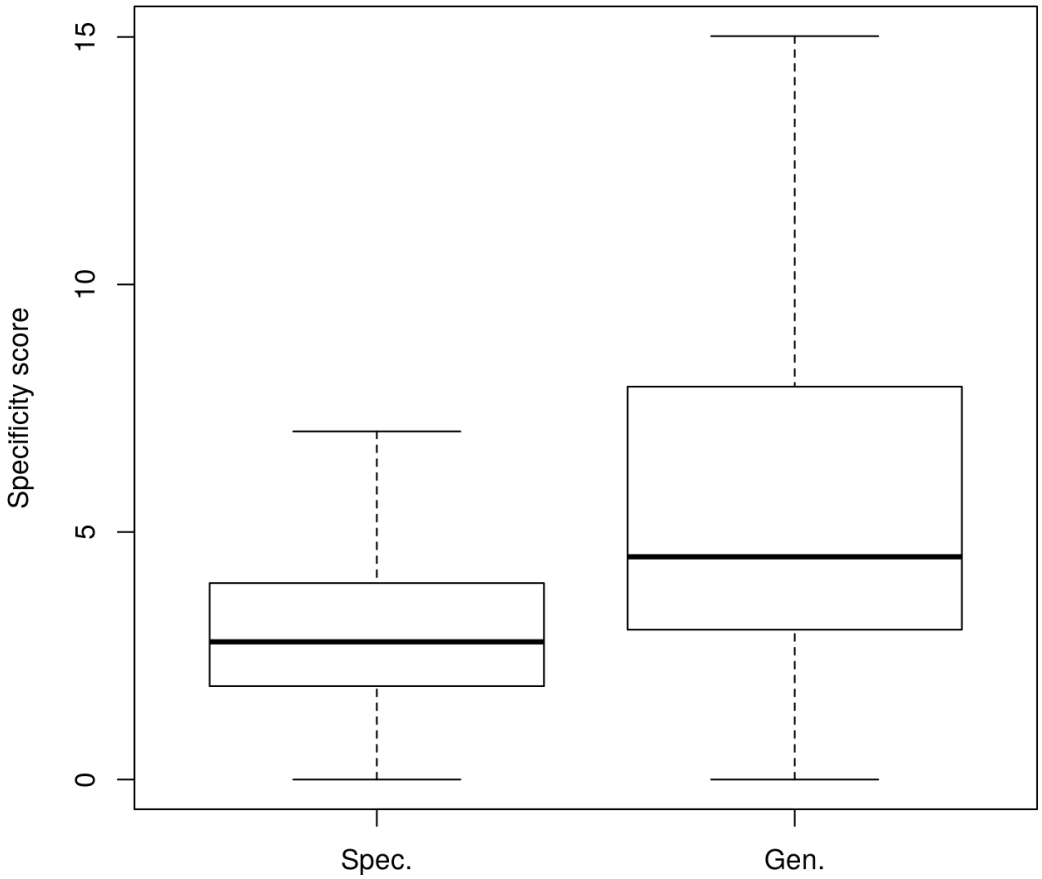
Supplementary Figure S3. Enumerated pathways for the production of ethylene glycol. Each pathway is depicted by a distinct color. The final step of the pathway engineered in (Liu et al., 2013) is in red (B, A). Enzymatic step 1.13.12.19 is also known as RXN-12538 (MetaCyc) and R09784 (Kegg). Compounds 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. Involved compounds: ethylene glycol (A), glycolaldehyde (B), ethylene oxide (C), ethylene (D) and oxoglutarate (E).

SUPPLEMENTARY FIGURE S4



Supplementary Figure S4. Receiving operating characteristic curves (ROC) curves for the rules of RetroPath2.0 of diameter d = 8.

SUPPLEMENTARY FIGURE S5



Supplementary Figure S5. Box plot comparing the distribution of reaction scores for specialist and generalist enzymes in *E. coli*.

SUPPLEMENTARY FIGURE S6

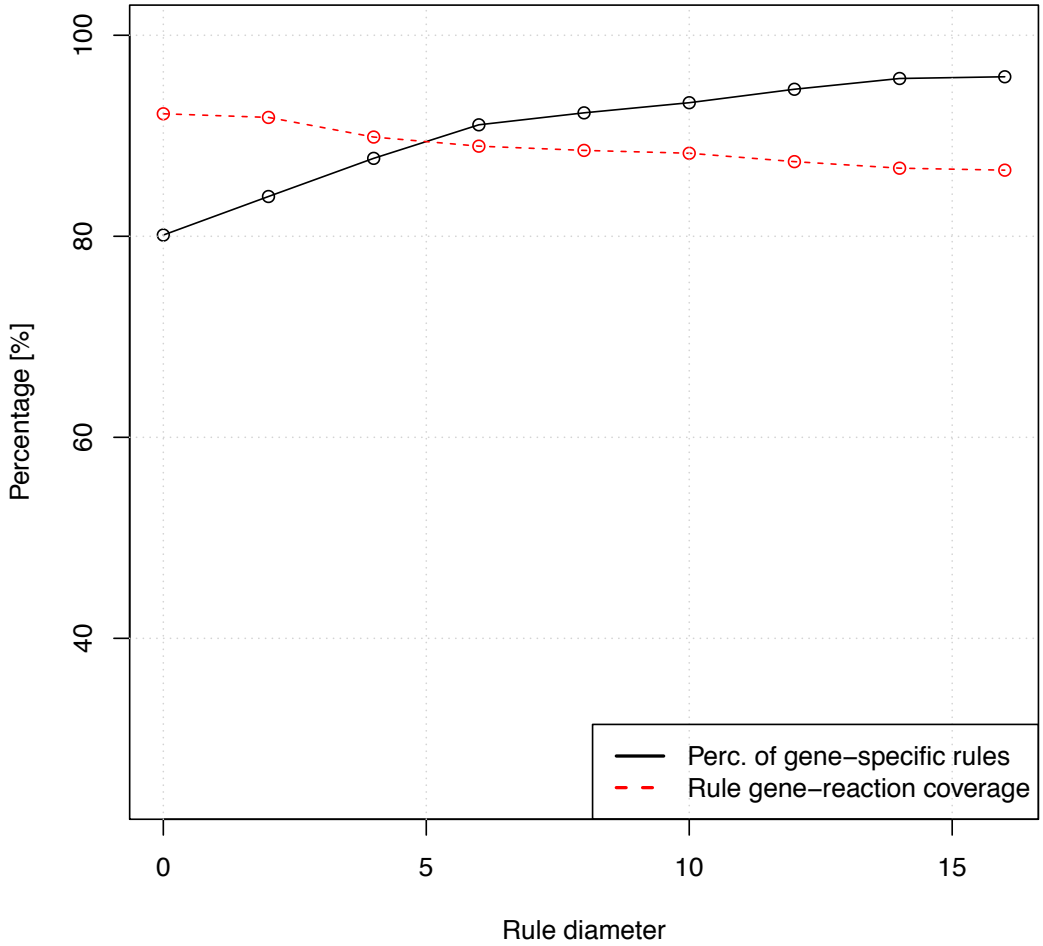


Figure S6. Specificity of rules vs. promiscuity recovery in function of diameter. Rules specificity is calculated as percentage of gene-specific rules (solid black line), which represents the percentage of the total rules at a given diameter that are associated with a single gene in *Escherichia coli*. 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.

SUPPLEMENTARY TABLES

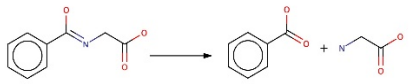
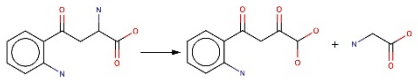
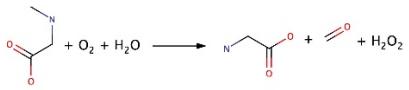
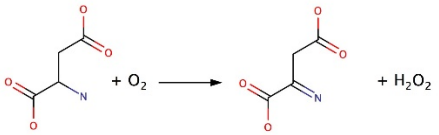
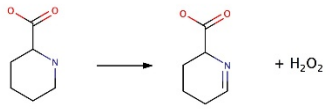
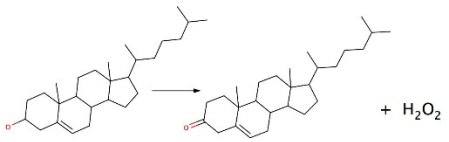
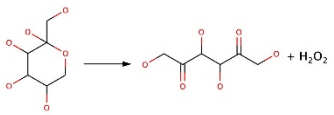
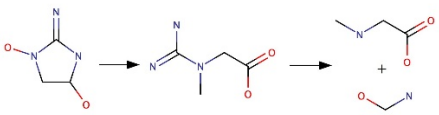
SUPPLEMENTARY TABLE S1

Supplementary Table S1 – 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.

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

SUPPLEMENTARY TABLE S2

Supplementary Table S2 – Examples of metabolic pathways enabling the detection of prostate cancer biomarkers. Supplementary Figure S2 illustrates the scope of sarcosine.

Biomarker	Metabolic Reaction	Effector(s)	References: Sample (S), Enzyme (E), Transcription-Factor (TF)*
Hippuric acid		Benzoic acid Glycine	S: Urine down-regulated (Struck-Lewicka et al., 2015) E: hippurate hydrolase (3.5.1.32) TF: benzoate BenR (Libis et al., 2016), glycine GcvR (Stauffer and Stauffer, 1994)
Kynurenine		Glycine Anthranilate	S: Serum, Urine, Tissue up-regulated (Sreekumar et al., 2009) E: glycine kynurenine-glyoxylate aminotransferase (2.6.1.63), anthranilate *kynureninase* (Jakoby and Bonner, 1953) TF: glycine GcvR (Stauffer and Stauffer, 1994) anthranilate AntR (Urata et al., 2004, p. 10)
Sarcosine		Glycine H_2O_2	S: Serum, Urine, Tissue up-regulated (Sreekumar et al., 2009) E: sarcosine oxydase (1.5.3.1) TF: glycine GcvR (Stauffer and Stauffer, 1994), H2O2: OxyR (Rubens et al., 2016)
N-acetyl-aspartate		H_2O_2	S: Serum, Urine, Tissue up-regulated (Sreekumar et al., 2009) E: aspartate oxydase (1.4.3.16) TF: OxyR (Rubens et al., 2016)
Pipecolate		H_2O_2	S: Serum, Urine, Tissue up-regulated (Sreekumar et al., 2009) E: pipecolate oxydase (1.5.3.7) TF: OxyR (Rubens et al., 2016)
Cholesterol		H_2O_2	S: Serum, Urine, Tissue down-regulated (Sreekumar et al., 2009) E: chlolesterol oxydase (1.1.3.6) TF: OxyR (Rubens et al., 2016)
L-Sorbose		H_2O_2	S: Urine down-regulated (Lima et al., 2016) E: sorbose oxydase (1.1.3.13) TF: OxyR (Rubens et al., 2016)
Creatinine		Urea H_2O_2 and glycine via sarcosine	S: Serum, Urine, Tissue down-regulated (Sreekumar et al., 2009) E: creatininase (1.1.3.13) followed by creatine amidinohydrolase (3.5.3.3) TF: urea NtcA (D'Orazio et al., 1996), glycine GcvR (Stauffer and Stauffer, 1994), H2O2: OxyR (Rubens et al.,

			2016)

* The column indicates the sample type (Serum, Urine, Metastasic 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 (Delépine et al., 2016).

SUPPLEMENTARY DATA

RETROPATH2.0 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`. Updates for RetroPath2.0 will be hosted at MyExperiment.org (https://www.myexperiment.org/workflows/4987.html).

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 of the set 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 4 and 5 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 of inhibiting transcription factors. Transcription factors are not provided but can easily be retrieved by entering the InChIs of the effectors in the SensiPath web server SensiPath.micalis.fr [1]. Table S2 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.

RP2PATHS OPEN SOURCE CODE FOR ENUMERATING PATHWAYS

The source code is available on GitHub in the "RP2paths" repository at https://github.com/brsynth/rp2paths). It is release under the MIT license (see https://opensource.org/licenses/MIT). Installation and documentation information are provided within the repository (INSTALL.txt and README.md files) as well as some examples (see examples folder).

In order to enumerate pathways from the LASER dataset, the following command was used (for a given compound):

```
python RP2paths.py all target_scope.csv --outdir pathways --notPathsStartingBy
$(cat list_currency_metabolites.txt)
```

where:

- *target_scope.csv* is a result scope file outputted by RetroPath2.0 that describes the network linking the targeted compound to the chassis.
- '*--outdir*' sets the path to the output folder
- '*--notPathsStartingBy*' defines a list of compounds that do not want as to use as starting chassis compound."

REFERENCES

Delépine, B., Libis, V., Carbonell, P., Faulon, J.-L., 2016. SensiPath: computer-aided design of sensing-enabling metabolic pathways. Nucleic Acids Res. 44, W226-231. doi:10.1093/nar/gkw305

D'Orazio, S.E., Thomas, V., Collins, C.M., 1996. Activation of transcription at divergent urea-dependent promoters by the urease gene regulator UreR. Mol. Microbiol. 21, 643–655.

Fernández-Peralbo, M.A., Gómez-Gómez, E., Calderón-Santiago, M., Carrasco-Valiente, J., Ruiz-García, J., Requena-Tapia, M.J., Luque de Castro, M.D., Priego-Capote, F., 2016. Prostate Cancer Patients-Negative Biopsy Controls Discrimination by Untargeted Metabolomics Analysis of Urine by LC-QTOF: Upstream Information on Other Omics. Sci. Rep. 6, 38243. doi:10.1038/srep38243

Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2010. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. Biotechnol. Bioeng. 106, 462–473. doi:10.1002/bit.22673

Huan, T., Troyer, D.A., Li, L., 2016. Metabolite Analysis and Histology on the Exact Same Tissue: Comprehensive Metabolomic Profiling and Metabolic Classification of Prostate Cancer. Sci. Rep. 6, 32272. doi:10.1038/srep32272

Jakoby, W.B., Bonner, D.M., 1953. Kynureninase from Neurospora: purification and properties. J. Biol. Chem. 205, 699–707.

Khalil, A.S., Collins, J.J., 2010. Synthetic biology: applications come of age. Nat. Rev. Genet. 11, 367–379. doi:10.1038/nrg2775

Li, Y., Qiu, S., Zhang, A.H., 2016. High-throughput metabolomics to identify metabolites to serve as diagnostic biomarkers of prostate cancer. Anal. Methods 8, 3284–3290. doi:10.1039/C6AY00127K

Libis, V., Delépine, B., Faulon, J.-L., 2016. Expanding Biosensing Abilities through Computer-Aided Design of Metabolic Pathways. ACS Synth. Biol. 5, 1076–1085. doi:10.1021/acssynbio.5b00225

Lima, A.R., Bastos, M. de L., Carvalho, M., Guedes de Pinho, P., 2016. Biomarker Discovery in Human Prostate Cancer: an Update in Metabolomics Studies. Transl. Oncol. 9, 357–370. doi:10.1016/j.tranon.2016.05.004

Liu, H., Ramos, K.R.M., Valdehuesa, K.N.G., Nisola, G.M., Lee, W.-K., Chung, W.-J., 2013. Biosynthesis of ethylene glycol in Escherichia coli. Appl. Microbiol. Biotechnol. 97, 3409–3417. doi:10.1007/s00253-012-4618-7

McDunn, J.E., Li, Z., Adam, K.-P., Neri, B.P., Wolfert, R.L., Milburn, M.V., Lotan, Y., Wheeler, T.M., 2013. Metabolomic signatures of aggressive prostate cancer. The Prostate 73, 1547–1560. doi:10.1002/pros.22704

Nam, H., Lewis, N.E., Lerman, J.A., Lee, D.-H., Chang, R.L., Kim, D., Palsson, B.O., 2012. Network Context and Selection in the Evolution to Enzyme Specificity. Science 337, 1101–1104. doi:10.1126/science.1216861

Rubens, J.R., Selvaggio, G., Lu, T.K., 2016. Synthetic mixed-signal computation in living cells. Nat. Commun. 7, 11658. doi:10.1038/ncomms11658

Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R.J., Li, Y., Nyati, M.K., Ahsan, A., Kalyana-Sundaram, S., Han, B., Cao, X., Byun, J., Omenn, G.S., Ghosh, D., Pennathur, S., Alexander, D.C., Berger, A., Shuster, J.R., Wei, J.T., Varambally, S., Beecher, C., Chinnaiyan, A.M., 2009. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457, 910–914. doi:10.1038/nature07762

Stauffer, L.T., Stauffer, G.V., 1994. Characterization of the gcv control region from Escherichia coli. J. Bacteriol. 176, 6159–6164.

Struck-Lewicka, W., Kordalewska, M., Bujak, R., Yumba Mpanga, A., Markuszewski, M., Jacyna, J., Matuszewski, M., Kaliszan, R., Markuszewski, M.J., 2015. Urine metabolic fingerprinting using LC-MS and GC-MS reveals metabolite changes in prostate cancer: A pilot study. J. Pharm. Biomed. Anal. 111, 351–361. doi:10.1016/j.jpba.2014.12.026

Tartaglia, L.A., Storz, G., Ames, B.N., 1989. Identification and molecular analysis of oxyR-regulated promoters important for the bacterial adaptation to oxidative stress. J. Mol. Biol. 210, 709–719.

Tedeschi, G., Nonnis, S., Strumbo, B., Cruciani, G., Carosati, E., Negri, A., 2010. On the catalytic role of the active site residue E121 of E. coli l-aspartate oxidase. Biochimie 92, 1335–1342. doi:10.1016/j.biochi.2010.06.015

Urata, M., Miyakoshi, M., Kai, S., Maeda, K., Habe, H., Omori, T., Yamane, H., Nojiri, H., 2004. Transcriptional regulation of the ant operon, encoding two-component anthranilate 1,2-dioxygenase, on the carbazole-degradative plasmid pCAR1 of Pseudomonas resinovorans strain CA10. J. Bacteriol. 186, 6815–6823. doi:10.1128/JB.186.20.6815-6823.2004

van der Meer, J.R., Belkin, S., 2010. Where microbiology meets microengineering: design and applications of reporter bacteria. Nat. Rev. Microbiol. 8, 511–522. doi:10.1038/nrmicro2392

Yim, H., Haselbeck, R., Niu, W., Pujol-Baxley, C., Burgard, A., Boldt, J., Khandurina, J., Trawick, J.D., Osterhout, R.E., Stephen, R., Estadilla, J., Teisan, S., Schreyer, H.B., Andrae, S., Yang, T.H., Lee, S.Y., Burk, M.J., Van Dien, S., 2011. Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol. Nat. Chem. Biol. 7, 445–452. doi:10.1038/nchembio.580

Zang, X., Jones, C.M., Long, T.Q., Monge, M.E., Zhou, M., Walker, L.D., Mezencev, R., Gray, A., McDonald, J.F., Fernández, F.M., 2014. Feasibility of detecting prostate cancer by ultraperformance liquid chromatography-mass spectrometry serum metabolomics. J. Proteome Res. 13, 3444–3454. doi:10.1021/pr500409q

Zhang, T., Watson, D.G., Wang, L., Abbas, M., Murdoch, L., Bashford, L., Ahmad, I., Lam, N.-Y., Ng, A.C.F., Leung, H.Y., 2013. Application of Holistic Liquid Chromatography-High Resolution Mass Spectrometry Based Urinary Metabolomics for Prostate Cancer Detection and Biomarker Discovery. PloS One 8, e65880. doi:10.1371/journal.pone.0065880