Metabolic Engineering 25 (2014) 140–158

Contents lists available at ScienceDirect

Metabolic Engineering

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

Generation of an atlas for commodity chemical production in *Escherichia coli* and a novel pathway prediction algorithm, GEM-Path

Miguel A. Campodonico [a,b], Barbara A. Andrews [a], Juan A. Asenjo [a], Bernhard O. Palsson [b,c], Adam M. Feist [b,c,*]

[a] Centre for Biotechnology and Bioengineering, CeBiB, University of Chile, Beauchef 850, Santiago, Chile
[b] Department of Bioengineering, University of California, 9500 Gilman Drive # 0412, San Diego, La Jolla, CA 92093-0412, USA
[c] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Lyngby, Denmark

ARTICLE INFO

Article history:
Received 7 March 2014
Received in revised form
17 July 2014
Accepted 21 July 2014
Available online 28 July 2014

Keywords:
Systems biology
Pathway predictions
Escherichia coli
Strain design

ABSTRACT

The production of 75% of the current drug molecules and 35% of all chemicals could be achieved through bioprocessing (Arundel and Sawaya, 2009). To accelerate the transition from a petroleum-based chemical industry to a sustainable bio-based industry, systems metabolic engineering has emerged to computationally design metabolic pathways for chemical production. Although algorithms able to provide specific metabolic interventions and heterologous production pathways are available, a systematic analysis for all possible production routes to commodity chemicals in *Escherichia coli* is lacking. Furthermore, a pathway prediction algorithm that combines direct integration of genome-scale models at each step of the search to reduce the search space does not exist. Previous work (Feist et al., 2010) performed a model-driven evaluation of the growth-coupled production potential for *E. coli* to produce multiple native compounds from different feedstocks. In this study, we extended this analysis for non-native compounds by using an integrated approach through heterologous pathway integration and growth-coupled metabolite production design. In addition to integration with genome-scale model integration, the GEM-Path algorithm developed in this work also contains a novel approach to address reaction promiscuity. In total, 245 unique synthetic pathways for 20 large volume compounds were predicted. Host metabolism with these synthetic pathways was then analyzed for feasible growth-coupled production and designs could be identified for 1271 of the 6615 conditions evaluated. This study characterizes the potential for *E. coli* to produce commodity chemicals, and outlines a generic strain design workflow to design production strains.
© 2014 International Metabolic Engineering Society. Published by Elsevier Inc. All rights reserved.

1. Introduction

The global chemical industry has been driven by petroleum feedstocks for the past 100 years, where synthetic organic chemistry played a key role. Today, the global chemical market landscape is beginning to change, based on new possibilities for bio-based product and process development. The renewed interest in industrial biotechnology is due to several reasons. First, the increases in petroleum prices squeeze commodity chemical production margins, increasing economically attractiveness of bio-based processes. Second, there is a strong socio-economic driver towards green chemistry and renewable feedstocks (Keasling, 2012). Third, due to technological developments, the past 20 years has seen the successful demonstration of metabolic engineering enabling the generation of microbial strains for the production of a wide range of chemical compounds (Atsumi and Liao, 2008; Lee et al., 2012; Peralta-Yahya et al., 2012). The availability of high-throughput technologies, the advances of computational methods, and emergence of genome-scale systems analysis to analyze large amount of omics data, has given rise to the concept of 'systems metabolic engineering' (Jang et al., 2012; Lee et al., 2012; Palsson and Zengler, 2010) where the focus has shifted from perturbing individual pathways to manipulating the organisms as a whole. Genome-scale models (GEMs) can now be used as query platforms to examine new strategies and interventions as they contain a parts list of cellular components and their interactions (Feist et al., 2007, 2009; Orth et al., 2011). By using constraint-based reconstruction and analysis (COBRA) approaches (Schellenberger et al., 2011), outcomes of cellular metabolism have been predicted successfully for the production of various compounds (Bordbar et al., 2014; Kim et al., 2008; Lee et al., 2012; McCloskey et al., 2013; Yim et al., 2011). Moreover, model-driven evaluations for the production potential for growth-coupled native products in *Escherichia coli* have been performed (Feist et al., 2010). However, a comprehensive computational analysis for the production of valuable non-native *E. coli*

* Corresponding author at: Department of Bioengineering, University of California, 9500 Gilman Drive #0412, San Diego, La Jolla, CA 92093-0412, USA. Fax: 858 822 3120.
E-mail address: afeist@ucsd.edu (A.M. Feist).

http://dx.doi.org/10.1016/j.ymben.2014.07.009
1096-7176/© 2014 International Metabolic Engineering Society. Published by Elsevier Inc. All rights reserved.

metabolites has not been performed. Therefore, we developed a systematic workflow in order to evaluate the production potential of 20 industrially relevant chemicals (Assary and Broadbelt, 2011; Curran and Alper, 2012; Fischer et al., 2008; Lee et al., 2012; Paster et al., 2003; Werpy et al., 2004; Zeng and Sabra, 2011) in *E. coli*, by integrating a combination of computational methods and developing a new pathway prediction algorithm, GEM-Path (Genome-scale Model Pathway Predictor).

Computational approaches for the prediction of non-native pathways exist, but are limited in their design and scope. Different approaches have been implemented for pathway prediction (Arita, 2000; Carbonell et al., 2011; Cho et al., 2010; Dale et al., 2010; Greene et al., 1999; Hatzimanikatis et al., 2005; Heath et al., 2010; Hou et al., 2003; McShan et al., 2003; Pharkya et al., 2004), where increasing attention has been focused mainly on retrosynthetic algorithms (Carbonell et al., 2011; Cho et al., 2010; Henry et al., 2010; Yim et al., 2011) based on Biochemical Reaction Operators (BROs). In these analyses, BROs are used to go from a target compound to a predefined set of metabolites in an iterative backward search. In summary, all of these methods shared basically the same workflow, first calculating all structurally possible pathways and then scoring them using different kinds of metrics. During the synthetic pathway calculation, these algorithms unnecessarily expand the reaction space, generating all possible pathways that link a specific metabolite to a final specific product without performing pathway integration with content known to exist in a given production host. Furthermore, previous algorithms do not integrate the bioprocessing condition-specific cofactor usage/generation, substrate usage, strain/oxygenation conditions, and related energy balances during the computation of pathways. In order to address these problems, we developed GEM-Path, by integrating retrosynthetic algorithms based on BROs and filtering procedures with GEMs at each iteration step. Furthermore, a novel reaction promiscuity analysis is introduced, which is based on known reaction substrate similarities. These two features distinguish GEM-Path from other computational approaches.

Once a synthetic pathway is successfully established, additional approaches can be taken to further engineer the host strain and synthetic pathways for enhanced production of a desired chemical. Adaptive laboratory evolution together with COBRA methods and organism-specific models has proven successful for the calculation of wild type *E. coli* optimal growth rates (Ibarra et al., 2002), native *E. coli* metabolite production through knock-outs (Fong et al., 2005), and for non-native *E. coli* metabolite production through heterologous pathway incorporation and knock-out implementations (Yim et al., 2011). Furthermore, the use of adaptive laboratory evolution together with growth-coupled knock-outs design, allows to select for strains with higher target compound production rates by coupling them to the selection for faster growth (Portnoy et al., 2011). Here, we integrate each of the predicted pathways under several different substrates/strain/oxygenation conditions with growth-coupled designs generated through reaction knock-outs by utilizing the RobustKnock (Tepper and Shlomi, 2010) and GDLS (Lun et al., 2009) algorithms. Finally, in order to characterize *E. coli's* potential production landscape for the studied compounds and for designs implementation purposes, a productivity analysis for maximum theoretical yield and maximum theoretical growth-coupled yield was performed.

2. Methods

2.1. Model and flux balance analysis

The metabolic reconstruction of *E. coli* iJO1366 was utilized as a basis for synthetic pathway calculations, yield analysis, and further strain designs. This model has been proven to be predictive for computations of growth rates and metabolite excretion rates on a range of substrates and genetic conditions (Feist et al., 2007; Orth et al., 2011). For all phenotype simulation, flux balance analysis (FBA) was used. The biomass objective function (BOF_{core}), maintenance energy, and basic constraints were set according to the reported values in the reconstruction. FBA used the assumption of steady-state metabolic flux as described elsewhere (Orth et al., 2010). All computations were performed using MATLAB® (The Mathworks Inc., Natick, MA, USA) and the COBRA Toolbox (Schellenberger et al., 2011) software packages with TOMLAB (Tomlab Optimization Inc., San Diego, CA, USA) solvers.

2.2. GEM-Path algorithm: chemoinformatics tools and techniques

Throughout the process of synthetic pathway generation, chemoinfomatic tools were essential for integrating computational chemical analysis into genome-scale model theory. In order to properly handle molecular structures, a range of chemoinfomatic techniques were incorporated into the COBRA Toolbox MATLAB® environment. For this purpose, in-house methods and functions, which are described below, were developed based on ChemAxon (ChemAxon Ltd., Budapest, Hungary) software package libraries.

Chemical representation: for compound and reaction representation MDL Molfiles (Dalby et al., 1992) were used. A Molfile contains information about the atoms, bonds, connectivity, and coordinates of a molecule. The Molfile consists of some header information, the connection table containing atom information, then bond connections and types, followed by sections for more complex information.

SMIRKS & SMARTS: for BRO representation, SMIRKS (James et al., 2004) was used as a language for describing generic reactions by using a SMARTS (James et al., 2004) representation of the reaction's substructures. A SMARTS pattern may include not only a specification of reaction center but also a specification of a local structure that must occur or is necessarily absent based on our best understanding of the relevant biochemistry (Silverman, 2002). BROs were constructed based on the smallest substructure related to the structural change of the main substrates and products in the reaction. Based on previous studies (Henry et al., 2010; Mu et al., 2011; Yim et al., 2011), a set of 443 irreversible BROs were defined to generate novel biochemical reactions and pathways. Approximately 76% of the reactions in KEGG (Kanehisa et al., 2006) and 72% of the reactions in BRENDA (Curran and Alper, 2012) involved a transformation captured in this defined BRO set. Furthermore, depending on the BROs's nature, three different types of metabolic transformation were defined: (i) '1-1' BROs simulate the substrate conversion without including any co-products and co-substrates in the BRO, (ii) '2-1' BROs simulate anabolic conversions, merging the substrate with a cosubstrate, and (iii) '1–2' BROs simulate catabolic conversions, where the substrate breaks into the corresponding product and a co-product. 2–2 transformations were ignored since they can be represented by a $2-1$ transformation followed by a 1–2 transformation. Co-products and co-substrates were selected from *E. coli's* metabolome information. This formulation allows a host-specific integration at the reaction prediction level.

Standardization and mass balance: since MDL Molfiles might come from different sources, a standardization procedure was performed. For each molecular structure, stereochemical information was removed and the major protonation form at pH 7 was determined. For each reaction, mass balance was performed using previously standardized molecular structures.

If hydrogen did not reach the balance, reaction stoichiometry was corrected.

Substrate fingerprint: substrates were represented by chemical fingerprints. A chemical fingerprint (CFP) is a simple record of the fragments present in a chemical structure. The chemical fingerprint (CFP) of a molecule is defined as CFP=(Fi), where Fi refers to a molecular fragment with real occurrence in a molecule. Fi is obtained by the molecular fragmentation method. Each Fi in the fingerprint is represented in bit string where each position of the sequence is represented by '1' or '0' digits, depending on the presence or absence of the structural pattern predefined by Fi. Previous studies have shown good results by using linear fragments from 5 to 6 bonds (Hu et al., 2012; Latino and Aires-de-Sousa, 2009). In this study linear fragmentation up to 6 bonds was used.

Tanimoto coefficient (*TC*): the premise of similarity searching is that similar structures have similar fingerprints. Here, we used the TC dissimilarity (TC_{diss}) metric to determine how similar two fingerprints were. Values of this metric are non-negative numbers. A zero dissimilarity value indicates that the two fingerprints are identical, and the larger the value of the dissimilarity coefficient the higher the difference between the two structures. In its original form, the Tanimoto metric is a similarity metric (TC_{sim}):

$$T_{sim} = \frac{B(a\&b)}{B(a)+B(b)-B(a\&b)}$$

where a and b are two binary fingerprints, & denotes binary bit-wise and-operator, | denotes bit-wise or-operator, and $B(x)$ is the number of 1 bits in any binary fingerprint x:

$$B(x) = |\{x_i = 1\}|x_i \in \{0,1\}; i = 1, \ldots\ldots, n\}| = \sum_{i=1}^{n} x_i$$

From that it is straightforward to obtain a dissimilarity measure:

$$T_{diss} = 1 - T_{sim}$$

It is worth noting that if the TC_{diss} between two fingerprints is 0, it means that both molecules share the exact same fingerprint. While this does not mean that both molecules are the same, it does mean that both molecules share the same bonds according to the fragmentation process, since the molecular fingerprint only represents the presence or absence of a given particular bond pattern.

Exact topology matching: molecular graphs consist of nodes and edges, with atoms corresponding to the nodes and bonds corresponding to the edges. When we compare structures represented as graphs, the graph patterns must match. The type of atoms and bonds must be similar during the structural search. In this study, no stereochemical information was used for matching compounds, only bond and atom connectivity for structural matching was analyzed. A full structure search solution in MolSearch is based on a substructure search algorithm (Ullmann's algorithm) combined with various heuristics and an additional check to verify that the number of heavy atoms are the same in the query and target molecules.

2.3. GEM-Path algorithm: databases

The *E. coli* metabolome was defined based on the GEM *i*JO1366. Metabolites were extracted from the model and downloaded from PubChem's (Bolton et al., 2008) compound database. Metabolites were saved as molfiles and named after their BiGG (Schellenberger et al., 2010) identifier. For reaction existence and reaction promiscuity analysis, the BRENDA (Scheer et al., 2011) database file and molecular structure molfiles were downloaded. Three digit EC number databases were generated by lumping together all reactions with similar third level EC numbers. Each entry in the database specifies the corresponding known biochemical reaction formula, the corresponding four digit EC number association, reaction-organism association, and substrate structure file. In cases where a specific reaction-organism association reported affinity for more than one substrate, an entry specifying all substrates was generated. For this purpose all reactions were assumed to be reversible and cofactors were not assigned as substrates.

2.4. GEM-Path algorithm: thermodynamic analysis

Thermodynamic analysis was performed by calculating the $\Delta_r G'$ (KJ/mol) where $\Delta_r G'^{\circ}$ was estimated based on the group contribution method (Jankowski et al., 2008). Intracellular concentrations were defined based on previous studies (Bennett et al., 2009). For unknown concentrations, estimations were calculated based on the non-polar surface area and compound charge (Bar-Even et al., 2011).

2.5. GEM-Path algorithm: promiscuity analysis

This analysis takes into account only substrate reaction promiscuity. Based on the similarity (TC) of the native and non-native substrates, a reaction promiscuity space can be generated and potential promiscuous activities determined depending on the distance between the promiscuous space and the metabolite to analyze. Thus, a similarity matrix based on the TC was calculated between every possible metabolite that the specific reaction-organism association could catalyze; cofactors were excluded from the matrix. Then, the reaction promiscuity space was defined by performing multilinear regression analysis on the similarity matrix, and an average distance between each native metabolite and space centroid was calculated. By dividing the potential promiscuous target substrate distance from the centroid over the average native distance from the centroid, the reaction promiscuity score (PS) was calculated. If the score was lower than 1.2, the reaction is considered to be promiscuous for the target substrate (Fig. 1). The reaction promiscuity score was tested and validated by using *E. coli's* promiscuous reaction information from *i*JO1366 (Supplementary Figs 1 and 2).

2.6. Theoretical analysis of the production potential in E. coli

To evaluate the production efficiency of each product under different metabolic conditions and to determine the most predominant metabolic subsystems that work as precursor sources for product formation, an initial theoretical analysis was performed calculating the maximum theoretical yield in *E. coli for* all predicted pathways. This analysis was executed by: (i) incorporating the heterologous pathways to the model, (ii) setting an uptake rate to 120 C-mmol $gDW^{-1}\,h^{-1}$ for each carbon source, 20 mmol $gDW^{-1}\,h^{-1}$ O_2 (Varma et al., 1993) when specified, (iii) setting the reactions CYTBDpp, CYTBD2pp, and CYTBO3_4pp to 0 mmol $gDW^{-1}\,h^{-1}$ for the ECOM strain (Portnoy et al., 2008), (iv) setting a minimal growth rate to sustain growth as 0.1 h^{-1} (as set by the amount of flux necessary through the BOF_{core}), and v) using FBA to maximize the flux through each of the exchange reactions in the model for the target compound. For each predicted pathway, phenotypic results were reported in terms of yield; specific product yield ($Y_{p/s}$) defined as the maximum amount of carbon product that can be generated per unit of

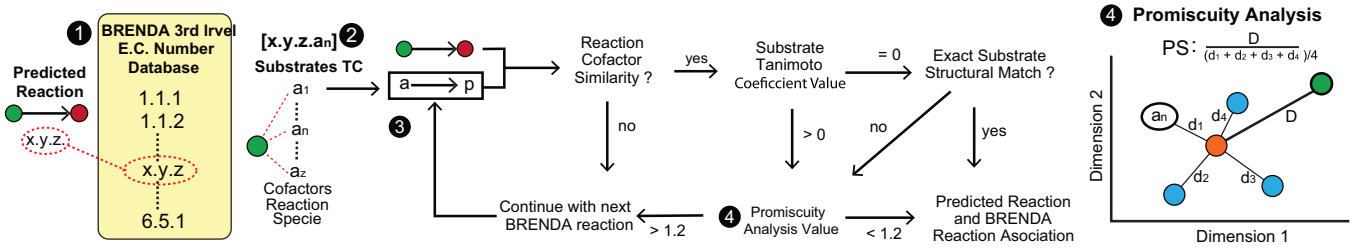

Fig. 1. Reaction existence and promiscuity analysis: the first three steps outline the main processes for reaction existence and promiscuity analysis, while the forth step shows specifically how the promiscuity analysis was performed. First, for a predicted reaction the third level BRENDA EC number database was identified (yellow box). For each reaction in the databases structural information regarding substrates, cofactor uses and species were determined. Second, the predicted reaction substrate (green circle) was compared to the corresponding third level BRENDA EC number database substrates by calculating the TC. From bottom to top, substrate pairs of TCs were sorted in decreasing order. Third, starting from the lowest TC (a1) until a predicted reaction and BRENDA reaction association was found (an), an iterative decision making algorithm determines whether the predicted reaction exists in BRENDA or there is any reaction in the database able to show promiscuous activity. Fourth, when a specific reaction is sent to promiscuous analysis, non-specific substrates (blue circles) for the reaction/species association are assigned according to BRENDA databases. By calculating the TC between all of the substrates a reaction promiscuity space was generated. From this space, distances from the centroid for each substrates and promiscuity score were calculated. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

carbon substrate.

$$Y_{p/s} = \frac{C_{\text{product}} * \textit{production rate}_{\text{product}}}{C_{\text{substrate}} * \textit{production rate}_{\text{substrate}}} \left(\frac{\text{cmmol}_{\text{product}}}{\text{cmmol}_{\text{substrate}}}\right)$$

where C is the number of carbons in the substrate and product. This metric provides a proper comparison between pathways productivities, since it standardizes the carbon consumption for each substrate.

2.7. Strain design computations

Before strain design, the model was preprocessed based on the problem formulation described by Feist et al. (2010). Preprocessing was condition specific and was performed for each pathway/substrate/oxygenation combination. The method utilizes six steps in which the model was reduced and target reactions were selected for knock-out simulations. By reducing the model and constraining the reaction set that could serve as a target for a reaction knock-out, computation time was effectively reduced when performing Robust Knock and GDLS algorithms.

RobustKnock and GDLS were implemented in the COBRA Toolbox framework as described in their original documentation. First, RobustKnock was utilized to design strains of *E. coli* for each target/substrate/oxygenation combination for a maximum of 2 and 3 reaction knock-outs. RobustKnock predicts reaction deletion strategies that lead to the over-production of compounds of interest by accounting for the presence of competing pathways in the network. Specifically, this method extends OptKnock to pinpoint specific enzyme-catalyzed reactions that should be removed from a metabolic network, such that the production of the desired product becomes an obligatory byproduct of biomass formation. The predicted set of reaction knock-outs eliminates all competing pathways that may hinder the chemical's production rate, resulting in more robust predictions than those obtained with OptKnock. This is achieved by searching for a set of reaction knock-outs under which the minimal guaranteed production rate of a chemical of interest is maximized, instead of simply assuming that the maximized production rate would be achieved by chance, as in OptKnock. The method is based on a bi-level max–min optimization problem that is efficiently solved via a transformation to a standard mixed-integer linear programming (MILP) problem. If the solution exists, this algorithm finds the global optima set of knock-outs that evaluate the maximum achievable yield for a specific target compound. Because of the nature of this algorithm and the large amount of combinations to simulate, a search with four knock-outs makes the computational time of the simulations intractable. Because of this, GDLS was used to evaluate the maximum theoretical yield for four knock-outs. GDLS is a scalable, heuristic, algorithmic method that employs an approach based on local search with multiple search paths ($k=2$), that results in an effective, low-complexity search of the space of genetic manipulations. Still, solutions found with this method do not assure a global optimum. Consumption rate for the main carbon substrate in each simulation was set to 120 C-mmol $gDW^{-1}\,h^{-1}$. If aerobic conditions were used, an oxygen uptake rate of 20 mmol $gDW^{-1}\,h^{-1}$ was also set. For the ECOM strain reactions CYTBDpp, CYTBD2pp, and CYTBO3_4pp were set to 0 mmol $gDW^{-1}\,h^{-1}$.

3. Results

A systematic workflow was developed and organized into three phases (Fig. 2). First, a synthetic pathway algorithm was developed which integrates GEMs directly into computation and industrially relevant target compounds for simulation were defined. Second, pathway production capabilities were examined in a number of production environments. Each pathway was incorporated into the *E. coli* GEM and analyzed in terms of maximal theoretical yield under different substrate, oxygenation, and strain conditions. Third, strain design computations was performed through a maximum yield analysis, utilizing the RobustKnock (Tepper and Shlomi, 2010) and GDLS (Lun et al., 2009) algorithms. The result was a compendium of candidate synthetic pathways leading to 20 large volume commodity chemicals and strain designs to couple their production to growth.

3.1. Synthetic pathway prediction algorithm development

GEM-Path combines and integrates different computational approaches (Supplementary Table 1). The motivation for generating this new framework was that no existing tool combined a comprehensive search of the biochemical space through reaction operators, a thermodynamic analysis of each step, and a filtering of possible reactions at each step through integration with a strain-specific GEM.

3.1.1. Biochemical Reaction Operators (BROs) formulation

An initial step in the design process was to define the set of Biochemical Reaction Operators (BROs) that accurately describes the biochemical reaction space. A total of 443 BROs were defined (see Section 2). For use in GEM-Path, each BRO was assigned a specific cofactor use based on the BiGG database (Schellenberger et al., 2010) terminology, and the corresponding third-level EC number.

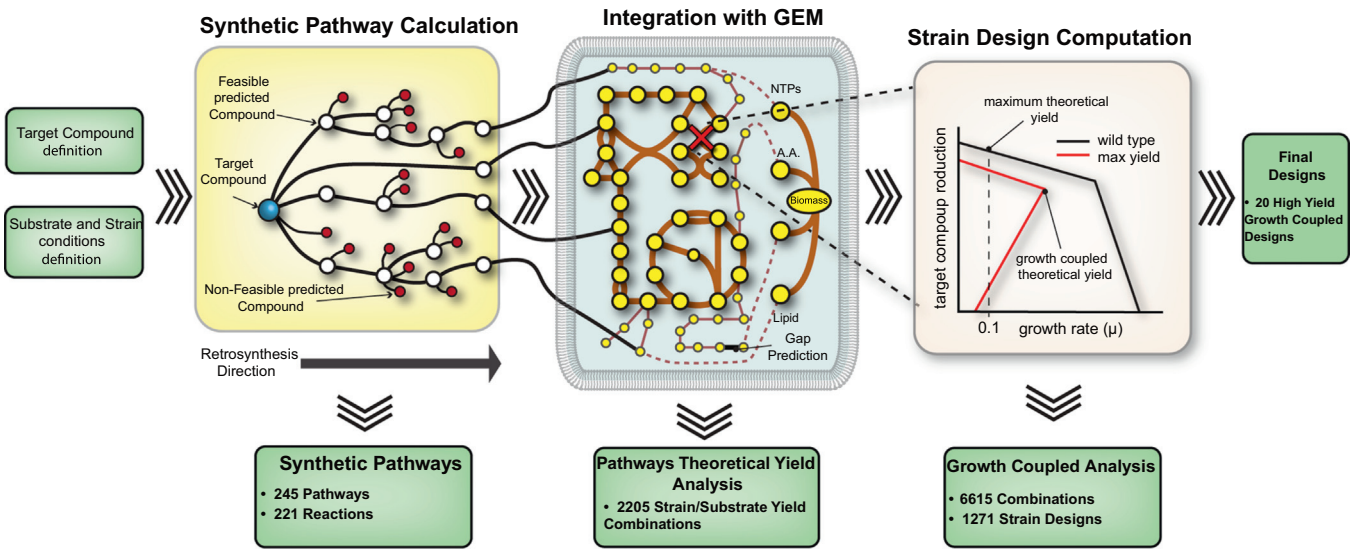

Fig. 2. Synthetic Pathway Calculation and strain design pipeline: this workflow outlines the integrated process of synthetic pathway prediction (yellow box), constraint-based modeling with the *E. coli* GEM (blue), and strain design computation with design algorithms (pink box). Green boxes represent framework inputs (entry arrows) and general result outputs (exit arrows). From the left, target compounds and substrates/strain conditions were defined to generate synthetic pathways. Synthetic pathways were calculated by using the developed GEM-Path algorithm integrated with GEM computation. Following GEM-Path, each pathway leading to a specific target compound was evaluated for growth-coupled feasibility under previously defined substrate/strain conditions. This workflow was used to outline the production routes from a distance of 4 reaction steps from *E. coli's* metabolome to 20 commodity chemicals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1.2. *Pathway predictor (GEM-Path) algorithm*

The pathway predictor algorithm was developed in an iterative manner (Supplementary Fig. 3). The process can be broken down into four major steps:

(a) Starting from the target metabolite, predictor constraints were set, such as maximal pathway length, metabolites to compute at each iteration, a thermodynamic threshold, and a reaction promiscuity threshold.
(b) Predefined BROs were applied to the target in a retrosynthetic manner for generating the corresponding substrates. After BROs application, the corresponding cofactors and third level EC numbers were assigned together with reaction structure files for further analysis. All predicted reactions were then checked for mass balance. If mass balance was not fulfilled, reactions were discarded from the process. Next, predicted metabolites were structurally compared against *E. coli's* metabolome. Substrate dissimilarities were sorted in terms of the TC (see Section 2), and an exact match analysis was performed for TCs equal to 0 since this does not necessarily mean that the compared molecules are the same. If the predicted metabolite matches any compound in the metabolome, FBA was performed in order to validate the potential production.
(c) A thermodynamic analysis was performed by calculating the $\Delta_r G$ (kJ/mol). Each predicted reaction was checked in terms of thermodynamic feasibility for existence of further reactions and potential promiscuity analysis. Reactions with $\Delta_r G$ lower than or equal to 25 kJ/mol were defined as feasible reactions and saved to continue the checking process. The threshold was set based on estimated variability calculated elsewhere (Henry et al., 2007).
(d) As shown in Fig. 1, in order to determine reaction existence, predicted reactions were compared against BRENDA. The database was structured by lumping together all reactions with similar third level EC numbers. Each level contains known biochemical reactions with the corresponding four digit EC number association, reaction-organism association, and substrate structure file. The third level EC number association for the predicted reaction facilitates the identification of the third level EC class BRENDA sub group for substrate comparison. By calculating TC, predicted substrates could be compared against all corresponding substrates present in the BRENDA subgroup. The results were sorted and analyzed starting with the most similar compound. Dissimilarities equal to 0 were structurally compared by performing an exact match comparison (see Section 2). If the substrates were structurally similar, reaction cofactors were compared. In cases where the predicted reaction matches a reaction in BRENDA, a specific reaction-organism association was assigned to the reaction and the pathway prediction procedure was continued. Otherwise, a substrate promiscuity analysis was performed by considering the reaction-organism association substrate information. If the reaction is considered to be promiscuous, the algorithm saves the reaction, otherwise, it proceeds by analyzing the potential promiscuity for the next sorted substrate. In order to decide whether a reaction might be promiscuous or not, a reaction promiscuity score was calculated based on the similarity between the reaction native substrate and the predicted substrate (Fig. 1, step 4). The reaction promiscuity score was calculated and analyzed by using *E. coli's* promiscuous reaction information from *i*JO1366 (Supplementary Figs. 1 and 2). Based on the previous analysis, the reaction promiscuity score threshold was set to 1.2.

After the filtering steps, only the 120 predicted compounds closest to *E. coli's* metabolome were allowed to continue the algorithm. This process was repeated 4 times, which means pathways of a maximal length of 4 were obtained. The GEM-Path algorithm overcame the disadvantages of previous methods by not setting a specific metabolite source for the target compound formation, instead leaving open the possibility to reach any metabolite in the metabolome. Furthermore, structural comparison gives the ability to focus on the retrosynthesis direction most similar to the corresponding region of the host metabolome. It should be noted that these characteristics could be extended to other organisms, predicting synthetic heterologous pathways in a host-context specific manner. After completion of this computational procedure, the resulting pathways were characterized and

used for theoretical yield analysis under different strain, oxygenation, and substrate conditions. All of the predicted pathways are given in Supplementary Fig. 9 and specified in Supplementary Table 8. Thus, a comprehensive list of feasible biochemical pathways leading to the target compound formation was established.

3.2. Description of substrate and product selection

Important production capabilities of the synthetic pathways predicted by GEM-Path were assessed using the *E. coli* GEM. For theoretical yield analysis, three primary substrates were evaluated based on the cost and availability of suitable feedstock (Sauer et al., 2008; Vickers et al., 2012), *E. coli's* metabolic capacity for catalyzing such carbon sources, and unique design potential (e.g., glucose and fructose are not unique and are examples of interconverted substrates with little to no cost to the cell, (Feist et al., 2010)). The first two substrates were glucose and xylose, five- and six-carbon sugars, present in lignocellulosic biomass, representing about 40–50% and 20–30% by dry weight of plant material, respectively (Wyman et al., 2005). The use of this type of feedstock is expected to increase with the incentive to produce biofuel and bio–based chemicals (Perlack and Stokes, 2011). The third substrate was glycerol, a three-carbon molecule and a byproduct of biodiesel production (Ma and Hanna, 1999), whose availability is expected to increase in the coming years (Yang et al., 2012). In addition, three different starting strains and oxygenation conditions were analyzed for each product during the synthetic pathway calculations procedure. These are a wild-type strain under aerobic conditions, a wild-type strain under anaerobic conditions, and the 'ECOM' (*E. coli* cytochrome oxidase mutant) strain under aerobic conditions (Portnoy et al., 2008). The ECOM strain has the advantage of "aerobic fermentation" as the strain cannot use oxygen as a terminal electron acceptor. The list of targeted overproduction metabolites included 20 different bulk chemicals with biological production potential and precursors for commercially valuable chemical production are shown in Fig. 3. The selection of the 20 target compounds was determined by evaluating reports generated by the US Department of Energy (Paster et al., 2003; Werpy et al., 2004), which includes chemicals that are currently being produced on an industrial scale (Zeng and Sabra, 2011) and metabolites that are described as precursors to or potential target biofuel compounds (Assary and Broadbelt, 2011; Curran and Alper, 2012; Fischer et al., 2008; Lee et al., 2012). By comparing the target compound list with *i*JO1366 *E. coli's* metabolome, 4 out of 20 products were assigned as native and 16 as non-native. Synthetic pathways for native products were calculated in order to explore the possibility of more productive pathways for their synthesis.

3.3. Predicted pathways and reaction specifications

The synthetic pathway calculation procedure using GEM-Path was applied to all selected target compounds of interest and

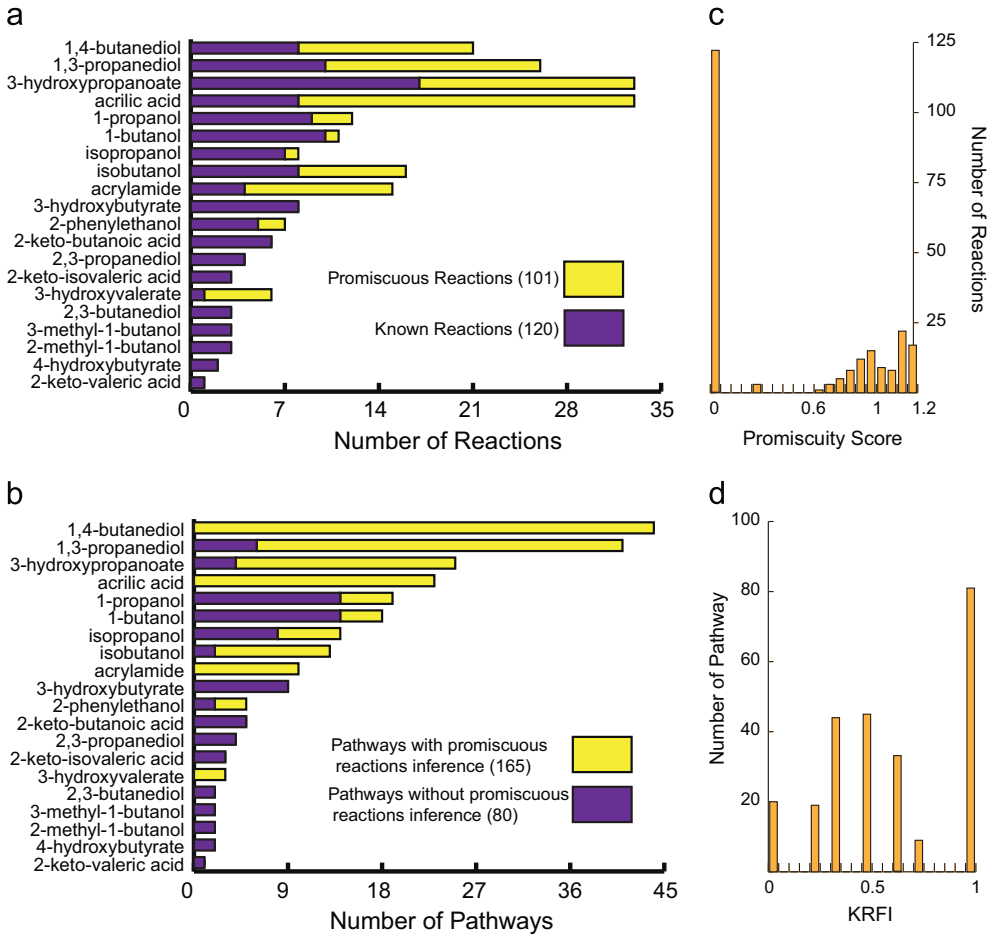

Fig. 3. Predicted reaction and pathway analysis: for each target compound, the total number of synthetically generated reactions (a) and pathways (b) were plotted. 'Promiscuous' predicted and 'known' reactions are differentiated with yellow and purple sub segments, respectively. In the case of pathways, those containing one or more promiscuous reactions and those with no promiscuous reactions involved were differentiated by corresponding yellow and purple sub segments. Reaction promiscuity score distribution (c) and Known Reaction Fractional Index (KRFI) for each pathway (d) are also shown. A value of 0 for the promiscuity score indicates 'known' reactions, and a value of 1 for KRFI indicates the predicted pathway is constituted by only 'known' reactions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

validated by comparing the output pathways with previous computationally calculated and experimentally implemented pathways. In summary, 245 pathways, 221 reactions, and 59 non-native intermediate metabolites were calculated after 4 iterations of the algorithm (i.e., a maximal pathway length of 4). In total, 25%, 39%, 28%, and 8% of the pathways were of length 4, 3, 2 and 1, respectively. For each product, pathways combining potential promiscuous reactions, already known reactions (i.e., in BRENDA), and different co-factor (i.e., using NAD^+ or $NADP^+$) uses were generated. In total, 44 different precursors from the native *E. coli* metabolome were determined that connected to the synthetic pathways. Furthermore, 42 gap-filling reactions interconnecting native *E. coli's* metabolites were identified which enabled production of a targeted compound. This set includes reactions which may be the reverse reaction of a native enzymatic step in the existing network or have completely unique chemistry acting on a native metabolite.

The number of reactions and pathways predicted using GEM-Path varied across the 20 target compounds analyzed. For 1,4-butanediol, 1,3-propanediol, 3-hydroxypropanoate, and acrylic acid, the number of reactions and pathways were the highest (Fig. 3). In total, approximately 51% of all predicted reactions were categorized as 'known', which means that each predicted reaction has an exact biochemical reaction association according to BRENDA (Fig. 3a purple sub-segments). Reactions represented by yellow sub-segments in Fig. 3 correspond to predicted biochemical reaction steps assigned as 'promiscuous' from the promiscuity analysis. Furthermore, for each of these reactions, a potential reaction from BRENDA that might carry flux through the synthetic pathway was assigned. The promiscuity score distribution is represented in Fig. 3c. As expected, all 112 'known' reactions were represented with a promiscuity index equal to 0 and 'promiscuous' reactions were distributed around 1.

A predicted pathway can be either entirely 'known' (Fig. 3b purple sub-segments), meaning every reaction in the pathway has an exact biochemical reaction association according to BRENDA, or partially known, where one or more reactions in the pathway were predicted as 'promiscuous' (Fig. 3b yellow subsegments). According to the classification in Fig. 3b, all of the pathways able to generate 1,4-butanediol, acrylic acid, acrylamide, and 3-hydroxyvalerate in *E. coli* contain at least one promiscuous reaction. In order to analyze the fraction of known reactions present in a pathway, we defined the Known Reaction Fractional Index (KRFI) between 0 and 1, where 1 means that the pathway has been completely reconstructed from 'known' reactions and 0 means that it has been completely reconstructed from 'promiscuous' reactions. Based on the previous definition, 30% of all predicted pathways were entirely 'known'. In Fig. 3d, entirely 'known' pathways are represented with a KRFI equal to 1, and the rest of the pathways correspond to partially known pathways distributed from 0 to 1. In total, approximately 40% of the predicted reactions were oxidoreductases, acting on hydroxyl or aldehyde groups with NAD^+ or $NADP^+$ acceptors. Carbon–oxygen and carbon–carbon lyases correspond to around 20%, and transferases, specifically CoA-transferases, were 8% of all of the reactions. This set of generic biochemical transformations details the chemical nature of the predicted reactions that most often enable the production of the targeted non-native compounds in *E. coli* (see Supplementary Table 3).

3.4. GEM-Path validation

In order to validate the proposed algorithm, previous work examining computational and experimentally implemented heterologous pathways in *E. coli* were compared to the GEM-Path calculated pathways. According to a bibliographic search, 14 out of 20 target compounds were found to be referenced and targeted by patents or scientific publications. The maximum theoretical yield calculated by GEM-Path for the targeted compounds was then compared to production levels from the bibliographic search set (Table 1). In order to determine the production potential for the novel pathway calculated using GEM-Path, a maximum theoretical production comparison was performed for experimentally and computationally reported pathways (Table 1). The analysis was performed by calculating the target production ratio between the highest flux carrying novel pathways predicted by GEM-Path over the experimentally and computationally reported pathways. Simulations were run under aerobic and anaerobic conditions, by using glucose, xylose, and glycerol as a carbon source. Values over 1 indicated that GEM-Path's novel pathways have higher production potential than already referenced pathways. Considerable improvements over experimentally implemented pathways were found in the GEM-Path set, specifically under anaerobic conditions, for 1,4-butanediol, 1,3-propanediol, isopropanol, and 3-hydroxybutyrate on various substrates. Distinct, but equal yield pathways were calculated for 1,3-propanediol and 1-butanol. In addition, already known implemented pathways for 1-propanol, 2-phenylethanol, 2,3-propanediol, 2,3-butanediol, 3-methyl-butanol, 2-methyl-butanol, and 4-hydroxybutyrate were found (Table 1). These findings revealed that GEM-Path calculated pathways contained experimentally-implemented pathways found in the literature screen and that the selected reaction rules were able to represent the known biochemistry and serve as validation of the approach.

For the synthetic design of biochemical pathways, considerable attention has been focused on BRO-based computational tools (Medema et al., 2012). As such, the pathways predicted from GEM-Path were compared against computationally-predicted pathways from three different BRO based algorithms; BioPath for the production of 1,4-butanediol (Yim et al., 2011), BNICE for the production of 3-hydroxypropanoate (Hatzimanikatis et al., 2005), and the one developed by Cho et al., for the production of several alcohols (Cho et al., 2010). The first comparison was for the synthetic pathway prediction of 1,4-butanediol by using the BioPath algorithm. When analyzing individual reactions, 91% off all reactions were able to be predicted by GEM-Path independently. Furthermore, through FBA analysis, novel pathways generated with GEM-Path were able to achieve higher theoretical productivity compared to BioPath reported pathways. Specifically for pathways 13 and 14 (see 1,4-butanediol pathways map in Supplementary Fig. 9), under aerobic condition and using glucose, xylose, and glycerol, a 10% theoretical productivity increase over BioPath predicted pathway was calculated. Moreover, by using the same substrates under anaerobic conditions, an approximately 30% increase over BioPath predicted pathways was calculated. The second case studied was for the synthetic pathway prediction of 3-hydroxypropanoate by using the BNICE algorithm. This framework is able to produce all thermodynamically feasible pathways from a source metabolite to a target compound. In this case, GEM-Path was able to generate 11% of all predicted pathways by this algorithm, and 87% of all reactions. This result was expected since both algorithms share similar BROs. By applying the reaction existence and promiscuity analysis based on BRENDA, GEM-Path was able to constrain the predicted pathways by reporting only a feasible subset of pathways. According to FBA simulations, novel pathways generated with GEM-Path were able to achieve the same maximum theoretical production rates compared to BNICE generated pathways, specifically for pathways 12, 13, and 3 (see 3-hydroxypropanoate pathways map in Supplementary Fig. 9). When using xylose and glucose as substrates, production rates were 76% higher than glycerol. Under aerobic conditions, no substantial increments in theoretical production rates between

Table 1
Comparison of GEM-Path predictions to previously identified pathways from literature.

Target compound	Experimental							Computational						
	Anaerobic			Aerobic			Ref.	Anaerobic			Aerobic			Ref.
	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol		Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol	
1,4-butanediol	1.3	1.3	1.3	1.1	1.1	1.1	(Yim et al., 2011)	1.3	1.3	1.3	1.1	1.1	1.1	(Yim et al., 2011)
1,3-propanediol	1.2	1.3	1	1	1	1	(Laffend et al., 1997; Nagarajan and Nakamura, 1998; Tang et al., 2009; Zeng and Sabra, 2011)	—	—	—	—	—	—	–
3-hydroxypropanoate	1	1	1	1	1	1	(Lynch, 2011; Suthers and Cameron, 2005; Wang et al., 2012)	1	1	1	1	1	1	(Henry et al., 2010)
1-propanol	1	0.9	0.9	1	1	1	(Pharkya, 2011; Shen and Liao, 2008, 2013)	2.4	2.8	3.3	1.2	1.2	1.1	(Cho et al., 2010)
1-butanol	1	1	1	1	1	1	(Atsumi et al., 2008; Bramucci et al., 2008; Lee and Park, 2008; Shen et al., 2011)	1	1	1	1	1	1	(Cho et al., 2010)
Isopropanol	1.2	1.2	1.9	1	1	1.1	(Hanai et al., 2007; Jojima et al., 2008; Pharkya, 2011)	—	—	—	—	—	—	–
Isobutanol	0.8	0.7	0.8	1	1	1	(Atsumi et al., 2010; Trinh, 2012)	0.8	0.7	0.8	1	1	1	(Cho et al., 2010)
3-hydroxybutyrate	1.2	1.2	1.5	1	1	1	(Tseng et al., 2009; Valentin and Dennis, 1997)	—	—	—	—	—	—	–
2-phenylethanol	0.9	0.9	0.9	1	1	1	(Hwang et al., 2009; Koma et al., 2012)	0.9	0.9	0.9	1	1	1	(Cho et al., 2010)
2,3-propanediol	1	1	1	1	1	1	(Altaras and Cameron, 1999; Soucaille et al., 2008)	—	—	—	—	—	—	–
2,3-butanediol	1	1	1	1	1	1	(Ji et al., 2011; Lu et al., 2012; Yan et al., 2009)	—	—	—	—	—	—	–
3-methyl-1-butanol	1	1	1	1	1	1	(Connor et al., 2010)	1	1	1	1	1	1	(Cho et al., 2010)
2-methyl-1-butanol	1	1	1	1	1	1	(Cann and Liao, 2008)	1	1	1	1	1	1	(Cho et al., 2010)
4-hydroxybutyrate	1	1	1	1	1	1	(Zhou et al., 2012)	—	—	—	—	—	—	–

For each target compound, the maximum theoretical productivity ratio between novel pathways generated by GEM-Path and experimentally implemented or computationally generated pathways is shown. Empty spaces (—) indicate that no referenced pathways for the corresponding target compound were found.

GEM-Path and previously generated BNICE pathways were identified. Finally, the third case analyzed after the synthetic pathway generation was Cho, et. al. Here, the author introduces a novel scoring algorithm in order to extract the most feasible pathways. The framework was validated for the production of 1-propanol, 1-butanol, 2-methyl-1-butanol, 3-methyl-1-butanol, isobutanol, and 2-phenylethanol from a variety of 2-ketoacids. When comparing the results between GEM-Path and Cho's predictions for each product, the same pathways were found for each case. Still, according to the simulations, none of the remaining pathways predicted by GEM-Path were able to achieve the production rates of pathways previously generated by Cho´s algorithm. Pathway and reaction prediction discrepancies were due to the filtering procedure, specifically during the promiscuity analysis, where only the most promising reactions were allowed to constitute a pathway in GEM-Path. However, of note is that the vast majority of reactions predicted in the referenced work was also predicted with GEM-Path. Specifically, GEM-Path was able to simulate 92% and 32% of all reactions and pathways, respectively. Furthermore, discrepancies arose due to a lack of connectivity between the host metabolic network and the predicted synthetic pathways in the referenced work and also from the predefined pathway length which allowed a maximum pathway length of four. A number of differences can be the result of the GEM-Path algorithm immediately stopping the search through each branch when it reaches the metabolome; the three other algorithms mentioned above do not have this stipulation.

When comparing GEM-Path with other computational tools (see Supplementary Table 1), its most characteristic features are its capability to shrink the biochemical reaction solution space by calculating the closest pathways to the metabolome and its ability to select mechanistically-feasible reactions from BRENDA. These properties rely on the integration of the promiscuity analysis and GEMS into the reaction prediction algorithm. Furthermore, GEM-Path is able to systematically integrate physiological conditions (e.g., carbon source and oxygenation) into the pathway generation procedure, allowing for the consideration of the active content under a given condition and not reactions or nodes that cannot be reached under a desired media condition. Furthermore, when comparing GEM-Path to previous tools, it shows a wider predictive capacity as it, (i) takes into account more cofactors, (ii) does not constrain the search to only one compound source, instead every metabolite in the metabolome might work as a source, and (iii) allows generation of anabolic and catabolic reactions. Nevertheless, some solutions might be hindered as not all nodes (i.e., predicted compounds) were allowed to continue through the prediction algorithm when compared to the *E. coli* metabolome. However, GEM integration into GEM-Path allows the algorithm to find more than one precursor present in the metabolic network without constraining the search to only one compound.

3.5. Theoretical yield analysis of the production potential in E. coli

The production potential landscape in *E. coli* was outlined by calculating and plotting the maximum theoretical yield for each target compound in terms of carbon moles captured (i.e., C-mol). Simulations were performed by combining all predicted pathways with the corresponding substrate utilization and oxygenation conditions (see Section 2). In total, 2205 flux balance analysis (FBA) combinations were calculated (Supplementary Fig. 4). Maximum theoretical yields (Fig. 4a–c) and the corresponding pathways were tabulated for each target compound (Table 2). Results were grouped together based on strain and oxygenation conditions and a yield interval was applied to plot the number of pathways for different substrates (Fig. 4a–c). Furthermore, in order to determine the most efficient subsystem for product formation, results were clustered in terms of yield and *E. coli's* precursor metabolic subsystems (Fig. 4d). The specific analysis for each strain/oxygenation condition can be found in Supplementary note. Overall, the average yields for WT/aerobic, ECOM/aerobic, and WT/anaerobic were 0.68, 0.53, and 0.38, respectively. By defining the ECOM/aerobic condition as an intermediate state of aerobiosis between WT/aerobic and WT/anaerobic, a correlation between the aerobiosis state of the cell and the production potential can be drawn. As shown in Fig. 4a–c, a pronounced displacement of maximum theoretical yield distributions towards lower yields is directly correlated with the extent of anaerobiosis. This trend is also shown in Fig. 4d, where a gradual shift from lower anaerobic yield to higher aerobic yields can be visualized. Furthermore, this pattern is shown together with a preference for glycerol as a substrate under aerobic conditions and for glucose under anaerobic conditions. Pathways near central carbon metabolism subsystems are able to achieve higher yields (Fig. 4d).

3.6. Strain design

Utilizing the synthetic pathways identified for each target compound, strain design simulations were performed to determine if reaction knock-outs could increase production. The predicted synthetic pathways were independently incorporated into the *E. coli* GEM and further model preprocessing was executed according to a previously developed approach (Feist et al., 2010). Growth-coupled designs, which couple the optimal production of biomass and energy generation to the production of the compound of interest, were chosen as objectives for the strain design performed here. A combination of the RobustKnock (Tepper and Shlomi, 2010) and the GDLS (Lun et al., 2009) algorithms with the conditioned model of *i*JO1366 (Orth et al., 2011) was used. First, RobustKnock was utilized to design strains of *E. coli* for each target/substrate/oxygenation combination for a maximum of two and three reaction knock-outs allowed. GDLS was used in order to decrease computational time and to evaluate the maximum theoretical growth-coupled yield for four knock-outs.

All reactions which were identified in the strain design process for elimination were collected and analyzed (see Supplementary Table 4). Growth-coupled designs could be found for 1271 different target/substrate/oxygenation/knock-out combinations (Supplementary Table 7). Overall, this number was 19% out of the 6615 possible conditions examined. The results of the design analysis are given in Table 3. Result landscapes of maximum growth-coupled yield for each target compound are shown in Fig. 5. Overall, production could be growth-coupled in 75% of the targeted compounds and 43% of all predicted pathways. Targets which could not be growth-coupled were 2,3-propanediol, 3-methyl-1-butanol, 2-methyl-1-butanol, 4-hydroxybutytare, and 2-phenylethanol. In total, 84 different reaction knock-outs were identified across all selected target reactions, some of them participating more frequently in strain designs. Pyruvate formate lyase and ATP synthase occurred 12 times more often than the average 44 knock-outs per reaction in all designs (1271). Pyruvate kinase occurred 7.4 times more and acetate kinase, pyruvate dehydrogenase, triose-phosphate isomerase, glucose-6-phosphate isomerase, ribulose 5-phosphate 3-epirase, glutamate dehydrogenase, alcohol dehydrogenase, and malate dehydrogenase occurred approximately 2.8 times more often than the average. As stated earlier (Feist et al., 2010), this uneven distribution of reaction knock-out occurrences suggests that certain reactions are critical for diverting carbon flux.

Approximately 8% of all designs were above a molar yield for carbon of 0.6, and this corresponded to designs for 9 out of 20 targeted compounds. When comparing different oxygenation conditions, most of the designs were calculated under wild type/anaerobic conditions (40%), followed by wild type/aerobic (33%), and ECOM/aerobic (23%). The highest average yield for all possible designs was calculated for wild type/anaerobic as being approximately 17% and 91% higher than the ECOM/aerobic and wild type

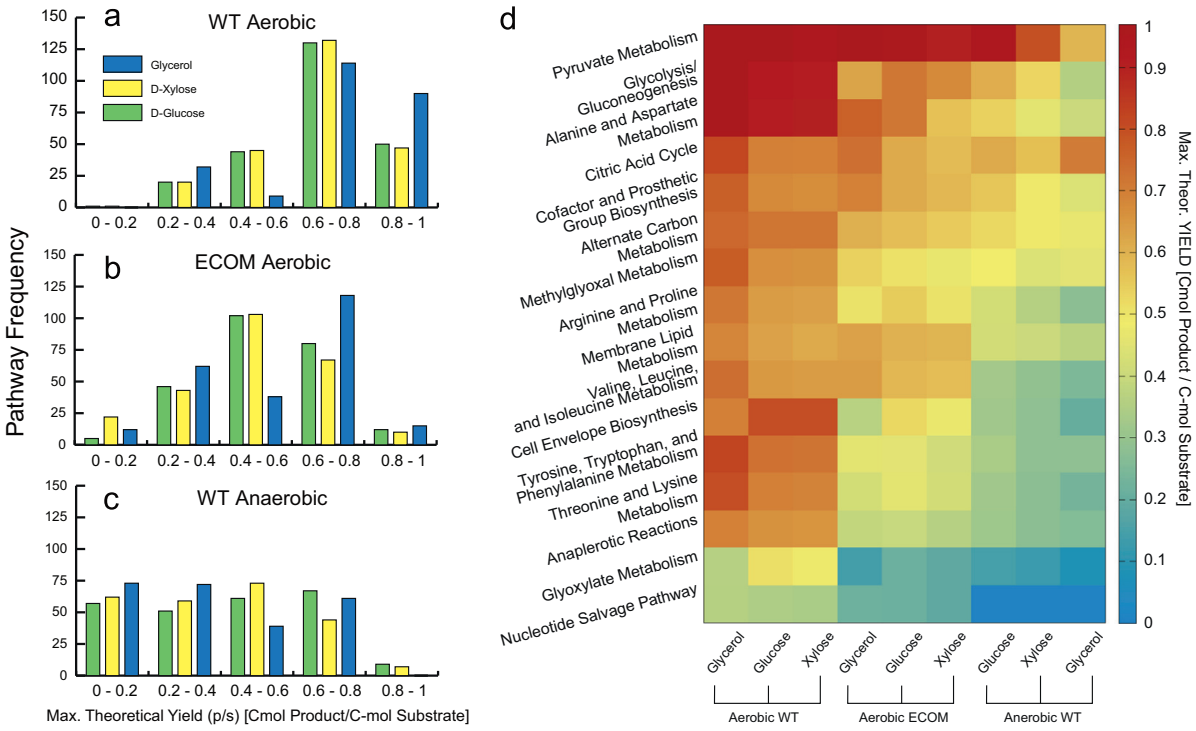

Fig. 4. Theoretical maximal yield distribution for different strain/substrate conditions and subsystems: FBA was performed for each predicted pathway and strain/substrate condition using the *E. coli* GEM. At each yield interval, the number of pathways was plotted for each specific substrate: glucose (green), xylose (yellow), and glycerol (blue). This analysis was performed for wild type/aerobic (a), ECOM/aerobic (b), and wild type/anaerobic (c). In total, 2205 simulations were performed. (d) The subsystem form which each precursor metabolite was determined by analyzing the reaction that connects the network with the corresponding synthetic pathway. The yield average was calculated for each precursor subsytem and clustered by each strain/substrate condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

aerobic, respectively. The predominant substrate for growth-coupled designs was glycerol (40%), then xylose (32%), and glucose (28%). The average growth-coupled yield distribution follows a similar trend where glycerol was 21% and 56% higher than xylose and glucose, respectively. According to the previous study (Feist et al., 2010), the larger the number of allowable knock-outs for a given target compound, the greater the maximum achievable yield. This trend was observed when comparing RobustKnock for two and three knock-out designs, where the average maximum growth-coupled yield was 21% higher for three versus two knock-outs. When comparing GDLS strain designs for four knock-outs to RobustKnock strain designs for three knock-outs, 37% of all designs were able to achieve higher C-mol yield when allowing four knock outs. Furthermore, for 9% of all designs, GDLS was able to find a growth-coupled design when RobustKnock could not (see Supplementary Fig. 6). However, when comparing GDLS output for four knock-outs to RobustKnock for 3 knock-outs, no increase in the average maximum achievable yield was observed (i.e., for 519 3-KO designs, the average C-mol yield was 0.35 whereas for 352 4-KO designs, the average yield was 0.34). This can be because the GDLS algorithm is not guaranteed to find an optimal solution (Lun et al., 2009), but this value could increase given a longer run time or different starting parameters.

Overall, the growth-coupled yield analysis revealed a positive correlation between the total number of strain designs and the number of predicted pathways for each target compound (Table 3). The same correlation was observed when comparing the number of independent growth-coupled pathways and the number of predicted pathways for each target compound. When examining specific targeted products, approximately 40% of all predicted pathways were able to couple the target compound production with growth across any of the predefined oxygenation/substrate/knock-out conditions. Strain design C-mol yield averages for the production of 1-butanol were higher than the corresponding medial yield for other target compounds. For acrylamide, acrylic acid, and 3-hydroxypropanoate, the average yield was higher, only when compared to other targets on xylose. 1-propanol, isopropanol, and 1,3-propanediol yield averages were higher under ECOM/aerobic, and 1,4-butanediol under wild type/anaerobic using glycerol as a substrate. Specifications regarding the number of predicted strain designs and average yield for each pathway are shown in Supplementary Table 4. As expected, depending on the pathway precursor, intermediates, stoichiometry, and cofactors involved, specific combinations for oxygenation/substrate/knock-out lead to different productivities. As shown in Fig. 5, for each target compound, most of the production potentials were under the maximum theoretical yield average. This behavior is due to the resulting strain designs being predicted as heterofermentative strains and also because some knock-outs significantly constrain the production potential. Still, some promising strain designs with growth-coupled yield between the average theoretical yield plus standard deviation and the highest theoretical yield values were found *in silico*. The pathways are outlined in Supplementary Fig. 6 and strain designs were specified in Supplementary Table 2. Specifically, pathways for the production of 1,3-propanediol, 1-butanol, 3-hydroxypropanoate, acrylic acid, and acrylamide were identified. Analysis on the potential experimental implementation is shown in Supplementary text.

In order to compare and determine the growth coupled production potential for the novel pathway calculated using GEM-Path and the already reported pathways (computationally or experimentally), an analysis was performed by calculating the ratio between the highest growth-coupled production for the novel pathways predicted by GEM-Path over the experimentally or computationally reported pathways. Results were displayed under aerobic, ECOM, and anaerobic conditions, by using glucose, xylose, and glycerol as a carbon source. Values over 1 indicated that GEM-Path's novel pathways have higher growth-coupled production potential than already referenced pathways

Table 2
Targeted compounds and theoretical maximum yield analysis.

Target compound	Native or non-native	No. of carbons	No. of computed pathways	No. of unique reactions in each pathway	Aerobic (C-mol yield/pathway ID)			Aerobic ECOM (C-mol yield/ pathway ID)			Anaerobic (C-mol yield/pathway ID)		
					Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol
1,4-butanediol	Non-native	4	44	21	70/13	69/13	82/13	70/13	66/13	78/13	70/13	66/13	78/14
1,3-propanediol	Non-native	3	41	26	69/16	69/14	84/7	57/14	54/16	79/34	57/16	52/16	79/7
3-hydroxypropanoate	Native	3	25	33	96/12	96/13	97/3	96/12	96/13	97/12	96/12	96/13	71/12
Acrilic acid	Non-native	3	23	33	96/2	96/2	97/1	96/2	96/2	97/2	96/1	96/2	71/1
1-propanol	Non-native	3	19	12	64/3	64/2	75/3	64/1	64/2	75/1	64/3	64/4	75/1
1-butanol	Non-native	4	18	11	64/5	64/15	75/10	64/4	64/4	75/10	64/11	64/5	75/5
Isopropanol	Non-native	3	14	8	63/4	62/4	72/1	61/3	59/3	63/4	58 /3	56/3	54/1
Isobutanol	Non-native	4	13	16	64/2	64/1	74/1	64/2	64/1	72/1	64/1	64/1	66/1
Acrylamide	Non-native	3	10	15	96/3	96/3	97/3	96/3	96/3	97/3	96/3	96/2	71/2
3-hydroxybutyrate	Non-native	4	9	8	83/5	82/1	93/5	79/2	77/1	82/1	73/2	68/1	49/1
2-phenylethanol	Non-native	8	5	7	73/9	73/9	83/9	47/1	44/2	50/6	36/5	31/1	36/6
2-keto-butanoic acid	Native	4	5	6	94/1	93/1	97/1	84/1	78/1	80/1	84/1	78/1	69/1
2,3-propanediol	Non-native	3	4	4	68/1	68/1	79/2	55/1	51/1	56/2	55/1	51/1	40/1
2-keto-isovaleric acid	Native	5	3	3	86/3	85/3	97/3	84/3	82/3	90/3	79/3	69/3	49/3
3-hydroxyvalerate	Non-native	5	3	6	75/2	74/2	85/2	54/2	51/2	49/2	44/2	38/2	27/2
2,3-butanediol	Non-native	4	2	3	70/1	69/1	79/1	70/1	68/1	74/1	69/1	66/1	53/1
3-methyl-1-butanol	Non-native	5	2	3	62/1	61/1	70/1	58/1	57/1	60/1	49/1	43/1	53/1
2 -methyl-1-butanol	Non-native	5	2	3	60/1	59/1	70/1	42/1	39/1	50/1	36/1	31/1	36/1
4-hydroxybutyrate	Native	4	2	2	78/1	77/1	88/1	64/1	58/1	45/1	44/1	39/1	26/1
2-keto-valeric acid	Non-native	5	1	1	88/1	88/1	97/1	88/1	87/1	92/1	88/1	87/1	70/1

For each target compound, maximum theoretical yields (C-mol) were reported for different strain and substrate conditions. Shown next to the yield is the corresponding pathway ID shown in Supplementary pathways.

(Table 4). Specifically, for 3-hydroxypropanoate, 1-propanol, and 1-butanol, considerable improvements were found on various substrates and oxygenation conditions. Furthermore, for 1,4-butanediol, isopropanol, isobutanol, and 3-hydroxybutarte, only growth-coupled designs associated to novel pathways from GEM-Path were found.

3.7. GEM-Path output example

3.7.1. Case I: production of 1,3-propanediol

A output example for the production of 1,3-propanediol using GEM-Path is shown in Fig. 6. Two different GEM-Path calculated pathways were outlined: Pathway #7 (reactions 6 and 3) that has already been experimentally implemented, and pathway #16 (reactions 17, 12, 16, and 3). For pathway #16, specific output relating to the existence of catalyzing reactions from BRENDA and the promiscuity analysis are shown. For reaction 3, 6, and 17, exact matches in the BRENDA database were found, sharing identical cofactors, substrates, and products. This is represented by a substrate TC equal to 0 during the search. Furthermore, the species and EC number were reported (Ishikura et al., 2005; Kajiura et al., 2007; Wang et al., 2012). It is worthwhile to note that for homologous enzymes, there was no ranking in terms of species shown to carry out a given reaction. The algorithm reports only the first hit, associated with the corresponding species and the predicted reaction. For experimental purposes, it may be necessary to use the predicted EC Number and a database such as BRENDA to find multiple possible enzyme options that can be evaluated. The concept of sorting in terms of species distances between the host and the species associated with the predicted reaction was not considered. Attempts have been made to mine content for homologous expression in *E. coli* (Bayer et al., 2009), but this is an active field of research. According to the iterative algorithm shown in Fig. 1, no exact substrate structural matches were found for reactions 12 and 16. Instead, a promiscuity analysis was performed, obtaining the corresponding reactions from BRENDA. The promiscuity space is represented by a multi-dimensional space, obtained from the multi-linear regression analysis. For simplicity, and in order to describe the promiscuity analysis output,, a two dimensional space was outlined in Fig. 6 and the native BRENDA reaction substrate's (blue circles) distances from the centroid (red circle) were normalized to 1, and the tested substrate (green circle) was outlined at a distance equal to the promiscuity score. For reaction 12 and 16, a PS equal to 1.1 and 0.69 was calculated, respectively. The BRENDA predicted reaction was reported showing the substrate difference in terms of the TC, the corresponding species, and EC number (Furuyoshi et al., 1991). In order to avoid extensive computation, the algorithm chooses and saves the first possible solution for the particular species and related promiscuity score, leaving behind a number of additional feasible solutions able to fulfill the PS threshold. Reaction 17 represents a reaction gap filled by the algorithm. As shown in Fig. 6, the only reaction that connects the D-malate (mal-D) metabolite to *E. coli* metabolism are transport reactions through the periplasmic and external membrane (iMALDt2_2pp and MALDtex, respectively). Since mal-D was not set as a media constituent, there was no option for it to be generated by the network. By calculating reaction 17, it was feasible to connect a heterologous pathway to central carbon metabolism, specifically to oxaloacetate (oaa). Note that *E. coli* does contain malate dehydrogenase (MDH) which reversibly converts L-malate to oaa, but it is not implied that it can convert it to D-malate (Sutherland and McAlister-Henn, 1985).

Following synthetic calculation of heterologous pathways for each target compound, strain design computations were performed to engineer host cell metabolism. Continuing with the current example for 1,3-propandiol production, production

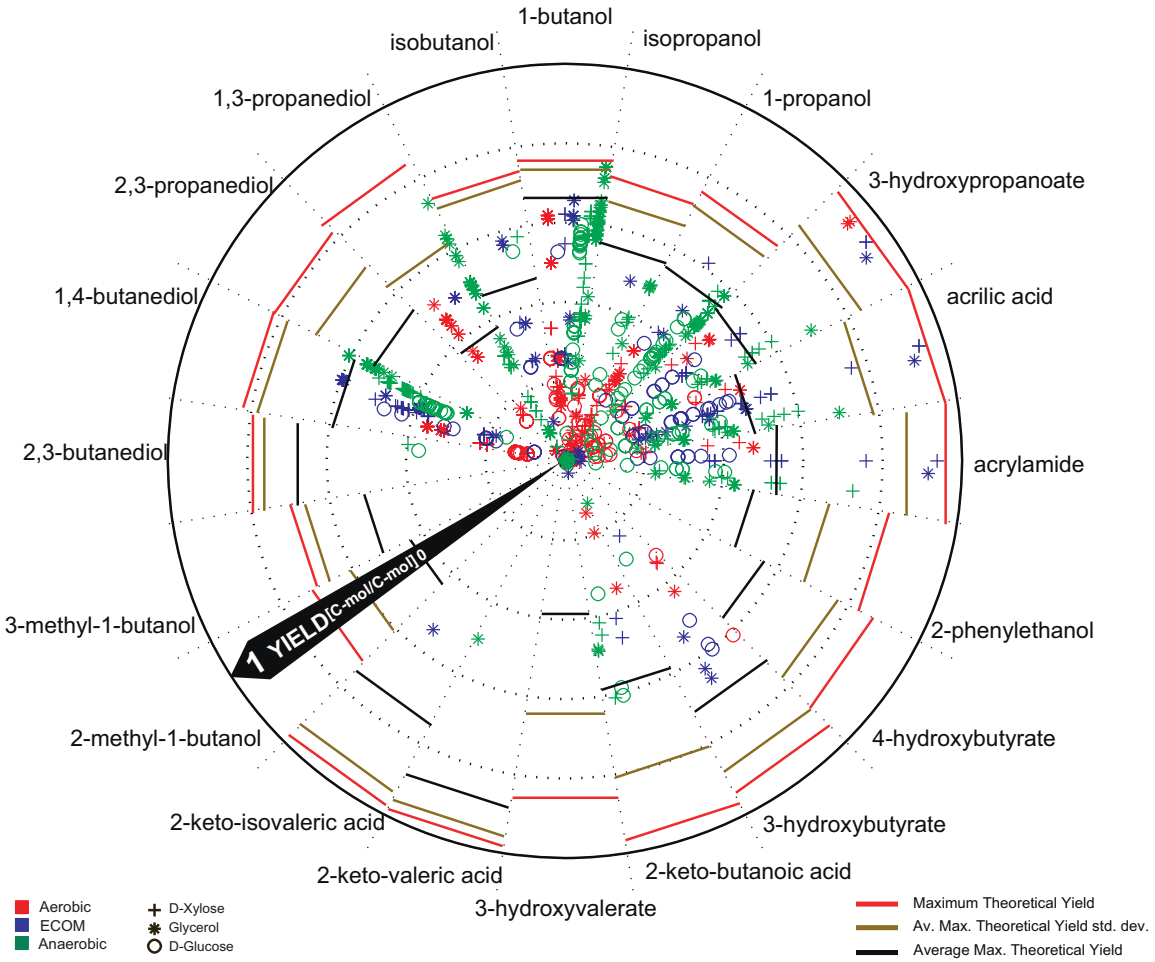

Fig. 5. Analysis of predicted yields for identified strain designs: the circumference of the plot is divided into 20 different segments, each corresponding to a specific target compound. The yield is represented along the radius, where the center and the perimeter represent C-mol yields equal to 0 and 1, respectively. Two different kinds of results are plotted in this diagram. First, the theoretical maximal growth-coupled yield for different knock-out and strain/substrate combinations were plotted for each target compound. Colored points represent the strain condition for wild type/aerobic (red), ECOM/aerobic (blue), and wild type/anaerobic (green). The shape defines a specific substrate use for xylose (+), glycerol (*), and glucose (o). The second set of results corresponds to the average maximal theoretical yield (black line) for each compound (each compound can have multiple predicted pathways) with the corresponding standard deviation (brown line) added to this mean. These values were calculated from the theoretical yield analysis, where all the simulations regarding strain/substrate conditions were taken into account. Finally, the highest maximal theoretical yield value is represented by the red line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

envelopes of growth-coupled designs for pathways #7 and #16 were outlined in Fig. 8a. Furthermore, a productivity analysis under different conditions was performed (Fig. 8b), where shaded areas represent the maximum theoretical production rate by setting the computational minimal growth rate to 0.1 h^{-1}, and solid areas represent the maximum growth-coupled production rate. As mentioned before, the overall trend shows that under aerobic conditions, maximum theoretical production is higher flux compared to anaerobic. Moreover, by using glycerol as a substrate instead of glucose, higher productivities were calculated for both aerobic and anaerobic conditions. Specifically, when comparing the maximum theoretical production for pathways #16 and #7 under aerobic conditions, an increase of 17% and 25% was observed, and under anaerobic conditions a 6% and 67% increase was observed over glucose, respectively.

Depending on the inserted heterologous pathway, different flux distributions were calculated. For pathways #16 and #7, the flux solution ratio for maximum theoretical production when using glycerol over glucose as substrates was calculated and qualitatively outlined (see Supplementary Fig. 7). For both pathways under aerobic conditions, there was a decrease in the carbon dioxide evolution when using glycerol as a substrate (approximately 50% less carbon dioxide was produced). Looking at the flux distributions, for pathway #16, no activity was predicted for the pentose phosphate pathway (PPP) when using glycerol as a substrate, and a higher target production rate when using glycerol was observed. This was due to the glycerol uptake metabolism, which is able to produce nadh and nadph similar to the PPP, but without generating carbon dioxide in the process, leading to more efficient carbon metabolism. For anaerobic conditions, a similar trend and a mixed acid fermentation behavior was observed. By comparing the maximum theoretical production for both pathways under the same conditions (same substrate and oxygenation), pathway #16 is able to achieve higher productivity by using glucose as a substrate, approximately 4% and 22% more under aerobic and anaerobic conditions, respectively. Still, by using glycerol as a carbon source, the productivity decreases approximately, 3% and 22% less under aerobic and anaerobic conditions, respectively. This result demonstrated that the novel GEM-Path predicted pathways #16 is more suited to implement linked to a glucose based fermentation process. According to Fig. 8a and b, growth-coupled designs were only found for glycerol under both anaerobic and aerobic conditions. No growth-coupled designs associated with glucose under anaerobic conditions were found, and under aerobic conditions, only a low productivity growth coupled design for pathway #16 was found.

Table 3
Predicted yields for growth-coupled strain designs by production interval and product.

	Total no. of strain designs	Oxygenation No. of designs/avg. C-mol yield			Substrate No. of designs/avg. C-mol yield			Knock outs No. of designs/avg. C-mol yield		
		Aerobic	ECOM	Anaerobic	Glucose	Xylose	Glycerol	2 KO	3 KO	4 KO
Yield interval for growth-coupled designs										
0.8–1.0	10	2/0.93	8/0.93	0/0	0/0	5/0.94	5/0.92	0/0	8/0.93	2/0.94
0.6–0.8	89	17/0.61	16/0.64	56/0.64	3/0.61	17/0.64	69/0.64	15/0.65	43/0.64	31/0.63
0.4–0.6	342	35/0.47	70/0.49	237/0.50	38/0.51	121 /0.48	183/0.51	105/0.5	146/0.50	91/0.48
0.2–0.4	461	123/0.29	130/0.32	208/0.32	169/0.31	138/0.31	154/0.32	133/0.31	185/0.31	143/0.32
0–0.2	369	248/0.1	59/0.05	62/0.08	149/0.1	121/0.10	99/0.06	147/0.07	137/0.11	85/0.09
Overall	1271	425/0.21	283/0.34	563/0.4	359/0.25	402/0.32	510/0.39	400/0.29	519/0.35	352/0.34
Percentage	19%	33%	22%	44%	28%	32%	40%	31%	41%	28%
Target compound										
Acrylamide	77	23/0.13	26/0.31	28/0.33	29/0.20	25/0.39	23/0.21	27/0.22	31/0.30	19/0.28
Acrilic acid	152	38/0.13	67/0.30	47/0.31	56/0.21	49/0.34	47/0.24	44/0.22	65/0.28	43/0.27
3-hydroxypropanoate	110	34/0.19	36/0.28	40/0.32	38/0.18	36/0.36	36/0.26	35/0.22	45/0.31	30/0.26
1-propanol	143	55/0.15	6/0.41	82/0.39	41/0.25	47/0.25	55/0.38	63/0.26	50/0.33	30/0.36
Isopropanol	34	18/0.11	1/0.48	15/0.39	10/0.19	11/0.22	13/0.30	12/0.19	12/0.26	10/0.28
1-butanol	211	62/0.35	25/0.44	124/0.49	53/0.36	56/0.39	102/0.51	76/0.41	96/0.43	39/0.52
Isobutanol	23	1/0.01	17/0.32	5/0.23	6/0.28	8/0.22	9/0.35	1/0.00	18/0.36	4/0.03
1,3-propanediol	106	40/0.26	3/0.50	63/0.35	4/0.14	30/0.20	72/0.38	31/0.30	41/0.37	34/0.28
2,3-propanediol	0	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00
1,4-butanediol	375	144/0.21	88/0.35	143/0.44	110/0.23	129/0.31	136/0.43	100/0.30	145/0.34	130/0.34
2,3-butanediol	2	0/0.00	0/0.00	2/0.38	1/0.37	1/0.40	0/0.00	0/0.00	2/0.38	0/0.00
3-methyl-1-butanol	0	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00
2-methyl-1-butanol	0	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00
2-keto-isovaleric acid	1	0/0.00	1/0.54	0/0.00	0/0.00	0/0.00	1/0.54	0/0.00	0/0.00	1/0.54
2-keto-valeric acid	1	0/0.00	0/0.00	1/0.50	0/0.00	0/0.00	1/0.50	0/0.00	1/0.50	0/0.00
3-hydroxyvalerate	2	0/0.00	2/0.00	0/0.00	0/0.00	0/0.00	2/0.00	1/0.00	0/0.00	1/0.00
2-keto-butanoic acid	17	5/0.14	3/0.31	9/0.50	4/0.39	6/0.39	7/0.31	4/0.36	7/0.34	6/0.38
3-hydroxybutyrate	17	5/0.41	8/0.53	4/0.11	7/0.42	4/0.36	6/0.40	6/0.30	6/0.43	5/0.46
4-hydroxybutyrate	0	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00
2-phenylethanol	0	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00	0/0.00

The number of strain designs and the corresponding average yield are separated by "/". Results were tabulated for each yield interval under different oxygenation/substrate/knock-out conditions. Overall values were added at the bottom. The number of strain designs and the corresponding average yield are separated by "/". Results were tabulated for target compounds under different oxygenation/substrate/knock-out conditions. Furthermore, the number of predicted strain designs and the number of growth-coupled pathway for each target compound were tabulated.

Table 4
Comparison of GEM-Path growth coupled design to previously identified pathways from literature.

Target compound	Aerobic			ECOM			Anaerobic		
	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol	Glucose	Xylose	Glycerol
1,4-butanediol	0.13*	0.22*	0.36*	0.47*	0.50*	0.60*	0.40*	0.51*	0.60*
1,3-propanediol	0.14*	0.16*	1,15^{E}	—	—	1,00^{E}	—	1,12^{E}	0,86^{E}
3-hydroxypropanoate	1,76^{C}	2,16^{C}	0,51^{E}	1,21^{C}	2,48^{C}	22,5^{C}	1,67^{C}	1,46^{C}	2,29^{C}
1-propanol	1,85^{E}	0,78^{E}	0,79^{E}	0.36*	0,65^{E}	0.48*	0,82^{E}	0,68^{E}	0,85^{E}
1-butanol	1,34^{E}	1,57^{E}	2,32^{E}	2,03$^{E/C}$	2,31$^{E/C}$	1,06^{E}	1,05$^{E/C}$	1,03^{E}	1,16^{E}
Isopropanol	0.18*	0.17*	0.21*	—	—	0.48*	0.38*	0.40*	0.50*
Isobutanol	—	—	0.01*	0.35*	0.37*	0.58*	0.54$^{\dagger E/C}$	0,01$^{E/C}$	—
3-hydroxybutyrate	0.61*	0.35*	0.43*	0.60*	0.52*	0.66*	0.30*	—	0.12*
2-phenylethanol	—	—	—	—	—	—	—	—	—
2,3-propanediol	—	—	—	—	—	—	—	—	—
2,3-butanediol	—	—	—	—	—	—	0.37$^{\dagger E}$	0.40$^{\dagger E}$	—
3-methyl-1-butanol	—	—	—	—	—	—	—	—	—
2-methyl-1-butanol	—	—	—	—	—	—	—	—	—
4-hydroxybutyrate	—	—	—	—	—	—	—	—	—

For each target compound, the growth-coupled ratio between novel pathways generated by GEM-Path and experimentally implemented (E) and/or computationally generated pathways (C) are shown. Empty spaces (—) indicate that no referenced pathways for the corresponding target compound were found.

* No experimentally implemented nor previous computationally predicted pathways were able to growth couple the target compound production. Maximum growth-coupled yield associated with new pathway predicted by GEM-Path is reported.

† Only experimentally or previous computationally predicted pathway were able to growth couple the target compound production. Maximum growth-coupled yield is reported.

3.7.2. *Case-study II: production of isopropanol*

GEM-Path predicted pathways for the production of isopropanol are shown in Fig. 7. For the production of isopropanol, two different GEM-Path calculated pathways were outlined: Pathway #13 (reactions 7, 8, and 1) which has been experimentally implemented, and Pathway #1 (reactions 1 and 3). For Pathway #1, specific output relating to a promiscuity analysis is shown. For reaction 1, an exact match in the BRENDA database was found,

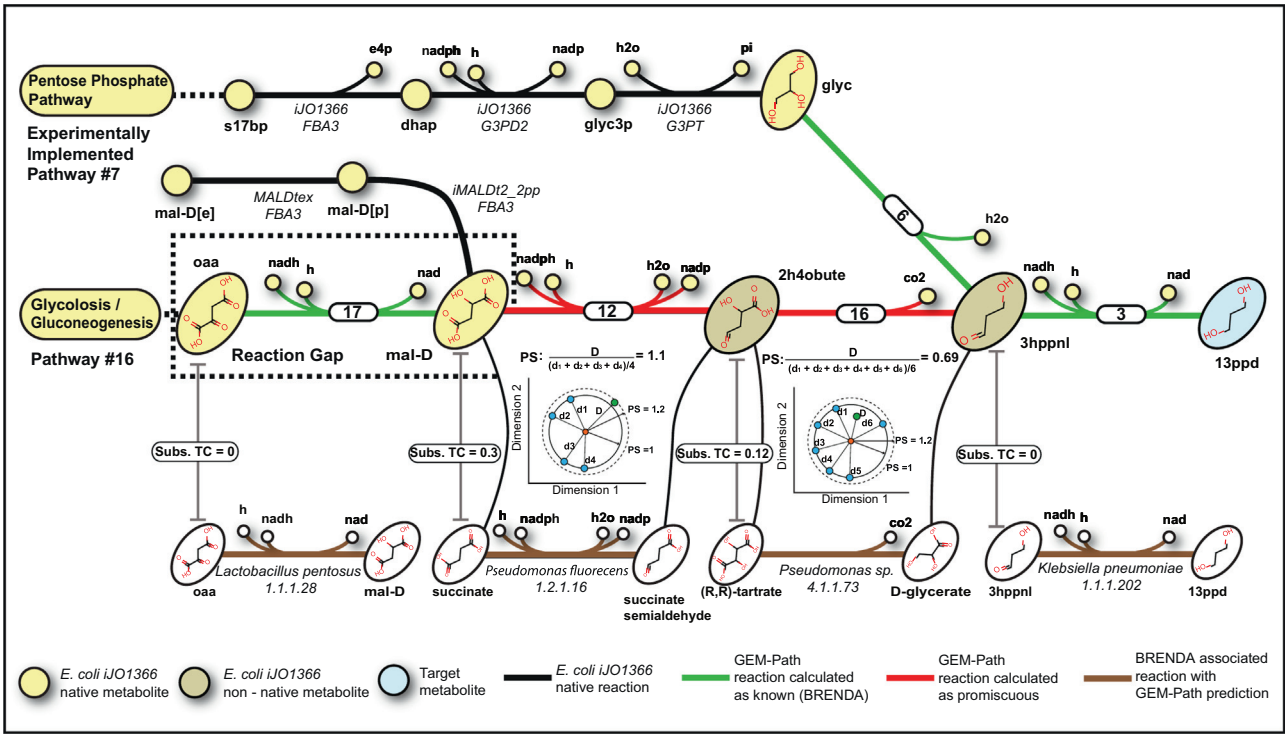

Fig. 6. GEM integrated synthetic pathway calculation (GEM-Path) output for 1,3-propanediol: two different GEM-Path calculated pathways are shown: pathway #7 (top, reactions 6 and 3) which has been experimentally implemented, and pathway #16 (bottom, reactions 17, 12, 16, and 3). For each pathway, reactions leading from the host metabolome are shown in black. Native and non-native *E. coli* metabolites are represented in yellow and brown, respectively. The corresponding target compound is shown in light blue. Reactions calculated as known in BRENDA and reaction calculated as promiscuous are shown in green and red, respectively. For each predicted reaction in pathway #16, specific values of the tanimoto coefficient (TC), promiscuity score (PS) and the corresponding BRENDA reaction (brown lines) are shown. For reactions predicted as promiscuous, the corresponding promiscuity space was outlined with the number of metabolites associated with the specific reaction found in BRENDA. For simplicity, a two dimensional space was plotted, where each of the native BRENDA metabolites (green) are separated from the centroid (red circle) in 1 dimensionless unit and the predicted metabolite is shown in green, with the corresponding distance equal to the PS. A promiscuity score threshold was plotted at a distance equal to 2. Each BRENDA reaction shows the corresponding associated EC number and a species known to catalyze the specific reaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

sharing the same cofactors, substrates, and products. This is represented by a substrate TC equal to 0 during the search. Furthermore, a species known to carry out this reaction and EC number were reported (Drewke and Ciriacy, 1988). Although the reaction was reported to proceed in the reverse direction, no evidence showing reaction irreversibility was found. Conversely, no exact substrate structural match was found for reaction 3. Thus, a promiscuity analysis was performed, obtaining the corresponding reaction from BRENDA. The promiscuity space is represented by 11 different native substrates ($n=11$). In order to describe the promiscuity analysis output and for simplicity, a two dimensional space was outlined (as described above). According to the results, a PS equal to 0.88 was calculated. The BRENDA predicted reaction was reported showing the substrate difference in terms of the TC along with a corresponding species and EC number. For the other pathway (i.e. #13), reaction 7 represents a reaction gap filled by the algorithm. As shown in Fig. 7, two different reactions connect acetoacetate (acac) to other metabolites in the network. The first reaction is ACACCT, which is an irreversible reaction on pathway #13 opposite the isopropanol production direction, and the second reaction is a transport reaction. Since acac was not set as a media constituent, there was no option for it to be generated by the network. By calculating Reaction 7, it was feasible to connect the heterologous pathway to central carbon metabolism, specifically to acetyl-coa (accoa). It should be noted that there is experimental evidence for the existence of a reversible ACACT1r reaction (Fujii et al., 2010; Gulevich et al., 2012), but there is also contradictory evidence indicating that operation in this direction could be highly unfavorable (Lan and Liao, 2012; McCloskey et al., 2014). Nonetheless, the GEM-Path algorithm uses the content as defined in the model (Orth et al., 2011) and curation is a helpful step after promising production pathways are identified. Beyond identifying pathways, growth coupled-designs utilizing pathways #1 and #13 were outlined in Fig. 8b. Furthermore, a max theoretical production analysis under different conditions was performed (Fig. 8d), where shaded areas represent the maximum theoretical production by setting the computational growth rate to 0.1 h^{-1}, and solid areas show the maximum growth-coupled productivity. As mention before, the overall trend shows that under aerobic conditions, pathways are capable of carrying higher theoretical flux as compared to anaerobic. Moreover, by using glycerol as a substrate instead of glucose, higher productivities were calculated for aerobic conditions. For anaerobic conditions and using glycerol as a substrate, only pathway #1 was able to achieve higher flux compared to glucose.

A maximum theoretical production analysis for isopropanol revealed differences in production potential when using glycerol or glucose as a substrate. For both pathways under aerobic conditions, a decrease in carbon dioxide evolution was observed when using glycerol as a substrate. For both pathways, approximately 25% less carbon dioxide was produced. Looking at the flux distribution (see Supplementary Fig. 8), for pathway #1 and #13, no activity in the PPP during glycerol consumption was observed, due to the same reason described in the first 1,3-propanediol case study. For anaerobic conditions, specifically for pathway #1, a mix acid fermentation behavior was observed. Higher productivity was observed compared to glycerol for pathway #13 under anaerobic conditions and using glucose as a substrate. Byproduct formation

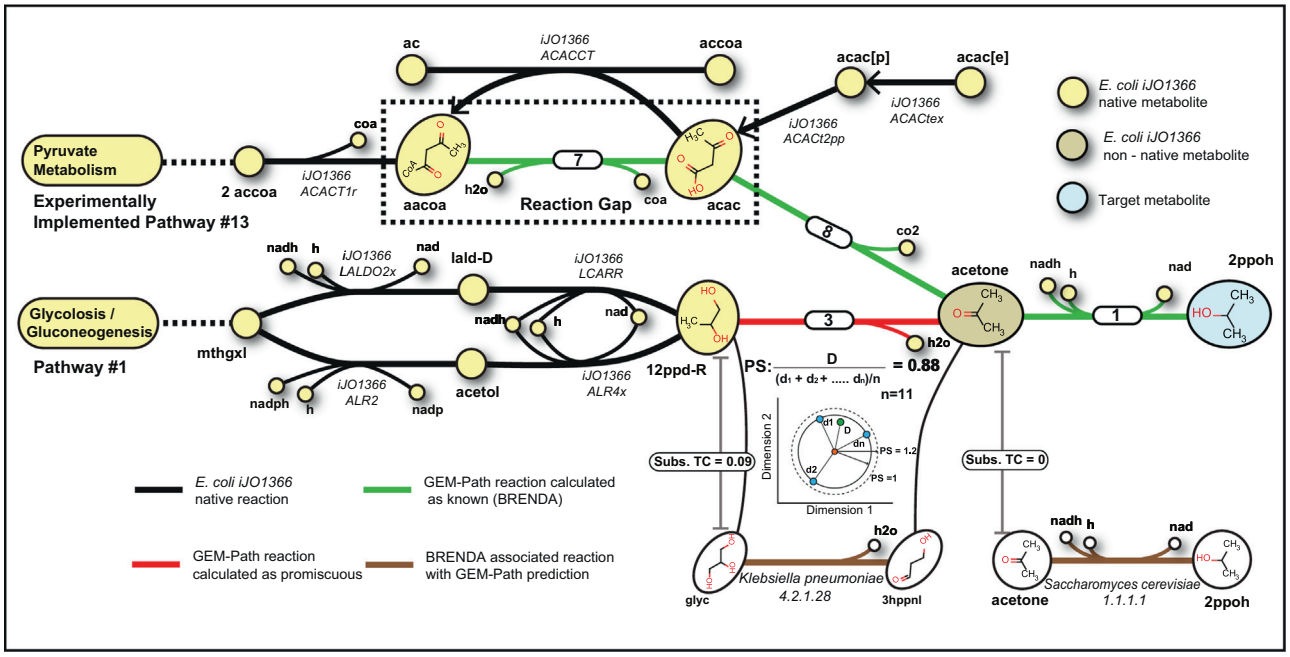

Fig. 7. GEM integrated synthetic pathway calculation (GEM-Path) output for GEM integrated synthetic pathway calculation (GEM-Path) output for isopropanol: two different GEM-Path calculated pathways are shown. Pathway #13 (reactions 7, 8, and 1), that has been experimentally implemented, and pathway #1 (reactions 1 and 3). For each pathway, reactions leading to the host cell metabolome are shown in black. Native and non-native *E. coli* metabolites are represented in yellow and brown, respectively. The corresponding target compound is shown in light blue. Reactions calculated as known in the BRENDA database and reactions calculated as promiscuous are shown in green and red, respectively. For each predicted reaction in pathway #1, specific values of the tanimoto coefficient (TC), promiscuity score (PS), and the corresponding BRENDA reaction (brown) are specified. For reactions predicted as promiscuous, the corresponding promiscuity space was outline with the number of metabolites associated with the specific reaction found in BRENDA. For simplicity, a two dimensional space was plot, where each of the native BRENDA metabolites (green) are separated from the centroid (red circle) in 1 dimensionless units and the predicted metabolite is shown in green, with the corresponding distance equal to the PS. A promiscuity scores threshold was plotted at a distance equal to 2. Each BRENDA reaction shows the corresponding associated EC number and species known to carry out the reaction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

during glycerol growth was critical in diminishing the productivity. By comparing the pathway's maximum theoretical production under the same conditions (same substrate and oxygenation), pathway #1 is able to achieve higher productivity by using glucose as a substrate, approximately 1% and 7% more under aerobic and anaerobic conditions, respectively. When utilizing glycerol as a carbon source, the productivity increases approximately 6% and 94% under anaerobic conditions, respectively. This result demonstrated that the novel GEM-Path predicted pathway #1 shows a higher theoretical potential when using a glycerol based fermentation process. As indicated in Fig. 8c and d, growth-coupled designs were only found for pathway #1. Growth-coupled productivities similar to the maximum theoretical achievable productivity were found using both glycerol and glucose as a substrate and under anaerobic conditions.

4. Discussion

The aim of this work was to outline the production potential for 20 industrially-relevant chemicals in *E. coli* and generate feasible designs for production strains. The enabling technology generated for the project was a computational pipeline including chemoinformatics, bioinformatics, constraint-based modeling, and GEMs to aid in the process of metabolic engineering of microbes for industrial bioprocessing purposes. The main results from this study are, (i) a comprehensive mapping from *E. coli's* native metabolome to commodity chemicals that are 4 reactions or less away from a natural metabolite, (ii) sets of metabolic interventions, specifically knock-outs and knock-ins, that coupled the target chemical production to growth rate, (iii) the development of a retrosynthetic based pathway predictor algorithm containing a novel integration with GEMs and reaction promiscuity analysis, and (iv) a complete strain design workflow integrating synthetic pathway prediction with growth-coupled designs for the production of non-native compounds in a target organism of interest.

For synthetic pathway predictions, considerable attention has been focused on retrosynthetic algorithms, where a backward search for synthetic pathways is performed by an iterative application of Biochemical Reaction Operators (BROs) from a target compound to a predefined source of metabolites (Medema et al., 2012). Based on 443 BROs included in this work, a retrosynthetic pathway predictor algorithm was developed which incorporates GEMs into the procedure. The GEM-Path algorithm is also coupled together with database analysis for reaction existence and reaction promiscuity inference. Predictions were compared to literature, and showed a good agreement with previously reported algorithms. Due to the filtering procedure at each iteration step, specifically the promiscuity analysis, the number of generated pathways was considerably lower as compared with previous algorithms, diminishing the candidates required for further experimental implementation. In total, GEM-Path generated 245 synthetic pathways for the production of 20 different compounds in *E. coli.* The majority of the predicted pathways involved at least one promiscuous reaction. Since the promiscuity analysis is based on E.C. reaction numbers instead of genes, an enzymatic validation step may be necessary to confirm the predicted functionality before introduction into a production host.

Theoretically, all synthetic pathways identified in this work are able to produce the target compound under a given substrate/oxygenation/strain conditions and in total, they characterize the production space. Novel pathways able to achieve high yield were found for a range of commodity chemicals. According to the theoretical maximum yield analysis, pathways implemented

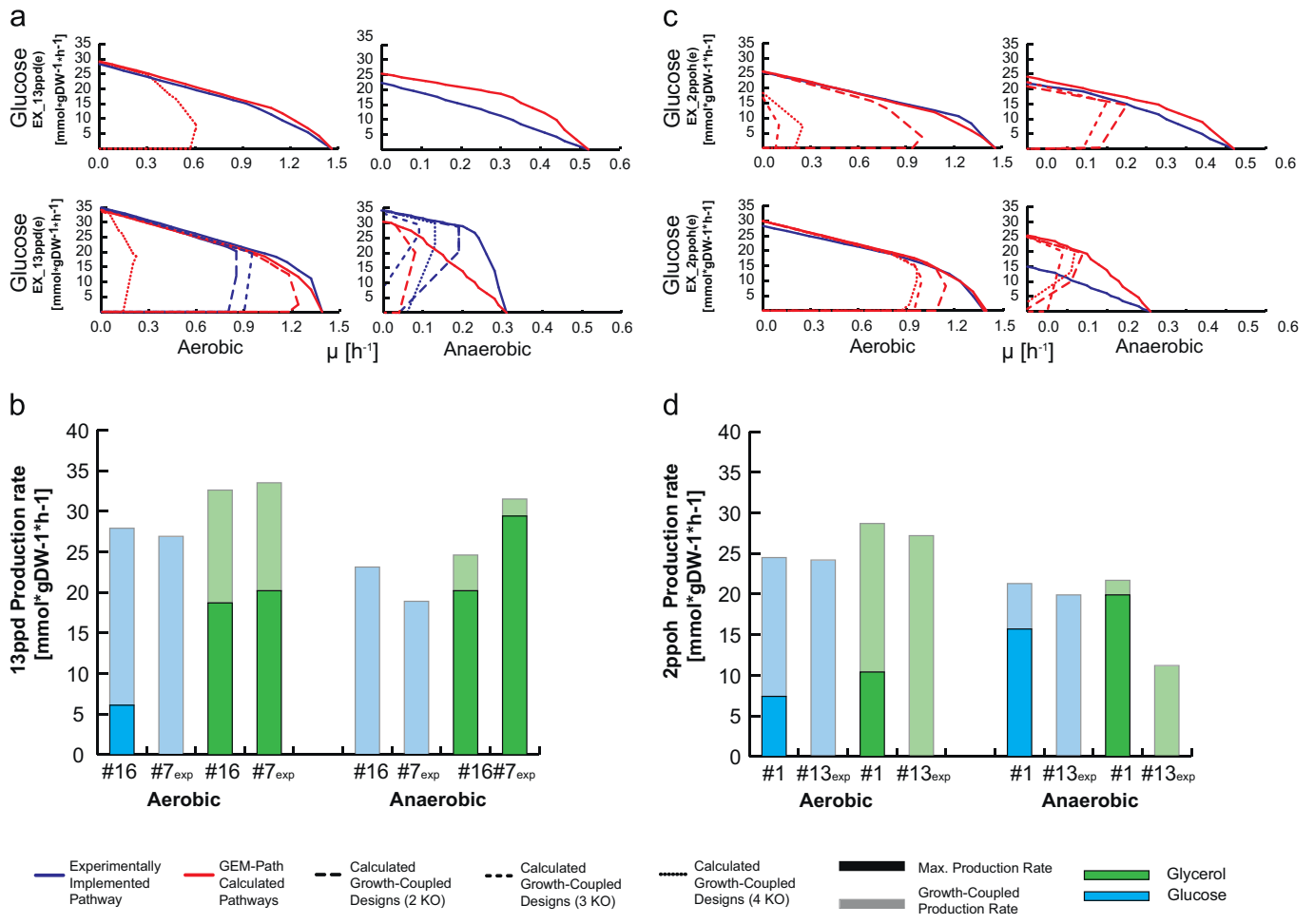

Fig. 8. Strain design productivity analysis for GEM-Path case studies: the production envelopes for strain designs of 2–4 gene knockouts were plot (a and c). Glucose and glycerol were used as substrate examples under aerobic and anaerobic conditions. Production envelopes for 1,3-propanediol (pathways #16 and #7) and ispopropanol (pathways #13 and #1) are shown in (a) and (c), respectively. Solid blue and red lines represent experimentally implemented pathways and novel GEM-Path calculated pathways, respectively. Production envelopes for growth-coupled designs are shown in dotted lines. Productivity analysis for the production of 1,3-propanediol (b) and isopropanol (d) were outlined. Results were grouped for aerobic and anaerobic conditions, associated with the corresponding pathway number. By using glucose (blue) and glycerol (green) as substrates, maximum theoretical production rate (shaded bars) and growth-coupled production rate (filled bars) were plotted. FBA was used to determine the maximum theoretical productivity by setting the growth rate to 0.1 h^{-1} and optimizing for the target compound production. Growth-coupled productivity was calculated by knocking out computationally identified reactions and optimizing for growth rate. The maximum value for each condition was reported. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

under wild type aerobic conditions tend to have a greater production potential compared to the other strain/oxygenation conditions. Furthermore, when changing the anaerobiosis threshold, the more anaerobic the condition of a strain, the less overall production could be achieved. Lower maximum theoretical yields observed under anaerobic conditions (vs. aerobic) are expected, as no oxygen is essentially an additional constraint, limiting the capability of the network (just like the removal of a key reaction in the network). Based on a C-mol Yield basis, under wild-type/aerobic conditions, glycerol is found to be the most efficient substrate for heterologous target compound production. However, for wild-type/anaerobic, xylose and glucose are the most efficient substrates. In addition, precursor yield analysis reveals that pathways having precursors closest to the central metabolism are able to achieve higher yields which agrees with logic as central metabolic reactions carry the most flux in the network (Almaas et al., 2004).

Growth-coupled production of a specific metabolite depends on the energy benefit that the cell can obtain through the pathway activation related to the growth-coupled metabolite. Growth-coupled design algorithms operate by knocking out reactions, thus generating an energy imbalance that is recovered by then coupling different pathways to growth. The final metabolite involved in these pathways works as a final electron acceptor, thus, under anaerobic conditions, pathways are more susceptible to coupling to growth. The ability to find growth-coupled designs preferentially under anaerobic conditions can be seen by analyzing the overall results, where growth-coupled designs under wild-type anaerobic conditions were found to be present more frequently and were able to achieve higher yields. Further, designs with glycerol as a substrate had the highest yields anaerobically. Thus, under anaerobic conditions, growth-coupled designs are easier to obtain compared to aerobic conditions. Furthermore, for most of the predicted reactions contributing to a growth-coupled design, approximately 40%, were oxidoredutases with NAD or NADP acceptors. Removal of these reactions facilitates growth-coupling as they shift the flow of electrons in metabolism (King and Feist, 2013).

Designs highlighted in this work were selected according to their production potential (i.e., yield, $Y_{p/s}$). Nevertheless, further improvements are needed for the design and production workflow to promote success in experimental implementation. For instance, toxicity due to product or co-product formation was not evaluated during the design pipeline; this might lead to the production of toxic compounds together with cell death. Due to the scope of the

GEMs used in this work (i.e., metabolic GEMs), key regulatory steps were not taken into account. Furthermore, the impact of low substrate affinity of predicted promiscuous enzymes might lead to false positive results, decreasing the *in vivo* maximum achievable yield. However, to generate non-native products, it is obvious that new production pathways are necessary and thus that was the focus of this work. Furthermore, growth coupled designs, such as those produced here, provide an extra tool for metabolic engineers by allowing for the use of selection pressure to achieve a desired production state. For reactions predicted as promiscuous, *in vitro* enzyme analysis might be necessary to identify and characterize the potential promiscuous activity. Moreover, in order to avoid undesirable metabolite sinks, a promiscuity analysis regarding native metabolites must be taken into account when reactions are incorporated into metabolism. Lastly, as in any production strain project, enzyme efficiency issues and heterologous codon optimization (Medema et al., 2012) must also be considered for product formation.

Taken together, the workflow presented here finds that the 20 major commodity chemicals are within 4 reactions from the metabolic network of *E. coli*. Further, it maps out all the feasible pathways linking the chemical structures of these commodity chemicals to the metabolic network of *E. coli* and their theoretical yields. It also maps out the chemical reactions and enzymatic requirements for building these pathways. Thus, in a way, we have generated a pathway atlas that can guide the global metabolic engineering and strain design efforts needed to convert the petroleum-based industry to a biomass-based industry, and thus forms the basis for a grand challenge undertaken by the community.

Acknowledgments

We would like to thank for Karsten Zengler, Joshua Lerman, Nikolaus Sonnenschein, Zachary King and Daniel Zielinski for their input and feedback on the project. Funding for this work was provided by the Novo Nordisk Foundation. Also we would like to thank MCESESUP2: Doctoral Scholarship for study abroad, the Conicyt Basal Centre Grant for the CeBiB FB0001 and Project UCH0717 National Doctoral Scholarship, Chile.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.ymben.2014.07.009.

References

Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N., Barabasi, A.L., 2004. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. Nature 427, 839–843.
Altaras, N.E., Cameron, D.C., 1999. Metabolic engineering of a 1,2-propanediol pathway in *Escherichia coli*. Appl. Environ. Microbiol. 65, 1180–1185.
Arita, M., 2000. Metabolic reconstruction using shortest paths. Simul. Pract. Theory 8, 109–125.
Arundel, A., Sawaya, D., 2009. The bioeconomy to 2030: Designing a policy agenda.
Assary, R.S., Broadbelt, L.J., 2011. 2-Keto Acids to branched-chain alcohols as biofuels: application of reaction network analysis and high-level quantum chemical methods to understand thermodynamic landscapes. Comput. Theor. Chem. 978 (1), 160–165.
Atsumi, S., Cann, A.F., Connor, M.R., Shen, C.R., Smith, K.M., Brynildsen, M.P., Chou, K.J., Hanai, T., Liao, J.C., 2008. Metabolic engineering of Escherichia coli for 1-butanol production. Metab. Eng. 10, 305–311.
Atsumi, S., Liao, J.C., 2008. Metabolic engineering for advanced biofuels production from *Escherichia coli*. Curr. Opin. Biotechnol. 19, 414–419.
Atsumi, S., Wu, T.Y., Eckl, E.M., Hawkins, S.D., Buelter, T., Liao, J.C., 2010. Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. Appl. Microbiol. Biotechnol. 85, 651–657.
Bar-Even, A., Noor, E., Flamholz, A., Buescher, J.M., Milo, R., 2011. Hydrophobicity and charge shape cellular metabolite concentrations. PLoS Comput. Biol. 7, e1002166.
Bayer, T.S., Widmaier, D.M., Temme, K., Mirsky, E.A., Santi, D.V., Voigt, C.A., 2009. Synthesis of methyl halides from biomass using engineered microbes. J. Am. Chem. Soc. 131, 6508–6515.
Bennett, B.D., Kimball, E.H., Gao, M., Osterhout, R., Van Dien, S.J., Rabinowitz, J.D., 2009. Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. Nat. Chem. Biol. 5, 593–599.
Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H., 2008. PubChem: integrated platform of small molecules and biological activities. Annu. Rep. Comput. Chem. 4, 217–241.
Bordbar, A., Monk, J.M., King, Z.A., Palsson, B.O., 2014. Constraint-based models predict metabolic and associated cellular functions. Nat. Rev. Genet. 15, 107–120.
Bramucci, M. G., Flint, D., Miller, E. S., Nagarajan, V., Sedkova, N., Singh, M., Van Dyk, T. K., 2008. Method for the production of 1-butanol. U.S. Patent Application 2/110503.
Cann, A.F., Liao, J.C., 2008. Production of 2-methyl-1-butanol in engineered *Escherichia coli*. Appl. Microbiol. Biotechnol. 81, 89–98.
Carbonell, P., Planson, A.G., Fichera, D., Faulon, J.L., 2011. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. BMC Syst. Biol. 5, 122.
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.
Connor, M.R., Cann, A.F., Liao, J.C., 2010. 3-Methyl-1-butanol production in Escherichia coli: random mutagenesis and two-phase fermentation. Appl. Microbiol. Biotechnol. 86, 1155–1164.
Curran, K.A., Alper, H.S., 2012. Expanding the chemical palate of cells by combining systems biology and metabolic engineering. Metab. Eng. 14, 289–297.
Dalby, A., Nourse, J.G., Hounshell, W.D., Gushurst, A.K., Grier, D.L., Leland, B.A., Laufer, J., 1992. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J. Chem. Inf. Comput. Sci. 32, 244–255.
Dale, J.M., Popescu, L., Karp, P.D., 2010. Machine learning methods for metabolic pathway prediction. BMC Bioinform. 11, 15.
Drewke, C., Ciriacy, M., 1988. Overexpression, purification and properties of alcohol dehydrogenase IV from Saccharomyces cerevisiae. Biochim. Biophys. Acta. 950, 54–60.
Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., Palsson, B.O., 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3, 121.
Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L., Palsson, B.O., 2009. Reconstruction of biochemical networks in microorganisms. Nat. Rev. Microbiol. 7, 129–143.
Feist, A.M., Zielinski, D.C., Orth, J.D., Schellenberger, J., Herrgard, M.J., Palsson, B.O., 2010. Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. Metab. Eng. 12, 173–186.
Fischer, C.R., Klein-Marcuschamer, D., Stephanopoulos, G., 2008. Selection and optimization of microbial hosts for biofuels production. Metab. Eng. 10, 295–304.
Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D., Palsson, B.O., 2005. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. Biotechnol. Bioeng. 91, 643–648.
Fujii, T., Ito, K., Katsuma, S., Nakano, R., Shimada, T., Ishikawa, Y., 2010. Molecular and functional characterization of an acetyl-CoA acetyltransferase from the adzuki bean borer moth Ostrinia scapulalis (Lepidoptera: Crambidae). Insect Biochem. Mol. Biol. 40, 74–78.
Furuyoshi, S., Nawa, Y., Kawabata, N., Tanaka, H., Soda, K., 1991. Purification and characterization of a new NAD(+)-dependent enzyme, L-tartrate decarboxylase, from Pseudomonas sp. group Ve-2. J. Biochem. 110, 520–525.
Greene, N., Judson, P.N., Langowski, J.J., Marchant, C.A., 1999. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR QSAR Environ. Res. 10, 299–314.
Gulevich, A.Y., Skorokhodova, A.Y., Sukhozhenko, A.V., Shakulov, R.S., Debabov, V.G., 2012. Metabolic engineering of Escherichia coli for 1-butanol biosynthesis through the inverted aerobic fatty acid beta-oxidation pathway. Biotechnol. Lett. 34, 463–469.
Hanai, T., Atsumi, S., Liao, J.C., 2007. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. Appl. Environ. Microbiol. 73, 7814–7818.
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.
Heath, A.P., Bennett, G.N., Kavraki, L.E., 2010. Finding metabolic pathways using atom tracking. Bioinformatics 26, 1548–1555.
Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2007. Thermodynamics-based metabolic flux analysis. Biophys. J. 92, 1792–1805.
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.
Hou, B.K., Wackett, L.P., Ellis, L.B., 2003. Microbial pathway prediction: a functional group approach. J. Chem. Inf. Comput. Sci. 43, 1051–1057.
Hu, Q.N., Zhu, H., Li, X., Zhang, M., Deng, Z., Yang, X., 2012. Assignment of EC numbers to enzymatic reactions with reaction difference fingerprints. PLoS One 7, e52901.

Hwang, J.Y., Park, J., Seo, J.H., Cha, M., Cho, B.K., Kim, J., Kim, B.G., 2009. Simultaneous synthesis of 2-phenylethanol and L-homophenylalanine using aromatic transaminase with yeast Ehrlich pathway. Biotechnol. Bioeng. 102, 1323–1329.
Ibarra, R.U., Edwards, J.S., Palsson, B.O., 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature 420, 186–189.
Ishikura, Y., Tsuzuki, S., Takahashi, O., Tokuda, C., Nakanishi, R., Shinoda, T., Taguchi, H., 2005. Recognition site for the side chain of 2-ketoacid substrate in d-lactate dehydrogenase. J. Biochem. 138, 741–749.
James, C. A., Weininger, D., Delany, J., 2004. Daylight theory manual. Daylight Chemical Information Systems Inc.. 3951.
Jang, Y.S., Park, J.M., Choi, S., Choi, Y.J., do.Y, Seung, Cho, J.H., Sang, Y.L, 2012. Engineering of microorganisms for the production of biofuels and perspectives based on systems metabolic engineering approaches. Biotechnol. Adv. 30, 989–1000.
Jankowski, M.D., Henry, C.S., Broadbelt, L.J., Hatzimanikatis, V., 2008. Group contribution method for thermodynamic analysis of complex metabolic networks. Biophys. J. 95, 1487–1499.
Ji, X.J., Huang, H., Ouyang, P.K., 2011. Microbial 2,3-butanediol production: a state-of-the-art review. Biotechnol. Adv. 29, 351–364.
Jojima, T., Inui, M., Yukawa, H., 2008. Production of isopropanol by metabolically engineered *Escherichia coli*. Appl. Microbiol. Biotechnol. 77, 1219–1224.
Kajiura, H., Mori, K., Shibata, N., Toraya, T., 2007. Molecular basis for specificities of reactivating factors for adenosylcobalamin-dependent diol and glycerol dehydratases. FEBS J. 274, 5556–5566.
Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M., 2006. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–D357.
Keasling, J.D., 2012. Synthetic biology and the development of tools for metabolic engineering. Metab. Eng. 14, 189–195.
Kim, T.Y., Sohn, S.B., Kim, H.U., Lee, S.Y., 2008. Strategies for systems-level metabolic engineering. Biotechnol. J. 3, 612–623.
King, Z.A., Feist, A.M., 2013. Optimizing cofactor specificity of oxidoreductase enzymes for the generation of microbial production strains—optswap. Ind. Biotechnol. 9, 236–246.
Koma, D., Yamanaka, H., Moriyoshi, K., Ohmoto, T., Sakai, K., 2012. Production of aromatic compounds by metabolically engineered *Escherichia coli* with an expanded shikimate pathway. Appl. Environ. Microbiol. 78, 6203–6216.
Laffend, L. A., Nagarajan, V., Nakamura, C. E., 1997. Bioconversion of a fermentable carbon source to 1, 3-propanediol by a single microorganism. U.S. Patent No. 5686276.
Lan, E.I., Liao, J.C., 2012. ATP drives direct photosynthetic production of 1-butanol in cyanobacteria. Proc. Natl. Acad. Sci. U.S.A 109, 6018–6023.
Latino, D.A., Aires-de-Sousa, J., 2009. Assignment of EC numbers to enzymatic reactions with MOLMAP reaction descriptors and random forests. J. Chem. Inf. Model. 49, 1839–1846.
Lee, J.W., Na, D., Park, J.M., Lee, J., Choi, S., Lee, S.Y., 2012. Systems metabolic engineering of microorganisms for natural and non-natural chemicals. Nat. Chem. Biol. 8, 536–546.
Lee, S. Y., Park, J. H., 2008. Enhanced butanol producing microorganisms and method for preparing butanol using the same. WO Patent WO/2008/072921.
Lu, M., Lee, S., Kim, B., Park, C., Oh, M., Park, K., Lee, S.Y., Lee, J., 2012. Identification of factors regulating Escherichia coli 2,3-butanediol production by continuous culture and metabolic flux analysis. J. Microbiol. Biotechnol. 22, 659–667.
Lun, D.S., Rockwell, G., Guido, N.J., Baym, M., Kelner, J.A., Berger, B., Galagan, J.E., Church, G.M., 2009. Large-scale identification of genetic design strategies using local search. Mol. Syst. Biol. 5, 296.
Lynch, M. D., 2011. Compositions and methods for 3-hydroxypropionate bio-production from biomass. US Patent 80486242011.
Ma, F., Hanna, M.A., 1999. Biodiesel production: a review. Bioresour. Technol. 70, 1–15.
McCloskey, D., Gangoiti, J.A., King, Z.A., Naviaux, R.K., Barshop, B.A., Palsson, B.O., Feist, A.M., 2014. A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. Biotechnol. Bioeng. 111, 803–815.
McCloskey, D., Palsson, B.O., Feist, A.M., 2013. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. Mol. Syst. Biol. 9, 661.
McShan, D.C., Rao, S., Shah, I., 2003. PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics 19, 1692–1698.
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.
Mu, F., Unkefer, C.J., Unkefer, P.J., Hlavacek, W.S., 2011. Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. Bioinformatics 27, 1537–1545.
Nagarajan, V., Nakamura, C. E., 1998. Production of 1, 3-propanediol from glycerol by recombinant bacteria expressing recombinant diol dehydratase. US Patent 5821092.
Orth, J.D., Conrad, T.M., Na, J., Lerman, J.A., Nam, H., Feist, A.M., Palsson, B.O., 2011. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. Mol. Syst. Biol. 7, 535.
Orth, J.D., Thiele, I., Palsson, B.Ø., 2010. What is flux balance analysis? Nat. Biotechnol. 28, 245–248.
Palsson, B., Zengler, K., 2010. The challenges of integrating multi-omic data sets. Nat. Chem. Biol. 6, 787–789.
Paster, M., Pellegrino, J. L., Carole, T.M., Energetics, I., U.S. Department of Energy, O. o. E. E., Renewable Energy, O. o. t. B. P., 2003. Industrial Bioproducts: Today and Tomorrow. Energetics, Incorporated.
Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B., Keasling, J.D., 2012. Microbial engineering for the production of advanced biofuels. Nature 488, 320–328.
Perlack, R. D., Stokes, B. J., 2011. US billion-ton update: biomass supply for a bioenergy and bioproducts industry. Oak Ridge National Laboratory.
Pharkya, P., 2011. Microorganisms and methods for the co-production of isopropanol eith primary alcohols, diols and acids. WO Patent WO/2011/031897.
Pharkya, P., Burgard, A.P., Maranas, C.D., 2004. OptStrain: a computational framework for redesign of microbial production systems. Genome Res. 14, 2367–2376.
Portnoy, V.A., Bezdan, D., Zengler, K., 2011. Adaptive laboratory evolution-harnessing the power of biology for metabolic engineering. Curr. Opin. Biotechnol. 22, 590–594.
Portnoy, V.A., Herrgard, M.J., Palsson, B.O., 2008. Aerobic fermentation of D-glucose by an evolved cytochrome oxidase-deficient *Escherichia coli* strain. Appl. Environ. Microbiol. 74, 7561–7569.
Sauer, M., Porro, D., Mattanovich, D., Branduardi, P., 2008. Microbial production of organic acids: expanding the markets. Trends Biotechnol. 26, 100–108.
Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., Stelzer, M., Thiele, J., Schomburg, D., 2011. BRENDA, the enzyme information system in 2011. Nucleic Acids Res. 39, D670–D676.
Schellenberger, J., Park, J.O., Conrad, T.M., Palsson, B.O., 2010. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinform. 11, 213.
Schellenberger, J., Que, R., Fleming, R.M., Thiele, I., Orth, J.D., Feist, A.M., Zielinski, D. C., Bordbar, A., Lewis, N.E., Rahmanian, S., Kang, J., Hyduke, D.R., Palsson, B.O., 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat. Protoc. 6, 1290–1307.
Shen, C.R., Lan, E.I., Dekishima, Y., Baez, A., Cho, K.M., Liao, J.C., 2011. Driving forces enable high-titer anaerobic 1-butanol synthesis in *Escherichia coli*. Appl. Environ. Microbiol. 77, 2905–2915.
Shen, C.R., Liao, J.C., 2008. Metabolic engineering of Escherichia coli for 1-butanol and 1-propanol production via the keto-acid pathways. Metab. Eng. 10, 312–320.
Shen, C.R., Liao, J.C., 2013. Synergy as design principle for metabolic engineering of 1-propanol production in *Escherichia coli*. Metab. Eng. 17, 12–22.
Silverman, R.B., 2002. The Organic Chemistry of Enzyme-Catalyzed Reactions. Academic Press.
Soucaille, P., Meynial, S. I., Voelker, F., Figge, R., 2008. Microorganisms and methods for production of 1, 2-propanediol and acetol. U.S. Patent Application 12/532469.
Sutherland, P., McAlister-Henn, L., 1985. Isolation and expression of the *Escherichia coli* gene encoding malate dehydrogenase. J. Bacteriol. 163, 1074–1079.
Suthers, P. F., Cameron, D. C., 2005. Production of 3-hydroxypropionic acid in recombinant organisms. US Patent 6852517.
Tang, X., Tan, Y., Zhu, H., Zhao, K., Shen, W., 2009. Microbial conversion of glycerol to 1,3-propanediol by an engineered strain of Escherichia coli. Appl. Environ. Microbiol. 75, 1628–1634.
Tepper, N., Shlomi, T., 2010. Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. Bioinformatics 26, 536–543.
Trinh, C.T., 2012. Elucidating and reprogramming *Escherichia coli* metabolisms for obligate anaerobic n-butanol and isobutanol production. Appl. Microbiol. Biotechnol. 95, 1083–1094.
Tseng, H.C., Martin, C.H., Nielsen, D.R., Prather, K.L., 2009. Metabolic engineering of Escherichia coli for enhanced production of (R)- and (S)-3-hydroxybutyrate. Appl. Environ. Microbiol. 75, 3137–3145.
Valentin, H.E., Dennis, D., 1997. Production of poly(3-hydroxybutyrate-co-4-hydroxybutyrate) in recombinant Escherichia coli grown on glucose. J. Biotechnol. 58, 33–38.
Varma, A., Boesch, B.W., Palsson, B.O., 1993. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. Appl. Environ. Microbiol. 59, 2465–2473.
Vickers, C.E., Klein-Marcuschamer, D., Kromer, J.O., 2012. Examining the feasibility of bulk commodity production in *Escherichia coli*. Biotechnol. Lett. 34, 585–596.
Wang, Q., Liu, C., Xian, M., Zhang, Y., Zhao, G., 2012. Biosynthetic pathway for poly (3-hydroxypropionate) in recombinant Escherichia coli. J. Microbiol. 50, 693–697.
Werpy T., Petersen G., Aden A., Bozell J., Holladay J., White J., Manheim A., Eliot D., Lasure L., Jones S. and Top Value Added, Chemicals from Biomass. Volume 1-Results of Screening for Potential Candidates from Sugars and Synthesis Gas, No. DOE/GO-102004-1992. Department of Energy Washington DC, 2004.
Wyman, C.E., Decker, S.R., Himmel, M.E., Brady, J.W., Skopec, C.E., Viikari, L., 2005. Hydrolysis of cellulose and hemicellulose, Polysaccharides: Structural Diversity and Functional Versatility, pp. 995–1033.
Yan, Y., Lee, C.C., Liao, J.C., 2009. Enantioselective synthesis of pure (R,R)-2,3-butanediol in Escherichia coli with stereospecific secondary alcohol dehydrogenases. Org. Biomol. Chem. 7, 3914–3917.
Yang, F., Hanna, M.A., Sun, R., 2012. Value-added uses for crude glycerol–a byproduct of biodiesel production. Biotechnol. Biofuels 5, 13.
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, G.Q., 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.

Zeng, A.P., Sabra, W., 2011. Microbial production of diols as platform chemicals: recent progresses. Curr. Opin. Biotechnol. 22, 749–757.

Zhou, X.Y., Yuan, X.X., Shi, Z.Y., Meng, D.C., Jiang, W.J., Wu, L.P., Chen, J.C., Chen, G.Q., 2012. Hyperproduction of poly(4-hydroxybutyrate) from glucose by recombinant *Escherichia coli*. Microb. Cell Fact. 11, 54.