A Duality-Based Method for Identifying Elemental Balance Violations in Metabolic Network Models

Hooman Zabeti
School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
hzabeti@sfu.ca

Tamon Stephen
Department of Mathematics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada.
tamon@sfu.ca

Bonnie Berger
Department of Mathematics and CSAIL, Massachusetts Institute of Technology, Cambridge, MA, 02139, United States of America.
bab@csail.mit.edu

Leonid Chindelevitch
School of Computing Science, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
leonid@sfu.ca

Abstract

Elemental balance, the property of having the same number of each type of atom on both sides of the equation, is a fundamental feature of chemical reactions. In metabolic network models, this property is typically verified on a reaction-by-reaction basis. In this paper we show how violations of elemental balance can be efficiently detected in an entire network, without the need for specifying the chemical formula of each of the metabolites, which enhances a modeler's ability to automatically verify that their model satisfies elemental balance.

Our method makes use of duality theory, linear programming, and mixed integer linear programming, and runs efficiently on genome-scale metabolic networks (GSMNs). We detect elemental balance violations in 40 out of 84 metabolic network models in the BiGG database. We also identify a short list of reactions that are candidates for being elementally imbalanced. Out of these candidates, nearly half turn out to be truly imbalanced reactions, and the rest can be seen as witnesses of elemental balance violations elsewhere in the network. The majority of these violations involve a proton imbalance, a known challenge of metabolic network reconstruction.

Our approach is efficient, easy to use and powerful. It can be helpful to metabolic network modelers during model verification. Our methods are fully integrated into the MONGOOSE software suite and are available at `https://github.com/WGS-TB/MongooseGUI3`.

2012 ACM Subject Classification Applied computing → Biological networks

Keywords and phrases Metabolic network analysis, elemental imbalance, linear programming, model verification

Digital Object Identifier 10.4230/LIPIcs.WABI.2018.1

Funding TS would like to acknowledge financial support from an NSERC Discovery Grant. LC would like to acknowledge financial support from a Sloan Foundation Fellowship and an NSERC Discovery Grant.

Acknowledgements The authors would like to thank Cedric Chauve, Michael Schnall-Levin, Kamyar Khodamoradi, Nafiseh Sedaghat and Reza Miraskarshahi for helpful discussions.

1 Introduction

Genome-scale metabolic network reconstructions (GSMNs) represent the collection of all metabolic reactions available to a specific organism, together with constraints on their direction. These GSMNs, coupled with the constraint-based analysis framework, have been successfully used for predicting the growth rates of various organisms under different environmental conditions [20], the minimal media necessary for growth [10], the essentiality and synthetic lethality of specific genes [2], as well as identifying promising intervention strategies to inhibit their growth [12].

A fundamental property of chemical reactions is elemental balance, the presence of the same number of each type of atom in the reactants (left-hand side of the reaction) and the products (right-hand side of the reaction). For instance, the ethanol combustion reaction,

$$C_2H_5OH + 3O_2 \rightarrow 3H_2O + 2CO_2,$$

is elementally balanced. Each reaction in a GSMN is expected to possess this property [18, 15]. Unfortunately, it turns out that elemental balance is frequently violated in published models [14]. Since the chemical formulas of the metabolites in a metabolic network are sometimes left unspecified, it is a challenge for modelers to use the facilities provided by platforms such as COBRA [18] to check elemental balance in their model directly.

In this paper we present a method for quickly and effectively identifying small groups of reactions in a metabolic network model such that at least one of them violates elemental balance, without the need to specify the elemental formulas of any metabolites. This method is based on the earlier observation that a set of elementally imbalanced reactions may be able to produce "something out of nothing" [14]. Our key theoretical result, first proved in Chindelevitch [7], shows that the ability to produce something out of nothing is mathematically equivalent to a violation of elemental balance; more specifically, if something cannot be produced out of nothing, then there is a set of chemical formulas that make all the reactions elementally balanced. Our method is fully integrated into the MONGOOSE software suite for metabolic network analysis in exact rational arithmetic [8, 13], and is available at `https://github.com/WGS-TB/MongooseGUI3`.

When we apply our method to the collection of existing GSMNs in the BiGG database [11], we find elemental balance violations in 40 out of 84 of them. When we compare these elemental balance violations to the formulas provided in some of the models, we observe that our tool directly identifies a subset of the elementally imbalanced reactions in 28 of the 40 GSMNs, and identifies small sets of reactions containing an imbalanced reaction, which we call "free lunches", in an additional 10 cases. We cannot ascertain which one of these options occurs in the remaining 2 cases since they do not have a formula specified for each metabolite. These results mean that our tool can help modelers efficiently identify elemental imbalances during the process of GSMN reconstruction, even in the absence of detailed chemical formulas for the metabolites.

Although previous authors have identified the presence of elemental imbalances in GSMNs [14], it has not been previously demonstrated, to the best of our knowledge, that this could be done without knowing the exact chemical formulas for the metabolites. Our results also show that our method fails to detect the elementally imbalanced reactions that are present in 32 out of 84 GSMNs in the BiGG database, so it does not guarantee that the network is elementally balanced. Nevertheless, with a detection rate exceeding 55%, this method can be an additional tool for model verification, and we expect it to be helpful to the community thanks to its versatility, transparency, and ease of use.

2 Methods

This section is organized as follows. We start by explaining the connection between elemental balance in a metabolic network and its ability to produce "something out of nothing". We then introduce our three-step method which consists of the following steps:

1. Verify whether the condition forbidding the production of something out of nothing is violated, and if so, identify any metabolites that can be produced out of nothing.
2. Identify a small set of reactions responsible for such a production being possible.
3. Find a lower bound on the number of imbalanced reactions in the network and identify reactions that are likely to be imbalanced.

All of these steps are carried out entirely from the stoichiometric matrix of the network, without any knowledge of the underlying chemical formulas of the metabolites. Steps 1 and 2 are carried out using linear programming, while step 3 uses integer linear programming.

Throughout this paper we use the standard convention that z can also denote the vector with all entries equal to z for any $z \in \mathbb{R}$, and that all vector inequalities (such as $x \geq y$) are interpreted component-wise except for $x \neq y$, which means that x and y are unequal vectors.

Elemental Balance and the No Free Lunch Condition

A GSMN is a collection of metabolic reactions available to an organism. If the GSMN has m metabolites and the n reactions, it is commonly represented by its *stoichiometric matrix* $S \in \mathbb{R}^{m \times n}$, which contains a row for each metabolite and a column for each reaction. The entry S_{ij} is the *stoichiometric coefficient* of metabolite i in reaction j, which is positive if reaction j produces metabolite i, negative if it consumes it, and zero otherwise.

The stoichiometric matrix S is *elementally balanced* with respect to some set of chemical elements (called a "closed system with atomic representation" in [16]) if and only if there exists a vector y with strictly positive components such that

$$y^T S = 0. \tag{1}$$

Indeed, one can think of y as containing the number of elements (atoms) in each metabolite, since every metabolite has a strictly positive (and integer) number of atoms, and the total number of atoms must be the same for the reactants and the products in any elementally balanced reaction.

Therefore, the non-existence of such a vector y implies the presence of an imbalanced reaction in the model. Theorem 2 shows that elemental balance is equivalent to the No *Free Lunch (FL)* condition, which postulates that something cannot be produced out of nothing by the net (overall) reaction of any linear combination of the reactions in the metabolic network. This condition can be written as

$$\mathcal{A} := \{v \mid Sv \geq 0 \text{ and } Sv \neq 0\} = \emptyset$$

The following theorem, known as Stiemke's Alternative [5], is one of several statements closely related to Farkas' Lemma.

▶ **Theorem 1** (Stiemke's Alternative). *Let $A \in \mathbb{R}^{m \times n}$. Then exactly one of the following is true.*

1. $\exists\ x \in \mathbb{R}^n$ *such that* $Ax \geq 0$ *and* $Ax \neq 0$
2. $\exists\ y \in \mathbb{R}^m$ *with* $y > 0$ *such that* $y^T A = 0$

▶ **Theorem 2.** *[7] Let S be a stoichiometric matrix of a metabolic network. Then S is elementally balanced with respect to some set of elements if and only if no linear combination of the reactions in S can result in the production of one or more metabolites out of no reagents.*

Proof. Assume that S is elementally balanced. Then there exists a vector y with strictly positive components such that

$$y^T S = 0.$$

In this case, according to Theorem 1, there is no vector $v \in \mathbb{R}^n$ such that

$$Sv \geq 0 \text{ and } Sv \neq 0.$$

Therefore, the No FL condition holds.
Similarly, if the No FL condition holds for the matrix S, by Theorem 1, we can conclude that there exists a strictly positive vector y such that

$$y^T S = 0.$$

Hence, the matrix S is elementally balanced. ◀

As a result of Theorem 2, we can decide whether a given model violates elemental balance by testing the No FL condition for it. That is, we can check whether there exists a non-negative non-zero vector in the column space of S (i.e. $\mathcal{A} \neq \emptyset$) [7].

Suppose the No FL condition is not satisfied for a stoichiometric matrix S (i.e. $\mathcal{A} \neq \emptyset$). Then there exists a linear combination of the reactions in S that can produce metabolites out of nothing, individually or in combination. For the definitions below we recall that the *support* of a vector $v \in \mathbb{R}^n$ is the set of its non-zero components, $supp(v) := \{i \in \{1, \ldots, n\} \mid v_i \neq 0\}$.

▶ **Definition 3** (Free Lunch)**.** Given a stoichiometric matrix S, we call a subset F of reactions a **free lunch** if there exists a non-zero vector $v \in \mathcal{A}$ with $supp(v) = F$. We call such a v a **vector corresponding** to the free lunch F.

▶ **Definition 4** (Free Lunch Metabolite)**.** For a free lunch F with a corresponding vector $v \in \mathcal{A}$, we call the metabolite t a **free lunch metabolite** if $t \in supp(Sv)$.

Step 1: No FL and all possible FL metabolites

In the first step of our method we introduce a test to verify the No FL condition for a given stoichiometric matrix S and identify the set of all possible FL metabolites. The verification of the No FL condition is performed by checking the feasibility of the linear program

$$Sv = w, w \geq 0, 1^T w \geq 1, \text{ where } v \in \mathbb{R}^n, w \in \mathbb{R}^m. \tag{2}$$

In addition, recall that $\|x\|_0 := |\{i \mid x_i \neq 0\}|$ denotes the size of the support of a vector x and t is a FL metabolite if $t \in supp(Sv)$ for some $v \in \mathcal{A}$. In the following lemma we show that the support of Sv for the vector v solving the non-linear program

$$\max_v \|Sv\|_0 \text{ subject to } Sv \geq 0, \ v \in \mathbb{R}^n \tag{3}$$

is the set of all possible FL metabolites for the matrix S. Note that if the optimum value of (3) is greater than zero, then (2) is feasible and the No FL Condition is not satisfied.

▶ **Lemma 5.** *Given a stoichiometric matrix $S \in \mathbb{R}^{m\times n}$, the support of Sv where v is a solution of the non-linear program in (3) corresponds to the set of all possible FL metabolites for the given matrix S.*

Proof. Let $\mathcal{T}_S$ be the set of all possible FL metabolites for S and $\hat{v}$ be a maximizer of (3). By definition of an FL metabolite we can observe that

$$supp(S\hat{v}) \subseteq \mathcal{T}_S.$$

Now, assume that $\mathcal{T}_S - supp(S\hat{v}) \neq \emptyset$ and let $t \in \mathcal{T}_S - supp(S\hat{v})$. Since t is a FL metabolite, there exists a vector $u \in \mathcal{A}$ such that $t \in supp(Su)$. Note that in this case $supp(S\hat{v}) \subsetneq supp(S(\hat{v}+u))$, which contradicts the maximality of $\|S\hat{v}\|_0$. Therefore, $\mathcal{T}_S = supp(\arg\max_v \|Sv\|_0)$. ◀

Although (3) is a non-linear program, we can in fact solve it with the following linear program.

▶ **Lemma 6.** *Given a matrix $S \in \mathbb{R}^{m\times n}$, the support of w, where w is a solution of*

$$\max_w (1^T w), \text{ subject to } Sv - w \geq 0,\ 0 \leq w \leq 1,\ (v,w) \in \mathbb{R}^{n+m} \tag{4}$$

corresponds to the set of all possible FL metabolites for the given matrix S.

Proof. Let (v,w) be a solution for (4). Also let p be a maximizer of (3) and define $q = Sp$. We show that

$$1^T w = \|q\|_0.$$

Note that since $0 \leq w \leq 1$, we have $1^T w = \|w\|_1 \leq \|q\|_0$. Now assume that $1^T w < \|q\|_0$. Let

$$\tilde{v} = \frac{v+p}{\min\{\min_{i\in supp(w)} w_i, \min_{j\in supp(q)} q_j\}}.$$

Also, define $\tilde{w} \in \mathbb{R}^m$ such that

$$\tilde{w}_i = \begin{cases} 1 & \text{if } \max\{w_i, q_i\} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Since the non-zero components of vector $S\tilde{v}$ are greater than or equal to 1, we found $(\tilde{v},\tilde{w}) \in \mathbb{R}^{n+m}$ that is in the feasible region of (4) and

$$1^T\tilde{w} \geq \|q\|_0 > 1^T w$$

which contradicts the maximality of $1^T w$. With a similar argument we can show that the solution of (4) is unique. Therefore, by Lemma 5, we are done. ◀

Step 2: Minimal Free Lunches

Theorem 2 shows that violation of the No FL condition implies violation of elemental balance in at least one reaction in the model. In the following section we continue the process in order to identify minimal subsets of reactions that contain such imbalanced reactions. We also show that, regardless of the chemical formulas assigned to the metabolites participating in the model, such a subset always contains an imbalanced reaction.

▶ **Definition 7** (Minimal Free Lunch)**.** The Free Lunch F is called a **Minimal Free Lunch**, if no proper subset of F is a free lunch.

A minimal FL can be computed with respect to any desired subset of FL metabolites. Let $\mathcal{T}$ be a non-empty subset of all FL metabolites in a model, and let us define the vector $w \in \mathbb{R}^m$ to be an indicator vector such that $w_i = 1$ if metabolite $i \in \mathcal{T}$ and $w_i = 0$ otherwise, for $i \in \{1, \ldots, m\}$. Then the smallest subset of reactions producing all FL metabolites in $\mathcal{T}$ can be computed by solving the following non-linear optimization problem:

$$\arg\min_{v} \|v\|_0 \text{ subject to } Sv - w \geq 0, v \in \mathbb{R}^n, w \in \mathbb{R}^m \tag{5}$$

This problem is NP-hard and challenging in practice [4], so we use the following steps to identify the minimal subsets instead. First we use the iterative reweighted ℓ_1 minimization algorithm, presented by Candès, Wakin and Boyd [6], to find an approximation for the non-convex minimization (5). We then reduce the (possibly non-minimal) FL found to a minimal FL by removing each reaction in turn and checking if the resulting combination is still a FL; if it is not, the reaction is restored. This procedure produces a minimal FL [7].

▶ **Lemma 8.** *Suppose $\mathcal{M}$ is a minimal FL. Then, for every assignment of chemical formulas to the metabolites involved in $\mathcal{M}$ there exists an elementally imbalanced reaction in $\mathcal{M}$. Furthermore, for each reaction $r \in \mathcal{M}$ there exists an assignment of chemical formulas to these metabolites that makes r imbalanced while making the reactions in $\mathcal{M} - \{r\}$ balanced.*

Proof. Assume $\mathcal{M}$ is a minimal FL. Let S' be the submatrix of the stoichiometric matrix S corresponding to the reactions in $\mathcal{M}$. Since $\mathcal{M}$ is a FL, by Theorem 2, there exits no strictly positive vector y for which (1) holds. Therefore, if we think of the vector y as containing the number of elements in each metabolite, regardless of the chemical formulas assigned to the metabolites in S', there exists at least one imbalanced reaction in $\mathcal{M}$.

In addition, since $\mathcal{M}$ is minimal, for each $r \in \mathcal{M}$, $\mathcal{M} - \{r\}$ cannot be a FL. Let S'' be the submatrix of S corresponding to the reactions in $\mathcal{M} - \{r\}$. Therefore, by Theorem 2, there exits a strictly positive vector y such that

$$y^T S'' = 0.$$

We can assume that the coefficients of y are rational, since S has rational entries. Let $\hat{y} = Ny$ be the vector obtained by scaling y by the least common multiple N of the denominators in y. Then $\hat{y}^T S'' = 0$ as well, and $\hat{y}$ contains positive integers. Thus we can set the formula of metabolite j to $C_{\hat{y}_j}$, and observe that (1) holds for $\mathcal{M} - \{r\}$ with these one-atom formulas. ◀

We solve the linear programs in Steps 1 and 2 using the QSOpt_ex solver [3], which checks that the solutions are correct in exact rational arithmetic. Some of our previous work [8, 13] has argued that this is necessary in order to ensure accuracy and reproducibility of the solutions. We use the QSOpt_ex API for Python [19] created by Jon Steffensen [17] to streamline the process.

Step 3: Free Lunch Witnesses

In the current section we aim to find the smallest number of reactions that are guaranteed to be involved in at least one FL each. We observe that the optimum of the minimization problem

$$\min_{y} \|y^T S\|_0, \text{ subject to } y > 0 \tag{6}$$

represents a lower bound on the minimum number of FL's over all possible metabolite formulas, and the minimizer represents a smallest set of reactions involved in FL's. Moreover,

the optimum value of (6) also gives us a lower bound on the number of imbalanced reactions in the model for an arbitrary or a specific set of metabolite formulas. In other words, if the optimum value of (6) is k, we know that at least k distinct reactions are elementally imbalanced, no matter what metabolite formulas we choose, and in particular, the metabolic network cannot be elementally balanced unless $k = 0$. This motivates the following

▶ **Definition 9** (Free Lunch Witness)**.** The reaction r is called a **Free Lunch Witness**, if $r \in supp(y^T S)$ for some $y \in \text{argmin}_{y>0} \|y^T S\|_0$.

Even though we cannot guarantee this, it turns out to be more likely than not in practice that a FL witness is elementally imbalanced for a given set of metabolite formulas, as we discuss in more detail in the Results section.

In order to solve (6), we can use the following mixed integer linear program:

$$\min_x \ 1^T x \text{ subject to } x \geq -y^T S, \ x \geq y^T S, y \geq \epsilon, \ x \in \{0,1\}^n, \ y \in \mathbb{R}^m \tag{7}$$

where ϵ is a small positive scalar (in practice, we use $\epsilon = 10^{-4}$). The lower bound of ϵ ensures that the vector $y = 0$ is not an admissible solution of (6), so that any y that is optimal for (7) can be scaled to a $\hat{y}$ that is optimal for (6) (with the same objective value), and vice versa.

We point out that the solution vector to the optimization problem (7) is in general not unique (even though the optimal value is), and may change based on the solver as well as the order of the reactions or metabolites in the stoichiometric network S. We used CPLEX 12.8.0 [1] via its API for Python [19] to solve the integer optimization problem (7).

3 Results

We analyzed a collection of 84 previously published metabolic network models in the BiGG database [11]. From our analysis we excluded any pseudo-reactions, namely, the biomass reactions (those containing the word "biomass" or "growth" in their name) as well as import and export reactions (those that have only positive or only negative stoichiometric coefficients, i.e. produce something out of nothing or reduce something to nothing, respectively). This ensures that all the reactions included in a stoichiometric matrix S were *bona fide* biochemical reactions, with both reactant and product sides non-empty. We expected all those reactions to be elementally balanced, as required in the model construction protocol [18].

Nevertheless, we found instances of elemental balance violation in 72 of the 84 network models by using the provided chemical formulas. An additional 7 of the models had no instance of elemental balance violation and 5 of them had no specified chemical formulas for the metabolites. Note that we only consider the following *bona fide* atoms in our decision about whether a reaction is imbalanced: C, Fe, H, Mg, Na, N, O, P, S, Z (the latter represents a photon, which can for instance be used by photosynthetic organisms). We do not consider symbols for functional groups or other moieties such as I, K, M, R, U, X, Y; in other words, if all the *bona fide* atoms were balanced, the reaction was considered to be balanced no matter whether these additional symbols were present on both sides of the reaction or only one.

The pie chart in Figure 1 shows the relative frequency of the atoms that are not balanced. There are a total of 810 imbalanced reactions, and 1492 atomic imbalances, so an average imbalanced reaction has just under two imbalanced atoms. Almost half of all atomic imbalances are proton or hydrogen imbalances, involving the H atom.

We applied the first step of our method to find instances of FL metabolites that result from elemental balance violations. We found that 40 out of 84 models include at least one FL metabolite. The number of FL metabolites ranged from 3 to 1791, with a mean just

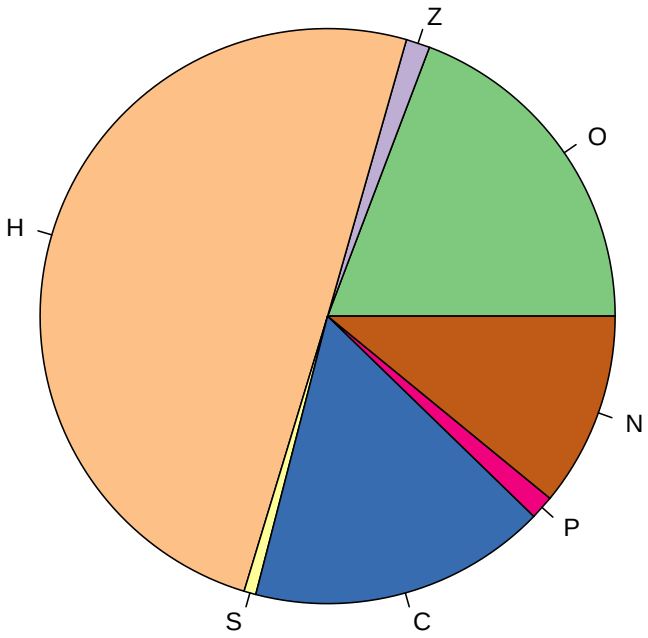

Figure 1 Frequency of imbalanced reactions in the BiGG networks by imbalanced atom.

under 139. There are a total of 5552 FL metabolites, of which 2912 are distinct; more than 90% of them happen to be in four of the models, namely, those for *Homo sapiens* (RECON1), *Mus musculus* (iMM1415), *Escherichia coli* (iECIAI1_1343), and *Salmonella enterica* (STM_v1_0). The vast majority of the other models only have 3 FL metabolites, which are predominantly protons (H) in three different compartments, namely, [c] (cytosol), [p] (periplasm), and [e] (extracellular space).

The pie chart in Figure 2 shows the relative frequency of the FL metabolites that are found in more than 5 of the metabolic networks we investigated. There are a total of 9 such metabolites, which all happen to be currency metabolites - H (protons), H_2O (water), H_2O_2 (hydrogen peroxide), O_2 (oxygen), and SO_2 (sulfur dioxide) - in the compartments listed above.

We also applied the second step to find a minimal FL with respect to one of the FL metabolites[1] in each of these 40 models, and their sizes ranged from 2 to 71, with a mean of 5.8. By using the specified chemical formulas, we were able to verify the existence of at least one imbalanced reaction in each minimal FL in 38 out of 40 models, in accordance with our theoretical results (the remaining 2 models, namely, iMM1415 and iECIAI1_1343, did not have any chemical formulas specified for their metabolites).

Finally, by applying the third step of our method, we found small sets of FL witnesses in all 40 models, with sizes ranging from 1 to 12, and a mean of 2.2. Just as for the second step, we compared the set of detected FL witnesses to the set of imbalanced reactions based on the chemical formulas provided in 38 of the models. We observed that, in 28 out of 38 models, at least one of the FL witnesses identified by our method was an imbalanced reaction. Furthermore, just over half (41 out of 79) of the FL witnesses were imbalanced reactions.

[1] We chose the FL metabolite in following way. First we found an optimum solution $(v, w) \in \mathbb{R}^{n+m}$ of $\min_v \|v\|_1$ subject to: $Sv - w \geq 0, w \geq 0, 1^T w \geq 1$ to roughly approximate the subset of FL metabolites for which we may have a small FL. We then chose the first index in $supp(w)$ as the desired FL metabolite.

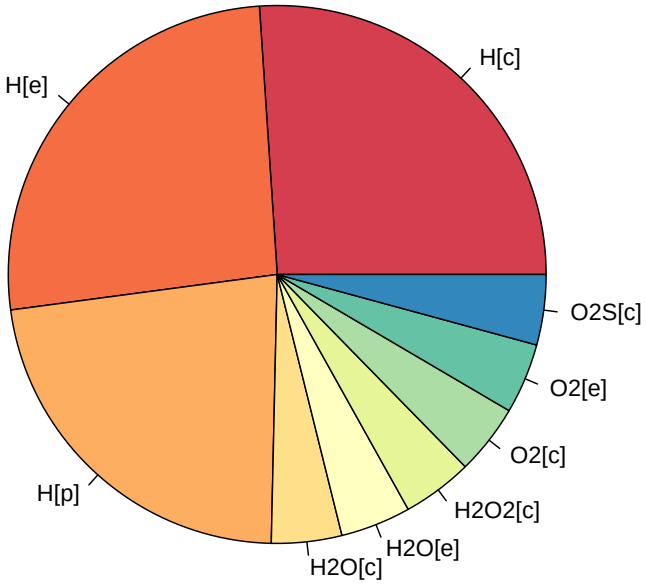

Figure 2 Relative frequency of the most common free lunch metabolites in the BiGG networks.

This suggests that FL witnesses can be good candidates for imbalanced reactions in the absence of chemical formulas for the metabolites in a network.

A detailed summary of the 40 models with at least one FL is presented in Table 1. We now discuss in detail the results of applying our algorithms to four particular models, chosen for illustration purposes. They are arranged in increasing order of complexity.

The first example is iECB_1328, a model for *Escherichia coli* B strain REL606 (the first model in Table 1). For this model, our method identifies 3 FL metabolites, which happen to be protons, H, in three different compartments. We also find a minimal FL that consists of two reactions - OPET decarboxylase and 5-carboxy-2-oxohept-3-enedioate decarboxylation:

$$C_8H_5O_7[c] \rightarrow CO_2[c] + C_7H_6O_5[c] \tag{8a}$$

$$H[c] + C_8H_5O_7[c] \rightarrow CO_2[c] + C_7H_6O_5[c] \tag{8b}$$

We can see that by subtracting reaction (8b) from reaction (8a), we get a net overall reaction that produces the FL metabolite H[c] out of nothing, while all other metabolites cancel out. The OPET decarboxylase reaction is also detected as the single FL witness in this model, and in fact, it is easy to see that the H atom is not balanced in this reaction (there are 5 H's on the reagent side and 6 on the product side). Thus, in this model, the FL witness is an imbalanced reaction.

The second example is iBWG_1329, another model for *Escherichia coli*, but a different strain, BW2952. Similarly to the first example, our method identifies 3 FL metabolites, which happen to be protons, H, in three different compartments. This time, the minimal FL consists of three reactions - nucleoside triphosphate pyrophosphorylase (atp), phenylacetate-CoA ligase, and phenylacetyl CoA thioesterase:

$$C_{10}H_{12}N_5O_{13}P_3[c] + H_2O[c] \rightarrow C_{10}H_{12}N_5O_7P[c] + H[c] + HO_7P_2[c] \tag{9a}$$

$$C_{10}H_{12}N_5O_{13}P_3[c] + C_{21}H_{32}N_7O_{16}P_3S[c] + C_8H_7O_2[c] \rightarrow$$
$$\rightarrow C_{10}H_{12}N_5O_7P[c] + C_{29}H_{38}N_7O_{17}P_3S[c] + HO_7P_2[c] \tag{9b}$$

$$H_2O[c] + C_{29}H_{38}N_7O_{17}P_3S[c] \rightarrow C_{21}H_{32}N_7O_{16}P_3S[c] + C_8H_7O_2[c] \tag{9c}$$

We can see that subtracting (9a) from the sum of (9b) and (9c) results in the production of the FL metabolite H[c] out of nothing. We can also see that the H atoms are not balanced in (9c), with 40 on the reagent side and 39 on the product side, while (9a) and (9b) are balanced. In this model, our method detects reaction (9b) as the unique FL witness. Thus, in this model, the FL witness is itself a balanced reaction, but is part of a FL of size 3.

The third example is iJN678, a model for *Synechocystis sp.* PCC 6803, a cyanobacterium that can grow by oxygenic photosynthesis [9]. For this model, our method once again identifies 3 FL metabolites, which this time are photons (Z) in three compartments. The minimal FL contains five reactions - ATP synthetase(u), cyanophycinase, cyanophycin sinthetase, cytochrome oxidase bd (plastocianine-8 2 protons) (lumen), and photosystem II, which we denote by R_1 through R_5 for convenience:

$$R_1 : 3C_{10}H_{12}N_5O_{10}P_2[c] + 3HO_4P[c] + 14H[u] \rightarrow \\ \rightarrow 3C_{10}H_{12}N_5O_{13}P_3[c] + 11H[c] + 3H_2O[c] \tag{10a}$$

$$R_2 : 2H_2O[c] + C_{10}H_{17}N_5O_4R[c] \rightarrow C_6H_{15}N_4O_2[c] + C_4H_6NO_4[c] + R[c] \tag{10b}$$

$$R_3 : C_6H_{15}N_4O_2[c] + C_4H_6NO_4[c] + 2C_{10}H_{12}N_5O_{13}P_3[c] + R[c] \rightarrow \\ \rightarrow 2C_{10}H_{12}N_5O_{10}P_2[c] + 2H[c] + 2HO_4P[c] + C_{10}H_{17}N_5O_4R[c] \tag{10c}$$

$$R_4 : C_8H_{10}O_2[u] + \frac{1}{2}O_2[u] \rightarrow C_8H_8O_2[u] + H_2O[u] \tag{10d}$$

$$R_5 : 2H[c] + C_8H_8O_2[u] + H_2O[u] + 2Z[c] \rightarrow 2H[u] + C_8H_{10}O_2[u] + \frac{1}{2}O_2[u] \tag{10e}$$

It is easy to check (but far from obvious to guess!) that $2R_1 + 3R_2 + 3R_3 + 14R_4 + 14R_5$ would result in the consumption of 28 units of the FL metabolite Z[c], while all other metabolites cancel out. Also, note that Z is only present on the reagent side of (10e). In this example, our method detects reaction (10e) as one of the 3 FL witnesses, and in this case, it is the one imbalanced FL witness.

For the final example we chose a model that does not have any chemical formulas assigned to its metabolites, namely, model iECIAI1_1343, which represents *Escherichia coli.* IAI1. For this model our method identifies 1279 FL metabolites and a minimal FL which contains two reactions - OPET decarboxylase and 5-carboxy-2-oxohept-3-enedioate decarboxylation:

$$\text{5-Carboxy-2-oxohept-3-enedioate}[c] \rightarrow \text{2-Hydroxyhepta-2-4-dienedioate}[c] + CO_2[c] \tag{11a}$$

$$H[c] + \text{5-Carboxy-2-oxohept-3-enedioate}[c] \rightarrow \text{2-Hydroxyhepta-2-4-dienedioate}[c] + CO_2[c] \tag{11b}$$

We can see that by subtracting reaction (11b) from reaction (11a), we get the FL metabolite H[c] out of nothing, while all other metabolites cancel out. Note that this is the same FL as the one identified in equations (8a) and (8b), although in this model they are not specified via chemical formulas. Moreover, our method identifies 4 FL witnesses; as a result of the absence of chemical formulas, we are not able to verify whether they are elementally balanced.

4 Conclusion

Our approach represents the first attempt to automatically verify elemental balance in a GSMN without access to the chemical formulas for the metabolites. It quickly and efficiently identifies small sets of reactions which must contain at least one imbalanced reaction, or determines that no such set exists. In addition, it can provide a lower bound on the number

Table 1 The summary of our results. n and m denote the number of reactions and metabolites, respectively. m_{FL} and n_{FL} is the number of FL metabolites and the size of a minimal FL we found, respectively. n_I and n_{FLW} are the numbers of imbalanced reactions and FL witness reactions, respectively. Finally, n_{IFLW} if the number of FL witness reactions that are imbalanced.

Model	Organism	n	m	m_{FL}	n_I	n_{FL}	n_{FLW}	n_{IFLW}
iECB_1328	E. coli B str. REL606	2748	1951	3	10	2	1	1
iNJ661	M. tuberculosis H37Rv	1025	825	192	10	4	4	1
iY75_1357	E. coli str. K-12	2759	1953	3	9	3	1	0
iMM1415	Mus musculus	3726	2775	1767	N/A	12	6	N/A
iECO103_1326	E. coli O26:H11 str. 11368	2758	1958	3	17	2	2	2
iECUMN_1333	E. coli UMN026	2740	1935	3	7	2	1	1
iEcolC_1368	E. coli ATCC 8739	2768	1969	3	16	2	2	1
RECON1	Homo sapiens	3741	2766	1791	5	40	4	1
iECO111_1330	E. coli O111:H- str. 11128	2760	1959	3	19	2	2	2
iSBO_1134	S. boydii Sb227	2591	1908	3	7	2	1	1
iRC1080	C. reinhardtii	2191	1706	21	61	6	12	0
iJB785	Synechococcus elongatus	849	768	16	11	14	4	0
iS_1188	S. flexneri 2a str. 2457T	2619	1914	3	4	2	1	1
STM_v1_0	S. enterica subsp. enterica	2545	1802	344	1	71	2	0
iECBD_1354	E. coli BL21	2748	1952	3	10	2	1	1
iECO26_1355	E. coli O26:H11 str. 11368	2780	1965	3	17	2	2	2
iWFL_1372	E. coli W	2782	1973	3	21	2	2	2
iECSE_1348	E. coli SE11	2768	1957	3	18	2	2	2
iECD_1391	E. coli BL21	2741	1943	3	10	2	1	1
iEcDH1_1363	E. coli DH1	2750	1949	3	8	3	1	1
iECDH1ME8569_1439	E. coli DH1	2755	1950	3	9	3	1	0
iECDH10B_1368	E. coli str. K-12	2742	1947	3	9	3	1	1
iB21_1397	E. coli BL21	2741	1943	3	10	2	1	1
iEcE24377_1341	E. coli O139:H28 str. E24377A	2763	1972	3	17	3	1	1
iUMNK88_1353	E. coli UMNK88	2777	1969	3	18	3	1	0
iETEC_1333	E. coli ETEC H10407	2756	1962	3	14	3	1	0
iECW_1372	E. coli W	2782	1973	3	21	2	2	2
iSF_1195	S. flexneri 2a str. 301	2630	1917	3	7	2	1	1
iSbBS512_1146	S. boydii CDC 3083-94	2591	1910	3	7	2	1	0
iEKO11_1354	E. coli KO11FL	2778	1972	3	21	2	2	2
iCHOv1	Cricetulus griseus	6663	4456	46	92	4	9	4
iSSON_1240	S. sonnei Ss046	2693	1936	3	9	2	1	1
iJN678	Synechocystis sp. PCC 6803	863	795	3	9	5	3	3
iSFxv_1172	S. flexneri 2002017	2638	1918	3	7	2	1	1
iBWG_1329	E. coli BW2952	2741	1949	3	9	3	1	0
iECIAI1_1343	E. coli IAI1	2765	1968	1279	N/A	2	4	N/A
iSFV_1184	S. flexneri 5 str. 8401	2621	1917	3	1	2	1	0
iLB1027_lipid	P. tricornutum	4456	2172	3	6	6	2	1
iEcHS_1320	E. coli HS	2753	1963	3	21	2	2	2
iEC55989_1330	E. coli 55989	2756	1953	3	14	3	1	1
Average				138.8	14.0	5.8	2.2	1

of imbalanced reactions present in the model. Lastly, it identifies all the metabolites that can be produced out of nothing by the overall reaction of a combination of model reactions.

The difference in the number of elementally imbalanced models examined with (72) and without (40) chemical formulas (out of a total of 84) can be addressed by examining equation (1). Since the vector y in (1) is not unique, different y vectors can represent different sets of metabolites. Therefore, although a model with stoichiometric matrix S may contain

imbalanced reactions for a specific set of chemical formulas, the same stoichiometric matrix S may be elementally balanced for a different set of formulas. Thus, our method can guarantee a violation of elemental balance by finding one or more FL metabolites, but the absence of such metabolites does not guarantee that the model is elementally balanced.

Nevertheless, we believe that our approach provides a complimentary way of ascertaining a model's soundness and point out potential issues, as part of the model verification process. It can - and should - be accompanied by a verification of the elemental balance of individual reactions whenever possible, i.e. when the chemical formulas are specified for its metabolites. We thus hope that our method is a useful contribution to metabolic network model verification.

References

1 CPLEX optimizer. URL: `www-01.ibm.com/software/integration/optimization/cplex-optimizer`.

2 V Acuña, F Chierichetti, V Lacroix, A Marchetti-Spaccamela, M-F Sagot, and L Stougie. Modes and cuts in metabolic networks: complexity and algorithms. *BioSystems*, 95:51–60, 2009.

3 D Applegate, W Cook, S Dash, and D Espinoza. Exact solutions to linear programming problems. *Operations Research Letters*, 35:693–699, 2007.

4 D Bertsimas, A King, and R Mazumder. Best subset selection via a modern optimization lens. *Ann. Statist.*, 44(2):813–852, 2016.

5 KC Border. Alternative linear inequalities. *Cal Tech Lecture Notes*, 2013.

6 EJ Candès, MB Wakin, and SP Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14(5-6):877–905, 2008.

7 L Chindelevitch. *Extracting Information from Biological Networks*. PhD thesis, MIT, 2010.

8 L Chindelevitch, J Trigg, A Regev, and B Berger. An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. *Nature Communications*, 5, 2014.

9 T Heidorn, D Camsund, H Huang, P Lindberg, P Oliveria, K Stensjo, and P Lindblad. Synthetic biology in cyanobacteria: Engineering and analyzing novel functions. In *Methods in Enzymology*, volume 497, pages 539–579. Academic Press, 2011.

10 M Imieliński, C Belta, H Rubin, and Á Halász. Systematic analysis of conservation relations in escherichia coli genome-scale metabolic network reveals novel growth media. *Biophysical Journal*, 90(8):2659–2672, 2006.

11 ZA King, J Lu, A Dräger, P Miller, S Federowicz, JA Lerman, A Ebrahim, BO Palsson, and NE Lewis. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research*, 44:D515–D522, 2015.

12 S Klamt and E Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20:226–234, 2004.

13 C Le and L Chindelevitch. The MONGOOSE rational arithmetic toolbox. In Marco Fondi, editor, *Metabolic Network Reconstruction and Modeling*, volume 1716 of *Methods in Molecular Biology*, pages 77–99. Humana Press, New York, NY, 2018.

14 J Monk, J Nogales, and B Palsson. Optimizing genome-scale network reconstructions. *Nature Biotechnology*, 32:447–452, 2014.

15 A Ravikrishnan and K Raman. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Briefings in Bioinformatics*, 16(6):1057–1068, 2015.

16 S Schuster and T Höfer. Determining all extreme semi-positive conservation relations in chemical reaction systems: a test criterion for conservativity. *Journal of the Chemical Society, Faraday Transactions*, 87:2561–2566, 1991.

17 JL Steffensen, K Dufault-Thompson, and Y Zhang. Psamm: A portable system for the analysis of metabolic models. *PLOS Computational Biology*, 12(2):e1004732, 2016.

18 I Thiele and B Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5:93–121, 2010.

19 G van Rossum. Python tutorial. Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.

20 A. Varma and B. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* w3110. *Appl. Environ. Microbiol.*, 60:3724–3731, 1994.