Structural conserved moiety splitting of a stoichiometric matrix

Ghaderi, S.; Haraldsdottir, H.S.; Ahookhosh, M.; Arreckx, S.; Fleming, R.M.T.

Journal of Theoretical Biology 499 (2020) 110276

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Structural conserved moiety splitting of a stoichiometric matrix

Susan Ghaderi[a], Hulda S. Haraldsdóttir[a], Masoud Ahookhosh[a,b], Sylvain Arreckx[a], Ronan M.T. Fleming[a,c,*]

[a] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 Avenue du Swing, Belvaux L-4362, Luxembourg
[b] Department of Electrical Engineering (ESAT-STADIUS)- KU Leuven, Kasteelpark Arenberg 10, Leuven 3001, Belgium
[c] Analytical Biosciences, Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Leiden, the Netherlands

ARTICLE INFO

Article history:
Received 12 November 2019
Revised 3 April 2020
Accepted 6 April 2020
Available online 23 April 2020

Keywords:
Reaction network
Stoichiometric matrix
Hypergraph
Conserved moiety
Moiety matrix splitting
Mathematical modelling

ABSTRACT

Characterising biochemical reaction network structure in mathematical terms enables the inference of functional biochemical consequences from network structure with existing mathematical techniques and spurs the development of new mathematics that exploits the peculiarities of biochemical network structure. The structure of a biochemical network may be specified by reaction stoichiometry, that is, the relative quantities of each molecule produced and consumed in each reaction of the network. A biochemical network may also be specified at a higher level of resolution in terms of the internal structure of each molecule and how molecular structures are transformed by each reaction in a network. The stoichiometry for a set of reactions can be compiled into a stoichiometric matrix $N \in \mathbb{Z}^{m \times n}$, where each row corresponds to a molecule and each column corresponds to a reaction. We demonstrate that a stoichiometric matrix may be split into the sum of $m - \text{rank}(N)$ moiety transition matrices, each of which corresponds to a subnetwork accessible to a structurally identifiable conserved moiety. The existence of this moiety matrix splitting is a property that distinguishes a stoichiometric matrix from an arbitrary rectangular matrix.

1. Introduction

Understanding biochemical networks is of great practical importance in systems biology. A variety of approaches for mathematical modelling of reaction networks have been developed, including topological (Barabási and Oltvai, 2004), stochastic, deterministic (Ingalls, 2013) and constraint-based modelling (Palsson, 2015). Before any biological application of any of these modelling approaches, an abstract representation of the relative quantities of molecules produced and consumed in each reaction of a reaction network is reconstructed from experimental literature. A key output of this reconstruction process is a *stoichiometric matrix,* where every row corresponds to a molecule, every column corresponds to a reaction, and each entry corresponds to the relative quantity of a molecule produced or consumed in a reaction. Typically, a stoichiometric matrix is the central mathematical object in any model of a reaction network for many biological, biotechnological and biomedical research applications. Therefore, characterising the mathematical properties of stoichiometric matrices is a fundamental problem in mathematical biology.

Although graph theory has been applied to the analysis of reaction networks (Klamt et al., 2009), thus far, this has required the application of approximations to underlying topology of the network. By labelling molecules as one type of vertex and reactions as another type of vertex it is possible to approximate biochemical network topology as a bipartite graph termed a *species-reaction graph* (Craciun and Feinberg, 2006). An appeal of this approximation is to facilitate the application of the extensive range of mathematical techniques that have arisen from the study of graphs. However, ultimately, the utility of the species-reaction graph concept is limited because the biochemical network of every living organism does contain hyperedges, so any representation as a single graph is an approximation. Furthermore, most hyperedges within a biochemical network consist of hyperedges between multisets, rather than sets, further limiting the range of established hypergraph theory techniques that could be applied to biochemical networks.

* Corresponding author at: Analytical Biosciences, Division of Systems Biomedicine and Pharmacology, Leiden Academic Centre for Drug Research, Leiden University, Einsteinweg 55, 2333 Leiden, the Netherlands.
E-mail addresses: ronan.mt.fleming@gmail.com, ronan.mt.fleming@nuigalway.ie (R.M.T. Fleming).

https://doi.org/10.1016/j.jtbi.2020.110276
0950-0618/

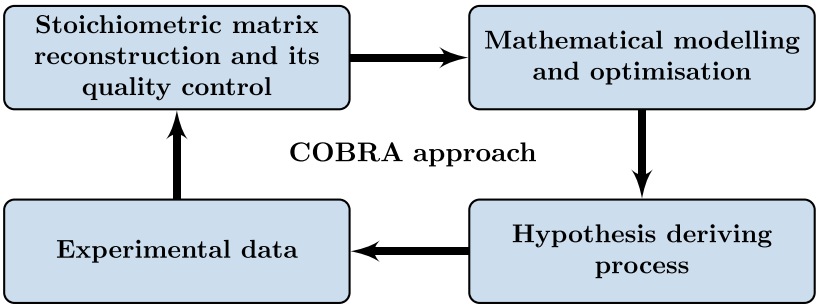


Fig. 1. Constraint-based reconstruction and analysis. Constraint-Based Reconstruction and Analysis (COBRA) is an example of a systems biology approach, carried out in an iterative cycle, where the aim is to increase the predictive accuracy of a constraint-based computational model. Quality controlled reconstruction of prior literature information generates a draft model, which includes a stoichiometric matrix (upper left). This is followed by mathematical modelling using optimisation methodologies (upper right), enabling hypothesis generation in the form of model predictions (lower right). These predictions are testing against experimental data (lower left) and any discrepancy is used to refinement the reconstruction. This iterative cycle is repeated until a desired accuracy is reached.

There is a pressing need for contributions from the graph and hypergraph theory community to establish connections between the form of hypergraph observed in applications to (bio)chemical reaction networks.

Among all modelling approaches for reaction networks, a particular emphasis of Constraint-Based Reconstruction and Analysis (COBRA, Fig. 1) (Palsson, 2015) is reconstruction and modelling of biochemical networks at genome-scale. Such models contain the majority of the known reactions in an organism, within a scope based on considerations of the application domain, and give rise to stoichiometric matrices with a large number of rows and columns. Almost every biochemical constraint-based modelling problem is posed as an optimisation problem involving a stoichiometric matrix (Palsson, 2015), and thus obtaining solutions to high-dimensional optimisation problems is essential to generate model predictions. This emphasis on optimisation has lead to an increasing interest in biochemical constraint-based modelling from the mathematical and numerical optimisation community (Ma et al., 2017).

A stoichiometric matrix may be distinguished from an arbitrary rectangular matrix by mathematical properties arising from its biochemical origins. This matrix is may be fully specified by the known biochemistry of an organism, cell, organelle or biochemical subsystem being modelled. The last universal common ancestor from which all organisms now living on Earth have common descent is hypothesised to have lived over three billion years ago. The complete biochemical network of this organism, and every descendant thereof, is not known. Therefore, we do not yet, and may never have, a complete mathematical classification that specifies the subset of rectangular matrices to which every stoichiometric matrix belongs. What we do have is a certainty that this class is restricted by the physicochemical and biological principles that govern all living systems. The main purpose of this paper is to emphasise certain special mathematical properties of stoichiometric matrices that arise from physicochemical principles.

To date, much of the focus has been on characterisation of mathematical properties shared by stoichiometric matrices and arbitrary rectangular matrices, e.g., (Papin et al., 2004). However, certain mathematical properties are known to distinguish a stoichiometric matrix from an arbitrary rectangular matrix. In chemistry, a moiety is a subunit of a molecule and conserved moiety is one that is invariant with respect to a defined set of chemical transformations. Clarke (1988) proposed that each basis vector for the left nullspace of a stoichiometric matrix corresponds to an independent conserved moiety. Famili and Palsson (2003) computed a convex basis, of extreme rays that may be linearly dependent, for the left null space and classified (conserved) moieties according to their relationship with cofactors and the boundary of the system. However, establishing a correspondence between each extreme ray and the structure of a moiety was not automatic. Householder QR factorisation (Vallabhajosyula et al., 2006) and sparse LU factorisation (Gill et al., 1987) are efficient methods for computing for basis vectors for the left nullspace of a large stoichiometric matrix but it is challengeing to interpret a linearly independent basis vector in terms of chemistry if it contains negative entries.

Plasson et al. (2008) defined a *reacton* (conserved moiety), as a subpart of a molecule that is never broken into smaller parts by any of the reactions composing the network. Based on this definition, it was proposed that a chemical reaction network be interpreted as simple recombinations of reactons, where each reaction could be represented by partial reactions, each one describing the transfer of reactons from one compound to another. Furthermore, examples were given of splitting a stoichiometric matrix into a sum of incidence matrices, each representing a directed graph of reacton transfers. Various approaches, none efficient at genome-scale, were considered to compute a non-negative basis for the left nullspace of a stoichiometric matrix, each of which was then manually identified with a reacton. Haraldsdóttir and Fleming (2016) defined a conserved moiety as a group of atoms that remains intact in all reactions of a network. They then showed that the structure of each conserved moiety and the corresponding non-negative left nullspace basis vector, could be efficiently identified at genome-scale by graph theoretical analysis of an atom transition graph, which required atom mappings for each reaction (Rahman et al., 2016). It is needed to clearly specify, in graph and hypergraph theoretical terms, the mathematical relationship between atom transition graphs, chemical reaction hypergraphs and conserved moieties. Furthermore, it is necessary to investigate the properties that this relationship endows on a stoichiometric matrix that distinguish it from an arbitrary rectangular matrix.

This paper has three objectives. The first objective, in Section 2, is to briefly introduces some basic concepts from graph and hypergraph theory (Voloshin, 2009). The second objective, in Sections 3–5, is to introduce the established concepts of a molecule, reaction and network, respectively, in terms of graph and hypergraph theory. This introduction is given at a high level in terms of a hypergraph where each vertex is a molecule, and each edge is a reaction, and also at a lower level in terms of graphs where each vertex is an atom embedded in a molecular structure and each edge is a bijection between atoms in separate molecules. The third objective, in Sections 6 and 7, is to introduce the concept of a conserved moiety in terms of graph and hypergraph theory, split a stoichiometric

matrix for a network into the sum of a set of subnetwork incidence matrices, each of which is an incidence matrix for a moiety subnetwork, then relate this to the mathematical properties of a stoichiometric matrix.

Notation

Throughout this paper, $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{m\times n}$ denote the field of real numbers, the vector space of n-tuples of real numbers, and the space of $m \times n$ matrices with entries in $\mathbb{R}$, respectively. Similarly, $\mathbb{Z}$, $\mathbb{Z}^n$, $\mathbb{Z}^{m\times n}$ stand for integer numbers, the vector space of n-tuples of integer number, and the space of matrices with entries in $\mathbb{Z}$, respectively. N^T denotes the transpose of a matrix N in $\mathbb{R}^{m\times n}$. $\mathbb{R}^n_+$ and $\mathbb{R}^n_{++}$ display non-negative real n-tuples and positive real n-tuples in $\mathbb{R}^n$, respectively, and $\mathbb{Z}^n_+$ and $\mathbb{Z}^n_{++}$ display non-negative integer n-tuples and positive integer n-tuples in $\mathbb{Z}^n$, respectively. Let $\mathbb{1}$ be the vector of all ones. For a matrix $A \in \mathbb{R}^{m\times n}$, A_i and $A_{:j}$ denote the ith row and the jth column of A, respectively, where $i \in 1,\dots,m$ and $j \in 1,\dots,n$. The exponential or natural logarithm of a vector is meant component-wise and $\exp(\log(0)) := 0$. Further, $[\ \cdot\ ,\ \cdot\ ]$ stands for the horizontal concatenation operator, and I denotes an identity matrix.

A calligraphic, uppercase, roman letter, e.g., $\mathcal{A}$, denotes a set, multiset or sequence, with $\{\ \cdot\ ,\ \cdot\ \}$ denoting an unordered pair, $(\ \cdot\ ,\ \cdot\ )$ denoting an ordered pair and $(\cdot,\dots,\cdot)$ denoting a sequence. Let $|\mathcal{A}|$ denote the cardinality of the set $\mathcal{A}$. A multiset is a modification of the concept of a set that, unlike a set, allows for multiple instances for each of its elements. In a multiset $\mathcal{M} := (\mathcal{A}, f)$, $\mathcal{A}$ is a set and $f : \mathcal{A} \to \mathbb{Z}_+$ is a function from $\mathcal{A}$ to the set of positive integers giving the multiplicity of the ith element $\mathcal{A}_i$ in the multiset as the number $f(\mathcal{A}_i)$. In multiset $\{a, a, b\}$, the element a has multiplicity 2, and b has multiplicity 1. The cardinality of a multiset is constructed by summing up the multiplicities of all its elements. The cardinality of sets, multisets and sequences is all assumed to be finite.

In illustrative examples, all metabolic species and reactions are annotated with their abbreviated identifier used in the Virtual Metabolic Human database (http://vmh.life), e.g., the *crn* abbreviation for the metabolite L-carnitine (crn).

2. Graph and hypergraph theory

There exist various excellent introductory textbooks on graph theory, e.g., Wilson (2020) and hypergraph theory, e.g., Voloshin (2009). Nevertheless, for completeness we introduce key terms in graph and hypergraph theory next. A *graph* $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is a mathematical object which consists of a set of *vertices* $\mathcal{V}$ and a set of *edges* $\mathcal{E}$, where $\mathcal{V} := \{\mathcal{V}_1,\dots,\mathcal{V}_m\}$ and $\mathcal{E} := \{\mathcal{E}_1,\dots,\mathcal{E}_n\}$. An edge $\mathcal{E}_j := \{\mathcal{V}_i, \mathcal{V}_k\} \in \mathcal{E}$ is an unordered pair of vertices $\mathcal{V}_i \in \mathcal{V}$ and $\mathcal{V}_k \in \mathcal{V}$, whence $\mathcal{V}_i$ and $\mathcal{V}_k$ are said to be *adjacent*. A directed edge $\mathcal{E}_j := (\mathcal{V}_i, \mathcal{V}_k) \in \mathcal{E}$ is an ordered pair of vertices $\mathcal{V}_i \in \mathcal{V}$ and $\mathcal{V}_k \in \mathcal{V}$, whence $\mathcal{E}_j$ is said to join the *head* vertex $\mathcal{V}_i$ to the *tail* vertex $\mathcal{V}_k$. An orientation of an undirected edge is an assignment of a direction to that edge, turning it into a directed edge. An inverted edge swaps the order of a pair of vertices in a directed edge. A subgraph $\mathcal{G}'$ of a graph $\mathcal{G}$ is a graph whose vertex set and edge set are subsets of those of $\mathcal{G}$.

A graph can be represented by an incidence matrix $B \in \mathbb{Z}^{m\times n}$, where each row corresponds to a vertex, each column corresponds to an edge and the entries are given by

$$B_{ij} := \begin{cases} -1 & \text{if} \quad \mathcal{V}_i \in \textit{tail}, \\ 1 & \text{if} \quad \mathcal{V}_i \in \textit{head}, \\ 0 & \text{otherwise}, \end{cases}$$

or by its adjacency matrix $A \in \mathbb{Z}^{m\times m}$ given by

$$A_{ij} := \begin{cases} 1 & \text{if} \quad \mathcal{V}_i \textit{ is adjacent to } \mathcal{V}_j, \\ 0 & \text{otherwise}, \end{cases}$$

where $i = 1,\dots,m$ and $j = 1,\dots,n$. An incidence matrix $B \in \mathbb{R}^{m\times n}$ is said to be *conserved* if the summation of each column of B vanishes, that is

$$\mathbb{1}^T B =: 0_n.$$

A labelled graph is a graph that associates each vertex with one of a set of vertex labels and associates each edge with one of a set of edge labels. A vertex-labelled graph is a graph that associates each vertex with one of a set of vertex labels. An edge-labelled graph is a graph that associates each edge with one of a set of edge labels. An *isomorphism* between two graphs $\mathcal{G}_1(\mathcal{V}^1, \mathcal{E}^1)$ and $\mathcal{G}_2(\mathcal{V}^2, \mathcal{E}^2)$ is a bijection $\psi : \mathcal{V}^1 \to \mathcal{V}^2$ and $\theta : \mathcal{E}^1 \to \mathcal{E}^2$. If the graphs are labelled, an isomorphism also preserves labelling. A set of graphs isomorphic to each other is called an *isomorphism class* of graphs. A *path* is a finite sequence of edges which connect a sequence of vertices. A pair of vertices is *connected* if there exists a path between them. A *component* of a graph is a subgraph with a path between any two of its vertices and without a path to any vertex in the remainder of the supergraph. A vertex with no incident edges is itself a component.

A *hypergraph* $\mathcal{H}(\mathcal{V}, \mathcal{S})$ is a generalisation of a graph in which the jth hyperedge $\mathcal{S}_j := \{\mathcal{A}_j, \mathcal{B}_j\} \in \mathcal{S}$ is a pair of multisets of vertices $\mathcal{A}_j \subset \mathcal{V}$ and $\mathcal{B}_j \subset \mathcal{V}$. A directed *hypergraph* $\mathcal{H}(\mathcal{V}, \mathcal{S})$ is a generalisation of a directed graph in which the jth directed hyperedge $\mathcal{S}_j := (\mathcal{F}_j, \mathcal{R}_j) \in \mathcal{S}$ is an ordered pair of subsets of vertices, where $\mathcal{F}_j \subset \mathcal{V}$ and $\mathcal{R}_j \subset \mathcal{V}$ denote subsets of vertices corresponding to the tail and head of the jth hyperedge. A *network* is either a graph or a hypergraph.

3. Molecules

Strictly speaking, a molecule is an electrically neutral group of two or more atoms held together by chemical bonds. However, henceforth, for the sake of simplicity, we stretch this definition to also encompass an electrically charged molecule (ion) and a molecule with one atom. This is akin allowing a single isolated vertex to be defined as a graph. A molecule may be represented at multiple levels of abstraction. First, Section 3.1 introduces a molecule at a high level of abstraction, where each molecule is only represented by a chemical formula. Then, Section 3.2 introduces a molecule at a low level of abstraction in terms of its topological structure.

3.1. Molecules

A high level abstract representation of a molecule is to associate it a unique label.

Definition 1. A*molecule* is a singular instance of a distinct chemical. A set of m molecules is denoted with $\mathcal{V} := \{\mathcal{V}_1,\dots,\mathcal{V}_m\}$, where $\mathcal{V}_i$ is the label associated with the ith molecule.

Unless otherwise specified, a molecule is assumed to mean a biochemical, that is, a chemical that is found in a biological system. A molecule could be a protein, a carbohydrate, an ion, a water molecule, or any other singular instance of a chemical found in a living being.

Definition 2. A*compartment*, is a distinct, finite, contiguous subdivision of the three-dimensional space of a biochemical system that is demarcated by a boundary that selectively permeable to certain molecules.

All biochemical systems occupy at least one compartment (Lane and Pariseau, 2016), and often multiple hierarchically embedded compartments. For instance, a eukaryotic cell consists of several compartments such as mitochondria, cytosol, nucleus and endoplasmic reticulum. A selectively permeable boundary prevents the diffusive exchange of certain molecules across the boundary of a compartment.

Definition 3. A*molecular species* is a finite set of identical molecules, labelled with a single compartment.

Unless otherwise specified, a molecule is assumed to mean biomolecule. Two molecules in separate compartments, that are otherwise identical, are still considered distinct species. Compartmentalisation is denoted with a bracketed suffix to the abbreviated species label, e.g., *crn*[*c*] and *crn*[*m*] are the labels for the molecule L-carnitine (crn) in the cytosolic [*c*] and mitochondrial [*m*] compartments, respectively.

3.2. Molecular graphs

Although there is a rich literature on the representation of chemistry in terms of graphs (Trinajstić, 1992), we only introduce some basic concepts in chemical graph theory here as our focus is on the mathematical structure of stoichiometric matrices, rather than the structure of individual molecules. Each molecule consists of a set of atoms. Each atom consists of a nucleus, with subatomic entities termed protons and neutrons, surrounded by electrons. Protons have positive electrical charge, neutrons have neutral charge and electrons have negative charge. We assume that biological systems conserve atomic nuclear structure, but they can change the number of electrons associated with an atomic nucleus, therefore each molecule is assigned a net electrical charge.

Definition 4. An *atom* is a singular instance of a chemical element.

Unless otherwise specified, an atom is assumed to mean an atom of an element that is found in a biological system. Of the ~ 118 known chemical elements only ~27 are known to be incorporated into biochemical systems.

Definition 5. A *molecular formula,* is the natural number of atoms of each element in a molecule.

For example, the molecular formula of a citrate molecule with charge -3 (cit) is $C_6H_5O_7$. That is, it consists of 6 carbon atoms (C), 5 hydrogen atoms (H) and 7 oxygen atoms (O). The mass of a molecule is given by the sum of the strictly positive masses of each of its constituent atoms. The (mono-isotopic) molecular mass of a citric acid molecule with charge -3 is 192.0270026 Da.

Definition 6. Given a molecule $\mathcal{V}_k$, its*atomic cardinalityn*$(\mathcal{V}_k)$ is sum of the number of atoms, irrespective of element label, in that molecule. Given a set of molecules $\mathcal{V}$ its atomic cardinality is the sum of the cardinality of each molecule, that is

$$n(\mathcal{V}) = \sum_{k=1}^{|\mathcal{V}|} n(\mathcal{V}_k).$$

For example, citrate has atomic cardinality 18, while the molecular formula of L-carnitine is $C_7H_{15}NO_3$ and therefore its atomic cardinality is 26, therefore the atomic cardinality of the set $\mathcal{A} = \{\text{citrate}, \text{L}-\text{carnitine}\}$ is 44.

Definition 7. A *chemical bond* is a singular instance of a pair of atoms.

In chemical terminology, a chemical bond is a lasting attraction between two atoms.

Definition 8. Given a set of molecules $\mathcal{V}$, the *molecular graph* of molecule $\mathcal{V}_k$ is a graph $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{V}_k)$ where each vertex $\mathcal{X}_i$ is an atom and each edge $\mathcal{Y}_j$ is a chemical bond in a molecule. A molecular graph represents the complete set of $|\mathcal{X}|$ atoms and $|\mathcal{Y}|$ bonds in a molecule as a single connected component. Each vertex is triply labelled, with (i) an element label, which is a type of chemical element, (ii) a molecular label, which uniquely identifies the molecule, and (iii) an atomic label $i \in 1 \dots n(\mathcal{V})$, which uniquely identifies each of the $n(\mathcal{V})$ atoms in $\mathcal{V}$. Each edge is labelled with a type of chemical bond.

In chemistry, a molecule must have at least one bond between two atoms. However, for the sake of consistency with graph theory, a chemical entity that consists of a single atom and no bond is also referred to as a molecule, as it corresponds to a graph with one vertex and no edge. Certain chemical assumptions are used to define the conditions for two molecules to be considered identical or distinct. These assumptions arise from topological and geometric considerations as to the structure of a molecule. However, with respect to the structural representation of a molecule considered here, it is to necessary and sufficient to consider that two molecules are of the same molecule if and only if both molecules are labelled with the same compartment and their corresponding molecular graphs are isomorphic.

3.2.1. Example molecule and molecular graph

.

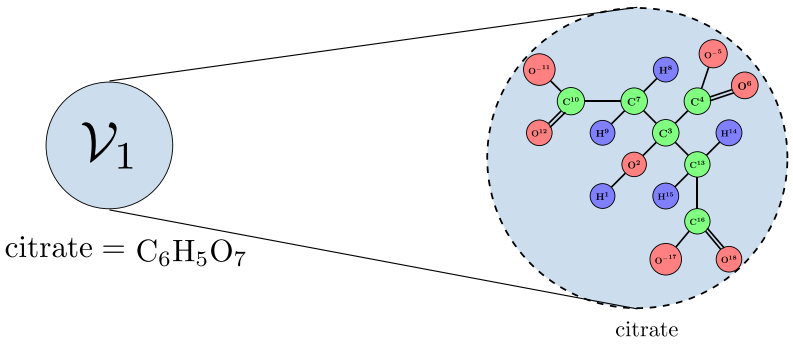


Fig. 2. A molecule of citrate represented as a vertex and a molecular structure. Citrate represented as a node with its molecular formula (left) and represented as a molecular graph (right). The three types of element are oxygen (red), carbon (green) and hydrogen (blue). The two types of bond are illustrated, single (–) and double (=). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Reactions

A reaction is a process that leads to the chemical transformation of one set of molecular entities to another. Unless otherwise specified, a reaction is assumed to be a biochemical reaction, that is, a reaction that is found in a biological system. This excludes reactions that involve changes to nuclear structure, e.g., nuclear fusion. A reaction may be represented at multiple levels of abstraction. First, Section 4.1 introduces a reaction at a high level of abstraction, in terms of molecules and reaction stoichiometry. Then Section 4.2 introduces a reaction at a low level of abstraction in terms of molecular structures and atom mappings.

4.1. Reaction stoichiometry

At a high level of abstraction, a reaction may be represented by a reaction equation, described below, which only specifies the quantities associated with each molecule involved and whether they are consumed or produced in the reaction. The concept of a hyperedge, as a pair of vertex subsets, is well established. However, before we mathematically define a reaction, we must generalise the concept of a vertex subset, to allow a natural number weight on each vertex in a subset. This permits a generalisation of the concept of a hyperedge, where each involved vertex is associated with a natural number weight.

Definition 9. A*chemical complex* $\mathcal{C}(\mathcal{V})$ is a subset or multiset of molecules, drawn from a set of molecules $\mathcal{V}$. A *stoichiometric number* is the multiplicity of molecules of a molecular species in a complex.

The term stoichiometry is derived from the ancient Greek origins of *stoicheion* meaning element and *metron* meaning measure. The cardinality of a chemical complex $|\mathcal{C}|$ is the sum of the multiplicities of each of its constituent molecule.

Definition 10. A *reaction* is hyperedge $\mathcal{H} := \{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\}$, formed from a pair of chemical complexes $\mathcal{P}(\mathcal{V})$ and $\mathcal{Q}(\mathcal{V})$, where $\mathcal{P} \neq \mathcal{Q}$ and $\mathcal{V}$ is a set of molecules.

The set of molecules $\mathcal{V}$ may be the same for both chemical complexes in a hyperedge, but in that case their multiplicity must differ. In a chemical complex, the entities may be distinct or identical, that is corresponding to distinct species, or a single species, respectively. If there is no molecule of a species in a complex, then the stoichiometric number is trivially zero. If a reaction involves a molecule with a multiplicity greater than one, then this is represented by multiple instances of the same molecule, rather than a single molecular species, as is often the approach taken in stoichiometric modelling (Palsson, 2015). We are interested in the relationship between mathematical modelling of a biochemical network at stoichiometric and atomic levels of resolution and atom mappings are between molecules, rather than molecular species, so we also represent reaction stoichiometry in terms of molecules.

In chemistry, thermodynamics dictates an to the complexes in reaction, leading to a directed reaction (hyperedge). In graph theory, an orientation of an (undirected) graph is an assignment of a direction to each edge, turning a graph into a directed graph.

Definition 11. A *directed reaction* is a directed hyperedge $\mathcal{Y} := (\mathcal{F}(\mathcal{V}), \mathcal{R}(\mathcal{V}))$, formed from an ordered pair of complexes, where $\mathcal{F}$ is the tail complex and $\mathcal{R}$ is the head complex.

By the principle of microscopic reversibility (Lewis, 1925), each reaction is reversible, therefore when representing a real reaction we always have a pair of symmetric directed hyperedges.

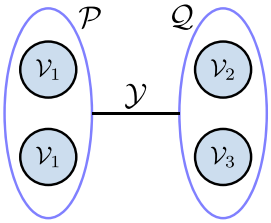


Fig. 3. An example reaction. A reaction $\mathcal{H} := \{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\}$, where the complexes are the multiset $\mathcal{P} = \{\mathcal{V}_1, \mathcal{V}_1\}$ and the set $\mathcal{Q} = \{\mathcal{V}_2, \mathcal{V}_3\}$, while the set of vertices is $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3\}$.

Definition 12. A reaction*equation* is of the form

$$\sum_{i\in\mathcal{F}} \mathcal{V}_i \rightleftharpoons \sum_{k\in\mathcal{R}} \mathcal{V}_k,$$

or more precisely in mathematical terms

$$\bigcup_{i\in\mathcal{F}} \mathcal{V}_i \equiv \bigcup_{k\in\mathcal{R}} \mathcal{V}_k.$$

In an undirected reaction, $\mathcal{V}_i$ denotes the ith molecule in complex $\mathcal{F}$ and $\mathcal{V}_k$ the kth molecule in the complex $\mathcal{R}$, whereas for a directed reaction, $\mathcal{F}$ and $\mathcal{R}$ are referred to as the tail and head complexes, respectively.

In a reaction equation the symbol $\rightleftharpoons$ signifies an equivalence relation. This is consistent with the chemistry literature. Once an orientation is chosen, it is conventional to write a directed reaction equation with the tail complex (substrate complex) to the left and the head complex (product complex) to the right. The use of a union symbol is mathematically correct, but the use of a summation symbol is far more commonly observed in the chemistry literature.

4.1.1. Example reactions

Consider the reaction $\mathcal{H} := \{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\}$, with reaction equation

$$2\mathcal{V}_1 \rightleftharpoons \mathcal{V}_2 + \mathcal{V}_3. \tag{1}$$

illustrated in Fig. 3. The complexes are the multiset $\mathcal{P} = \{\mathcal{V}_1, \mathcal{V}_1\}$, and the set $\mathcal{Q} = \{\mathcal{V}_2, \mathcal{V}_3\}$, and the set of vertices is $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3\}$. In complex $\mathcal{Q}$, the stoichiometric number is 2 for $\mathcal{V}_1$, and in complex $\mathcal{Q}$ the stoichiometric number is 1 for $\mathcal{V}_2$ and 1 for $\mathcal{V}_3$.

Fig. 4 provides a toy biochemical example, consisting of a cell with one sub-cellular compartment, several molecules and one directed reaction whose equation is

$$cit[m] \rightleftharpoons h2o[m] + cisa[m]. \tag{2}$$

Each molecule corresponds to a vertex, and the set of vertices is $\{cit[m],\ h2o[m], cisa[m]\}$. The forward complex is the tail set of vertices $\mathcal{F} := \{cit[m]\}$ and reverse complex is the head set of vertices $\mathcal{R} := \{h2o[m],\ cisa[m]\}$. This reaction (citrate hydro-lyase, link) takes place in the mitochondrial compartment, hence the $[m]$ suffix, and transforms the molecule citrate (cit) into the molecule water (h2o) and the molecule cis-aconitic acid (cisa).

4.2. Atom mappings

At a low level of abstraction, a reaction can be represented as a mapping between pairs of atoms, where one atom is in a substrate complex and another atom is in a product complex.

Definition 13. Given a set of molecules $\mathcal{V}$ and a chemical complex $\mathcal{C}(\mathcal{V})$, a *complex graph* $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{C}(\mathcal{V}))$ is the disjoint union of a multiset of $|\mathcal{C}|$ molecular graphs, where each molecular graph corresponds to one molecule $\mathcal{V}_k \in \mathcal{C}$. Let $n(\mathcal{V}_k)$ denote the atomic cardinality molecule of molecule $\mathcal{V}_k$, then the total number of vertices

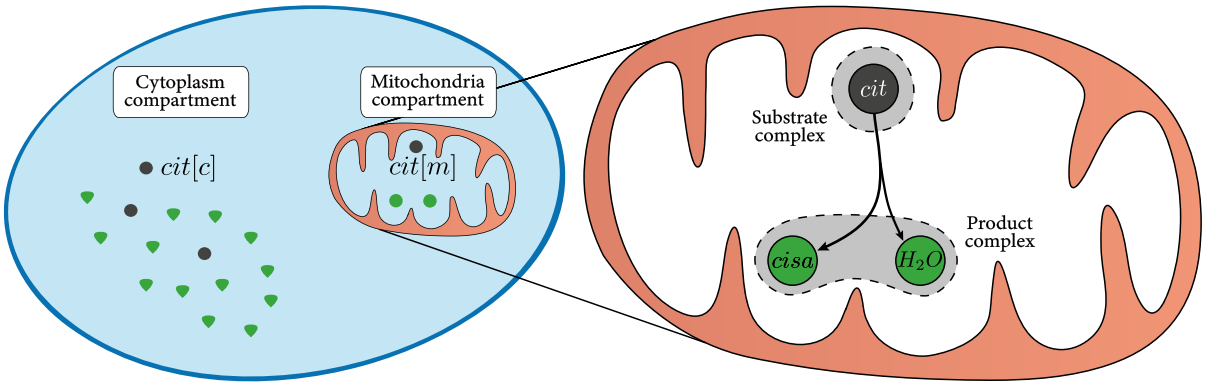


Fig. 4. Molecules, molecular species, complexes, a reaction and compartments. An illustration of a faux cell (left) with two compartments. The mitochondrial compartment [*m*] is embedded in the larger cytoplasmic compartment [*c*]. Molecules of citrate (*cit*, gray dots) in the cytoplasm are denoted *cit*[*c*]. Molecules of citrate in the mitochondria are denoted *cit*[*m*] which is considered distinct from *cit*[*c*]. To the right is an enlarged view of the mitochondrial compartment with a single reaction. The substrate complex is consumed in the reaction and consists of a single citrate molecule and the product complex is produced in the reaction and consists of one cis-aconitate molecule (*cisa*) and one water molecule (*h2o*).

in complex graph $\mathcal{C}$ is

$$|\mathcal{X}| = \sum_{\mathcal{V}_k \in \mathcal{C}} n(\mathcal{V}_k).$$

Each vertex is triply labelled with (i) an element label, (ii) a molecular label, and (iii) an atomic label.

The number of connected components of a complex graph is equal to the number of molecules in that complex. For example, a complex graph will contain two connected components that are isomorphic up to vertex labelling, if the complex consists of two identical molecules, that is a molecular species with stoichiometric number (multiplicity) two.

Definition 14. Given a reaction $\mathcal{H} := \{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\}$, an *atom transition* is a labelled edge $\mathcal{E} := \{\mathcal{X}_i, \mathcal{X}_j\}$ that joins vertex $\mathcal{X}_i$ of molecule $\mathcal{V}_k$ in complex graph $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{P})$ with vertex $\mathcal{X}_j$ of molecule $\mathcal{V}_l$ in complex graph $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{Q})$. The edge is labelled with a reaction label, which uniquely identifies a reaction. Both vertices must have the same element label, but the molecular and atomic labels may be different.

The element label of the vertex $\mathcal{X}_i \in \mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{P})$ is the same as the element label of the vertex $\mathcal{X}_k \in \mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{Q})$. That is, an atom transition is an edge between a pair of atoms of the same element, one in each of the pair of complexes involved in a reaction. Therefore, in a reaction, the total number of atoms of each element in both complexes is the same. For example, in reaction (1) the molecular formula of citrate is $C_6H_5O_7$ while the molecular formula of water is H_2O and the molecular formula of cis-aconitic acid is $C_6H_3O_6$. The element specific sum of atoms in the latter two molecules is $C_6H_3O_7$, which is the same as the molecular formula for citrate. The atomic label of both vertices is generally not the same, because typically reactions involve transformation of one set of molecules into another set of molecules and, within a given set of molecules, atomic labels are unique, by definition.

Definition 15. Given a set of molecules $\mathcal{V}$ and a reaction $\mathcal{H} := \{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\}$, an *atom mapping* is a graph $\mathcal{G}(\mathcal{X}, \mathcal{Y}, \mathcal{H}\{\mathcal{P}(\mathcal{V}), \mathcal{Q}(\mathcal{V})\})$ formed by the disjoint union of the set of

$$|\mathcal{Y}| := \sum_{\mathcal{V}_k \in \mathcal{P}} n(\mathcal{V}_k) = \sum_{\mathcal{V}_k \in \mathcal{Q}} n(\mathcal{V}_k)$$

atom transitions, between

$$|\mathcal{X}| := \sum_{\mathcal{V}_k \in \mathcal{P}} n(\mathcal{V}_k) + \sum_{\mathcal{V}_k \in \mathcal{Q}} n(\mathcal{V}_k) = 2|\mathcal{Y}|$$

vertices. Each edge is labelled with an identical reaction label. Each vertex is labelled with an element label, a molecular label and an atomic label.

Note that an atom mapping consists of $|\mathcal{Y}|$ connected components, each of which contains one edge and two vertices with identical element labels. That is, all edges of the molecular graphs of each molecule in $\mathcal{V}$ are omitted. One reaction may correspond to multiple alternate atom mappings, e.g., if a molecular structure has a symmetrical subgraph, this may permit multiple alternate atom mappings that are equivalent with respect to element vertex labelling, but not with respect to atomic vertex labelling.

4.2.1. Example complex graph and atom mapping

Fig. 5 illustrates an atom mapping for the citrate hydro-lyase reaction (link).

5. Networks

A biochemical network consists of a set of molecules that are chemically transformed into one another by a set of reactions. A biochemical network may be represented at multiple levels of abstraction. First, Section 4.1 introduces a biochemical network at a high level of abstraction, in terms of molecules and reaction stoichiometry. Then Section 4.2 introduces a biochemical network at a low level of abstraction in terms of molecular graphs and atom mappings.

5.1. Stoichiometric hypergraphs

A stoichiometric hypergraph is a network of reactions expressed in terms of molecules and reaction stoichiometry.

Definition 16. *A stoichiometric hypergraph* is a hypergraph $\mathcal{H}(\mathcal{V}, \mathcal{Y}\{\mathcal{F}, \mathcal{R}\})$ that consists of a set of m vertices $\mathcal{V} := \{\mathcal{V}_1, \ldots, \mathcal{V}_m\}$, each corresponding to one molecule, and a set of n hyperedges $\mathcal{Y} := \{\mathcal{Y}_1, \ldots, \mathcal{Y}_n\}$, each corresponding to one reaction. The jth hyperedge $\mathcal{Y}_j\{\mathcal{F}_j, \mathcal{R}_j\}$ is composed of pair of complexes $\mathcal{F}_j(\mathcal{V})$ and $\mathcal{R}_j(\mathcal{V})$ where $\mathcal{F}_j \neq \mathcal{R}_j$.

Note that in this definition of a stoichiometric network, each vertex corresponds to a molecule, which is a singular instance of a distinct chemical, rather than a molecular species, which is a finite set of identical molecules. If one replaces each reaction with a symmetric pair of directed hyperedges then a stoichiometric network can also be represented by a directed hypergraph.

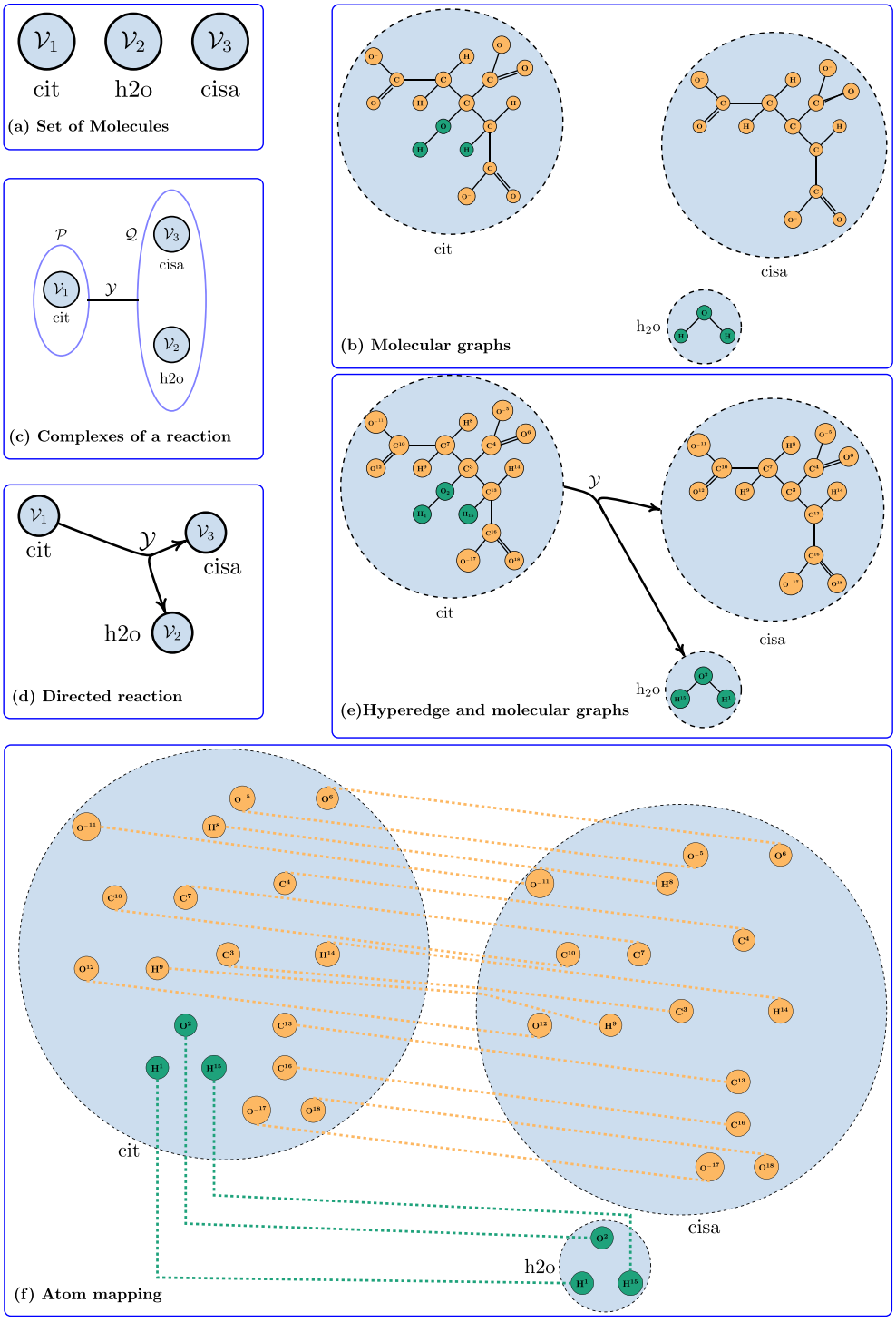


Fig. 5. The citrate hydro-lyase reaction. The chemical conversion of citrate into water and cis-aconitic acid represented as a hyperedge (reaction) and a set of edges (atom mapping). The citrate hydro-lyase reaction (link) involves three molecules (a), each of which may be represented by a molecular graph (b), where each vertex is an atom and each edge is a chemical bond between atoms. The reaction is between two complexes (c), one complex $\mathcal{P}$ consisting of citrate (left, cit) and one complex $\mathcal{Q}$ consisting of water (middle, h2o) and cis-aconitic acid (right, cisa). A reaction may be considered as a hyperedge $\mathcal{Y}$, where each vertex is defined by a molecular graph (e). Each atom in complex $\mathcal{P}$ (citrate) corresponds to one atom in complex $\mathcal{Q}$ (water and cis-aconitic acid) and together they constitute the atom mapping (corresponding atoms are individually labelled with numerical superscripts and connected by dotted lines) that represents the hyperedge $\mathcal{Y}$ as a set of disconnected edges.

Definition 17. *A directed stoichiometric hypergraph* $\mathcal{H}(\mathcal{V}, \mathcal{Y}(\mathcal{F}, \mathcal{R}))$ is an oriented stoichiometric hypergraph, that consists of a set of m vertices $\mathcal{V} := \{\mathcal{V}_1, \ldots, \mathcal{V}_m\}$, and sequence of n directed hyperedges $\mathcal{Y} := (\mathcal{Y}_1, \ldots, \mathcal{Y}_n)$. In the jth reaction $\mathcal{Y}_j := (\mathcal{F}_j, \mathcal{R}_j)$ the tail complex is

$$\mathcal{F}_j := \sum_{i=1}^{m} F_{i,j}\mathcal{V}_i$$

and the head complex is

$$\mathcal{R}_j := \sum_{i=1}^{m} R_{i,j}\mathcal{V}_i$$

where $F \in \mathbb{Z}_+^{m\times n}$ is a *forward stoichiometric matrix*, $R \in \mathbb{Z}_+^{m\times n}$ is a *reverse stoichiometric matrix*, with $\mathcal{F}$ and $\mathcal{R}$ being two sequences of cardinality n.

The entry $F_{i,j}$ is the stoichiometric number of molecule i consumed in the jth directed reaction and the entry $R_{i,j}$ is the stoichiometric number of molecule i produced in the jth directed reaction. If the ith molecule is neither produced, nor consumed in the jth directed reaction, then $F_{i,j} = R_{i,j} = 0$. If $F_{j,j} = R_{i,j} > 0$ then the ith molecule is termed a *catalyst* of the jth directed reaction as it is chemically invariant with respect to that chemical transformation.

Let us know introduce the main mathematical object that is the main focus of attention in this paper.

Conjecture 18. *Given a directed stoichiometric hypergraph* $\mathcal{H}(\mathcal{V}, \mathcal{Y}(\mathcal{F}, \mathcal{R}))$ *with m molecules and n reactions, its* stoichiometric matrix$N \in \mathbb{Z}^{m\times n}$ *is*

$$N := R - F$$

where $F_{i,j}$ *and* $R_{i,j}$ *are the stoichiometric numbers of the ith molecule consumed and produced in the jth directed reaction, respectively. A* stoichiometric coefficient$N_{i,j}$ *is a signed stoichiometric number, with a negative or positive sign if a molecule is consumed or produced in a directed reaction, respectively.*

If and only if the ith molecule is a catalyst in the jth directed reaction then $N_{i,j} = 0$ yet $F_{j,j} = R_{i,j} > 0$. If the ith molecule does not participate in the jth directed reaction then $N_{i,j} = F_{j,j} = R_{i,j} = 0$. Therefore, N can be defined in terms of F and R while the opposite is not the case. We have introduced a stoichiometric matrix with a conjecture, rather than a definition as the construction of a complete mathematical definition of a stoichiometric matrix is an open problem. Note that in this definition of a stoichiometric hypergraph, each vertex corresponds to a molecule, which is a singular instance of a distinct chemical, rather than a molecular species, which is a finite set of identical molecules. Therefore, a stoichiometric matrix is a sign matrix, i.e., $N \in \{-1, 0, 1\}^{m\times n}$, while forward and reverse stoichiometric matrices are binary matrices F, $R \in \{0, 1\}^{m\times n}$.

Certain key topological features of a stoichiometric hypergraph can be discerned from its stoichiometric matrix (Palsson, 2015). Given a stoichiometric matrix $N \in \mathbb{R}^{m\times n}$, its *zero pattern* $\widetilde{N}\{0, 1\}^{m\times n}$ is the binary matrix obtained by replacing each non-zero entry of N by 1. The number of non-zero entries in each column, $\widetilde{N}^T\mathbb{1}$, gives the *molecular cardinality* for each reaction. The number of non-zero entries in each row, $\widetilde{N}\mathbb{1}$, gives the *reaction cardinality* for each molecule. The*molecularadjacency matrix* is given by $B := \widetilde{N}\widetilde{N}^T$. Each diagonal element of the molecule adjacency matrix gives the reaction cardinality of a molecule, and each off-diagonal element gives the number of reactions in which two molecules participate together. The *reactionadjacency matrix* is given by $A := \widetilde{N}^T\widetilde{N}$. Each diagonal element of the reaction adjacency matrix gives the number of molecules that participate in a reaction while each off-diagonal element gives the number of molecules shared by two reactions. Therefore, the vectors $\widetilde{N}^T\mathbb{1}$ and $\widetilde{N}\mathbb{1}$ and matrices $A := \widetilde{N}^T\widetilde{N}$ and $B := \widetilde{N}\widetilde{N}^T$ can provide us valuable information about the sparsity pattern of stoichiometric matrices. Such issues become especially important in practical applications involving numerical computing with high dimensional stoichiometric matrices.

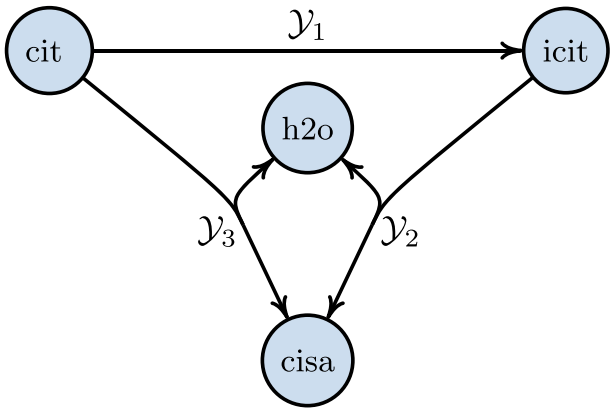


Fig. 6. A directed stoichiometric hypergraph. The four molecules (vertices) are citrate (cit, $C_6H_5O_7$), isocitrate (icit, $C_6H_5O_7$), cis-aconitic acid (cisa, $C_6H_3O_6$) and water (h2o, H_2O). In biochemical terms, the reactions (black hyperedges) are $\mathcal{Y}_1$: aconitate hydratase (ACONTm), $\mathcal{Y}_2$: citrate hydro-lyase (link) and $\mathcal{Y}_3$: isocitrate hydro-lyase (link). Although each reaction is, in principle, reversible, the directions of each hyperedge are given in the conventional orientation, consistent with the corresponding stoichiometric matrix.

5.1.1. A directed stoichiometric hypergraph with three reactions

Consider a directed stoichiometric hypergraph with 4 molecules $\mathcal{V} = (cit, \text{icit}, \text{cisa}, \text{h2o})$ and 3 reactions $\mathcal{Y} = (\mathcal{Y}_1, \mathcal{Y}_2, \mathcal{Y}_3)$, a planar representation of which is illustrated in Fig. 6. The 3 reaction equations are

$$\begin{aligned} \mathcal{Y}_1 &: \ \text{cit} \rightleftharpoons \text{icit}, \\ \mathcal{Y}_2 &: \ \text{cit} \rightleftharpoons \text{h2o} + \text{cisa}, \\ \mathcal{Y}_3 &: \ icit \rightleftharpoons \text{h2o} + \text{cisa}. \end{aligned} \tag{3}$$

$\mathcal{Y}_2$ is the citrate hydro-lyase reaction introduced in Section 4.1. The forward and reverse stoichiometric matrix of this network are

$$F := \begin{array}{c} \\ \text{h2o} \\ cit \\ icit \\ \text{cisa} \end{array} \begin{array}{c} \begin{array}{ccc} \mathcal{Y}_1 & \mathcal{Y}_2 & \mathcal{Y}_3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{array} \qquad R := \begin{array}{c} \\ \text{h2o} \\ cit \\ icit \\ \text{cisa} \end{array} \begin{array}{c} \begin{array}{ccc} \mathcal{Y}_1 & \mathcal{Y}_2 & \mathcal{Y}_3 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array}$$

In these matrices, each row is labelled with the corresponding molecule and each column is labelled with the corresponding reaction. The net stoichiometric matrix of the mitochondrial subnetwork is

$$N := \begin{array}{c} \begin{matrix} \mathcal{Y}_1 & \mathcal{Y}_2 & \mathcal{Y}_3 \end{matrix} \\ \begin{bmatrix} 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} \text{h2o} \\ cit \\ icit \\ \text{cisa} \end{matrix}$$

where again rows and columns correspond to molecules and reactions, respectively. The molecular adjacency matrix is

$$NN^T = \begin{array}{c} \begin{matrix} \text{h2o} & cit & icit & \text{cisa} \end{matrix} \\ \begin{bmatrix} 2 & -1 & -1 & 2 \\ -1 & 2 & -1 & -1 \\ -1 & -1 & 2 & -1 \\ 2 & -1 & -1 & 2 \end{bmatrix} \end{array} \begin{matrix} \text{h2o} \\ cit \\ icit \\ \text{cisa} \end{matrix}$$

while the reaction adjacency matrix is

$$N^T N = \begin{array}{c} \begin{matrix} \mathcal{Y}_1 & \mathcal{Y}_2 & \mathcal{Y}_3 \end{matrix} \\ \begin{bmatrix} 2 & 1 & -1 \\ 1 & 3 & 2 \\ -1 & 2 & 3 \end{bmatrix} \end{array} \begin{matrix} \mathcal{Y}_1 \\ \mathcal{Y}_2 \\ \mathcal{Y}_3 \end{matrix}$$

5.1.2. A stoichiometric hypergraph of human metabolism

Metabolism refers to the set of reactions necessary to sustain the life of a single organism. Metabolism extracts energy and material precursors from food, and uses them to synthesises the macromolecules, e.g., proteins, that make up an organism. Metabolism also degrades macromolecules and eliminates waste. A stoichiometric hypergraph of metabolism is a reaction network representing metabolism where the molecules are metabolites (low molecular mass organic chemicals) and the reactions are metabolic reactions (Fig. 7).

The latest comprehensive reconstruction of human metabolism, Recon3D (Brunk et al., 2018), accounts for 17% of the functionally annotated genes in the human genome, and consists of 5,835 rows (molecular species) and 10,600 columns (reactions) in 9 compartments. The 9 compartments are extracellular [e], cytosol [c], mitochondria [m], mitochondrial intermembrane space [i], endoplasmic reticulum [r], lysosome [l], peroxisome [x], golgi apparatus [g], and nucleus [n]. Certain key topological features of the human metabolic network can be discerned from analysis of its corresponding stoichiometric matrix. The sparsity pattern for the stoichiometric matrix of the Recon3D reconstruction is illustrated in Fig. 8. Reaction cardinality can vary widely depending on the molecular species concerned, with some molecular species participating in many reactions and others at least two, but perhaps only two reactions. For all genome-scale metabolic networks known there is an approximately linear relationship between the logarithm of reaction cardinality and the rank ordered reaction cardinality. That is, reaction cardinality approximates a *power law distribution* (Palsson, 2015). Figure 9 illustrates the molecular and reaction cardinality of Recon3D. Fig. 10 illustrates the molecular and reaction adjacency matrices of Recon3D.

5.2. Atom transition graphs

An atom transition graph is a representation of a reaction network in terms of atoms and atom mappings.

Definition 19. Given a set of molecules $\mathcal{V}$ and a stoichiometric hypergraph $\mathcal{H}(\mathcal{X}, \mathcal{Y}\{\mathcal{F}(\mathcal{V}), \mathcal{R}(\mathcal{V})\})$, an *atom transition graph* is a graph $\mathcal{G}(\mathcal{X}, \mathcal{E}, \mathcal{H})$ formed by uniting a set of $|\mathcal{Y}|$ atom mappings, each of which corresponds to a reaction. The union merges vertices of atom mappings that have identical elemental and atomic labels. Each of the $p := |\mathcal{X}|$ vertices corresponds to an atom of an element in one of the $m := |\mathcal{V}|$ molecules. Each of the $q := |\mathcal{E}|$ edges corresponds to an atom transition in an atom mapping corresponding to one of the $n := |\mathcal{Y}|$ reactions. Each vertex is labelled with elemental, molecular and atomic labels, while each edge is labelled with a reaction label.

In a molecular graph, each vertex is triply labelled, with (i) an element label, which is a type of chemical element, (ii) a molecular label, which uniquely identifies the molecule, and (iii) an atomic label $i \in 1 \ldots n(\mathcal{V})$, which uniquely identifies each of the $n(\mathcal{V})$ atoms in $\mathcal{V}$. If a pair of reactions share at least one molecule in common, then they share vertices from the same molecular graph, therefore the corresponding pair of atom mappings can be united in an atom transition graph, by merging vertices with identical elemental and atomic labelling, but possibly different molecular labels. In an atom transition graph, each edge is labelled with a reaction label.

Definition 20. A *directed atom transition graph* $\mathcal{G}(\mathcal{V}, \mathcal{E}, \mathcal{H})$ is an oriented atom transition graph, with $p := |\mathcal{X}|$ vertices and $q := |\mathcal{E}|$ directed edges, with topology represented by the incidence matrix $A \in \{-1, 0, 1\}^{p \times q}$.

5.2.1. An example of an atom transition graph

We now provide an example of an atom transition graph corresponding to the 3 reaction biochemical network introduced in Section 5.1.1. This atom transition graph is formed by uniting identical vertices from an atom mapping for $\mathcal{Y}_2$, which is the citrate hydro-lyase reaction illustrated in Fig. 5, and with identical vertices from atom mappings for $\mathcal{Y}_1$ and $\mathcal{Y}_3$. The atom transition graph is illustrated in Fig. 11.

6. Moieties

Each connected component of an atom transition graph corresponds to a set of atoms that have identical elemental labels, but may have different molecular and atomic labels. Each path in a connected component of an atom transition graph corresponds to the trajectory that a single instance of an atom could take, via a sequence of atom transitions, each of which corresponds to a reaction. It is of interest to group connected components that are the same throughout an atom transition network, because they identify conserved molecular substructures, as defined below. Herein we assume a time invariant representation of a reaction network at atomic resolution, that is, every chemical transformation corresponding to a reaction has occurred sufficiently that an atom in every position of every molecule of a substrate complex has been mapped to every chemically feasible position of every molecule in a product complex.

6.1. Conserved moieties

Definition 21. A *conserved moiety* is a set of atoms, where each atom belongs to one connected component of an isomorphism class of connected components of an atom transition graph $\mathcal{G}(\mathcal{X}, \mathcal{E}, \mathcal{H})$(Haraldsdóttir and Fleming, 2016). The isomorphism class is of maximum cardinality and the isomorphism is label preserving with respect to molecular labelling of vertices and reaction labelling of edges.

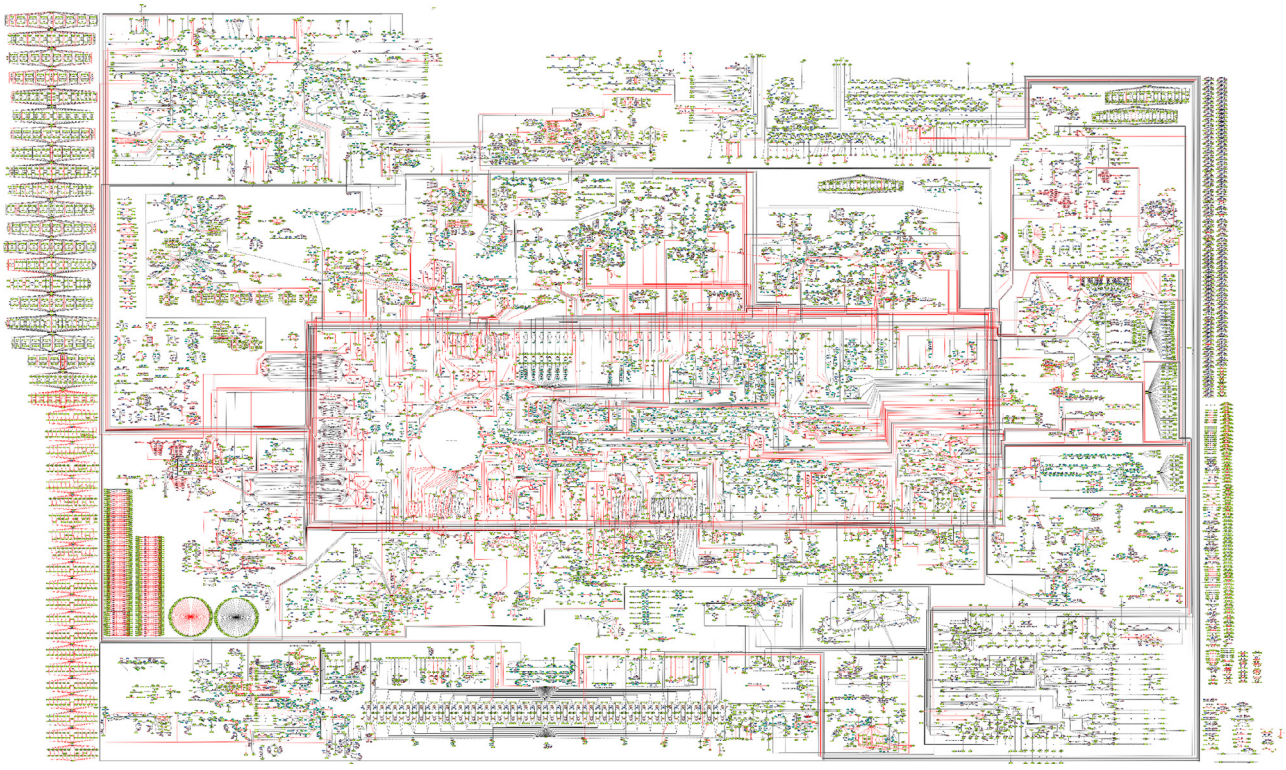


Fig. 7. A planar visualisation of a stoichiometric hypergraph human metabolism. A planar visualisation of the stoichiometric hypergraph of Recon3D (Brunk et al., 2018), termed Recon3Map (Noronha et al., 2017), which was manually drawn using the network layout editor CellDesigner (version 4.4) (Funahashi et al., 2008). To avoid excessive crossing of hyperedges, certain molecules that are involved in many reactions have been duplicated at different positions in the network.

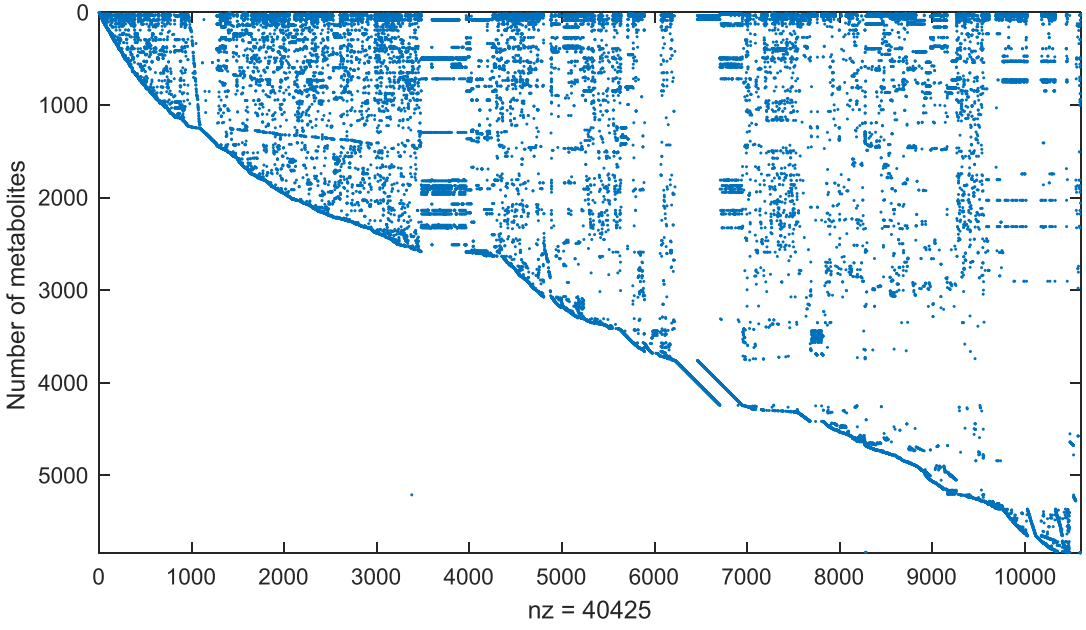
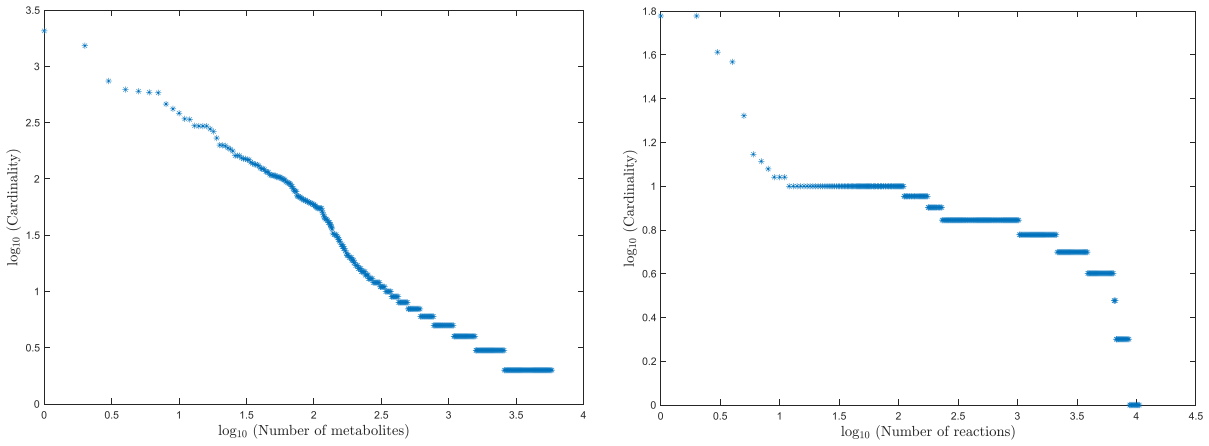


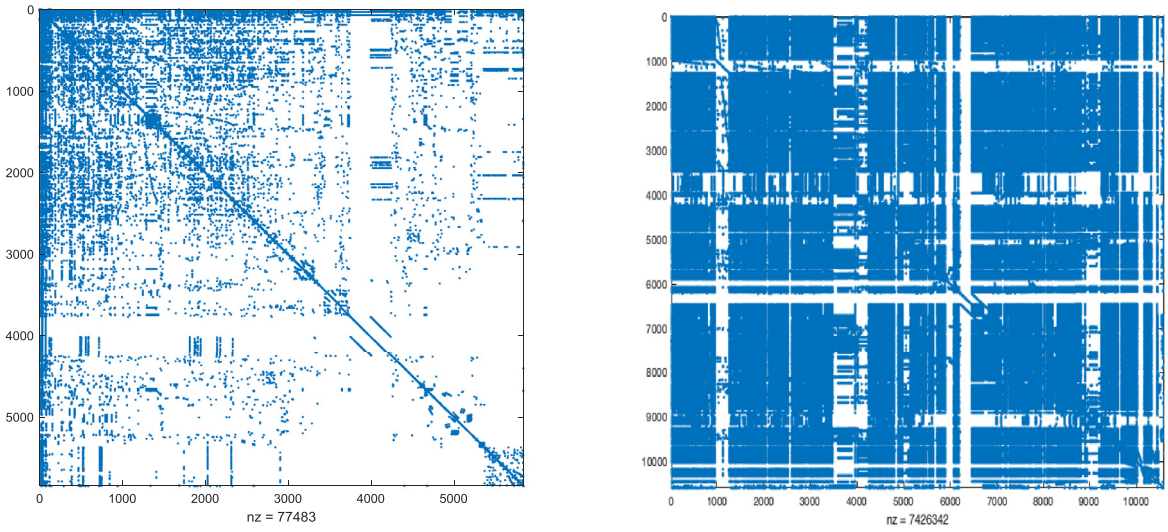
Fig. 8. The stoichiometric matrix of Recon3D. This stoichiometric matrix consists of 5835 rows (molecular species *not* molecules) and 10,600 columns (reactions). Only 0.065% (40, 425/61, 851, 000) of entries are non-zero (nz). The approximate upper diagonal appearance is due to the ordering of the reactions, rather than an intrinsic feature of a stoichiometric matrix. For genome-scale biochemical networks, stoichiometric matrices are sparse because molecular cardinality is typically less than 10 for most reactions.

Each connected component of an atom transition graph consists of vertices with the same elemental label. However, a pair of connected components of an atom transition graph may still be isomorphic with respect to Definition 21 even though they might correspond to different elements. For example, one connected component might correspond to an oxygen atom, while another connected component might correspond to a carbon atom. The atoms of a conserved moiety always corresponds to a subgraph of a molecular substructure. Often this subgraph consists of a single connected component, but it may consist of multiple connected components. For example, a pair of isomorphic connected components may correspond to a pair of atoms in the same molecule



(a) Rank-ordered molecular cardinality (b) Rank-ordered reaction cardinality

Fig. 9. The rank-ordered molecular cardinality molecular (species) and reaction cardinality of Recon3D.



(a) Molecular adjacency matrix NN^T (b) Reaction adjacency matrix N^TN

Fig. 10. The molecular and reaction adjacency matrices of a stoichiometric hypergraph. The sparsity patterns of the molecular (species) adjacency matrix (NN^T, left) and the reaction adjacency matrix (N^TN, right) for Recon3D, have 0.23% and 6.61% non-zero elements (nz), respectively. The fraction of blue is an overestimate of the actual sparsity pattern due as the minimum size of a coloured pixel is greater than the size of an element. Nevertheless, one can observe that it is less common for a pair of molecular species to participate in the same reaction (off-diagonals in NN^T, left) than it is for a pair of reactions to involve the same molecular species (off-diagonals in N^TN, right). . (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

but without a bond between them and therefore the corresponding conserved moiety consists of a set of vertices, but more than one connected component.

Definition 22. Given an atom transition graph $\mathcal{G}(\mathcal{X}, \mathcal{E}, \mathcal{H})$ between a set of molecules $\mathcal{V}$, where $m := |\mathcal{V}|$, a *conserved moiety vector* $L_k \in \mathbb{Z}_+^{1\times m}$ is a non-negative integer (row) vector, where $L_{k,i}$ is the number of instances of the kth conserved moiety in molecule $\mathcal{V}_i$. A set of t conserved moiety vectors can be concatenated to form a *conserved moiety matrix* $L \in \mathbb{Z}_+^{t\times m}$.

If $L_{k,i} = 0$ then the kth conserved moiety is not incident in molecule $\mathcal{V}_i$. There may be more than one instance of a conserved moiety in a molecule, so $L_{k,i} \in \mathbb{Z}_+$ rather than $L_{k,i} \in \{0, 1\}$. To see this, consider a connected component in an atom transition graph that is incident more than once in the same molecule. In this case there will be more than one instance of the corresponding conserved moiety in the same molecule, and therefore $L_{k,i} > 1$.

Corollary 23. *Let* $N \in \mathbb{Z}^{m\times n}$ be a stoichiometric matrix corresponding to a directed stoichiometric hypergraph $\mathcal{H}(\mathcal{V}, \mathcal{Y}(\mathcal{A}, \mathcal{B}))$ with m molecules and n reactions. The conserved moiety matrix $L \in \mathbb{Z}_+^{t\times m}$ derived from the corresponding atom transition graph $\mathcal{G}(\mathcal{X}, \mathcal{E}, \mathcal{H})$ is orthogonal to $\mathcal{R}(N)$, that is $L \cdot N = 0$.

Proof. In an atom mapping, the number of atoms of each element is the same in both tail and head complexes (cf (15)). Therefore, the number of instances of the kth conserved moiety in tail complex of the jth reaction $L_k \cdot F_{:j}$ is the same as the number of instances of the identical conserved moiety in a molecule of a head complex $L_k \cdot R_{:j}$, that is $L_k \cdot F_{:j} = L_k \cdot R_{:j}$, so $L_k \cdot N_{:j} = 0$. Each column of N represents the transformation of a tail complex into a

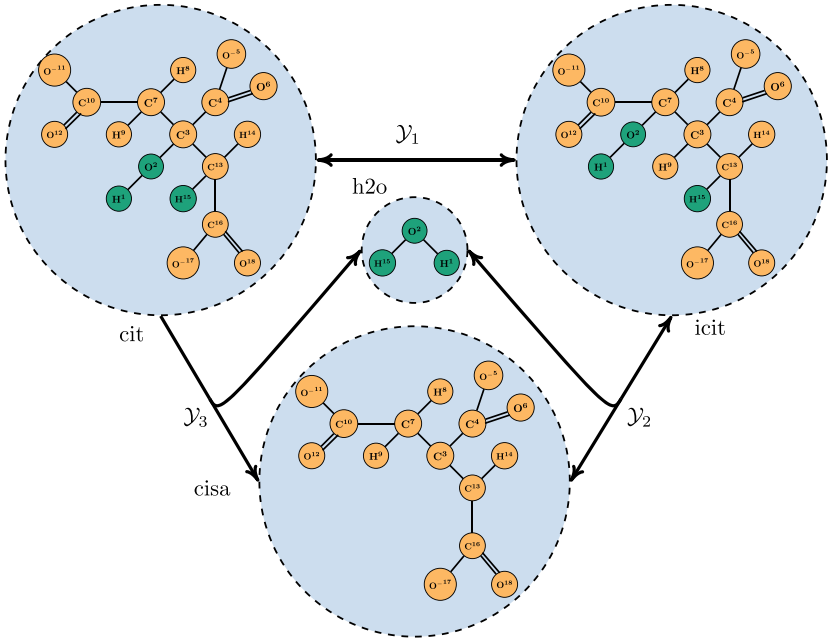


Fig. 11. An atom transition graph for 3 reactions. The molecular structures of each molecule (blue disks) are those of citrate (left, cit), isocitrate (right, cit), water (middle, h2o) and cis-aconitic acid (bottom, cisa). Each atom in each complex is individually labelled (numerical superscripts). The labelling of each atom is invariant with respect to each atom transition. This is a sufficient condition to ensure that an atom transition is always between atoms of the same element. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

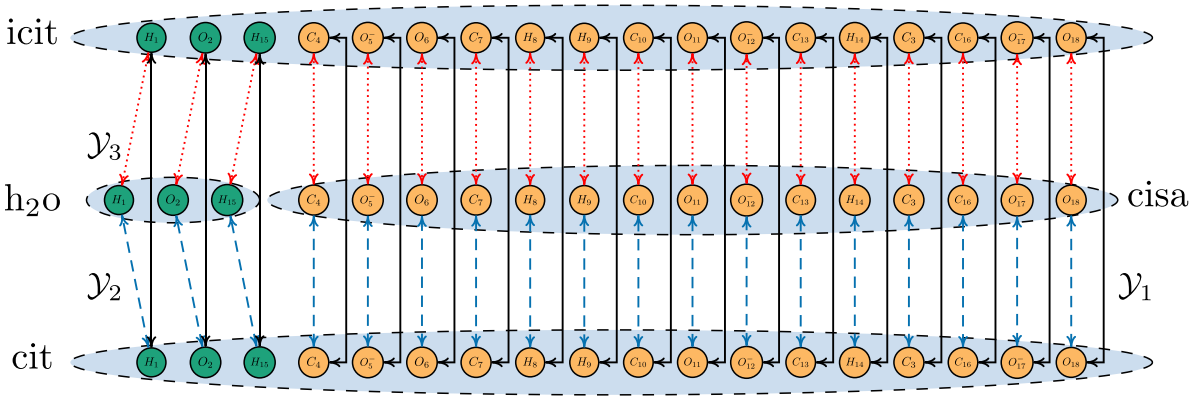


Fig. 12. Connected components and conserved moieties of an atom transition graph. Each molecule in $\{icit, h2o, cit, cisa\}$ is displayed as a set of atoms. Atom transitions are labelled with colours corresponding to reactions $\mathcal{R}_1$(black), $\mathcal{R}_2$(blue) and $\mathcal{R}_3$ (red). Connected components corresponding to atoms 1, 2 and 15 (green, also in Fig. 11) belong to one isomorphism class, that is label preserving with respect to molecular labelling of vertices and reaction labelling of edges. The set of atoms $\{H_1, O_2, H_{15}\}$ are therefore a conserved moiety. Connected components corresponding to atoms 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 3, 16, 17, and 18 (yellow, also in Fig. 11) belong to a different isomorphism class and make up a second conserved moiety. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

head complex, leaving the number of each conserved moiety invariant in every reaction, that is $L \cdot N = 0$. □

$$L := \begin{array}{c} \begin{matrix} h2o[m] & cit[m] & \text{icit}[m] & cisa[m] \end{matrix} \\ \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} L_1 \\ L_2 \end{matrix}$$

6.2. Example conserved moieties

Fig. 12 illustrates the connected components and conserved moieties of the 3 reaction biochemical network introduced in Section 5.1.1. In the atom transition graph introduced in Section 5.2.1, there are two moieties and each of their atoms are labelled green and yellow in Fig. 11 and as sets of connected components in Fig. 12. The conserved moiety matrix corresponding to Fig. 12 is

The first and second conserved moiety vectors, L_1 and L_2 correspond to two isomorphism classes (green and yellow) in Fig. 11. The invariance of the number of moieties with respect to each reaction is illustrated with

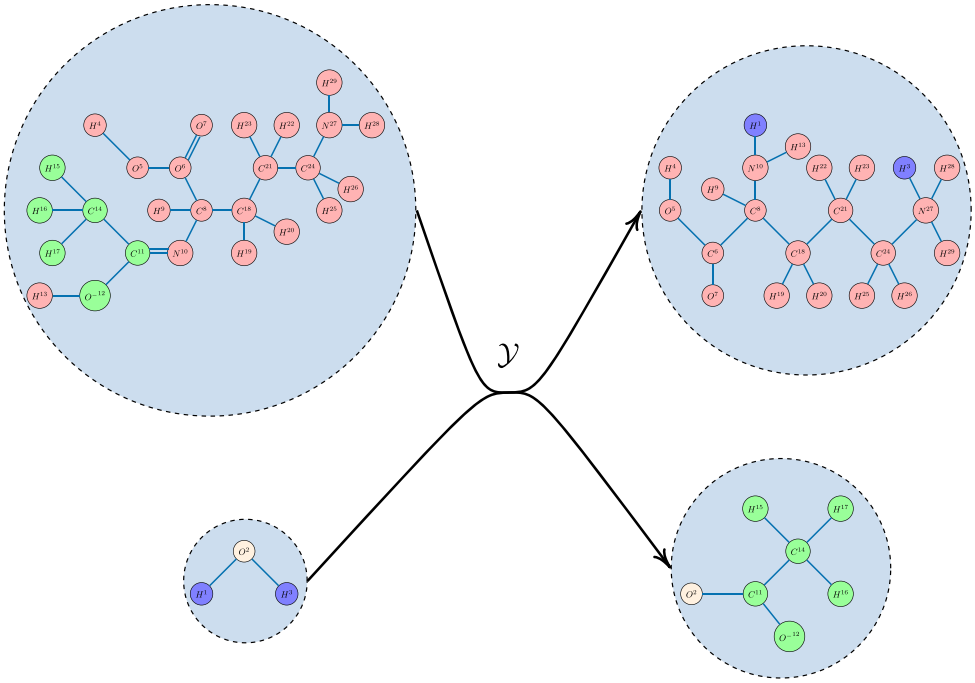


Fig. 13. An organic reaction of the form $ab + cd \rightarrow ac + bd$ where a, b, c and d are moieties. The reaction is acetylornithine deacetylase (ACODA) and the chemical formulas of the moieties are $a = O$, $b = H_2$, $c = C_2H_3O$ and $d = C_5H_{11}N_2O_2$.

Table 1
(a) The stoichiometric matrix $N \in \mathbb{Z}^{4\times 1}$ for a reaction of the form $ab + cd \rightarrow ac + bd$. (b) The conserved moiety matrix $L \in \mathbb{Z}_+^{4\times 4}$ for a reaction of the form $ab + cd \rightarrow ac + bd$ where a, b, c and d are moieties. The matrix has the conserved moiety vectors for a, b, c and d as columns.

	N		l_a	l_b	l_c	l_d
ab	−1	ab	1	1	0	0
cd	−1	cd	0	0	1	1
ac	1	ac	1	0	1	0
bd	1	bd	0	1	0	1
(a)		(b)				

$$LN = \begin{array}{c} \\ L_1 \\ L_2 \end{array} \begin{array}{c} \begin{array}{cccc} h2o[m] & cit[m] & \text{icit}[m] & cisa[m] \end{array} \\ \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \end{array} \cdot \begin{array}{c} \begin{array}{ccc} \mathcal{Y}_1 & \mathcal{Y}_2 & \mathcal{Y}_3 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} \end{array} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

.

6.3. Redundancy of conserved moiety vectors

The following example illustrates that there may exist more than $m - \text{rank}(N)$ conserved moiety vectors orthogonal to $\mathcal{R}(N)$ that are linearly dependent. Consider a single reaction of the form

$$ab + cd \rightharpoonup ac + bd,$$

where a, b, c and d are moieties. The stoichiometric matrix $N \in \mathbb{Z}^{m\times n}$ and conserved moiety matrix $L \in \mathbb{Z}_+^{t\times m}$, respectively, are given in Table 1a and b. The number of moieties is $k = 4$, the number of molecules is $m = 4$ and $\text{rank}(N) = 1$. Therefore, $t > m - \text{rank}(N)$. A moiety basis for N can be formed by selecting any three of the four conserved moiety vectors in L, giving a total of four possible combinations. A real example reaction of this form is shown in Fig. 13.

7. Moiety splitting

Given a stoichiometric hypergraph and its corresponding atom transition graph, subject to certain assumptions, we now show how to split a stoichiometric matrix into a non-negative sum of incidence matrices, each of which corresponds to a compartmental network.

7.1. Moiety splitting of a stoichiometric matrix

Theorem 24. *(Moiety splitting) Let $N \in \mathbb{Z}^{m\times n}$ be a stoichiometric matrix, with $r = \text{rank}(N)$, such that there exists an $L \in \mathbb{Z}_+^{m-r\times m}$ and $LN = 0$, where each L_k is a moiety vector, for all $k \in 1, \ldots, m - r$, then the following matrix splitting exists*

$$N = \text{diag}^{-1}\left(L^T\mathbb{1}\right)\sum_{k=1}^{m-r} N(k), \tag{4}$$

where $N(k) \in \mathbb{Z}^{m\times n}$ is a moiety transition matrix, given by

$$N(k) := \text{diag}(L_k)N \tag{5}$$

Proof. Substituting (4) into (5), it is enough to show $\ell := L^T\mathbb{1} \in \mathbb{Z}_{++}^m$ and that

$$\text{diag}\left(L^T\mathbb{1}\right) = \sum_{k=1}^{m-r}\text{diag}(L_k).$$

The expression on the left sums each column of L then places it on the diagonal of an $m \times m$ matrix. The expression on the

Table 2
Moiety splitting of a stoichiometric matrix.

$$N(1) \coloneqq \operatorname{diag}(L_1)N = \begin{array}{|ccc|}\hline \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 \\ \hline 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \\ \hline \end{array} = \begin{array}{|cccc|}\hline h2o & cit & \text{icit} & cisa \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ \hline \end{array} \cdot \begin{array}{|ccc|}\hline & & \\ \hline 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ \hline \end{array} \begin{array}{c} \\ h2o \\ cit \\ icit \\ cisa \end{array}$$

$$N(2) \coloneqq \operatorname{diag}(L_2)N = \begin{array}{|ccc|}\hline \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 \\ \hline 0 & 0 & 0 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ \hline \end{array} = \begin{array}{|cccc|}\hline h2o & cit & \text{icit} & cisa \\ \hline 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline \end{array} \cdot \begin{array}{|ccc|}\hline & & \\ \hline 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ \hline \end{array} \begin{array}{c} \\ h2o \\ cit \\ icit \\ cisa \end{array}$$

$$N = \operatorname{diag}^{-1}\left(L^T\mathbb{1}\right)\left(N(1)+N(2)\right) = \begin{array}{|ccc|}\hline \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 \\ \hline 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ \hline \end{array} =$$

$$= \begin{array}{|ccc||cccc|}\hline \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 & h2o & cit & \text{icit} & cisa \\ \hline 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ \hline \end{array} \cdot \left(\begin{array}{|ccc|}\hline & & \\ \hline 0 & 1 & 1 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \\ \hline \end{array} + \begin{array}{|ccc|}\hline & & \\ \hline 0 & 0 & 0 \\ -1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & 1 \\ \hline \end{array} \begin{array}{c} \\ h2o \\ cit \\ icit \\ cisa \end{array} \right)$$

right places each row of L on the diagonal of a matrix, and sums the matrices, which is equivalent to the expression on the left as the operations involved are commutative. Each entry of L is non-negative so $\ell \geq 0$, therefore it remains to show that $L^T\mathbb{1} \in \mathbb{Z}^m_{++}$. By Definition 19, an atom transition graph $\mathcal{G}(\mathcal{X}, \mathcal{E}, \mathcal{H})$ is formed by joining all atom mappings corresponding to a stoichiometric hypergraph $\mathcal{H}(\mathcal{X}, \mathcal{Y}\{\mathcal{A}, \mathcal{B}\})$. Every molecule is therefore part of some atom transition graph, and therefore some isomorphism class, so $L^T\mathbb{1} \in \mathbb{Z}^m_{++}$, giving the desired result. □

It is an open question as to the biochemically interpretable conditions required to be satisfied for there to exist an $L \in \mathbb{Z}_+^{m-r\times m}$ and $LN = 0$, where each L_k is a moiety vector, for all $k \in 1, \ldots, m-r$. In general, each stoichiometric matrix for a moiety subnetwork $N(k)$ is significantly more sparse than N. Since a molecule may contain more than one type of conserved moiety, some of rows, or multiplications thereof, are often repeated in several moiety transition matrices $N(k)$. The splitting formalised in Theorem 24 exists for any matrix $N \in \mathbb{Z}^{m\times n}$ such that, there exists an $L \in \mathbb{Z}_+^{t\times m}$ satisfying $L^T\mathbb{1} \in \mathbb{Z}^m_{++}$.

7.2. Example of moiety splitting

Moiety splitting of a stoichiometric matrix, by application of Theorem 24, for the three reaction Eq. (3), using the conserved moiety vectors given in Section 6.2, is illustrated in Table 2.

The two corresponding moiety subnetworks are both graphs, as illustrated in Fig. 14.

7.3. Moiety transition matrices

We next provide some technical properties of the matrices $N(k)$, $k \in 1, \ldots, m-r$, which shows that $N(k)$ is conserved and the rank of $N(k)$ is related to the number of components of the associated subnetwork. Recall that a vertex without any incident edges is considered a (trivial) component.

Theorem 25. *Let $N \in \mathbb{Z}^{m\times n}$ be a stoichiometric matrix, with $r = \operatorname{rank}(N)$, such that there exists an $L \in \mathbb{Z}_+^{m-r\times m}$ with $LN = 0$, where each $L_k \in \mathbb{Z}_+^{1\times m}$ is a moiety vector and $N_k \coloneqq \operatorname{diag}(L_k)N$ is an incidence matrix for a moiety subnetwork, for all $k \in 1, \ldots, m-r$, then the following assertions hold:*

(i) *each matrix $N(k)$ is conserved;*
(ii) *each moiety subnetwork is one connected component (graph or hypergraph);*
(iii) *if c denotes the number of components of the subnetwork $N(k)$, then*

$$\operatorname{rank}(N(k)) = m - c;$$

(iv) *if $\mathcal{N}(N(k))$ and $\mathcal{N}(N(k)^T)$ denote the nullspace and the left nullspace of $N(k)$, then*

$$\dim(\mathcal{N}(N(k))) = n - (m - c),$$
$$\dim(\mathcal{N}(N(k)^T)) = c.$$

.

Proof. By the definition of $N(k)$, we get

$$\mathbb{1}N(k) = \mathbb{1}\operatorname{diag}(L_k)N = L_kN = 0,$$

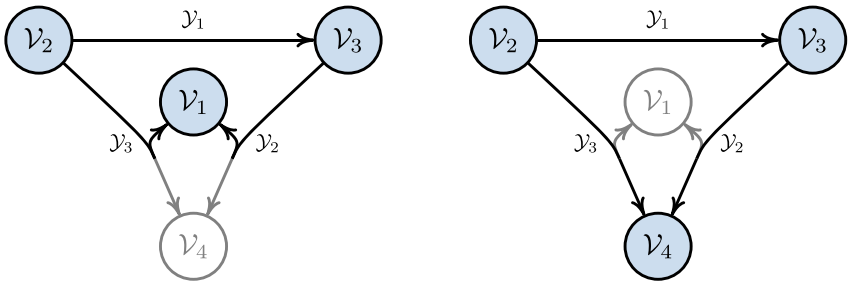


Fig. 14. A stoichiometric hypergraph split into two moiety graphs. The three reaction network introduced in Section 5.1.1, is split into two moiety subnetworks. The $N(1)$ moiety subnetwork (a) is a graph with that omits the vertex $\mathcal{V}_4$, while the $N(2)$ moiety subnetwork (b) is a graph with that omits the vertex $\mathcal{V}_1$ from the stoichiometric hypergraph N.

giving Assertion (i). To prove Assertion (ii), recall that each moiety corresponds to an isomorphism class of connected components in an atom transition graph, so each moiety subnetwork is a connected component, therefore in $N(k)$ there is only one connected component which could be a graph or a hypergraph. In order to prove Assertion (iii), without loss of generality by a suitable reordering the rows of $N(k)$, we rewrite $N(k)$ in the following form,

$$N(k) = \begin{bmatrix} d_1 \\ \vdots \\ d_{m-c+1} \\ \vdots \\ d_m \end{bmatrix},$$

where d_i for $i \in 1, \ldots, m-c+1$ corresponding to nonzero rows of $N(k)$ (representing the incidence of the connected component) and $d_i = 0$ for $i \in m-c+2, \ldots, m$ (standing for unconnected component that is a single vertex without any edge). For sake of simplicity, we denote the first $m-c+1$ rows of $N(k)$ as $N(k)^{(1)}$.

If the connected component $N(k)^{(1)}$ is a graph, since there is just one $+1$ and just one -1 in each column of $N(k)^{(1)}$, it follows that the sum of the rows of $N(k)^{(1)}$ is the zero row vector, and that the rank of $N(k)^{(1)}$ is at most $m-c$. On the other hand, if $N(k)^{(1)}$ represents a hypergraph, also because of being conserved moiety the sum of the rows of $N(k)^{(1)}$ is the zero row vector, and the rank of $N(k)^{(1)}$ is at most $m-c$. We now show, the rank of $N(k)^{(1)}$ is exactly $m-c$. To do so, suppose we have a linear relation $\sum \gamma_j d_j = 0$, where the summation is over all rows of $N(k)^{(1)}$, and not all of the coefficients γ_j are zero. Choose a row d_k for which $\gamma_k \neq 0$. If $N(k)^{(1)}$ represents a graph then this row has non-zero entries in those columns corresponding to the directed edges incident with v_k. For each such column, there is just one other row d_l with a non-zero entry in that column, and in order that the given linear relation should hold, we must have $\gamma_l = \gamma_k$. Thus, if $\gamma_k \neq 0$, then $\gamma_l = \gamma_k$ for all vertices v_l adjacent to v_k. Since $N(k)^{(1)}$ is a connected component, it follows that all coefficients γ_j are equal, i.e., the given linear relation is just multiple of $\sum d_j = 0$. Consequently, the rank of $N(k)^{(1)}$ is $m-c$. Consequently, the rank of $N(k)^{(1)}$ is $m-c$. From the structure of matrix $N(k)$, it is evident that

$$\text{rank}(N(k)) = \text{rank}N(k)^{(1)} = m-c,$$

which proves Assertion (ii). The results of Assertion (iv) are straightforward from (iii). □

Note that the connected component of $N(k)$ given in this theorem can be a graph or a hypergraph. Given N with $r = \text{rank}(N)$, and assuming there exists an $L \in \mathbb{Z}_+^{m-r\times m}$ satisfying $LN = 0$, and where each L_k is a moiety vector, for all $k \in 1, \ldots, m-r$, it is an open question as to the biochemically interpretable conditions required to be satisfied for $N(k) := \text{diag}(L_k)N$ to always result in an incidence matrix for a graph, as opposed to a hypergraph.

Table 3
A set of reaction equations for part of human dopamine synthesis.

$\mathcal{R}_1$:	Phe + BH_4+O_2	→	Tyr + BH_2+H_2O,
$\mathcal{R}_2$:	Tyr+BH_4+O_2	→	$L-DOPA + BH_2 + H_2O$,
$\mathcal{R}_3$:	$L-DOPA + H^+$	→	$DA + CO_2$
$\mathcal{R}_4$:	$Formate + BH_2 + H^+$	→	$CO_2 + BH_4$.

Theorem 25 (i) clearly implies that the vector of $\mathbb{1}$ is in the left nullspace of $N(k)$, i.e., $\mathbb{1} \in \mathcal{N}(N(k)^T)$, which has been a known result for incidence matrix of a graph; however, the subnetwork associated to $N(k)$, $k \in 1, \ldots, m-r$, can be a hypergraph, but $N(k)$ is still conserved.

7.4. Example moiety transition matrices

Consider the matrices $N(1)$ and $N(2)$ in Table 2, where it can be seen that the summation of elements of each column is zero, consistent with Theorem 25(i). In Fig. 14(a) and (b), the $N(1)$ and $N(2)$ moiety graphs each consists of 3 vertices and one component therefore $\text{rank}(N(1)) = m-c = 3-1 = 2$.

For a slightly larger example, consider the directed stoichiometric hypergraph corresponding to part of human dopamine synthesis, investigated in Haraldsdóttir and Fleming (2016). It consists of four reactions and eleven molecules given in Table 3. The stoichiometric matrix and a conserved moiety basis for this network is given in Table 4. Since $N \in \mathbb{Z}^{4\times 11}$ and $\text{rank}(N) = 4$, it follows that $\dim(\mathcal{N}(N)) = 11-4 = 7$, so $L \in \mathbb{Z}_+^{7\times 11}$ and therefore this stoichiometric hypergraph may be split into 7 moiety subnetworks. Consider the $N(6)$ moiety subnetwork in Fig. 15. The sum of elements of each column of the $N(6)$ stoichiometric matrix is zero, consistent with Theorem 25(i). There is only one non-trivial component, which is a hypergraph, since it contains one directed hyperedge $(\mathcal{F}\{\mathcal{V}_2, \mathcal{V}_9\}, \mathcal{R}\{\mathcal{V}_3\})$. From Fig. 15(b), one observes $c = 8$ components, $m = 11$ vertices, and $n = 3$ (2 edges and one hyperedge). Hence, Theorem 25(iii) implies $\text{rank}(N(k)) = m-c = 11-8 = 3$, $\dim(\mathcal{N}(N(k))) = n-(m-c) = 4-(11-8) = 1$ and $\dim(\mathcal{N}(N(k)^T)) = c = 8$.

7.5. Moieties in thermodynamically closed and open systems

In all of the examples presented thus far, we assume we are given a stoichiometric matrix $N \in \mathbb{Z}^{m\times n}$, with $r = \text{rank}(N)$, such that there exists an $L \in \mathbb{Z}_+^{m-r\times m}$ and $LN = 0$, where each L_k is a moiety vector, for all $k \in 1, \ldots, m-r$. As a consequence, we are only considering reactions that are mass balanced. This begs the question, with these assumptions how one can model the chemical reaction network of a living system, which is a thermodynamically open system, if every reaction in the system is mass balanced? In most of the literature on mathematical modelling of living systems,

Table 4
The stoichiometric matrix, a moiety basis matrix and its column sum, for part of human dopamine synthesis.

$$N := \begin{array}{c} \\ Phe \\ Tyr \\ L-DOPA \\ DA \\ CO2 \\ Formate \\ BH4 \\ BH2 \\ O2 \\ H2O \\ H+ \end{array} \begin{array}{c} \begin{array}{cccc} \mathcal{R}_1 & \mathcal{R}_2 & \mathcal{R}_3 & \mathcal{R}_4 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & -1 \\ -1 & -1 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 \end{bmatrix} \end{array}$$

$$L := \begin{array}{c} \\ Phe \\ Tyr \\ L-DOPA \\ DA \\ CO_2 \\ Formate \\ BH_4 \\ BH_2 \\ O_2 \\ h2O \\ H+ \end{array} \begin{array}{c} \begin{array}{ccccccc} l_1 & l_2 & l_3 & l_4 & l_5 & l_6 & l_7 \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix} \end{array}, \quad L^T\mathbb{1} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 4 \\ 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}$$

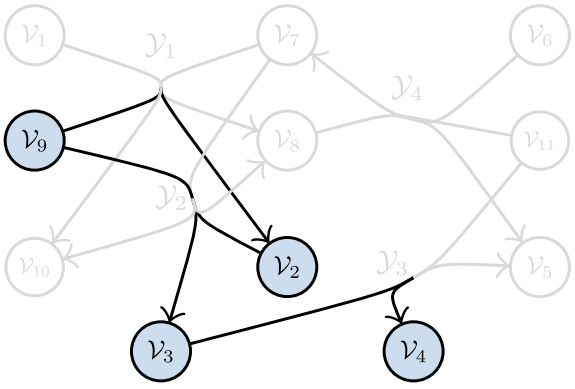
thermodynamic forcing by the environment is represented by mass imbalanced reactions that inject mass across the boundary of a model. That is, one can inject mass across the boundary of a model by augmenting a stoichiometric matrix with a set of faux reactions that are mass imbalanced. This condition is sufficient but not necessary for modelling a living system. Alternatively, it is sufficient if all reactions are mass balanced, but a subset of faux elementary reactions are given thermodynamically infeasible kinetic parameters (Fleming and Thiele, 2012). In (Fleming and Thiele, 2012), such reactions were termed *perpetireactions,* where *perpeti* is from the latin *perpes* meaning perpetual. Any vector in the range of N may be used to represent the stoichiometry of a perpetireaction (Fleming and Thiele, 2012). With respect to conserved moieties, the following *cyclic stoichiometric matrix*

$$C := \begin{bmatrix} N & -I \\ 0 & L \end{bmatrix}$$

augments a stoichiometric matrix $N \in \mathbb{Z}^{m\times n}$ with m perpetireactions, each of which consumes one molecule and produces its constituent set of conserved moieties, corresponding to one column of the conserved moiety basis $L \in \mathbb{Z}_+^{m-r\times m}$. In an electrical network, there is one moiety, corresponding to an electron, and electrical networks are shown as closed loops, where the circulation of electrons must be driven by some energy source. Likewise, a living system and the part of its environment that it exchanges mass with, may be considered as a set of $m-r$ cycles, one for each of its constituent moieties. Ultimately these cycles must each be driven by an external energy input. Since $[L, I]$ is a non-negative left nullspace basis for C, if N admits a moiety splitting, C satisfies the conditions for Theorem 24 and so also admits a moiety splitting.

8. Discussion

While many biological discoveries have been made possible by applying established mathematical theory and algorithms, the reverse is also true. That is, the biology itself can also inspire the development of novel mathematical theory and algorithms. It is therefore important to keep in mind the principles that govern the behaviour of a biological system being modelled. In biology, mathematical and computational modelling play complimentary roles. Given certain assumptions, a mathematical model of a reaction network allows one to reason about biochemical networks in general. However, each mathematical model depends on a set of assumptions and, even if each of these assumptions are appropriate, it is hard to be sure that any set of assumptions is sufficient. Computational examples of real world biochemical networks complement mathematical models of generic biochemical networks by suggesting new assumptions that might otherwise not be consid-



$$N(6) := \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 2 & -2 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \mathcal{V}_1 & Phe \\ \mathcal{V}_2 & Tyr \\ \mathcal{V}_3 & L-DOPA \\ \mathcal{V}_4 & DA \\ \mathcal{V}_5 & CO2 \\ \mathcal{V}_6 & Formate \\ \mathcal{V}_7 & BH4 \\ \mathcal{V}_8 & BH2 \\ \mathcal{V}_9 & O2 \\ \mathcal{V}_{10} & H2O \\ \mathcal{V}_{11} & H+ \end{matrix}$$

Fig. 15. A moiety subnetwork that is a hypergraph. A moiety subnetwork of a directed stoichiometric hypergraph corresponding to part of human dopamine synthesis (Haraldsdóttir and Fleming, 2016). There are 8 components $\{\mathcal{V}_1\}$, $\{\mathcal{V}_5\}$, $\{\mathcal{V}_6\}$, $\{\mathcal{V}_7\}$, $\{\mathcal{V}_8\}$, $\{\mathcal{V}_{10}\}$, $\{\mathcal{V}_{11}\}$, and $\{\mathcal{V}_9, \mathcal{V}_2, \mathcal{V}_3, \mathcal{V}_4\}$. The latter component corresponds to the hypergraph since $(\mathcal{F}\{\mathcal{V}_2, \mathcal{V}_9\}, \mathcal{R}\{\mathcal{V}_3\})$ is a directed hyperedge.

ered. For example, it is often assumed that a stoichiometric matrix is an arbitrary rectangular matrix with integer entries, that is $N \in \mathbb{Z}^{m\times n}$. However, from the perspective of this paper, which is motivated by example biochemical networks, that assumption is insufficient. Stoichiometric matrices are central in a number of modelling paradigms for reaction networks. These include optimisation, ordinary differential equation, and stochastic models. Since the complexity of networks for biological, medical, and technological applications is already straining existing solution algorithms and is only expected to grow further, special properties of the stoichiometric matrix that are rooted in the underlying biochemistry can inform the design of efficient ad hoc algorithms to overcome this hurdle.

In this paper, and in previous efforts (Fleming, 2016), we have focussed on mathematical properties that seem to be peculiar to stoichiometric matrices. These properties should not be considered the definitive set of properties. We are still in the process of identifying the mathematical properties of stoichiometric matrices and there are very likely to be additional properties to discover. Furthermore, it is an open problem to establish which known properties are fundamental, and which are the consequences of other properties, known or unknown. With each new insight into the properties of stoichiometric matrices, there will be more opportunities to use their special structure in mathematical modelling of reaction networks. This has been the motivation for exploring these properties that we believe will help us to work towards a full definition and thereby characterisation of biochemical network topology.

The size of genome-scale biochemical network models has been growing exponentially, in response to the big data revolution in molecular systems biology in the last two decades. The generation of biological understanding from multiple omic datasets requires integrative analysis that compares this new data with hypotheses derived from prior information, in the form of computational model predictions. As discussed, the key element in developing mathematical models for biochemical networks is the stoichiometric matrix constructed from such biological data, which has considerable effects on properties of mathematical models. Therefore, emerging big data in biology brings many opportunities as well as many challenges to scientists in the fields of statistics, applied mathematics, and system biology. This increasingly demands developing new mathematical models that are biologically meaningful and computationally tractable. To this end, the specific structure of stoichiometric matrices should be taken into account to design novel context-specific algorithmic methodologies for such big data problems.

Several classes of optimisation models can be developed for responding to different problems arising in system biology such as linear (Ma et al., 2017), quadratic (Segré et al., 2002), convex (Fleming et al., 2012), duplomonotone (Artacho and Fleming, 2014) and other nonlinear problems (Ahookhosh et al., 2019, Ahookhosh et al. 2020.). For each class of optimisation models, the specific structure of the stoichiometric matrices might be used to decompose the original problem to several computationally tractable subproblems using the moiety matrix splitting described in Section 7. This may lead to algorithms with lower analytic and computational complexities. One of the main properties of stoichiometric matrices that can be effectively used in development of novel optimisation methodology is their *sparsity* pattern, i.e., most of the elements of large stoichiometric matrices are zero. Here, we emphasise that for implementations of optimisation methodologies, one needs to know about function values and gradients (subgradients in nonsmooth cases) of the objective function of the optimisation problems, which require some matrix-vector products. In order to decrease the computational complexity of these matrix-vector products, one can exploit the sparsity of stoichiometric matrices combined with high-performance computing to efficiently provide required information for the corresponding methodology. We assert that by combining existing techniques from these two approaches, and developing new tailored techniques that mix elements of these two approaches, will provide a secure path toward solving high dimensional optimisation problems arising in the study of biochemical networks.

9. Conclusion

A biochemical network with m molecules and n reactions may be expressed as a hypergraph $\mathcal{H}(\mathcal{V}, \mathcal{Y})$ that consists of a set of m vertices $\mathcal{V} := \{\mathcal{V}_1, \ldots, \mathcal{V}_m\}$, each corresponding to one molecule, and a set of n hyperedges $\mathcal{Y} := \{\mathcal{Y}_1, \ldots, \mathcal{Y}_n\}$, each corresponding to one reaction. Once an orientation is chosen for each reaction, topology of such a network may be represented by a stoichiometric matrix $N \in \{-1, 0, 1\}^{m\times n}$, with $r = \text{rank}(N)$, where $N_{i,j} < 0$ if molecule i is consumed in reaction j and $N_{i,j} > 0$ if molecule i is produced in reaction j. However, N is not an arbitrary rectangular sign matrix because there exists a set of non-negative vectors $L \in \mathbb{Z}^{m-r\times m}$ such that (i) $L^T\mathbb{1} > 0$, (ii) $LN = 0$, (iii) $N(k) := \text{diag}(L_k)N$ corresponds to an incidence matrix for a graph, or a hypergraph, with one connected component with the property that $\mathbb{1}N(k) = 0$

and (iv) the following matrix splitting exists

$$N = \text{diag}^{-1}\left(L^T \mathbb{1}\right) \sum_{k=1}^{m-r} N(k). \tag{6}$$

Fundamentally, these properties arise due to the fact that, at a higher level of resolution: (i) a molecule may be represented as a molecular graph where each vertex represents an atom and each edge represents a chemical bond between a pair of atoms in a molecule, and (ii) a reaction may be represented as a graph where each vertex represents an atom and each edge represents a transition between an atom in a substrate complex and an atom of the same element in a product complex. It is an open problem to establish whether these properties are particular to stoichiometric matrices alone, or whether they have been studied in graph or hypergraph theory but not yet applied to biology.

CRediT authorship contribution statement

Susan Ghaderi: Formal analysis, Visualization, Writing - review & editing. **Hulda S. Haraldsdóttir:** Conceptualization, Writing - review & editing. **Masoud Ahookhosh:** Writing - review & editing. **Sylvain Arreckx:** Writing - review & editing. **Ronan M.T. Fleming:** Conceptualization, Funding acquisition, Supervision, Validation, Writing - original draft, Writing - review & editing.

Acknowledgments

This work was supported by the U.S. Department of Energy, Offices of Advanced Scientific Computing Research and the Biological and Environmental Research as part of the Scientific Discovery Through Advanced Computing program, grant #DE-SC0010429 and by the EU Marie Sklodowska-Curie Innovative Training Network, grant #812616.

References

Ahookhosh, M., et al., 2019. Local convergence of the Levenberg–Marquardt method under Hölder metric subregularity. Adv. Comput. Math 45, 2771–2806.
Ahookhosh, M., Fleming, R.M.T., Vuong, P.T., 2020. Finding zeros of Hölder metrically Subregular mappings via globally convergent Levenberg–Marquardt methods. Optim. Methods Softw. 1–37. doi:10.1080/10556788.2020.1712602.
Artacho, F.J.A., Fleming, R.M.T., 2014. Globally convergent algorithms for finding zeros of duplomonotone mappings. Optim. Lett. 9 (569), 1–16.
Barabási, A.-L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5 (2), 101–113.
Brunk, E., et al., 2018. Recon3D enables a three-dimensional view of gene variation in human metabolism. Nat. Biotechnol. 36, 272.
Clarke, B.L., 1988. Stoichiometric network analysis. Cell Biophys. 12, 237–253.
Craciun, G., Feinberg, M., 2006. Multiple equilibria in complex chemical reaction networks: II. The species-reaction graph. SIAM J. Appl. Math. 66 (4), 1321–1338.
Famili, I., Palsson, B.Ø., 2003. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. Biophys. J. 85 (1), 16–26.
Fleming, R.M., Thiele, I., 2012. Mass conserved elementary kinetics is sufficient for the existence of a non-equilibrium steady state concentration. J. Theor. Biol. 314, 173–181.
Fleming, R.M.T., et al., 2012. A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks. J. Theor. Biol. 292, 71–77.
Fleming, R.M.T., et al., 2016. Conditions for duality between fluxes and concentrations in biochemical networks. J. Theor. Biol. 409, 1–10.
Funahashi, A., et al., 2008. Celldesigner 3.5: A versatile modeling tool for biochemical networks. Proc. IEEE 96 (8), 1254–1265. 00336.
Gill, P.E., et al., 1987. Maintaining LU factors of a general sparse matrix. Linear Algebra Appl. 88–89, 239–270.
Haraldsdóttir, H.S., Fleming, R.M.T., 2016. Identification of conserved moieties in metabolic networks by graph theoretical analysis of atom transition networks. PLOS Comput. Biol. 12 (11), e1004999.
Ingalls, B.P., 2013. Mathematical Modeling in Systems Biology: An Introduction. MIT Press, p. 423.
Klamt, S., Haus, U.-U., Theis, F., 2009. Hypergraphs and cellular networks. PLoS Comput. Biol. 5 (5), e1000385. doi:10.1371/journal.pcbi.1000385.
Lane, N., Pariseau, K., 2016. The vital question: energy, evolution, and the origins of complex life. MP3 una edition. Audible Studios on Brilliance Audio, Mar. 15.
Lewis, G.N., 1925. A new principle of equilibrium. Proc. Natl. Acad. Sci. USA 11 (3), 179–183.
Ma, D., et al., 2017. Reliable and efficient solution of genome-scale models of metabolism and macromolecular expression. Sci. Rep. 7, 00003. srep40863.
Noronha, A., et al., 2017. Reconmap: an interactive visualization of human metabolism. Bioinformatics 33 (4), 605–607.
Palsson, B.O., 2015. Systems Biology: Constraint-Based Reconstruction and Analysis. Cambridge University Press, Cambridge, England, p. 550.
Papin, J.A., et al., 2004. Comparison of network-based pathway analysis methods. Trends Biotechnol. 22 (8), 400–405.
Plasson, R., Bersini, H., Brandenburg, A., 2008. Decomposition of complex reaction networks into reactions. https://arxiv.org/abs/0803.1385.
Rahman, S.A., et al., 2016. Reaction decoder tool (RDT): extracting features from chemical reactions. Bioinformatics 32 (13), 2065–2066.
Segré, D., Vitkup, D., Church, G.M., 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc. Natl. Acad. Sci. USA 99 (23), 15112–15117.
Trinajstić, N., 1992. Chemical graph theory. In: Mathematical Chemistry Series. CRC Press, Boca Raton, p. 322.
Vallabhajosyula, R.R., Chickarmane, V., Sauro, H.M., 2006. Conservation analysis of large biochemical networks. Bioinformatics 22 (3), 346–353.
Voloshin, V.I., 2009. Introduction to Graph and Hypergraph Theory. Nova Science Publishers.
Wilson, R.J., 2010. Introduction to Graph Theory, Fifth ed. Pearson, Harlow, United Kingdom.