Manuscript Submission

Journal of Computational Biology

Cayley Graphs of Semigroups Applied to Atom Tracking in Chemistry

Authors:

Nikolai Nøjgaard[1], Walter Fontana[3], Marc Hellmuth[2], and Daniel Merkle[1*]

[1] Department of Mathematics and Computer Science
University of Southern Denmark, Odense M DK-5230, Denmark
Tel: +45 6550 2387
Email: {merkle,nojgaard}@imada.sdu.dk

[2] School of Computing
University of Leeds, Leeds, UK
Email: mhellmuth@mailbox.org

[3] Harvard Medical School
Boston, MA 02115, United States
Email: walter_fontana@hms.harvard.edu

Submitted: Friday 30th October, 2020

Cayley Graphs of Semigroups Applied to Atom Tracking in Chemistry

Abstract

While atom tracking with isotope-labeled compounds is an essential and sophisticated wet-lab tool in order to, e.g., illuminate reaction mechanisms, there exists only a limited amount of formal methods to approach the problem. Specifically when large (bio-)chemical networks are considered where reactions are stereo-specific, rigorous techniques are inevitable. We present an approach using the right Cayley graph of a monoid in order to track atoms concurrently through sequences of reactions and predict their potential location in product molecules. This can not only be used to systematically build hypothesis or reject reaction mechanisms (we will use the ANRORC mechanism "Addition of the Nucleophile, Ring Opening, and Ring Closure" as an example), but also to infer naturally occurring subsystems of (bio-)chemical systems. Our results include the analysis of the carbon traces within the TCA cycle and infer subsystems based on projections of the right Cayley graph onto a set of relevant atoms.

Keywords— Computational Biology, Graph Transformations, Double Pushout, Chemical Reaction Networks, Algorithmic Cheminformatics

1 Introduction

Traditionally, atom tracking is used in chemistry to understand the underlying reactions and interactions behind some chemical or biological system. In practice, atoms are usually tracked using isotopic labeling experiments. In a typical isotopic labeling experiment, one or several atoms of some educt molecule of the chemical system we wish to examine are replaced by an isotopic equivalent (e.g. ^{12}C is replaced with ^{13}C). These compounds are then introduced to the system of interest, and the resulting product compounds are examined, e.g. by mass spectrometry (Chahrour, Cobice, and Malone 2015) or nuclear magnetic resonance (Deev et al. 2019). By determining the positions of the isotopes in the product compounds, information about the underlying reactions might then be derived. From a theoretical perspective, characterizing a formal framework to track atoms through

reactions is an important step to understand the possible behaviors of a chemical or biological system. In this contribution, we introduce such a framework based on concepts rooted in semigroup theory. Semigroup theory can be used as a tool to analyze biological systems such as metabolic and gene regulatory networks (Nehaniv et al. 2015; Egri-Nagy and Nehaniv 2008). In particular, Krohn-Rhodes theory (Rhodes and Nehaniv 2009) was used to analyze biological systems by decomposing a semigroup into simpler components. The networks are modeled as state automatas (or ensembles of automatas), and their characteristic semigroup, i.e., the semigroup that characterizes the transition function of the automata (Mikolajczak 1991), is then decomposed using Krohn-Rhodes decompositions or, if not computationally feasible, the holonomy decomposition variant (Egri-Nagy and Nehaniv 2015). The result is a set of symmetric natural subsystems and an associated hierarchy between them, that can then be used to reason about the system. In (Andersen, Merkle, and Rasmussen 2019) algebraic structures were employed for modeling atom tracking: graph transformation rules are iteratively applied to sets of undirected graphs (molecules) in order to generate the hyper-edges (the chemical reactions) of a directed hypergraph (the chemical reactions network) (Andersen et al. 2016; Andersen et al. 2013). A semigroup is defined by using the (partial) transformations that naturally arise from modeling chemical reactions as graph transformations. Utilizing this particular semigroup so-called pathway tables can be constructed, detailing the orbit of single atoms through different pathways to help with the design of isotopic labeling experiments.

In this work, we show that we can gain a deeper understanding of the analyzed system by considering how atoms move in relation to each other. To this end, we briefly introduce useful terminology in Section 2, found in graph transformation theory as well as semigroup theory. In Section 3 we show how the possible trajectories of a subset of atoms can be intuitively represented as the (right) Cayley graph (Dénes 1966) of the associated semigroup

of a chemical network. Moreover, we define natural subsystems of a chemical network in terms of reversible atom configurations and show how they naturally relate to the strongly connected components of the corresponding Cayley graph. We show the usefulness of our approach in Section 4.1 by using the constructions defined in Section 3 to differentiate chemical pathways, based on the atom trajectories derived from each pathway. We then show how the Cayley graph additionally provides a natural handle for the analysis of cyclic chemical systems such as the TCA cycle (Harvey and Ferrier 2010).

2 Preliminaries

Graphs: In this contribution we consider directed as well as undirected connected graphs $G = (V, E)$ with vertex set $V(G) \coloneqq V$ and edge set $E(G) \coloneqq E$. A graph is vertex or edge labeled if its vertices or edges are equipped with a labeling function respectively. If it is both vertex and edge labeled, we simply call the graph labeled. We write $l(x)$ for the vertex labels ($x \in V(G)$) and edge labels ($x \in E(G)$).

Given two (un)directed graphs G and G' and a bijection $\varphi : V(G) \to V(G')$, we say that φ is edge-preserving if $(v, u) \in E(G)$ if and only if $(\varphi(v), \varphi(u)) \in E(G')$. Additionally, if G and G' are labeled, φ is label-preserving if $l(v) = l(\varphi(v))$ for any $v \in V(G)$ and $l(v, u) = l(\varphi(v), \varphi(u))$ for any $(v, u) \in E(G)$. The bijection φ is an isomorphism if it is edge-preserving and, in the case that G and G' are labeled, label-preserving. If $G = G'$, then φ is also an automorphism.

Given a (directed) graph G we call G (strongly) connected if there exists a path from any vertex u to any vertex v. We call the subgraph H of G a (strongly) connected component if H is a maximal (strongly) connected subgraph.

Since the motivation of this work is rooted in chemistry, sometimes it is more natural to talk about the undirected labeled graphs as molecules, their vertices as atoms (with labels defining the atom type), and their edges as bonds (whose labels distinguish single, double, triple, and aromatic bonds, for instance), while still using common graph terminology for mathematical precision.

Graph Transformations: As molecules are modeled as undirected labeled graphs, it is natural to think of chemical reactions as graph transformations, where a set of educt graphs are transformed into a set of product graphs. We model such transformations using the double pushout (DPO) approach. For a detailed overview of the DPO approach and its variations see (Habel, Müller, and Plump 2001). Here, we will use DPO as defined in (Andersen et al. 2016) that specifically describes how to model chemical reactions as rules in a DPO framework.

A rule p describing a transformation of a graph pattern L into a graph pattern R is denoted as a span $L \overset{l}{\leftarrow} K \overset{r}{\rightarrow} R$, where K is the subgraph of L remaining unchanged during rewriting and l and r are the subgraph morphism K to L and R respectively. The rule p can be applied to a graph G if and only if (i) L can be embedded in G (i.e., L is subgraph monomorphic to G) and (ii) the graphs D and H exists such that the diagram depicted in Fig. 1 commutes.

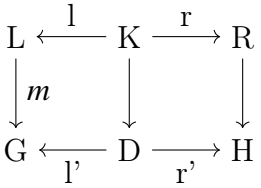


Figure 1: A direct derivation.

The graphs D and H are unique if they exist (Habel, Müller, and Plump 2001). The graph H is the resulting graph obtained by rewriting G with respect to the rule p. We call

the application of p on G to obtain H via the map $m : L \to G$, a direct derivation and denote it as $G \overset{p,m}{\Longrightarrow} H$ or $G \overset{p}{\Longrightarrow} H$, if m is not important. We note, that m is not necessarily unique, i.e., there might exist a different map m' such that $G \overset{p,m'}{\Longrightarrow} H$.

For a DPO rule p to model chemistry, we follow the modeling in (Andersen et al. 2013), and impose 3 additional conditions that p must satisfy. (i) All graph morphisms must be injective (i.e., they describe subgraph isomorphisms). (ii) The restriction of graph morphisms l and are r to the vertices must be bijective, ensuring atoms are conserved through a reaction. (iii) Changes in charges and edges (chemical bonds) must conserve the total number of electrons.

In the above framework, a chemical reaction is a direct derivation $G \overset{p,m}{\Longrightarrow} H$, where each connected component of G and H corresponds to the educt and product molecules, respectively. Condition (i) and (ii), ensures that l and r, and by extension l' and r' are bijective mappings when restricted to the vertices. As a consequence we can track each atom through a chemical reaction modeled as a direct derivation by the map $l'^{-1} \circ r'$. We note, that like m, l' and r' might not be unique for a given direct derivation $G \overset{p}{\Longrightarrow} H$. We define the set of all such maps $l'^{-1} \circ r'$ for all possible maps l' and r' obtained from $G \overset{p}{\Longrightarrow} H$ as $tr(G \overset{p}{\Longrightarrow} H)$. An example of a direct derivation representing a chemical reaction is depicted in Fig. 2.

Chemical Networks: We consider a directed hypergraph where each edge $e = (e^+, e^-)$ is a pair of subsets of vertices. Moreover, we let $Y_e = e^+ \cup e^-$ denote the set of vertices that are comprised in the start-vertex e^+ and end-vertex e^- of e. In short, a chemical network CN is a hypergraph where each vertex is a connected graph representing a molecule and each hyper-edge a rule application corresponding to a chemical reaction. Hence, every hyper-edge e of CN corresponds to a set of direct derivations transforming the in-going vertices of e into its out-going vertices. For a given set of edges E of CN, let $\mathcal{D}$ be the set of all direct derivations

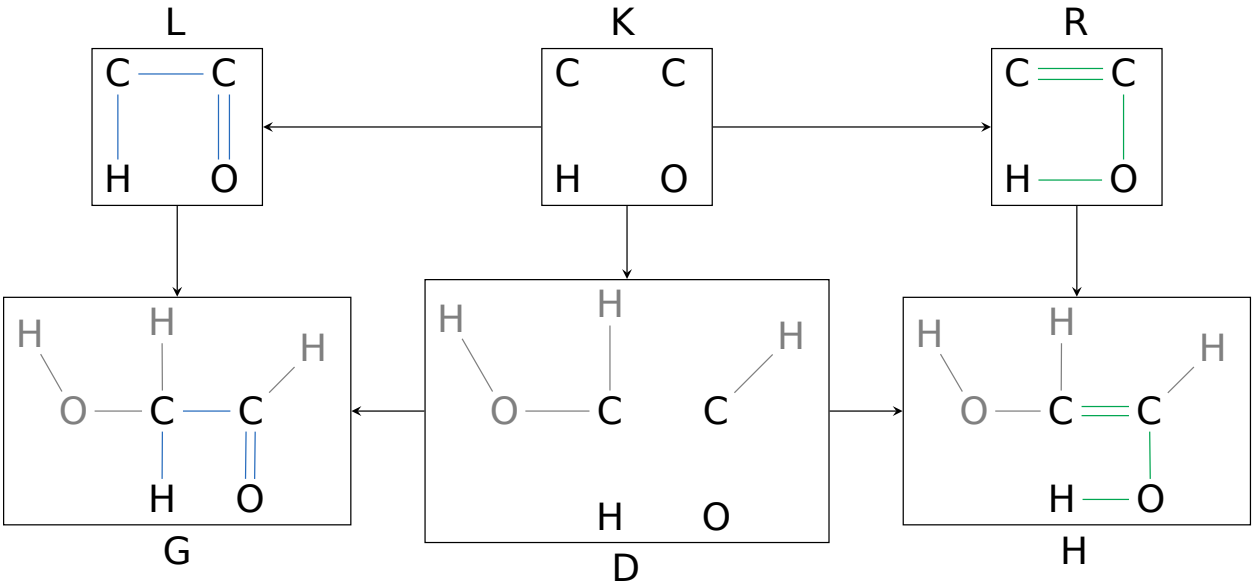


Figure 2: An example of a direct derivation. The mapping l, r, l' and r' is implicitly given by the depicted positions of the atoms. Given a chemical network, each hyper-edge directly corresponds to such a direct derivation.

that can be obtained from E. Then, $tr(E) = \bigcup_{G \overset{p}{\Longrightarrow} H \in \mathcal{D}} tr(G \overset{p}{\Longrightarrow} H)$ and $tr(\mathrm{CN}) = tr(E(CN))$.

Semigroups and transformation semigroups: A *semigroup* is a pair $(S, \circ)$, where S is a set and $\circ : S \times S \to S$ an associative binary operator on S. We often write ab for the product $a \circ b$. A semigroup that contains the identity element 1 (i.e., $s1 = s = 1s$ for all $s \in S$) is a *monoid*. The *order* of a semigroup S is its cardinality $|S|$. A subset $A \subseteq S$ is said to *generate* S or called a generating set for S, $\langle A \rangle = S$, if all elements of S can be expressed as a finite product of elements in A.

Given a non-empty finite set X, a *transformation* on X is an arbitrary map $f : X \to X$ that assigns to *every* element $x \in X$ some element $f(x) \in X$. The identity of a transformation on X is denoted 1_X. A transformation monoid is a transformation semigroup with identity. If $X = \{1, \ldots, n\}$, we often use the notation $(i_1, i_2, \ldots, i_n)$ for the transformation $f(j) = i_j$,

$1 \leq j \leq n$. Note, the elements $i_1, i_2, \ldots, i_n$ need not necessarily be pairwise distinct. Let T be the set of all possible transformations on X. If $S \subseteq T$ and S is closed under function composition $\circ$, then $(S, \circ)$ forms a semigroup, also called a *transformation semigroup*. To emphasize that S is a collection of transformations on X, we will use the notation (X, S) for transformation semigroups and say that S *acts on* X. Given a tuple $\bar{z} = (z_1, z_2, \ldots, z_n)$ of n distinct elements of X and a transformation semigroup (X, S), the orbit of $\bar{z}$ is defined as $O(\bar{z}, S) = \{(s(z_1), \ldots, s(z_n) \mid s \in S\}$. In what follows, we use the notion $y \in t = (i_1, i_2, \ldots, i_n)$ to indicate that $y = i_j$ for some j, $1 \leq j \leq n$.

Given a transformation semigroup (X, S) with generating set A, in symbols $S = \langle A \rangle$, we will employ the *(right) Cayley graph* $\mathrm{Cay}(S, A)$ of S and A with vertex set S and edge set $E(\mathrm{Cay}(S, A)) = \{(s, sa) \mid s \in S, a \in A\}$. In addition, every edge (s, sa) of $\mathrm{Cay}(S, A)$ obtains label l_a, that is, the unique label that is associated to each generator a in A. Similarly, the *projected* Cayley graph $\mathrm{PCay}(S, A, \bar{z})$ is defined for tuples $\bar{z}$: It has vertex set $O(\bar{z}, S)$ and for all $s \in O(\bar{z}, S)$ and for all $a \in A$ there is an edge (s, sa) with label l_a. A *free semigroup* Σ^+, is the semigroup containing all finite sequences of strings constructed from the alphabet Σ with concatanation as the associative binary operator. Adding the empty string ϵ results in the free monoid $\Sigma^* = \Sigma^+ \cup \{\epsilon\}$.

3 Chemical Networks and their Algebraic Structures

3.1 Characteristic Monoids

Assume we are given some chemical network CN that is some hypergraph modeling some chemistry. As we are interested in tracking the possible movements of atoms in CN, we are inherently interested in the reactions of CN, i.e., in its edge set $E(\mathrm{CN})$. Indeed, atoms can

only reconfigure to construct new molecules under the execution of some reaction. We will refer to the execution of a reaction as an *event*. The possible reconfigurations of atoms caused by a single event, is given by the set of atom maps $tr(\mathrm{CN})$ constituting a set of (partial) transformations on $X = \bigcup_{M \in V(\mathrm{CN})} V(M)$. Note, the vertex $M \in V(\mathrm{CN})$ corresponds to an entire molecule for which $V(M)$ denotes the set of atoms (=labeled vertices). A transformation t on X describes the position (i.e., in what molecule and where in the molecule the atom is found) of each atom in X when X is transformed by t. In what follows, we will sometimes refer to such transformations on X as *atom states*, as the transformations encapsulates the "state" of the network, i.e., the position of each atom. To track the possible movement of atoms through a chemical network, we must consider sequences of events.

Definition 1 (Event Traces)**.** *Let Σ be an alphabet containing a unique identifier t for each atom map in $tr(\mathrm{CN})$. Then, an* event trace *is an element of the free monoid Σ^*.*

The free monoid Σ^* contains all possible sequences of events that can move the atoms of X. Note, Σ^* does not track the actual atoms through event traces. For this, we use the following structure:

Definition 2 (Characteristic Monoids)**.** *Let the characteristic monoid of* CN *be defined as the transformation monoid $\mathcal{S}(\mathrm{CN}) = (X, \langle tr(\mathrm{CN}) \cup 1_X \rangle)$. Moreover, given a set of edges $E \subseteq E(\mathrm{CN})$, and the set of atoms $Y \subseteq X$ found in E (that is $Y = \cup_{e \in E} Y_e$), we let the characteristic monoid of E be defined as $\mathcal{S}(E) = (Y, \langle tr(E) \cup 1_Y \rangle)$.*

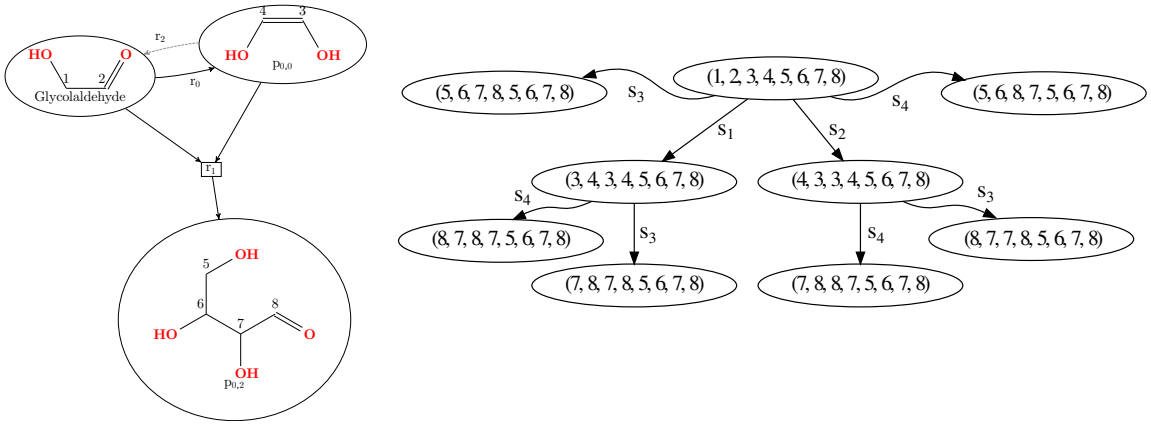
Let $\sigma : \Sigma \to tr(\mathrm{CN})$ be the function, that maps all identifiers of Σ to their corresponding atom map in $tr(\mathrm{CN})$. Given an event trace $t = t_1 t_2 \dots t_n \in \Sigma^*$, we let the events of t refer to their corresponding transformations in $tr(\mathrm{CN})$ when acting on an element $s \in \mathcal{S}(\mathrm{CN})$, i.e., $st = s\sigma(t_1)\sigma(t_2)\dots\sigma(t_n) \in \mathcal{S}(\mathrm{CN})$. Every event trace $t \in \Sigma^*$ gives rise to a member $\mathcal{S}(\mathrm{CN})$,

in particular the transformation $1_X t$, that represents the resulting atom state obtained from moving atoms according to t. Hence, there is a homomorphism from Σ^* to $\mathcal{S}(\mathrm{CN})$, meaning that $\mathcal{S}(\mathrm{CN})$ captures all possible movements of atoms through reactions of CN.

Often, we are only interested in tracking the movement of a small number of atoms. Let $\bar{z}$ be a tuple of distinct elements from X that we want to track. Then, there is again a homomorphism from Σ^* and $\mathcal{O}(\bar{z}, \mathcal{S}(\mathrm{CN}))$. Namely, for a given event trace $t \in \Sigma^*$, we can track the atoms of $\bar{z}$ as the atom state $1_{\{x \mid x \in \bar{z}\}} t$ corresponding to an element in the orbit $\mathcal{O}(\bar{z}, \mathcal{S}(\mathrm{CN}))$, if we treat the element as a (partial) transformation. As a result, $\mathcal{O}(\bar{z}, \mathcal{S}(\mathrm{CN}))$ characterizes the possible movements of the atoms in $\bar{z}$, and we will refer to its elements as atom states similarly to elements in $\mathcal{S}(\mathrm{CN})$ as they conceptually represent the same thing.

We note, the above definitions are not unlike some of the core definitions within algebraic automata theory (Mikolajczak 1991). Here, the possible inputs of an automata is often defined in terms of strings obtained from the free monoid on the alphabet of the automata. The characteristic semigroup is then defined as the semigroup that characterizes the possible state transitions. In the same vein, we can view our notion of event traces as the possible "inputs" to our chemical network CN that moves some initial configuration of atoms 1_X. The characteristic monoid of CN then characterize the possible movements of atoms through event traces.

In what follows we let $\mathrm{Cay}(\mathrm{CN})$ denote the Cayley graph $\mathrm{Cay}(\mathcal{S}(\mathrm{CN}), tr(\mathrm{CN}) \cup 1_X)$. Similarly, given a tuple of atoms $\bar{z}$, we let $\mathrm{PCay}(\mathrm{CN}, \bar{z})$ denote the projected Cayley graph $\mathrm{PCay}(\mathcal{S}(\mathrm{CN}), tr(\mathrm{CN}) \cup 1_X, \bar{z})$. We note, that by Def. 2, $\mathcal{S}(\mathrm{CN})$ is constructed from the generating set $\langle tr(\mathrm{CN}) \cup 1_X \rangle$, and hence $\mathrm{Cay}(\mathrm{CN})$ and $\mathrm{PCay}(\mathrm{CN}, \bar{z})$ are well defined. Since the transformation 1_X will always result in a loop on every vertex of the (projected) Cayley graph, and conveys no meaningful information, we will refrain from including any edge arising



(a) (b)

Figure 3: (a) A small example using molecules and reactions found in the Formose reaction. The carbon atoms of each molecule are labeled with a unique identifier for easy reference. (b) The Cayley graph Cay(CN) of Fig. 3a from the example of Sec. 3.1. From the graph, we see the longest path from 1_X has length 2, meaning that any event trace can at most transform 1_X meaningfully twice. In fact, only two types of event traces are of interest: Either the tracked atoms are immediately moved by the reaction r_1 to $p_{0,2}$, or the atoms of glycolaldehyde are first moved to $p_{0,0}$ using r_0, and then moved to $p_{0,2}$.

from 1_X.

We can illustrate the relation between atom states using the Cayley graph Cay(CN). More precisely, there exists an edge between two atom states $a, b \in \mathcal{S}(\mathrm{CN})$ with label t, if it is possible to move the atoms in a to b using t. It is natural to relate Σ^* to Cay(CN). Namely, any path in Cay(CN) corresponds directly to an event trace in Σ^*. Hence, where Σ^* encapsulates the "inputs" of the chemical network and $\mathcal{S}(\mathrm{CN})$ contains the possible atom states derived from Σ^*, the Cayley graph Cay(CN) captures *how* atom states from $\mathcal{S}(\mathrm{CN})$ can be created by event traces.

Example: As an illustrative example, consider the reaction network CN depicted in Fig. 3a. For simplicity we will use reactions r_0 and r_1 involved in the so-called Formose reaction. We restrict ourselves to only consider the carbon atoms of all molecules, and have labeled them with a corresponding unique id for easy reference. Here, the underlying set

$X = \{1, 2, \ldots, 8\}$ corresponds to the eight elements labeled by $1, 2, \ldots, 8$ in Fig. 3a. From $tr(\mathrm{CN})$ we get 4 transformations: $s_1 = [3, 4, 3, 4, 5, 6, 7, 8]$, $s_2 = [4, 3, 3, 4, 5, 6, 7, 8]$ (both obtained from r_0), and $s_3 = [5, 6, 7, 8, 5, 6, 7, 8]$, $s_4 = [5, 6, 8, 7, 5, 6, 7, 8]$ (both obtained from r_1) with the corresponding alphabet $\Sigma = \{s_1, s_2, s_3, s_4\}$. For a reaction, the corresponding transformation(s) maps the atoms of the educt molecules to the atoms of the product molecules while all other atoms are mapped with the identity. The transformations describe how carbon atoms are rearranged into different configurations when an event is fired. s_1 and s_2 describe how the carbon atoms of a glycolaldehyde molecule are arranged in the molecule $p_{0,0}$ when transformed via the reaction r_0. In the case of s_1, we see that the carbons are rearranged such that $s_1(1) = 3$ and $s_1(2) = 4$. Of course, due to the symmetries in the molecule $p_{0,0}$, reaction r_0 also results in the mirrored transformation of s_1, i.e., $s_2(1) = 4$ and $s_2(2) = 3$. The characteristic monoid of CN, $\mathcal{S}(\mathrm{CN})$, has an order of 9. We illustrate the movement of atoms in CN by its Cayley graph Cay(CN) which is depicted in Fig. 3b. Any path originating from the identity element corresponds to an event trace, e.g. we can track the atoms 1 and 2 through the event trace $s_1 s_3$ as the corresponding path and realize $s_1 s_3(1) = 8$ and $s_1 s_3(2) = 7$. Assume now, that we were only interested in tracking the carbon atoms found in the glycolaldehyde molecule. To this end, we can examine $\mathcal{O}(\bar{z}, \mathcal{S}(\mathrm{CN}))$, which contains 6 elements, meaning there exists 6 unique atom states for the atoms in a glycolaldehyde molecule. Again, we can study these movements using the projected Cayley graph $\mathrm{PCay}(\mathrm{CN}, (1, 2))$. The resulting graph is depicted in Fig. 4a.

3.2 Natural Subsystems of Atom States

In the intersection between group theory and systems biology, attempts to formalize the notion of natural subsystems and hierarchical relations within such systems have been done

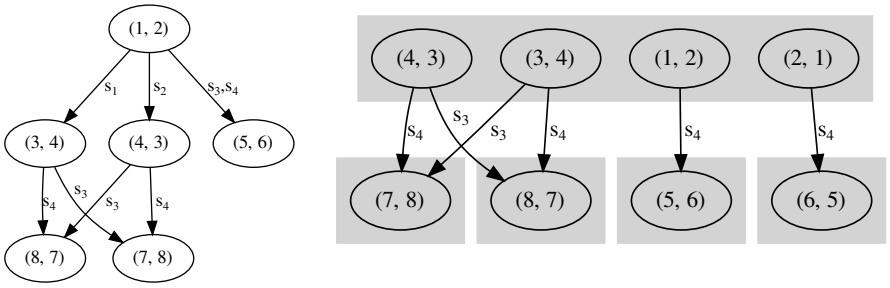


Figure 4: (a) The projected Cayley graph PCay(CN, (1, 2)) from the example of Section 3.1. Like for Cay(CN), we see there are only two types of event traces of interest. However, since we are only tracking the atoms of the glycolaldehyde molecule, some atom states are effectively coalesced compared to Cay(CN). (b) The projected Cayley graph PCay(CN, (1, 2)) from the example of Sec. 3.2. The graph shows the natural subsystems of the carbon atoms of a glycolaldehyde molecule. Vertices in the same box constitutes vertices that are in the same natural subsystem. Note, edges between vertices in the same natural subsystem are not depicted (e.g, one of the eight hidden edges in the top-level subsystem is $(3, 4) \to (1, 2)$ with label s_5).

by works such as (Nehaniv et al. 2015). Here, natural subsystems are defined as symmetric structures arising from a biological system. Such symmetries manifests as permutation groups of the associated semigroup representing said system. In such a model the Krohn-Rhodes decomposition or the holonomy decomposition (Egri-Nagy and Nehaniv 2015) can be used to construct a hierarchical structure on such natural subsystems of the biological system. In terms of atom tracking, however, defining natural subsystems in terms of the permutation groups in $\mathcal{S}(\mathrm{CN})$ does not have an immediately useful interpretation. Similarly, the hierarchical structure obtained from methods such as holonomy decomposition are not intuitive to interpret. Instead, when talking about natural subsystems in terms of atom tracking, we are interested in systems of reversible *event traces*, i.e., event traces that do not change the original configuration of atoms. To this end, it is natural to define natural subsystems of $\mathcal{S}(\mathrm{CN})$ in terms of Green's relations (Clifford and Preston 1967). For elements $s_1, s_2 \in \mathcal{S}(\mathrm{CN})$, we define the reflexive transitive relation $\succeq_{\mathcal{R}}$ as $s_1 \succeq_{\mathcal{R}} s_2$, if there exists an

event trace $t \in \Sigma^*$ such that $s_1 t = s_2$. In addition, we define an equivalence relation $\mathcal{R}$, where s_1 is equivalent to s_2, in symbols $s_1 \mathcal{R} s_2$ whenever $s_1 \succeq_{\mathcal{R}} s_2$ and $s_2 \succeq_{\mathcal{R}} s_1$.

Definition 3 (Natural Subsystems)**.** *The natural subsystems of* $\mathcal{S}(\mathrm{CN})$ *is the set of equivalence classes induced by the* $\mathcal{R}$*-relation.*

The equivalence classes correspond to the strongly connected components of the Cayley graph Cay(CN) (Froidure and Pin 1997). We note, that for a tuple of atoms $\bar{z}$, the natural extension to natural subsystems of the orbit $O(\bar{z}, \mathcal{S}(\mathrm{CN}))$ is simply the strongly connected components of its projected Cayley graph $\mathrm{PCay}(\mathrm{CN}, \bar{z})$. The $\mathcal{R}$ relation is interesting, as the equivalence classes on $\mathcal{S}(\mathrm{CN})$ induced by the $\mathcal{R}$ relation forms pools of reversible event traces. More precisely, let $s_1 \mathcal{R} s_2$ for some $s_1, s_2 \in \mathcal{S}(\mathrm{CN})$, where $s_1 \cdot t_{12} = s_2$ and $s_2 \cdot t_{21} = s_1$ for some $t_{12}, t_{21} \in \Sigma^*$. Then, the event traces t_{12} and t_{21} are reversible, i.e. we can re-obtain s_1 as $s_1 t_{12} t_{21} = s_1$ and s_2 as $s_2 t_{21} t_{12} = s_2$. Additionally, the quotient graph of the equivalence classes of the $\mathcal{R}$ relation on the Cayley graph Cay(CN) naturally forms a hierarchical relation on the atom states of $\mathcal{S}(\mathrm{CN})$ that has a useful interpretation from the point of view of chemistry as we will see in Sec. 4.3.

Example: Again, consider the reaction network obtained from the formose reaction depicted in Fig. 3a. We will include the transformations obtained from reaction r_2 in additions to the transformations listed in Sec. 3.1: $s_5 = [1, 2, 1, 2, 5, 6, 7, 8]$ and $s_6 = [1, 2, 2, 1, 5, 6, 7, 8]$ (both obtained from r_2). Assume we are interested in determining how carbon atoms of a glycolaldehyde molecule can reconfigure into different molecules. The projected Cayley graph $\mathrm{PCay}(\mathrm{CN}, (1, 2))$ shows such configurations and is depicted in Fig. 4b. Here, the atom states belonging to the same gray box are strongly connected and hence belong to the same natural subsystem. For clarity, we have removed edges between atom states in the same subsystem, since any atom state in a subsystem can be transformed into any other state in the same

subsystem.

Notably, we see from Fig. 4b that the atoms 1 and 2 in the glycolaldehyde molecule can swap positions. We could of course also realize that such a swap was possible by noticing the symmetries in the glycolaldehyde molecule and the fact that we can convert glycolaldehyde to the $p_{0,0}$ molecule and vice versa. However, such patterns becomes immediately obvious from the projected Cayley graph. Finally, we can derive from Fig. 4b, that it is only possible to leave the original subsystem by applying transformation s_3 or s_4, corresponding to reaction r_1.

4 Results

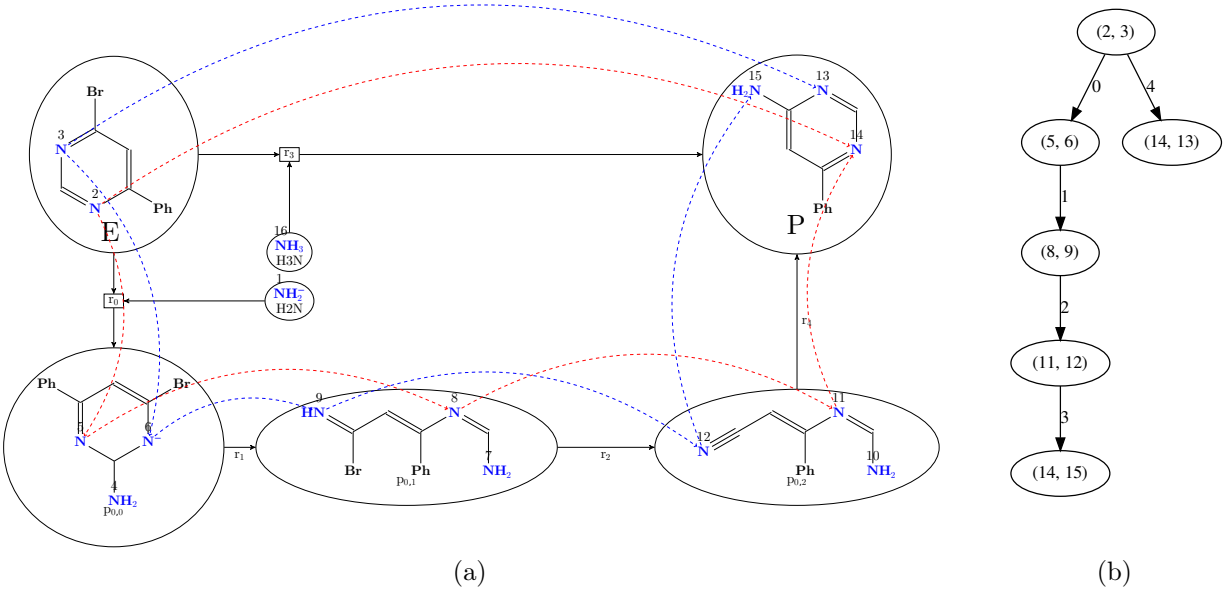
4.1 Implementation

To test the practicality of the structures introduced in the previous section, we implemented the construction of the projected Cayley graph of a set of atoms in a chemical network. The resulting implementation can be found at `https://github.com/Nojgaard/cat` All code is written in python and uses the software package MØD (Andersen et al. 2016) and NetworkX (Hagberg, Schult, and Swart 2008) to construct the chemical networks and find the transformations used for the characteristic monoid. All figures in the following section were constructed with said implementation, and each run finished within seconds on an 8 core Intel Core i9 CPU with 64 GB memory. The most time consuming part of the implementation was the computation of the transformations obtained from each hyper-edge in the chemical network. In contrast, the construction time of the projected Cayley graph proved to be negligible.

4.2 Differentiating Pathways

In this section, we will explore the possibilities of using the characteristic monoids of chemical networks to determine if it is possible to distinguish between two pathways P_1 and P_2, based on their atom states of their respective characteristic monoids. The motivation stems from methods such as isotope labeling. Here, a "labeled" atom, is a detectable isotope whose position is known in some initial molecule and can then be detected, along with its exact position, in the product molecules of some pathway. In contrast to (Andersen, Merkle, and Rasmussen 2019), we will not focus on the orbits of atoms in isolation, as we lose the ability to reason about atom positions in relation to each other. Moreover, as we will see here, the Cayley graph of the chemical network can be used to identify the exact event two pathways split.

Given a chemical network CN, a pathway P is a set of hyper-edges (i.e. reactions) from CN equipped with a set of input and output molecules. We think of a pathway as a process that consumes a set of input molecules to construct a set of output molecules, using the reactions specified by P. In our case, a "labeled" atom is a point in $\mathcal{S}(\text{CN})$. Given two pathways P_1 and P_2, we can characterize the possible movement of atoms as the characteristic monoids $\mathcal{S}(P_1)$ and $\mathcal{S}(P_2)$. In practice, it might not be feasible to track every atom in CN, e.g. we are only able to replace a few atoms with its corresponding detectable isotope, and hence it becomes useful to consider the orbits $O(\bar{z}, \mathcal{S}(P_1))$ and $O(\bar{z}, \mathcal{S}(P_2))$ where $\bar{z}$ is the atoms from the input molecules we can track. Clearly, of the atom states in $O(\bar{z}, \mathcal{S}(P_1))$ and $O(\bar{z}, \mathcal{S}(P_2))$, we can only expect to observe, e.g. in an isotope labeling experiment, the atom states that locates the tracked atoms in the output molecules. As a result, we arrive at the following observation:



(a) (b)

Figure 5: (a) The chemical network for the creation of P from E using ammonia. The dotted red and blue lines shows the possible atom trajectories for the atoms 2 and 3 respectively. (b) The projected Cayley graph $\mathrm{PCay}(\mathrm{CN}, (2,3))$.

Observation 1. *Let $Y_i \subseteq O(\bar{z}, \mathcal{S}(P_i))$, $i \in \{1,2\}$, be the atom states we can hope to observe after some isotope labeling experiment. Then, we can always distinguish between P_1 and P_2 if $Y_1 \cap Y_2 = \emptyset$.*

Example: Consider the network CN depicted in Fig. 5a modelling the creation of product 4-phenyl-6-aminopyrimidine (denoted P) from the educt 4-(benzyloxy)-6-bromopyrimidine (denoted E) using ammonia. This well investigated and widely used substitution mechanism (ANRORC) (Plas 1978) was proven to non-trivially function via ring opening and ring closure (and an accompanied carbon replacement) via isotope labeling. Two possible pathways are modelled: the input molecules for the two pathways are the molecules E, NH_3, NH_2, while the output is the single molecule P. The first, seemingly correct but wrong, pathway $P_1 = \{r_3\}$ converts E and an NH_3 molecule directly into P, by replacing the Br atom with NH_2.

The second pathway consists of the reactions $P_2 = \{r_0, r_1, r_2, r_4\}$ and models the ANRORC mechanism.

Assume we wanted to device a strategy to decide what pathway is executed in reality. By replacing the nitrogen atoms of the E molecule with the isotope ^{13}N we would be able to observe where the atoms are positioned in the produced P molecule. Since we, by assumption, only label the nitrogen atoms of the E molecule, i.e., the atoms 3 and 2, we can look at the orbits of the characteristic monoids $\mathcal{O}((2,3), \mathcal{S}(P_1))$ and $\mathcal{O}((2,3), \mathcal{S}(P_2))$ with the order of 5 and 2 respectively. We see that both orbits only contains a single element locating $(2,3)$ in the P molecule, namely the element $(14,15)$ for $\mathcal{O}((2,3), \mathcal{S}(P_1))$ and $(14,13)$ for $\mathcal{O}((2,3), \mathcal{S}(P_2))$. As the possible configurations are different for P_1 and P_2, it is hence possible to always identify if the P molecule was created by P_1 or P_2.

This fact, also becomes immediately obvious by looking at the projected Cayley graph PCay(CN, $(2,3)$) depicted in Fig. 5b, that shows the immediate divergence of atom states of the two pathways.

4.3 Natural Subsystems in the TCA Cycle

The citric acid cycle, also known as the tricarboxylic (TCA) cycle or the Krebs cycle, is at the heart of many metabolic systems. The cycle is used by aerobic organisms to release stored energy in the form of ATP by the oxidation of acetyl-CoA into water and CO_2. The details for the TCA cycle can be found in any standard chemistry text book, e.g. (Harvey and Ferrier 2010). In (Smith and Morowitz 2016), the trajectories of different carbon atoms in the TCA cycle was examined to explain the change of their oxidation states. It is well known that there is an enzymatic differentiation of the two carboxymethyl groups in citrate, which requires a rigorous stereochemical modeling of the graph grammar rules used (Andersen et al. 2017).

Ignoring such stereochemical modeling would lead to atom mappings not occurring in nature. We will provide a formal handle to analyze theoretically possible carbon trajectories using the algebraic constructs provided in this paper. As we will see, such structures provides intuitive interpretations for the TCA cycle. More precisely, assume we are interested in answering the following questions: What are the possible trajectories of the carbons of an oxaloacetate (OAA) molecule within the TCA cycle while i.) ignoring the enzymatic differentiation of the two carboxymethyl groups in citrate (denoted TCA-□), or ii.) not ignoring (denoted TCA-◬). To answer these questions, we will decompose the characteristic monoid of the TCA cycle into its natural subsystems and examine them using the projected Cayley graph.

In our setting, the TCA cycle is the chemical network CN,depicted in Fig. 6, giving rise to transformations of the underlying monoid. The network is made up of 13 reactions, however, some of the reactions are not shown for simplicity. Of these 13 reactions, 7 of them yields exactly 1 transformation each while the remaining 6 yields 2 possible transformations each, resulting in a total of 19 transformations found. The reactions containing multiple transformations are due to automorphisms in molecules such as citrate and fumarate. When the enzymatic differentiation of the carboxymethyl group in citrate is not ignored, only 4 of the 13 reactions yield 2 possible transformations, as the carbon traces to and from citrate are more constrained. In short, while both TCA-□ and TCA-◬ are modeled by the same network, the obtained transformations differ. More precisely, $|tr(\mathrm{CN})| = 19$ wrt. TCA-□ and $|tr(\mathrm{CN})| = 17$ wrt. TCA-◬.

To start the cycle, an Acetyl-CoA molecule is condensed with an OAA molecule, executing a cycle of reactions that ends up regenerating the OAA molecule while expelling two CO_2 and water on the way. When an original atom is expelled from the cycle, we will consider it permanently lost. The carbon atoms of the OAA molecule that we are interested

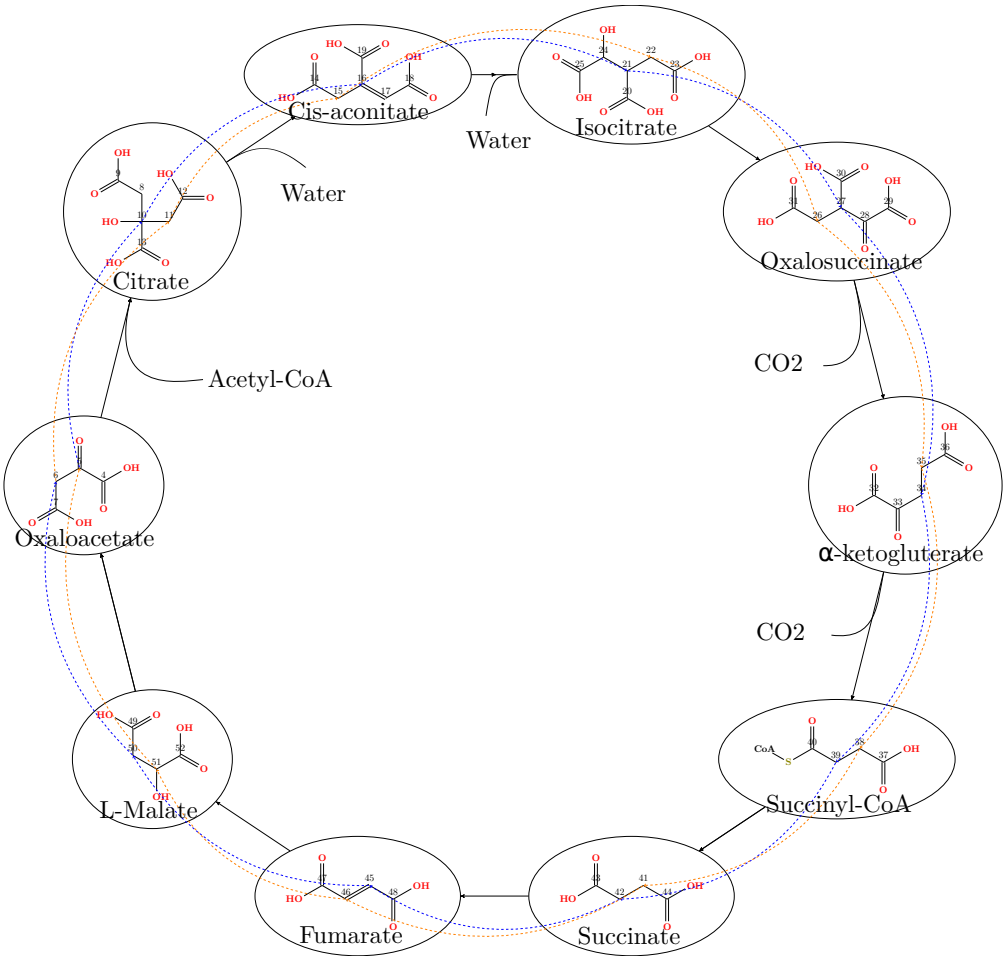
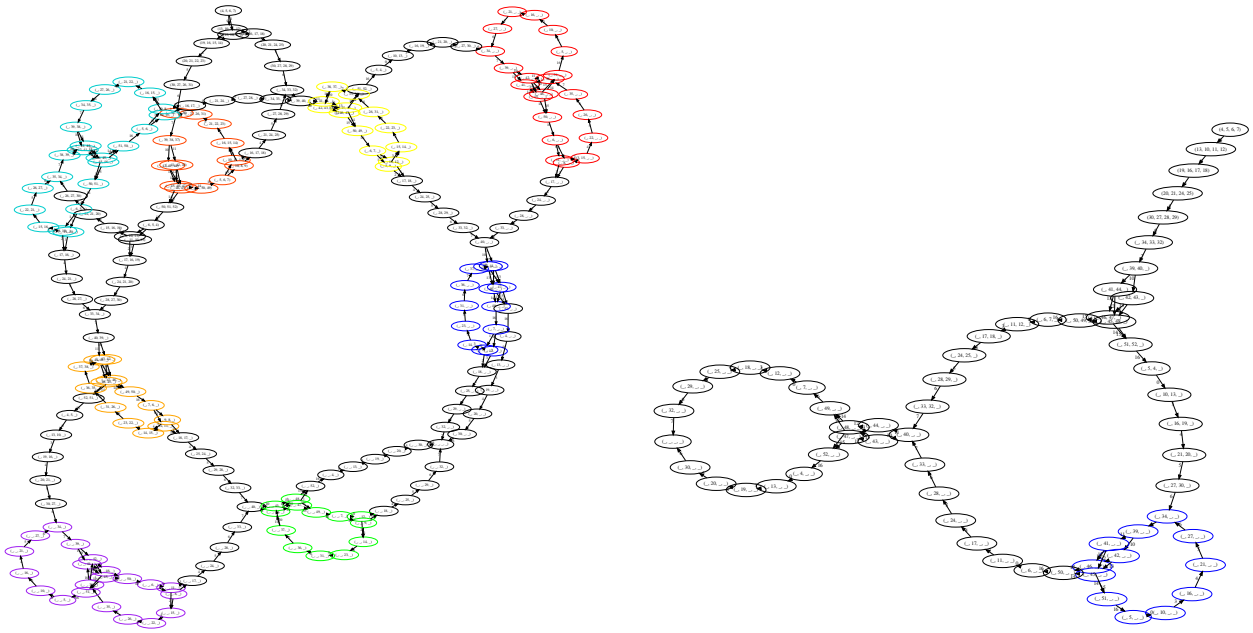


Figure 6: A (simplified) chemical network modelling the TCA cycle. Note, any molecules not containing carbon atoms are modelled, but not depicted here. Each carbon atom is equipped with a unique id for easy reference.



(a) (b)

Figure 7: (a) The projected Cayley graph PCay(CN, (4, 5, 6, 7)) wrt. TCA-□. The non black colored vertices of the same color, correspond to atom states that are part of the same strongly connected component. (b) The projected cayley graph PCay(CN, (4, 5, 6, 7) wrt. TCA-⏢. The non black colored vertices of the same color, correspond to atom states that are part of the same strongly connected component.

in tracking are annotated with the ids 4, 5, 6, and 7. Let $\bar{z} = (4, 5, 6, 7)$. The projected Cayley graph of PCay(CN, $\bar{z}$) wrt. TCA-□ (resp. TCA-⏢) , consists of 213 (resp. 67) vertices. The full Cayley graphs are depicted in Fig. 7a and 7b respectively. When a carbon atom leaves the TCA cycle we denote it by "_". E.g. the atom state $(\_, 7, 6, \_)$ should be read as the original carbon atoms with ids 4 and 7 has been expelled, while the carbon atoms with ids 5 and 6 are located at the atoms with id 7 and 6 respectively.

We can find the natural subsystems of CN as the strongly connected components of PCay(CN, $\bar{z}$). In TCA-□ (resp. TCA-⏢) we find 92 (resp. 51) strongly connected components of which 8 (resp. only 1) are non-trivial. Any non-trivial strongly connected component

must invariably contain at least one tour around the TCA cycle, since this is the only way the original atoms of the OAA molecules can be reused to create another OAA molecule. Moreover, any non-trivial strongly connected component represents a sequence(s) of reactions that uses (some of the) original atoms of the OAA molecule. To simplify PCay(CN, $\bar{z}$) such that only the information on carbon traces of the atoms of OAA are depicted, we will construct the simplified projected Cayley graph, denoted SCay(CN, $\bar{z}$), as follows: collapse any vertex in PCay(CN, $\bar{z}$) that is part of a trivial strongly connected component and whose atoms are not located in an OAA molecule. Moreover, for any non-trivial strongly connected component, hide the edges between atom states in the same strongly connected component, and finally only include atom states if the atoms are located in a OAA molecule. The resulting graphs for TCA-□ and TCA-⏢ are depicted in Fig. 8. Each box in the figure represents a natural subsystem that contains an atom state where every atom is either expelled or located in an OAA molecule. When ignoring the stereochemical formation of citrate, $(\_, 5, 6, 7)$ is a grey node in SCay(CN, $\bar{z}$) (i.e., a representative of a strongly connected component PCay(CN, $\bar{z}$)), i.e., there is a trajectory where three of the four original carbons of OAA are re-used at the same location after a TCA-□ cycle turnover. However in TCA-⏢ only $(\_, 5, \_, \_)$ is a representative of a strongly connected component, i.e., only the carbon with id 5 of OAA can be kept at the same location when a multitude of TCA-⏢ turnovers are executed. If that carbon changes location it will leave the TCA cycle after exactly two more turnovers (the natural subsystems reachable from $(\_, 5, \_, \_)$ do not correspond to strongly connected components) via positions $5 \to 6 \to 4 \to \_$ or via $5 \to 6 \to 7 \to \_$. To the best of our knowledge such investigations have not been executed formally before.

Interestingly, SCay(CN, $\bar{z}$), as depicted in Fig. 8c, allows us to closely examine each of the possible carbon trajectories of TCA−□. E.g. the fact that the atom state $(\_, 6, 7, \_)$ is

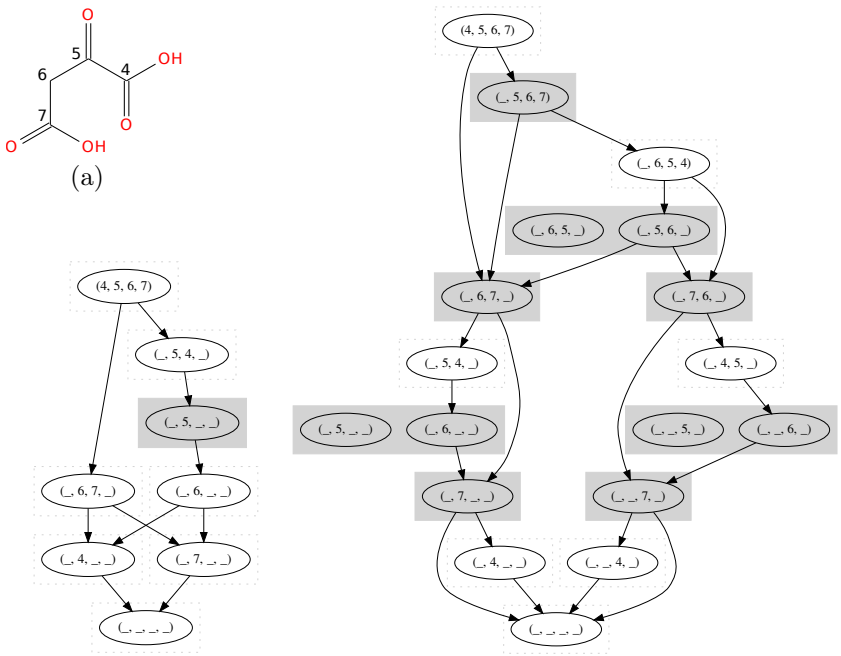


Figure 8: (a) The oxaloacetate molecule. The carbon atoms are equipped with the ids 4, 5, 6, and 7. (b) The simplified projected Cayley graph $\mathrm{SCay}(\mathrm{CN}, (4,5,6,7))$, when adjusting for stereospecific citrate in $tr(\mathrm{CN})$. (c) The simplified projected Cayley graph $\mathrm{SCay}(\mathrm{CN}, (4,5,6,7))$ when not considering stereospecificity.

present in $\mathrm{SCay}(\mathrm{CN}, \bar{z})$ wrt. TCA-$\square$, means that there exists a sequence of reactions that expels the carbons with ids 4 and 7, but re-uses the carbon atoms with id 5 and 6 to create a new OAA atom, where 5 is located at 6 and 6 is located at 7. Structurally the atoms 4 and 7 corresponds to the the outer atoms in the carbon backbone in the OAA molecule, while the atoms 5 and 6 correspond to the inner atoms in the carbon backbone. In other words the presence of $(\_, 6, 7, \_)$, means there exists a sequence of reactions that expels the outer atoms of the carbon backbone while recycling the inner atoms.

Fig. 8c, gives us a rough road map to determine exactly what sequence of events must have taken place in order to end up in the atom state $(\_, 6, 7, \_)$. We start with the atom

state $(4, 5, 6, 7)$ and see there is an edge directly to $(\_, 6, 7, \_)$, meaning that we can expel the two outer atoms in a single cycle. This is, however, not the only way we can end up with the atom state $(\_, 6, 7, \_)$. E.g. after one cycle we can expel the carbon with id 4 and end up with the atom state $(\_, 5, 6, 7)$, i.e., all other atoms are still in their original positions. After another cycle we can end up in the atom state $(\_, 6, 7, \_)$ or $(\_6, 5, 4)$. Note, that $(\_, 5, 6, 7)$ is part of a non-trivial strongly connected component, meaning that there exists a sequence of reactions in the TCA cycle that ends up in the exact same atom state. i.e., we expel the carbon atom at position 4 (which is already expelled) while keeping all other atoms at their original position. In contrast, the atom state $(\_, 6, 5, 4)$ is part of a trivial strongly connected component, meaning that any sequence of reaction in the TCA cycle will have to change the atom state.

If any non-trivial strongly connected component in Fig. 8c contains more than one vertex, it means that we can swap between atom states after a tour in the TCA cycle. As an example, consider the atom state $(\_, 6, 5, \_)$ and $(\_, 5, 6, \_)$ that are both part of the same strongly connected component. The fact that they are part of the same strongly connected component, means it is possible to swap the inner atoms of the carbon backbone during a TCA cycle. If we would be interested in the exact sequence of transformations that lead to the swap, we simply examine the subgraph of $\mathrm{PCay}(\mathrm{CN}, \bar{z})$ wrt. TCA-$\square$ corresponding to that natural subsystem of $\mathrm{SCay}(\mathrm{CN}, \bar{z})$ wrt. TCA-$\square$ as illustrated in Fig. 9. The figure depicts all possible ways to swap the positions of atoms with ids 5 and 6 as the possible paths between $(\_, 5, 6, \_)$ and $(\_, 6, 5, \_)$. Fig. 6, shows one such path traversing the TCA cycle without expelling any of the remaining carbon atoms.

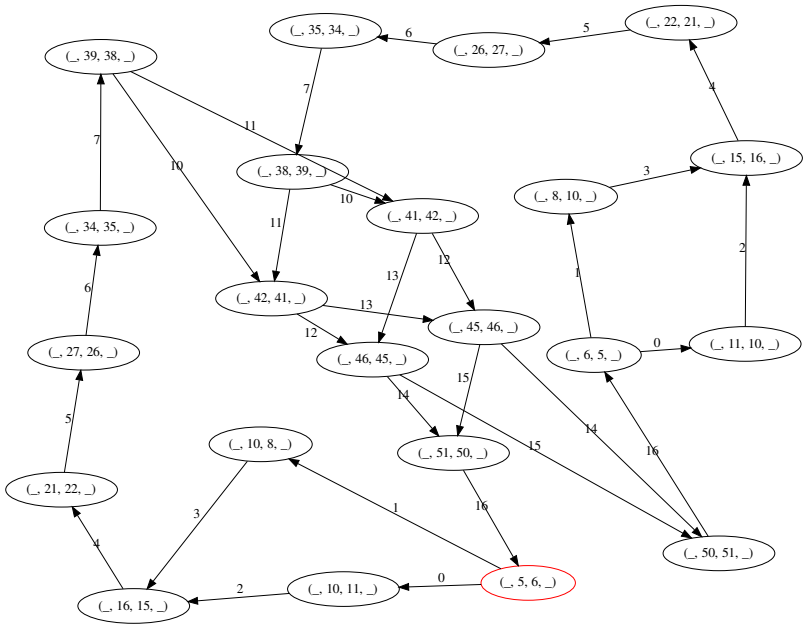


Figure 9: The strongly connected component of PCay(CN, $(4, 5, 6, 7)$) wrt. TCA-□ containing the state $(\_, 6, 5, \_)$ and $(\_, 5, 6, \_)$.

5 Conclusion

In this work we have extended the insights provided by (Andersen, Merkle, and Rasmussen 2019), by showing the natural relationship between event traces, the characteristic monoid and its corresponding Cayley graph. The projected Cayley graph provides valuable insights into local substructures of reversible event traces.

We see future steps for this approach to branch in at least two directions. On one hand, these methods shows obvious applications in isotopic labeling design. To this end, it is natural to extend the system to model the actual process of such experiments. E.g. when doing isotopic labeling experiments with mass spectrometry, molecules are broken into fragments and the weight of such fragments are deduced to determine the topology of the fragment. Using our model to track where the atoms might end up in such fragments and how it affects their weight seems like a natural next step. On the other hand, a more rigorous investigation of the fundamental properties derived from semigroup theory of the characteristic monoid seems appealing. As we have shown here, understanding such relations might grant insights into the nature of the examined system.

Acknowledgements

This work is supported in by Novo Nordisk Foundation grant NNF19OC0057834 and by the Independent Research Fund Denmark, Natural Sciences, grant DFF-0135-00420B.

Author Disclosure Statement

Nothing to declare.

References

Andersen, J. L, Flamm, C., et al. (2013). "Inferring chemical reaction patterns using rule composition in graph grammars". In: *Journal of Systems Chemistry* 4.1, p. 4.

Andersen, J. L., Flamm, C., et al. (2016). "A Software Package for Chemically Inspired Graph Transformation". In: *Graph Transformation - 9th International Conference, ICGT 2016, Proceedings*. Vol. 9761. LNCS. Springer, pp. 73–88.

Andersen, J. L., Flamm, C., et al. (2017). "Chemical Graph Transformation with Stereo-Information". In: *Graph Transformation - 10th International Conference, ICGT*. LNCS, pp. 54–69.

Andersen, J. L., Merkle, D., et al. (2019). "Graph Transformations, Semigroups, and Isotopic Labeling". In: *International Symposium on Bioinformatics Research and Applications*. Springer, pp. 196–207.

Chahrour, Osama, Cobice, Diego, et al. (2015). "Stable isotope labelling methods in mass spectrometry-based quantitative proteomics". In: *Journal of pharmaceutical and biomedical analysis* 113, pp. 2–20.

Clifford, Al.H. and Preston, G.B. (1967). *The algebraic theory of semigroups, Volume II*. Vol. 2. American Mathematical Soc.

Deev, Sergey L., Khalymbadzha, Igor A., et al. (2019). "15N labeling and analysis of 13C–15N and 1H–15N couplings in studies of the structures and chemical transformations of nitrogen heterocycles". In: *RSC Adv.* 9 (46), pp. 26856–26879. URL: `http://dx.doi.org/10.1039/C9RA04825A`.

Dénes, József (1966). *Connections Between Transformation-semigroups and Graphs*. Hungarian Academy of Sciences Central Research Institute for Physics.

Egri-Nagy, A. and Nehaniv, C. L (2008). "Hierarchical coordinate systems for understanding complexity and its evolution, with applications to genetic regulatory networks". In: *Artificial Life* 14.3, pp. 299–312.

Egri-Nagy, A. and Nehaniv, C.L. (2015). "Computational holonomy decomposition of transformation semigroups". In: *arXiv preprint arXiv:1508.06345*.

Froidure, V. and Pin, J.-E. (1997). "Algorithms for computing finite semigroups". In: *Foundations of Computational Mathematics*. Springer, pp. 112–126.

Habel, A., Müller, J., et al. (2001). "Double-pushout graph transformation revisited". In: *Mathematical Structures in Computer Science* 11.5, pp. 637–688.

Hagberg, Aric A., Schult, Daniel A., et al. (2008). "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, et al. Pasadena, CA USA, pp. 11 –15.

Harvey, R. and Ferrier, D. (2010). *Biochemistry (Lippincott's illustrated reviews series)*. Lippincott Williams & Wilkins, Baltimore, MD and Philadelphia, PA, USA.

Mikolajczak, Boleslaw (1991). *Algebraic and structural automata theory*. Elsevier.

Nehaniv, C. L, Rhodes, J., et al. (2015). "Symmetry structure in discrete models of biochemical systems: natural subsystems and the weak control hierarchy in a new model of computation driven by interactions". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 373.2046, p. 20140223.

Plas, Henk C. Van der (1978). "The SN(ANRORC) mechanism: a new mechanism for nucleophilic substitution". In: *Acc. Chem. Res.* 11.12, pp. 462–468.

Rhodes, J. and Nehaniv, C.L. (2009). *Applications of automata theory and algebra*. World Scientific.

Smith, Eric and Morowitz, Harold J (2016). *The origin and nature of life on earth: the emergence of the fourth geosphere*. Cambridge University Press.