Disconnection Rules are Complete for Chemical Reactions

Ella Gale
University of Bristol, UK
ella.gale@bristol.ac.uk

Leo Lobski
University College London, UK
leo.lobski.21@ucl.ac.uk

Fabio Zanasi
University College London, UK
University of Bologna, Italy
f.zanasi@ucl.ac.uk

Graph-based models of chemical processes typically come at two different levels of abstraction: formal *reactions* and *disconnection rules*. The former are combinatorial rearrangements of atoms and charge, and are used for reaction prediction and storage of reaction data. The latter 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, 9, 5, 8, 10].
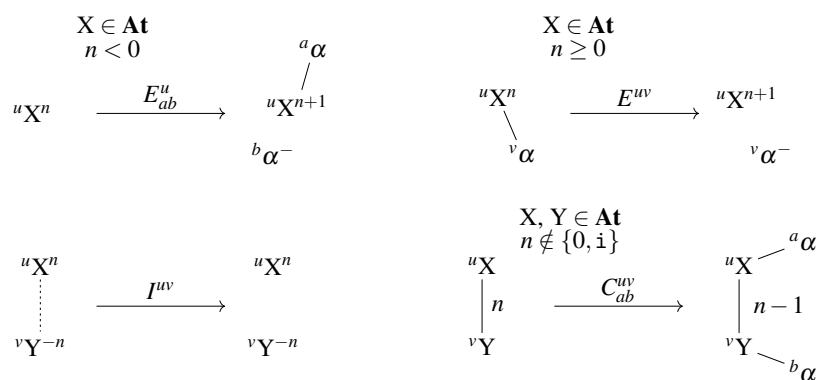
Whereas reactions have been studied formally before, a mathematical description of disconnection rules has received far less attention (see, however [4, 7]). In this work, we provide a categorical formalisation of the two levels, and show that they are tightly linked: the disconnection rules are complete and universal with respect to the reactions (Corollary 2). This is a consequence of the functor from the category of disconnection rules to the category of reactions being a faithful opfibration (Theorem 1). 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.

Molecules are represented as certain labelled graphs (called *chemical graphs*), whose vertices are labelled with atoms and charge, and edges with ionic or covalent bonds. Reactions are partial bijections between chemical graphs which preserve the atoms and the net charge, with the additional requirement that the complement of the domain is isomorphic to the complement of the codomain as a labelled graph. The idea is that nothing outside of the domain of the bijection is changed during the reaction. We denote by **React** the category whose objects are chemical graphs and whose morphisms are reactions.

This way of representing reactions is strongly connected to double pushout graph rewriting [1, 2, 3]: in fact, every morphism in **React** can be represented as a double pushout diagram in the category of chemical graphs. The composition, however, is not that of composing rewrite rules, which is non-deterministic in the sense that it results in a set of all possible combinations of the rules being composed [3]. The composition in **React** can be though of as composition of two fixed instances of rewrite rules. We give an example of a reaction (glucose phosphorylation) below, where the numbered vertices are in the (co)domain of the bijection, with vertices with the same superscript being mapped to each other:
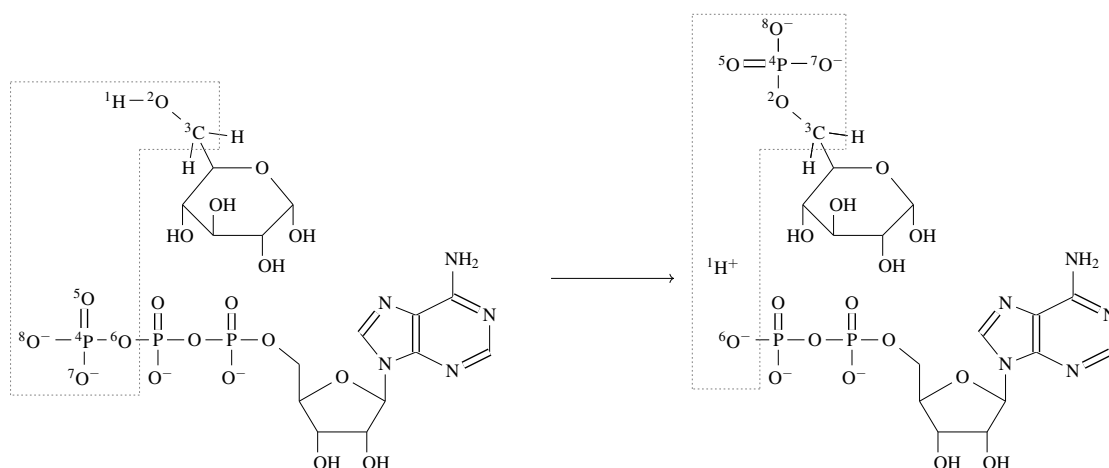
We define four classes of disconnection rules, all of which have a clear chemical significance: two versions of *electron detachment*, *ionic bond breaking* and *covalent bond breaking*. The converse of a disconnection rule is a *connection rule*. We represent the rules as the following graph transformations (here α denotes an unpaired electron and α^- a free electron):

$$X \in \mathbf{At} \quad n < 0$$
$$^u X^n \xrightarrow{\quad E^u_{ab} \quad} \begin{array}{c} {}^a\alpha \\ {}^u X^{n+1} \\ {}^b\alpha^- \end{array}$$

$$X \in \mathbf{At} \quad n \geq 0$$
$$\begin{array}{c} {}^u X^n \\ {}^v\alpha \end{array} \xrightarrow{\quad E^{uv} \quad} \begin{array}{c} {}^u X^{n+1} \\ {}^v\alpha^- \end{array}$$

$$\begin{array}{c} {}^u X^n \\ \vdots \\ {}^v Y^{-n} \end{array} \xrightarrow{\quad I^{uv} \quad} \begin{array}{c} {}^u X^n \\ {}^v Y^{-n} \end{array}$$

$$X, Y \in \mathbf{At} \quad n \notin \{0, \mathtt{i}\}$$
$$\begin{array}{c} {}^u X \\ \Big| n \\ {}^v Y \end{array} \xrightarrow{\quad C^{uv}_{ab} \quad} \begin{array}{c} {}^u X \diagup {}^a\alpha \\ \Big| n-1 \\ {}^v Y \diagdown {}^b\alpha \end{array}$$

Our first observation (universality) is highly intuitive: the (dis)connection rules generate all the morphisms in **React** (up to an isomorphism). This can be thought of as a consistency result for reactions: the definition captures exactly those rearrangements of chemical graphs which result from local, chemically motivated rewrite rules. On the other hand, this establishes the basic units of retrosynthetic analysis as first-class citizens of reaction representation. The second result is perhaps slightly more surprising: treating the (dis)connection rules as terms with types (the source and the target graphs), the terms can be endowed with equations such that the terms describing the same reaction are identified. This axiomatisation is essentially a set of commutativity conditions: the identities say when two rules can be applied in either order. We denote the resulting category by **Disc**, and note that there is a functor to **React** that simply keeps track of the vertices on which the (dis)connections act on.

Theorem 1. *The functor* **Disc** $\to$ **React** *is a faithful opfibration.*

Corollary 2 (Universality and completeness). *For every reaction $r : A \to B$, there is a unique (up to equations in* **Disc***) sequence of disconnection rules $d : A \to B'$ and an isomorphism $f : B' \xrightarrow{\sim} B$ such that $f \circ R(d) = r$.*

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*, doing retrosynthesis or storing reaction data in a systematic way.

References

[1] Jakob L Andersen, Christoph Flamm, Daniel Merkle & Peter F Stadler (2013): *Inferring chemical reaction patterns using rule composition in graph grammars*. Journal of systems chemistry 4(1), pp. 1–4.

[2] Jakob L. Andersen, Christoph Flamm, Daniel Merkle & Peter F. Stadler (2017): *An intermediate level of abstraction for computational systems chemistry*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 375(2109), p. 20160354, doi:10.1098/rsta.2016.0354.

[3] Nicolas Behr, Jean Krivine, Jakob L. Andersen & Daniel Merkle (2021): *Rewriting Theory for the Life Sciences: A Unifying Theory of CTMC Semantics*. Theor. Comput. Sci. 884(C), p. 68–115, doi:10.1016/j.tcs.2021.07.026.

[4] Olivier Bournez, Liliana Ibănescu & Hélène Kirchner (2006): *From Chemical Rules to Term Rewriting*. Electronic Notes in Theoretical Computer Science 147(1), pp. 113–134, doi:https://doi.org/10.1016/j.entcs.2005.06.040. Available at https://www.sciencedirect.com/science/article/pii/S1571066106000375. Proceedings of the 6th International Workshop on Rule-Based Programming (RULE 2005).

[5] Jonathan Clayden, Nick Greeves & Stuart Warren (2012): *Organic chemistry*, second edition. edition. Oxford University Press, Oxford.

[6] E. J. Corey & Xue min Cheng (1989): *The logic of chemical synthesis*. John Wiley, New York.

[7] Ella Gale, Leo Lobski & Fabio Zanasi (2023): *A Categorical Approach to Synthetic Chemistry*. In Erika Ábrahám, Clemens Dubslaff & Silvia Lizeth Tapia Tarifa, editors: *Theoretical Aspects of Computing – ICTAC 2023*, Springer Nature Switzerland, Cham, pp. 276–294.

[8] Yijia Sun & Nikolaos V. Sahinidis (2022): *Computer-aided retrosynthetic design: fundamentals, tools, and outlook*. Current Opinion in Chemical Engineering 35, p. 100721.

[9] Stuart Warren & Paul Wyatt (2008): *Organic synthesis : the disconnection approach*, 2nd ed. edition. Wiley, Hoboken, N.J.

[10] Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Tingjun Hou & Mingli Song (2024): *Recent advances in deep learning for retrosynthesis*. Wiley Interdisciplinary Reviews: Computational Molecular Science 14(1), p. e1694.