Disconnection Rules are Complete for Chemical Reactions

Ella Gale[1], Leo Lobski[2]✉, and Fabio Zanasi[2,3]

[1] University of Bristol, UK ella.gale@bristol.ac.uk
[2] University College London, UK leo.lobski.21@ucl.ac.uk f.zanasi@ucl.ac.uk
[3] University of Bologna, Italy

Abstract. We provide a category theoretical framework capturing two approaches to graph-based models of chemistry: formal *reactions* and *disconnection rules*. We model a translation from the latter to the former as a functor, which is faithful, and full up to isomorphism. This allows us to state, as our main result, that the disconnection rules are sound, complete and universal with respect to the reactions. Concretely, this means that every reaction can be decomposed into a sequence of disconnection rules in an essentially unique way. This provides a uniform way to store and compare reaction data, and gives an algorithmic interface between (forward) reaction prediction and (backward) reaction search or retrosynthesis.

Keywords: chemical reactions, disconnection rules, completeness

1 Introduction

Graph-based models of chemical processes typically come at two different levels of abstraction: formal *reactions* and *disconnection rules*. Formal reactions are combinatorial rearrangements of atoms and charge, and are used for reaction prediction and storage of reaction data (see eg. the rightmost part of Figure 1 below). Disconnection rules constitute hypothetical bond breaking in the direction opposite to a reaction, and are used for designing synthetic pathways and reaction search, known as *retrosynthesis* [6,15,5,14,17] (see eg. Figure 2 below). Retrosynthetic analysis starts with a target molecule we wish to produce but do not know how. The aim is to "reduce" the target molecule to known starting molecules in such a way that when the starting molecules react, the target molecule is obtained as a product. This is done by (formally) partitioning the target molecule into functional parts referred to as *synthons*, which are replaced by actual molecules acting as the new targets [7,6,15,5]. We refer the reader to our previous work on formalising retrosynthesis [9] for the details. In the current work, we are merely interested in the first step of this process: our key observation is that the bond breaking disconnection rules (together with their converse rules), as found in the theory and practise of retrosynthesis (see Figure 2), are, in fact, enough to capture all reactions.

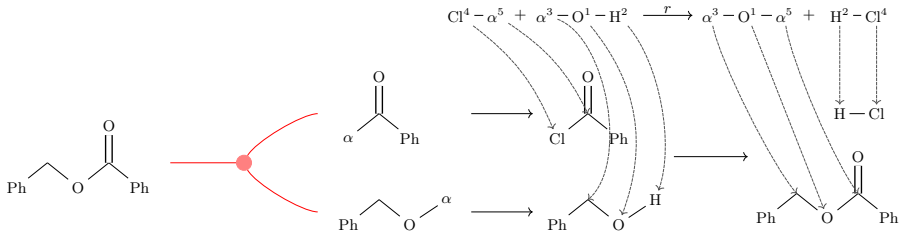


Fig. 1. A simple retrosynthetic sequence. A molecule (far left) is disconnected at the O-COPh bond giving rise to two synthons (left) which can be mapped to precursor molecules (right) which can react to give the product (far right).

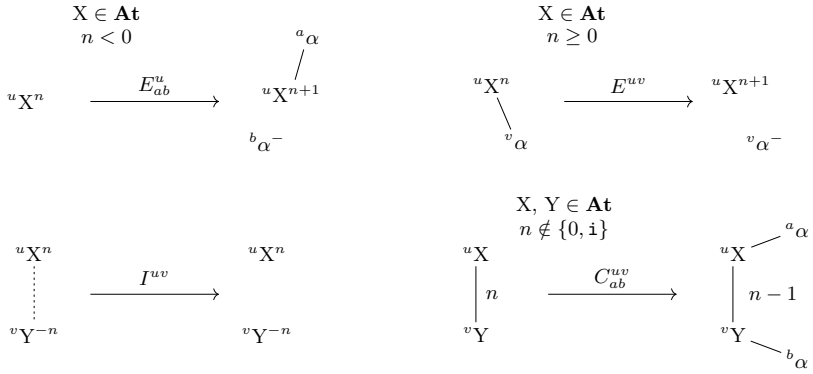


Fig. 2. The four disconnection rules: *electron detachment* (negative and nonnegative charge versions), *ionic bond breaking* and *covalent bond breaking.*

Whereas reactions have been studied formally before, a mathematical description of disconnection rules has received far less attention [4,9]. Our approach takes a novel perspective on the basic units of retrosynthetic analysis – the disconnection rules – by making them first-class citizens of reaction representation. The mathematical and conceptual justification for doing so lies is that, as we show, both disconnection rules and reactions can be arranged into (monoidal) categories [9], such that there is a functor taking each sequence of disconnection rules to a reaction. Our main result states that, under a certain axiomatisation of the disconnection rules, the functor is faithful and full up to isomorphism. Such a categorical perspective provides a precise mathematical meaning to the claim that disconnection rules are sound, complete and universal with respect to the reactions. This implies that every reaction can be decomposed into a sequence of disconnection rules (universality) in an essentially unique way (completeness). More broadly, our contribution incorporates disconnection rules within the framework of applied category theory [8], which emphasises compositional modelling as a means to uniformly study systems across various disciplines of science.

The rest of the paper is structured as follows. Section 2 defines the chemical graphs and the category of reactions. Section 3 defines the disconnection rules and their category. It is moreover shown that any sequence of well-typed disconnection rules has a normal form. Section 4 defines a functor from the disconnection category to the reaction category, and proves completeness and universality. Section 5 concludes.

2 Chemical Graphs and Reactions

We first define the chemical graphs in Subsection 2.1, which form the objects both in the category of reactions (Definition 4) and the disconnection category (Definition 9). Chemical reactions are modelled as certain combinatorial rearrangements of chemical graphs that preserve matter and charge in the appropriate sense. The core of the section is Definition 4, the semantic domain for our interpretation of disconnection rules in Section 4.

2.1 Chemical Graphs

Let us fix a finite set of *vertex labels* $\mathbf{At}$, containing the special symbol α. Formally, the only assumptions we make about $\mathbf{At}$ are (1) it is finite, (2) it contains the special symbol α, and (3) $\mathbf{At} \setminus \{\alpha\}$ has at least two elements. However, in all the examples we shall assume that $\mathbf{At}$ contains a symbol for each main-group element of the periodic table: $\{H, C, O, P, \dots\} \subseteq \mathbf{At}$. For this reason we will also refer to $\mathbf{At}$ as the *atom labels*. The special symbol α may be thought of as representing an unpaired electron. Similarly, we fix a *valence function* $\mathbf{v} : \mathbf{At} \to \mathbb{N}$ with the only formal assumption that $\mathbf{v}(\alpha) = 1$, but shall assume in the examples that the valence of an element symbol is the number of electrons in its outer electron shell.

Remark 1. The reason for choosing such level of generality for the atom labels and their valencies is the ability to model elements which exhibit different valence depending on the context. For instance, one could have separate atom labels for nitrogen whose valence is 5 (all outer shell electrons are shared or take part in a reaction) or 3 (two of the outer shell electrons pair with each other).

Let $\mathbf{Lab} \coloneqq \{0, 1, 2, 3, 4, \mathtt{i}\}$ denote the set of *edge labels*, where the integers stand for a covalent bond, and $\mathtt{i}$ for an ionic bond. We further define maps $\mathtt{cov}, \mathtt{ion} : \mathbf{Lab} \to \mathbb{N}$: for $\mathtt{cov}$, assign to each edge label 0, 1, 2, 3, and 4 the corresponding natural number and let $\mathtt{i} \mapsto 0$, while for $\mathtt{ion}$, let $0, 1, 2, 3, 4 \mapsto 0$ and $\mathtt{i} \mapsto 1$. Finally, let us fix a countable set $\mathbf{VN}$ of *vertex names*; we usually denote the elements of $\mathbf{VN}$ by lowercase Latin letters $u, v, w, \ldots$.

Definition 1 (Chemically labelled graph). *A* chemically labelled graph *is a triple* (V, τ, m)*, where* $V \subseteq \mathbf{VN}$ *is a finite set of* vertices*,* $\tau : V \to \mathbf{At} \times \mathbb{Z}$ *is a* vertex labelling function*, and* $m : V \times V \to \mathbf{Lab}$ *is an* edge labelling function *satisfying* $m(v, v) = 0$ *and* $m(v, w) = m(w, v)$ *for all* $v, w \in V$*.*

Thus, a chemically labelled graph is irreflexive (we interpet the edge label 0 as no edge) and symmetric, and each of its vertices is labelled with an element of $\mathbf{At}$, together with an integer indicating the charge. Given a chemically labelled graph A, we write (V_A, τ_A, m_A) for its vertex set and the labelling functions. We abbreviate the vertex labelling function followed by the first projection as τ_A^{At}, and similarly we write $\tau_A^{\mathtt{Crg}}$ for composition with the second projection.

Given a chemically labelled graph A and vertex names $u, v \in \mathbf{VN}$ such that $u \in V_A$ but $v \notin V_A \setminus \{u\}$, we denote by $A(u \mapsto v)$ the chemically labelled graph whose vertex set is $(V_A \setminus \{u\}) \cup \{v\}$, and whose vertex and edge labelling functions agree with those of A, treating v as if it were u. Further, we define the following special subsets of vertices:

- α*-vertices*, whose label is the special symbol: $\alpha(A) \coloneqq \tau_A^{-1}(\alpha, \mathbb{Z})$,
- *chemical vertices*, whose label is not α: $\mathtt{Chem}\,(A) \coloneqq V_A \setminus \alpha(A)$,
- *neutral vertices*, whose charge is zero: $\mathtt{Neu}\,(A) \coloneqq \tau_A^{-1}(\mathbf{At}, 0)$,
- *charged vertices*, which have a non-zero charge: $\mathtt{Crg}\,(A) \coloneqq V_A \setminus \mathtt{Neu}\,(A)$,
- *negative vertices*, which have a negative charge:

$$\mathtt{Crg}^{-}\,(A) \coloneqq \{v \in V_A : \tau_A^{\mathtt{Crg}}(v) < 0\},$$

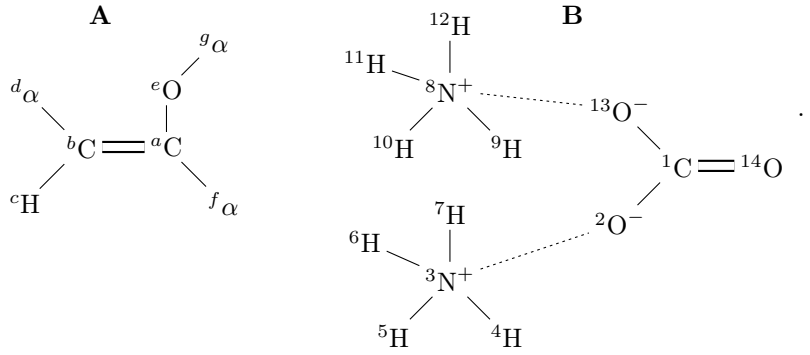
- *positive vertices*, which have a positive charge:

$$\mathtt{Crg}^{+}\,(A) \coloneqq \{v \in V_A : \tau_A^{\mathtt{Crg}}(v) > 0\}.$$

The *net charge* of a subset $U \subseteq V_A$ is the integer $\mathtt{Net}\,(U) \coloneqq \sum_{v \in U} \tau_A^{\mathtt{Crg}}(v)$.

Example 1. We give examples of two chemically labelled graphs: A (ethenone) and B (ammonium carbonate). We adopt the following conventions: (1) the vertex label from $\mathbf{At}$ is drawn at the centre of a vertex, (2) the vertex name is drawn as a superscript on the left (so within a single graph, no left superscript

appears twice), (3) the charge is drawn as a superscript on the right, (4) n-ary covalent bonds are drawn as n parallel lines, and (5) ionic bonds are drawn as dashed lines. Moreover, for the sake of legibility, we shall omit the vertex names (left superscripts), unless the precise name plays a role.



Below we give a table with different kinds of vertex subsets for the graphs A and B:

	A	B
α-vertices	$\{d, f, g\}$	$\varnothing$
chemical vertices	$\{a, b, c, e\}$	V_B
neutral vertices	V_A	$\{1, 4, 5, 6, 7, 9, 10, 11, 12, 14\}$
charged vertices	$\varnothing$	$\{2, 3, 8, 13\}$
negative vertices	$\varnothing$	$\{2, 13\}$
positive vertices	$\varnothing$	$\{3, 8\}$

The net charge of both A and B is zero, while the net charge of the subset $\{1, 2, 13, 14\}$ of B (the carbonate anion) is -2.

Definition 2 (Neighbours). *Given a chemically labelled graph A and a vertex $u \in V_A$, we define the sets of* neighbours $\mathtt{N}_A(u)$, covalent neighbours $\mathtt{CN}_A(u)$ *and* ionic neighbours $\mathtt{IN}_A(u)$ *of u as follows:*

$$\begin{aligned}\mathtt{N}_A(u) &\coloneqq \{v \in V_A : m_A(u, v) \neq 0\},\\ \mathtt{CN}_A(u) &\coloneqq \{v \in V_A : \mathtt{cov}(m_A(u, v)) \neq 0\},\\ \mathtt{IN}_A(u) &\coloneqq \{v \in V_A : \mathtt{ion}(m_A(u, v)) \neq 0\}.\end{aligned}$$

Definition 3 (Chemical graph). *A* chemical graph $A = (V_A, \tau_A, m_A)$ *is a chemically labelled graph satisfying the following additional conditions:*

1. *for all $v \in V_A$, we have*

$$\left|\tau_A^{\mathtt{Crg}}(v)\right| + \sum_{u \in V_A} \mathtt{cov}\left(m_A(u, v)\right) = \mathbf{v}\tau_A^{\mathrm{At}}(v),$$

2. *for all $v \in \alpha(A)$ and $w \in V_A$ we have*

   (a) $\tau_A^{\mathtt{Crg}}(v) \in \{-1, 0\}$,
   (b) $m_A(v, w) \in \{0, 1\}$,
   (c) $\mathtt{N}_A(u)$ *has at most one element, and if* $w \in \mathtt{N}_A(u)$*, then* $w \in \mathtt{Chem}(A)$,
3. *for all* $v \in \mathtt{Chem}(A)$*, the set* $\mathtt{IN}_A(v)$ *has at most one element, and if* $u \in \mathtt{IN}_A(v)$*, then* $\tau_A^{\mathtt{Crg}}(v) = -\tau_A^{\mathtt{Crg}}(u)$.

Condition 1 states that the sum of incident covalent bonds together with the absolute value of the charge must equal the valence of the vertex. Conditions 2a-2c say that a vertex labelled by α is either neutral or has charge -1, has at most one neighbour, which is necessarily chemical and to which it is connected via a single covalent bond. Conditions 3 says that an edge with label `i` only connects charged chemical vertices with equal net charges of opposite signs.

A *synthon* is a chemical graph which is moreover connected. The collection of chemical graphs is, therefore, generated by the disjoint unions of synthons. A *molecular graph* is a chemical graph with no α-vertices. A *molecular entity* is a connected molecular graph.

Example 2. Both chemically labelled graphs A and B in Example 1 are, in fact, chemical graphs with the standard valences of the atoms (i.e. $\mathbf{v}(\mathrm{C}) = 4$, $\mathbf{v}(\mathrm{O}) = 2$, $\mathbf{v}(\mathrm{H}) = 1$ and $\mathbf{v}(\mathrm{N}) = 5$). Moreover, A is a synthon and B is a molecular entity.

2.2 Category of Reactions

We define reactions between chemically labelled graphs as partial bijections preserving the atom labels of chemical vertices whose domain and image have the same net charge, with the additional condition that the complements of the domain and image are isomorphic. The intuition is that electrons and atoms' charge may appear and disappear in the course of a reaction in such a way that the overall charge is preserved. This way of representing reactions is motivated by and formally connected to *double pushout graph rewriting* [1,2,3]: in fact, every reaction can be represented as a double pushout diagram in the category of chemical graphs.

Definition 4 (Category of reactions). *We denote by* **React** *the* category of reactions*, whose*

- *objects are chemical graphs,*
- *morphisms* $A \to B$ *are tuples* (U_A, U_B, b, i)*, where*
  - $U_A \subseteq V_A$ *and* $U_B \subseteq V_B$ *are subsets with* $\mathtt{Net}(U_A) = \mathtt{Net}(U_B)$,
  - $b : \mathtt{Chem}(U_A) \to \mathtt{Chem}(U_B)$ *is a bijection preserving the atom labels,*
  - $i : V_A \setminus U_A \to V_B \setminus U_B$ *is an isomorphism of labelled graphs,*

  such that for all $u \in \mathtt{Chem}(U_A)$ *and* $a \in V_A \setminus U_A$ *we have*

$$m_A(u, a) = m_B(bu, ia),$$

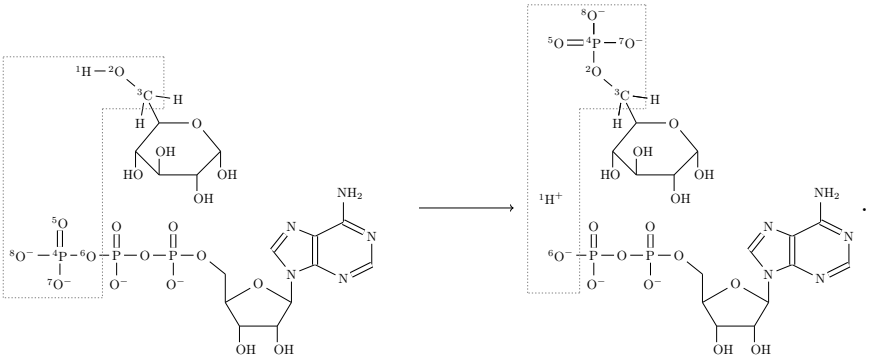
– *the composition of* $(U_A, U_B, b, i) : A \to B$ *and* $(W_B, W_C, c, j) : B \to C$ *is given by*

$$(Z_A, Z_C, (c+j)(b+i), ji) : A \to C,$$

where $Z_A \coloneqq U_A \cup i^{-1}(W_B \cap (V_B \setminus U_B))$ *and* $Z_C \coloneqq W_C \cup j(U_B \cap (V_B \setminus W_B))$,
– *for a molecular graph* A*, the identity is given by* $(\varnothing, \varnothing, !, \mathrm{id}_A)$*, where* $!$ *is the unique endomorphism on the empty set.*

Note that the composition in **React** is *not* the composition in the usual category of partial bijections: instead, it crucially relies on the fact that there is an isomorphism between the "unchanged" parts of the graph. The category **React** has a dagger structure [12,10]: the dagger of $(U_A, U_B, b, i) : A \to B$ is given by $(U_B, U_A, b^{-1}, i^{-1}) : B \to A$. Given a morphism $r \in \mathbf{React}$, we will denote its dagger by $\overline{r}$.

Example 3. We give an example of a reaction (glucose phosphorylation) below. We have indicated the subsets on which the bijection is defined by dashed lines, and the bijection is the identity in this case.



3 Disconnection Rules

A *disconnection rule* is a partial endofunction on the set of chemical graphs. We define four classes of disconnection rules, all of which have a clear chemical significance: two versions of *electron detachment*, *ionic bond breaking* and *covalent bond breaking*. While Definition 5 is rather technical, Figure 2 gives a better idea of what is going on.

Definition 5 (Disconnection rules). *Let* $u, v, a, b \in \mathbf{VN}$ *be pairwise distinct vertex names. Let* $U \in \{u, uv\}$ *and* $D \in \{\varnothing, ab\}$ *range over the specified lists of vertex names. The four* disconnection rules *are defined by the table in Figure 3 as follows: a chemical graph* A *is in the domain of* d^U_D *if* $U \subseteq V_A$ *but* $D \cap V_A = \varnothing$*, and the additional conditions of the first column hold; the output chemical graph* $d(A)$ *has the vertex set* $V_A \cup D$*, and the labelling functions on* $U \cup D$ *are defined by the remaining columns, while the labelling functions agree with those of* A *on* $V_A \setminus U$*.*

d_D^U	$A \in dom(d)$	$\tau_{d(A)}^{\mathtt{Crg}}(u)$	$\tau_{d(A)}^{\mathtt{Crg}}(v)$	$\tau_{d(A)}(a)$	$\tau_{d(A)}(b)$	$m_{d(A)}(u,v)$	$m_{d(A)}(u,a)$	$m_{d(A)}(v,b)$
E_{ab}^{u}	$u \in \mathtt{Chem}(A)$ $u \in \mathtt{Crg}^{-}(A)$	$\tau_A^{\mathtt{Crg}}(u)+1$	N/A	$(\alpha,0)$	$(\alpha,-1)$	N/A	1	N/A
E^{uv}	$u \in \mathtt{Chem}(A)$ $u \notin \mathtt{Crg}^{-}(A)$ $v \in \alpha(A)$ $m_A(u,v)=1$	$\tau_A^{\mathtt{Crg}}(u)+1$	-1	N/A	N/A	0	N/A	N/A
I^{uv}	$m_A(u,v)=\mathtt{i}$ $u \in \mathtt{Crg}^{+}(A)$ $v \in \mathtt{Crg}^{-}(A)$	$\tau_A^{\mathtt{Crg}}(u)$	$\tau_A^{\mathtt{Crg}}(v)$	N/A	N/A	0	N/A	N/A
C_{ab}^{uv}	$u,v \in \mathtt{Chem}(A)$ $m_A(u,v) \notin \{0,\mathtt{i}\}$	$\tau_A^{\mathtt{Crg}}(u)$	$\tau_A^{\mathtt{Crg}}(v)$	$(\alpha,0)$	$(\alpha,0)$	$m_A(u,v)-1$	1	1

Fig. 3. The disconnection rules defined as partial functions.

Note that the disconnection rules look a lot like (a subset of) morphisms in **React**, except that we keep track of the precise vertex names, and a rule applies to a whole set of chemical graphs. We make this connection precise in Section 4. We further observe that each disconnection rule is injective (as a partial function), and hence has an inverse partial function.

Definition 6 (Terms). *The set of* terms *with types is generated by the following recursive procedure:*

- *for every chemical graph A, let* $\mathrm{id} : A \to A$ *be a term,*
- *for every chemical graph A and every $u \in V_A$, let $S^u : A \to A$ be a term,*
- *for every chemical graph A, every $u \in \alpha(A)$ and every $v \in \mathbf{VN}$ such that $v \notin V_A \setminus \{u\}$, let $R^{u \mapsto v} : A \to A(u \mapsto v)$ be a term,*
- *for every disconnection rule d and every chemical graph A in the domain of d, both $d : A \to d(A)$ and $\bar{d} : d(A) \to A$ are terms,*
- *if $\mathtt{t} : A \to B$ and $\mathtt{s} : B \to C$ are terms, then $\mathtt{t};\mathtt{s} : A \to C$ is a term.*

We refer to the terms of the form $d : A \to B$ and $\bar{d} : B \to A$ generated by the fourth item as *disconnections* and *connections*, respectively. More specifically, we use the symbols $E^{<0}$, $E^{\geq 0}$, I and C to denote the disconnections corresponding to the specific disconnection rules, and similarly the symbols $\bar{E}^{<0}$, $\bar{E}^{\geq 0}$, $\bar{I}$ and $\bar{C}$ refer to the corresponding connections. Finally, S and R refer to the classes of terms generated by the second and third items. The same letters in the typewriter type font ($\mathtt{E}^{<0}$, $\mathtt{E}^{\geq 0}$, $\mathtt{I}$, $\mathtt{C}$, $\bar{\mathtt{E}}^{<0}$, $\bar{\mathtt{E}}^{\geq 0}$, $\bar{\mathtt{I}}$, $\mathtt{S}$ and $\mathtt{R}$) are used to denote a sequence of terms of the corresponding kind.

Let us define the endofunction $\overline{()}$ on terms by the following recursion:

- $(\mathrm{id} : A \to A) \mapsto (\mathrm{id} : A \to A)$,
- $(S^u : A \to A) \mapsto (S^u : A \to A)$,
- $(R^{u \mapsto v} : A \to A(u \mapsto v)) \mapsto (R^{v \mapsto u} : A(u \mapsto v) \to A)$,
- $(d : A \to B) \mapsto \big(\bar{d} : B \to A\big)$,
- $\big(\bar{d} : A \to B\big) \mapsto (d : B \to A)$,
- $\overline{\mathtt{t}; \mathtt{s}} \coloneqq \overline{\mathtt{s}}; \overline{\mathtt{t}}$.

For defining equations, it will be useful to allow untyped terms.

Definition 7 (Untyped terms, well-typedness). *An* untyped term *is an element of the free monoid on the set*

$$\{\mathrm{id}, S^u, R^{a \mapsto b}, E^{ua}, E^u_{ab}, C^{uv}_{ab}, I^{uv}, \bar{E}^{ua}, \bar{E}^u_{ab}, \bar{C}^{uv}_{ab}, \bar{I}^{uv} : u, v, a, b \in \mathbf{VN}\},$$

where we use the symbol ; *to indicate the multiplication of the monoid.*

Given an untyped term $\mathtt{t}$ *and chemical graphs* A *and* B*, we say that the expression*

$$\mathtt{t} : A \to B$$

is well-typed *if it is in fact a term, that is, if it can be constructed using the recursive procedure of Definition 6.*

We define the binary relation $\leq$ *on the set of untyped terms by letting* $\mathtt{t} \leq \mathtt{s}$ *if whenever* $\mathtt{t} : A \to B$ *is well-typed, then so is* $\mathtt{s} : A \to B$*.*

We define the endofunction $\overline{()}$ on the untyped terms in exactly the same way as for the terms with types, simply ignoring the types. Note that $\mathtt{t} : A \to B$ is well-typed if and only if $\overline{\mathtt{t}} : B \to A$ is. Moreover, observe that $\leq$ defines a preorder on the untyped terms. Consequently, we have $\mathtt{t} \leq \mathtt{s}$ if and only if $\overline{\mathtt{t}} \leq \overline{\mathtt{s}}$.

Given an untyped term $\mathtt{t}$, there are either no chemical graphs such that $\mathtt{t} : A \to B$ is well-typed, or there are infinitely many such graphs. The latter case is the reason for introducing the untyped terms: we want certain equalities to hold *whenever* both sides are well-typed.

Definition 8 (Term equality). *Let* $\simeq$ *be an equivalence relation on the set of untyped terms. This induces the equivalence relation* $\equiv$ *on the set of terms as follows: for two terms* $\mathtt{t}, \mathtt{s} : A \to B$ *with the same type, we let* $\mathtt{t} \equiv \mathtt{s}$ *if either* $\mathtt{t} \simeq \mathtt{s}$ *or* $\overline{\mathtt{t}} \simeq \overline{\mathtt{s}}$ *as untyped terms. We introduce the following shorthand binary relations on the untyped terms:*

– $\mathtt{t} \leqq \mathtt{s}$ *if* $\mathtt{t} \simeq \mathtt{s}$ *and* $\mathtt{t} \leq \mathtt{s}$,
– $\mathtt{t} = \mathtt{s}$ *if* $\mathtt{t} \leqq \mathtt{s}$ *and* $\mathtt{s} \leqq \mathtt{t}$.

Definition 9 (Disconnection category). *The* disconnection category **Disc** *has as objects the chemical graphs. Given chemical graphs A and B, the set of morphisms* $\mathbf{Disc}(A, B)$ *is given by the terms of type $A \to B$, subject to the usual associativity and unitality equations of a category, together with the identities induced (in the sense of Definition 8) by the equivalence relation defined in Figure 4.*

$$R^{u\mapsto z}; R^{z\mapsto w} \leqq R^{u\mapsto w} \quad (1)$$
$$R^{u\mapsto z}; R^{v\mapsto w} = R^{v\mapsto w}; R^{u\mapsto z}, \quad z \neq v, u \neq w \quad (2)$$
$$R^{u\mapsto u} \leqq S^u \quad (3)$$
$$R^{b\mapsto z}; R^{a\mapsto b} \simeq S^b; R^{a\mapsto z} \quad (4)$$
$$R^{u\mapsto v}; S^w = S^w; R^{u\mapsto v}, \quad w \notin \{u, v\} \quad (5)$$
$$R^{u\mapsto v}; S^v = S^u; R^{u\mapsto v} = R^{u\mapsto v} \quad (6)$$

$$R^{u\mapsto v}; d^U_D = d^U_D; R^{u\mapsto v}, \quad u, v \notin U, D \quad (7)$$
$$R^{u\mapsto v}; E^{wv} = E^{wu}; R^{u\mapsto v} \quad (8)$$
$$d^U_{D[u]}; R^{u\mapsto v} = d^U_{D[v/u]} \quad (9)$$
$$d^{U'}_{ij}; \bar{h}^U_{ab}; R^{i\mapsto c}; R^{j\mapsto d} \leqq \bar{h}^U_{ab}; d^{U'}_{cd} \quad (10)$$

$$d^U_{ab}; \bar{d}^U_{cd} \leqq S^U; R^{c\mapsto a}; R^{d\mapsto b}, \quad c \neq a, d \neq b \quad (11)$$
$$d^U_{ab}; \bar{d}^U_{cb} \leqq S^U; R^{c\mapsto a}, \quad c \neq a \quad (12)$$
$$d^U_{ab}; \bar{d}^U_{ad} \leqq S^U; R^{d\mapsto b}, \quad d \neq b \quad (13)$$
$$d^U_D; \bar{d}^U_D \leqq S^U \quad (14)$$
$$\bar{d}^U_D; d^U_D \leqq S^U \quad (15)$$
$$E^{ua}; \bar{E}^{ub} \simeq S^u; R^{a\mapsto z}; R^{b\mapsto a}; R^{z\mapsto b}, \quad a \neq b \quad (16)$$
$$\bar{d}^{uv}; d^{wz} = d^{wz}; \bar{d}^{uv}, \quad \{u, v\} \neq \{w, z\} \quad (17)$$

$$S^u; S^v = S^v; S^u \quad (18)$$
$$S^u; S^u = S^u \quad (19)$$
$$S^u; d^U_D \leqq d^U_D; S^u \quad (20)$$
$$d^{U[v]}_D; S^v = d^{U[v]}_D \quad (21)$$
$$C^{uv}_{ab} = C^{vu}_{ba} \quad (22)$$

$$d^U_D; d^{U'}_{D'} = d^{U'}_{D'}; d^U_D \quad (23)$$
$$C^{uv}_{ab}; I^{wz} = I^{wz}; C^{uv}_{ab} \quad (24)$$
$$E^u_{ab}; I^{wz} \leqq I^{wz}; E^u_{ab} \quad (25)$$
$$E^{uv}; I^{wz} \leqq I^{wz}; E^{uv} \quad (26)$$
$$\bar{E}^{uv}; I^{wz} \leqq I^{wz}; \bar{E}^{uv} \quad (27)$$
$$\bar{E}^u_{ab}; I^{wz} \leqq I^{wz}; \bar{E}^u_{ab} \quad (28)$$
$$\bar{C}^{uv}_{ab}; I^{wz} \leqq I^{wz}; \bar{C}^{uv}_{ab} \quad (29)$$
$$E^u_{ab}; C^{wz}_{cd} = C^{wz}_{cd}; E^u_{ab} \quad (30)$$
$$E^{uv}; C^{wz}_{cd} \leqq C^{wz}_{cd}; E^{uv} \quad (31)$$
$$\bar{E}^{uv}; C^{wz}_{cd} = C^{wz}_{cd}; \bar{E}^{uv} \quad (32)$$
$$E^{uv}; E^w_{cd} \leqq E^w_{cd}; E^{uv} \quad (33)$$
$$\bar{E}^{uv}; E^w_{cd} = E^w_{cd}; \bar{E}^{uv} \quad (34)$$

Fig. 4. The equivalence relation inducing the identities in the disconnection category. Here d and h range over $\{E, C, I\}$, while S^U stands for the sequence $S^u; S^w$ if $U = uw$. Given vertex names $a, b \in \mathbf{VN}$, the notation $D[a]$ means a occurs in D, and $D[b/a]$ means the occurrence of a in D is replaced with b.

Note that the assignment $\overline{()} : \mathbf{Disc} \to \mathbf{Disc}$ is functorial, thus making **Disc** a dagger category [12,10].

Proposition 1. *The following identities are derivable in* **Disc***:*

$$d^U_{D[a]}; S^a = d^U_{D[a]}, \tag{35}$$

$$\bar{d}^U_{ab}; d^U_{cd} \leqq S^U; R^{a\mapsto c}; R^{b\mapsto d}, \tag{36}$$

$$R^{z\mapsto c}; R^{w\mapsto d}; d^U_{ab} \simeq R^{z\mapsto a}; R^{w\mapsto b}; d^U_{cd}, \tag{37}$$

$$R^{z\mapsto c}; d^U_{ab} \simeq R^{z\mapsto a}; d^U_{cb}, \tag{38}$$

$$R^{w\mapsto d}; d^U_{ab} \simeq R^{w\mapsto b}; d^U_{ad}, \tag{39}$$

$$d^U_{ab}; d^U_{cd} = d^U_{ad}; d^U_{cb}. \tag{40}$$

3.1 Normal Form

In this subsection, we define a normal form (Definition 12), and show that every term is equal to a term in a normal form under the equalities of **Disc** (Proposition 3). We also identify a class of syntactic manipulations of terms in a normal form (Definition 13) that both keep the normal form and preserve equality (Lemma 2). These results are used in the next section to prove completeness.

Definition 10 (*ICE*-form). *We say that a term is in an* ICE-form *if it is either an identity term, or if it has the following structure:*

$$\mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \bar{\mathtt{E}}^{\geq 0}; \bar{\mathtt{E}}^{<0}; \bar{\mathtt{C}}; \bar{\mathtt{I}}; \mathtt{R}; \mathtt{S},$$

where every letter is a sequence of generating terms of the corresponding kind.

Proposition 2. *Any term is equal to a term in an* ICE*-form.*

Proof (sketch). The proof proceeds by repeated inductions: one first shows that all I-terms can always be commuted to the left, then that all C-terms can be commuted to the left of anything that is not an I-term, and so on.

Definition 11 (Renaming form). *A well-typed sequence of renaming terms* $\mathtt{R} : H \to G$ *is in a* renaming form *if there are sets of vertex names* $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_n\}$, $C = \{c_1, \ldots, c_m\}$ *and* $D = \{d_1, \ldots, d_m\}$ *such that (1)* $\mathtt{R}$ *can be split into two sequences* $\mathtt{R} = \mathtt{A}; \mathtt{B}$ *with*

$$\mathtt{A} = R^{a_1\mapsto b_1}; \ldots; R^{a_n\mapsto b_n} \quad \textit{and} \quad \mathtt{B} = R^{c_1\mapsto d_1}; \ldots; R^{c_m\mapsto d_m},$$

where $\mathtt{B}$ *can be possibly empty, (2) the vertex names in each of the sets are pairwise distinct, (3)* $A \cap B = \varnothing$*, (4)* $C \subseteq B$*, (5)* $D \subseteq A$*, and (6) if* $c_i \in C$ *and* $b_j \in B$ *is the unique element such that* $b_j = c_i$*, then* $\mathtt{N}_H(a_j) \neq \mathtt{N}_H(d_i)$*.*

Lemma 1. *Any well-typed sequence of renaming terms is equal to a term* $\mathtt{R}; \mathtt{S}$*, where* $\mathtt{R} = \mathtt{A}; \mathtt{B}$ *is in a renaming form and* $\mathtt{S}$ *is a sequence of* S*-terms.*

Proof (sketch). The idea is that if a vertex a is to be renamed to b, then $a \in A$, and we have two cases: (1) b does not already occur in the original chemical graph, and (2) b does occur in the original graph. If (1), then $R^{a \mapsto b} \in \mathtt{A}$ and $b \in B \setminus C$. If (2), then we first rename a using some "dummy" name c, so that $R^{a \mapsto c} \in \mathtt{A}$, $R^{c \mapsto b} \in \mathtt{B}$, $c \in C$ and $b \in D$. Note that condition (5) of the renaming form is satisfied, as b must itself be renamed in order for the vertex name become free. Any term of the form $R^{a \mapsto a}$ is replaced by S^a. The formal proof proceeds by induction on the length of the original sequence.

A term is said to be in an $ICER$-form if it is in an ICE-form whose sequence of renaming terms is in a renaming form (or is empty).

Definition 12 (Normal form). *Let*

$$\mathtt{t} = \mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \bar{\mathtt{E}}^{\geq 0}; \bar{\mathtt{E}}^{<0}; \bar{\mathtt{C}}; \bar{\mathtt{I}}; \mathtt{A}; \mathtt{B}; \mathtt{S}$$

be a term in an $ICER$-form. Let us denote the sets of vertex names in the renaming form by $A_\mathtt{t}$, $B_\mathtt{t}$, $C_\mathtt{t}$ and $D_\mathtt{t}$. Let us additionally define the following sets of vertex names occurring in $\mathtt{t}$*:*

- $D^{add}_\mathtt{t} \coloneqq \left\{ a \in \mathbf{VN} : d^U_{D[a]} \in \mathtt{t} \right\}$ *– the vertex names appearing as subscripts in the disconnections,*
- $D^{remove}_\mathtt{t} \coloneqq \left\{ a \in \mathbf{VN} : \bar{d}^U_{D[a]} \in \mathtt{t} \right\}$ *– the vertex names appearing as subscripts in the connections,*
- $U_\mathtt{t} \coloneqq \left\{ v \in \mathbf{VN} : d^{U[v]}_D \in \mathtt{t} \text{ or } \bar{d}^{U[v]}_D \in \mathtt{t} \right\}$ *– the vertex names appearing as superscripts of the (dis)connections,*
- $S_\mathtt{t} \coloneqq \{u \in \mathbf{VN} : S^u \in \mathtt{t}\}$ *– the vertex names appearing in the S-terms.*

We say that a term $\mathtt{t}$ *is in a* normal form *if it is in an $ICER$-form as above, and additionally the following conditions hold:*

(1) for every $u \in S_\mathtt{t}$, the term S^u occurs in $\mathtt{t}$ exactly once,
(2) $(U_\mathtt{t} \cup A_\mathtt{t} \cup B_\mathtt{t}) \cap S_\mathtt{t} = \varnothing$,
(3) $D^{add}_\mathtt{t} \setminus D^{remove}_\mathtt{t} \subseteq A_\mathtt{t} \setminus D_\mathtt{t}$,
(4) $D^{add}_\mathtt{t} \cap B_\mathtt{t} = \varnothing$,
(5) if a connection $\bar{d}^U_{D[a]} : A \to B$ and a renaming term $R^{z \mapsto a}$ both occur, then A is not in the domain of $\bar{d}^U_{D[z/a]}$,
(6) if $d \neq I$ and a disconnection d^U_D occurs in $\mathtt{t}$, then the connections $\bar{d}^U_F$ and $\bar{d}^{U^r}_F$ do not occur in $\mathtt{t}$ for any F,
(7) if the disconnection E^{uv} occurs in $\mathtt{t}$, then for any vertex name $w \in \mathbf{VN}$, the connection $\bar{E}^{uw}$ does not occur in $\mathtt{t}$,
(8) if the disconnection I^{uv} and the connection $\bar{I}^{uv}$ both occur in $\mathtt{t}$, then one of the terms E^v_D, $\bar{E}^v_D$, E^{va} or $\bar{E}^{va}$ occurs in $\mathtt{t}$.

Proposition 3. *Any term is equal to a term in normal form.*

Proof (sketch). By Proposition 2 and Lemma 1, every terms is equal to a term in an $ICER$-form. Conditions (1) and (2) are obtained by absorbing the "excess" S-terms into other terms using equations (6), (19) and (21). Conditions (3) and (4) are obtained by treating all the vertex names in $D_{\mathtt{t}}^{add}$ as "dummy" names, which are removed either by a connection or a renaming term. Condition (5) is obtained by equations (38) and (39). Conditions (6) and (7) are minimal change conditions: if the number of bonds (or electrons) is decreased, then it should not be increased again. These are obtained by cancelling the redundant disconnection-connection pairs. Finally, condition (8) says that an ionic bond is first removed and then reintroduced between the same pair of vertices just in case the charge of the vertices has to be changed.

Definition 13 (Normal form equivalence). *Let*

$$\mathtt{t} = \mathtt{I};\mathtt{C};\mathtt{E}^{<0};\mathtt{E}^{\geq 0};\bar{\mathtt{E}}^{\geq 0};\bar{\mathtt{E}}^{<0};\bar{\mathtt{C}};\bar{\mathtt{I}};\mathtt{A};\mathtt{B};\mathtt{S}$$

be a term in a normal form. Define the following syntactic manipulations of $\mathtt{t}$*:*

1. *commuting the terms inside each of the named sequences in the normal form,*
2. *permuting vertex names in C-terms: if the term C_{ab}^{uv} occurs, we may substitute it with C_{ba}^{vu},*
3. *if $d \in \{C, E, \bar{C}, \bar{E}\}$ such that $d_{ab}^U; d_{cd}^U$ occurs, we may substitute $d_{ab}^U; d_{cd}^U \mapsto d_{ad}^U; d_{cb}^U$,*
4. *renaming of vertices that are introduced and removed: if $a \in D_{\mathtt{t}}^{add} \cup C_{\mathtt{t}}$ and $z \in \mathbf{VN}$ does not occur in $\mathtt{t}$ or its domain, then we may substitute $\mathtt{t} \mapsto \mathtt{t}[z/a]$,*
5. *exchanging vertex names between $\bar{E}^{<0}$-, $\bar{C}$- and R-terms: if $d \in \{E, C\}$, and $\bar{d}_{D[a]}^U : A \to B$ and $R^{z \mapsto b}$ both occur such that A is in the domain of $\bar{d}_{D[z/a]}^U$, then we may substitute $\bar{d}_{D[a]}^U \mapsto \bar{d}_{D[z/a]}^U$ and $R^{z \mapsto b} \mapsto R^{a \mapsto b}$.*

We say that two terms $\mathtt{t}$ *and* $\mathtt{s}$ *in a normal form are* equivalent*, written* $\mathtt{t} \sim \mathtt{s}$*, if one can be obtained from the other by a sequence of the syntactic manipulations defined above.*

Observing that each syntactic manipulation in Definition 13 is reversible, we see that $\sim$ is an equivalence relation on the set of terms in normal form.

Lemma 2. *Let* $\mathtt{t}$ *and* $\mathtt{s}$ *be terms in normal forms such that* $\mathtt{t} \sim \mathtt{s}$*. Then* $\mathtt{t} = \mathtt{s}$*.*

Proof. This follows by noticing that every syntactic manipulation of Definition 13 keeps the term in a normal form, and moreover preserves equality:

1. the terms may be commuted by equations (23), (2) and (18),
2. vertex names in C-terms may be permuted by (22),
3. the indices in repeated (dis)connections may be exchanged by (40),
4. if $a \in D_{\mathtt{t}}^{add}$, so that there is a disconnection $d_{D[a]}^U$, we use equation (9) to obtain $d_{D[a]}^U = d_{D[b/a]}^U; R^{b \mapsto a}$ to introduce the desired fresh variable b; the renaming term can then be absorbed into the second occurrence of a, hence replacing a with b (the case when $a \in C_{\mathtt{t}}$ is similar),
5. the last syntactic manipulation is obtained by equations (38) and (39).

4 From Disconnections to Reactions, Functorially

We define a function R from terms to morphisms in **React** as follows. Given a term $\mathtt{t} : A \to B$, the morphism $R(\mathtt{t}) : A \to B$ has the form

$$R(\mathtt{t}) = (R_1(\mathtt{t}), R_2(\mathtt{t}), \mathrm{id}, \mathrm{id}),$$

where $\mathrm{id} : \mathtt{Chem}\,(R_1(\mathtt{t})) \to \mathtt{Chem}\,(R_2(\mathtt{t}))$ and $\mathrm{id} : V_A \setminus R_1(\mathtt{t}) \to V_B \setminus R_2(\mathtt{t})$ are both identity maps. Since all the terms are mapped to morphisms whose bijection and isomorphism parts are the identities, we omit these, and write $R(\mathtt{t}) = (R_1(\mathtt{t}), R_2(\mathtt{t}))$. The recursive definition of this mapping is given below:

- $R(\mathrm{id}_A) \coloneqq (\varnothing, \varnothing)$,
- $R\,(S^u) \coloneqq (\{u\}, \{u\})$,
- $R\,(R^{u \mapsto v}) \coloneqq (\{u\}, \{v\})$,
- $R(E^u_{ab}) \coloneqq (\{u\}, \{u, a, b\})$,
- $R(E^{uv}) \coloneqq (\{u, v\}, \{u, v\})$,
- $R(I^{uv}) \coloneqq (\{u, v\}, \{u, v\})$,
- $R(C^{uv}_{ab}) \coloneqq (\{u, v\}, \{u, v, a, b\})$,
- $R\left(\bar{d}^{uv}_{ab}\right) \coloneqq \overline{R\,(d^{uv}_{ab})}$,
- $R(\mathtt{t}; \mathtt{s}) \coloneqq R(\mathtt{t}); R(\mathtt{s})$.

Observe that for all the disconnections we have $R(d^U_D) = (U, U \cup D)$.

Soundness of disconnection rules with respect to reactions is expressed as functoriality:

Proposition 4. *The assignment $R : \mathbf{Disc} \to \mathbf{React}$ is a dagger functor.*

Proof (sketch). Functoriality and preservation of dagger structure follow immediately by construction. We have to show that R preserves the equations of Figure 4. Most of these follow immediately by applying R to expressions on both sides of the equality. We give an example derivation of equation (10) below. To further simplify the notation, we omit the curly brackets of set-builder notation as well as the commas separating vertex names from each other: so e.g. $(uv, uvab)$ stands for $(\{u, v\}, \{u, v, a, b\})$.

Denote $R(d^{U'}_{ij}) = (u'v', u'v'ij)$, $R(d^{U'}_{cd}) = (u'v', u'v'cd)$ and $R(\bar{h}^U_{ab}) = (uvab, uv)$. Note that d and h are not $E^{\geq 0}$-terms, whence it follows that $i, j \notin U$ and $a, b \notin U'$. From the fact that the left-hand side is defined, we obtain that $\{i, j\}$ and $\{a, b\}$ are disjoint. The left-hand side is thus translated to

$$\begin{aligned}
(u'v', u'v'ij); (uvab, uv); (i, c); (j, d) &= (u'v'uvab, uvu'v'ij); (i, c); (j, d) \\
&= (u'v'uvab, cuvu'v'j); (j, d) \\
&= (u'v'uvab, dcuvu'v') \\
&= (uvu'v'ab, uvu'v'cd) \\
&= (uvab, uv); (u'v', u'v'cd),
\end{aligned}$$

which we recognise as the translation of the right-hand side.

Recall the syntactic manipulations of terms in normal form of Definition 13. We have seen that these manipulations preserve equality (Lemma 2). The following lemma is the core of the completeness argument.

Lemma 3. *Let* $\mathtt{t}$ *and* $\mathtt{s}$ *be terms in a normal form such that* $R(\mathtt{t}) = R(\mathtt{s})$. *Then* $\mathtt{t} \sim \mathtt{s}$.

Proof (sketch). The minimal change conditions of a normal form ((6), (7) and (8)) imply that the terms have the same (dis)connections up to renaming the α-vertices. The case analysis of the renaming terms then shows that $\mathtt{t}$ and $\mathtt{s}$ have the same renaming forms, up to renaming and exchange allowed by the normal form equivalence.

Combining the above lemma with the results from the previous section, we conclude that the functor $R : \mathbf{Disc} \to \mathbf{React}$ is faithful. We spell this out in detail in the following:

Theorem 1 (Completeness). *For all terms* $\mathtt{t}$ *and* $\mathtt{s}$, *we have* $\mathtt{t} = \mathtt{s}$ *in* **Disc** *if and only if* $R(\mathtt{t}) = R(\mathtt{s})$ *in* **React**.

Proof. The 'only if' direction is functoriality (Proposition 4). The 'if' direction follows from the fact that every term is equal to a term in normal form (Proposition 3) and from Lemmas 13 and 2.

The argument for universality turns out to be much simpler than that for completeness. However, in combination with Theorem 1, it gives a rather strong representation result for reactions: not only can every reaction be decomposed into a sequence of disconnection rules, but this sequence is also unique, up to changing the vertex names and up to the equations in **Disc**. In abstract terms, the statement of universality is that the functor $R : \mathbf{Disc} \to \mathbf{React}$ is full up to isomorphism in **React**. As for completeness, we spell out the details in the following:

Theorem 2 (Universality). *Given a reaction* $r : A \to C$, *there is a term* $\mathtt{t} : A \to B$ *and an isomorphism* $\iota : B \xrightarrow{\sim} C$ *such that* $R(\mathtt{t}); \iota = r$.

Proof (sketch). Observe that every reaction $r : A \to C$ factorises as

$$(U_A, U_B, \mathrm{id}, \mathrm{id}); (\varnothing, \varnothing, !, \iota),$$

where $(U_A, U_B) : A \to B$ is some reaction and $\iota : B \to C$ is an isomorphism of labelled graphs. Now, we may disconnect all possible bonds inside U_A, and then connect all possible bonds to obtain U_B. The fact that U_A and U_B have the same atom vertices and the same net charge guarantee that this can always be done.

5 Conclusion

We have formalised and axiomatised the retrosynthetic rules as a category, showing that there is a sound, complete and universal translation into the category of reactions – a more familiar object to study in computational chemistry.

Universality can be thought of as a consistency result for reactions: their definition captures exactly those rearrangements of chemical graphs which result from local, chemically motivated rewrite rules. Completeness says that there is no redundancy in the representation: treating the (dis)connection rules as terms, the terms can be endowed with equations such that the terms describing the same reaction are identified. As the decomposition of a reaction into a sequence of (dis)connection rules is algorithmic, these results can be used to automatically break a reaction (or its part) into smaller components: the purpose can be, *inter alia*, retrosynthetic analysis or storing reaction data in a systematic way.

5.1 Future Work

Chemical Questions. An important part of chemical data is stereochemistry, that is, spatial orientation of the molecule: many molecules of interest (like pharmaceuticals) possess chiral enantiomers (i.e. molecules that have the same atoms and connectivity, but are mirror images of each other due to spatial orientation) which have different properties. We therefore wish to incorporate spatial data into our categorical representation. While stereochemistry is relatively straightforward to account for on the level of chemical graphs and reactions [9], it is unclear how to do this for the disconnection rules, as they only operate at one or two vertices at a time. A more straightforward extension of the formalism presented here would introduce energy and dynamics into the disconnection rules by quantifying how much energy each (dis)connection (in a particular context) requires to occur.

Computational Questions. Given the algorithmic nature of both completeness and universality proofs, the next step is to implement both. The first algorithm would take an arbitrary reaction as an input, and output a sequence of disconnection rules representing it. The second algorithm would decide whether two terms are equal or not, implementing the normalisation procedure. Another direction for connecting this work with more standard approaches to computational chemistry would be translating our formalism to a widely used notation such as SMILES [16,13].

Mathematical Questions. An important mathematical development is to introduce monoidal terms in the disconnection category, so as to allow parallel reactions, as well as usage of graphical calculi for monoidal categories [11]. Another mathematical question is whether the categories **Disc** and **React** have any interesting categorical structure, apart from being dagger categories. Finally, we would like to make precise the connection between the category of reactions and double pushout rewriting [1,2,3].

References

1. Andersen, J.L., Flamm, C., Merkle, D., Stadler, P.F.: Inferring chemical reaction patterns using rule composition in graph grammars. Journal of systems chemistry **4**(1), 1–4 (2013). https://doi.org/10.1186/1759-2208-4-4
2. Andersen, J.L., Flamm, C., Merkle, D., Stadler, P.F.: An intermediate level of abstraction for computational systems chemistry. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **375**(2109), 20160354 (2017). https://doi.org/10.1098/rsta.2016.0354
3. Behr, N., Krivine, J., Andersen, J.L., Merkle, D.: Rewriting theory for the life sciences: A unifying theory of ctmc semantics. Theor. Comput. Sci. **884**(C), 68–115 (2021). https://doi.org/10.1016/j.tcs.2021.07.026
4. Bournez, O., Ibănescu, L., Kirchner, H.: From chemical rules to term rewriting. Electronic Notes in Theoretical Computer Science **147**(1), 113–134 (2006). https://doi.org/10.1016/j.entcs.2005.06.040, proceedings of the 6th International Workshop on Rule-Based Programming (RULE 2005)
5. Clayden, J., Greeves, N., Warren, S.: Organic chemistry. Oxford University Press, Oxford, 2nd edn. (2012)
6. Corey, E.J., Cheng, X.: The logic of chemical synthesis. John Wiley, New York (1989)
7. Corey, E.J.: General methods for the construction of complex molecules. Pure and Applied chemistry **14**(1), 19–38 (1967). https://doi.org/10.1351/pac196714010019
8. Fong, B., Spivak, D.I.: An Invitation to Applied Category Theory: Seven Sketches in Compositionality. Cambridge University Press, Cambridge (2019)
9. Gale, E., Lobski, L., Zanasi, F.: A categorical approach to synthetic chemistry. In: Ábrahám, E., Dubslaff, C., Tarifa, S.L.T. (eds.) Theoretical Aspects of Computing – ICTAC 2023. pp. 276–294. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-47963-2_17
10. Heunen, C., Vicary, J.: Categories for Quantum Theory: An Introduction. Oxford University Press (11 2019). https://doi.org/10.1093/oso/9780198739623.001.0001
11. Piedeleu, R., Zanasi, F.: An introduction to string diagrams for computer scientists (2023), https://arxiv.org/abs/2305.08768
12. Selinger, P.: Dagger compact closed categories and completely positive maps: (extended abstract). Electronic Notes in Theoretical Computer Science **170**, 139–163 (2007). https://doi.org/10.1016/j.entcs.2006.12.018, proceedings of the 3rd International Workshop on Quantum Programming Languages (QPL 2005)
13. SMILES tutorial. Website, https://daylight.com/dayhtml_tutorials/languages/smiles/index.html
14. Sun, Y., Sahinidis, N.V.: Computer-aided retrosynthetic design: fundamentals, tools, and outlook. Current Opinion in Chemical Engineering **35**, 100721 (2022). https://doi.org/10.1016/j.coche.2021.100721
15. Warren, S., Wyatt, P.: Organic synthesis : the disconnection approach. Wiley, Hoboken, N.J, 2nd ed. edn. (2008)
16. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences **28**(1), 31–36 (1988). https://doi.org/10.1021/ci00057a005

17. Zhong, Z., Song, J., Feng, Z., Liu, T., Jia, L., Yao, S., Hou, T., Song, M.: Recent advances in deep learning for retrosynthesis. Wiley Interdisciplinary Reviews: Computational Molecular Science **14**(1), e1694 (2024). https://doi.org/10.1002/wcms.1694

A Proofs

We place the proofs here for completeness and for the purposes of peer review. These will not be part of the proceedings publication due to the lack of space. We intend to publish full proofs in a journal version.

A.1 Inductions for *ICE*-form

Lemma 4. *Let* $\mathtt{t}$ *be a term such that the term* $\mathtt{t}; I^{uv}$ *is defined. Then there exists a term* $\mathtt{t}'$ *such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* I*-terms, and one of the following holds:*

(1) $\mathtt{t}; I^{uv} = \mathtt{t}'$*, or*
(2) there is a disconnection I^{ab} *such that* $\mathtt{t}; I^{uv} = I^{ab}; \mathtt{t}'$*.*

Proof. We proceed by induction on the structure of $\mathtt{t}$.

Base cases:

$$
\begin{aligned}
\mathrm{id}; I^{uv} &= I^{uv}; \mathrm{id},\\
S^w; I^{uv} &= I^{uv}; S^w,\\
R^{w\mapsto z}; I^{uv} &= I^{uv}; R^{w\mapsto z},\\
I^{wz}; I^{uv} &\quad \text{is already in the right form}\\
C^{wz}_{ab}; I^{uv} &= I^{uv}; C^{wz}_{ab},\\
E^{w}_{ab}; I^{uv} &= I^{uv}; E^{w}_{ab},\\
E^{wz}; I^{uv} &= I^{uv}; E^{wz},\\
\bar{I}^{wz}; I^{uv} &= \begin{cases} S^u; S^v \text{ if } \{w,z\} = \{u,v\} \\ I^{uv}; \bar{I}^{wz} \text{ otherwise,} \end{cases}\\
\bar{C}^{wz}_{ab}; I^{uv} &= I^{uv}; \bar{C}^{wz}_{ab}\\
\bar{E}^{w}_{ab}; I^{uv} &= I^{uv}; \bar{E}^{w}_{ab},\\
\bar{E}^{wz}; I^{uv} &= I^{uv}; \bar{E}^{wz}.
\end{aligned}
$$

Inductive case: Let $\mathtt{t} : A \to B$ and $\mathtt{s} : B \to C$ be terms such that the statement of the lemma holds. Suppose that the term $\mathtt{t}; \mathtt{s}; I^{uv}$ is defined. Then also the term $\mathtt{s}; I^{uv}$ is defined, so by the inductive hypothesis for $\mathtt{s}$, there is a term $\mathtt{s}'$ with the same number of I-terms as $\mathtt{s}$ such that either (1) $\mathtt{s}; I^{uv} = \mathtt{s}'$, or (2) $\mathtt{s}; I^{uv} = I^{ab}; \mathtt{s}'$ for some I-term I^{ab}. In the first case, we have

$$\mathtt{t}; \mathtt{s}; I^{uv} = \mathtt{t}; \mathtt{s}',$$

and since $\mathtt{t};\mathtt{s}'$ has the same number of I-terms as $\mathtt{t};\mathtt{s}$, it is the sought-after term for the inductive case satisfying (1). In the second case, we have

$$\mathtt{t};\mathtt{s};I^{uv} = \mathtt{t};I^{ab};\mathtt{s}',$$

so that $\mathtt{t};I^{ab}$ is defined. By the inductive hypothesis for $\mathtt{t}$, there is a term $\mathtt{t}'$ with the same number of I-terms as $\mathtt{t}$ such that either (1) $\mathtt{t};I^{ab} = \mathtt{t}'$, or (2) $\mathtt{t};I^{ab} = I^{wz};\mathtt{t}'$ for some I-term I^{wz}. In the first case, we get

$$\mathtt{t};\mathtt{s};I^{uv} = \mathtt{t}';\mathtt{s}',$$

satisfying (1) for the inductive case, as $\mathtt{t}';\mathtt{s}'$ and $\mathtt{t};\mathtt{s}$ have the same number of I-terms. In the second case, we obtain

$$\mathtt{t};\mathtt{s};I^{uv} = I^{wz};\mathtt{t}';\mathtt{s}',$$

satisfying (2) for the inductive case.

Corollary 1. *Any term is equal to a term of the form* $\mathtt{I};\mathtt{t}$*, where* $\mathtt{I}$ *is a sequence of* I*-terms, and the term* $\mathtt{t}$ *contains no* I*-terms.*

Lemma 5. *Let* $\mathtt{t}$ *be a term not containing any* I*-terms such that the term* $\mathtt{t};C^{uv}_{ab}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* C*-terms, and one of the following holds:*

(1) $\mathtt{t};C^{uv}_{ab} = \mathtt{t}'$*, or*
(2) there is a disconnection C^{wz}_{cd} *such that* $\mathtt{t};C^{uv}_{ab} = C^{wz}_{cd};\mathtt{t}'$*.*

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id};C^{uv}_{ab} &= C^{uv}_{ab};\mathrm{id},\\
S^w;C^{uv}_{ab} &= C^{uv}_{ab};S^w,\\
R^{w\mapsto z};C^{uv}_{ab} &= \begin{cases} C^{uv}_{ab};R^{w\mapsto z} \text{ if } w\neq\{a,b\},\\ C^{uv}_{kb};R^{a\mapsto z};R^{k\mapsto a} \text{ if } w=a,\\ C^{uv}_{ak};R^{b\mapsto z};R^{k\mapsto b} \text{ if } w=b,\end{cases}\\
C^{wz}_{cd};C^{uv}_{ab} &\quad \text{is already in the right form},\\
E^w_{cd};C^{uv}_{ab} &= C^{uv}_{ab};E^w_{cd},\\
E^{wz};C^{uv}_{ab} &= C^{uv}_{ab};E^{wz},\\
\bar{I}^{wz};C^{uv}_{ab} &= C^{uv}_{ab};\bar{I}^{wz},\\
\bar{C}^{wz}_{cd};C^{uv}_{ab} &= C^{uv}_{ij};\bar{C}^{wz}_{cd};R^{i\mapsto a};R^{j\mapsto b},\\
\bar{E}^w_{cd};C^{uv}_{ab} &= C^{uv}_{ij};\bar{E}^w_{cd};R^{i\mapsto a};R^{j\mapsto b},\\
\bar{E}^{wz};C^{uv}_{ab} &= C^{uv}_{ab};\bar{E}^{wz}.
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 2. *Any term is equal to a term of the form* $\mathtt{I};\mathtt{C};\mathtt{t}$*, where* $\mathtt{I}$ *and* $\mathtt{C}$ *are sequences of* I*-terms and* C*-terms, and the term* $\mathtt{t}$ *contains no* I*-terms or* C*-terms.*

Lemma 6. *Let* $\mathtt{t}$ *be a term not containing any* I*- or* C*-terms such that the term* $\mathtt{t};E^u_{ab}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*- or* C*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $E^{<0}$*-terms, and one of the following holds:*

(1) $\mathtt{t};E^u_{ab}=\mathtt{t}'$*, or*
(2) there is a disconnection E^w_{cd} *such that* $\mathtt{t};E^u_{ab}=E^w_{cd};\mathtt{t}'$*.*

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id};E^u_{ab} &= E^u_{ab};\mathrm{id},\\
S^w;E^u_{ab} &= E^u_{ab};S^w,\\
R^{w\mapsto z};E^u_{ab} &= \begin{cases} E^u_{ab};R^{w\mapsto z} \text{ if } w\neq\{a,b\},\\ E^u_{kb};R^{a\mapsto z};R^{k\mapsto a} \text{ if } w=a,\\ E^u_{ak};R^{b\mapsto z};R^{k\mapsto b} \text{ if } w=b,\end{cases}\\
E^w_{cd};E^u_{ab} &\quad \text{is already in the right form},\\
E^{wz};E^u_{ab} &= E^u_{ab};E^{wz},\\
\bar{I}^{wz};E^u_{ab} &= E^u_{ab};\bar{I}^{wz},\\
\bar{C}^{wz}_{cd};E^u_{ab} &= E^u_{ij};\bar{C}^{wz}_{cd};R^{i\mapsto a};R^{j\mapsto b},\\
\bar{E}^w_{cd};E^u_{ab} &= E^u_{ij};\bar{E}^u_{cd};R^{i\mapsto a};R^{j\mapsto b},\\
\bar{E}^{wz};E^u_{ab} &= E^u_{ab};\bar{E}^{wz}.
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 3. *Any term is equal to a term of the form* $\mathtt{I};\mathtt{C};\mathtt{E}^{<0};\mathtt{t}$*, where* $\mathtt{I}$*,* $\mathtt{C}$ *and* $\mathtt{E}^{<0}$ *are sequences of* I*-,* C*-, and* $E^{<0}$*-terms, and the term* $\mathtt{t}$ *contains no* I*-,* C*-, or* $E^{<0}$*-terms.*

Lemma 7. *Let* $\mathtt{t}$ *be a term not containing any* I*-,* C*-, or* $E^{<0}$*-terms such that the term* $\mathtt{t};E^{uv}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-,* C*-, or* $E^{<0}$*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $E^{\geq 0}$*-terms, and one of the following holds:*

(1) $\mathtt{t};E^{uv}=\mathtt{t}'$*, or*
(2) there is a disconnection E^{wz} *such that* $\mathtt{t};E^{uv}=E^{wz};\mathtt{t}'$*.*

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id}; E^{uv} &= E^{uv}; \mathrm{id},\\
S^w; E^{uv} &= E^{uv}; S^w,\\
R^{w\mapsto z}; E^{uv} &= \begin{cases} E^{uv}; R^{w\mapsto z} \text{ if } z \neq v,\\ E^{uw}; R^{w\mapsto v} \text{ if } z = v, \end{cases}\\
E^{wz}; E^{uv} &\quad \text{is already in the right form},\\
\bar{I}^{wz}; E^{uv} &= E^{uv}; \bar{I}^{wz},\\
\bar{C}^{wz}_{cd}; E^{uv} &= E^{uv}; \bar{C}^{wz}_{cd},\\
\bar{E}^{w}_{cd}; E^{uv} &= E^{uv}; \bar{E}^{w}_{cd},\\
\bar{E}^{wz}; E^{uv} &= \begin{cases} S^u; S^v \text{ if } w = u \text{ and } z = v,\\ E^{uv}; \bar{E}^{wz} \text{ otherwise.} \end{cases}
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 4. *Any term is equal to a term of the form* $\mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \mathtt{t}$*, where* $\mathtt{I}$*,* $\mathtt{C}$*,* $\mathtt{E}^{<0}$ *and* $\mathtt{E}^{\geq 0}$ *are sequences of* I*-,* C*-,* $E^{<0}$*, and* $E^{\geq 0}$*-terms, and the term* $\mathtt{t}$ *contains no* I*-,* C*-,* $E^{<0}$*, or* $E^{\geq 0}$*-terms.*

Lemma 8. *Let* $\mathtt{t}$ *be a term not containing any* I*-,* C*- or* E*-terms such that the term* $\mathtt{t}; \bar{E}^{uv}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-,* C*- or* E*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $\bar{E}^{\geq}$*-terms, and one of the following holds:*

(1) $\mathtt{t}; \bar{E}^{uv} = \mathtt{t}'$*, or*
(2) there is a connection $\bar{E}^{wz}$ *such that* $\mathtt{t}; \bar{E}^{uv} = \bar{E}^{wz}; \mathtt{t}'$*.*

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id}; \bar{E}^{uv} &= \bar{E}^{uv}; \mathrm{id},\\
S^w; \bar{E}^{uv} &= \bar{E}^{uv}; S^w,\\
R^{w\mapsto z}; \bar{E}^{uv} &= \begin{cases} \bar{E}^{uv}; R^{w\mapsto z} \text{ if } z \neq v,\\ \bar{E}^{uw}; R^{w\mapsto v} \text{ if } z = v, \end{cases}\\
\bar{I}^{wz}; \bar{E}^{uv} &= \bar{E}^{uv}; \bar{I}^{wz},\\
\bar{C}^{wz}_{cd}; \bar{E}^{uv} &= \bar{E}^{uv}; \bar{C}^{wz}_{cd},\\
\bar{E}^{w}_{cd}; \bar{E}^{uv} &= \bar{E}^{uv}; \bar{E}^{w}_{cd},\\
\bar{E}^{wz}; \bar{E}^{uv} &\quad \text{is already in the right form.}
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 5. *Any term is equal to a term of the form* $\mathtt{I};\mathtt{C};\mathtt{E}^{<0};\mathtt{E}^{\geq 0};\bar{\mathtt{E}}^{\geq 0};\mathtt{t}$, *where the term* $\mathtt{t}$ *contains no* I*-,* C*-,* E*-, or* $\bar{E}^{\geq 0}$*-terms.*

Lemma 9. *Let* $\mathtt{t}$ *be a term not containing any* I*-,* C*-,* E*-, or* $\bar{E}^{\geq 0}$*-terms such that the term* $\mathtt{t};\bar{E}^u_{ab}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-,* C*-,* E*-, or* $\bar{E}^{\geq 0}$*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $\bar{E}^{<0}$*-terms, and one of the following holds:*

(1) $\mathtt{t};\bar{E}^u_{ab} = \mathtt{t}'$*, or*
(2) there is a connection $\bar{E}^w_{cd}$ *such that* $\mathtt{t};\bar{E}^u_{ab} = \bar{E}^w_{cd};\mathtt{t}'$.

Proof. By induction on $\mathtt{t}$.

Base cases:

$$
\begin{aligned}
\mathrm{id};\bar{E}^u_{ab} &= \bar{E}^u_{ab};\mathrm{id},\\
S^w;\bar{E}^u_{ab} &= \bar{E}^u_{ab};S^w,\\
R^{w\mapsto z};\bar{E}^u_{ab} &= \begin{cases} \bar{E}^u_{ab};R^{w\mapsto z} \text{ if } z\notin\{a,b\},\\ \bar{E}^u_{wb} \text{ if } z=a,\\ \bar{E}^u_{aw} \text{ if } z=b,\end{cases}\\
\bar{I}^{wz};\bar{E}^u_{ab} &= \bar{E}^u_{ab};\bar{I}^{wz},\\
\bar{C}^{wz}_{cd};\bar{E}^u_{ab} &= \bar{E}^u_{ab};\bar{C}^{wz}_{cd},\\
\bar{E}^w_{cd};\bar{E}^u_{ab} &\quad \text{is already in the right form.}
\end{aligned}
$$

The inductive case is very similar to that of Lemma 4.

Corollary 6. *Any term is equal to a term of the form* $\mathtt{I};\mathtt{C};\mathtt{E}^{<0};\mathtt{E}^{\geq 0};\bar{\mathtt{E}}^{\geq 0};\bar{\mathtt{E}}^{<0};\mathtt{t}$, *where the term* $\mathtt{t}$ *contains no* I*-,* C*-,* E*- or* $\bar{E}$*-terms.*

Lemma 10. *Let* $\mathtt{t}$ *be a term not containing any* I*-,* C*-,* E*- or* $\bar{E}$*-terms such that the term* $\mathtt{t};\bar{C}^{uv}_{ab}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-,* C*-,* E*- or* $\bar{E}$*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $\bar{C}$*-terms, and one of the following holds:*

(1) $\mathtt{t};\bar{C}^{uv}_{ab} = \mathtt{t}'$*, or*
(2) there is a connection $\bar{C}^{wz}_{cd}$ *such that* $\mathtt{t};\bar{C}^{uv}_{ab} = \bar{C}^{wz}_{cd};\mathtt{t}'$.

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id}; \bar{C}^{uv}_{ab} &= \bar{C}^{uv}_{ab}; \mathrm{id}, \\
S^w; \bar{C}^{uv}_{ab} &= \bar{C}^{uv}_{ab}; S^w, \\
R^{w\mapsto z}; \bar{C}^{uv}_{ab} &= \begin{cases} \bar{C}^{uv}_{ab}; R^{w\mapsto z} \text{ if } z \notin \{a, b\}, \\ \bar{C}^{uv}_{wb} \text{ if } z = a, \\ \bar{C}^{uv}_{aw} \text{ if } z = b, \end{cases} \\
\bar{I}^{wz}; \bar{C}^{uv}_{ab} &= \bar{C}^{uv}_{ab}; \bar{I}^{wz}, \\
\bar{C}^{wz}_{cd}; \bar{C}^{uv}_{ab} &\quad \text{is already in the right form.}
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 7. *Any term is equal to a term of the form* $\mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \bar{\mathtt{E}}^{\geq 0}; \bar{\mathtt{E}}^{<0}; \bar{\mathtt{C}}; \mathtt{t}$, *where the term* $\mathtt{t}$ *contains no* I*-,* C*-,* E*-,* $\bar{E}$*- or* $\bar{C}$*-terms.*

Lemma 11. *Let* $\mathtt{t}$ *be a term not containing any* I*-,* C*-,* E*-,* $\bar{E}$*- or* $\bar{C}$*-terms such that the term* $\mathtt{t}; \bar{I}^{uv}$ *is defined. Then there exists a term* $\mathtt{t}'$ *not containing any* I*-,* C*-,* E*-,* $\bar{E}$*- or* $\bar{C}$*-terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* $\bar{I}$*-terms, and one of the following holds:*

(1) $\mathtt{t}; \bar{I}^{uv} = \mathtt{t}'$*, or*
(2) there is a connection $\bar{I}^{ab}$ *such that* $\mathtt{t}; \bar{I}^{uv} = \bar{I}^{ab}; \mathtt{t}'$.

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned}
\mathrm{id}; \bar{I}^{uv} &= \bar{I}^{uv}; \mathrm{id}, \\
S^w; \bar{I}^{uv} &= \bar{I}^{uv}; S^w, \\
R^{w\mapsto z}; \bar{I}^{uv} &= \bar{I}^{uv}; R^{w\mapsto z}, \\
\bar{I}^{wz}; \bar{I}^{uv} &\quad \text{is already in the right form.}
\end{aligned}$$

The inductive case is very similar to that of Lemma 4.

Corollary 8. *Any term is equal to a term of the form* $\mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \bar{\mathtt{E}}^{\geq 0}; \bar{\mathtt{E}}^{<0}; \bar{\mathtt{C}}; \bar{\mathtt{I}}; \mathtt{t}$, *where the term* $\mathtt{t}$ *contains only* S*-,* R*-, and identity terms.*

Lemma 12. *Let* $\mathtt{t}$ *be a term containing only* S*-,* R*-, and identity terms such that the term* $S^u; \mathtt{t}$ *is defined. Then there exists a term* $\mathtt{t}'$ *containing only* S*-,* R*-, and identity terms such that* $\mathtt{t}$ *and* $\mathtt{t}'$ *have the same number of* S*-terms, and one of the following holds:*

(1) $S^u; \mathtt{t} = \mathtt{t}'$*, or*
(2) there is a term S^v *such that* $S^u; \mathtt{t} = \mathtt{t}'; S^v$.

Proof. By induction on $\mathtt{t}$.

Base cases:

$$\begin{aligned} S^u; \mathrm{id} &= \mathrm{id}; S^u, \\ S^u; S^w &\quad \text{is already in the right form,} \\ S^u; R^{w \mapsto z} &= \begin{cases} R^{w \mapsto z}; S^u \text{ if } u \neq w, \\ R^{u \mapsto z} \text{ if } u = w. \end{cases} \end{aligned}$$

The inductive case is very similar to that of Lemma 4.

A.2 Proofs for Section 3

Proposition 1. *The following identities are derivable in* **Disc***:*

$$d^U_{D[a]}; S^a = d^U_{D[a]}, \tag{35}$$

$$\bar{d}^U_{ab}; d^U_{cd} \leqq S^U; R^{a \mapsto c}; R^{b \mapsto d}, \tag{36}$$

$$R^{z \mapsto c}; R^{w \mapsto d}; d^U_{ab} \simeq R^{z \mapsto a}; R^{w \mapsto b}; d^U_{cd}, \tag{37}$$

$$R^{z \mapsto c}; d^U_{ab} \simeq R^{z \mapsto a}; d^U_{cb}, \tag{38}$$

$$R^{w \mapsto d}; d^U_{ab} \simeq R^{w \mapsto b}; d^U_{ad}, \tag{39}$$

$$d^U_{ab}; d^U_{cd} = d^U_{ad}; d^U_{cb}. \tag{40}$$

Proof. We compute (35) as

$$d^U_{D[a]}; S^a = d^U_{D[a]}; R^{a \mapsto a} = d^U_{D[a]}.$$

Equality (36) is derived as follows:

$$\begin{aligned} \bar{d}^{uw}_{ab}; d^{uw}_{cd} &\simeq d^{uw}_{ij}; \bar{d}^{uw}_{ab}; R^{i \mapsto c}; R^{j \mapsto d} && \text{(by (10))} \\ &\leqq S^u; R^{a \mapsto i}; R^{b \mapsto j}; R^{i \mapsto c}; R^{j \mapsto d} && \text{(by (11))} \\ &\leqq S^u; R^{a \mapsto c}; R^{b \mapsto d}. && \text{(by (1))} \end{aligned}$$

For (37), we first use (36) to get

$$d^{uv}_{cd}; \bar{d}^{uv}_{zw}; d^{uv}_{ab} \simeq d^{uv}_{cd}; S^u; R^{z \mapsto a}; R^{w \mapsto b},$$

where $z \neq c$ and $w \neq d$. Observing that (11) applies on the left-hand side, and simplifying using (5), (21) and (7), we obtain precisely (37). Identities (38) and (39) are derived similarly, by letting $w = d$ and $z = c$, respectively.

Identity (40) is derived as follows:

$$\begin{aligned} d^U_{ab}; d^U_{cd} &= d^U_{ad}; R^{d \mapsto b}; d^U_{cd} && \text{(by (9))} \\ &= d^U_{ad}; R^{d \mapsto d}; d^U_{cb} && \text{(by (39))} \\ &= d^U_{ad}; d^U_{cb}. && \text{(by (9))} \end{aligned}$$

Proposition 2. *Any term is equal to a term in an ICE-form.*

Proof. This follows from the sequence of Lemmas and Corollaries 4-12 in Subsection A.1.

Lemma 1. *Any well-typed sequence of renaming terms is equal to a term* $\mathtt{R};\mathtt{S}$*, where* $\mathtt{R} = \mathtt{A};\mathtt{B}$ *is in a renaming form and* $\mathtt{S}$ *is a sequence of* S*-terms.*

Proof. We proceed by induction on the length of the sequence.

The term $R^{a\mapsto b}$ is equal to S^a if $a = b$, or is already in a renaming form by taking $A = \{a\}$, $B = \{b\}$ and $C = D = \varnothing$ if $a \neq b$.

Suppose that the statement of the lemma holds for all sequences of renaming terms of length at most n. Let $\mathtt{R}$ be such a sequence of length n such that $\mathtt{R};R^{a\mapsto b}$ is defined. By the inductive hypothesis, we may assume that $\mathtt{R} = \mathtt{A};\mathtt{B};\mathtt{S}$ where $\mathtt{A};\mathtt{B}$ is a renaming form with vertex name sets A, B, C and D as in Definition 11. Using the equations for S- and R-terms, we may commute $R^{a\mapsto b}$ past $\mathtt{S}$, possibly changing the vertex name a, so that it suffices to show that the lemma holds for $\mathtt{A};\mathtt{B};R^{a\mapsto b}$. If $a = b$, the sequence is equal to $\mathtt{A};\mathtt{B};S^a$ and we are done; hence assume that $a \neq b$. Note that it follows that $a \notin A \setminus D$ and $a \notin C$, as every vertex name in $A \setminus D$ or C is removed, without being reintroduced. Similarly, we have that $b \notin D$ and $b \notin B \setminus C$. Moreover, if $b \in C$, rename the occurrence of b in both $\mathtt{A}$ and $\mathtt{B}$ with a fresh vertex name, updating the sets C and B accordingly. Thus we may assume that $b \notin B$. The remaining cases are as follows.

Case 1: $a \notin A \cup B$.

Subcase 1.1: $b \notin A$. We rewrite the term to $\mathtt{A};R^{a\mapsto b};\mathtt{B}$ and update the sets $A \mapsto A \cup \{a\}$ and $B \mapsto B \cup \{b\}$.

Subcase 1.2: $b \in A$. It follows that $b \in A \setminus D$, so that $R^{b\mapsto z} \in \mathtt{A}$ and a, b do not appear in $\mathtt{B}$. If $\mathtt{N}(a) \neq \mathtt{N}(b)$, let c be a fresh vertex name. We rewrite the term to $\mathtt{A};R^{a\mapsto c};R^{c\mapsto b};\mathtt{B}$ and update the sets $A \mapsto A \cup \{a\}$, $B \mapsto B \cup \{c\}$, $C \mapsto C \cup \{c\}$ and $D \mapsto D \cup \{b\}$. If $\mathtt{N}(a) = \mathtt{N}(b)$, we use equation (4) to rewrite $R^{b\mapsto z};R^{a\mapsto b}$ to $S^b;R^{a\mapsto z}$, which reduces the number of R-terms to n, so that the inductive hypothesis applies.

Case 2: $a \in A$. It follows that $a \in D$. Now $R^{a\mapsto b}$ commutes with all other terms in $\mathtt{B}$ except for the unique term $R^{c_i\mapsto d_i}$ where $d_i = a$. But $R^{c_i\mapsto a};R^{a\mapsto b} = R^{c_i\mapsto b}$, which reduces the length of the sequence to n, so it is has a renaming form by the inductive hypothesis.

Case 3: $a \in B$. It follows that $a \in B \setminus C$.

Subcase 3.1: $b \notin A$. Now $R^{a\mapsto b}$ commutes with all the terms in $\mathtt{B}$, and with all other terms in $\mathtt{A}$ except for the unique term $R^{a_i\mapsto b_i}$ where $b_i = a$. But $R^{a_i\mapsto a};R^{a\mapsto b} = R^{a_i\mapsto b}$, which reduces the length of the sequence to n, so it is has a renaming form by the inductive hypothesis.

Subcase 3.2: $b \in A$. It follows that $b \in A \setminus D$. Now $R^{a\mapsto b}$ commutes with all the terms in $\mathtt{B}$, and with all other terms in $\mathtt{A}$ except for the terms $R^{a_i\mapsto a}$ and $R^{b\mapsto b_j}$. There are two options: (1) $a_i = b$ and $b_j = a$, so that these are the same term, (2) the terms are distinct, in which case they commute. In both cases, we use the substitution $R^{a_i\mapsto a};R^{a\mapsto b} = R^{a_i\mapsto b}$ to reduce the length of the sequence, so that the inductive hypothesis applies.

This completes the induction.

Proposition 3. *Any term is equal to a term in normal form.*

Proof. By Proposition 2 and Lemma 1, every terms is equal to a term in an $ICER$-form: let us fix such a term

$$\mathtt{t} = \mathtt{I}; \mathtt{C}; \mathtt{E}^{<0}; \mathtt{E}^{\geq 0}; \bar{\mathtt{E}}^{\geq 0}; \bar{\mathtt{E}}^{<0}; \bar{\mathtt{C}}; \bar{\mathtt{I}}; \mathtt{A}; \mathtt{B}; \mathtt{S}.$$

Conditions (1) and (2) are obtained by absorbing the "excess" S-terms into other terms using equations (6), (19) and (21). Conditions (3) and (4) are obtained by treating all the vertex names in $D_{\mathtt{t}}^{add}$ as "dummy" names, which are removed either by a connection or a renaming term.

For (5), suppose that $\bar{d}^U_{D[a]} : A \to B$ and $R^{z \mapsto a}$ both occur, and moreover A is in the domain of $\bar{d}^U_{D[z/a]}$. We commute the renaming term to the left to obtain $\bar{d}^U_{D[a]}; R^{z \mapsto a}$. But this is equal to $\bar{d}^U_{D[z/a]}; R^{a \mapsto a}$ by equations (38) and (39), which gets rid of the renaming term by $R^{a \mapsto a} = S^a$.

For (6) and (7), we consider the three cases separately.

Case 1: E^{uv} and $\bar{E}^{uw}$ occur in $\mathtt{t}$. We commute the terms so that they occur one after the other $E^{uv}; \bar{E}^{uw}$, which by equations (14) and (14) is equal to some combination of S- and R-terms. Noticing that the rewriting procedure to obtain an ICE-form either commutes or absorbs S- and R-terms (specifically, the results from Lemma 8 onwards apply), we conclude that the term has an ICE-form without E^{uv} and $\bar{E}^{uw}$.

Case 2: E^u_{ab} and $\bar{E}^u_{cd}$ occur in $\mathtt{t}$. In combination with Case 1, it follows that the terms E^{uv} and $\bar{E}^{uw}$ do not occur, so that there are no obstruction for commuting E^u_{ab} next to $\bar{E}^u_{cd}$, obtaining $E^u_{ab}; \bar{E}^u_{cd}$. By equations (11), (12) and (13), this is equal to some combination of S- and R-terms, which we commute to the right as in Case 1.

Case 3: C^{uv}_{ab} and either $\bar{C}^{uv}_{cd}$ or $\bar{C}^{vu}_{cd}$ occur in $\mathtt{t}$. The latter case simply reduces to the former by equation (22). Thus suppose that $\bar{C}^{uv}_{cd}$ occurs. First, we use equation (30) to commute C^{uv}_{ab} to the right past all the $E^{<0}$-terms, and $\bar{C}^{uv}_{cd}$ to the left past all the $\bar{E}^{<0}$-terms. Next, we commute any terms of the form E^{ui} and E^{vj} to the right past the $\bar{E}^{\geq 0}$-terms and $\bar{C}^{uv}_{cd}$ using the fact that by Case 1 the terms $\bar{E}^{ui}$ and $\bar{E}^{vj}$ do not occur, together with the equations (17) and (32). Now there are no obstructions for commuting C^{uv}_{ab} to the right past all the $E^{\geq 0}$- and $\bar{E}^{\geq 0}$-terms, obtaining $C^{uv}_{ab}; \bar{C}^{uv}_{cd}$. As in Case 2, equations (11), (12) and (13) yield that this is equal to some combination of S- and R-terms, which we commute to the right as in Case 1. Finally, we return the terms of the form E^{ui} and E^{vj} back to the left past all the $\bar{E}^{\geq 0}$-terms.

For (8), suppose that both I^{uv} and $\bar{I}^{uv}$ occur in $\mathtt{t}$ such that no E- or $\bar{E}$ term containing u occurs. It follows that no E- or $\bar{E}$ term containing v occurs either: application of $\bar{I}^{uv}$ requires for u and v to have equal and opposite charge, so if the charge of u is unchanged, so is the charge of v; moreover, by (6) and (7), we may assume that no change can be reversed, so that we indeed cannot have any E- or $\bar{E}$ term containing v. But now there are no obstructions for commuting I^{uv} all to the right until we obtain $I^{uv}\bar{I}^{uv}$, which is equal to S^u.

A.3 Proofs for Section 4

Proposition 4. *The assignment* $R : \mathbf{Disc} \to \mathbf{React}$ *is a dagger functor.*

Proof. Functoriality and preservation of dagger structure follow immediately from the definition. We have to show that R preserves the equations of Figure 4. Most of these follow immediately by applying R to expressions on both sides of the equality, hence we only give the cases that are less obvious or require more computation. To further simplify the notation, we omit the curly brackets of set-builder notation as well as the commas separating vertex names from each other: so e.g. $(uv, uvab)$ stands for $(\{u, v\}, \{u, v, a, b\})$.

For (7), let us write $R(d^U_D) = (R_1, R_2)$. Since $u, v \notin U \cup D$, we have that $u, v \notin R_1, R_2$. We use this to show that both sides of the equality evaluate to the same map:

$$(u, v); (R_1, R_2) = (\{u\} \cup R_1, R_2 \cup \{v\}) = (R_1, R_2); (u, v).$$

For (10), denote $R(d^{U'}_{ij}) = (u'v', u'v'ij)$, $R(d^{U'}_{cd}) = (u'v', u'v'cd)$ and $R(\bar{h}^U_{ab}) = (uvab, uv)$. Note that d and h are not $E^{\geq 0}$-terms, whence it follows that $i, j \notin U$ and $a, b \notin U'$. From the fact that the left-hand side is defined, we obtain that $\{i, j\}$ and $\{a, b\}$ are disjoint. The left-hand side is thus translated to

$$\begin{aligned}(u'v', u'v'ij); (uvab, uv); (i, c); (j, d) &= (u'v'uvab, uvu'v'ij); (i, c); (j, d)\\ &= (u'v'uvab, cuvu'v'j); (j, d)\\ &= (u'v'uvab, dcuvu'v')\\ &= (uvu'v'ab, uvu'v'cd)\\ &= (uvab, uv); (u'v', u'v'cd),\end{aligned}$$

which we recognise as the translation of the right-hand side.

For (16), we start from the translation of the right-hand side:

$$\begin{aligned}(u, u); (a, z); (b, a); (z, b) &= (ua, zu); (bz, ba)\\ &= (uab, bau)\\ &= (uab, uba)\\ &= (ua, ua); (ub, ub),\end{aligned}$$

which we recognise as the translation of the left-hand side.

For (23), write $R(d^U_D) = (U, U \cup D)$ and $R(d^{U'}_{D'}) = (U', U' \cup D')$, so that we get

$$R(d^U_D; d^{U'}_{D'}) = (U \cup U', U \cup U' \cup D \cup D') = R(d^{U'}_{D'}; d^U_D).$$

Lemma 13. *Let* $\mathtt{t}$ *and* $\mathtt{s}$ *be terms in a normal form such that* $R(\mathtt{t}) = R(\mathtt{s})$. *Then* $\mathtt{t} \sim \mathtt{s}$.

Proof. Let us write

$$\mathtt{t} = \mathtt{I}_\mathtt{t}; \mathtt{C}_\mathtt{t}; \mathtt{E}_\mathtt{t}^{<0}; \mathtt{E}_\mathtt{t}^{\geq 0}; \bar{\mathtt{E}}_\mathtt{t}^{\geq 0}; \bar{\mathtt{E}}_\mathtt{t}^{<0}; \bar{\mathtt{C}}_\mathtt{t}; \bar{\mathtt{I}}_\mathtt{t}; \mathtt{A}_\mathtt{t}; \mathtt{B}_\mathtt{t}; \mathtt{S}_\mathtt{t},$$
$$\mathtt{s} = \mathtt{I}_\mathtt{s}; \mathtt{C}_\mathtt{s}; \mathtt{E}_\mathtt{s}^{<0}; \mathtt{E}_\mathtt{s}^{\geq 0}; \bar{\mathtt{E}}_\mathtt{s}^{\geq 0}; \bar{\mathtt{E}}_\mathtt{s}^{<0}; \bar{\mathtt{C}}_\mathtt{s}; \bar{\mathtt{I}}_\mathtt{s}; \mathtt{A}_\mathtt{s}; \mathtt{B}_\mathtt{s}; \mathtt{S}_\mathtt{s}.$$

Similarly, let us denote the vertex name sets in the respective renaming forms by $A_\mathtt{t}, B_\mathtt{t}, C_\mathtt{t}, D_\mathtt{t}$ and $A_\mathtt{s}, B_\mathtt{s}, C_\mathtt{s}, D_\mathtt{s}$. Let us denote the morphism $R(\mathtt{t}) = R(\mathtt{s})$ by $(R_1, R_2) : A \to B$.

First, we observe that if $E^{ua} \in \mathtt{E}_\mathtt{t}^{\geq 0}$, then condition (7) of a normal form implies that the charge of u cannot be increased, whence there is a vertex name $b \in \mathbf{VN}$ such that $E^{ub} \in \mathtt{E}_\mathtt{s}^{\geq 0}$. Similarly, by condition (6), if $E^u_{ab} \in \mathtt{E}_\mathtt{t}^{<0}$, then $E^u_{cd} \in \mathtt{E}_\mathtt{s}^{<0}$ for some $c, d \in \mathbf{VN}$; and if $C^{uv}_{ab} \in \mathtt{C}_\mathtt{t}$, then $C^{uv}_{cd} \in \mathtt{C}_\mathtt{s}$ for some $c, d \in \mathbf{VN}$. By condition (8), if $I^{uv} \in \mathtt{I}_\mathtt{t}$ then $I^{uv} \in \mathtt{I}_\mathtt{s}$. Since the connections cannot be undoing the disconnections, a similar inclusion up to α-vertices holds for them. Thus we obtain that the sequences of disconnections and connections must coincide, up to renaming the α-vertices.

Next, suppose that $R^{a \mapsto b} \in \mathtt{A}_\mathtt{t}$, so that $a \in A_\mathtt{t}$ and $b \in B_\mathtt{t}$. There are four cases.

Case 1: $a \notin D^{add}_\mathtt{t}$ and $b \notin C_\mathtt{t}$. This implies that $a \in R_1$ and $b \in R_2$. Moreover, if $b \in R_1$, then condition (5) yields that $\mathtt{N}(a) \neq \mathtt{N}(b)$. It follows that either $R^{a \mapsto b} \in \mathtt{A}_\mathtt{s}$, or both $\bar{d}^U_{D[a]}$ and $R^{z \mapsto b}$ occur in $\mathtt{s}$ such that $\bar{d}^U_{D[z/a]}$ is defined. But in the latter case the vertex names a and z may be exchanged by syntactic manipulation 5, so that we may assume $R^{a \mapsto b} \in \mathtt{A}_\mathtt{s}$.

Case 2: $a \notin D^{add}_\mathtt{t}$ and $b \in C_\mathtt{t}$. This means that $R^{b \mapsto d} \in \mathtt{B}_\mathtt{t}$ for some $d \in D_\mathtt{t}$, and for some $x \in \mathbf{VN}$, we have $R^{d \mapsto x} \in \mathtt{A}_\mathtt{t}$. Condition (3) implies that $d \notin D^{add}_\mathtt{t}$, so that we have $a \in R_1$ and $d \in R_1 \cap R_2$. If $x \notin C_\mathtt{t}$, then by Case 1, $R^{d \mapsto x} \in \mathtt{A}_\mathtt{s}$, so that also $R^{a \mapsto z} \in \mathtt{A}_\mathtt{s}$ and $R^{z \mapsto d} \in \mathtt{B}_\mathtt{s}$ for some $z \in \mathbf{VN}$. If $x \in C_\mathtt{t}$, then we inductively repeat Case 2.

Case 3: $a \in D^{add}_\mathtt{t}$ and $b \notin C_\mathtt{t}$. Thus there is a disconnection $d^U_{D[a]} \in \mathtt{t}$, so that $d^U_{D[x/a]} \in \mathtt{s}$. This implies $a \notin R_1 \cup R_2$ and $b \in R_2$. As in Case 1, it follows that $R^{z \mapsto b} \in \mathtt{A}_\mathtt{s}$ for some $z \in \mathbf{VN}$. Moreover, in this case we must have $\mathtt{N}(x) = \mathtt{N}(z)$, so that we may assume that $R^{x \mapsto b} \in \mathtt{A}_\mathtt{s}$ where $x \in D^{add}_\mathtt{s}$.

Case 4: $a \in D^{add}_\mathtt{t}$ and $b \in C_\mathtt{t}$. This means that $R^{b \mapsto d} \in \mathtt{B}_\mathtt{t}$ for some $d \in D_\mathtt{t}$, so that $R^{d \mapsto x} \in \mathtt{A}_\mathtt{t}$ for some $x \in \mathbf{VN}$. Condition (3) implies that $d \notin D^{add}_\mathtt{t}$, so that either Case 1 or Case 2 applies. In both cases we conclude that $R^{d \mapsto z} \in \mathtt{A}_\mathtt{s}$ for some $z \in \mathbf{VN}$. Since there is a disconnection $d^U_{D[a]} \in \mathtt{t}$, we have $d^U_{D[y/a]} \in \mathtt{s}$. Thus $R^{w \mapsto d} \in \mathtt{B}_\mathtt{s}$ for some $w \in \mathbf{VN}$, and consequently we may assume that $R^{y \mapsto w} \in \mathtt{A}_\mathtt{s}$ with $y \in D^{add}_\mathtt{s}$ and $w \in C_\mathtt{s}$.

Thus, the sequences of renaming terms in $\mathtt{t}$ and $\mathtt{s}$ have the same structure: (1) $A_\mathtt{t} \setminus D^{add}_\mathtt{t} = A_\mathtt{s} \setminus D^{add}_\mathtt{s}$, (2) $\left|A_\mathtt{t} \cap D^{add}_\mathtt{t}\right| = \left|A_\mathtt{s} \cap D^{add}_\mathtt{s}\right|$, (3) $B_\mathtt{t} \setminus C_\mathtt{t} = B_\mathtt{s} \setminus C_\mathtt{s}$, (4) $|C_\mathtt{t}| = |C_\mathtt{s}|$. In particular, it follows that $D^{remove}_\mathtt{t} \setminus D^{add}_\mathtt{t} = D^{remove}_\mathtt{s} \setminus D^{add}_\mathtt{s}$.

If $S^u \in \mathtt{S}_\mathtt{t}$, then $u \in R_1 \cap R_2$ and, by conditions (1) and (2), u does not occur anywhere else in $\mathtt{t}$. The argument so far entails that $u \notin U_\mathtt{s} \cup A_\mathtt{s} \cup B_\mathtt{s}$, so that $S^u \in \mathtt{S}_\mathtt{s}$. Thus $\mathtt{S}_\mathtt{t} = \mathtt{S}_\mathtt{s}$.

Now the only difference left between $\mathtt{t}$ and $\mathtt{s}$ is in which order the vertex names are introduced and removed. This is taken care of precisely by syntactic manipulations 1 and 3.

Theorem 2 (Universality). *Given a reaction $r : A \to C$, there is a term $\mathtt{t} : A \to B$ and an isomorphism $\iota : B \xrightarrow{\sim} C$ such that $R(\mathtt{t}); \iota = r$.*

Proof. Observe that every reaction $r : A \to C$ factorises as

$$(U_A, U_B, \mathrm{id}, \mathrm{id}); (\varnothing, \varnothing, !, \iota),$$

where $(U_A, U_B) : A \to B$ is some reaction and $\iota : B \to C$ is an isomorphism of labelled graphs. The sought-after term $\mathtt{t} : A \to B$ is given by

$$\prod_{\substack{u \in \mathtt{Crg}^+(U_A) \\ v \in \mathtt{Crg}^-(U_A)}} (I^{uv})^{\mathtt{ion}(m_A(u,v))}; \prod_{u,v \in \mathtt{Chem}(U_A)} \prod_{i=1}^{\mathtt{cov}(m_A(u,v))} C^{uv}_{a_i b_i};$$

$$\prod_{u \in \mathtt{Crg}^-(U_A)} \prod_{i=1}^{\tau_A^{\mathtt{Crg}}(u)} E^u_{a_i b_i}; \prod_{u \in \mathtt{Chem}(U_A)} \prod_{i=1}^{\mathbf{v}\tau_A^{\mathrm{At}}(u) - \max\left(\tau_A^{\mathtt{Crg}}(u), 0\right)} E^{u a_i};$$

$$\prod_{u \in \mathtt{Chem}(U_B)} \prod_{i=1}^{\mathbf{v}\tau_B^{\mathrm{At}}(u) - \max\left(\tau_B^{\mathtt{Crg}}(u), 0\right)} \bar{E}^{u a_i}; \prod_{u \in \mathtt{Crg}^-(U_B)} \prod_{i=1}^{\tau_B^{\mathtt{Crg}}(u)} \bar{E}^u_{a_i b_i};$$

$$\prod_{u,v \in \mathtt{Chem}(U_B)} \prod_{i=1}^{\mathtt{cov}(m_B(u,v))} \bar{C}^{uv}_{a_i b_i}; \prod_{\substack{u \in \mathtt{Crg}^+(U_B) \\ v \in \mathtt{Crg}^-(U_B)}} \left(\bar{I}^{uv}\right)^{\mathtt{ion}(m_B(u,v))};$$

$$\prod_{a \in \alpha(U_B)} R^{b_a \mapsto a} \prod_{u \in U_B} S^u.$$

Note that while the term we obtain is in an ICE-form, it will not, in general, be in normal form.