*Bioinformatics*, 33(22), 2017, 3603–3609
doi: 10.1093/bioinformatics/btx453
Advance Access Publication Date: 14 July 2017
Original Paper



Systems biology

# Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models

**Siu H. J. Chan[1,*], Jingyi Cai[2], Lin Wang[1], Margaret N. Simons-Senftle[1] and Costas D. Maranas[1,*]**

[1]Department of Chemical Engineering, The Pennsylvania State University, University Park, PA 16801, USA and [2]Beijing Key Lab of Bioprocess, College of Life and Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 6, 2017; revised on June 4, 2017; editorial decision on July 9, 2017; accepted on July 11, 2017

## Abstract

**Motivation:** In a genome-scale metabolic model, the biomass produced is defined to have a molecular weight (MW) of 1 g $mmol^{-1}$. This is critical for correctly predicting growth yields, contrasting multiple models and more importantly modeling microbial communities. However, the standard is rarely verified in the current practice and the chemical formulae of biomass components such as proteins, nucleic acids and lipids are often represented by undefined side groups (e.g. X, R).

**Results:** We introduced a systematic procedure for checking the biomass weight and ensuring complete mass balance of a model. We identified significant departures after examining 64 published models. The biomass weights of 34 models differed by 5–50%, while 8 models have discrepancies >50%. In total 20 models were manually curated. By maximizing the original versus corrected biomass reactions, flux balance analysis revealed >10% differences in growth yields for 12 of the curated models. Biomass MW discrepancies are accentuated in microbial community simulations as they can cause significant and systematic errors in the community composition. Microbes with underestimated biomass MWs are overpredicted in the community whereas microbes with overestimated biomass weights are underpredicted. The observed departures in community composition are disproportionately larger than the discrepancies in the biomass weight estimate. We propose the presented procedure as a standard practice for metabolic reconstructions.

**Availability and implementation:** The MALTAB and Python scripts are available in the Supplementary Material.

**Contact:** costas@psu.edu or joshua.chan@connect.polyu.hk

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Genome-scale metabolic (GSM) networks and the constraint-based reconstruction and analysis (COBRA) framework have proved to be valuable tools in modeling cellular metabolism (Lewis *et al.*, 2012; Price *et al.*, 2004; Schellenberger *et al.*, 2011) and answering important biological questions (McCloskey *et al.*, 2014). In flux balance analysis (FBA), one of the fundamental constraints is the steady-state mass balance equation (Orth *et al.*, 2010), which quantifies the conservation of component balance. As a prerequisite for quantitative predictions, all reactions must be component and charge balanced. This principle of mass balance also applies to the biomass reaction, which expresses biomass as a defined ratio of macromolecules synthesized from metabolites

(Feist and Palsson, 2010). A GSM model describes in quantitative terms the substrate-to-biomass conversion, from mmol of substrates to gram dry cell weight of cells. By definition the biomass produced must have a molecular weight (MW) of 1 g $mmol^{-1}$ in order to quantitatively compare biomass formation with the observed growth yields or specific growth rates (Feist *et al.*, 2010; Fong and Palsson, 2004; Ibarra *et al.*, 2002; Lewis *et al.*, 2010). FBA identifies optimal solutions that strike a balance between biomass formation and ATP production, thus any discrepancy in biomass weight may tilt the balance to have a disproportionate influence on FBA derived flux predictions. This becomes even more critical when comparing multiple GSM models or modeling microbial communities that combine multiple organisms (Harcombe *et al.*, 2014; Heinken and Thiele, 2015a,b; Khandelwal *et al.*, 2013; Klitgord and Segrè, 2010; Magnúsdóttir *et al.*, 2016; Shoaie *et al.*, 2013, 2015; Zhang *et al.*, 2016; Zhuang *et al.*, 2011; Zomorrodi and Maranas, 2012; Zomorrodi *et al.*, 2014; Zhang *et al.*, 2016; Magnúsdóttir *et al.*, 2016; Chan *et al.*, 2017). Predictions from optimization-based analyses such as FBA will introduce a bias towards lightly biomass-weighted microbes in the overall community abundance. Despite its importance, we found that ensuring that the MW of biomass is exactly equal to 1 g $mmol^{-1}$ is not always enforced in practice. We introduce a systematic procedure named 'minimum inconsistency under parsimony' (MIP) for determining the MW of the biomass and simultaneously ensuring the complete elemental balance in every single reaction in a metabolic network. MIP formulates the elemental balance as an optimization problem that solves for the chemical formulae of generic metabolites by minimizing the inconsistencies using the information of known metabolites. The MIP solution provides guidelines for resolving all imbalances in a model. After examining 64 published models, 20 models were manually curated to ensure complete mass-and-charge balance and a standardized biomass weight. The impact of standardizing the biomass weight on FBA-based simulations was assessed by comparing the biomass yield per carbon using the original versus corrected biomass reactions for the 20 curated models.

## 2 Materials and methods

### 2.1 Minimum inconsistency under parsimony

An optimization problem for computing the biomass MW based on the elemental balance of internal reactions is presented. Let $\mathbf{I}$ and $\mathbf{J}$ be the sets of metabolites and reactions, respectively. $\mathbf{S} = [S_{ij}]_{i\in\mathbf{I},j\in\mathbf{J}}$ denotes the stoichiometric matrix and $\mathbf{I}^k$ the set of known metabolites with defined chemical formulae. Consequently, $\mathbf{I}^u = \mathbf{I}–\mathbf{I}^k$ is the set of metabolites with unknown molecular groups. $\mathbf{I}^u$ includes biomass, macromolecules and any other metabolites with generic side groups. We define $\mathbf{E}$ as the set of chemical elements and $\mathbf{J}^{ne}$ as the set of non-exchange reactions. $\mathbf{J}^{ne}$ consists of all biochemical reactions but excludes all exchange reactions that represent the net system inputs/outputs (reactions solely producing or consuming single metabolites, e.g. glc-D[e] <=>). By invoking elemental balances of each reaction in $\mathbf{J}^{ne}$, the unknown molecular groups of metabolites in $\mathbf{I}^u$ can be determined by solving the following system of linear equations:

$$\sum_{i\in I^u} S_{ij}m_{ie} = -\sum_{i\in I^k} S_{ij}m_{ie}, \quad \forall j\in J^{ne}, e\in E \tag{1}$$

$$m_{ie} \geq 0, \quad \forall i \in \mathbf{I}^u, e \in \mathbf{E}$$

where $m_{ie}$ is the stoichiometry of element $e$ in metabolite $i$. However, if any elemental imbalance exists in any reaction, the system of equations becomes infeasible. To accommodate imbalances in Equation (1), a variable $x_{je}$ is added in the optimization problem to quantify the elemental imbalance in reaction $j$ for element $e$. An objective function is thus defined that minimizes the sum of the absolute elemental imbalances. The result of this minimization problem is the minimum possible elemental imbalances that can be used to identify defined chemical formulae for metabolites in $\mathbf{I}^u$ that balance as many reactions as possible. We find that most imbalances are caused by differences in elemental hydrogen balances that can be resolved through the addition of missing protons. The optimization model, therefore, determines the stoichiometry of protons or other small molecules in each reaction that minimizes imbalances. The set of metabolites used for correcting imbalances is denoted by $\mathbf{I}^f$, which in this study contains only protons. For each $i$ in $\mathbf{I}^f$, a variable $A_{ij}$ representing the adjustment of the stoichiometry is defined for each internal reaction $j$ in $\mathbf{J}^{ne}$. The complete procedure termed MIP is a three-step optimization approach. First, the total inconsistency is minimized (Step 1):

$$\min \sum_{j\in\mathbf{J}^{ne}} \sum_{e\in\mathbf{E}} \left| x_{je} \right|$$

subject to

$$\sum_{i\in\mathbf{I}^u} S_{ij}m_{ie} + \sum_{i\in\mathbf{I}^f} A_{ij}m_{ie} + x_{je} = -\sum_{i\in\mathbf{I}^k} S_{ij}m_{ie}, \quad \forall j \in \mathbf{J}^{ne}, e \in \mathbf{E}$$

$$m_{ie} \geq 0, \quad \forall i \in \mathbf{I}^u, e \in \mathbf{E}$$

$$x_{je}, \quad A_{ij} \in R, \quad \forall i \in \mathbf{I}^f, j \in \mathbf{J}^{ne}, e \in \mathbf{E}$$

Then, the total inconsistency is bounded above by its minimum and the total adjustment by $A_{ij}$ is minimized (Step 2):

$$\min \sum_{i\in\mathbf{I}^f} \sum_{j\in\mathbf{J}^{ne}} \left| A_{ij} \right|$$

subject to Constraints in Step 1

$$\sum_{j\in\mathbf{J}^{ne}} |x_{je}| \leq \sum_{j\in\mathbf{J}^{ne}} |x^*_{je}|, \quad \forall e \in \mathbf{E}$$

where $x^*_{je}$ is the inconsistency determined in Step 1. Having $A_{ij} > 0$ implies that adding metabolite $i$ as a product to reaction $j$ relieves the inconsistency whereas $A_{ij} < 0$ denotes that metabolite $i$ should be added as a substrate. Any remaining non-zero $x_{je}$ represents an inconsistency requiring manual resolution. Finally, a set of minimal formulae is obtained by minimizing $m_{ie}$ with the total inconsistency and adjustment bounded above by (Step 3):

$$\min \sum_{i\in\mathbf{I}^u} \sum_{e\in\mathbf{E}} m_{ie}$$

subject to Constraints in Step 2

$$A_{ij} = A^*_{ij}, \quad \forall i \in \mathbf{I}^f, j \in \mathbf{J}^{ne}$$

where $A^*_{ij}$ is the adjustment determined in Step 2. We find that if the formulae for all known metabolites used are correct and the elemental imbalances among reactions are rare then the imbalance $x_{je}$ identified by MIP generally correctly represents the true underlying inconsistency. The veracity of MIP predictions thus relies on the accuracy of the chemical formulae for the known metabolites. Flagging as many metabolites as possible as 'known' and providing their correct elemental composition can help reveal inconsistencies. Ideally, inconsistencies should be fixed so that the objective values in Steps 1 and 2 are both zero in the final model. In this case, unique formulae for metabolites not involved in the transfer of any conserved moieties (discussed in the Section 2.2) can be obtained. Alternatively, in the presence of inconsistency, a range for the minimum and maximum possible $m_{ie}$ of an interested metabolite $i$ (e.g. biomass) can be calculated. See SI Methods for more details.

### 2.2 Conserved moieties involved in metabolites with non-unique formulae

Another issue that frequently arises is that a non-zero null space of $\mathbf{S}^T$ will result in non-unique solutions when solving the formulation MIP. For example, oxidized and reduced ferredoxins always appearing as a pair in reactions are in the null space of $\mathbf{S}^T$. Although MIP finds the set of the simplest chemical formulae, the general case can be modeled using mass conservation vectors in metabolic networks studied previously by (Fleming *et al.*, 2016; Gevorgyan *et al.*, 2008). A mass conservation vector is any vector $\mathbf{n} \geq \mathbf{0}$ satisfying $\mathbf{S}^T\mathbf{n} = \mathbf{0}$. The space of all conservation vectors $\mathbf{N}^+$ forms a convex polytope:

$$\mathbf{N}^+ = \{\, \mathbf{n} \quad | \quad \mathbf{S}^T\mathbf{n} = 0, \quad \mathbf{n} \geq 0 \,\}$$

The set of extreme rays of $\mathbf{N}^+$ is finite and computable using the existing methods for finding elementary modes of metabolic networks (Schuster *et al.*, 1999; Terzer and Stelling, 2008). Denote the set of all extreme rays by:

$$\mathbf{P} = \{\, \mathbf{p}_t = [p_{it}]_{i\in\mathbf{I}} \in \mathbf{N}^+ \quad | \quad t = 1, \ldots, T \,\}$$

where $T$ is the number of extreme rays. Each extreme ray $\mathbf{p}_t$ defines a minimal unit of mass conservation and $\mathbf{P}$ characterizes all possible conserved moieties transferred between metabolites. If $\mathbf{p}_t$ describes a metabolite in $\mathbf{I}^k$, then the corresponding moiety is defined. Otherwise, if all metabolites containing the conserved moiety are in $\mathbf{I}^u$ (i.e. $i \in \mathbf{I}^u$ whenever $p_{it} > 0$) then their chemical formulae are not unique and can be a variety of solutions formed by adding or subtracting a multiple of $\mathbf{p}_t$. For example, MIP usually returns the formulae for the oxidized and reduced ferredoxins as 'none' and '$H_2$', respectively as they only differ by two hydrogen atoms. Any pair of formulae with this difference is thus a valid solution even '$C_{100}$' and '$C_{100}H_2$'. By adding a hypothetical elemental component $R_{p_{it}}$ to the chemical formula calculated by MIP for each metabolite $i$, the expression can be generalized to chemical formulae that contain the unknown moiety corresponding to $\mathbf{p}_t$. In this formalism, the formulae for the oxidized and reduced ferredoxins become 'R' and '$H_2$R', respectively. One unique hypothetical "element" was used for each unknown moiety corresponding to each $\mathbf{p}_t$ in this study. This is similar to the current practice of assigning a 'R' or 'X' group in GSM reconstruction, except unknown groups here are constrained by the metabolic network structure (i.e. the space $\mathbf{N}^+$) and therefore do not affect the mass-and-charge balance. Figure 1 shows an example of the conversion between palmitate, palmitoyl-CoA and palmitoyl-[acyl-carrier protein (ACP)]. In this example, the set of unknown metabolites $\mathbf{I}^u$ includes ACP and palmitoyl-[ACP] because the chemical formula of ACP is usually not defined. Three conserved moieties can be identified from the extreme ray matrix $\mathbf{P}$. $\mathbf{p}_1$ and $\mathbf{p}_2$ respectively represent the transfer of the palmitoyl-group and the coenzyme A, both having defined chemical formulae while $\mathbf{p}_3$ represents the transfer of ACP. Both palmitoyl-[ACP] and ACP involved in $\mathbf{p}_3$ have undefined chemical formulae. The generic element 'R' is therefore added into the minimal formulae for metabolites involved in $\mathbf{p}_3$.

### 2.3 Summary of the procedure

The complete procedure is summarized as follows:

i. Identify the set of non-exchange reactions $\mathbf{J}^{ne}$, the set of known metabolites $\mathbf{I}^k$ and unknown metabolites $\mathbf{I}^u$.
ii. Solve MIP and resolve conflicts in $\mathbf{S}$ and $\mathbf{m}$.
iii. Compute the set of extreme rays $\mathbf{P}$ for the space of conservation vectors $\mathbf{N}^+$. For each extreme ray $\mathbf{p}_t$, if $p_{it} = 0$ for all $i$ in $\mathbf{I}^k$, add a hypothetical elemental component (e.g. akin to –R or –X) $R_{p_{it}}$ for each $i$ with $p_{it} > 0$ to the chemical formula.

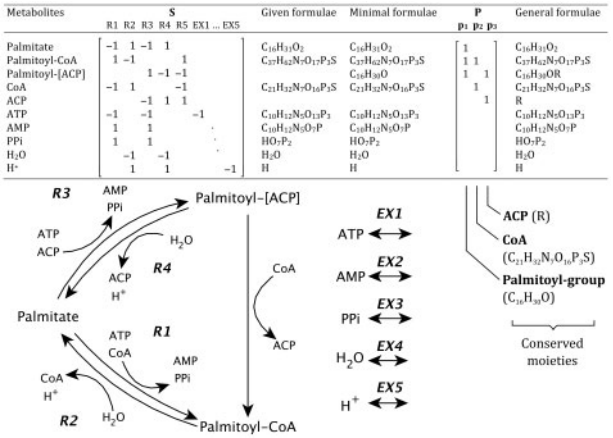


**Fig. 1.** Identification of defined or generic conserved moieties. S is the stoichiometric matrix. P is the extreme ray matrix identified from S. The minimal formulae are the results from solving MIP. Adding the generic conserved moieties into the minimal formulae results in the general formulae

The scripts and examples for the procedure are available in Supplementary Material.

### 2.4 Standardization of biomass reactions

After running the procedure in Section 2.3, the MWs of the biomass and macromolecules can be calculated. A biomass reaction is not standardized if (i) the MW of the formula is not equal to 1 g $mmol^{-1}$, or (ii) the formula contains any of the conserved moieties flagged by the extreme ray calculation (e.g. ACP, ferredoxin). The stoichiometric coefficients in the biomass and macromolecule reactions need to be rescaled according to the original biomass composition so that the biomass has a MW of 1 g $mmol^{-1}$. For the 20 models curated in this study, the chemical formulae for most known metabolites ($\mathbf{I}^k$) used were adopted from MetaCyc, in which all metabolites are protonated to the physiological pH 7.3 (Caspi *et al.*, 2016). The BiGG (King *et al.*, 2016) or the

SEED databases (Henry *et al.*, 2010) were used for those not found in MetaCyc. See Supplementary Table S2 for the curation of biomass reactions and Supplementary Table S3 for the conserved moieties identified for each of the curated models.

### 2.5 Flux balance analysis

FBA simulations were performed under the condition of carbon and ATP limitation. Community simulations were performed by treating the microbial community as one multi-compartment metabolic model with the unweighted sum of the biomass reaction fluxes of individual organisms as the objective function.

## 3 Results

### 3.1 Determination of *in silico* biomass weight

48 models from the GSM model collection in the openCOBRA community (https://github.com/opencobra/m_model_collection), together with 16 models from the literature, were taken to form a group of 64 models (Supplementary Table S1). We applied MIP to determine the MW of biomass within each GSM model (see Section 2). Only external metabolites with defined chemical formulae were used to define the set of known metabolites $\mathbf{I}^k$. Only a few models

strictly comply with the standard of 1 g mmol$^{-1}$ for the MW of the biomass (Fig. 2). They include models of *Escherichia coli* (Archer *et al.*, 2011; Feist *et al.*, 2007; Orth *et al.*, 2011), *Salmonella typhimurium* (AbuOun *et al.*, 2009), *Saccharomyces cerevisiae* (Förster *et al.*, 2003) and *Mycobacterium tuberculosis* (Bordbar *et al.*, 2010) (≤1% deviation from 1 g mmol$^{-1}$). Approximately 42 of 64 models exhibited >5% discrepancies and 20 models had >20% discrepancies. We selected 20 models, which span model and industrially important organisms for correction and further analysis (see Table 1, models are available in Supplementary Material).

## 3.2 Curation of biomass reactions

For the 20 corrected models that are now mass-and-charge balanced, the biomass reaction of each model was further modified to produce biomass with a MW that is normalized to 1 g mmol$^{-1}$. This entails renormalizing the coefficient of each biomass component according to its mass fraction provided in the original biomass composition data. If the data was not available, the coefficient for each component was divided by the biomass MW calculated by MIP (in g mmol$^{-1}$), ensuring that the updated biomass reaction produces biomass with a MW equal to 1 g mmol$^{-1}$.

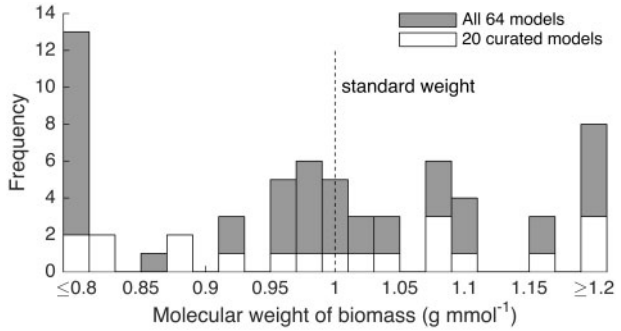


**Fig. 2.** MW of the *in silico* biomass of each model examined in this study calculated using MIP

We identified three primary sources leading to inaccuracies in the biomass MW (Fig. 3). First, a subset of models uses biomass reactions that were generated by automated platforms or adapted from other models (e.g. the models for *Bacteroides thetaiotaomicron*, *Faecalibacterium prausnitzii*, *Lactococcus lactis*, *Streptococcus thermophilus* and Yeast 7). In the absence of experimental data, the mass fractions for the biomass reaction were simply obtained by uniformly normalizing over all biomass components. Second, for some models we found inconsistent stoichiometric coefficients in the biomass reaction because the MWs of macromolecules used for calculating the coefficients were not the same as the actual MWs implied by their elemental balance. For example, in the *Yarrowia lipolytica* model, our MIP procedure calculated MWs for phospholipids that were ∼100× larger than the MWs used in the original model construction (Pan and Hua, 2012) yielding a biomass MW of 30 g mmol$^{-1}$. The reason for this was that the model lipid building blocks such as 1-acyl-sn-glycerol 3-phosphate were synthesized as polymers with 100-mers (e.g. in the reaction for glycerol 3-phosphate acyltransferase), instead of monomers as in other models. Inconsistent stoichiometric coefficients were also found in the models for *Corynebacterium glutamicum*, *Clostridium acetobutylicum* and *Eubacterium rectale*. A probable reason for the errors is the lack of the application of a procedure to ensure complete mass balance and verify the biomass MW. Some GSM models included metabolites with undefined side-groups (e.g. acyl groups in lipids) that

**Table 1.** Genome-scale reconstructions curated in this study

| Organism | Model/References | Biomass MW (g mmol$^{-1}$) | No. of conserved moieties (generic) | Error 1 | Error 2 | Error 3 |
|---|---|---|---|---|---|---|
| *B. thetaiotaomicron* VPI-5482 | *i*AH991 (Heinken *et al.*, 2013) updated in (Heinken and Thiele, 2015b) | 1.44 | 33 (7) | X | | |
| *Bifidobacterium adolescentis* L2-32 | *i*Bif452 (El-Semman *et al.*, 2014) | 1.04 | 96 (25) | X | | |
| *E. faecalis* V583 | V583 (Veith *et al.*, 2015) | 1.08 | 46 (21) | X | | |
| *E. coli* K-12 MG1655 | *i*JO1366 (Orth *et al.*, 2011) | 1.00 | 39 (35) | | | |
| *E. rectale* ATCC 33656 | *i*Ere400 (Shoaie *et al.*, 2013) | 0.62 | 32 (15) | | X | X |
| *F. prausnitzii* A2-165 | *i*Fpraus_v1.0 (Heinken *et al.*, 2014) | 1.44 | 40 (4) | X | | |
| *Klebsiella pneumoniae* MGH 78578 | *i*YL1228 (Liao *et al.*, 2011), updated in (Heinken and Thiele, 2015b) | 0.97 | 44 (30) | X | | |
| *Lactobacillus casei* ATCC 334 | *i*Lca12A_640 (Vinay-Lara *et al.*, 2014) | 1.02 | 95 (15) | X | | |
| *P. gingivalis* W83 | *i*VM679 (Mazumdar *et al.*, 2009) | 1.08 | 73 (27) | X | | |
| *S. thermophilus* LMG 18311 | *i*MP429 (Pastink *et al.*, 2009), updated in (Heinken and Thiele, 2015b) | 0.81 | 44 (24) | X | | |
| *C. acetobutylicum* ATCC 824 | *i*Cac802 (Dash *et al.*, 2014) | 1.16 | 94 (22) | | X | X |
| *L. lactis* MG1363 | (Flahaut *et al.*, 2013) | 0.83 | 52 (24) | X | | |
| *B. subtilis* 168 | *i*Bsu1103V2 (Tanaka *et al.*, 2013) | 1.10 | 67 (10) | | X | X |
| *C. glutamicum* ATCC 13032 | (Kjeldsen and Nielsen, 2009) | 0.79 | 5 (4) | | X | X |
| Consensus model of yeast | Yeast7 (Aung *et al.*, 2013) | 0.88 | 90 (60) | X | X | |
| *Y. lipolytica* | *i*YL619 (Pan and Hua, 2012) | 30.76 | 17 (12) | | X | |
| *P. pastoris* | *i*MT1026 (Tomàs-Gamisans *et al.*, 2016) | 0.92 | 110 (71) | | | X |
| *Synechococcus elongatus* PCC 7942 | (Mueller *et al.*, 2017) | 0.87 | 181 (58) | | X | |
| *Cyanothece* sp. ATCC 51142 | *i*Cyt773 (Saha *et al.*, 2012) | 1.08 | 47 (29) | | X | |
| *Methanosarcina acetivorans* | *i*MAC868 (Nazem-Bokaee *et al.*, 2016) | 0.99 | 40 (24) | | X | X |

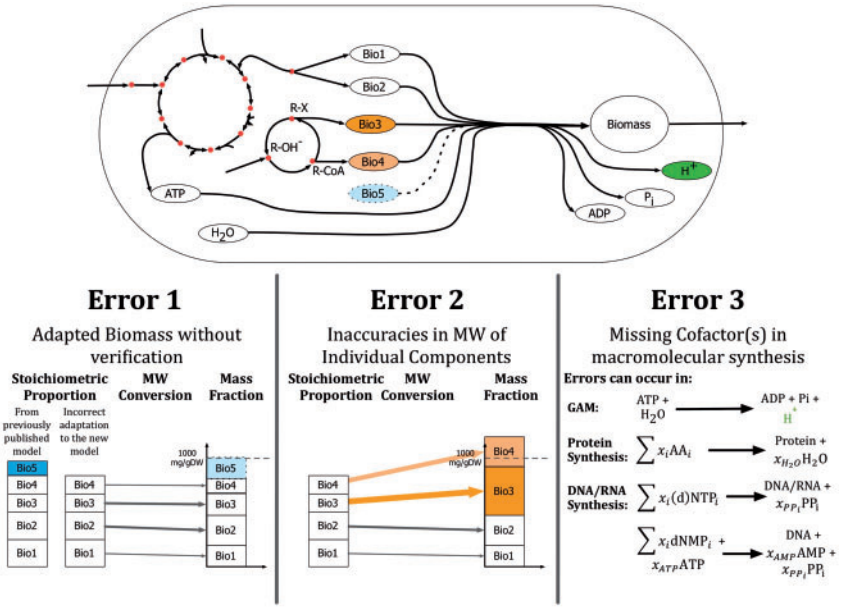


**Fig. 3.** Sources of errors in the biomass reactions. Three sources of errors in the biomass reactions: (i) biomass reactions generated by automated platforms or adapted from other models with biomass components deleted ('Bio5') or newly added; (ii) inaccurate stoichiometric coefficients in the biomass reaction ('Bio3', 'Bio4') partially due to the existence of undefined side-groups (e.g., 'R' and 'X'); and (iii) missing cofactors in macromolecular synthetic reactions, such as proton in GAM ('$H^+$'), water in protein synthesis and pyrophosphate in DNA and RNA syntheses

complicate the estimation of the MWs of macromolecules. Among the models examined, *Bacillus subtilis*, *C. acetobutylicum*, *Enterococcus faecalis*, *E. rectale*, *Porphyromonas gingivalis* and Yeast 7 included metabolites with chemical formulae containing generic 'R' or 'X' groups. Finally, small molecules in macromolecular synthesis reactions were sometimes missing, (e.g. missing proton in the growth-associated maintenance (GAM), $H_2O$ in protein synthesis, and pyrophosphate in DNA or RNA synthesis). This was observed in the models for *B. subtilis*, *C. acetobutylicum*, *C. glutamicum*, *E. rectale* and *Pichia pastoris*. Figure 3 pictorially illustrates the sources of error in the calculation of biomass composition. A detailed discussion regarding mass-and-charge imbalance and how biomass reactions were corrected is included in Supplementary Material S1 and Supplementary Table S2.

### 3.3 Impact on FBA and community simulation

To access the impact of the non-standardized biomass reactions, FBA was performed for each one of the 18 curated models to optimize the original and the corrected biomass reactions (matching 1 g $mmol^{-1}$) for 1000 sets of randomly sampled uptake rates on carbon sources. Half of the models have $\geq$10% difference (on average) in biomass yield over the range of the uptake sets (Fig. 4A). We observed that significant errors for even a single component in the biomass reaction could have a profound effect on FBA calculations. For example, in the *B. subtilis* model, proton production due to GAM was missing resulting in a 10% increase in the MW of biomass. After correction, the predicted biomass yield increased by 17% because the extra proton can drive additional ATP production to support growth.

We also constructed a proxy gut microbial community of three bacteria using the curated models for *B. thetaiotaomicron*, *C. acetobutylicum* and *E. rectale*. The growth of the community was simulated using FBA by maximizing the uniform sum of individual biomass formations (El-Semman *et al.*, 2014; Heinken *et al.*, 2013; Heinken and Thiele, 2015a,b; Stolyar *et al.*, 2007). Diets composed of various ratios of total carbohydrate and dietary fiber to amino acids (CF/AAs) were tested in the simulations (Supplementary Table S4). Differences were observed upon optimizing the sum of the original versus corrected biomass reactions (Fig. 4B). Using the original biomass reactions, FBA predicted the dominance of *E. rectale* at high CF/AA ratio and no growth of *B. thetaiotaomicron* under all conditions. Using the corrected biomass reactions, the expected dominance of *B. thetaiotaomicron* at high CF/AA ratios was predicted. The co-growth of Bacteroidetes and Firmicutes experimentally observed in gut microbiota (Huttenhower *et al.*, 2012), was also predicted at CF/AA ratios ranging from 0.2 to 1 and from 4 to 6. This confirms that non-standardized biomass reactions in microbial community simulations can cause a significant and systematic biased preference toward the more lightly weighted microbes (*E. rectale* in this case). The quantitative impact of 10–40% discrepancies in biomass MWs for the community members causes a disproportionate and often dramatic change in community composition.

## 4 Discussion

### 4.1 A method to ensure complete mass balance in GSM networks

Analyzing the biomass MWs for 64 GSM models suggests that standardizing the biomass reactions in models is not yet common practice. This can lead to significant errors, especially for microbial community models. A possible reason for this omission is the difficulty of computing the *in silico* biomass MWs in the presence of metabolites with undefined side-groups (e.g. alkyl group) and inconsistencies in mass balances. These observations support the necessity of a systematic procedure to ensure complete elemental balance and standardized biomass reactions, in line with the recent call for clearer standards (Ebrahim *et al.*, 2015). The proposed procedure determines the biomass MW and reveals imbalances among reactions based on a set of defined metabolites which must at least include external metabolites sufficient for biomass production. It can accommodate both generic

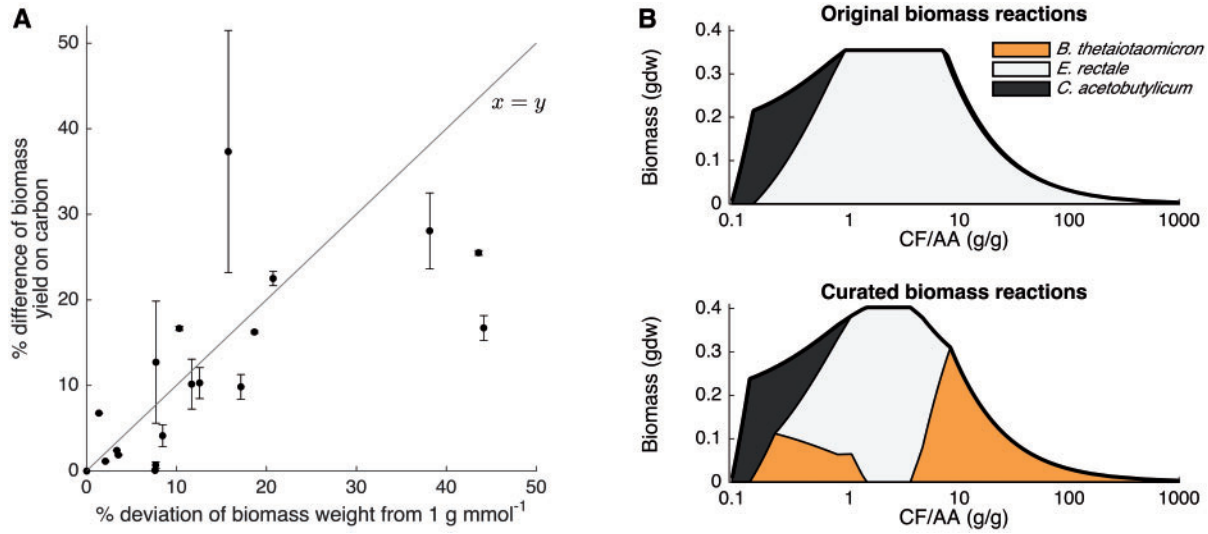


**Fig. 4.** Discrepancies in FBA and community simulations caused by non-standardized biomass reactions. **(A)** Correlation between the difference in FBA prediction and the deviation of the biomass weight from 1 g mmol$^{-1}$ for the 20 curated models. One outlying data point (2900% deviation in weight, 98% difference in biomass yield) is not plotted. **(B)** Predicted growth of the three-membered community by maximizing the sum of individual biomass. Predictions made using the original biomass reactions and the curated biomass reactions are shown in the upper and lower panel respectively. The total community biomass (black curves) and the biomass of individual microbes (filled areas) are shown

metabolites containing conserved moieties (e.g. ferredoxin) and metabolites with defined formulae (e.g., AAs) using the network structure encoded in the **S** matrix. The procedure provides a systematic way of testing for biomass MW consistency. The veracity of the results from the procedure relies on the correct chemical formulae for known metabolites. This implies the need to determine/verify the correct protonation state of each known metabolite at the compartment-specific physiological pH based on its chemical structure (Flamholz *et al.*, 2012; Noor *et al.*, 2013).

## Funding

This work has been supported by the U.S. Department of Energy (DOE, http://www.energy.gov/) grant no. DE-SC0008091.

*Conflict of Interest*: none declared.

## References

AbuOun,M. *et al.* (2009) Genome scale reconstruction of a salmonella metabolic model: comparison of similarity and differences with a commensal Escherichia coli strain. *J. Biol. Chem.*, **284**, 29480–29488.

Archer,C.T. *et al.* (2011) The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics*, **12**, 9.

Aung,H.W. *et al.* (2013) Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol.*, **9**, 215–228.

Bordbar,A. *et al.* (2010) Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. *Mol. Syst. Biol.*, **6**, 422.

Caspi,R. *et al.* (2016) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **44**, D471–D480.

Chan,S.H.J. *et al.* (2017) SteadyCom: Predicting microbial abundances while ensuring community stability. *PLOS Comput. Biol.*, **13**, e1005539.

Dash,S. *et al.* (2014) Capturing the response of Clostridium acetobutylicumto chemical stressors using a regulated genome-scale metabolic model. *Biotechnol. Biofuels*, **7**, 144.

Ebrahim,A. *et al.* (2015) Do genome-scale models need exact solvers or clearer standards?. *Mol. Syst. Biol.*, **11**, 831.

El-Semman,I.E. *et al.* (2014) Genome-scale metabolic reconstructions of Bifidobacterium adolescentis L2-32 and Faecalibacterium prausnitzii A2-165 and their interaction. *BMC Syst. Biol.*, **8**, 41.

Feist,A.M. *et al.* (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.*, **3**, 121.

Feist,A.M. *et al.* (2010) Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. *Metab. Eng.*, **12**, 173–186.

Feist,A.M. and Palsson,B.O. (2010) The biomass objective function. *Curr. Opin. Microbiol.*, **13**, 344–349.

Flahaut,N.A.L. *et al.* (2013) Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. *Appl. Microbiol. Biotechnol.*, **97**, 8729–8739.

Flamholz,A. *et al.* (2012) EQuilibrator - The biochemical thermodynamics calculator. *Nucleic Acids Res.*, **40**,

Fleming,R.M.T. *et al.* (2016) Conditions for duality between fluxes and concentrations in biochemical networks. *J. Theor. Biol.*, **409**, 1–10.

Fong,S.S. and Palsson,B.Ø. (2004) Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat. Genet.*, **36**, 1056–1058.

Förster,J. *et al.* (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.*, **13**, 244–253.

Gevorgyan,A. *et al.* (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*, **24**, 2245–2251.

Harcombe,W.R. *et al.* (2014) Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.*, **7**, 1104–1115.

Heinken,A. *et al.* (2014) Functional metabolic map of Faecalibacterium prausnitzii, a beneficial human gut microbe. *J. Bacteriol*, **196**, 3289–3302.

Heinken,A. *et al.* (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*, **4**, 28–40.

Heinken,A. and Thiele,I. (2015a) Anoxic conditions promote species-specific mutualism between gut microbes in silico. *Appl. Environ. Microbiol.*, **81**, 4049–4061.

Heinken,A. and Thiele,I. (2015b) Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes*, **6**, 120–130.

Henry,C.S. *et al.* (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.*, **28**, 977–982.

Huttenhower,C. *et al*. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Ibarra,R.U. *et al*. (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, **420**, 186–189.

Khandelwal,R. a. *et al*. (2013) Community flux balance analysis for microbial consortia at balanced growth. *PLoS One*, **8**, e64567.

King,Z.A. *et al*. (2016) BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*., **44**, D515–D522.

Kjeldsen,K.R. and Nielsen,J. (2009) In silico genome-scale reconstruction and validation of the Corynebacterium glutamicum metabolic network. *Biotechnol. Bioeng*., **102**, 583–597.

Klitgord,N. and Segrè,D. (2010) Environments that Induce Synthetic Microbial Ecosystems. *PLoS Comput. Biol*., **6**, e1001002.

Lewis,N.E. *et al*. (2012) Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol*., **10**, 291–305.

Lewis,N.E. *et al*. (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol*., **6**, 390.

Liao,Y.-C. *et al*. (2011) An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. *J. Bacteriol*., **193**, 1710–1717.

Magnúsdóttir,S. *et al*. (2016) Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat. Biotechnol*., **35**, 81–89.

Mazumdar,V. *et al*. (2009) Metabolic Network Model of a Human Oral Pathogen. *J. Bacteriol*., **191**, 74–90.

McCloskey,D. *et al*. (2014) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol*., **9**, 661–661.

Mueller,T.J. *et al*. (2017) Identifying the Metabolic Differences of a Fast-Growth Phenotype in Synechococcus UTEX 2973. *Sci. Rep*., **7**, 41569.

Nazem-Bokaee,H. *et al*. (2016) Assessing methanotrophy and carbon fixation for biofuel production by Methanosarcina acetivorans. *Microb. Cell Fact*, **15**, 10.

Noor,E. *et al*. (2013) Consistent Estimation of Gibbs Energy Using Component Contributions. *PLoS Comput. Biol*., **9**, e1003098.

Orth,J.D. *et al*. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol. Syst. Biol*., **7**, 535.

Orth,J.D. *et al*. (2010) What is flux balance analysis?. *Nat. Biotechnol*., **28**, 245–248.

Pan,P. and Hua,Q. (2012) Reconstruction and In Silico Analysis of Metabolic Network for an Oleaginous Yeast, Yarrowia lipolytica. *PLoS One*, **7**, e51535.

Pastink,M.I. *et al*. (2009) Genome-scale model of Streptococcus thermophilus LMG18311 for metabolic comparison of lactic acid bacteria. *Appl. Environ. Microbiol*., **75**, 3627–3633.

Price,N.D. *et al*. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol*., **2**, 886–897.

Saha,R. *et al*. (2012) Reconstruction and comparison of the metabolic potential of cyanobacteria Cyanothece sp. ATCC 51142 and Synechocystis sp. PCC 6803. *PLoS One*, **7**, e48285.

Schellenberger,J. *et al*. (2011) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc*., **6**, 1290–1307.

Schuster,S. *et al*. (1999) Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*., **17**, 53–60.

Shoaie,S. *et al*. (2015) Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell Metab*., **22**, 320–331.

Shoaie,S. *et al*. (2013) Understanding the interactions between bacteria in the human gut through metabolic modeling. *Sci. Rep*., **3**, 2532.

Stolyar,S. *et al*. (2007) Metabolic modeling of a mutualistic microbial community. *Mol. Syst. Biol*., **3**, 92.

Tanaka,K. *et al*. (2013) Building the repertoire of dispensable chromosome regions in Bacillus subtilis entails major refinement of cognate large-scale metabolic model. *Nucleic Acids Res*., **41**, 687–699.

Terzer,M. and Stelling,J. (2008) Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics*, **24**, 2229–2235.

Tomàs-Gamisans,M. *et al*. (2016) Integration and validation of the genome-scale metabolic models of Pichia pastoris: A comprehensive update of protein glycosylation pathways, lipid and energy metabolism. *PLoS One*, **11**, 1–24.

Veith,N. *et al*. (2015) Using a genome-scale metabolic model of Enterococcus faecalis V583 to assess amino acid uptake and its impact on central metabolism. *Appl. Environ. Microbiol*., **81**, 1622–1633.

Vinay-Lara,E. *et al*. (2014) Genome -scale reconstruction of metabolic networks of Lactobacillus casei ATCC 334 and 12A. *PLoS One*, **9**, e110785.

Zhang,L. *et al*. (2016) Farnesoid X Receptor Signaling Shapes the Gut Microbiota and Controls Hepatic Lipid Metabolism. *mSystems*., **1**, e00070–16.

Zhuang,K. *et al*. (2011) Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *ISME J*., **5**, 305–316.

Zomorrodi,A.R. *et al*. (2014) d-OptCom: Dynamic Multi-level and Multi-objective Metabolic Modeling of Microbial Communities. *ACS Synth. Biol*., **3**, 247–257.

Zomorrodi,A.R. and Maranas,C.D. (2012) OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput. Biol*., **8**, e1002363.