# Knowledge-based generalization of metabolic models

Anna Zhukova*

Inria Bordeaux Sud-Ouest/University of Bordeaux/CNRS UMR 5800

joint project-team Magnome

351, cours de la Libération F-33405 Talence cedex, France

David James Sherman†

Inria Bordeaux Sud-Ouest/University of Bordeaux/CNRS UMR 5800

joint project-team Magnome

351, cours de la Libération F-33405 Talence cedex, France

*corresponding author - anna.zhukova@inria.fr

†david.sherman@inria.fr

**Abstract**

**Background:** Genome-scale metabolic model reconstruction is a complicated process beginning with (semi-)automatic inference of the reactions participating in the organism's metabolism, followed by many iterations of network analysis and improvement. Despite advances in automatic model inference and analysis tools, reconstruction may still miss some reactions or add erroneous ones. Consequently a human expert's analysis of the model will continue to play an important role in all the iterations of the reconstruction process. This analysis is hampered by the size of the genome-scale models (typically thousands of reactions), which makes it hard for a human to understand them.

**Results:** To aid human experts in curating and analyzing metabolic models, we have developed a method for *knowledge-based generalization* that provides a higher-level view of a metabolic model, masking its inessential details while presenting its essential structure. The method groups biochemical species in the model into semantically equivalent classes based on the ChEBI ontology, identifies reactions that become equivalent with respect to the generalized species, and factors those reactions into *generalized reactions*.

**Conclusions:** Generalization allows curators to quickly identify divergences from the expected structure of the model, such as alternative paths or missing reactions, that are the priority targets for further curation. We have applied our method to genome-scale yeast metabolic models and shown that it improves understanding by helping to identify both specificities and potential errors.

**Keywords:** metabolic modeling; generalization; genome-scale reconstruction.

# Introduction

Genome-scale metabolic models are complex networks that describe the thousands of reactions and molecular species that participate in an organism's metabolism. The complexity of these networks makes it difficult for human curators to understand, analyze, and verify them, since the individual reactions that require their attention are hidden among the reactions that are correctly described. The priority targets are gaps in pathways, and reactions that are in some sense *unusual*, because they represent shortcuts or alternative paths. Efficient curation of genome-scale models requires analysis and exploration tools that synthesize high-level views of the network and focus curator attention on these small subsets of unusual reactions.

Curation is performed after the automatic inference of the draft metabolic model. Inference methods combine databases of reactions and pathways with genomic information and existing models for similar organisms (Swainston et al., 2011). Genomic data for the new organism is compared to the data of the reference organism, to find genomic evidence such as the presence of catalysing enzymes for the reactions conserved in the new organism. Starting from the inference of a draft model, the model refinement process includes several iterations of model analysis, error detection, and improvement (Thiele and Palsson, 2010). The models produced at each iteration are intended for computer simulation, and so describe all the reactions thought to participate in the organism's metabolism. Although automatic model inference tools and genomic comparison methods are becoming steadily more sophisticated, they may still leave gaps in the model or add erroneous reactions. Curation by human experts is necessary.

Much of the complexity of the reaction network comes from biochemically similar reactions that operate on slightly different substrates. For example, in the peroxisome compartment of *Yarrowia lypolitica* model (*MODEL1111190000* (Loira et al., 2012)) six *acetyl-CoA oxidase* reactions are present, transforming *fatty acyl-CoA*s differing in their carbon chain length (*decanoyl-CoA*, *lauroyl-CoA*, etc.) into the corresponding *unsaturated fatty acyl-CoA*s. There are also several similar reactions for other steps of the *β-oxidation of fatty acids* pathway (Metzler, 2001). Although all of these details are needed for accurate computer simulation, and are common to many models, not all of them are interesting for a curator. It is instead the differences from the common pattern that demand attention. They may be errors in the model, such as missing steps

or erroneous connections between pathways, or they may be organism-specific differences such as alternative pathways that are biologically interesting.

To aid human understanding of genome-scale models, while keeping the details needed for a computer simulation, we propose a 3-level *zoomable* approach:

- The most abstract level represents compartmentalization of the model, and focusses on such questions as: Are all the compartments present? Are they well connected by transport reactions?
- The second level shows the modules inside of each of the compartments. The questions to be addressed on this level include: Are all the essential processes present? Is the structure of each process correct? Is there any organism-specific adaptation of the structure?
- The most detailed level is intended for computer simulation and represents the inner structure of each of the modules with all the species, reactions and their kinetics, stoichiometry and constraints.

The two abstract levels are intended for a human expert, and the last one for the computer.

In this study we focus on the second level of abstraction, that represents the modules inside compartments. A fair amount of work has been done on identifying reusable modules. These approaches can be divided into two groups: *series* and *parallel*. A *series* approach operates on chains of reactions, and generalizes them as a series, consequently hiding the structure of the network. An example of a *series* approach is representing the network as a set of metabolic pathways (KEGG (Kanehisa et al., 2012), MetaCyC (Caspi et al., 2012)), that can be further divided, for example, into reaction modules (conserved sequences of reactions along the metabolic pathways) (Muto et al., 2013).

The other type of approach operates on reactions that are *parallel*, keeping the steps and preserving the general view of the network. An example of this approach is grouping reactions based on EC (Enzyme Commission) numbers (Tohsato et al., 2000). The drawback of this approach is that it is not applicable to networks with no EC number assigned or reactions with no catalysing enzymes identified. We have developed another *parallel*-reaction method for knowledge-based generalization of metabolic models, which does not depend on enzyme information. It provides a higher-level view of a model while keeping its essential structure

and omitting the details.

**Definition 1** *The* model generalization *process groups chemical species present in the model into equivalence classes, and merges each class into a generalized chemical species. Reactions that involve same generalized chemical species are then factored together into a generalized reaction.*

By applying the model generalization process, we can build a simplified model that focusses on the high level relationships. The simplified model can be further divided into pathways.

# Mathematical basis

## Basic definitions

We represent a *metabolic model* $M$ as a pair of two sets: a set $S$ of biochemical species, and a set $R$ of reactions between them:

$$\begin{aligned} M &= \langle S, R \rangle && \text{- model,} \\ S &= \{s_1, \ldots, s_n\} && \text{- species set,} \\ R &= \{r_1, \ldots, r_m\} && \text{- reaction set.} \end{aligned}$$

We represent each *reaction* $r \in R$ as a pair of sets of species: its reactants and products. A chemical reaction may be represented by a balanced chemical equation, showing the formulae of the reactants and products, and the changes that take place (Clugston and Flemming, 2000). This definition leads to restriction (1) that all the species participating in the reaction must be different.

$$\begin{aligned} r = \quad & \langle \{s_1^{(rs)}, \ldots, s_k^{(rs)}\}, \{s_1^{(ps)}, \ldots, s_l^{(ps)}\} \rangle \in R \subset \langle 2^S \times 2^S \rangle, \\ & \text{where } s_1^{(rs)} \neq \ldots \neq s_k^{(rs)} \neq s_1^{(ps)} \neq \ldots \neq s_l^{(ps)} \end{aligned} \tag{1}$$

To perform the model generalization, we define an *equivalence operation* $\sim$ on the species set, and group species into equivalence classes: $[s]^{\sim} = \{\tilde{s} \in S | \tilde{s} \sim s\}$.

Species equivalence imposes reaction equivalence: two reactions are equivalent if their corresponding reactant and product species sets are pairwise equivalent.

$$\begin{aligned} \forall r, \tilde{r} \in R \quad & r = \langle \{s_1^{(rs)}, \ldots, s_k^{(rs)}\}, \{s_1^{(ps)}, \ldots, s_l^{(ps)}\} \rangle, \\ & \tilde{r} = \langle \{\tilde{s}_1^{(rs)}, \ldots, \tilde{s}_{\tilde{k}}^{(rs)}\}, \{\tilde{s}_1^{(ps)}, \ldots, \tilde{s}_{\tilde{l}}^{(ps)}\} \rangle \\ & k = \tilde{k}, l = \tilde{l}, \\ r \sim \tilde{r} \iff \quad & \wedge \ \forall i \ 0 \leq i \leq k \ \exists! \tilde{i} \ 0 \leq \tilde{i} \leq \tilde{k} : \ s_i^{(rs)} \sim \tilde{s}_{\tilde{i}}^{(rs)}, \\ & \quad \forall j \ 0 \leq j \leq l \ \exists! \tilde{j} \ 0 \leq \tilde{j} \leq \tilde{l} : \ s_j^{(ps)} \sim \tilde{s}_{\tilde{j}}^{(ps)}. \end{aligned}$$

Equivalent reactions are factored together into a generalized reaction that operates with generalized species (i.e. species equivalence classes): $[r]^{\sim} = \langle\{[s_1^{(rs)}]^{\sim}, \ldots, [s_k^{(rs)}]^{\sim}\}, \{[s_1^{(ps)}]^{\sim}, \ldots, [s_l^{(ps)}]^{\sim}\}\rangle$.

In order to maintain the number of distinct species participating in a reaction, restriction (1'), analogous to restriction (1), must be satisfied:

$$[s_1^{(rs)}]^{\sim} \neq \ldots \neq [s_k^{(rs)}]^{\sim} \neq [s_1^{(ps)}]^{\sim} \neq \ldots \neq [s_l^{(ps)}]^{\sim} \quad (1')$$

In order to avoid creation of paths in the generalized model that are not based on the evidence from the initial model, we introduce restriction (2): Species that do not participate in any pair of equivalent reactions and do not have any common equivalent species must not be grouped together.

$$\forall s, \tilde{s} \in S \; s \sim \tilde{s} \;\Rightarrow\; \vee \begin{array}{l} \exists r \sim \tilde{r} \in R : s \in reactants(r) \wedge \tilde{s} \in reactants(\tilde{r}) \\ \exists r \sim \tilde{r} \in R : s \in products(r) \wedge \tilde{s} \in products(\tilde{r}) \\ \exists \dot{s} \in S : \; s \tilde{\sim} \dot{s} \wedge \dot{s} \tilde{\sim} \tilde{s}. \end{array} \quad (2)$$

Note that restriction (2) can be reformulated as maximizing the number of species equivalence classes while keeping the reaction equivalence classes unchanged.

The *generalized model* $M/\sim$ is a pair of generalized species and reaction sets (quotient sets):

$$\begin{array}{rcll} M/\sim & = & \langle S/\sim, R/\sim\rangle & \text{- generalized model,} \\ S/\sim & = & \{[s_1]^{\sim}, \ldots, [s_{\tilde{n}}]^{\sim}\} & \text{- quotient species set,} \\ R/\sim & = & \{[r_1]^{\sim}, \ldots, [r_{\tilde{m}}]^{\sim}\} & \text{- quotient reaction set.} \end{array}$$

The generalized model is a **zoom out** of the initial model. It provides a higher-level view by including less species and reactions, but more generic ones. For example, *3-oxodecanoyl-CoA*, *3-oxolauroyl-CoA*, and *3-oxohexanoyl-CoA* species of the initial model can be generalized into *oxo-fatty acyl-CoA*.

Every reaction of the generalized model corresponds to at least one reaction of the initial model, having the same topology (number of distinct reactant and product species) and operating on species that can

be zoomed out into those participating in the generalized reaction. The appropriate level of abstraction is defined with respect to the initial model as the most general one that satisfies restrictions (1') and (2).

The method and restrictions are described in figure 1.

### Specific and ubiquitous species

We say that a *ubiquitous species* is one that participates in many reactions (more than some threshold), such as *water*, *hydrogen*, *oxygen*, etc. Grouping such species would increase the number of reactions each group would participate in, beyond the sharing already common to most of the models, and decrease readability; in fact, during visualisation these species are often even duplicated to improve readability (Rohn et al., 2012). Consequently we do not generalize ubiquitous species. In the generalized model each forms a trivial equivalence class:

$$S^{(ub)} = \{s_1^{(ub)}, \ldots, s_{\breve{n}}^{(ub)}\} \subset S : \forall i \, [s_i^{(ub)}]^{\sim} = \{s_i^{ub}\}$$

*Specific species* are the others, which we divide into non-trivial equivalence classes and generalize accordingly.

## Model generalization problem

**Problem 1** *Given a metabolic model* $M = \langle S, S^{(ub)} \subset S, R \rangle$ *that describes* $n$ *species (including* $\breve{n} \leq n$ *ubiquitous ones) and* $m$ *reactions, find an equivalence operation* $\sim$ *that obeys restrictions (1') and (2), and minimizes the number of reaction equivalence classes* $\sharp R/\sim$.

We will solve model generalization problem 1 in three steps:

1. Define the most general equivalence operation $\mathring{\sim}$ (having minimal number of species equivalence classes $\sharp S/\mathring{\sim}$), that does not take into account the restrictions;
2. Modify the current equivalence operation to satisfy the restriction (1');
3. Modify the current equivalence operation to satisfy the restriction (2).

**Step 1. Equivalence operation $\mathring{\sim}$.**

**Definition 2** *Given a model* $M = \langle S, S^{(ub)} \subset S, R\rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$, *we define an* equivalence operation $\mathring{\sim}$ *on the species set* $S$ *as forming* $\breve{n}+1$ *equivalence classes in the quotient set* $S/\mathring{\sim}$*: one for each of the ubiquitous species, and one for all the other species:*

$$\forall s^{(ub)} \in S^{(ub)} \quad [s^{(ub)}]^{\mathring{\sim}} = \{s^{(ub)}\},$$
$$\forall s, \tilde{s} \in S \backslash S^{(ub)} \quad [s]^{\mathring{\sim}} = [\tilde{s}]^{\mathring{\sim}} = S \backslash S^{(ub)}.$$

**Lemma 1** *For any equivalence operation* $\sim$ *on the model* $M = \langle S, S^{(ub)} \subset S, R\rangle$*, the corresponding quotient species set* $S/\sim$ *and quotient reaction set* $R/\sim$ *are partitions of, respectively, the quotient species set* $S/\mathring{\sim}$ *and the quotient reaction set* $R/\mathring{\sim}$ *induced by* $\mathring{\sim}$*:*

$$\forall \text{ equivalence operation } \sim \text{ defined on } \langle S, S^{(ub)}, R\rangle$$
$$\forall s \in S \quad [s]^{\sim} \subset [s]^{\mathring{\sim}}$$
$$\forall r \in R \quad [r]^{\sim} \subset [r]^{\mathring{\sim}}$$

**Data**: $M = \langle S, S^{(ub)} \subset S, R \rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$ - metabolic model describing $n$ species, $\breve{n}$ among them being ubiquitous, and $m$ reactions.

**Result**: $\mathring{\sim}$ - equivalence operation described in Lemma 1, $M/\mathring{\sim} = \langle S/\mathring{\sim}, S^{(ub)}/\mathring{\sim} \subset S/\mathring{\sim}, R/\mathring{\sim} \rangle$ - corresponding generalized model.

$S/\mathring{\sim} \leftarrow \emptyset$ ; `// resultant quotient species set` $S/\mathring{\sim} \subset 2^S$

$S^{(ub)}/\mathring{\sim} \leftarrow \emptyset$ ; `// resultant quotient ubiquitous species set` $S^{(ub)}/\mathring{\sim} \subset 2^{S^{(ub)}}$

$R/\mathring{\sim} \leftarrow \emptyset$ ; `// resultant quotient reaction set` $R/\mathring{\sim} \subset 2^R$

$\mathring{\sim} \leftarrow \emptyset$ ; `// resultant equivalence operation` $\mathring{\sim} : S \cup R \rightarrow S/\mathring{\sim} \cup R/\mathring{\sim}$

`/* Generalize ubiquitous species */`

**for** $s^{(ub)} \in S^{(ub)}$ **do**

  $[s^{(ub)}]^{\mathring{\sim}} \leftarrow \{s^{(ub)}\}$ ; `// map` $s^{(ub)}$ `to its equivalence class`

**end**

$S^{(ub)}/\mathring{\sim} \leftarrow \{[s^{(ub)}]^{\mathring{\sim}} | s^{(ub)} \in S^{(ub)}\}$;

`/* Generalize specific species */`

**for** $s \in S \backslash S^{(ub)}$ **do**

  $[s]^{\mathring{\sim}} \leftarrow S \backslash S^{(ub)}$ ; `// map` $s$ `to its equivalence class`

**end**

$S/\mathring{\sim} \leftarrow S^{(ub)}/\mathring{\sim} \cup \{S \backslash S^{(ub)}\}$;

`/* Generalize reactions */`

`// map a reaction to its generalized version that operates with generalized species`

$gen \leftarrow \lambda r. \langle \{[s]^{\mathring{\sim}} | s \in reactants(r)\}, \{[s]^{\mathring{\sim}} | s \in products(r)\} \rangle$;

**for** $r \in R$ **do**

  $[r]^{\mathring{\sim}} \leftarrow \{\tilde{r} \in R | gen(\tilde{r}) = gen(r)\}$;

**end**

$R/\mathring{\sim} \leftarrow \{[r]^{\mathring{\sim}} | r \in R\}$;

**return** $\mathring{\sim}, \langle S/\mathring{\sim}, S^{(ub)}/\mathring{\sim}, R/\mathring{\sim} \rangle$

**Algorithm 1:** Compute$\mathring{\sim}$

The Compute$\overset{\circ}{\sim}$ algorithm forms the equivalence classes for ubiquitous and then specific species as in Definition 2 and then computes the generalized reactions.

## Step 2. Stoichiometry preserving property obedience

**Problem 2** *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R\rangle$ find an equivalence operation $\breve{\sim}$ that obeys restriction (1') and induces a quotient species set $S/\breve{\sim}$ of minimal size $\sharp S/\breve{\sim}$, such that $S/\breve{\sim}$ is a partition of the quotient species set $S/\sim$ induced by $\sim$, i.e., $\forall s \in S\ [s]^{\breve{\sim}} \subset [s]^{\sim}$.*

### Algorithm

We start with the given equivalence operation $\sim^0 = \sim$, and iteratively improve it, until the stoichiometry preserving property (1') is obeyed. We denote the equivalence operation obtained at the $i$-th iteration step as $\sim^i$.

At each iteration, if there exists a species equivalence class that violates the stoichiometry preserving property (1'), i.e.:

$$\exists s \neq \tilde{s} \in S, r \in R : s \in species(r) \wedge \tilde{s} \in species(r) \wedge [s]^{\sim^i} = [\tilde{s}]^{\sim^i},$$

we partition this species equivalence class $[s]^{\sim^i} = [\tilde{s}]^{\sim^i}$ into two: $[s]^{\sim^{i+1}} \vee [\tilde{s}]^{\sim^{i+1}} = [s]^{\sim^i} = [\tilde{s}]^{\sim^i}$ to form a new approximation $\sim^{i+1}$ of the equivalence operation. When no species equivalence class violating the restriction (1') can be found, the current equivalence operation is returned as result.

At each iteration one equivalence species class is partitioned. In the worst case, the equality operation $=$ (each species is equivalent only to itself) will be achieved. As it obeys restriction (1'), the process will terminate.

**Data**: $\sim$ - equivalence operation defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$,

$M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim \rangle$ - corresponding generalized model.

**Result**: $\breve{\sim}$ - equivalence operation described in Problem 3, $M/\breve{\sim} = \langle S/\breve{\sim}, S^{(ub)}/\breve{\sim} \subset S/\breve{\sim}, R/\breve{\sim} \rangle$ - corresponding generalized model.

$S/\breve{\sim} \leftarrow S/\sim$ ; `// resultant quotient species set` $S/\breve{\sim} \subset 2^S$

$S^{(ub)}/\breve{\sim} \leftarrow S^{(ub)}/\sim$ ; `// resultant quotient ubiquitous species set` $S^{(ub)}/\breve{\sim} \subset 2^{S^{(ub)}}$

$R/\breve{\sim} \leftarrow \emptyset$ ; `// resultant quotient reaction set` $R/\breve{\sim} \subset 2^R$

$\breve{\sim} \leftarrow \sim$ ; `// resultant equivalence operation` $\breve{\sim} : S \cup R \rightarrow S/\breve{\sim} \cup R/\breve{\sim}$

`/* Partition quotient species that do not obey restriction (1') */`

**for** $S^{(gen)} \in \{\tilde{S}^{(gen)} \in S/\breve{\sim} | \exists s \neq \tilde{s} \in \tilde{S}^{(gen)}, r \in R : s \in species(r) \wedge \tilde{s} \in species(r)\}$ **do**

  $\Pi = Partition(S^{(gen)})$;

  $S/\breve{\sim} \leftarrow \Pi \cup S/\breve{\sim} \backslash \{S^{(gen)}\}$ ; `// Update` $S/\breve{\sim}$

  **for** $\tilde{S}^{(gen)} \in \Pi$ **do**

    **for** $s \in \tilde{S}^{(gen)}$ **do**

      $[s]^{\breve{\sim}} \leftarrow \tilde{S}^{(gen)}$ ; `// Update` $\breve{\sim}$

    **end**

  **end**

**end**

`/* Generalize reactions */`

$gen \leftarrow \lambda r.\langle \{[s]^{\breve{\sim}} | s \in reactants(r)\}, \{[s]^{\breve{\sim}} | s \in products(r)\} \rangle$ `// map a reaction to its generalized version that operates with generalized species`

**for** $r \in R$ **do**

  $[r]^{\breve{\sim}} \leftarrow \{\tilde{r} \in R | gen(\tilde{r}) = gen(r)\}$;

**end**

$R/\breve{\sim} \leftarrow \{\breve{\sim}(r) | r \in R\}$;

**return** $\breve{\sim}, \langle S/\breve{\sim}, S^{(ub)}/\breve{\sim}, R/\breve{\sim} \rangle$

**Algorithm 2:** PreserveStoichiometry

**Species equivalence class partition**

In a species equivalence class that violates the restriction (1') there are usually only a few conflicts present, and multiple solutions of the partition problem exist.

**Species ontology**

In order to make the choice of the species equivalence classes biologically meaningful, we use an ontology that describes hierarchical *is_a* relationships (more specific to more general) between biochemical species.

**Definition 3** *A term $t$ is a* model term *if it corresponds to a specific species in the metabolic model.*

We assume that no two model terms are connected by a descendant-ancestor (more specific–more general) relationship in the ontology; otherwise, we mark the ancestor term ubiquitous:

$$\forall t, T \in terms\ (\exists\, species(t), species(T) \in S \wedge t \in descendants(T) \Rightarrow t = T).$$

We iteratively remove all the leaf terms that are not model terms from the ontology, so that all the model terms become leaves, and all the leaves become model terms.

For each species equivalence class that needs to be partitioned, we first find the least common ancestor $T$ of the ontological terms corresponding to its species. If the ontology allows for multiple inheritance, and there are several such least common ancestors, we pick the first one. Then we look among the $T$-th descendant terms for those that are compatible (to avoid multiple inheritance).

**Definition 4** *Terms $t_1, \ldots, t_k$ are* compatible *if and only if their descendant model terms do not intersect: $t_1, \ldots, t_k$ are compatible $\iff \forall i \neq j \in \{1, \ldots, k\}\ descendants(t_i) \cap descendants(t_k) \cap leaves(T) = \emptyset$.*

**Problem 3** *Given a term $T$, find a compatible term set among its descendants, such that it has minimal*

*size, covers all the $T$-th descendant leaf terms, and satisfies the stoichiometry preserving property (1"):*

$$
? \; t_1, \ldots, t_k \in descendants(T) : \; \wedge \quad
\begin{array}{l}
k = k_{min}, \\
t_1, \ldots, t_k \text{ are compatible}, \\
leaves(T) \subset descendants(t_1) \cup \ldots \cup descendants(t_k), \\
\forall i \neq j \in \{1, \ldots, k\} \; \forall t \in leaves(t_i), \tilde{t} \in leaves(t_j) \\
\quad \forall r \in R \; \{species(t), species(\tilde{t})\} \not\subset species(r).
\end{array}
\tag{1''}
$$

To do so, we first exclude all the terms that violate the stoichiometry preserving property (1"). We thus obtain an exact set cover problem.

**Problem 4 (Set cover)** *Given a set $X$ and a collection of its finite subsets $\Psi$, such that $\bigcup_{S \in \Psi} S = X$, find a minimum-size subset $\Pi \subset \Psi$ whose members cover all of $X$: $\bigcup_{S \in \Pi} S = \bigcup_{S \in \Psi} S = X$.*

**Remark 1** *Set cover is NP-complete (Karp, 1972).*

**Problem 5 (Exact set cover)** *As in* Set cover *problem, except that here the sets used in the cover are not allowed to intersect.*

**Remark 2** *Exact cover is NP-complete (Goldreich, 2008).*

**Exact set cover applied to ontological terms**

Each ontological term $t$ defines a set $S(t)$ of its descendant leaf terms (including $t$ if it is a leaf). The instance consists of a set $X$ of the model terms of interest, and a collection $\Psi$ of all sets defined by their common ancestor $T$, its descendant terms, and their relative complements with respect to $X$: $\forall S \in \Psi \; X \backslash S \in \Psi$, excluding all the sets that violate the stoichiometry preserving property (1"). We look for a minimum-size exact cover of $X$.

Note, that in this case an exact cover always exists, e.g. the one formed by all the leaf terms.

**Choice of the ontology**

We assume that any term that violates property (1") is removed from the ontology. Note that the term $T$ is also removed.

If the ontology has no multiple inheritance, i.e. $\forall S, \tilde{S} \in \Psi \; S \cap \tilde{S} \neq \emptyset \Rightarrow S \subseteq \tilde{S} \vee \tilde{S} \subseteq S$, the problem becomes trivial: the set of the root terms forms the solution. The size of the solution, though, depends on the characteristics of the ontology, e.g. for a completely flat ontology (i.e., with no relationships) the solution consists of singleton equivalence classes.

If multiple inheritance is allowed, any $\Psi \subseteq 2^X$ becomes possible, and the problem becomes *NP*-complete.

We use the ChEBI ontology (de Matos et al., 2010) of chemical compounds, as it is *de facto* a standard for species annotation in metabolic models. ChEBI consists of three main branches: *chemical entity*, *role*, and *subatomic particle*. The *chemical entity* branch describes terms useful for annotation of biochemical species in a metabolic model.

The level of detail in the ChEBI hierarchy is not uniform: some sub-branches are more developed than others, so equally precise terms may be placed unequally deep in the hierarchical tree. For example, both *hydrogen peroxide (CHEBI:16240)* and *decanoyl-CoA (CHEBI:28493)* terms describe precise chemical molecules; but *hydrogen peroxide* is only 5 terms away from the *chemical entity* in the ChEBI hierarchy, while *decanoyl-CoA* is 11 terms away.

Besides that, different types of classification are combined together in the hierarchical tree, leading to multiple inheritance. For example, in the *fatty-acid (CHEBI:35366)* sub-branch, several classification types are present, including:

- classification based on the length of the carbon chain:
  - *short-chain fatty acid (CHEBI:26666)*: 2-4 carbons;
  - *medium-chain fatty acid (CHEBI:59554)*: 6-12 carbons;
  - etc.
- classification based on the presence of double bonds in the carbon chain:

  - *saturated fatty acid (CHEBI:26607)*: no double bonds;
  - *unsaturated fatty acid (CHEBI:27208)*: one or more double bonds;
- classification based on substituent groups:
  - *hydroxy fatty acid (CHEBI:24654)*: one or more hydroxy substituents;
  - *oxo fatty acid (CHEBI:59644)*: at least one aldehydic or ketonic group;
  - etc.

Moreover, using only hierarchical relationships in the ChEBI ontology is not always enough. Examples show, that similar reactions can happen to the acid and the base in a conjugate acid-base pair. A conjugate acid-base pair is two species, one an acid and one a base, that differ from each other through the loss or gain of a proton (Stoker, 2012). For instance, in the Rhea database of chemical reactions (Alcántara et al., 2012), the *acyl-CoA oxidase (RHEA:28354)* reaction: *decanoyl-CoA+FAD+H+* $\rightarrow$ *trans-dec-2-enoyl-CoA+*$FADH_2$ is found for both *decanoyl-CoA (CHEBI:28493)* and its conjugate base *decanoyl-CoA(4-) (CHEBI:61430)*. But hierarchically these species are very far from each other in the ChEBI ontology: Their least common ancestor is *molecular entity (CHEBI:23367)*, a direct descendant of the root *chemical entity*. To establish a conjugate acid-base pair correspondence in the ChEBI ontology, not the hierarchical (*is_a*) but the special *is_conjugate_base_of* / *is_conjugate_acid_of* relationships are used. To maximize the chances of a conjugate acid-base pair being in the same quotient species set, we generalize the hierarchical relationship.

**Definition 5** *Term $t$ is a* generalized direct descendant/ancestor *of a term $T$ if and only if $t$ or a conjugate base or acid of $t$ is a direct descendant/ancestor of $T$ or of a conjugate base or acid of $T$.*

**Definition 6** *Term $t$ is a* generalized descendant/ancestor *of a term $T$ if and only if $t$ is a generalized direct descendant/ancestor of $T$ or of any generalized descendant/ancestor of $T$.*

We extend $\Psi$ so that it is closed under the operation of relative complement: $\forall S, \tilde{S} \in \Psi \; S \backslash \tilde{S} \in \Psi$. This allows for solving the set cover problem instead of the exact cover one: As $\Psi$ is closed under the operation of complement intersection, we can obtain an exact set cover $\tilde{C}$ from any set cover

$C = \{S_1, S_2, \ldots, S_m\}$ by replacing its elements with their relative complements with the previous elements of $C$: $\tilde{C} = \{S_1, S_2 \backslash S_1, \ldots, S_m \backslash \bigcup_{i=1}^{m-1} S_i\}$.

To approximate the solution of the set cover problem, we use a greedy algorithm.

**Greedy Algorithm**

Among the available subset candidates $S_i \in \Psi$, pick the one of the largest size and add it to the resulting set cover $\Pi$. Repeat this operation until all elements of $X$ are covered.

**Data**: $X$ - set of interest, $\Psi \subseteq 2^X$ - set of subsets of $X$

**Result**: $\Pi \subseteq \Psi$ - set cover of $X$

$\Pi \leftarrow \emptyset$ ; `// resultant cover`

**while** $X \neq \emptyset$ **do**

`// select` $S \in \Psi$ `that covers maximum elements of` $X$

$S^{(max)} \leftarrow max(\Psi, criterion = \lambda S.\sharp(S \cap X))$;

$\Psi \leftarrow \Psi \backslash \{S^{(max)}\}$;

$X \leftarrow X \backslash S^{(max)}$;

$\Pi \leftarrow \Pi \cup \{S^{(max)}\}$;

**end**

**return** $\Pi$

**Algorithm 3:** GreedySetCover

Greedy set cover is a polynomial time approximation algorithm that achieves an approximation ratio of $H(\sharp X)$, where $H(n)$ is the $n$-th harmonic number: $H(n) = \sum_{i=1}^{n} \frac{1}{i} \leq \ln n + 1$ (Chvatal, 1979). It is the best possible polynomial time approximation algorithm for set cover, under plausible complexity assumptions (Feige, 1998).

## Step 3. Species equivalence class number maximization

**Problem 6** *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R\rangle$, find an equivalence operation $\tilde{\sim}$ that obeys restriction (2) and does not change the reaction equivalence classes: $R/\sim = R/\tilde{\sim}$.*

### Algorithm

To satisfy restriction (2) we associate each species $s$ in the initial model with a pair of sets of reaction equivalence classes in the quotient reaction set $R/\sim$, induced by reactions where it participates as a reactant or product.

$s \to \langle R_s^{(rs)} = \{[r_1^{(rs)}]^\sim, \ldots, [r_o^{(rs)}]^\sim\}, R_s^{(ps)} = \{[r_1^{(ps)}]^\sim, \ldots, [r_t^{(ps)}]^\sim\}\rangle$.

**Definition 7** *Given an equivalence operation $\sim$ defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R\rangle$, we define an* equivalence operation $\tilde{\sim}$ *as forming a separate species equivalence class for each of the ubiquitous species, and putting $\sim$-equivalent specific species that intersect in their product or reactant reaction classes in the same equivalence class:*

$$\forall s^{(ub)} \in S^{(ub)}, s \in S \;\; s^{(ub)} \tilde{\sim} s \quad \Longleftrightarrow \quad s^{(ub)} = s,$$

$$\forall s, \tilde{s} \in S \backslash S^{(ub)} \;\; s \tilde{\sim} \tilde{s} \quad \Longleftrightarrow \quad \begin{array}{l} s \sim \tilde{s} \\ \wedge \\ (R_s^{(rs)} \cap R_{\tilde{s}}^{(rs)} \neq \emptyset) \vee (R_s^{(ps)} \cap R_{\tilde{s}}^{(ps)} \neq \emptyset) \vee (\exists \dot{s} \in S : \; s \tilde{\sim} \dot{s} \wedge \dot{s} \tilde{\sim} \tilde{s}). \end{array}$$

Any further partition of the quotient species set would imply the partition of the quotient reaction set. Hence the number of species equivalence classes is maximal for the current number of reaction equivalence

classes, and restriction (2) is satisfied.

**Data**: $\sim$ - equivalence operation defined on a metabolic model $M = \langle S, S^{(ub)} \subset S, R \rangle$,

$M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim \rangle$ - corresponding generalized model.

**Result**: $\tilde{\sim}$ - equivalence operation described in Problem 2, $M/\tilde{\sim} = \langle S/\tilde{\sim}, S^{(ub)}/\tilde{\sim} \subset S/\tilde{\sim}, R/\tilde{\sim} \rangle$ - corresponding generalized model.

$S/\tilde{\sim} \leftarrow \emptyset$ ; `// resultant quotient species set` $S/\tilde{\sim} \subset 2^S$

$S^{(ub)}/\tilde{\sim} \leftarrow S^{(ub)}/\sim$ ; `// resultant quotient ubiquitous species set` $S^{(ub)}/\tilde{\sim} \subset 2^{S^{(ub)}}$

$R/\tilde{\sim} \leftarrow R/\sim$ ; `// resultant quotient reaction set` $R/\tilde{\sim} \subset 2^R$

$\tilde{\sim} \leftarrow \sim$ ; `// resultant equivalence operation` $\tilde{\sim} : S \cup R \rightarrow S/\tilde{\sim} \cup R/\tilde{\sim}$

`/* Update specific species generalization */`

`// Map a species to a set of its` $\sim$`-equivalent species that participate in` $\sim$`-equivalent reactions`

$r\_sim \leftarrow \lambda s.\{\tilde{s} \sim s | \exists r, \tilde{r} \in R : s \in reactants(r) \wedge \tilde{s} \in reactants(\tilde{r}) \wedge r \sim \tilde{r}\}$;

$p\_sim \leftarrow \lambda s.\{\tilde{s} \sim s | \exists r, \tilde{r} \in R : s \in products(r) \wedge \tilde{s} \in products(\tilde{r}) \wedge r \sim \tilde{r}\}$;

$sim \leftarrow \lambda s.r\_sim(s) \cup p\_sim(s)$;

$S/\tilde{\sim} \leftarrow S^{(ub)}/\tilde{\sim} \cup \{sim(s) | s \in S \backslash S^{(ub)}\}$;

`// Merge all quotient species sets that intersect`

**while** $\exists S^{(gen)} \neq \tilde{S}^{(gen)} \in S/\tilde{\sim} : S^{(gen)} \cap \tilde{S}^{(gen)} \neq \emptyset$ **do**

  $S/\tilde{\sim} \leftarrow (S/\tilde{\sim} \backslash \{S^{(gen)}, \tilde{S}^{(gen)}\}) \cup \{S^{(gen)} \cup \tilde{S}^{(gen)}\}$;

**end**

**for** $S^{(gen)} \in S/\tilde{\sim}$ **do**

  **for** $s \in S^{(gen)}$ **do**

    $[s]^{\tilde{\sim}} \leftarrow S^{(gen)}$ ; `//` map $s$ `to its equivalence class`

  **end**

**end**

**return** $\tilde{\sim}, \langle S/\tilde{\sim}, S^{(ub)}/\tilde{\sim}, R/\tilde{\sim} \rangle$

**Algorithm 4:** Maximize

## Complete algorithm

**Data**: $M = \langle S, S^{(ub)} \subset S, R\rangle : \sharp S = n, \sharp S^{(ub)} = \breve{n} \leq n, \sharp R = m$ - metabolic model describing $n$ species, $\breve{n}$ among them being ubiquitous, and $m$ reactions.

**Result**: $\sim$ - approximation of the equivalence operation described in Problem 0, $M/\sim = \langle S/\sim, S^{(ub)}/\sim \subset S/\sim, R/\sim\rangle$ - corresponding generalized model.

$\mathring{\sim}, M/\mathring{\sim} \leftarrow Compute\mathring{\sim}(M)$;

$\breve{\sim}, M/\breve{\sim} \leftarrow PreserveStoichiometry(\mathring{\sim}, M/\mathring{\sim})$;

$\sim, M/\sim \leftarrow Maximize(\breve{\sim}, M/\breve{\sim})$;

**return** $\sim, M/\sim = \langle S/\sim, S^{(ub)}/\sim, R/\sim\rangle$

**Algorithm 5:** Compute$\sim$

## Applications

To illustrate the model generalization method we show its application to the genome-scale metabolic network of the lipid-accumulating yeast *Yarrowia lipolytica* (*MODEL1111190000* (Loira et al., 2012)). The generalized model is attached as additional files 1, 2. The generalization has created 100 non-trivial quotient species and 217 non-trivial quotient reactions, and reduced the total number of species from 1847 to 1072 and of reactions from 2002 to 893.

The generalization method shows the best performance if the model is well-annotated with ChEBI. The species lacking ChEBI annotations are forced to form trivial quotient species in the generalized model as there is no evidence of their biochemical similarity with any other species in ChEBI. For 430 species in the *Y. lipolytica* model no appropriate ChEBI annotation was found, thus they could not be grouped with other species.

*Peroxisome* is an example of a well-annotated compartment in the *Y. lipolytica* model: only two species have no ChEBI annotations: *YLR043C disulphide* and *TRX1*. The generalization process reduced the number of reactions in peroxisome from 65 to 27. Figures 2 and 3 represent the peroxisome before and after the generalization and were produced using Tulip graph visualisation tool (Auber, 2004).

The model before the grouping of equivalent reactions and species into generalized ones is shown on figure 2: different colors correspond to different equivalence classes. The same color code is used in figure 3 representing the generalized model that operates with quotient species and reactions. For example, the violet *unsaturated FA-CoA* node is a quotient of 8 species: *hexadec-2-enoyl-CoA*, *oleoyl-CoA*, *tetradecenoyl-CoA*, *trans-dec-2-enoyl-CoA*, *trans-dodec-2-enoyl-CoA*, *trans-hexacos-2-enoyl-CoA*, *trans-octadec-2-enoyl-CoA*, and *trans-tetradec-2-enoyl-CoA* (colored violet in figure 2). In a similar manner, the light-green *acCoA oxidase* quotient reaction, that converts *fatty acyl-CoA* (yellow) into *unsaturated FA-CoA* (violet), generalizes 6 corresponding light-green reactions of the initial model (figure 2).

The generalized model describes the *β-oxidation of fatty acids* pathway (Metzler, 2001) happening inside the *Y. lypolitica* peroxisome in a generic way: as a transformation of *fatty acyl-CoA* (yellow) into *unsaturated FA-CoA* (violet), then into *hydroxy FA-CoA* (green), *3-oxo FA-CoA* (magenta), and back to *fatty acyl-CoA*

(with a shorter carbon chain); while the initial model describes the same process in more details, specifying those reactions for each of the *fatty acyl-CoA* species present in the organisms' cell (e.g. *decanoyl-CoA*, *dodecanoyl-CoA*, etc.). That is why the *beta-oxidation* chain of the reactions in the initial model, transforming step-by-step the *fatty-acyl-CoA* with the longest carbon chain into the one with the shortest chain, in the generalized model appears as a cycle (generalizing all the *fatty-acyl-CoA*s into one species, regardless the chain-length).

The more precise model is needed for simulation, while the more general one is clearer to a human, and reveals the main properties of the model. For example, the generalized model highlights the fact that there is a particularity concerning *C24:0-CoA (tetracosanoyl-CoA)* (red, inside the cycle in figure 3): there exists a "short-cut" reaction, producing it directly from another *fatty acyl-CoA* (yellow), avoiding the usual four-reaction beta-oxidation chain, used for other *fatty acyl-CoA*s.

Another application of model generalization is metabolic model comparison. The generalization brings the models to the same level of abstraction and highlights the differences such as gaps. Examples can be found in (Zhukova and Sherman, 2013).

## Discussion

We have developed a method that provides a semantically zoomed-out view of a metabolic model, that keeps its essential structure but hides the details.

We have implemented our method as a Python program, that is available for download from http://metamogen.gforge.inria.fr. It takes an SBML model as an input, annotates its species with ChEBI terms (if the annotations are not present in the model) and generalizes it. It produces two SBML files, as an output. The first output file contains the generalized model. The second output file uses groups(Hucka, 2012) extension of SBML, and contains the initial model plus the groups representing quotient species and reaction sets.

We have applied our method to genome-scale metabolic models of yeasts. We have illustrated it here on the lipid-accumulating yeast *Y. lipolytica* and have shown that generalization helps finding gaps and peculiarities, as well as compresses the information stored in the model, which can be used for model visualisation and model comparison. In the example, the chain of *β-oxidation of fatty acids* reactions in the constitutive peroxisome of *Y. lipolytica* is generalized to a cycle of reactions, highlighting the alternative path for *C24:0-CoA (tetracosanoyl-CoA)*.

Currently the generalization method depends on the ChEBI ontology. It cannot generalize species that lack ChEBI annotations. In future work we will overcome this limitation.

The method zooms out a model to the most general level of abstraction that is consisted with the model structure, i.e. does not violate the restrictions (1') and (2). It remains to be seen whether there are intermediate levels of abstraction that can be useful for model analysis. In particular it may be interesting to define the maximal generalization for a group of organisms, in order to highlight the specific differences of the individual models with respect to a common generalization.

## Acknowledgements

The authors would like to thank Drs. Romain Bourqui and Antoine Lambert of the LaBRI MABioVis team for advice on graph layout.

AZ was supported by a CORDI-S doctoral fellowship from Inria.

## Author Disclosure Statement

No competing financial interests exist.

## References


Alcántara, R., Axelsen, K.B., Morgat, A., et al. 2012. Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Research* 40(Database issue), D754–60.

Auber, D. 2004. Tulip A Huge Graph Visualization Framework. *In* M. Jünger, P. Mutzel, G. Farin, H.C. Hege, D. Hoffman, C.R. Johnson, K. Polthier, and M. Rumpf, eds., *Graph Drawing Software*, Mathematics and Visualization, 105–126. Springer Berlin Heidelberg.

Caspi, R., Altman, T., Dreher, K., et al. 2012. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research* 40(Database issue), D742–53.

Chvatal, V. 1979. A Greedy Heuristic for the Set-Covering Problem. *Mathematics of Operations Research* 4(3), 233–235.

Clugston, M. and Flemming, R. 2000. *Advanced Chemistry (Advanced Science)*. OUP Oxford.

de Matos, P., Alcántara, R., Dekker, A., et al. 2010. Chemical Entities of Biological Interest: an update. *Nucleic Acids Research* 38(suppl 1), D249–D254.

Feige, U. 1998. A threshold of ln n for approximating set cover. *Journal of the ACM* 45(4), 634–652.

Goldreich, O. 2008. *Computational Complexity: A Conceptual Perspective*. Cambridge University Press, Cambridge.

Hucka, M. 2012. Groups Proposal. URL http://sbml.org/Community/Wiki/SBML_Level_3_Proposals/Groups_Proposal_Updated_%282012-06%29.

Kanehisa, M., Goto, S., Sato, Y., et al. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40(0305-1048 (Linking)), D109–14.

Karp, R.M. 1972. Reducibility Among Combinatorial Problems. *In* R.E. Miller and J.W. Thatcher, eds., *Complexity of Computer Computations*, 85–103. Plenum Press.

Loira, N., Dulermo, T., Nicaud, J.M., et al. 2012. A genome-scale metabolic model of the lipid-accumulating yeast Yarrowia lipolytica. *BMC Systems Biology* 6(1), 35.

Metzler, D.E. 2001. *Biochemistry: The Chemical Reactions of Living Cells*. No. v. 1 in Biochemistry: The Chemical Reactions of Living Cells. Elsevier Science.

Moodie, S., Le Novere, N., Demir, E., et al. 2011. Systems Biology Graphical Notation: Process Description language Level 1. *Nature Precedings* .

Muto, A., Kotera, M., Tokimatsu, T., et al. 2013. Modular Architecture of Metabolic Pathways Revealed by Conserved Sequences of Reactions. *Journal of chemical information and modeling* .

Rohn, H., Junker, A., Hartmann, A., et al. 2012. VANTED v2: a framework for systems biology applications. *BMC systems biology* 6(1), 139.

Stoker, H.S. 2012. *General, Organic, and Biological Chemistry*. Textbooks Available with Cengage YouBook Series. Brooks/Cole.

Swainston, N., Smallbone, K., Mendes, P., et al. 2011. The SuBliMinaL Toolbox: automating steps in the reconstruction of metabolic networks. *Journal of integrative bioinformatics* 8(2), 186.

Thiele, I. and Palsson, B.O. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols* 5(1), 93–121.

Tohsato, Y., Matsuda, H., and Hashimoto, A. 2000. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* 8, 376–83.

Zhukova, A. and Sherman, D.J. 2013. Knowledge-based generalization of metabolic networks: a practical study. Tech. rep., Inria. URL http://hal.inria.fr/hal-00906911.

# Additional Files

## Additional file 1 — The generalized SBML level 2 version 4 model of *Yarrowia lipolytica* (derived from *MODEL1111190000*)

The model contains generalized reactions and species.

## Additional file 2 — The SBML level 3 version 1 model of *Yarrowia lipolytica* (derived from *MODEL1111190000*) with groups extension representing equivalent species and reactions

The model contains all the elements (reactions, species, etc.) of the initial *MODEL1111190000* model, and is enriched with groups representing species and reaction equivalence information.

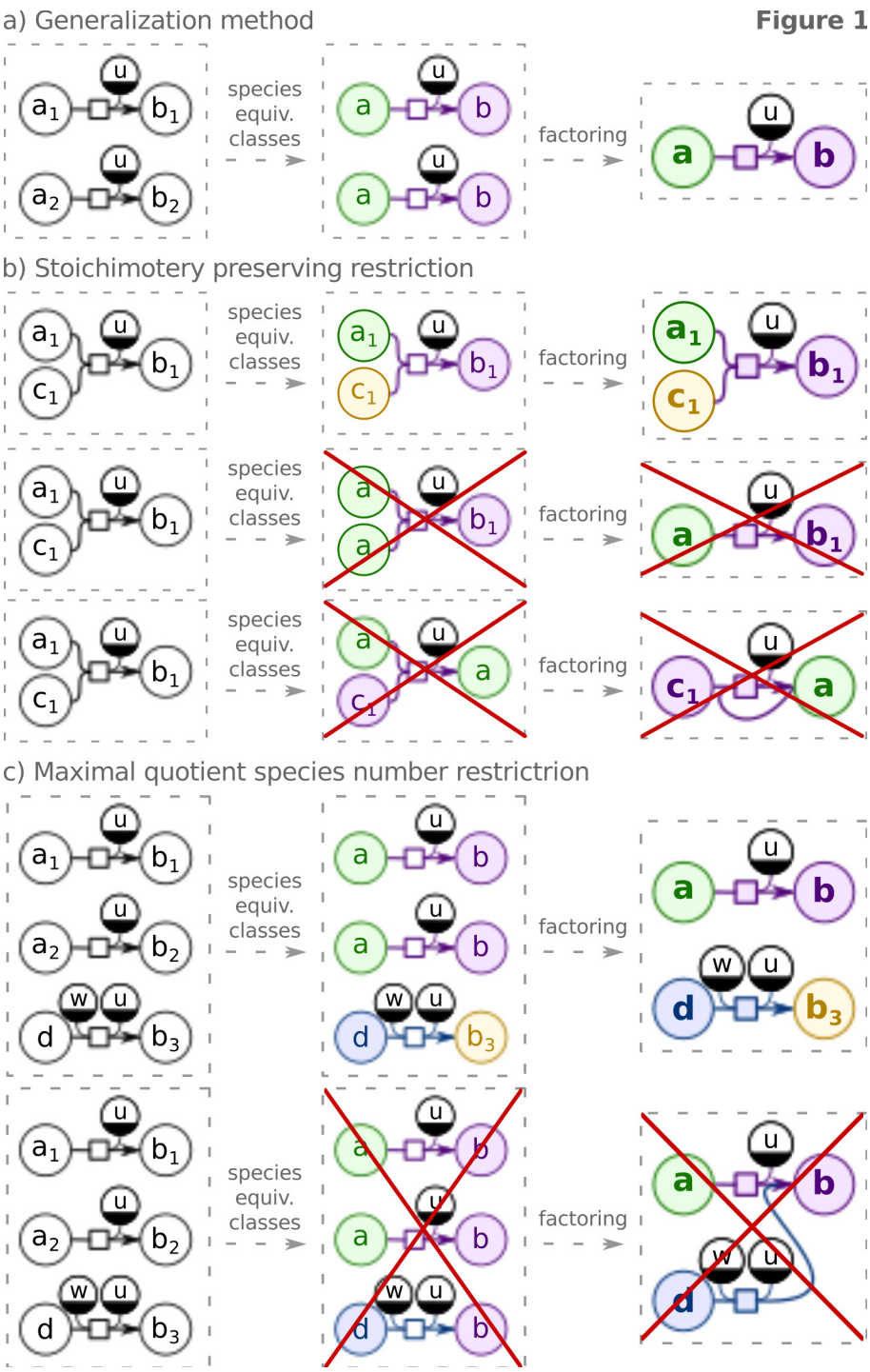


Figure 1: **Model generalization method and restrictions. a)** *Generalization* first groups the species into equivalence classes, and then factors them into generalized species. The reaction equivalence classes and factoring are inferred from the species classes. **b)** *Restriction (1').* The top part shows a correct generalization that obeys the restriction. Two bottom parts show generalizations that would change the reaction stoichiometry, and thus are not allowed. **c)** *Restriction (2).* The top part shows a correct generalization that obeys the restriction. The bottom part violates the restriction as there is no evidence in the model (i.e. no equivalent reaction) of the species $b_3$ belonging to the same equivalence class as $b_1$ and $b_2$.

Figure 2

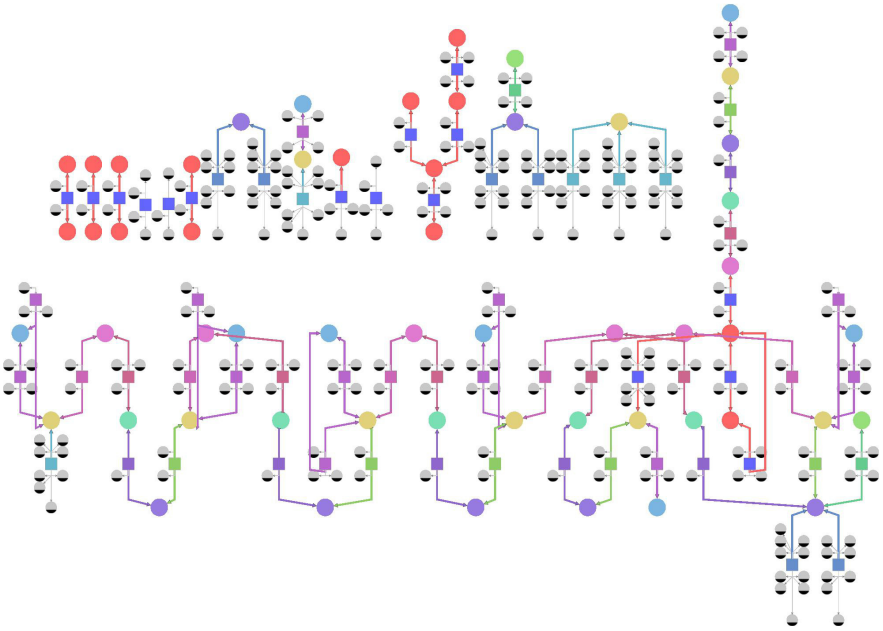

Figure 2: **Peroxisome of the *Y. lypolitica* model (*MODEL1111190000*).** Species are represented as circular nodes, and the reactions as square ones, connected by edges to their reactants/products, according to SBGN notation (Moodie et al., 2011). Ubiquitous species are of smaller size and colored gray. Specific species are divided into six non-trivial equivalence classes, and colored accordingly (violet, light-blue, yellow, green, light-green, magenta). The specific species that form trivial equivalence classes are all colored red. Reactions are divided into fifteen non-trivial equivalence classes, also represented by different colors. Reactions that form trivial equivalence classes are all colored blue. The size of the model does not allow for readable species labels, so they are omitted.

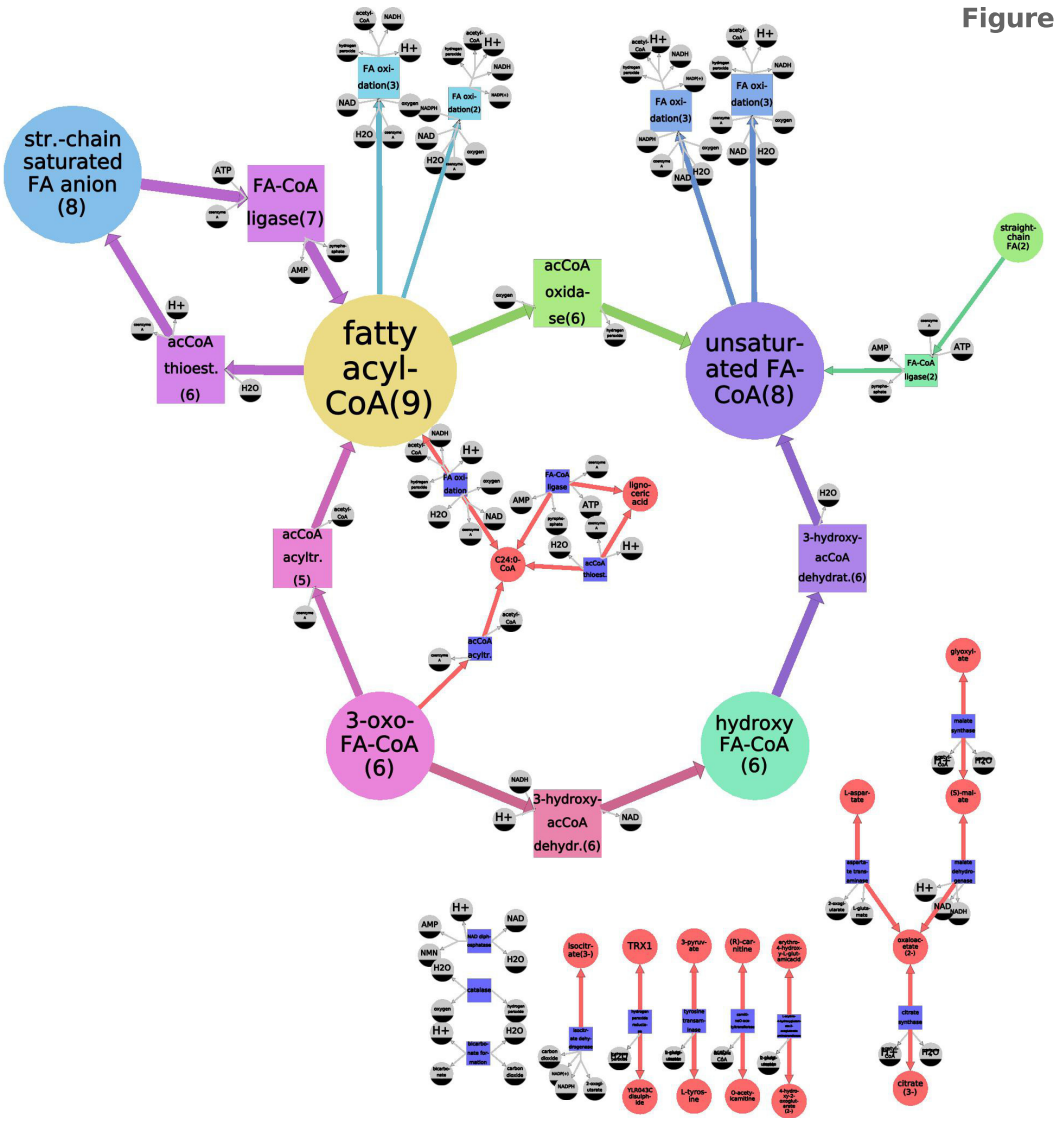


Figure 3: **The generalization of the peroxisome of the *Y. lypolitica* model** (see figure 2). The generalized model operates on quotient species and reactions. The number given in parentheses and the size of each node indicates how many entities it generalizes.