# Improved phenotypic predictions of metabolic models by integrating regulatory constraints

Gaoyang Li[1,2,†], Chunman Zuo[1,4,†], Qi Liu[2], Ping Wang[2], Wei Du[1*] and Huansheng Cao[3*]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of the Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, 130012, China

[2] Tongji University Cancer Center, Shanghai Tenth People's Hospital, School of Medicine, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China

[3] Biodesign Center for Fundamental and Applied Micobiomics, Arizona State University, AZ 85287, USA

[4] Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong, China

†These authors contributed equally

* Corresponding authors

Gaoyang Li: lgyzngc@tongji.edu.cn; Chunman Zuo: zuocm15@mails.jlu.edu.cn;

Qi Liu: qiliu@tongji.edu.cn; Ping Wang: wangp@tongji.edu.cn;

Wei Du: weidu@jlu.edu.cn; Huansheng Cao: hshcao@asu.edu

Running title: Regulation constrained metabolic model.

# Abstract

Physiological states-derived metabolic regulation network determines cellular phenotype but is not considered in current constraint-based models (CBMs). Here, we proposed a novel metabolic flux prediction model, Decrem, which integrated comprehensive regulatory contexts—topologically coupled reactions-derived coactivation regulation and global cell state regulators-derived enzyme kinetics—into canonical CBMs to approximate biologically feasible fluxes. A unique advantage of Decrem is it relies only on the concentrations of identified metabolites. When applied to three model organisms, *Escherichia coli*, *Saccharomyces cerevisiae* and *Bacillus subtilis*, Decrem demonstrated high accuracy in metabolic flux prediction across various conditions, particularly for incomplete metabolic models. Growth analysis by Decrem on yeast knockout strains revealed specific overflowed pathway for several low growth rate strains, which are consistent with experiments but not identified by previous methods. Additionally, the Pearson correlation coefficients between experimental and predicted growth rates on 1030 *E. coli* genome-scale deletion strains were increased to 0.743, compared to the 0.127, 0.103 and 0.281, respectively. This method is expected to accurately simulate dynamic metabolic regulation and phenotype prediction for synthetic biology.



**Keywords**: genome-scale growth rate analysis / hierarchical metabolic regulation model / linearized michaelis-menten kinetics / metabolic flux prediction / sparse linear basis decomposition


# Introduction

Metabolism is the primary cellular machinery sustaining life activities and adjusting cell state

to genetic (internal) and environmental (external) perturbations. Hence, it is a central goal for metabolic modeling to accurately quantify the cellular responses, e.g., metabolic regulation and flux distribution, integrating these internal and external conditions. Given the complex nature of metabolism and composition of different groups of heterogeneous components (e.g., enzymes, metabolites, and regulators), such goal has remained a major challenge in computational/experimental systems biology (Feist & Palsson, 2008; Hackett et al, 2016; Millard et al, 2017; Raman & Chandra, 2009; Reznik et al, 2017; Steuer et al, 2006). In the past two decades, CBMs have provided a working theoretic framework in connecting the observed cellular phenotype and genetic regulatory mechanism (Bordbar et al, 2014a; Covert & Palsson, 2002; Feist & Palsson, 2008; Raman & Chandra, 2009). These models optimize the prescribed objective function (e.g., flux of biomass reaction) to explore optimum metabolic flux distribution at steady state. Several methods have been developed to work on CBMs and applied in various areas such as bioengineering (Alper et al, 2005; Lee et al, 2007), drug discovery (Chavali et al, 2012; Li et al, 2011), gap filling (Latendresse, 2014; Satish Kumar et al, 2007), and the genome-scale synthetic biology (Fell & Small, 1986; Fong, 2014; Stalidzans et al, 2018). For example, the flux balance analysis (FBA) predicts metabolic flux distribution by maximizing the fluxes toward the biomass production based on experimentally measured nutrient uptake (Fell & Small, 1986; Orth et al, 2010), while the minimization of metabolic adjustments (MOMA) and regulatory on/off minimization (ROOM) minimize the divergence between the reference fluxes and the prediction of perturbed fluxes for mutant strains (Segrè et al, 2002; Shlomi et al, 2005). Furthermore, multi-omics experimental data such as transcriptomes and proteomes are now integrated into CBMs to

reinforce the algorithm's performance, and such condition-specific cell state information helps to shrink the feasible regions of CBMs and identify the activated reactions (Kim & Reed, 2012; Long & Reed, 2016; Sánchez et al, 2017; Tian & Reed, 2018; Xie et al, 2017).

Nevertheless, many intrinsic weaknesses of CBMs limit their performance, such as the discrepancy between predicted and experimentally measured growth and flux distribution in mutant stains, in which metabolic fluxes are often adjusted and even some reactions are rewired through complex metabolic regulation (Chen et al, 2011; Khodayari et al, 2015; Khodayari & Maranas, 2016; O'Brien et al, 2015). For example, the TCA cycle, a major bioenergetics pathway for energy production and some precursor metabolites, is empirically measured to have fully coactivated fluxes through all reactions across diverse strains of *E. coli*; but when maximizing the growth rates in FBA, some are predicted to have zero fluxes, as they do not affect the optimal biomass function (Kim et al, 2016) (Appendix File 1 Fig S1). Furthermore, a systematic evaluation uncovered integration of transcriptomes and proteomes into current CBMs only marginally improves the prediction of metabolic fluxes (Machado & Herrgård, 2014; Tian & Reed, 2018). These observations reflect the fact that current CBMs have inadequate constraints in modeling *in vivo* metabolism and cannot fully utilize the omics information defining cellular state in question. In particular, recent studies revealed the coordinated regulation of topologically highly-coupled reactions in central metabolism influences empirical flux distribution (Pandit et al, 2017; Venayak et al, 2018), and some metabolites also play important roles in the metabolic flux regulation (Hackett et al, 2016; Kochanowski et al, 2017; Park et al, 2016). However, neither of these constraints has been exploited in current CBM methods. Therefore, novel computational methods as

well as comprehensive exploration/incorporation of metabolic regulation are needed for CBMs to better approximate metabolic state and dynamics.

One underutilized constraint is the topology of metabolic networks. Microbial metabolic networks have been revealed as hierarchically organized scale-free networks, with a huge number of closely-coupled hubs (reactions/metabolites) forming a bowtie structure (catabolism→core→anabolism) (Li et al, 2018; Ravasz et al, 2002). We previously found the highly-coupled reactions of central metabolism (the core of bowtie) involve a large number of precursors and cofactors, with fast interconversion among corresponding metabolites and complex interactions between their enzymes and regulators (Li et al, 2018). These characteristics all contribute to buffering the concentration of essential precursors, quick responses to diverse intra/extra-cellular perturbations, so as to maintain global metabolic optimum under various growth conditions (Braakman & Smith, 2012; Hackett et al, 2016; Kochanowski et al, 2017; Li et al, 2018; McCloskey et al, 2018; Okada & Mochizuki, 2017; Park et al, 2016; Venayak et al, 2018). For example, the dynamic trade-off between TCA cycle and glycolysis is regulated by both the ADP/ATP ratio and the concentrations of pyruvate, the starting metabolite of energy metabolism (Berg, 2007a; Berg, 2007b); the reaction fluxes in pentose phosphate pathway (PPP) are primarily determined by the ratio of NAPD/NADPH and the ribose 5-phosphate, the main precursor for the synthesis of nucleotides (Berg JM, 2002). These observations suggest the topologically coupled reactions from the same pathways in central metabolism appear to be determined by the global optimal state and then co-regulate their fluxes through shared key metabolite regulators (Millard et al, 2017), which is confirmed by the co-activation of reactions in the minimal metabolic pathways

derived from metabolic topological decomposition (Bordbar et al, 2014c). However, existing linear programming derived canonical CBMs cannot utilize these constraints as the topology (pathway) conservation isn't directly integrated into those topology-free models, therefore incorrect predictions of flux distribution by CBMs will occur, especially in cases with metabolic network rewiring.

Metabolite concentration has been recognized to not only affect the affinity of catalytic enzymes, but also determine the expression of growth-associated enzyme-coding genes through transcription factor (TF)-metabolite interactions (Kochanowski et al, 2017; Zelezniak et al, 2018). Gene knockout experiments indicate that the concentrations of metabolites in the composition of biomass (e.g., nucleic acids and amino acids), by regulating TFs, have a dominant role in determining the promoter activities of the genes in central metabolism in the multi-scale regulatory architecture of *E. coli* [34]. And the regulatory metabolites in many metabolic pathways adjust the upstream fluxes by controlling the activities of their cognate (often allosteric) enzymes in a feedback loop [30].Such regulatory metabolites have been utilized to complement the Michaelis-Menten kinetics to improve the prediction of isolated fluxes in yeast central metabolism [29]. These observations highlight the advantages of building complete metabolic regulation kinetics models to capture the precise distribution and dynamic regulation of metabolic fluxes (Khodayari & Maranas, 2016). However, these kinetic models are not available in practice, due to the scarcity of key kinetic parameters such as $K_m$ and $K_{cat}$, the incompleteness of current regulation-integrated metabolic network models and the lack of corresponding multi-omics data (e.g., paired metabolite and enzyme quantity under multiple growth conditions).

In this paper, we present a computational framework to integrate hierarchical regulation of metabolism into CBMs and applied it to approximate biologically feasible fluxes (Fig 1). We first incorporated the identified topological profiles of metabolic networks into the canonical FBA, by replacing the optimized synchronously coactivated (coregulated) reactions with their sparse linear basis of the corresponding stoichiometric matrix, and reconstructed a **de**coupled **c**oactivated-**re**actions constrained metabolic **m**odel (Decrem). Next, a kinetic model was created to enhance Decrem by incorporating linearized Michaelis-Menton equation for essential regulatory reactions. A unique feature of the kinetic Decrem model is regulatory metabolites identified with Bayesian model selection and reformulated Michaelis-Menton kinetics, which depends only on the associated metabolite concentrations and avoids the unavailable kinetic parameters and the requirement for paired multi-omics data. The performance of Decrem was validated in three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae* (yeast) and *Bacillus subtilis.* The optimal flux distributions predicted by Decrem highly correlated with experimentally measured fluxes in various cellular behaviors across multiple strains, in the genotype-guided growth rate, the flux distribution on different carbon sources, and the metabolic alternative path rewiring. In particular, the accurate predicted growth rates by kinetic Decrem in *E. coli* genome-scale knockout strains revealed the intracellular perturbations are mainly 'buffered' by the highly coupled reactions, which reinforce the co-regulation and robustness properties of the central metabolism. To the best of our knowledge, this is the first comprehensive computational platform available to integrate metabolic topology and regulation into CBMs.

## Results

## The topologically decoupled reconstruction of metabolic network

In this section, we developed a topologically decoupled linear representation of metabolic networks to characterize the coactivated regulation of highly-coupled reactions with three major steps.

Firstly, to explore the topologically coordinated pattern of highly coupled reactions, compact substructures of coupled reactions in metabolic network were identified from its bipartite graph representation with a graph clustering algorithm (see Materials and Methods and Appendix file 1 Fig S2) (Meilă & Pentney), and the similarity metric of clustering analysis was the number of the simple cycles between any two reaction nodes of bipartite graph (see Materials and Methods). Thus, the reactions involved in multiple simple cycles were treated as the candidate elements of highly-coupled substructures and classified into several clusters to further explore their linear decoupled representation. Specifically, two coupled-reaction clusters were identified, which included 927 of the total 2382 reactions in the *E. coli i*AF1260 model (Appendix file 2 Table S1). Statistical analyses revealed the central metabolism, such as the TCA cycle, PPP, glycolysis, amino acid and glycerophospholipid pathways, primarily consisted of coupled reactions, and included 70% (141 of 201) of reversible reactions. In contrast, the biomass, tRNA, membrane lipid biosynthesis and membrane transports pathways mainly consist of (non-coupled) linear reaction chains (Fig 2a). Additionally, the $K_m$ values (0.023 mM) of the coupled reactions in the clusters were smaller than those (0.036 mM) in the (non-coupled) linear reaction chains (Wilcoxon test, $p$ value = 5.74E-4; Fig 2b). Together, these highly-coupled structure and kinetic enzyme parameters indicated the reactions in central metabolism tend to closely interact with each both in proximity (short

path length) and at high priority (low $K_m$), for fast generation of essential metabolites (precursors) (Braakman & Smith, 2012; Li et al, 2018). This fast interconversion is consistent with the concentration coordination among the 12 essential precursors in *E. coli* according to the minimal reaction walks (Braakman & Smith, 2012; Noor et al, 2010), as opposed to directly connecting input nutrients to biomass with shortest reaction paths (which is prioritized in canonical CBMs). We then validated the coordination of the identified coupled reactions with publicly available fluxes experimentally measured in 24 *E. coli* knockout strains. The flux correlations among the coupled reactions from the same pathways were higher than those for non-coupled reactions from different pathways. For example, the average correlation coefficients were 0.913, 0.975, and 0.794 for glycolysis, pentose phosphate pathway and citric acid cycle, respectively, in contrast with the 0.505, 0.267, and 0.421 for each control reaction set (t test, *p* values = 2.33E-9, 3.06E-43 and 2.25E-4, respectively; Fig 2c), which indicated the topologically coupled reactions also tend to coordinate their fluxes.

Secondly, we identified the minimal independent components of highly coupled reactions to characterize their coordinated metabolic fluxes (regulation). Topologically, we represented the identified coupled-reactions-based clusters as sparse linear basis (SLB) vectors of their corresponding null-space of stoichiometry matrix, and treated them as the decoupled components of metabolic network, that is, each SLB consisted of the least number of coupled reactions corresponding to an indivisible independent flux (see Materials and Methods). Like EFMs (elementary flux modes), arbitrary metabolic fluxes under real growth conditions could be decomposed as the weighted linear combination of the identified SLBs. Opposite to the possibly infinite groups of ordinary linear basis, there was a

unique group of SLBs in the null-space of the stoichiometry matrix, which was an elegant way to define the mutually independent components for the dense biological interaction-associated reactions (Bordbar et al, 2014c). Remarkably, unlike the NP-hard problem of enumerating EFMs, our decomposition only produced, at most, r SLBs, r being the rank of stoichiometry matrix (see Materials and Methods and Appendix file 1 Method S1). We then validated the coordinated activation of reactions within the SLBs through the gene expression in the 24 *E. coli* knockout strains, and found that the correlations (mean $r$ = 0.447) of gene expression among the element reactions of the same SLB were higher than those (mean $r$ = 0.28) between the SLBs (t test, $p$ value = 2.05E-33; Fig 2d). This suggested that the reactions from the same SLB tended to be coactivated while the reactions SLBs were more independently regulated.

Finally, we reconstructed a decoupled linear representation of the metabolic network (Decrem) by combining the "linear basis reactions", the merged form of the element reactions of each SLB by the reallocated stoichiometric coefficients, to the non-coupled reactions of the original metabolic network (see Materials and Methods and Appendix file 1 Fig S2). To make sure this decoupled representation has the same flux variability as the original metabolic network, flux variability analysis (FVA) was conducted in *E. coli* and *S. cerevisiae* which confirmed the preservation of solution space (Fig 2e and Appendix file 1 Fig S3) (Gudmundsson & Thiele, 2010). Our model reassigned the trivial flux while effectively preserving the primary flux capability of the original metabolic model. This Decrem model served as our working model for subsequent functional analysis of the CBMs.

**Prediction of metabolic fluxes in response to environmental perturbation**

In this section, we applied Decrem to metabolic flux prediction, in comparison with three other methods: FBA, pFBA (parsimonious FBA) and RELATCH (Kim & Reed, 2012; Lewis et al, 2010; Orth et al, 2010), on three model organisms: *E. coli* (*iAF1260*) *S. cerevisiae* (*iMM904*) and *B. subtilis* (*iYO844*), which are composed of 2382, 1577, and 1250 reactions and 766, 558 and 332 "linear basis reactions" in the corresponding decoupled reconstructed metabolic models, respectively (see Data and software availability and Appendix file 2 Table S1).

We first calculated the metabolic flux distribution in wildtype *E. coli* K12 MG1655 under four conditions in MOPS medium supplemented with glucose or xylose and under aerobic or anaerobic respiration (Gonzalez et al, 2017) (Appendix file 1 Material S1), with input nutrients as the sole constraints on CBMs (Appendix file 3 Table S2). Using three evaluation metrics—Pearson correlation coefficient *r*, mean squared error (MSE) and the number of activated reactions, the predictions by Decrem had higher correlations with the experimentally measured $^{13}$C MFA fluxes (Appendix file 4 Table S3) than the predictions by three other methods under all four conditions (Fig 3). One prominent feature of Decrem was that it produced much less discrepancies (higher accuracy thereof) between predicted and empirical fluxes than the other methods (Fig 3a). Furthermore, the fluxes predicted by Decrem have the smallest MSE (maximum upper bound of flux being 1000 mM/gDW/hr), and most of the activated reactions (with non-zero flux) agreed well with the experimental $^{13}$C MFA fluxes (Fig 3b and c). A prominent example was that the activated reactions in the TCA cycle, when Decrem was compared to FBA: the Decrem predictions well agreed with the experimentally measured $^{13}$C MFA fluxes (with one exception; Appendix file 1 Fig S1), whereas four activated reactions were predicted as inactive (with zero fluxes) by FBA.

Consistently, Decrem performed better than the three other methods in flux prediction in two other model organisms: *S. cerevisiae* and *B. subtilis*, each having 32 and 84 experimentally measured $^{13}C$ MFA fluxes (Moxley et al, 2009; Rühl et al, 2012) (Appendix file 1 Material S1, Appendix file 2 Table S1 and Appendix file 3 Table S2) (Fig 3). Particularly, it appeared Decrem performed well with the highly complex eukaryote yeast: *r* are 0.696, 0.3, 0.284 and 0.3 for Decrem, FBA, pFBA and RELATCH, respectively, in which reactions often were strongly coupled based on cellular compartments. This was more extraordinary since current *i*MM904 model of *S. cerevisiae* was not complete. Therefore, the results above highlight two advantages of Decrem in metabolic flux prediction: (1) high accuracy in predicting the activeness of reactions under diverse growth conditions; and (2) high precision in quantifying the actual flux distribution of the activated reactions, especially for the incomplete metabolic models; this is significant as most of current metabolic models are incomplete.

## Metabolic flux prediction in single-gene deletion strains

We further evaluated Decrem in predicting flux response to internal perturbation, single-gene deletions. Two metabolic models of *S. cerevisiae* were used, *i*DN750 (1059 metabolites and 1266 reactions) and *i*MM904 (1226 metabolites and 1577 reactions) were used. And 38 single-gene deletion mutants were collected with experimental $^{13}C$-labled MFA fluxes, growth rates, nutrient uptake and several extracellular exchange fluxes (Blank et al, 2005). These genes encoded 28 of the 51 flexible and highly active reactions in PPP, TCA cycle, glyoxylate cycle, polysaccharide synthesis, mitochondrial transporters, and byproduct formation (Appendix file 1 Material S2 and Appendix file 5 Table S4) (Blank et al, 2005).

Decrem and four other methods—pFBA, FBA, RELATCH and REPPS (Long & Reed, 2016) were applied to flux prediction in these mutants, with the specified extracellular exchange fluxes and actual nutrient concentrations constraining the feasible region (the upper/lower bounds of corresponding reactions).

The predictions from Decrem showed the highest correlations with the experimentally measured fluxes in almost all strains for both metabolic models (Fig 4 and Appendix file 6 Table S5). Specifically, the average Spearman and Pearson correlation coefficients of Decrem for the *i*DN750 were 0.72 and 0.763 among the 38 mutant strains, compared to the mean of 0.615, 0.607, 0.556, 0.462 and 0.733, 0.681, 0.707, 0.595 for pFBA, FBA, RELATCH and REPPS, respectively (one way ANOVA, *p* values= 5.52E-20 and 4.21E-07, respectively). With the more complete yeast *i*MM904 model, the average Spearman and Pearson's correlation were 0.782，0.765, 0.656, 0.750, 0.645 and 0.936，0.747, 0.915, 0.925, 0.841 for Decrem, pFBA, FBA, RELATCH and REPPS, respectively, which were significantly different (one way ANOVA, *p* values = 4.35E-06 and 8.26E-06, respectively). The differential accuracy of flux predictions between the *i*DN750 and *i*MM904 model showed that Decrem performed better than other methods particularly on metabolic models which are less complete (e.g., *i*DN750 vs *i*MM904), while the predictions of FBA, pFBA and RELATCH were apparently influenced by network completeness (Fig 4). Further analysis revealed that such differences were due to the fact that a larger proportion, 47% (146 of 311), of new reactions(compared to the *i*DN750 model) in the *i*MM904 model were highly-coupled reactions, vs average 35% ( 558 of 1577) for the *i*MM904 (fisher's exact test, *p* = 5.427e-04; Appendix file 2 Table S1), which clearly refined the solution space of metabolic flux prediction and changed the optimal reaction paths in

the *i*MM904 model. Therefore, Decrem was a more general and flexible method for predicting the flux distribution of less complete metabolic models, as are nearly all current metabolic networks.

We then applied Decrem and pDecrem (Parsimonious Decrem, see Appendix file 1 Method S2) to estimate the growth rate of the 38 mutant strains of *S. cerevisiae*, with the four other methods using the *i*MM904 model. PCA analysis of the predicted fluxes from the six methods showed the top two principal components (PCs) could explain more than 99% the variance of all predicted fluxes. The two PCs of Decrem predictions were very consistent and highly correlated ($r$>0.7) with experimentally measured growth rates, while the other methods had, at most, one PC which merely showed moderate correlation ($r$~0.5) (Appendix file 6 Table S5). Furthermore, a PCA regression revealed Decrem produced the best flux variance to explain the observed growth rates, as shown by the coefficient of determination, $R^2$ = 0.74, 0.731, 0.731, 0.841, 0.9 and 0.9 respectively for the six methods (Decrem and pDecrem are our methods; Fig 5a). Most importantly, our Decrem successfully identified all the mutations with large effects leading to growth <0.5 $h^{-1}$: ALD6, FUM1, PDA1, RPE1 and ZWF1 (Fig 5b). Moreover, Decrem-derived flux distribution through the specific pathways could explain those low growth rates. For examples, the ZWF1 mutant strain had a different flux distribution from those of the ALD6 and RPE1 (Fig 5c). Interestingly, the exceptionally high fluxes of the transport mitochondrial pathway and citric acid cycle pathway in ZWF1 mutant strains well agreed with the experimental observation that the NADPH, $NADP^+$-dependent mitochondrial malic enzyme flux was significantly increased in the zwf1 mutant (Blank et al, 2005). The other methods only identified some of these of large-effect

mutations on growth rate, leading to a high false positive rate. Comparatively, our Decrem demonstrated high accuracy and low false positive rates in assessing the growth rate of mutants and well approximated the real cell state perturbed by the internal genetic changes.

**Metabolite concentration uncovering the metabolic regulation-enhanced kinetic model**

External or internal perturbations disrupt global metabolic optima, which are quantitatively manifested by the empirical measured- or the Michaelis-Menten kinetics derived- metabolic fluxes of the key regulated reactions, reflecting cell state. For example, growth rate-associated experimental exchange fluxes have been integrated into canonical CBMs to constrain the flux allocation of reactions involved in essential mass/energy metabolism (this strategy is used in all six methods at above section on a medium-scale deletion strains). Given the scarcity of paired multi-omics data even in model organisms, unavailable kinetic parameters and complicated experimental design of the $^{13}C$ MFA on a genome-scale scope (Machado & Herrgård, 2014; Tian & Reed, 2018), here we built the growth rate-associated metabolic kinetic model dependent only on the physiologic concentrations of regulator metabolites (Hackett et al, 2016; Kochanowski et al, 2017; Park et al, 2016), incorporating highly-coupled topology-derived coactivated regulations, to enhance the CBMs to characterize the genome-scale cell state specific metabolic flux response (see Appendix file 1 Fig S4).

Firstly, given the metabolic regulation of 30 TFs in central carbon metabolism by global growth rate and local pathway-specific essential metabolites, e.g., AMP, Fructose-1,6p, Fructose-1p in *E. coli* (Kochanowski et al, 2017), we hypothesized there are quantitative

relationships between the expression levels of metabolic genes activated by these 30 TFs and the biomass composition (growth rate being equivalent to change in biomass), as well as some essential metabolites and precursors. Then we tested this hypothesis with a multi-omics dataset of 24 *E. coli* single-gene knockout strains, which included the expression of 85 metabolic genes in central metabolism, more than 100 metabolite concentrations and 51 $^{13}C$ MFA fluxes for each strain (Ishii et al, 2007). Through extensive literature survey (Chang et al, 2016; Kothari et al, 2016), 45 essential metabolites (and biomass composition) from the dataset were selected as potential metabolic regulators. After that, they were classified into two groups using a hierarchical clustering method according to their concentration profiles across the strains: biomass-constituent group (BG), and precursors or essential metabolites group (PG) (Appendix file 1 Fig S5 and Appendix file 7 Table S6). This classification indicated different concentration patterns between the global (biomass) and local (pathway-specific) regulators (Appendix file 1 Material S3). Next, we conducted a partial least squares regression (PLS) to fit the observed expression profiles of the 85 genes using the concentrations of the potential regulatory metabolites of either group in the 24 mutant strains (see Materials and Methods). By taking restrict multi-thresholds (total regression correlation > 0.84 and the correlation of first PC > 0.38 according PLS; Appendix file 1 Material S3), 32 of 85 genes were identified as the regulatory targets of the 23 BG metabolites, whereas no genes were identified as the regulatory targets of the PG metabolites (Fig 6a and Appendix file 1 Fig S6). These identified metabolite-gene regulations were highly consistent with quantitative proteome-derived prediction of metabolite concentrations and the target genes of the 30 metabolite-regulated TFs (Appendix file 7

Table S6) (Kochanowski et al, 2017; Zelezniak et al, 2018). The significance of the identified metabolite-gene regulations was verified with the correlation between measured and predicted gene expression levels using a wide range of metabolites selected by 10,000 random samplings from the 45 potential metabolic regulators (see Appendix file 1 Material S3). A *p* value of 3.1E-3 for the identified 23 growth-associated metabolites were observed against the randomly selected metabolites (Fig 6b), while the *p* value was 0.48 for the PG metabolites. Those identified 32 genes were primarily located in the PPP and pyruvate metabolism in KEGG pathways (Fig 7), which were associated with the global cell state for biosynthesis (growth): generating NADPH and pentoses toward nucleotide and amino acid biosynthesis (Voet & Voet, 2010), instead of being in energy-producing pathways (TCA and glycolysis). These results together suggested that the expression of the genes whose products (enzymes) participated in growth-associated pathways of *E. coli* could be represented as a linear combination of concentrations of biomass composition.

Then, we constructed a regulation-enhanced kinetic model for those growth-associated metabolic reactions. Specifically, we combined the PG metabolites-derived enzyme catalytic kinetic regulators and BP-represented gene (enzyme) expression to reformulate a linear Michaelis-Menten equation (see Materials and Methods). This kinetic regulatory model depended only on the concentrations of associated metabolites and supported by the following two sources of information. First, the concentrations of primary carbon metabolites in central metabolism were at levels approximate to or higher than their Michaelis constants: $K_m$ (Park et al, 2016), which would allow derivation of linearized approximation of Michaelis-Menten kinetics (see Materials and Methods). Second, the

identified PG metabolites could act as potential metabolic enzyme enhancers or inhibitors in the Michaelis-Menten kinetics (Hackett et al, 2016); therefore, these essential metabolites or precursors directly regulated the catalytic activity of their cognate enzymes to control the reaction fluxes. We verified this through a genome-scale map of protein-metabolite interactions based on experimentally measured data [52] (Appendix file 7 Table S6). Based on these two information (linearized Michaelis-Menton kinetics and metabolite-enzyme interaction), we used a Bayesian model selection method to identify enzyme regulators for cell state-associated reactions, and quantify the corresponding kinetic parameters, according to the empirical $^{13}C$ FMA fluxes and metabolite concentrations (see Materials and Methods). The resulting *E. coli* regulation integrated central metabolic network can be found in Fig 7.

Finally, we evaluated the constructed kinetic model by optimizing the growth-associated key regulated fluxes in *E. coli* central metabolism. The produced optimal kinetic fluxes displayed high consistency with experimental measurements through $^{13}C$ isotope tracing (Fig 6c), with $r$ being 0.95, 0.97 and 0.93, respectively, for the reactions of glucose transport, alcohol dehydrogenase and acetate kinase. Topologically, those complex regulated reactions were located at critical branch points of the metabolic network and dictated the flux allocation toward downstream reactions. Such a regulatory scheme could be utilized to improve the performance of Decrem by constraining key regulated reactions to their kinetic fluxes to correct/adjust the allocation of intracellular mass and energy within metabolic model. The resulting model, kinetic Decrem, now had combined advantages of locally coordinated reaction activation with regulation by global cell state (self-adaption strategy toward intra/extra-cellular perturbations) (see Materials and Methods). And we applied

kinetic Decrem for genome-scale growth rate estimation in the next section.

## Growth estimation for *E. coli* genome-scale deletion strains with kinetic Decrem

We applied kinetic Decrem to 3,800 *E. coli* single-gene deletion mutants, which have experimentally measured concentrations of over 7,000 metabolites and corresponding growth rates [55]. After several steps of data cleaning and preprocessing (see Appendix file 1 Material S4), 1,030 mutants with genes involved in metabolism were selected for downstream analysis. We first examined the predicted growth rates by the kinetics-free methods, the pFBA, MOMA and Decrem without any external flux constraints. As expected, poor predictions were produced, leading to low correlations with experimentally measured growth rates, $r$ being 0.127, 0.103 and 0.281, respectively (Appendix file 1 Fig S7). This underscored how challenging it is to estimate growth rate in the absence of appropriate constraints about metabolic state in gene knockout strains, as the complex metabolic regulation compensates for the internal genetic perturbations.

Next, we constructed the kinetic fluxes of 5 key regulated reactions in central metabolism with predictions (based on metabolite concentrations) for each of 1,030 mutants to approximate the specific metabolic state, e.g., the branch points of glycolysis and PPP, the allocation of downstream flux of pyruvate metabolism, and the growth-associated secretion (Fig 7 and Appendix file 8 Table S7). On the CBMs integrated with kinetic fluxes, the growth rates of the 1,030 mutants were estimated with the six kinetic enhanced methods: Decrem, pDecrem, FBA, pFBA, RELATCH and REPPS. Results showed all the six kinetic methods significantly improved their predictions, compared to the kinetic-free methods (Fig 8a). Among them, kinetic Decrem (and pDecrem) presented the highest correlations with the

experimental growth rates among all the methods used ($r$ = 0.731 and 0.743 vs the 0.421, 0.685, 0.474, 0.509) (Appendix file 8 Table S7). This result clearly showed the outstanding advantage of our kinetic Decrem in approximating real cell states.

We then demonstrated the explanatory power of kinetic Decrem in interpreting the observed growth rates of mutants with corresponding (altered) flux distributions. The adjusted distribution of pathway-specific fluxes defined as the accumulation flux of each pathway (the accumulated sum of all non-zero fluxes within a pathway) (Appendix file 1 Material S4), was predicted by the five methods (excluding pDecrem). The correlations between the accumulated pathway-specific fluxes and growth rate showed that kinetic Decrem quantified the most growth-related pathways we curated from literature. For instance, many well-known important pathways of cell growth—the Glutamate, Nucleic acid and most of Amino acid metabolisms —were recognized only by kinetic Decrem with large correlation coefficients (Fig 8b). Interestingly, although the kinetic pFBA method could predict the growth rates with relatively high accuracy, the corresponding accumulated pathway-specific fluxes only included few of the curated growth-associated pathways, which suggested the potential local optimal solution of the pFBA. Moreover, the analysis for the pathway-specific accumulated growth rates suggested kinetic Decrem-predicted distribution of accumulated growth rates through different pathways highly agreed with the real experimental distribution. FBA reached similar level of accuracy (Fig 8c), but the predictions by kinetic pFBA often involved pathways in which the growth rates were little influenced by gene knockouts and shrank the accumulated growth rate to zero for the pathways containing knockout genes of large effects (Fig 8c). Such biases were an intrinsic property of the L1 norm-based pFBA method (Guyon et al, 2003; Tibshirani, 1996), despite the relatively good fit

of correlations.

We further examined the flux variance distribution in each mutant for the changed growth rates of mutants. PCA analysis of the fluxes predicted by kinetic Decrem in the 1,030 mutants was shown in Fig 8d (Appendix file 1 Material S4). The distribution of the two top PCs indicated that the primary flux variances came from the decoupled “linear basis reactions”, compared to the non-coupled reactions:1.461 vs 1.10 on average on PC1 direction, t test, *p* value = 3.04E-20; 1.558 vs 1.07 on average on PC2 direction, t test, *p* value = 2.86E-25. This was consistent with the high robustness of the central metabolism (Braakman & Smith, 2012; Li et al, 2018) (“linear basis reactions” primarily consists of central metabolism). Further, we suspected that the effects of deleted genes encoding enzymes for the reactions within the SLBs would be more pronounced than those encoding enzymes for non-coupled reactions. It was confirmed by the analysis of the reaction type-based growth rates: the complex “linear basis reaction” (the number of element reactions of associated linear basis vector > 1), the simple “linear basis reaction” (the number of element reactions of associated linear basis vector = 1) and the non-coupled reactions (see Appendix file 1 Material S4), and the average growth were 0.666, 0.773 and 0.813 $h^{-1}$ (one way ANOVA, *p* value = 7.71E-28) (Fig 8e). Once again, this result suggested our reconstructed “linear basis reactions” well agreed with their biological functions. Finally, we examined the cause for the observed differences among the flux variance of “linear basis reactions”, in the number of simple cycle and the enzyme properties. The flux variance were primarily explained by the multi-subunit enzymes and the topologically highly-connected “linear basis reactions”: SLBs were involved in a large number of simple cycles and included limited number of element reactions (Fisher's Exact Test, *p* values = 2.34E-21 and 0.0051, respectively) (Fig 8f). Therefore, the

topological vulnerability of these reactions will result in functional variability.

## Discussion

We proposed Decrem, a hierarchical metabolic regulation framework to enhance CBMs for metabolic fluxes prediction. By rebuilding the decoupled metabolic network with the sparse linear basis decomposition and approximate the kinetic fluxes of key regulated reactions with the associated metabolites concentrations to constrain the feasible region of optimal flux, our Decrem effectively integrated the metabolic regulation information and reduced the required multi-omics data for the genome-scale metabolic kinetic modeling. Compared to existing methods which are either topologically unconstrained or constrained with empirical omics data, Decrem demonstrated superior performance in predicting metabolic fluxes in three model organisms and growth rate in a genome-scale deletion strains of *E .coli*. So, it is an effective strategy for accurately depicting the metabolic response and exploring the self-adapted regulation mechanism of cellular perturbation.

By applying the sparse linear basis decomposition, the (coupled) element reactions within identified SLBs displayed high co-expression among multiple growth conditions, indicating coordinated activation of topologically highly-coupled reactions in maintaining the synchronous collaboration of the essential metabolites (precursors). Interestingly, this characteristic also has been utilized to identify the nonredundant functional units of metabolism, i.e. the minimal metabolic pathway or flux tope (Bordbar et al, 2014c; Gerstl et al, 2018). A topological orthogonality principles was successfully used to designed the bioengineering strains with minimal interaction between desired product-associated pathways and metabolic components related to biomass synthesis (Pandit et al, 2017). In

addition, several topological constraints, such as removing the thermodynamically infeasible loops, decoupling two desired phenotypes, have been applied to the CBM methods to improve their metabolic production in recent studies (Venayak et al, 2018; Wang et al, 2018). All those observed metabolic topological properties can be concluded as: the network function is mainly determined by its topology (Ahnert & Fink, 2016), the fundamental principle of complex network theory.

Recently, metabolic kinetic models have attracted great attention as they can combine the fluxomics, metabonomics and transcriptomics into a unified framework to exploit the condition-specific cell state (Khodayari & Maranas, 2016; Khodayari et al, 2014; Stalidzans et al, 2018). However, currently constructing the genome-scale metabolic kinetic model is still very difficult because the limited knowledge about kinetic parameters, e.g., $K_m$, $K_{cat}$, scarcity of metabolic regulators and paired genome-scale multi-omics data (Hackett et al, 2016; Khodayari & Maranas, 2016). So far, the largest metabolic kinetic model, k-ecoli457 of *E. coli*, only contains 457 reactions, 337 metabolites and 295 substrate-level regulatory interactions according to the computationally predicted kinetic parameters (Khodayari & Maranas, 2016). Alternatively, a type of indirect kinetic method is approximating the possible flux boundary using its transcriptomics (proteomics) data to describe coarse-grained feasible region of reactions though stoichiometric matrix (Long & Reed, 2016; Sánchez et al, 2017; Tian & Reed, 2018), which has shown limited improvement (Machado & Herrgård, 2014). In addition, a recent study found experimental fluxes cannot be accurately predicted from Michaelis-Menten kinetics without considering metabolic regulators (Hackett et al, 2016). By integrating the metabolite-TF regulatory regime, kinetic Decrem can predict growth rates

and corresponding fluxes with high accuracy.

A unique advantage of our kinetic Decrem is it relies only on the experimental concentrations of identified essential metabolites, which serves as the key indicators of metabolic state, and directly regulate the enzyme activity or gene expression. Although some metabolites (e.g., those in glycolysis and TCA) have long been known to regulate enzyme activities in biochemistry textbooks, not a lot is known on the genome scale. On the one hand, the regulatory metabolites of specific pathways in central metabolism prefer to frequently interact with the catalyzing enzyme by blocking or activating the function domain, to synchronously adjust the upstream fluxes (Hackett et al, 2016; Piazza et al, 2018). Furthermore, the relative metabolite concentration-derived thermodynamics can determine the directions of reverse reactions (Flamholz et al, 2013). On the other hand, several studies revealed that about 70% of the total variance of promoters activity of metabolic genes in central metabolism of *E. coli* can be explained by the growth rate-derived global transcriptional regulation across multiple mutant strains (Kochanowski et al, 2017; Reznik et al, 2017). These findings suggest a potential relationship between the concentration of biomass-constituent metabolites and the expression of metabolic genes involved in the regulons of transcriptional factors. This relationship is verified by a recent study, in which the identified metabolites concentrations are predicted by quantitative proteome data [44]. In addition, a comprehensive analysis of metabolite concentrations in central metabolism shows that most $K_m$ values, the threshold of the enzyme activation, are satisfied in three microbes (Park et al, 2016), which provides a linearized approximation of the complex Michaelis-Menten kinetics. Overall, both the topology of metabolic networks and regulatory

metabolites are utilized to identify coactivated or key regulated reactions, which produces a minimal set of regulatory constraints to develop a genome-scale kinetic model.

Compared to other methods on three model organisms, Decrem not only best recovers the real flux distribution with a high accuracy in a wide range of strains, but also present excellent predictions and annotation of the observed growth rate. This demonstrated strong capability of Decrem can approximate real intracellular state and be used in designing high-yield mutant strains in bioengineering. In contrast, pFBA method minimizes the total fluxes using L1 regularization constrained by an optimal objective reaction, i.e. biomass synthesis, and seems over-optimized by yielding the high correlation of reactions associated experimental large fluxes, while forcing the zero-fluxes to majority of reactions with empirically non-zero fluxes (Kim et al, 2016). Even for several growth critical pathways, i.e. the Pentose Phosphate Pathway, Glutamate metabolism and Purine and Pyrimidine Biosynthesis, pFBA also predicted them to be non-activated with small experimental growth rates (Fig 8b and c), obviously, this is unreasonable to describe the optimal cell state and its regulation. Decrem shows that there is a large influence of metabolic network topology to the prediction of fluxes distribution, this phenomenon is found both in the reconstructed “linear basis reaction” of Decrem and in the two versions of yeast metabolic models. A possible explanation of this observation is that the optimal fluxes distribution of real metabolic network is greatly determined by its topology, e.g., the rewiring or perturbed network structure will lead to changed feasible region. Moreover, Decrem could be applied to elucidate important regulatory branchpoints and the self-adaption of regulation mechanism for the knockout strains, which is helpful to accurately predict the potential

target genes/reactions for the design of bioengineering strains.

The main limitation of Decrem is that the medium-scale matched $^{13}C$ MFA flux and metabolite concentration data are needed to construct reasonable kinetic models. Alternatively, the experimentally measured $K_{cat}$ and turnover number also can be used instead of the $^{13}C$ MFA flux. However, the well constructed kinetic models are convenient to transfer to any other applications. Overall, we propose, Decrem, a hierarchical regulation network reinforced metabolic analysis model, to accurately predict the phenotypes and uncover the complex regulation of cell metabolism.

# Materials and Methods

## The topology-decoupled reconstruction of metabolic model

The original metabolic network consists of both coupled coactivated reaction cycles (e.g., TCA cycle) and simple linear chain reactions (e.g., biosynthetic reaction chains). We developed the Decrem method to capture the contribution of the coactivated coupled reactions while preserving the consistency of the linear components. The detailed framework for our method, which is shown in Appendix file 1 Fig S2, is given as follows:

***Step1: Identifying reaction cycle-based coupled subnetworks***

Defining a bipartite graph derived from the metabolic network as $G(V_{m+n}, E)$, where the node set $V_{m+n}$ includes both the $m$ metabolite nodes and $n$ reaction nodes, and the edge set $E$ is composed of all the interactions between the metabolite and reaction nodes. We firstly built the similarity matrix $A_{n\times n}$ for $n$ metabolic reactions based on the number of topological simple cycles of $G$. Specifically, the element $a_{ij}$ of $A_{n\times n}$, which indicates the similarity between reactions $i$ and $j$ in the metabolic network, is measured by the number

of simple directed cycles between the paired reaction nodes $(v_i, v_j)$ of $G$. According to the similarity matrix $A_{n\times n}$ above, the BestWCut clustering algorithm (Meilă & Pentney, 2007) was then used to identify the dense subnetworks (also known as network communities) consist of highly coupled reaction cycles. Here, the subnetworks are denoted as $C = \{C_k | k = 1 \dots K\}$, where $C_k$ is a subset of reaction index set $[n] = \{1, 2, \dots, n\}$, $k$ is the index of subnetwork $C_k$, and $K$ is the total number of subnetworks. If we define $D_i = \sum_{j\in[n]} a_{ij}$ as the weighted out-degree of reaction node $v_i$ , $i \in [n]$, then the cluster degree $D_k$ for subnetwork $C_k$, could be defined as:

$$D_k = \sum_{i\in C_k} D_i \tag{1}$$

Now the generalized weighted cut (WCut) associated with $C$ is obtained by minimizing $WCut(C)$:

$$WCut(C) = \sum_{k=1}^{K} \sum_{k'\neq k} \frac{Cut(C_k, C_{k'})}{D_k} \tag{2}$$

where

$$Cut(C_k, C_{k'}) = \sum_{i\in k} \sum_{j\in k'} a_{i,j} \tag{3}$$

and $C_{k'}$ is the complement set of the cluster $C_k$

***Step2: Reconstructing decoupled representation of identified subnetwork using sparse linear basis***

Inspired by the minimal metabolic pathways (Bordbar et al, 2014c), we represented the highly-coupled subnetworks with the SLB of null space of the corresponding stoichiometric matrix, which is biologically explained as the minimal and indecomposable coupled components, and satisfy the constraint of thermodynamic and mass balance of element reactions. Different from infinite ordinary linear basis vectors of null space of stoichiometric

matrix, there is a unique and globally optimal sparsest basis group of null space (Bordbar et al, 2014c; Saa & Nielsen, 2016). Briefly, the orthonormal null space $N_{C_k}$ is initially defined by singular value decomposition (SVD) for the stoichiometric matrix $S_{C_k}$ of subnetwork $C_k$, here additional artificial exchange reactions are introduced in $S_{C_k}$ to maintain the mass balance of reactions in subnetwork $C_k$ (more details can be found in Appendix file 1 Method S1). Then, the column vectors of the orthonormal null space $N_{C_k}$ are iteratively replaced by the minimal element reactions that span the removed vectors' subspace. This process is repeated until all the number of non-zero entries in $N_{C_k}$ has converged on a minimum (Bordbar et al, 2014b). Here we utilized the advantage of sparse regularization of null space $N_{C_k}$ to solve the minimum L1-norm of null space $N_{C_k}$ of $S_{C_k}$ (Saa & Nielsen, 2016). And the detailed process is showcased in Appendix file 1 Method S1, which finds $N_{C_k}^S$, a minimal sparse basis representation of $N_{C_k}$ in at most $2r_k$ linear programming optimization runs (where $r_k = l_k - rank(S_{C_k})$, $l_k$ is the number of columns (reactions) in $S_{C_k}$), and then each of linear programming problem can be formulated as follows:

$$min_{(x,v_{C_k}^m)} \sum_i x_i \tag{4}$$

$$\text{s.t. } S_{C_k}.v_{C_k}^m = 0 \tag{5}$$

$$v_{C_k}^m \leq x \tag{6}$$

$$-v_{C_k}^m \leq x \tag{7}$$

$$lb \leq v_{C_k}^m \leq ub \tag{8}$$

$$w^T.P_{C_k}^{N_m}.v_{C_k}^m > \zeta \ \vee \ w^T.P_{C_k}^{N_m}.v_{C_k}^m < -\zeta \tag{9}$$

$$x \in \mathrm{R}^n$$

where the $v_{C_k}^m$ is the SLB of null space of $S_{C_k}$at the $m$th runs, the $lb$ and $ub$ are lower and upper bound of $v_{C_k}^m$, respectively. The constraint of formula (9) ensures the $v_{C_k}^m$ is linearly independent of the previous $m-1$ SLBs, and w represents a vector of random weights. Here we employed uniform random weights, and $\zeta$ is a small positive constant, e.g., $1.0^{-3}$. $P_{C_k}^{N_m}$ is a projection matrix onto the sparse null space $N_{C_k}^{S_m}$, and $N_{C_k}^{S_m} = [v_{C_k}^1, v_{C_k}^2, \dots, v_{C_k}^{m-1}]$. More details of this process are provided in Appendix file 1 Method S1.

The SLBs of $S_{C_k}$, assumed as $N_{C_k}^S = [v_{C_k}^1, v_{C_k}^2, \dots, v_{C_k}^{r_k}]$, can be solved by above algorithm. Then, the highly-coupled subnetwork $C_k$ can be substituted to construct the mutually independent "linear basis reactions" with the following formula:

$$S_{C_k}^{IBR} = S_{C_k}^{*}.N_{C_k}^{S^*} \tag{10}$$

where $S_{C_k}^{*}$ and $N_{C_k}^{S^*}$ are derived from $S_{C_k}$ and $N_{C_k}^{S}$ after removing the artificial exchange reactions, respectively, and the $IBR$ means the reconstructed independent "linear basis reactions".

***Step3: Establishing prediction for the decoupled metabolic model***

In this step, we reformulated FBA to adapt the reconstructed decoupled metabolic network $S^{IR}$. The central problem is to determine the flux bounds for each "linear basis reactions":

$$\max_{v^{IR}} cv^{IR} \tag{11}$$

$$\text{s.t.} \quad S^{IR}.v^{IR} = 0 \tag{12}$$

$$S^{IR} = \left[S^{NC}, S_{C_1}^{IBR}, \dots, S_{C_K}^{IBR}\right] = \left[S^{NC}, S_{C_1}^{*}N_{C_1}^{S^*}, \dots, S_{C_K}^{*}N_{C_K}^{S^*}\right] \tag{13}$$

$$v^{IR} = \left[v^{NC}, v_{C_1}^{IBR}, \dots, v_{C_K}^{IBR}\right]^T \tag{14}$$

$$v_{C_k}^{IBR} = \left[v_{C_k}^1, \dots, v_{C_k}^{r_k}\right]^T \tag{15}$$

$$lb^{NC} \leq v^{NC} \leq ub^{NC} \tag{16}$$

$$\max\left(f\left(lb_{C_k}^i, NZ\left(N_{C_k}^i\right)\right)./f_N\right) \leq v_{C_k}^i \leq \min\left(f\left(ub_{C_k}^i, NZ\left(N_{C_k}^i\right)\right)./f_N\right) \tag{17}$$

$$f_N = f\left(N_{C_k}^i, NZ\left(N_{C_k}^i\right)\right) \tag{18}$$

$$i = 1, \dots, r_k; k = 1, \dots, K$$

where $c$ and $v^{IR}$ represent the objective function and optimal metabolic flux of $S^{IR}$, respectively. The superscript $IR$ represents the linearly independent reactions-derived metabolic network, and $NC$ means the non-coupled reactions (which are composed of linear reaction chains) of the original metabolic network. $r_k$ is the number of columns of $N_{C_k}^{S^*}$ and $K$ is the total number of highly-coupled reactions subnetworks identified in *step* 1. $N_{C_k}^i$ is the $i$th column of $N_{C_k}^{S^*}$, and $lb_{C_k}^i$ ,$ub_{C_k}^i$ are the lower bounds and upper bounds of the reaction indicated by $N_{C_k}^i$, respectively. $v_{C_k}^{IBR}$ is the flux vector of all the "linear basis reactions" of subnetwork $C_k$, and $v_{C_k}^i$ is the $i$th flux of $v_{C_k}^{IBR}$. Then the fluxes of reactions in the original metabolic network (element reactions) of linear basis vectors will be recovered by the formula of $N_{C_k}^{S^*}.v_{C_1}^{IBR}$ according the optimal solution of Decrem outlined above.

The function $NZ(.)$ takes the index of non-zero elements of the input vector, and the function $f(v, I)$ takes elements indexed by input indicator $I$, from the input vector $v$. So $f(N_{C_k}^i, NZ\left(N_{C_k}^i\right))$ represents taking the non-zero partition coefficient of each element reactions composed of the $i$th SLB of subnetwork $C_k$, which is indicated by the non-zero terms of $N_{C_k}^i$. The $f(lb_{C_k}^i, NZ(N_{C_k}^i)$ represents taking the lower bounds of the element reactions composed of the $i$th SLB of subnetwork $C_k$ and $f(ub_{C_k}^i, NZ\left(N_{C_k}^i\right)$ taking the upper bounds. In summary, Decrem forces the metabolic fluxes of highly-coupled reactions

to be incorporated in optimization by representing them as independent linear basis vectors.

**Gene expression estimation**

45 candidate global and local regulating metabolites of *E. coli* were collected through KEGG pathway analysis and literatures survey (Chang et al, 2016; Kothari et al, 2016; Tanabe et al, 2015). These candidates are then categorized into two clusters by hierarchical clustering analysis over the expression profile of their concentrations across 24 mutant strains. Then, the possible regulation relationship between two identified metabolites groups and expression of 85 genes are inferred by the partial least squares regression (PLS) (Wold et al, 2001), which selects the non-redundant independent factors to maximize the correlation of response variables, using the stepwise principal components regression. Furthermore, PLS is used to discover the fundamental quantitative relations between two (expression) observation variable set, and the general underlying model of multivariate PLS is:

$$X = TP^T + E \tag{19}$$

$$Y = UQ^T + F \tag{20}$$

where $X$ is an $n \times m$ matrix of predictors (each group metabolites concentrations), $Y$ is an $n \times p$ matrix of responses (genes expression); $T$ and $U$ are $n \times l$ matrices and projections of $X$ (the $X$ score, component or factor matrix) and projections of $Y$ (the $Y$ scores), respectively. $P$ and $Q$ are $m \times l$ and $p \times l$ orthogonal loading matrices, respectively. Matrices $E$ and $F$ are the error terms, assumed to be independent and identically distributed random normal variables. The decompositions of $X$ and $Y$ are made so as to maximize the covariance between $T$ and $U$.

Finally, the significant metabolite profile and corresponding explicable

genes-metabolites regulatory relationships are filtered by setting proper correlation threshold and bulid statistical test based on random sampling (see Appendix file 1 Material S3).

## The metabolite concentration derived linearized kinetic model

***Step 1: The metabolic kinetic model***

In this section, we derived a complete reversible rate law for arbitrary reactant stoichiometries. When consider the constraint of thermodynamics and metabolite regulator (Flamholz et al, 2013; Hackett et al, 2016), we can rewrite the Michaelis–Menten kinetics for a reversible reaction S <=> P as:

$$\mathrm{v}([S],[P],[A],[I],[E]) = [E]\frac{k^{+}[S]/k_m^s - k^{-}[P]/k_m^p}{1+[S]/k_m^s+[P]/k_m^p}\prod_u \frac{[A_u]/k_A^u}{1+[A_u]/k_A^u}\Bigg/\prod_v \frac{1}{1+[I_v]/k_I^v} \quad (21)$$

where $[E]$ is the concentration of enzyme active sites, and $[S]$ and $[P]$ are the concentrations of substrates and products, $k_m^s$ and $k_m^p$ are the affinities of the reactants for this enzyme, and $k^+$ and $k^-$ are the maximal forward and reverse catalytic rate constants, $[A]$ and$[I]$ are the concentrations of activators and inhibitors, and $k_a^{A_u}$ and $k_i^{I_v}$ are the corresponding affinities of them. The positive and negative term in the numerator are associated with the forward and backward rate, respectively.

We next applied the metabolic thermodynamics constraints given by the Haldane relationship to simplify the term for the backward rate:

$$k^{eq} = \frac{k^{+}k_m^p}{k^{-}k_m^s} = \frac{[P_0]}{[S_0]} = e^{-\Delta_r G'^{\circ}/RT} \quad (22)$$

where $\Delta_r G'^{\circ}$ is the standard Gibbs energy of the reaction (and does not depend on the enzyme parameters). Using this equality with the above rate law, we can get:

$$\mathrm{v}([S],[P],[A],[I],[E])$$

$$= ([E]k^{+})\frac{[S]/k_m^s\,(1-[P]/[S]\,e^{-\Delta_r G'^{\circ}/RT})}{1+[S]/k_m^s+[P]/k_m^p}\prod_u \frac{[A_u]/k_A^u}{1+[A_u]/k_A^u}\Bigg/\prod_v \frac{1}{1+[I_v]/k_I^v} \tag{23}$$

***Step 2: The optimal strategy of linearized kinetic model***

In this section, we seek the simplified representation of the equation (23) based only on the associated metabolites concentrations. Firstly, we took the negative logarithmic operation of the both side of equation (23) and reorganized the right hand terms:

$$-\log\big(\mathrm{v}([S],[P],[A],[I],[E])\big) = \underbrace{-\log([E])}_{\text{Enzyme term}} \underbrace{-\log\big(1-[P]/[S]\,e^{-\Delta_r G'^{\circ}/RT}\big)+\sum_v \log(1+[I_v])}_{\text{Metabolite associated terms}}$$

$$\underbrace{-\log(k^{+})-\sum_v k_I^v}_{\text{Kinetic constants}} + \underbrace{\sum_u \log\left(1+\frac{k_A^u}{[A_u]}\right)+\log\left(1+\frac{k_m^s}{[S]}\left(1+\frac{[P]}{k_m^p}\right)\right)}_{\text{Nonlinear terms}} \tag{24}$$

This model can be solved by collecting the corresponding kinetic parameters, enzyme expression and metabolite concentrations, however, those matched data often are not available in practice, also, the metabolic regulators often are unknown. An alternative is approximate the optimal parameters according the machine learning method. Specifically, the cell growth state regulated enzymes expression can be represented as the linear combination of the concentrations of biomass composition (see Appendix file 1 Material S3), which can be marked as: $[E]=\sum_{i=1}^{m} w_i[S_i]+b$, $S_i$ is the concentration of $i$th indicator metabolite of biomass identified in the section of "Gene expression estimation". In addition, we reexamine the nonlinear terms of equation (24) based on the knowledge that systemic experimental analysis revealed that there are $[S] \geq k_m^s$ for almost of the metabolites in

central metabolism of three model organisms (Park et al, 2016), hence we have $\frac{k_m^s}{[S]} \leq 1$.

Situation 1: For the growth state regulated irreversible reaction , we have $\frac{[P]}{k_m^p} \to 0$, and the $\frac{k_A^u}{[A_u]} \leq 1$, then the Taylor expansion of equation (24) is:

$$-\log\big(\mathrm{v}([S],[P],[A],[I],[E])\big)$$

$$= -\log\left(\sum_{i=1}^{m} w_i[S_i] + b\right) - \log\big(1 - [P]/[S]\, e^{-\Delta_r G'^{\circ}/RT}\big) + \sum_{v} \log(1 + [I_v])$$

$$-\log(\boldsymbol{k}^{+}) - \sum_{v} \boldsymbol{k_I^v} + \frac{\boldsymbol{k_m^s}}{[S]} + \sum_{u}\left(\frac{\boldsymbol{k_A^u}}{[A_u]}\right) + O\left(\frac{k_m^s}{[S]}, \frac{k_A^u}{[A_u]}\right) \tag{25}$$

Situation 2: For the generalization reversible reaction in which the enzyme is substrate or product saturated (or both): $\frac{k_m^s}{[S]} \leq 1$ and $\frac{k_m^p}{[P]} \leq 1$. Assuming $v > 0$,we have:

$$-\log(k^{+}) + \log\left(1 + \frac{k_m^s}{[S]}\left(1 + \frac{[P]}{k_m^p}\right)\right) \approx \log\left(\frac{1}{k^{+}} + \frac{1}{k^{+}}\frac{[P]}{[S]}\frac{k_m^s}{k_m^p}\right) = \log\left(\frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}}\right) \tag{26}$$

Here, $[P]/[S] < k^{eq}$ according the reaction Gibbs energy, $\frac{1}{k^{+}} \leq 1$ and $\frac{1}{k^{-}} \leq 1$ for the significant regulated reactions, and $\frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}} - 1 < 1$, then the Taylor expansion of (26) is:

$$\log\left(\frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}}\right) = \log\left(1 + \frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}} - 1\right)$$

$$= \frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}} - 1 + O\left(\frac{1}{k^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{k^{-}}\right) \tag{27}$$

$$-\log\big(\mathrm{v}([S],[P],[A],[I],[E])\big)$$

$$\approx -\log\left(\sum_{i=1}^{m} w_i[S_i] + b\right) - \log\big(1 - [P]/[S]\, e^{-\Delta_r G'^{\circ}/RT}\big) + \sum_{v} \log(1 + [I_v])$$

$$-\sum_{v} \boldsymbol{k_I^v} + \sum_{u}\left(\frac{\boldsymbol{k_A^u}}{[A_u]}\right) + \frac{1}{\boldsymbol{k}^{+}} + \frac{[P]/[S]}{k^{eq}}\frac{1}{\boldsymbol{k}^{-}} - 1 \tag{28}$$

Where all the bold variables in equation (25) and (28) are kinetic parameters. Similarly, when $v > 0$, we have $-\log(k^{-}) + \log\left(1 + \frac{k_m^p}{[P]}\left(1 + \frac{[S]}{k_m^s}\right)\right) \approx \log\left(\frac{1}{k^{-}} + \frac{[S]/[P]}{k^{eq}}\frac{1}{k^{+}}\right)$. And this result can be expanded to multi-substrate/multiproduct reactions.

Overall, when we neglect the infinitesimal of higher order, the identified regulators and optimal kinetic parameters in model (25) and (28) can be solved with a Bayesian linear regression derived model selection method, i.e. the SIMMER package (Hackett et al, 2016), which firstly build the regulator-free kinetic model, and then use the BIC criteria to identify the real regulators from the candidate metabolites set (PG metabolites), by evaluating the additional ability to account for the empirical fluxes when integrating a new candidate regulator to the kinetic model in an incremental way (see Appendix file 1 Methods S3 ), after, the optimal model is expand to any another application.

***Step 3: The kinetic Decrem***

Finally, the parameterized kinetic model is utilized to describe the growth-associated key regulated reactions in central metabolism. And the genome-scale flux distribution is predicted by the kinetic regulated fluxes constrained Decrem method. i.e.:

$$\max cv_{obj} \tag{29}$$

$$\text{s.t.}\ \ S.v\ = 0 \tag{30}$$

$$lb^{NC} \leq v^{NC} \leq ub^{NC} \tag{31}$$

$$\max\left(f\left(lb_{C_k}^i, NZ\left(N_{C_k}^i\right)\right)./ff_N\right) \leq v_{C_k}^i \leq \min\left(f\left(ub_{C_k}^i, NZ\left(N_{C_k}^i\right)\right)./f_N\right) \tag{32}$$

$$f_N = f\left(N_{C_k}^i, NZ\left(N_{C_k}^i\right)\right) \tag{33}$$

$$v_j^{KF} - \delta \leq v_j \leq v_j^{KF} + \delta \tag{34}$$

$$i = 1, \dots, r_k; j = 1, \dots, \mathrm{m}\,; k = 1, \dots, K$$

Where the $v_j^{KF}$ is the $j$-th kinetic flux of the $m$ key regulation reactions, and the $\delta$ is the tolerance of kinetic fluxes. And the $i$ ,$k$ ,$NC$, $lb_{C_k}^i$, $NZ\left(N_{C_k}^i\right)$ can be found in *step* 3 of the section “The topology-decoupled reconstruction of metabolic model”.

## Acknowledgements

The authors thank Schellenberger J's open-source Matlab package: Cobra, and Saa PA's open-source Matlab package: Fast_SNP for the integration of Decrem. This work was supported by the National Natural Science Foundation of China (61872418) and Natural Science Foundation of Jilin Province (20180101050JC).

## Author contributions

GY.L. and HS.C. conceived and designed the work. GY.L. carried out computer implementation and data analysis. GY.L., CM.Z., HS.C, Q.L and W.D. interpreted the simulation results. GY.L., HS.C. and CM.Z. wrote the original manuscript, and P.W., Q.L., W.D., CM.Z. and HS.C. contributed to the writing of final manuscript.

## Conflict of interest

The authors declare they have no competing interests.

## Data and software availability

The original and reconstructed metabolic models, data used in this work, and the developed computational method are all available online:

- Original metabolic models: all the used metabolic models are available at BIGG models http://bigg.ucsd.edu/models.
- Reconstructed metabolic models: the three reconstructed models, *i*AF1260, *i*MM904, iDN750, are available at https://github.com/lgyzngc/Decrem.git.
- Genome-scale LS-MS data: https://www.ebi.ac.uk/biostudies/.
- Decrem method: https://github.com/lgyzngc/Decrem.git.

## References

Ahnert SE, Fink TMA (2016) Form and function in gene regulatory networks: the structure of network motifs determines fundamental properties of their dynamical state space. *Journal of The Royal Society Interface* **13:** 20160179

Alper H, Jin Y-S, Moxley JF, Stephanopoulos G (2005) Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metabolic Engineering* **7:** 155-164

Berg JM TJ, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; (2002) The Metabolism of Glucose 6-Phosphate by the Pentose Phosphate Pathway Is Coordinated with Glycolysis. *Biochemistry 5th edition*

Berg JM, Tymoczko, J. L., and Stryer, L. (2007a) The citric acid cycle. *In Biochemistry (6th ed)***:** 2

Berg JM, Tymoczko, J. L., and Stryer, L. (2007b) Glycolysis and gluconeogenesis. *In Biochemistry (6th ed)***:** 2

Blank LM, Kuepfer L, Sauer U (2005) Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biology* **6:** R49

Bordbar A, Monk JM, King ZA, Palsson BO (2014a) Constraint-based models predict metabolic and associated cellular functions. *Nature Reviews Genetics* **15:** 107

Bordbar A, Nagarajan H, Lewis NE, Latif H, Ebrahim A, Federowicz S, Schellenberger J, Palsson BO (2014b) Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular Systems Biology* **10**

Bordbar A, Nagarajan H, Lewis NE, Latif H, Ebrahim A, Federowicz S, Schellenberger J, Palsson

BO (2014c) Minimal metabolic pathway structure is consistent with associated biomolecular interactions. *Molecular Systems Biology* **10:** 737

Braakman R, Smith E (2012) The compositional and evolutionary logic of metabolism. *Physical Biology* **10:** 011001

Chang A, Schomburg I, Tillack J, Jeske L, Ulbrich M, Placzek S, Schomburg D (2016) BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Research* **45:** D380-D388

Chavali AK, D'Auria KM, Hewlett EL, Pearson RD, Papin JA (2012) A metabolic network approach for the identification and prioritization of antimicrobial drug targets. *Trends in Microbiology* **20:** 113-123

Chen X, Alonso AP, Allen DK, Reed JL, Shachar-Hill Y (2011) Synergy between 13C-metabolic flux analysis and flux balance analysis for understanding metabolic adaption to anaerobiosis in E. coli. *Metabolic Engineering* **13:** 38-48

Covert MW, Palsson BØ (2002) Transcriptional Regulation in Constraints-based Metabolic Models of Escherichia coli. *Journal of Biological Chemistry* **277:** 28058-28064

Feist AM, Palsson BØ (2008) The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nature biotechnology* **26:** 659-667

Fell DA, Small JR (1986) Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochemical Journal* **238:** 781-786

Flamholz A, Noor E, Bar-Even A, Liebermeister W, Milo R (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proceedings of the National Academy of Sciences* **110:** 10039-10044

Fong SS (2014) Computational approaches to metabolic engineering utilizing systems biology

and synthetic biology. *Computational and Structural Biotechnology Journal* **11:** 28-34

Gerstl MP, Zanghellini J, Müller S, Regensburger G (2018) Flux tope analysis: studying the coordination of reaction directions in metabolic networks. *Bioinformatics* **35:** 266-273

Gonzalez JE, Long CP, Antoniewicz MR (2017) Comprehensive analysis of glucose and xylose metabolism in Escherichia coli under aerobic and anaerobic conditions by 13C metabolic flux analysis. *Metabolic Engineering* **39:** 9-18

Gudmundsson S, Thiele I (2010) Computationally efficient flux variability analysis. *BMC bioinformatics* **11:** 489-489

Guyon I, Andr, #233, Elisseeff (2003) An introduction to variable and feature selection. *J Mach Learn Res* **3:** 1157-1182

Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, Gibney PA, Botstein D, Storey JD, Rabinowitz JD (2016) Systems-level analysis of mechanisms regulating yeast metabolic flux. *Science* **354:** aaf2786

Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, Ho PY, Kakazu Y, Sugawara K, Igarashi S, Harada S, Masuda T, Sugiyama N, Togashi T, Hasegawa M, Takai Y et al (2007) Multiple High-Throughput Analyses Monitor the Response of E. coli to Perturbations. *Science* **316:** 593-597

Khodayari A, Chowdhury A, Maranas CD (2015) Succinate Overproduction: A Case Study of Computational Strain Design Using a Comprehensive Escherichia coli Kinetic Model. *Frontiers in bioengineering and biotechnology* **2:** 76-76

Khodayari A, Maranas CD (2016) A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communications* **7:** 13806

Khodayari A, Zomorrodi AR, Liao JC, Maranas CD (2014) A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data. *Metabolic Engineering* **25:** 50-62

Kim J, Fabris M, Baart G, Kim MK, Goossens A, Vyverman W, Falkowski PG, Lun DS (2016) Flux balance analysis of primary metabolism in the diatom Phaeodactylum tricornutum. *The Plant Journal* **85:** 161-176

Kim J, Reed JL (2012) RELATCH: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biology* **13:** R78

Kochanowski K, Gerosa L, Brunner SF, Christodoulou D, Nikolaev YV, Sauer U (2017) Few regulatory metabolites coordinate expression of central metabolic genes in Escherichia coli. *Molecular Systems Biology* **13:** 903

Kothari A, Fulcher C, Weaver D, Keseler IM, Latendresse M, Krummenacker M, Subhraveti P, Ong Q, Billington R, Caspi R, Paley S, Karp PD, Mackie A, Paulsen I, Santos-Zavaleta A, Bonavides-Martínez C, Velázquez-Ramírez DA, Collado-Vides J, Muñiz-Rascado L, Peralta-Gil M et al (2016) The EcoCyc database: reflecting new knowledge about Escherichia coli K-12. *Nucleic Acids Research* **45:** D543-D550

Latendresse M (2014) Efficiently gap-filling reaction networks. *BMC Bioinformatics* **15:** 225

Lee KH, Park JH, Kim TY, Kim HU, Lee SY (2007) Systems metabolic engineering of Escherichia coli for L-threonine production. *Molecular systems biology* **3:** 149-149

Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, Weitz KK, Eils R, König R, Smith RD, Palsson BØ (2010) Omic data from evolved E. coli are consistent with computed optimal growth from

genome-scale models. *Molecular Systems Biology* **6:** 390

Li G, Xu Y, Cao H (2018) Structural and functional analyses of microbial metabolic networks reveal novel insights into genome-scale metabolic fluxes.

Li Z, Wang R-S, Zhang X-S (2011) Two-stage flux balance analysis of metabolic networks for drug target identification. *BMC Systems Biology* **5:** S11

Long MR, Reed JL (2016) Improving flux predictions by integrating data from multiple strains. *Bioinformatics* **33:** 893-900

Machado D, Herrgård M (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLOS Computational Biology* **10:** e1003580

McCloskey D, Xu S, Sandberg TE, Brunk E, Hefner Y, Szubin R, Feist AM, Palsson BO (2018) Evolution of gene knockout strains of E. coli reveal regulatory architectures governed by metabolism. *Nature Communications* **9:** 3796

Meilă M, Pentney W (2007) Clustering by weighted cuts in directed graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pp 135-144.

Millard P, Smallbone K, Mendes P (2017) Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli. *PLOS Computational Biology* **13:** e1005396

Moxley JF, Jewett MC, Antoniewicz MR, Villas-Boas SG, Alper H, Wheeler RT, Tong L, Hinnebusch AG, Ideker T, Nielsen J, Stephanopoulos G (2009) Linking high-resolution metabolic flux phenotypes and transcriptional regulation in yeast modulated by the global regulator Gcn4p. *Proceedings of the National Academy of Sciences* **106:** 6477-6482

Noor E, Eden E, Milo R, Alon U (2010) Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. *Molecular Cell* **39:** 809-820

O'Brien Edward J, Monk Jonathan M, Palsson Bernhard O (2015) Using Genome-scale Models to Predict Biological Capabilities. *Cell* **161:** 971-987

Okada T, Mochizuki A (2017) Sensitivity and network topology in chemical reaction systems. *Physical Review E* **96:** 022322

Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? *Nature Biotechnology* **28:** 245

Pandit AV, Srinivasan S, Mahadevan R (2017) Redesigning metabolism based on orthogonality principles. *Nature Communications* **8:** 15188

Park JO, Rubin SA, Xu Y-F, Amador-Noguez D, Fan J, Shlomi T, Rabinowitz JD (2016) Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nature Chemical Biology* **12:** 482

Piazza I, Kochanowski K, Cappelletti V, Fuhrer T, Noor E, Sauer U, Picotti P (2018) A Map of Protein-Metabolite Interactions Reveals Principles of Chemical Communication. *Cell* **172:** 358-372.e323

Raman K, Chandra N (2009) Flux balance analysis of biological systems: applications and challenges. *Briefings in Bioinformatics* **10:** 435-449

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási A-L (2002) Hierarchical Organization of Modularity in Metabolic Networks. *Science* **297:** 1551-1555

Reznik E, Christodoulou D, Goldford JE, Briars E, Sauer U, Segrè D, Noor E (2017) Genome-Scale Architecture of Small Molecule Regulatory Networks and the Fundamental

Trade-Off between Regulation and Enzymatic Activity. *Cell Reports* **20:** 2666-2677

Rühl M, Le Coq D, Aymerich S, Sauer U (2012) 13C-flux Analysis Reveals NADPH-balancing Transhydrogenation Cycles in Stationary Phase of Nitrogen-starving Bacillus subtilis. *Journal of Biological Chemistry* **287:** 27959-27970

Saa PA, Nielsen LK (2016) Fast-SNP: a fast matrix pre-processing algorithm for efficient loopless flux optimization of metabolic models. *Bioinformatics* **32:** 3807-3814

Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology* **13:** 935

Satish Kumar V, Dasika MS, Maranas CD (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* **8:** 212

Segrè D, Vitkup D, Church GM (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences* **99:** 15112-15117

Shlomi T, Berkman O, Ruppin E (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America* **102:** 7695-7700

Stalidzans E, Seiman A, Peebo K, Komasilovs V, Pentjuss A (2018) Model-based metabolism design: constraints for kinetic and stoichiometric models. *Biochemical Society Transactions*: BST20170263

Steuer R, Gross T, Selbig J, Blasius B (2006) Structural kinetic modeling of metabolic networks. *Proceedings of the National Academy of Sciences* **103:** 11868-11873

Tanabe M, Furumichi M, Kanehisa M, Kawashima M, Sato Y (2015) KEGG as a reference

resource for gene and protein annotation. *Nucleic Acids Research* **44:** D457-D462

Tian M, Reed JL (2018) Integrating proteomic or transcriptomic data into metabolic models using linear bound flux balance analysis. *Bioinformatics* **34:** 3882-3888

Tibshirani R (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B (Methodological)* **58:** 267-288

Venayak N, von Kamp A, Klamt S, Mahadevan R (2018) MoVE identifies metabolic valves to switch between phenotypic states. *Nature Communications* **9:** 5332

Voet D, Voet JG (2010) *Biochemistry, 4th Edition*: John Wiley & Sons.

Wang L, Dash S, Chan SHJ, Maranas CD (2018) Accelerating flux balance calculations in genome-scale metabolic models by localizing the application of loopless constraints. *Bioinformatics* **34:** 4248-4255

Wold S, Sjöström M, Eriksson L (2001) PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58:** 109-130

Xie Z, Zhang T, Ouyang Q (2017) Genome-scale fluxes predicted under the guidance of enzyme abundance using a novel hyper-cube shrink algorithm. *Bioinformatics* **34:** 502-510

Zelezniak A, Vowinckel J, Capuano F, Messner CB, Demichev V, Polowsky N, Mülleder M, Kamrad S, Klaus B, Keller MA, Ralser M (2018) Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts. *Cell Systems* **7:** 269-283.e266

## Figures

**Figure 1. An overview of kinetic Decrem in central metabolism model.**

The composite regulations of central metabolism: (1) the schematic representation of gene (enzyme) expression predicted from the metabolite concentrations; (2) the schematic of identification of metabolic regulators according to Bayesian model selection; (3) the schematic decomposition of topologically-coupled reactions. (4) minimal independent reaction components (the sparse linear basis (SLB) of corresponding metabolic stoichiometric matrix) derived from topologically coupled reactions (highlighted yellow at the bottom right) are utilized to constrain the coactivation fluxes in Decrem (highlighted yellow on the bottom left); the enzyme (genes) expression (the light red box at the top right) derived from global cell state-associated metabolite concentrations, and the essential metabolites-associated metabolic regulators (the light blue at the middle right), are utilized to reconstructed the kinetic fluxes of the key regulatory reactions to constrain the canonical CBMs on the left plot (dark blue and red).

**Figure 2. The statistical analysis of decomposed topological components of *E. coli i*AF1260 model.**

(a) The distribution of proportion of coupled reactions (green bar) among the 41 metabolic pathways.

(b) the comparison of accumulated distribution of kinetic constant, $K_m$, between the coupled reactions (red line) and non-coupled reactions (green line), the vertical bars are the mean $K_m$ values of the two groups.

(c) the correlation distribution of $^{13}C$ MFA fluxes within the each of three pathways of central metabolism, compared to the corresponding distantly-coupled reactions (examined

reactions from different pathways) in 24 strains.

(d) the correlation distribution of gene expression among the element reactions of sparse linear basis (SLB) vectors against the gene expression among the non-coupled reactions; (e) the accumulated flux distribution of reconstructed decomposition metabolic model vs the corresponding original metabolic model according the flux variability analysis.

**Figure 3. Comparison of the predictions in three model microorganisms by FBA, pFBA, RELATCH and Decrem.**

(a) Metabolic flux distribution between the experimental measurements and predictions by the four methods in *E. coli i*AF1260 with four conditions: glucose (Gly) or xylose (Oxy) under aerobic (Aer) or anaerobic (Ana) respiration, as well as in the *S. cerevisiae i*MM904 model and *B. subtilis i*YO844 model with glucose under aerobic respiration. *r* means the Pearson correlation coefficient between the measured $^{13}C$ MFA fluxes and the predicted fluxes.

(b) Comparison of MSE among the four test methods across the 6 growth conditions.

(c) Comparison of activated reactions among the four test methods across the 6 growth conditions.

**Figure 4. Comparison of the flux predictions in two *S. cerevisiae* metabolic models.**

The Pearson and Spearman correlation between the $^{13}C$ MFA measured and predicted metabolic fluxes across each of the five test methods (FBA, pFBA, RELATCH, REPPS and Decrem) among 38 strains in the *i*MM904 model and *i*DN750 model of *S. cerevisiae*.

**Figure 5. Cell state predictions for *S. cerevisiae* though FBA, pFBA, RELATCH, REPPS, Decrem and pDecrem.**

(a) Correlation of the experimentally measured (*y*-axis) and predicted (*x*-axis) growth rates with PCA regression of the top two PCs from the six test methods.

(b) The top two principal components (PCs) of predicted flux variances among the 38 mutant strains, according to the six test methods. The deleted genes with high negative effects on growth are labeled in red and the genes with no significant effects on growth are labeled in black.

(c) The heatmap of predicted fluxes by Decrem through 30 pathways revealed different responding mechanisms among the ZWF1 and the ALD6 mutant, as well as the RPE1 (highlighted pathways in green box indicates the ZWF1 deletion strain specific flux distribution).

**Figure 6. The analysis of metabolites-based kinetic regulation model of *E.coli* in 24 mutant strains.**

(a) The predictions of 16 genes expression using the 23 biomass-associated metabolites by the partial least squares regression.

(b) The distribution of correlation between the measured and predicted gene expression using the metabolites from 10,000 random sampling (red curve), as well as the prediction significance according to the biomass metabolites group (green line) and the precursor metabolites group (blue line).

(c) Predicted kinetic fluxes for three key regulatory reactions according to

metabolites-determined Michaelis-Menten kinetics.

**Figure 7. Annotation of regulators and genes representable by the biomass in the central metabolism.**

The meaning of different backgrounds and colors: red and green metabolites are metabolic inhibitors and activators; the blue genes present the identified 32 global growth state associated genes. The brown background indicates the identified key regulated reactions which were used to build the kinetic Decrem.

**Figure 8. The growth rate analysis on the genome-scale deletion strains.**

(a) The comparison of measured and predicted growth rates by the kinetic pDecrem, Decrem, FBA, pFBA, RELATCH and REPPS, on the 1,030 mutants.

(b) The correlation between the measured growth rates and the predicted accumulated fluxes though the identified growth-associated pathways.

(c) The comparison of the predicted accumulated growth rates through the pathways where the mutants belong, according to the 6 test methods.

(d) The top two PCs revealed difference of flux variances between the coupled reactions (cyan dots) and the non-coupled reactions (purple dots) which are predicted by the Decrem according to the 1,030 strains.

(e) The comparison of growth rates distribution among topological components where the mutants belong: the complex “linear basis reaction” (the number of element reactions of corresponding sparse linear basis vector > 1), the simple “linear basis reaction” (the number

of element reactions of corresponding sparse linear basis vector = 1) and the non-coupled reactions.

(f) The flux variance of "linear basis reactions" based on the multimeric classification: the multi-subunit enzymes (green dots) and single-unit enzymes (red dots), the number of simple cycles and the length (the number of element reactions) of "linear basis reactions" (Fisher's Exact Test, *p* values = 2.34E-21 and 0.0051).

## Appendix files

**Appendix file 1: Figure S1**: Comparison of activated reactions of TCA cycle among 13C MFA, Decrem and FBA; **Figure S2**: The workflow of linear flux balance analysis (Decrem); **Figure S3**: The accumulation distribution of Michaelis constant Km according the Yeast iMM904 model and reconstructed decoupled model; **Figure S4**: The workflow of kinetic regulation model; **Figure S5**: The concentration profiles of 45 candidate regulating metabolites according to the hierarchical clustering, under 24 mutant strains; **Figure S6**: The consistency of 32 genes between measured and predicted gene expression levels by the PLS methods; **Figure S7:** The growth rate distribution of the 1030 metabolic genes knockout strains predicted by the conventional methods: FBA, MOMA and Decrem, respectively; **Material S1**: the data processing in the section of "Prediction of metabolic fluxes in response to environmental perturbation"; **Material S2**: the data processing in the section of "Metabolic flux prediction in single-gene deletion strains"; **Material S3**: the workflow and data processing in the section of "Metabolite concentration uncovering the metabolic regulation-enhanced kinetic model"; **Material S4**: the workflow and data processing in the section of "Growth estimation for E.

coli genome-scale deletion strains with kinetic Decrem"; **Method S1**: the Decrem algorithm. **Method S2**: The Parsimonious Decrem; **Method S3**: identifying the high confidence regulators by Bayesian model section method.

**Appendix file 2**: **Table S1**. The annotation of reconstruction decoupled metabolic models.

**Appendix file 3**: **Table S2**. Nutrient uptake reactions for three microorganisms under six growth conditions.

**Appendix file 4**: **Table S3**. The 13C MFA fluxes for three microorganisms under six growth conditions.

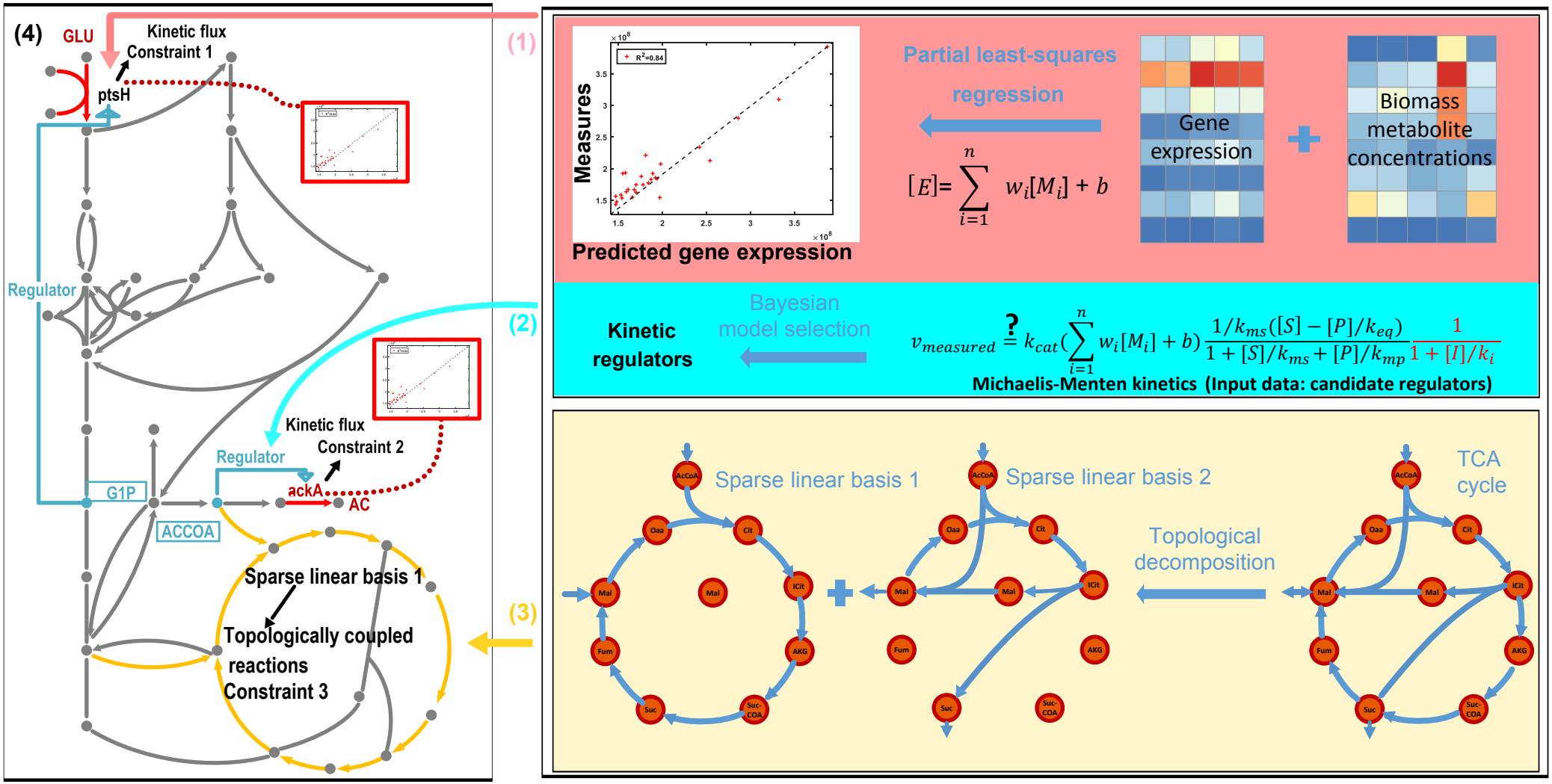
**Appendix file 5**: **Table S4**. The dataset used to predict the fluxes and growth rates for the Yeast 38 deletion strains.

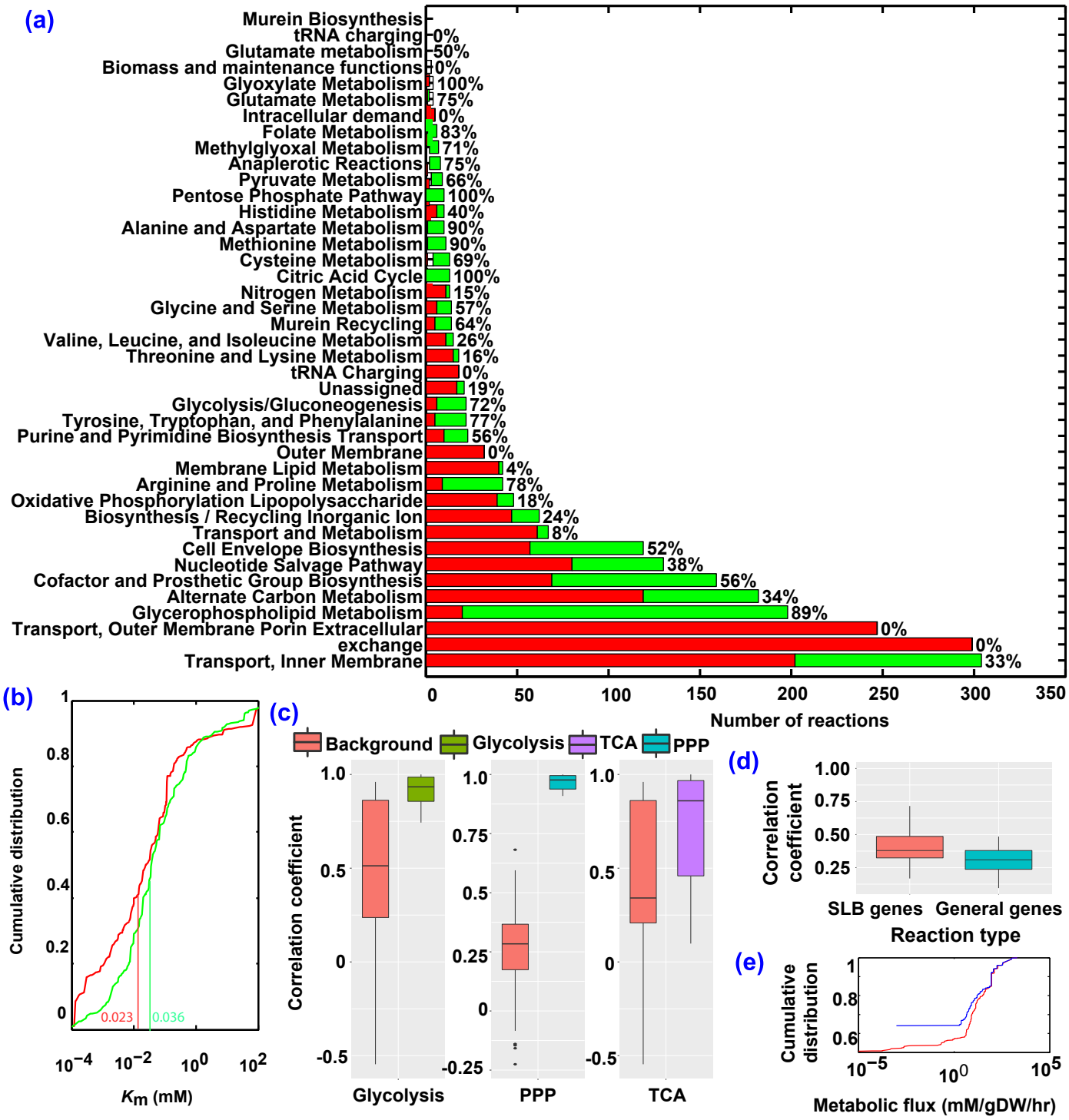
**Appendix file 6**: **Table S5**. The correlation statistic between the experimental and predicted metabolic fluxes and growth rates for the Yeast 38 deletion strains.

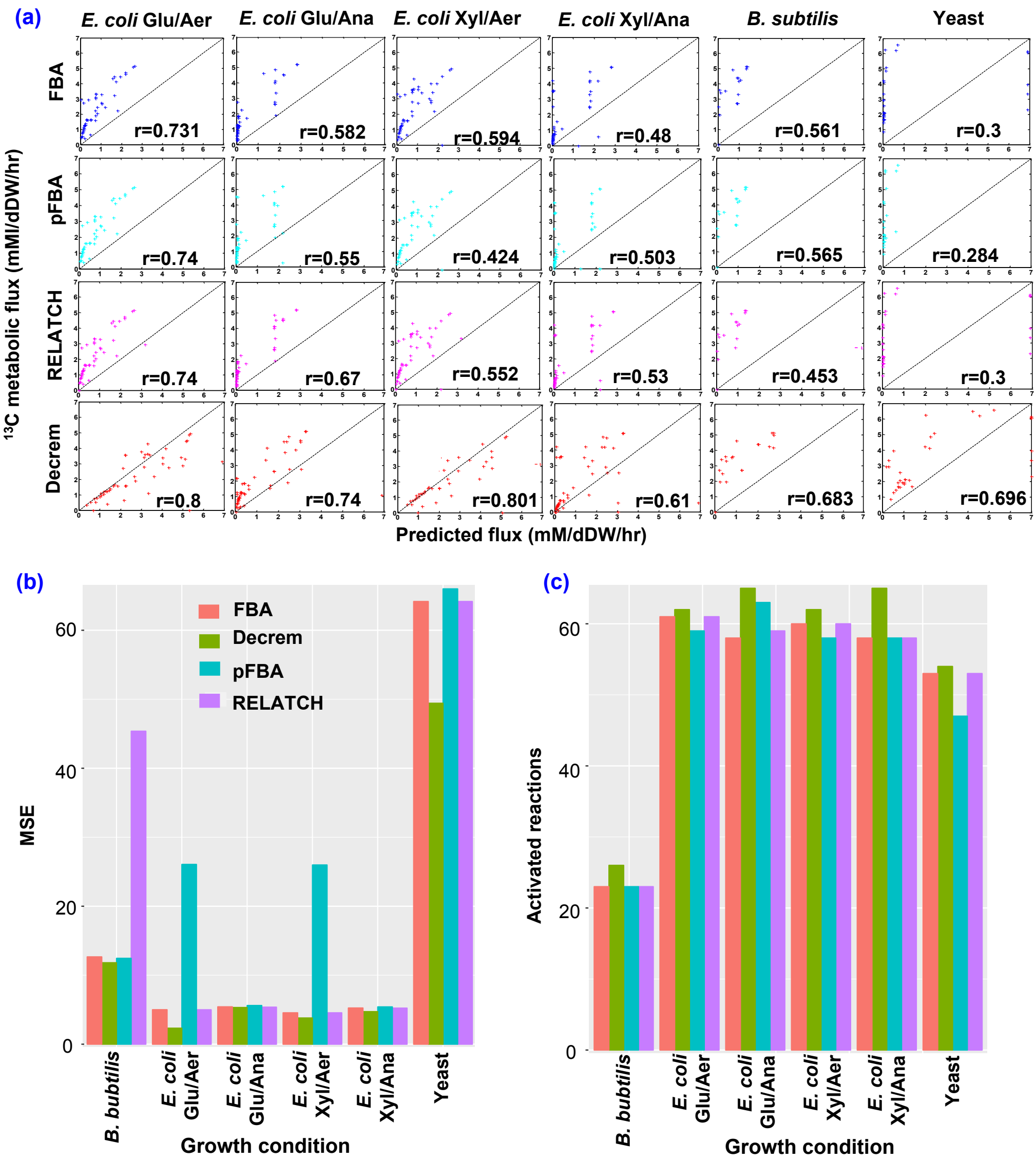
**Appendix file 7**: **Table S6**. The dataset used to build and validate the hierarchical regulation model.

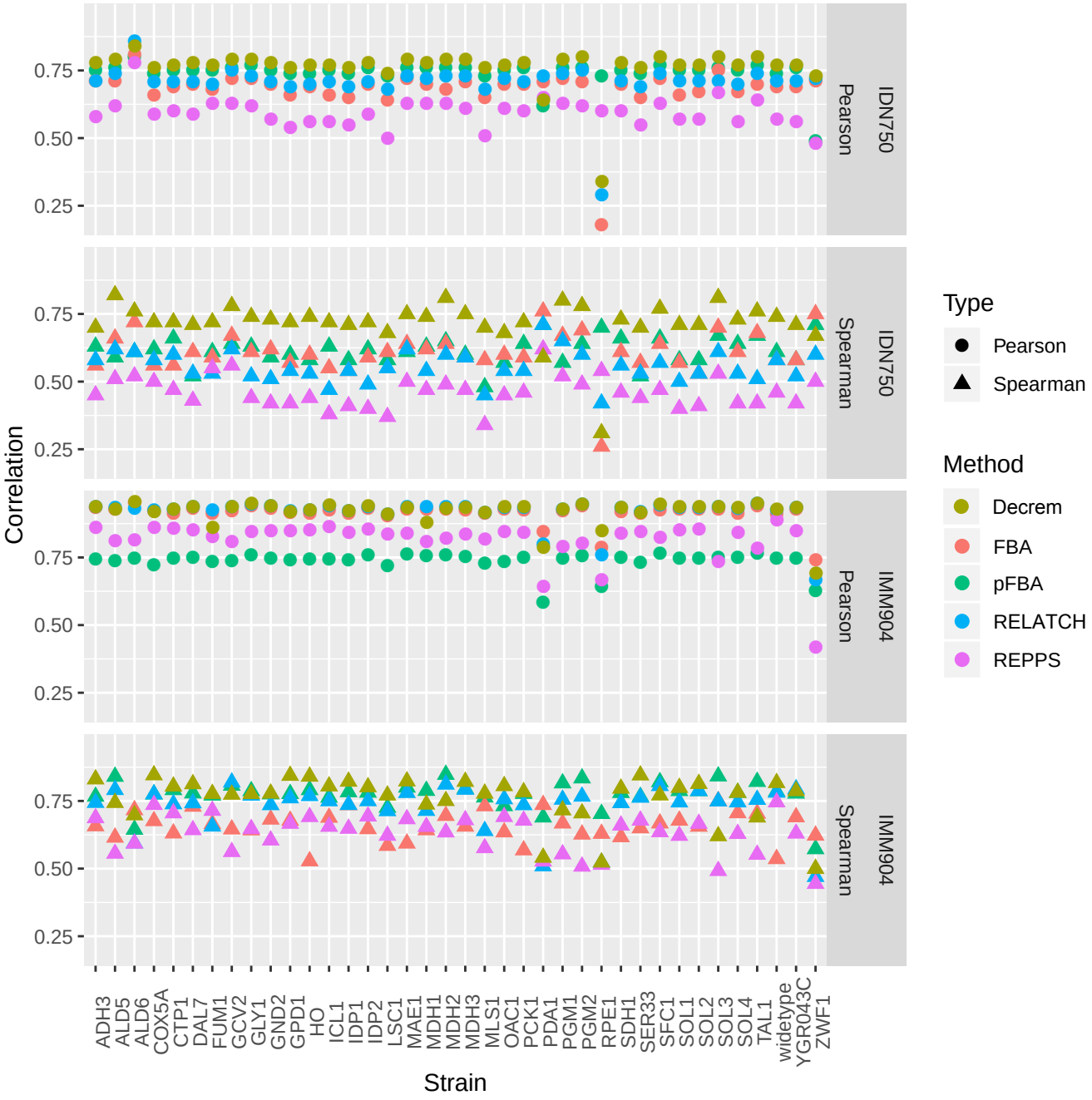
**Appendix file 8**: **Table S7**. The dataset used to construct kinetic Decrem to predict the

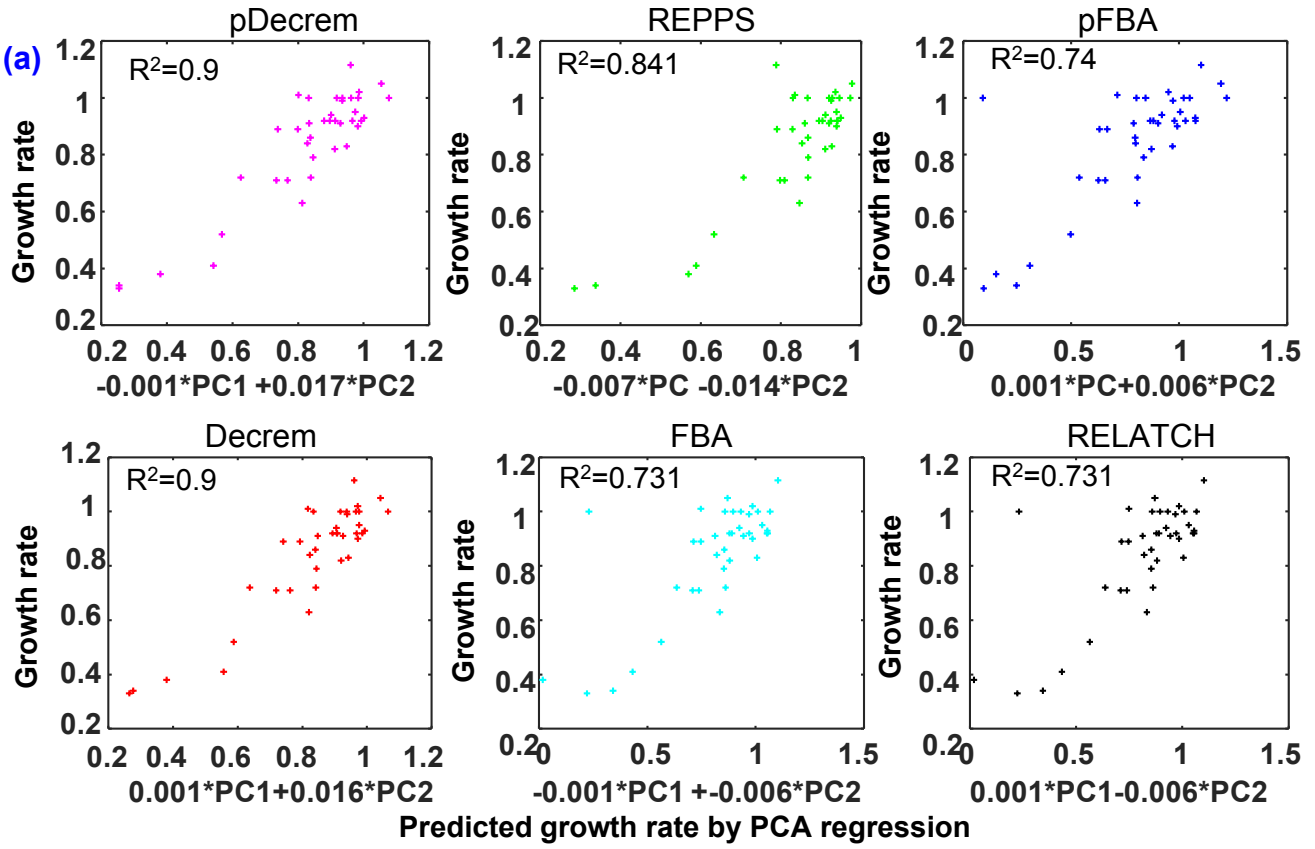
1 growth rates on genome-scale deletion strains.

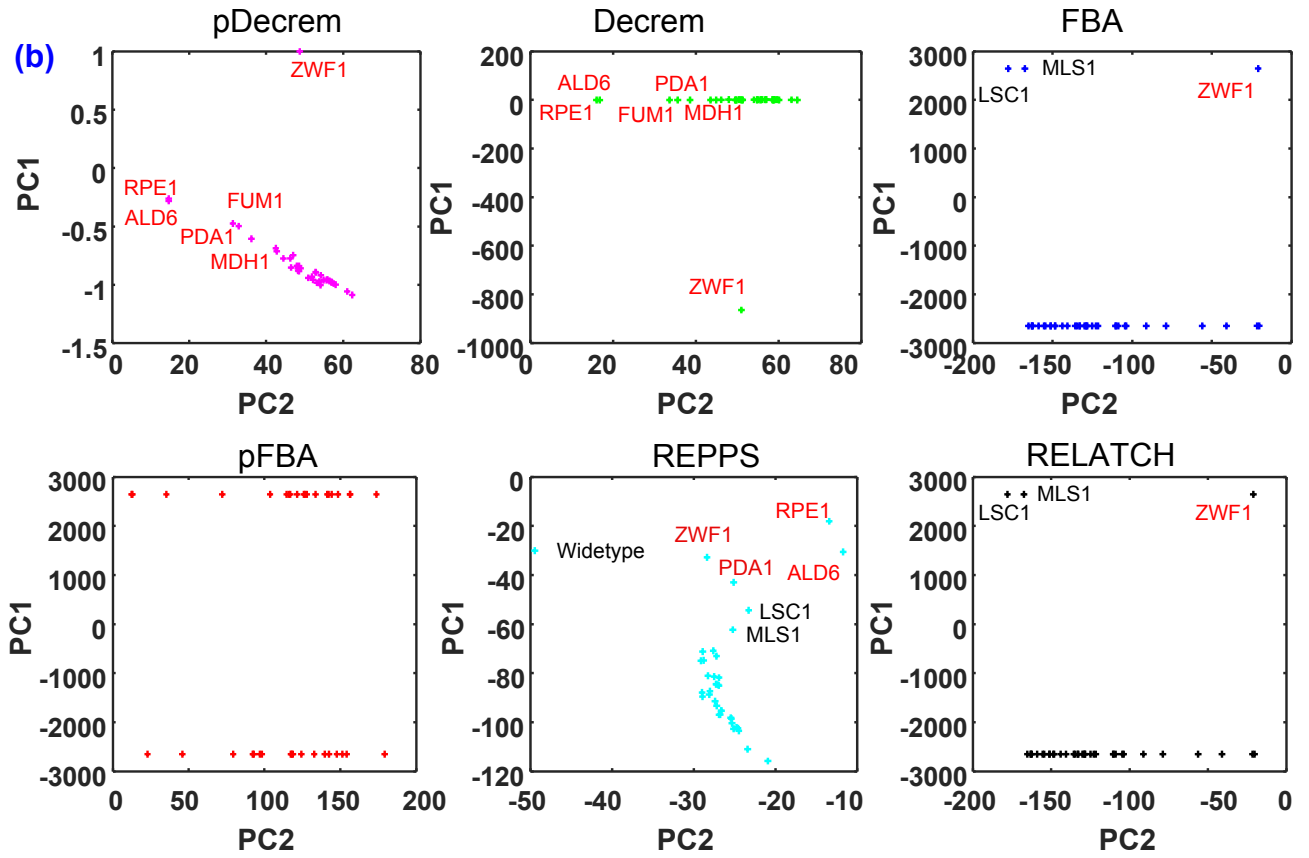

(4)
GLU
Kinetic flux
Constraint 1
ptsH
Regulator
Kinetic flux
Constraint 2
Regulator
G1P
ackA
AC
ACCOA
Sparse linear basis 1
Topologically coupled
reactions
Constraint 3
(1)
(2)
(3)
Measures
Predicted gene expression
Partial least-squares
regression
$[E]=\sum_{i=1}^{n} w_i[M_i] + b$
Gene
expression
Biomass
metabolite
concentrations
Kinetic
regulators
Bayesian
model selection
$v_{measured} \overset{?}{=} k_{cat}(\sum_{i=1}^{n} w_i[M_i] + b)\frac{1/k_{ms}([S]-[P]/k_{eq})}{1+[S]/k_{ms}+[P]/k_{mp}}\frac{1}{1+[I]/k_i}$
Michaelis-Menten kinetics (Input data: candidate regulators)
Sparse linear basis 1
Sparse linear basis 2
Topological
decomposition
TCA
cycle
AcCoA
Oaa
Cit
ICit
AKG
Suc-
COA
Suc
Fum
Mal

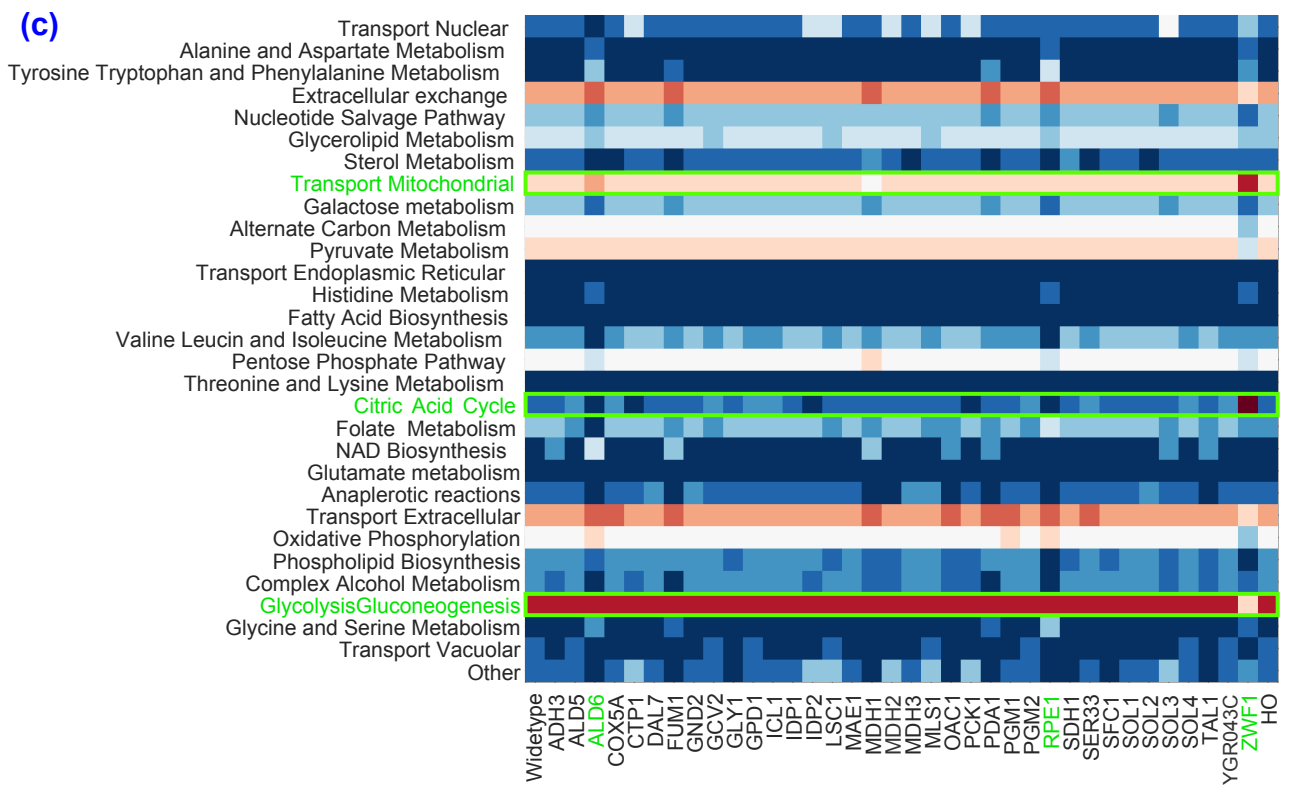

(a)
Murein Biosynthesis
tRNA charging 0%
Glutamate metabolism 50%
Biomass and maintenance functions 0%
Glyoxylate Metabolism 100%
Glutamate Metabolism 75%
Intracellular demand 0%
Folate Metabolism 83%
Methylglyoxal Metabolism 71%
Anaplerotic Reactions 75%
Pyruvate Metabolism 66%
Pentose Phosphate Pathway 100%
Histidine Metabolism 40%
Alanine and Aspartate Metabolism 90%
Methionine Metabolism 90%
Cysteine Metabolism 69%
Citric Acid Cycle 100%
Nitrogen Metabolism 15%
Glycine and Serine Metabolism 57%
Murein Recycling 64%
Valine, Leucine, and Isoleucine Metabolism 26%
Threonine and Lysine Metabolism 16%
tRNA Charging 0%
Unassigned 19%
Glycolysis/Gluconeogenesis 72%
Tyrosine, Tryptophan, and Phenylalanine 77%
Purine and Pyrimidine Biosynthesis Transport 56%
Outer Membrane 0%
Membrane Lipid Metabolism 4%
Arginine and Proline Metabolism 78%
Oxidative Phosphorylation Lipopolysaccharide 18%
Biosynthesis / Recycling Inorganic Ion 24%
Transport and Metabolism 8%
Cell Envelope Biosynthesis 52%
Nucleotide Salvage Pathway 38%
Cofactor and Prosthetic Group Biosynthesis 56%
Alternate Carbon Metabolism 34%
Glycerophospholipid Metabolism 89%
Transport, Outer Membrane Porin Extracellular 0%
exchange 0%
Transport, Inner Membrane 33%
0 50 100 150 200 250 300 350
Number of reactions
(b)
Cumulative distribution
0.023
0.036
$10^{-4}$ $10^{-2}$ $10^{0}$ $10^{2}$
$K_m$ (mM)
(c)
Background Glycolysis TCA PPP
Correlation coefficient
Glycolysis
PPP
TCA
(d)
Correlation coefficient
SLB genes
General genes
Reaction type
(e)
Cumulative distribution
$10^{-5}$ $10^{0}$ $10^{5}$
Metabolic flux (mM/gDW/hr)

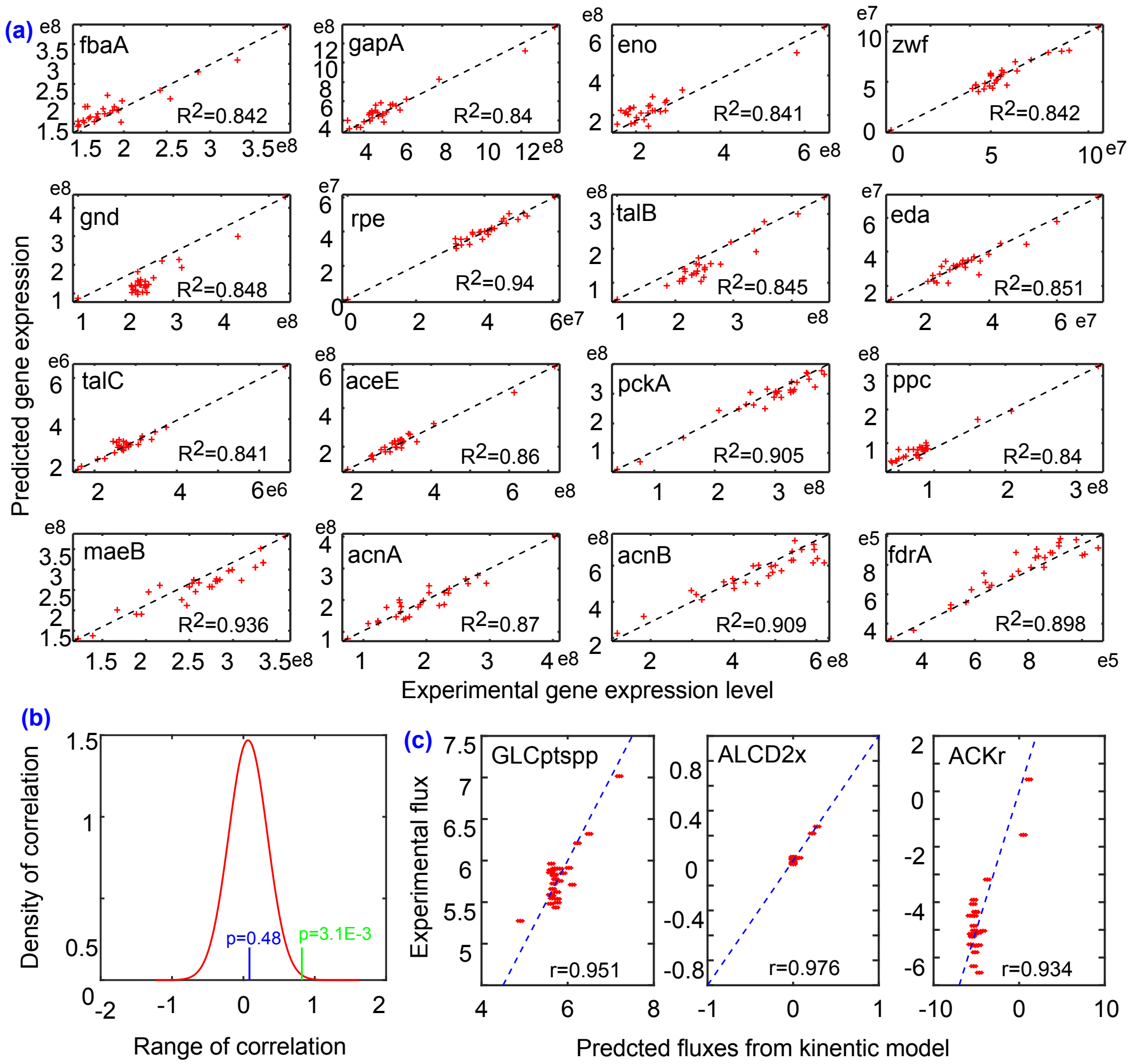

(a)
E. coli Glu/Aer
E. coli Glu/Ana
E. coli Xyl/Aer
E. coli Xyl/Ana
B. subtilis
Yeast
FBA
pFBA
RELATCH
Decrem
r=0.731
r=0.582
r=0.594
r=0.48
r=0.561
r=0.3
r=0.74
r=0.55
r=0.424
r=0.503
r=0.565
r=0.284
r=0.74
r=0.67
r=0.552
r=0.53
r=0.453
r=0.3
r=0.8
r=0.74
r=0.801
r=0.61
r=0.683
r=0.696
13C metabolic flux (mMl/dDW/hr)
Predicted flux (mM/dDW/hr)
(b)
FBA
Decrem
pFBA
RELATCH
MSE
B. bubtilis
E. coli Glu/Aer
E. coli Glu/Ana
E. coli Xyl/Aer
E. coli Xyl/Ana
Yeast
Growth condition
(c)
Activated reactions
B. bubtilis
E. coli Glu/Aer
E. coli Glu/Ana
E. coli Xyl/Aer
E. coli Xyl/Ana
Yeast
Growth condition

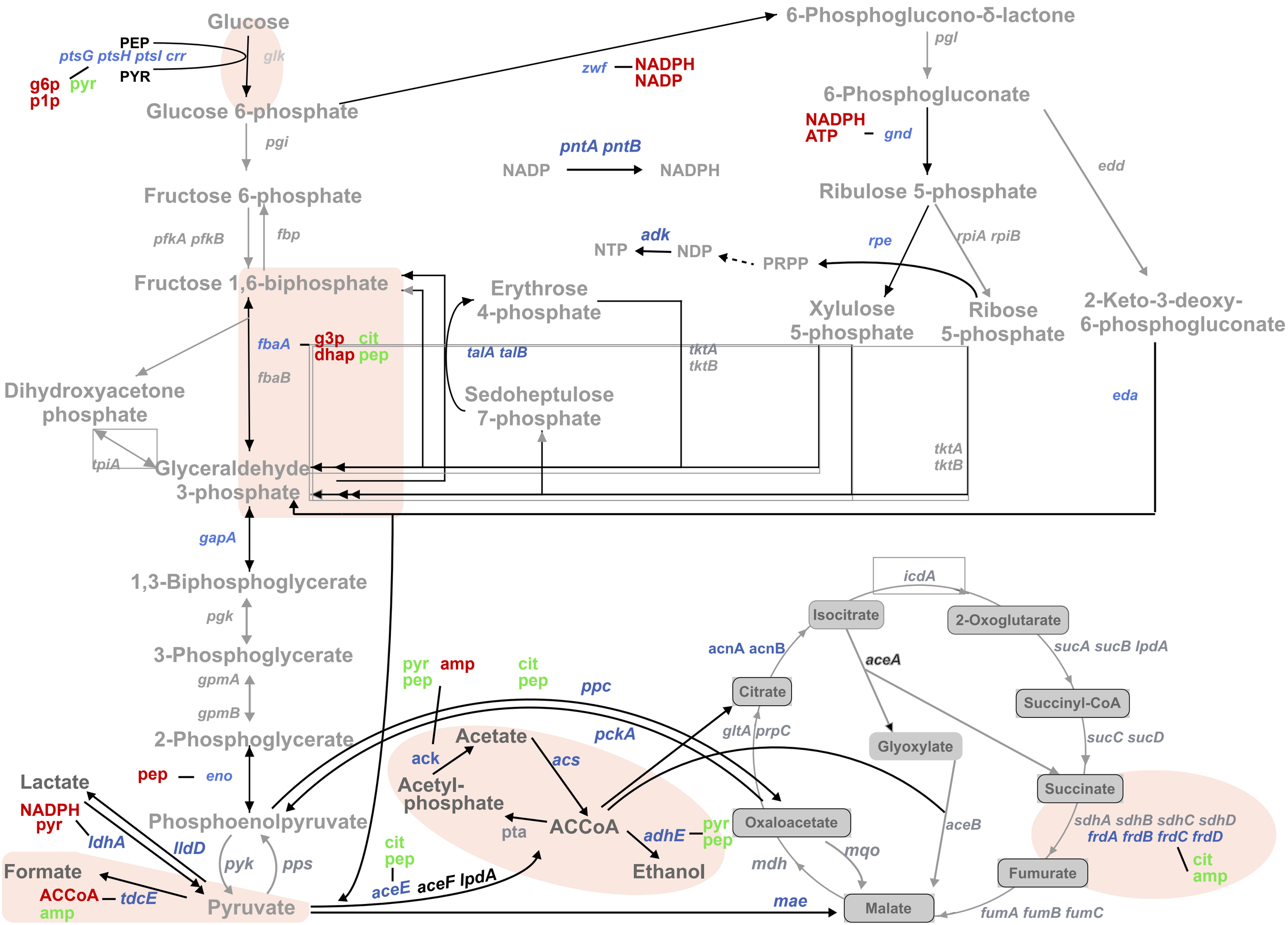

Correlation
0.75
0.50
0.25
IDN750
Pearson
IDN750
Spearman
IMM904
Pearson
IMM904
Spearman
ADH3
ALD5
ALD6
COX5A
CTP1
DAL7
FUM1
GCV2
GLY1
GND2
GPD1
HO
ICL1
IDP1
IDP2
LSC1
MAE1
MDH1
MDH2
MDH3
MLS1
OAC1
PCK1
PDA1
PGM1
PGM2
RPE1
SDH1
SER33
SFC1
SOL1
SOL2
SOL3
SOL4
TAL1
widetype
YGR043C
ZWF1
Strain
Type
Pearson
Spearman
Method
Decrem
FBA
pFBA
RELATCH
REPPS

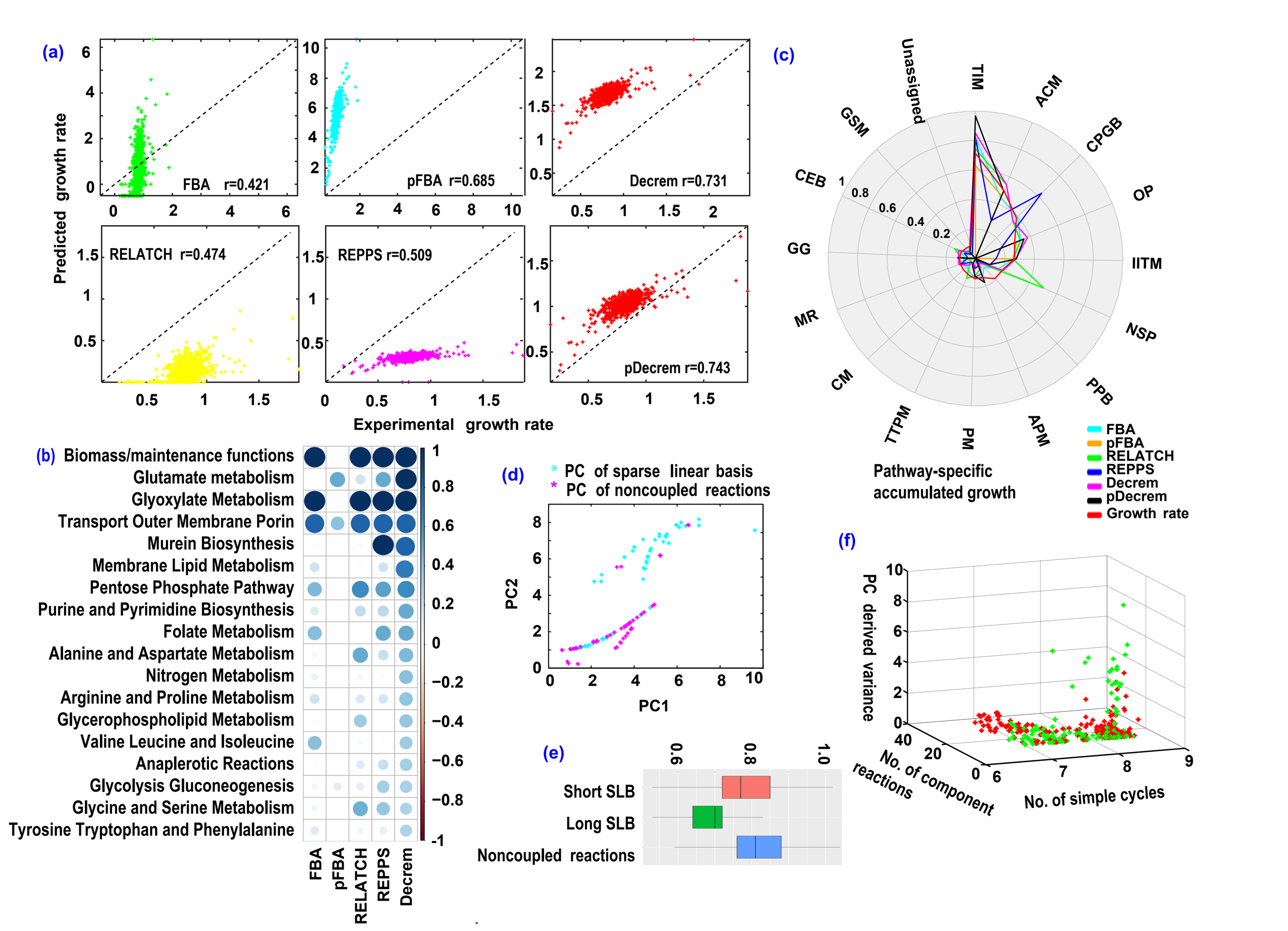

(a)
pDecrem
R²=0.9
Growth rate
-0.001*PC1 +0.017*PC2
REPPS
R²=0.841
-0.007*PC -0.014*PC2
pFBA
R²=0.74
0.001*PC+0.006*PC2
Decrem
R²=0.9
0.001*PC1+0.016*PC2
FBA
R²=0.731
-0.001*PC1 +-0.006*PC2
RELATCH
R²=0.731
0.001*PC1-0.006*PC2
Predicted growth rate by PCA regression


(b)
pDecrem
ZWF1
RPE1
ALD6
FUM1
PDA1
MDH1
Decrem
ALD6
PDA1
RPE1
FUM1
MDH1
ZWF1
FBA
MLS1
LSC1
ZWF1
pFBA
REPPS
RPE1
ZWF1
Widetype
PDA1
ALD6
LSC1
MLS1
RELATCH
MLS1
LSC1
ZWF1
PC1
PC2


(c)
Transport Nuclear
Alanine and Aspartate Metabolism
Tyrosine Tryptophan and Phenylalanine Metabolism
Extracellular exchange
Nucleotide Salvage Pathway
Glycerolipid Metabolism
Sterol Metabolism
Transport Mitochondrial
Galactose metabolism
Alternate Carbon Metabolism
Pyruvate Metabolism
Transport Endoplasmic Reticular
Histidine Metabolism
Fatty Acid Biosynthesis
Valine Leucin and Isoleucine Metabolism
Pentose Phosphate Pathway
Threonine and Lysine Metabolism
Citric Acid Cycle
Folate Metabolism
NAD Biosynthesis
Glutamate metabolism
Anaplerotic reactions
Transport Extracellular
Oxidative Phosphorylation
Phospholipid Biosynthesis
Complex Alcohol Metabolism
GlycolysisGluconeogenesis
Glycine and Serine Metabolism
Transport Vacuolar
Other
Widetype
ADH3
ALD5
ALD6
COX5A
CTP1
DAL7
FUM1
GND2
GCV2
GLY1
GPD1
ICL1
IDP1
IDP2
LSC1
MAE1
MDH1
MDH2
MDH3
MLS1
OAC1
PCK1
PDA1
PGM1
PGM2
RPE1
SDH1
SER33
SFC1
SOL1
SOL2
SOL3
SOL4
TAL1
YGR043C
ZWF1
HO


(a)
fbaA
R2=0.842
gapA
R2=0.84
eno
R2=0.841
zwf
R2=0.842
gnd
R2=0.848
rpe
R2=0.94
talB
R2=0.845
eda
R2=0.851
talC
R2=0.841
aceE
R2=0.86
pckA
R2=0.905
ppc
R2=0.84
maeB
R2=0.936
acnA
R2=0.87
acnB
R2=0.909
fdrA
R2=0.898
Predicted gene expression
Experimental gene expression level
(b)
Density of correlation
p=0.48
p=3.1E-3
Range of correlation
(c)
GLCptspp
r=0.951
ALCD2x
r=0.976
ACKr
r=0.934
Experimental flux
Predcted fluxes from kinentic model


Glucose
PEP
ptsG ptsH ptsI crr
PYR
g6p
pyr
p1p
glk
Glucose 6-phosphate
pgi
Fructose 6-phosphate
pfkA pfkB
fbp
Fructose 1,6-biphosphate
fbaA
g3p
cit
dhap
pep
fbaB
Dihydroxyacetone phosphate
tpiA
Glyceraldehyde 3-phosphate
gapA
1,3-Biphosphoglycerate
pgk
3-Phosphoglycerate
gpmA
gpmB
2-Phosphoglycerate
pep
eno
Phosphoenolpyruvate
Lactate
NADPH
pyr
ldhA
lldD
pyk
pps
Formate
ACCoA
amp
tdcE
Pyruvate
6-Phosphoglucono-δ-lactone
pgl
6-Phosphogluconate
zwf
NADPH
NADP
NADPH
ATP
gnd
pntA pntB
NADP
NADPH
Ribulose 5-phosphate
edd
adk
NTP
NDP
PRPP
rpe
rpiA rpiB
Erythrose 4-phosphate
Xylulose 5-phosphate
Ribose 5-phosphate
2-Keto-3-deoxy-6-phosphogluconate
talA talB
tktA
tktB
Sedoheptulose 7-phosphate
eda
tktA
tktB
pyr
pep
amp
cit
pep
ppc
Acetate
ack
Acetyl-phosphate
acs
pckA
ACCoA
pta
adhE
pyr
pep
Ethanol
cit
pep
aceE
aceF lpdA
mae
icdA
Isocitrate
2-Oxoglutarate
acnA acnB
aceA
sucA sucB lpdA
Citrate
Succinyl-CoA
gltA prpC
Glyoxylate
sucC sucD
Succinate
Oxaloacetate
aceB
sdhA sdhB sdhC sdhD
frdA frdB frdC frdD
mdh
mqo
Fumurate
cit
amp
Malate
fumA fumB fumC


(a)
Predicted growth rate
Experimental growth rate
FBA r=0.421
pFBA r=0.685
Decrem r=0.731
RELATCH r=0.474
REPPS r=0.509
pDecrem r=0.743
(b)
Biomass/maintenance functions
Glutamate metabolism
Glyoxylate Metabolism
Transport Outer Membrane Porin
Murein Biosynthesis
Membrane Lipid Metabolism
Pentose Phosphate Pathway
Purine and Pyrimidine Biosynthesis
Folate Metabolism
Alanine and Aspartate Metabolism
Nitrogen Metabolism
Arginine and Proline Metabolism
Glycerophospholipid Metabolism
Valine Leucine and Isoleucine
Anaplerotic Reactions
Glycolysis Gluconeogenesis
Glycine and Serine Metabolism
Tyrosine Tryptophan and Phenylalanine
FBA
pFBA
RELATCH
REPPS
Decrem
(c)
Unassigned
TIM
ACM
CPGB
OP
IITM
NSP
PPB
APM
PM
TTPM
CM
MR
GG
CEB
GSM
Pathway-specific accumulated growth
FBA
pFBA
RELATCH
REPPS
Decrem
pDecrem
Growth rate
(d)
PC of sparse linear basis
PC of noncoupled reactions
PC2
PC1
(e)
Short SLB
Long SLB
Noncoupled reactions
(f)
PC derived variance
No. of component reactions
No. of simple cycles