

# Exploration of Chemical Space

*Formal, chemical and historical aspects*

Der Fakultät für Mathematik und Informatik  
der Universität Leipzig  
angenommene

## DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium  
(Dr. rer. nat.)

im Fachgebiet

Informatik

vorgelegt von

Bachelor of Science in Chemistry and Bachelor of Science in Mathematics Wilmer  
Leal  
geboren am 10.05.1989 in Pamplona, Kolumbien

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler (Universität Leipzig) & Prof. Dr. Jürgen Jost  
(Max-Planck-Institut für Mathematik in den Naturwissenschaften)
2. Prof. Dr. Rolf Fagerberg (University of Southern Denmark)

Die Verleihung des akademischen Grades erfolgt mit Bestehen  
der Verteidigung am 5. Dezember 2022 mit dem Gesamtprädikat *Summa cum laude*



To my wife Angie



# Bibliographic Description

Title:	Exploration of Chemical Space
Subtitle:	Formal, chemical and historical aspects
Type:	Dissertation
Author:	Wilmer Leal
Year:	2022
Professional discipline:	Computer Science
Language:	English
Pages in the main part:	91
Chapter in the main part:	6
Number of Figures:	44
Number of Tables:	19
Number of Appendices:	4
Number of Citations:	159
Key Words:	Chemical space, categorical chemistry, hypernetworks, complex systems, directed hypergraphs, discrete curvature, Forman-Ricci curvature, chemical similarity, system of chemical elements, history of chemistry, computational history of chemistry, classification in chemistry, Reaxys, Petri nets, Dialectica categories, Dialectica Petri nets, symmetric monoidal closed categories, categorical models of chemical reaction networks, categorical chemistry.

## This thesis is based on the following publications.

- [1] Eidi, M., Farzam, A., Leal, W., Samal, A., and Jost, J. "Edge-based analysis of networks : curvatures of graphs and hypergraphs". In: *Theory in biosciences* 139.4 (2020), pp. 337–348. *ISSN*: 1431-7613. *DOI*: 10.1007/s12064-020-00328-0.
- [2] Lavore, E. D., Leal, W., and Paiva, V. de. "Dialectica Petri nets". ArXiv. (Submitted). 2021.
- [3] Leal, W., Eidi, M., and Jost, J. "Curvature-based analysis of directed hypernetworks". In: *Complex networks 2019: the 8th international conference on complex networks and their applications; December 10 - 12, 2019 Lisbon, Portugal; book of abstracts*. Ed. by H. Cherifi. [s.l.]: International conference on complex networks and their applications, 2019, pp. 32–34. *ISBN*: 978-2-9557050-3-2.

- [4] Leal, W., Eidi, M., and Jost, J. "Ricci curvature of random and empirical directed hypernetworks". In: *Applied network science* 5.1 (2020), p. 65. *ISSN*: 2364-8228. *DOI*: 10.1007/s41109-020-00309-8.
- [5] Leal, W., Llanos, E. J., Bernal, A., Stadler, P. F., Jost, J., and Restrepo, G. "The expansion of chemical space in 1826 and in the 1840s prompted the convergence to the periodic system". (Accepted in *Proceedings of the National Academy of Sciences of the United States of America*). 2022.
- [6] Leal, W. and Restrepo, G. "Formal structure of periodic system of elements". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2224 (2019), p. 20180581. *ISSN*: 1364-5021. *DOI*: 10.1098/rspa.2018.0581.
- [7] Leal, W., Restrepo, G., Stadler, P. F., and Jost, J. "Forman-Ricci curvature for hypergraphs". In: *Advances in Complex Systems* 24.01 (2021), p. 2150003. *ISSN*: 0219-5259. *DOI*: 10.1142/S021952592150003X.
- [8] Llanos, E. J., Leal, W., Bernal, A., Restrepo, G., Jost, J., and Stadler, P. F. "A Network model of the Chemical Space provides similarity structure to the system of chemical elements". In: *Complex networks 2019: the 8th international conference on complex networks and their applications; December 10 - 12, 2019 Lisbon, Portugal; book of abstracts*. Ed. by H. Cherifi. [s.l.]: International conference on complex networks and their applications, 2019, pp. 308–310. *ISBN*: 978-2-9557050-3-2.
- [9] Llanos, E. J.\*, Leal, W.\*, Luu, D. H., Jost, J., Stadler, P. F., and Restrepo, G. "Exploration of the chemical space and its three historical regimes". In: *Proc. Natl. Acad. Sci. U.S.A.* 116.26 (2019), pp. 12660–12665. *ISSN*: 0027-8424. *DOI*: 10.1073/pnas.1816039116.
- [10] Llanos, E. J., Leal, W., Restrepo, G., and Stadler, P. F. "Computational approach to the history of chemical reactivity: Exploring Reaxys database". In: *Abstracts of papers of the American Chemical Society* 254 (2017).

---

\*The authors share first authorship.

# Abstract

Starting from the observation that substances and reactions are the central entities of chemistry, I have structured chemical knowledge into a formal space called a directed hypergraph, which arises when substances are connected by their reactions. I call this hypernet chemical space. In this thesis, I explore different levels of description of this space: its evolution over time, its curvature, and categorical models of its compositionality.

The vast majority of the chemical literature focuses on investigations of particular aspects of some substances or reactions, which have been systematically recorded in comprehensive databases such as Reaxys for the last 200 years. While complexity science has made important advances in physics, biology, economics, and many other fields, it has somewhat neglected chemistry. In this work, I propose to take a global view of chemistry and to combine complexity science tools, modern data analysis techniques, and geometric and compositional theories to explore chemical space. This provides a novel view of chemistry, its history, and its current status.

We argue that a large directed hypergraph, that is, a model of directed relations between sets, underlies chemical space and that a systematic study of this structure is a major challenge for chemistry. Using the Reaxys database as a proxy for chemical space, we search for large-scale changes in a directed hypergraph model of chemical knowledge and present a data-driven approach to navigate through its history and evolution. These investigations focus on the mechanistic features by which this space has been expanding: the role of synthesis and extraction in the production of new substances, patterns in the selection of starting materials, and the frequency with which reactions reach new regions of chemical space. Large-scale patterns that emerged in the last two centuries of chemical history are detected, in particular, in the growth of chemical knowledge, the use of reagents, and the synthesis of products, which reveal both conservatism and sharp transitions in the exploration of the space. Furthermore, since chemical similarity of substances arises from affinity patterns in chemical reactions, we quantify the impact of changes in the diversity of the space on the formulation of the system of chemical elements.

In addition, we develop formal tools to probe the local geometry of the resulting directed hypergraph and introduce the Forman-Ricci curvature for directed and undirected hypergraphs. This notion of curvature is characterized by applying it to social and chemical networks with higher order interactions, and then used for the investigation of the structure and dynamics of chemical space.

The network model of chemistry is strongly motivated by the observation that the compositional nature of chemical reactions must be captured in order to build a model of chemical reasoning. A step forward towards categorical chemistry, that is, a formalization of all the flavors of compositionality in chemistry, is taken by the construction of a categorical model of directed hypergraphs. We lifted the structure from a lineale (a poset version of a symmetric monoidal closed category) to a

category of Petri nets, whose wiring is a bipartite directed graph equivalent to a directed hypergraph. The resulting construction, based on the Dialectica categories introduced by Valeria De Paiva, is a symmetric monoidal closed category with finite products and coproducts, which provides a formal way of composing smaller networks into larger in such a way that the algebraic properties of the components are preserved in the resulting network. Several sets of labels, often used in empirical data modeling, can be given the structure of a lineale, including: stoichiometric coefficients in chemical reaction networks, reaction rates, inhibitor arcs, Boolean interactions, unknown or incomplete data, and probabilities. Therefore, a wide range of empirical data types for chemical substances and reactions can be included in our model.

# Zusammenfassung

Ausgehend von der Beobachtung, dass Substanzen und Reaktionen die zentralen Einheiten der Chemie sind, habe ich chemisches Wissen in einen formalen Raum namens gerichteter Hypergraph strukturiert, der entsteht, wenn Substanzen durch ihre Reaktionen verbunden werden. Ich nenne dieses Hypernetz chemischen Raum. In dieser Arbeit untersuche ich verschiedene Ebenen der Beschreibung dieses Raums: seine zeitliche Entwicklung, seine Krümmung und kategoriale Modelle seiner Kompositionalität.

Der überwiegende Teil der chemischen Literatur konzentriert sich auf Untersuchungen zu bestimmten Aspekten einiger Substanzen oder Reaktionen, die in umfassenden Datenbanken wie Reaxys seit 200 Jahren systematisch erfasst werden. Während die Komplexitätswissenschaft wichtige Fortschritte in Physik, Biologie, Wirtschaftswissenschaften und vielen anderen Bereichen gemacht hat, hat sie die Chemie etwas vernachlässigt. In dieser Arbeit schlage ich vor, einen globalen Blick auf die Chemie zu werfen und komplexitätswissenschaftliche Werkzeuge, moderne Datenanalysetechniken sowie geometrische und Zusammensetzungstheorien zu kombinieren, um den chemischen Raum zu erforschen. Dies bietet einen neuartigen Blick auf die Chemie, ihre Geschichte und ihren aktuellen Status.

Wir argumentieren, dass ein großer gerichteter Hypergraph, d. h. ein Modell gerichteter Beziehungen zwischen Mengen, dem chemischen Raum zugrunde liegt und dass eine systematische Untersuchung dieser Struktur eine große Herausforderung für die Chemie darstellt. Unter Verwendung der Reaxys-Datenbank als Proxy für den chemischen Raum suchen wir nach großräumigen Änderungen in einem gerichteten Hypergraphenmodell des chemischen Wissens und präsentieren einen datengesetzten Ansatz, um durch seine Geschichte und Entwicklung zu navigieren. Diese Untersuchungen konzentrieren sich auf die mechanistischen Merkmale, durch die sich dieser Raum erweitert hat: die Rolle der Synthese und Extraktion bei der Herstellung neuer Substanzen, Muster bei der Auswahl von Ausgangsmaterialien und die Häufigkeit, mit der Reaktionen neue Bereiche des chemischen Raums erreichen. Großräumige Muster, die in den letzten zwei Jahrhunderten der Chemiegeschichte entstanden sind, lassen sich insbesondere beim Wachstum des chemischen Wissens, der Verwendung von Reagenzien und der Synthese von Produkten nachweisen, die sowohl Konservatismus als auch scharfe Übergänge in der Erforschung der Chemie erkennen lassen Platz. Da die chemische Ähnlichkeit von Substanzen aus Affinitätsmustern in chemischen Reaktionen entsteht, quantifizieren wir außerdem den Einfluss von Änderungen in der Vielfalt des Raums auf die Formulierung des Systems chemischer Elemente.

Darüber hinaus entwickeln wir formale Werkzeuge, um die lokale Geometrie des resultierenden gerichteten Hypergraphen zu untersuchen, und führen die Forman-Ricci-Krümmung für gerichtete und ungerichtete Hypergraphen ein. Dieser Begriff der Krümmung wird charakterisiert, indem er auf

soziale und chemische Netzwerke mit Wechselwirkungen höherer Ordnung angewendet und dann für die Untersuchung der Struktur und Dynamik des chemischen Raums verwendet wird.

Das Netzwerkmodell der Chemie ist stark motiviert durch die Beobachtung, dass die Zusammensetzung chemischer Reaktionen erfasst werden muss, um ein Modell des chemischen Denkens zu erstellen. Ein Schritt vorwärts in Richtung kategorialer Chemie, d. h. einer Formalisierung aller Geschmacksrichtungen der Kompositionalität in der Chemie, wird durch die Konstruktion eines kategorialen Modells gerichteter Hypergraphen unternommen. Wir haben die Struktur von einem Lineale (einer Poset-Version einer symmetrischen monoidalen geschlossenen Kategorie) zu einer Kategorie von Petri-Netzen angehoben, deren Verdrahtung ein zweigeteilter gerichteter Graph ist, der einem gerichteten Hypergraphen entspricht. Die resultierende Konstruktion, basierend auf den von Valeria De Paiva eingeführten Dialectica-Kategorien, ist eine symmetrische monoidale geschlossene Kategorie mit endlichen Produkten und Nebenprodukten, die eine formale Möglichkeit bietet, kleinere Netzwerke so zu größeren zusammenzusetzen, dass die algebraischen Eigenschaften der Komponenten im resultierenden Netzwerk erhalten. Mehrere Sätze von Labels, die häufig bei der empirischen Datenmodellierung verwendet werden, können mit der Struktur eines Lineale versehen werden, darunter: stöchiometrische Koeffizienten in chemischen Reaktionsnetzwerken, Reaktionsraten, Inhibitorbögen, Boolesche Wechselwirkungen, unbekannte oder unvollständige Daten und Wahrscheinlichkeiten. Daher kann ein breites Spektrum an empirischen Datentypen für chemische Substanzen und Reaktionen in unser Modell aufgenommen werden.

# Acknowledgment

I would like to thank Prof. Peter F. Stadler for being a kind and generous Doktorvater! I thank him for having me as a guest in his group before starting my PhD; for his total support which translated into a successful DAAD scholarship; and for his advice, training, and *immense* patience during my Ph.D. I look forward to continuing working with and learning from him.

Thanks to Prof. Jost for having received me so generously into his group, for all the time he devoted to my training, for his support, and for the wonderful ways in which he promoted my career.

Thanks to my three masters for the training, advice, and unwavering support. They played key roles in my preparation for this PhD. Alphabetically: Eugenio Llanos, Juan Carlos López Carreño, and Guillermo Restrepo.

I am indebted to Joachim Schummer, whose philosophical reflections on the nature of chemistry have inspired many of my ideas and motivated my search for mathematical tools to formalize them.

Thanks to Paolo Perrone for my first course in category theory and for the valuable discussions. Thanks to Prof. Valeria De Paiva for the training, for being an inspiring mentor, and for keeping an eye on us in trying times. Thanks to Angie González, Alejandro Moncada, and Esteban Rivas for their enthusiasm and great work in our category theory seminar.

For their brilliant work in making easy any technical, administrative, or bureaucratic task, I thank Petra Pregel, Jens Steuck (Bioinformatics, University of Leipzig), Antje Vandenberg, Kenny Puder, Heike Rackwitz, Britta Schneemann, Rainer Kleinrensing, Lisa Braun and Oliver Heller (MPI MiS). Also, thanks to Fabian Gärtner for the latex class for this thesis.

Thanks to the German Academic Exchange Service (DAAD) for the financial support that made possible this thesis.

Thanks to my dear friends Rituparno Sen and Deisy Gysi for their warm welcome to Leipzig and for our adventures together.

Thanks to Asad, Ayu, Ali, Mesa, and Sola, for the family that we started in Berlin.

I am deeply grateful to my friend Antje Vandenberg for her infinite support.

Thanks to my dear friend Adrián for his incredible energy! He has meant the world to us.

Thanks to my friends Eugenio, Esteban, Tjorben, Daniel, Santiago, Andrés Bernal, Andrés Marulanda, Edith, Lukas, Mascha, Shuhan, Nadia, Theresa, Haya, Hannaneh, Leonardo, Nathaly, Oscar, Mónica, and Simón for the wonderful time we had together.

Thanks to Marzieh and Zach for their advice and friendship, which made this experience so much better.

Thanks to my friends Margot, Barbara, Enrico, Doris, Annet, Ibeth, and Corina, for our wonderful Tuesday nights.

I am grateful to have met Jochen Guck and to have him as an advisor and source of wisdom.

Thanks to my family Mary, Marco, Fabián, Leidy, Jhon, Lucha, Ene, Jenny, Álvaro, Jonathan, Pacho, Andrés, Rosalba, Juan Carlos, and Roberto for being a constant source of motivation and strength.

I found no phrase that begins to describe the role that Angie has played in my academic and personal life. I have wondered for years in the vast chemical space in the hope of finding appropriate words hidden somewhere. My exploration, whose results are reported in this thesis, is dedicated to her.

# Contents

<b>I</b>	<b>Introduction</b>	<b>1</b>
<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Motivation . . . . .	4
1.2	Outline . . . . .	6
<b>II</b>	<b>Computational History of Chemistry</b>	<b>9</b>
<b>2</b>	<b>Looking for large scale patterns</b>	<b>12</b>
2.1	The role of synthesis . . . . .	12
2.2	On the selection of substrates . . . . .	12
2.3	Reaching regions of the chemical space . . . . .	16
2.4	Historical changes in the structure of the chemical space: growth of substances and reactions . . . . .	18
2.5	Chemical space density changes . . . . .	22
<b>3</b>	<b>Chemical space and similarity in the system of chemical elements</b>	<b>26</b>
3.1	Similarity, atomic weights, and the expanding chemical space . . . . .	26
3.2	Evolution of the chemical space (1800-1868) . . . . .	26
3.3	Evolution of the system of chemical elements (1800-1868) . . . . .	30
<b>III</b>	<b>Geometry and Compositionality in Chemical Space</b>	<b>41</b>
<b>4</b>	<b>Curvature of hypergraphs and chemical space</b>	<b>44</b>
4.1	Introduction and outline . . . . .	44
4.2	A very brief story of Forman-Ricci curvature . . . . .	45
4.3	Forman-Ricci curvature of graphs . . . . .	46
4.4	Forman-Ricci curvature of undirected hypergraphs . . . . .	48
4.5	Curvature of directed hypergraphs . . . . .	51
4.6	Curvature of chemical space . . . . .	63
<b>5</b>	<b>A categorical model of chemical reaction networks</b>	<b>72</b>
5.1	Dialectica Petri nets . . . . .	72
5.2	Petri nets via Dialectica Categories . . . . .	73

5.3	The category $M_L\text{Set}$ and its structure . . . . .	74
5.4	A category of Petri nets . . . . .	77
5.5	Different lineales . . . . .	79
<b>IV</b>	<b>Final remarks</b>	<b>85</b>
<b>6</b>	<b>Conclusions and further work</b>	<b>88</b>
6.1	Discussion and conclusions . . . . .	88
6.2	Further work . . . . .	91
	<b>Appendices</b>	<b>92</b>
<b>A</b>	<b>Supplementary data of Chapter 2</b>	<b>95</b>
<b>B</b>	<b>Supplementary data of Chapter 3</b>	<b>99</b>
<b>C</b>	<b>Supplementary data of Chapter 4</b>	<b>117</b>
<b>D</b>	<b>Proofs for Chapter 5</b>	<b>123</b>
	<b>Bibliography</b>	<b>131</b>
	<b>Curriculum Scientiae</b>	<b>143</b>

## Part I

# Introduction



CHAPTER 1 

Introduction

Contents

1.1	Motivation . . . . .	4
1.2	Outline . . . . .	6

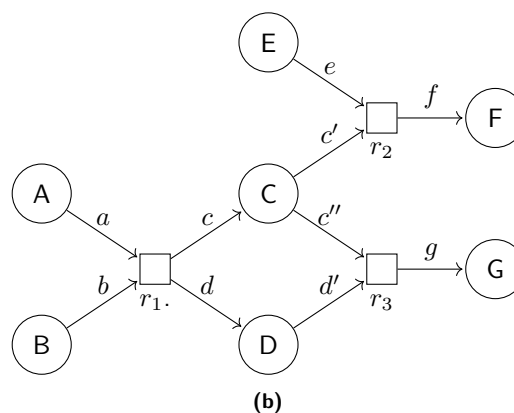
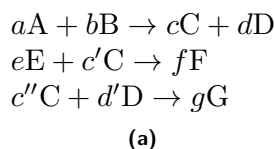
## 1.1 Motivation

In 1905, Alfred Werner, the Swiss 1913 Nobel Laureate in Chemistry, said "Die Chemie muss zur Astronomie der Molekularwelt werden". A century later, this metaphor appeared on the cover of *Nature* under the fruitful name of *Chemical Space*. Although there is no unique definition of Chemical Space, the different notions seem to agree that it is a conceptual space where substances lie described by their properties (see for instance Kirkpatrick and Ellis [63], Lilienfeld, Müller, and Tkatchenko [82], and Llanos et al. [84]). In practice, realizations of this space depend on the aspects that are relevant for its explorers. In drug design, for instance, emphasis is placed on (small) molecules and on molecular descriptors that may be related to their potential biological activity [28]. The question then is whether we can find a fundamental structure underlying the Chemical Space, one whose investigation may bring insights to explorers of every region of this space. Chemical substances are undoubtedly part of the structure we are looking for; it remains to make a decision about which properties we shall use to structure such a space.

Philosopher of chemistry Joachim Schummer investigated this question and argues that among all material properties of substances, reactivity (the affinity of a set of substances to combine, change, and produce new ones), stands out [137]. The arguments are manifold, but two rise to prominence: reactivity information provides a classification scheme (the chemical identity of a substance is given by both the reactions it undergoes and by the reactions that produce it); also, chemists reason on the network structure that emerges when reactions are connected to make their concurrency explicit, i.e., since chemical reactions are compositional in nature, one may for instance investigate what substances can be reached from a given set of substrates (e.g., from the network in Figure 1b, wired using reactions from 1a, we easily see that A, B, and E is the smallest set needed to reach every other substance in this tiny corner of the space), or plan an appropriate synthesis route to get to a target (the network in Figure 1b offers an alternative path to synthesise F in case C is absent but A, B, and E are in our lab's shelf: composing reactions  $r_1$  and  $r_2$ ) [69].

The knowledge above cannot be faithfully derived from isolated substances and molecular descriptors, but instead, it is drawn from the structure of the resulting network, which, according to Schummer, "forms the chemical core of experimental chemistry" [137]. Far from dismissing other properties, we shall imagine the current Chemical Space as Schummer's network (see Figure 2), with known substances and reactions as nodes and arcs, respectively, along with all their properties recorded throughout history, e.g., chemical names, molecular formulas, and molecular structure for substances; and names, thermodynamic conditions, and mechanisms for reactions. We have the structure we want to investigate, and now we have to decide what formal space to use for modeling the Chemical Space.

Systems made of objects and relations are represented mathematically by graphs or hypergraphs. A model is chosen for a given system based on both the symmetry and arity of the relation. Symmetric binary relations are modeled as undirected graphs, which are collections of two-element sets  $\{i, j\}$  called edges, one for each pair of objects  $i$  and  $j$  standing in relation (see definition 1), like two locations connected by a road, or Facebook friendship [25]. Sometimes the relation is not symmetric since the roles of  $i$  and  $j$  are not interchangeable, and the system is modeled as a directed graph instead, i.e., a collection of pairs  $(i, j)$  of elements called directed edges, whose order determines two roles generically called *tail* and *head*, respectively; examples include links from one web page to another one in the the World Wide Web [1], or a paper citing another one [139]. Relations that may involve more than two elements, like co-authorship in a paper, or having voted in the same Wikipedia election [78], have traditionally been approximated by binary relations (see for instance



**Figure 1:** Representation of chemical reaction data: as a list (a) and as a network (b). A, B, ..., H are substances and  $a, b, \dots, h$  are stoichiometric coefficients that indicate the proportion in which they combine.

M. E. J. Newman and Girvan [104]), perhaps to simplify computations and to take advantage of the large number of tools available for graphs [36]. But formally, moving from binary relations to higher order ones simply translates into generalizing graphs to a structure capable of faithfully encoding interactions among any number of elements. Hypergraphs generalize graphs precisely in that direction, as they consist of a collection of sets called hyperedges. In a coauthorship hypernetwork, for instance, there is a hyperedge of coauthors for each paper, as in Savić, Ivanović, and Jain [133]. Just as well, we may want to represent the following relation "the authors of paper *A* cite the authors of paper *B*". As in directed graphs, we have two roles, citing and cited authors, but this time around tail and head are not elements but sets. Directed hypergraphs emerge to deal with these kind of systems by allowing the tail and head to be arbitrary sets of vertices. Other examples include antecedents and consequents of a Horn formula in propositional logic [138] and reactants and products of a chemical reaction [40].

Directed hypergraphs quite naturally model chemical reaction networks for they are models of concurrency of directed relations. These models provide a rich semantic basis on which to interpret questions that arise in chemistry [69], such as: what substances can be synthesized from a given set of starting materials? [144]. Given a target substance, which synthetic routes are known, and which starting materials are needed to reach the target? [56]. Do chemical reactions turn targets into key precursors? [31]. How many synthetic routes pass through a given reaction? These questions can be answered by investigating the topology and geometry of the wiring of the network. The first two questions are answered by defining suitable closure operators [144, 145]. The last two questions can be addressed by computing the curvature of the edges of the network [78]. More abstractly, the network model of chemistry, and thus the questions above, are inspired and motivated by the compositional nature of chemical reactions. The question then arises: what categorical constructions underlie the structure of chemical reasoning and what are their formal connections? In this thesis, we investigate the formal structures of Chemical Space, geometric features of its underlying directed hypergraph, how they have changed throughout the history of chemistry, and their impact on similarity among elements. Also, as the first step towards my quest for flavors of compositionality in chemistry and the connections of their models, we investigate a categorical (compositional) network model of chemical reaction networks and its structure.

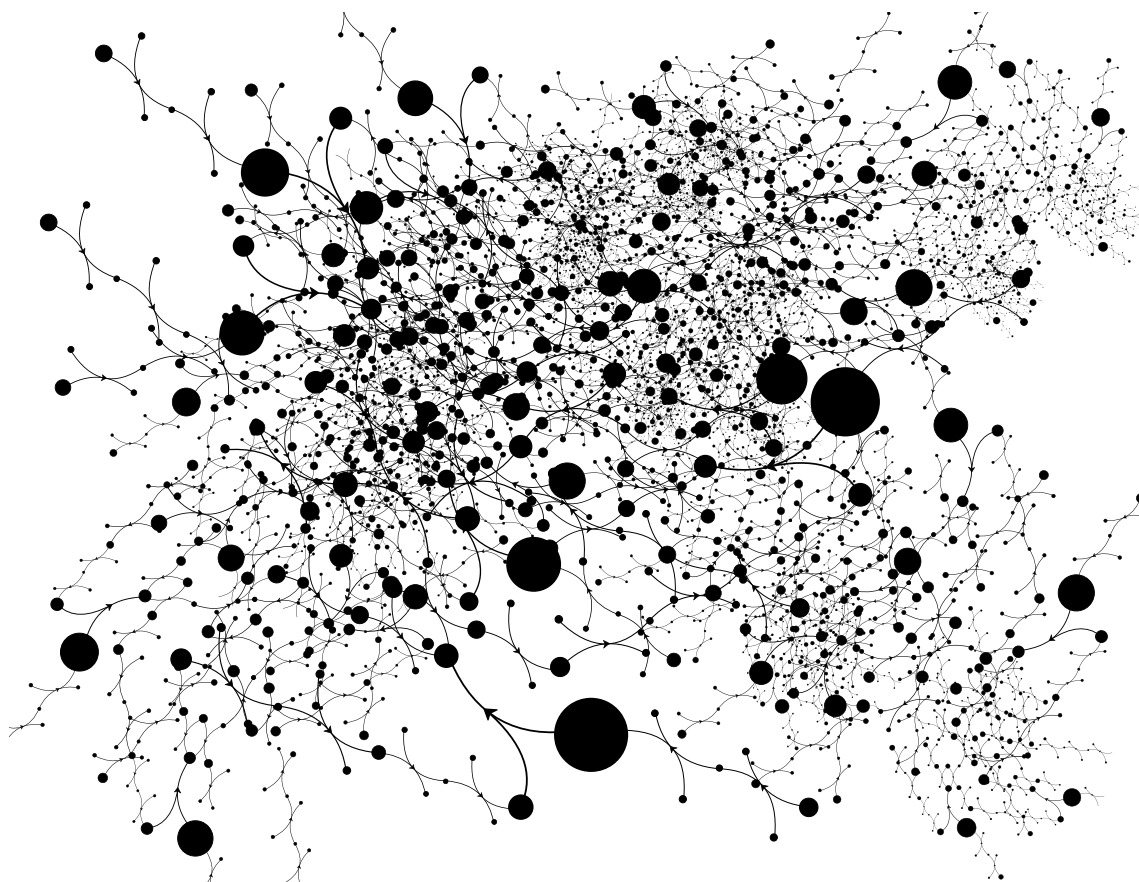


Figure 2: Chemical Space as a directed hypergraph

## 1.2 Outline

The remaining of this thesis is divided into three parts:

- In the first part we present a data-driven approach to the history of chemistry or *computational history of chemistry*: we use Reaxys database (a large corpus of chemical data on substances and reactions), to search for large scale changes in our directed hypergraph model of chemical space. This part is divided into two chapters:
  - *Chapter 2*: We present the results of our exploration of the entire chemical space for the period 1800-2015. The first three sections focus on three mechanistic features by which the space has been expanding: the role of synthesis and extraction in the production of new substances, patterns in the selection of starting materials, and frequency with which reactions reach new regions of the space. In the fourth section, we quantify the growth of chemical space (1800-2015) and relate growth changes with scientific and social historical

events. In the last section, we show how the density of chemical space has changed over the history of chemistry. The first four sections of this chapter are based on [84, 85] and the last section is ongoing research.

- *Chapter 3:* we quantify the impact of changes in chemical space on the formulation of the system of chemical elements by 1869. For this, we report the diversity changes of the space from 1800 to 1869 (first section). Then, we compute how the system evolved from 1800 until its formulation (second section). Finally, we reconstruct systems of elements for several nineteenth-century chemists (using two approaches, a contemporary view and a historical reconstruction), and quantify their differences from that of 1869. This chapter is based on [74, 76, 83].
- The second part is dedicated to curvature and compositionality in chemical space. The content of the second part is divided into two chapters:
  - *Chapter 4:* First, we give a brief account of Forman's definition of curvature for simplicial complexes. Then, we introduce the Forman-Ricci curvature for directed and undirected hypergraphs and characterize this notion by applying it to two empirical networks, one directed and the other undirected (both with higher order interactions). In the last section, we use our definition of Forman-Ricci curvature to investigate the structure of chemical space. This chapter is based on [31, 72, 73, 78].
  - *Chapter 5:* We produce a categorical model of bipartite directed graphs, which are equivalent to directed hypergraphs, called Petri nets. First, we lift the structure of a lineale (a poset version of a symmetric monoidal closed category) to an intermediate category from which the category of Petri nets is built. The result is a symmetric monoidal closed category with finite products and coproducts, providing a compositional way to put together smaller nets into bigger ones. We then show how various sets of labels/weights used in chemical reaction networks can be given the structure of a lineale and thus be handled by our model. This chapter is based on [69].
- In the last part, we draw some conclusions and state some open questions.



## **Part II**

# **Computational History of Chemistry**



## CHAPTER 2

# Looking for large scale patterns

**Contents**

---

2.1	The role of synthesis . . . . .	12
2.2	On the selection of substrates . . . . .	12
2.3	Reaching regions of the chemical space . . . . .	16
2.4	Historical changes in the structure of the chemical space: growth of substances and reactions . . . . .	18
2.5	Chemical space density changes . . . . .	22

---

Millions of substances and chemical reactions have been reported throughout the history of chemistry expanding the structure of chemical space. Such empirical information, accounting for more than two centuries of chemical exploration, is now systematically recorded in Reaxys, a database built from the Beilstein and Gmelin Handbooks and the Patent Chemistry Database, which covers data from 16,400 journals and patents [70]. In this work, we use Reaxys data as a proxy for the chemical space. By January 2017 Reaxys reported 42,782,394 chemical reactions and 20,669,217 associated substances, the first record dates back to 1771. Given that single-step reactions are often contained in multi-step reactions, that there are few entries for the 18th century and to avoid those of the latter months of 2016, which were still under curation and annotation, our analysis starts in 1800 and runs until 2015. The data for this study comprises 16,356,012 reactions and 14,341,955 compounds. We defined the publication year of a compound as its earliest report in a reaction.

In this chapter, we present the results of our exploration of the entire chemical space for the period 1800-2015 using data from Reaxys database [84, 85]. The first three sections focus on three mechanistic features by which the space has been expanding: the role of synthesis and extraction in the production of new substances, patterns in the selection of starting materials, and frequency with which reactions reach new regions of the space. In the fourth section, we quantify the growth of chemical space (1800-2015) and relate growth changes with scientific and social historical events. In the last section, we show how the density of chemical space has changed over the history of chemistry.

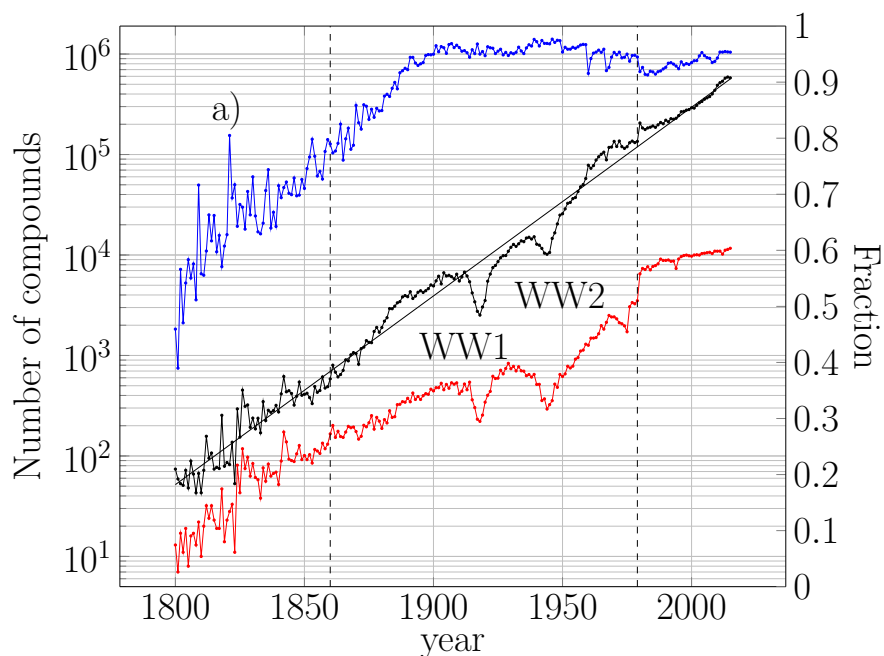
## 2.1 The role of synthesis

Wöhler's synthesis of urea in 1828 has become a legend [92]. Although historians of science have argued that it is just a myth [115], the claim that organic synthesis kicked off after Wöhler's synthesis of urea is part of the folklore of chemistry [106, 114]. To our knowledge, there is no quantitative account on this subject nor on the role that synthesis has played throughout history in the expansion of the space. We address that question here. We computed the annual ratio of synthesized compounds over reported ones, following the rule: if  $A + B \rightarrow C + D$  is the reaction where  $A$  and/or  $C$  is for the first time reported, we say that  $A$  is *extracted* and  $C$  is *synthesized*. We found that only during the first four years of the 19th century the percentage of new substances coming from synthesis was slightly lower than 50%. Afterward more than 50% of them are due to synthesis as shown in Figure 3, blue shows that, over history, more than half of the new substances have come from synthesis. At the time of Wöhler's synthesis, new substances containing C, H, N, O were about 50%, and so, organic synthesis was already well established (Figure 3a and Table 1). Moreover, the percentage of synthesized compounds increased to 90% by 2000 and has remained at such levels ever since.

In the next section, we turn our attention to the selection of substrates and to the products reached through chemical reactions. Our aim is to address such questions as whether chemists prefer to work with well-known substrates, and whether those preferred substrates change over time. Also, whether there is a higher diversity of products than of substrates, that is, whether chemists typically recombine sets of known substrates or they combine known ones with new ones, in order to reach new regions of the space.

## 2.2 On the selection of substrates

By counting how often a substance is used as a reactant, i.e., the outdegree of substances, we obtained the heavy tailed distributions presented in Figure 4a, indicating that there is a relatively



**Figure 3:** Growth of compounds. Annual number of new compounds (black). WW1 and WW2 indicate the World War periods and the vertical dotted lines the change in production regime. Annual number of new compositions (left axis; red) and the fraction of new synthesized compounds to the total of new ones (right axis; blue).

small set of substances preferred as starting materials but most chemicals are used as substrates only once. We give further details of these preferences below.

The analysis of the organic chemistry space (where a similar distribution was obtained) claims that the outdegree distribution follows a power-law type [39]. Our analysis rejects the power-law hypothesis. Therefore, the mechanism underlying the selection of substrates cannot be modelled by multiplicative or Yule processes only, or any other mechanisms generating these type of distributions [105]. Other distributions were statistically tested and likewise rejected (Tables 2 and 3). The selection of substrates results then from a different and perhaps more complex process, whose study bears the weight of further investigation.

Figure 4a shows two jumps in the distribution of outdegree, first between 1860-1879 and 1880-1899, and then between 1960-1979 and 1980-1999. The first one marks the transition to a prolific period of production of new compounds from about 1800 to 1990, seen in Figure 3 (black) as one of the regions where the number of new compounds is higher than the trend suggests. The second jump coincides with a transition to a prolific (Figure 3, black) and much more diverse period (Figure 3, red).

The 10 substances with higher outdegree, i.e., more frequently used as starting material, are shown in Table 4. Strong acids and bases appear as some of the most used substrates during the first half of the nineteenth century. Afterward, gradually, organic chemistry substrates took over. Ethanol

**Table 1:** List of the 10 most frequent combinations of elements in some particular years organized by lexicographic order. The second half of the table lists the remaining 10 most frequent combinations including metals. Non-carbon combinations are in red.

1800	1810	1820	1830	1840	1850	1860
CHNaO	CHAlO	CHNO	CHNO	CHO	CHN	CHO
CHNO	CHNO	FeO	HOSZn	CHNO	CHNO	CHNO
CHCuO	CHKO	CFeN	HIKO	CHCINO	CHCINPt	CHN
OS	OPb	AuHO	CIHHgN	CHCIO	AsCaHO	CHCIO
CuHOS	CBaO	CBaN	OPZn	FeHKOS	HNOPT	CHOS
S	Pb	Fe	CIHIKO	CCIO	CHFeNO	CHCI
OSn	KTe	FeS	CuOS	CHOS	CHO	IKSb
ClCuHO	CFeNS	CHFeNO	AsO	CHBrNO	CHOS	CHBrP
ClCu	NaTe	CrO	CIHO	CHCuNO	CHCIN	IRbSb
NaOS	HTe	CCoO	CHO	NaOSb	CHNOPt	HINSb
CHgNO	CHNaO	HNaOS	CHNZn	CHKO	HNaOSe	ISbTI
CIHNaOPt	CHCuO	FeS	CHNiO	CHCIOPt	BHOSr	CHOPb
-	CuHOS	Fe	HNaOP	CIHNOPt	AsBaHO	CuHOS
-	OSn	CaCIHO	HNiO	CIHNPT	BHMgO	CHOSn
-	ClCuHO	AgOS	CuHOS	BHMgO	CFeNO	CHCaO
-	ClCu	P	FeHOS	HNaOP	CHCuNO	NaOSi
-	NaOS	MgO	CoNO	CHFeNO	CIHNOPt	CaHOSe
-	CHgNO	CKO	HNOZn	CHCuO	CIHNPT	CHSn
-	CIHNaOPt	CrS	ClCoHN	CHAgNOS	HNOPtS	CHISn
-	-	CHKMgO	BaHOP	CHOPb	CoS	HNaOS

**Table 2:** Definition of power-law, normal, exponential, and Poisson distributions. For each distribution, the basic functional form  $f(R)$  is given along with the normalization constant  $C$  such that  $\sum_{R=R_{min}}^{\infty} Cf(R) = 1$ , where  $R_{min}$  is the minimum  $R$  value from which the distributions begins to apply. The general expression for these distributions is  $p(R) = Cf(R)$ . erfc stands for the complementary error function [44].

Name	$f(R)$	$C$
Power law	$R^{-\alpha}$	$1 / \sum_{n=0}^{\infty} (n + R_{min})^{-\alpha}$
Exponential	$e^{-\lambda R}$	$(1 - e^{-\lambda}) e^{\lambda R_{min}}$
Poisson	$\mu^R / R!$	$[e^{\mu} - \sum_{k=0}^{R_{min}-1} \frac{\mu^k}{k!}]^{-1}$
Log-normal	$\frac{1}{R} \exp[-\frac{(\ln R - \mu)^2}{2\sigma^2}]$	$\sqrt{\frac{2}{\pi\sigma^2}} [\text{erfc}(\frac{\ln R_{min} - \mu}{\sqrt{2}\sigma})]^{-1}$

**Table 3:** Parameters and  $p$ -values for the distributions shown in Table 2, when applied to the distribution of substrates. The  $p$ -values were calculated by running 1,000 simulations.

Period	Parameters			
	Power-law	Exponential	Poisson	Log-normal
Before 1860	$R_{min} = 2$ $\alpha = 1.963773$ $p = 0.01$	$R_{min} = 13$ $\lambda = 0.03506312$ $p = 0$	$R_{min} = 13$ $\mu = 41.02849$ $p = 0$	$R_{min} = 2$ $\mu = -5.196269$ $\sigma = 2.827289$ $p = 0.374$
1860-1879	$R_{min} = 1$ $\alpha = 2.016267$ $p = 0$	$R_{min} = 1$ $\lambda = 0.7320951$ $p = 0$	$R_{min} = 771$ $\mu = 1106$ $p = 0.089$	$R_{min} = 1$ $\mu = -6.733486$ $\sigma = 2.985701$ $p = 0.291$
1880-1899	$R_{min} = 9$ $\alpha = 2.113$ $p = 0.49$	$R_{min} = 1$ $\lambda = 0.6836599$ $p = 0$	$R_{min} = 4398$ $\mu = 4815.667$ $p = 0.587$	$R_{min} = 2$ $\mu = -15.148219$ $\sigma = 4.012288$ $p = 0$
1900-1919	$R_{min} = 3$ $\alpha = 2.177336$ $p = 0$	$R_{min} = 1$ $\lambda = 0.6561324$ $p = 0$	$R_{min} = 3839$ $\mu = 6352.571$ $p = 0.038$	$R_{min} = 18$ $\mu = -50.422936$ $\sigma = 7.343984$ $p = 0.587$
1920-1939	$R_{min} = 2$ $\alpha = 2.160733$ $p = 0$	$R_{min} = 1$ $\lambda = 0.6735098$ $p = 0$	$R_{min} = 4177$ $\mu = 8774.154$ $p = 0$	$R_{min} = 27$ $\mu = -19.421922$ $\sigma = 4.956338$ $p = 0.837$
1940-1959	$R_{min} = 18$ $\alpha = 2.011466$ $p = 0.11$	$R_{min} = 1$ $\lambda = 0.7368087$ $p = 0$	$R_{min} = 8638$ $\mu = 13026.17$ $p = 0.029$	$R_{min} = 23$ $\mu = -15.601899$ $\sigma = 4.544417$ $p = 0.916$
1960-1979	$R_{min} = 1$ $\alpha = 2.340214$ $p = 0$	$R_{min} = 2$ $\lambda = 0.8119482$ $p = 0$	$R_{min} = 3574$ $\mu = 7080.636$ $p = 0.001$	$R_{min} = 50$ $\mu = -0.3703384$ $\sigma = 2.3445863$ $p = 0.927$
1980-1999	$R_{min} = 1$ $\alpha = 2.341685$ $p = 0$	$R_{min} = 1$ $\mu = 0.8174186$ $p = 0$	$R_{min} = 23031$ $\mu = 41872.17$ $p = 0.064$	$R_{min} = 94$ $\mu = -3.513336$ $\sigma = 3.218282$ $p = 0.101$
2000-2015	$R_{min} = 5$ $\alpha = 2.114837$ $p = 0$	$R_{min} = 2$ $\mu = 0.8376273$ $p = 0$	$R_{min} = 46242$ $\mu = 61491.5$ $p = 0.052$	$R_{min} = 1$ $\mu = -1035.13808$ $\sigma = 28.84289$ $p = 0.064$

(EtOH) was at the top of the list for almost a century, and then went down and finally got off the list by 2000-2015. An opposite trend was found for methanol (MeOH), which showed up in the list at the beginning of the twentieth century, and gradually became the second most used substrate in chemistry. A remarkable substrate is acetic anhydride; predicted theoretically in 1851, synthesized the next year [114], becoming the fifth most used substrate by 1880-1889 and the top substrate since 1940-1959; it is mainly used in acetylation reactions [52]. Methyl iodide, an important methylating agent [88], appears in the list in almost every period.

We turn now our attention to the degree of sets of substrates. We will focus on reactions of 1, 2, or 3 substrates since they account for 87.4% of the space and involve 6,081,963 compounds. A third of the reactions report a single substrate, while half of them (48.7%) two, and only 5.7% three substrates.

Figure 4c shows how often substrates that have participated in 1-substrate reactions are used in other reactions. We found that most reactions (about 87%) follow a log-log decay, but it is followed by a final surge). About half of the reactions have a substrate that has been used only once. At about 30 uses the remaining 13% of the 1-substrate reactions spread in a smile fashion over higher uses, indicating that these substances are also frequently used in reactions with more than one substrate. In particular, in the right uppermost part, a few extremely used substrates are observed, namely, Acetic anhydride ((Ac<sub>2</sub>O)), methyl iodide (MeI), methanol (MeOH), ethanol (EtOH), water (H<sub>2</sub>O), and formaldehyde (CH<sub>2</sub>O) (Figure 4c).

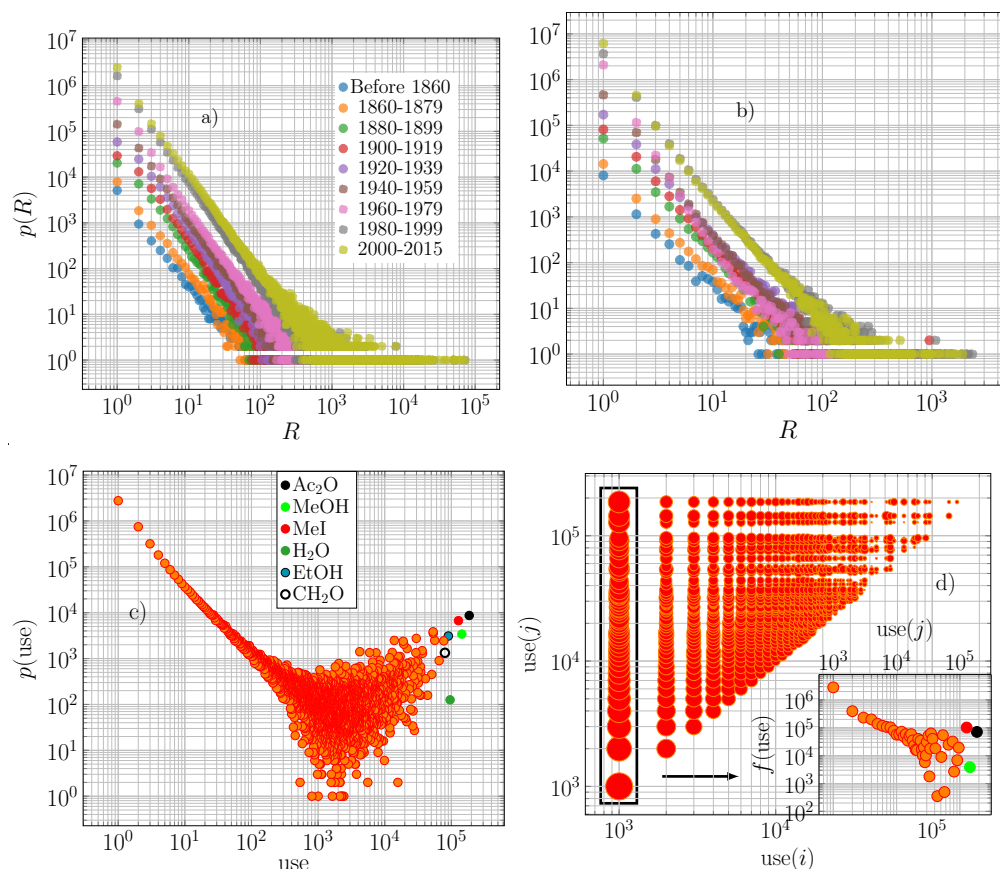
About half of the reported reactions use two starting materials ( $A + B \rightarrow$ ). Interestingly, in 94.3% of the reactions, one of these substances is of very low use (outdegree  $\leq 1000$ ). We investigated the outdegree values of the second one (see the portion of the distribution inside a rectangle in Figure 4d, which is presented in Figure 4d, inset). The most frequent outdegree values are also the lower ones, and frequency decreases monotonically as approaching larger values, except for some particular substrates that chemists often prefer to use (Figure 4d, inset), e.g. acetic anhydride, methanol, and methyl iodide. These recurrent substrates are part of the chemical toolkit for synthesis and for unveiling the chemical properties of new substances. Thus, chemists prefer the control of letting react less explored substances with better-known substrates. We call this the *fixed-substrate approach* to the exploration of the chemical space.

For reactions of three substrates ( $A + B + C \rightarrow$ ) we found that reactions including a substrate of low use (less than a thousand uses) and two substrates of any use account for 82.7% of the 3-substrate reactions. Among those, reactions that have two substrates of low use account for only 36.5%. As for the 2-substrate reactions, frequent substrates of these reactions are part of the chemical toolkit aforementioned.

## 2.3 Reaching regions of the chemical space

Figure 4b shows the distribution of indegree of substances. The distributions are heavy-tailed, as those for substrates, indicating that chemists have often synthesized some few products, while the majority of products is synthesized only once. As for substrates, the distributions of Figure 4b do not follow a power-law (Tables 2 and 6).

The distributions of indegree for products overlap for the periods 1980-1999 and 2000-2015 (Figure 4b), in spite of the exponential growth of new compounds. The stagnation in the report of new products in 2000-2015 was compensated by a higher fraction of new products than in 1980-1999.



**Figure 4:** Use and production of compounds. Frequency distributions of participation of compounds in  $R$  different reactions as a) substrates and b) products. The left hand side of the distributions corresponds to the many compounds appearing in few reactions, whereas the right-hand side to the few compounds appearing in many reactions. c) Frequency distribution of uses of substrates that have participated in a 1-substrate reaction. The following frequently used substrates are pin-pointed: Acetic anhydride ( $\text{Ac}_2\text{O}$ ), methyl iodide ( $\text{MeI}$ ), methanol ( $\text{MeOH}$ ), ethanol ( $\text{EtOH}$ ), water ( $\text{H}_2\text{O}$ ), formaldehyde ( $\text{CH}_2\text{O}$ ). d) Distribution of uses of substrates  $i$  and  $j$  that have participated in a 2-substrate reaction. The size of each point is proportional to the frequency of use of the couple  $\{i, j\}$  in reactions. Insert: Frequency distribution of use of  $j$  in 2-substrate reactions whose  $\text{use}(i) \leq 10^3$  is fixed, where some frequently used substrates are shown (see panel c this figure).

**Table 4:** Top-10 substrates over the history of chemistry. Abbreviations are found in Materials and Methods.

	Top	Before 1860	1860- 1879	1880- 1889	1900- 1919	1920- 1939	1940- 1959	1960- 1979	1980- 1999	2000- 2015
1	H <sub>2</sub> O	H <sub>2</sub> O	HCl	EtOH	EtOH	Ac <sub>2</sub> O	Ac <sub>2</sub> O	Ac <sub>2</sub> O	Ac <sub>2</sub> O	
2	NH <sub>3</sub>	HCl	EtOH	HCl	AcOH	EtOH	MeOH	MeOH	MeOH	
3	HNO <sub>3</sub>	EtOH	H <sub>2</sub> O	AcOH	HCl	AcOH	CH <sub>2</sub> N <sub>2</sub>	MeI	H <sub>2</sub> O	
4	HCl	H <sub>2</sub> SO <sub>4</sub>	AcOH	H <sub>2</sub> O	Ac <sub>2</sub> O	H <sub>2</sub> O	MeI	CH <sub>2</sub> N <sub>2</sub>	MeI	
5	H <sub>2</sub> SO <sub>4</sub>	HNO <sub>3</sub>	Ac <sub>2</sub> O	Ac <sub>2</sub> O	H <sub>2</sub> O	MeOH	CH <sub>2</sub> O	CH <sub>2</sub> O	PhCHO	
6	EtOH	Br <sub>2</sub>	H <sub>2</sub> SO <sub>4</sub>	H <sub>2</sub> SO <sub>4</sub>	Et <sub>2</sub> O	HCl	EtOH	PhCHO	CH <sub>2</sub> O	
7	Cl <sub>2</sub>	AcOH	MeI	MeI	H <sub>2</sub> SO <sub>4</sub>	C <sub>6</sub> H <sub>6</sub>	Morph	CuO	CO	
8	Na <sub>2</sub> CO <sub>3</sub>	NH <sub>3</sub>	PhNH <sub>2</sub>	Et <sub>2</sub> O	C <sub>6</sub> H <sub>6</sub>	Et <sub>2</sub> O	PhNH <sub>2</sub>	EtOH	TFA	
9	KOH	PhNH <sub>2</sub>	HNO <sub>3</sub>	PhNH <sub>2</sub>	MeOH	CH <sub>2</sub> O	DMA	BzCl	PhAcet	
10	I <sub>2</sub>	MeI	Br <sub>2</sub>	MeOH	Br <sub>2</sub>	MeI	PhCHO	CO	BnBr	

This is confirmed in Figure 3, where for 1980-1999 the percentage of synthetic products had an average of 91.1%, which rose to 93.5% for 2000-2015.

We found two jumps in Figure 4b, one between 1860-1879 and 1880-1899, the other between 1940-1959 and 1960-1979. They indicate that products in those transitions were more often obtained than the average trend (Figure 39 in Appendix A). The first jump coincides with the discussed first jump for substrates. The second one marks the transition from the WW2 recovery to the prolific period in the production of new compounds, from about 1960 up to 2000 (Figure 3, black).

The list with the 10 most synthesized compounds over history is shown in Table 7. To actually assess whether chemists have synthetic targets, we looked at the numbers of products in their reports. 81.6% of the reactions report no more than two products, and in fact, 74.2% report only a single product. It seems that chemists aim at synthesizing complex compounds, which are typically the heaviest products of reactions, i.e. the probability of picking up the right target then is 74.2%. The distribution of appearances of targets in reactions is shown in Figure (Appendix 38) and it follows the same trend as Figure 4b (Tables 2 and 5).

Table 8 lists the most synthesized targets per period. Organic acids such as oxalic and benzoic ones occur frequently; they are often used as synthetic intermediates [121]. Hydrogen sulfide and uranium hexafluoride are examples of targets motivated by nuclear research in the post-WW2 period [108, 125]. Likewise, organotin compounds and ferrocene evidenced the interest organometallics arose, especially in materials science and homogeneous catalysis [149]. We also want to draw attention to the presence of metallic oxides in the last period; they are often used in the synthesis of catalysts and nano-materials [159].

## 2.4 Historical changes in the structure of the chemical space: growth of substances and reactions

Reaxys is a corpus suitable for historical analyses of large scale patterns of expansion of chemical space. The growth in the number of new substances, for example, has been addressed in two previous studies for samples of the chemical space. The first study analyzed the growth in the number of new substances per year in the period 1800-1995; data was manually sampled from indices of eight

**Table 5:** Parameters and  $p$ -values for the distributions shown in Table 2, when applied to the distribution of targets. The  $p$ -values were calculated by running 1,000 simulations.

Period	Parameters			
	Power-law	Exponential	Poisson	Log-normal
Before 1860	$R_{min} = 2$ $\alpha = 2.417702$ $p = 0.186$	$R_{min} = 17$ $\mu = 0.06030108$ $p = 0.171$	$R_{min} = 56$ $\mu = 63.30087$ $p = 0.721$	$R_{min} = 1$ $\mu = -30.626118$ $\sigma = 4.792057$ $p = 0.186$
1860-1879	$R_{min} = 1$ $\alpha = 2.532562$ $p = 0.189$	$R_{min} = 12$ $\mu = 0.08234484$ $p = 0.003$	$R_{min} = 1$ $\mu = 1.269463$ $p = 0$	$R_{min} = 1$ $\mu = -7.279544$ $\sigma = 2.480881$ $p = 0.197$
1880-1899	$R_{min} = 3$ $\alpha = 2.658096$ $p = 0.223$	$R_{min} = 84$ $\mu = 0.01255286$ $p = 0.67$	$R_{min} = 1$ $\mu = 1.216067$ $p = 0$	$R_{min} = 5$ $\mu = -5.920010$ $\sigma = 2.291455$ $p = 0.584$
1900-1919	$R_{min} = 3$ $\alpha = 2.795804$ $p = 0.001$	$R_{min} = 47$ $\mu = 0.01875625$ $p = 0.064$	$R_{min} = 1$ $\mu = 1.197611$ $p = 0$	$R_{min} = 7$ $\mu = -20.806110$ $\sigma = 3.752783$ $p = 0.001$
1920-1939	$R_{min} = 7$ $\alpha = 2.636928$ $p = 0.063$	$R_{min} = 90$ $\mu = 0.01074621$ $p = 0.514$	$R_{min} = 1$ $\mu = 1.094216$ $p = 0$	$R_{min} = 11$ $\mu = -11.385738$ $\sigma = 3.071208$ $p = 0.738$
1940-1959	$R_{min} = 2$ $\alpha = 3.339249$ $p = 0$	$R_{min} = 68$ $\mu = 0.0181202$ $p = 0.886$	$R_{min} = 1$ $\mu = 0.5954598$ $p = 0$	$R_{min} = 1$ $\mu = -2.467482$ $\sigma = 1.333142$ $p = 0$
1960-1979	$R_{min} = 1$ $\alpha = 4.198562$ $p = 0$	$R_{min} = 1$ $\mu = 2.454067$ $p = 0$	$R_{min} = 1$ $\mu = 0.1825004$ $p = 0$	$R_{min} = 1$ $\mu = -6.231822$ $\sigma = 1.609130$ $p = 0$
1980-1999	$R_{min} = 1$ $\alpha = 3.479212$ $p = 0$	$R_{min} = 2$ $\mu = 1.154287$ $p = 0$	$R_{min} = 1$ $\mu = 0.390401$ $p = 0$	$R_{min} = 1$ $\mu = -6.062233$ $\sigma = 1.795392$ $p = 0$
2000-2015	$R_{min} = 1$ $\alpha = 3.855568$ $p = 0$	$R_{min} = 2$ $\mu = 0.954044$ $p = 0$	$R_{min} = 1$ $\mu = 0.2859205$ $p = 0$	$R_{min} = 1$ $\mu = -118.983190$ $\sigma = 7.210826$ $p = 0$

**Table 6:** Parameters and  $p$ -values for the distributions shown in Table 2, when applied to the distribution of products. The  $p$ -values were calculated running 1,000 simulations.

Period	Parameters			
	Power-law	Exponential	Poisson	Log-normal
Before 1860	$R_{min} = 2$ $\alpha = 2.314181$ $p = 0.049$	$R_{min} = 59$ $\mu = 0.0187567$ $p = 0.718$	$R_{min} = 161$ $\mu = 203.6254$ $p = 0.207$	$R_{min} = 1$ $\mu = -718.58135$ $\sigma = 23.31244$ $p = 0.049$
1860-1879	$R_{min} = 1$ $\alpha = 2.451058$ $p = 0$	$R_{min} = 68$ $\mu = 0.01125429$ $p = 0.941$	$R_{min} = 1$ $\mu = 1.671426$ $p = 0$	$R_{min} = 1$ $\mu = -16.007938$ $\sigma = 3.667497$ $p = 0.011$
1880-1899	$R_{min} = 5$ $\alpha = 2.479312$ $p = 0.513$	$R_{min} = 2$ $\mu = 0.8289619$ $p = 0$	$R_{min} = 1$ $\mu = 1.511293$ $p = 0$	$R_{min} = 1$ $\mu = -3.909269$ $\sigma = 1.944246$ $p = 0$
1900-1919	$R_{min} = 8$ $\alpha = 2.467412$ $p = 0.476$	$R_{min} = 2$ $\mu = 0.8198701$ $p = 0$	$R_{min} = 1$ $\mu = 1.435699$ $p = 0$	$R_{min} = 1$ $\mu = -1.929820$ $\sigma = 1.488447$ $p = 0$
1920-1939	$R_{min} = 11$ $\alpha = 2.388553$ $p = 0.881$	$R_{min} = 2$ $\mu = 0.8359276$ $p = 0$	$R_{min} = 1$ $\mu = 1.333986$ $p = 0$	$R_{min} = 20$ $\mu = -17.738687$ $\sigma = 4.029873$ $p = 0.881$
1940-1959	$R_{min} = 1$ $\alpha = 2.947494$ $p = 0$	$R_{min} = 2$ $\mu = 1.148441$ $p = 0$	$R_{min} = 1$ $\mu = 0.7163911$ $p = 0$	$R_{min} = 1$ $\mu = -3.183353$ $\sigma = 1.522464$ $p = 0$
1960-1979	$R_{min} = 1$ $\alpha = 4.126991$ $p = 0$	$R_{min} = 63$ $\mu = 0.0152121$ $p = 0.13$	$R_{min} = 1$ $\mu = 0.2046896$ $p = 0$	$R_{min} = 1$ $\mu = -10.352775$ $\sigma = 2.075666$ $p = 0$
1980-1999	$R_{min} = 1$ $\alpha = 3.211598$ $p = 0$	$R_{min} = 2$ $\mu = 0.9182153$ $p = 0$	$R_{min} = 1$ $\mu = 0.576151$ $p = 0$	$R_{min} = 1$ $\mu = -8.801885$ $\sigma = 2.259969$ $p = 0$
2000-2015	$R_{min} = 1$ $\alpha = 3.672836$ $p = 0$	$R_{min} = 2$ $\mu = 0.9465415$ $p = 0$	$R_{min} = 1$ $\mu = 0.3613914$ $p = 0$	$R_{min} = 1$ $\mu = -206.291472$ $\sigma = 9.765858$ $p = 0$

**Table 7:** Most synthesized products. Abbreviations can be found in Appendix A.

	Before 1860	1860-1879	1880-1889	1900-1919	1920-1939	1940-1959	1960-1979	1980-1999	2000-2015
1	NH <sub>3</sub>	NH <sub>3</sub>	NH <sub>3</sub>	NH <sub>3</sub>	BZA	H <sub>2</sub> O	H <sub>2</sub> O	PhCHO	Glc
2	H <sub>2</sub> O	CO <sub>2</sub>	CO <sub>2</sub>	CO <sub>2</sub>	NH <sub>3</sub>	CO <sub>2</sub>	H <sub>2</sub>	CO <sub>2</sub>	CO <sub>2</sub>
3	CO <sub>2</sub>	AcOH	MAC	BZA	CO <sub>2</sub>	CH <sub>2</sub> O	H <sub>2</sub> S	Ethene	PhCHO
4	S	HCl	BZA	H <sub>2</sub> O	MAC	Methane	O <sub>2</sub>	Methane	Ph <sub>2</sub>
5	HCl	BZA	H <sub>2</sub> O	PhNH <sub>2</sub>	AcOH	AcOH	CO	BZA	CuO
6	O <sub>2</sub>	MAC	PhNH <sub>2</sub>	AcOH	H <sub>2</sub> O	BZA	CO <sub>2</sub>	C <sub>6</sub> H <sub>6</sub>	H <sub>2</sub>
7	Cl <sub>2</sub>	H <sub>2</sub> O	AcOH	OA	OA	Acetone	B(OH) <sub>3</sub>	PhAc	ZnO
8	Hg	OA	OA	MAC	FA	HCl	Ag	CO	PhAc
9	SO <sub>2</sub>	H <sub>2</sub> S	HCl	PhCHO	CH <sub>2</sub> O	MAC	N <sub>2</sub>	Ph <sub>2</sub>	BZA
10	H <sub>2</sub>	S	EtOH	HCl	HCl	NH <sub>3</sub>	FC	Acetone	CO

**Table 8:** Top-10 targets over the history of chemistry. Abbreviations are found in Appendix A

	Top	Before 1860	1860-1879	1880-1889	1900-1919	1920-1939	1940-1959	1960-1979	1980-1999	2000-2015
1	I <sub>2</sub>	MAC	MAC	BZA	MAC	MAC	H <sub>2</sub> S	TPPO	Glc	
2	OA	BZA	BZA	CO <sub>2</sub>	BZA	CO <sub>2</sub>	TBTC	CO <sub>2</sub>	CuO	
3	CO <sub>2</sub>	I <sub>2</sub>	PhA	MAC	CO <sub>2</sub>	BZA	TBTB	BZA	ZnO	
4	Hg	OA	NH <sub>3</sub>	NH <sub>3</sub>	I <sub>2</sub>	HCl	Ag	PhCHO	NiO	
5	Cl <sub>2</sub>	AcOH	OA	I <sub>2</sub>	OA	Acetone	FC	Ph <sub>2</sub> CO	CO <sub>2</sub>	
6	H <sub>2</sub>	DHBZA	H <sub>2</sub> O	OA	AcOH	CH <sub>2</sub> O	B(OH) <sub>3</sub>	PhAc	CoO	
7	MAC	NH <sub>3</sub>	CO <sub>2</sub>	PhNH <sub>2</sub>	NH <sub>3</sub>	MeCHO	Ag <sub>2</sub> S	NPhOH	MBPh	
8	BZA	PhA	PhNH <sub>2</sub>	PhCHO	HCl	AcOH	H <sub>2</sub> O	Ph <sub>2</sub> S <sub>2</sub>	BZA	
9	HgO	H <sub>2</sub> S	I <sub>2</sub>	H <sub>2</sub> O	MeCHO	H <sub>2</sub> O	TMTc	EDBB	Pd	
10	NH <sub>3</sub>	HCl	AcOH	AcOH	CH <sub>2</sub> O	H <sub>2</sub> S	UF <sub>6</sub>	CuO	Ph <sub>2</sub>	

printed sources, including handbooks of organic and inorganic chemistry [136]. It was found that new substances were being reported exponentially at a rate of  $r = 5.5\%$  during that period. A second study considered only organic substances from Beilstein database (now part of Reaxys) for the period 1850-2004; they found that the organic chemical space was also growing exponentially, at a rate of  $r = 8.3\%$  before 1900 and  $r = 4.4\%$  afterward [39].

For the entire chemical space, we observed that the annual number of new compounds grows exponentially (Figure 5), following a switching model [87]. The growth fitting was calculated by linear regression methods and the equation for the annual number of new compounds is given by:

$$s_t = 8.18 \times 10^{-33} e^{0.04324t}, \quad (2.1)$$

where  $t$  is a year between 1800 and 2015.

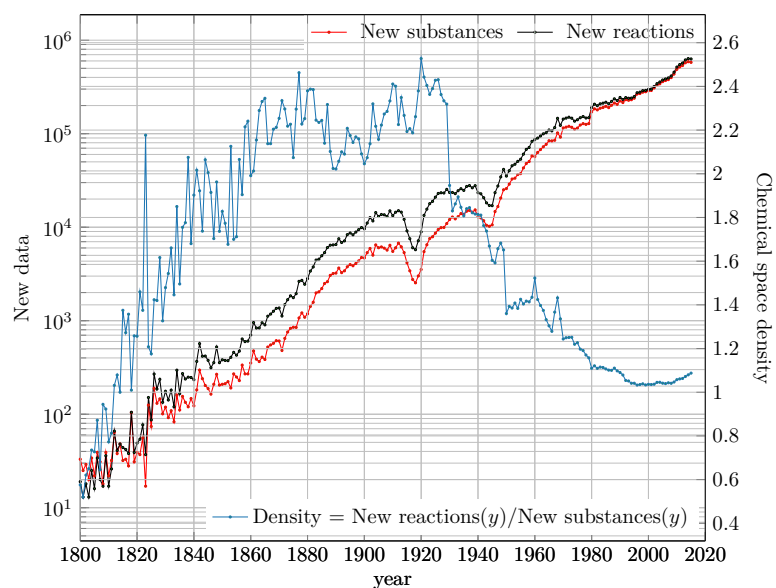
A salient feature is the effect of the World Wars, which appears as two dips in Figure 5. At its lowest point, WW1 set the production of new substances to numbers reported 37 years back, while WW2 set-back production by 16 years. The devastating effect of WW1 is due to the concentration of the chemical industry around Germany in pre-WW1 times [42]. In fact, WW1 led to the rapid strengthening of chemistry in the USA and other non-German countries [42], which is likely the reason why the drop in production during WW2 was not as dramatic. These results are the first quantitative account of the effect of World Wars on substance production.

## 2.5 Chemical space density changes

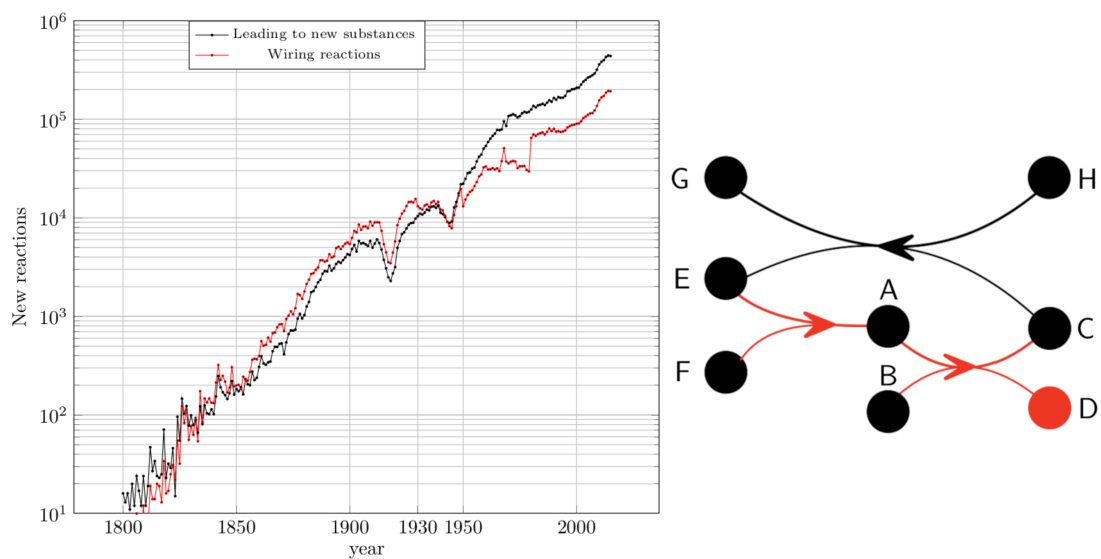
The chemical space is expanded by the discovery of new reactions or new substances. Chemical reactions serve at least two purposes: to make or use new substances through new reactions, or to discover a new reaction involving only known substances. We wonder what was the role of these two dynamics of space expansion throughout the history of chemistry. If we define the density of the chemical space in the year  $y$ , as the number of new reactions over the number of new substances reported that year, we see that the first type, which we call "new substance" reactions, can stall (if it involves only one new substance) or decrease the density of the chemical space (if there is more than one new substance per reaction). While the second type, or "wiring reactions", increases density. Figure 5, shows an increase in density over nearly a century (1800-1883), suggesting that wiring reactions were reported at a faster rate than the new substances reactions during this period. To verify this, we calculated the number of reactions of each type, whose evolution over time is shown in Figure 6. Indeed, the slope of the wiring reactions is greater than that of the reactions of new substances for this period. From 1883 and for five decades, there is a fall followed by a stagnation of density. This is due to a decrease in the rate at which wiring reactions were reported during this period, as Figure 6 shows. Around 1930, there is a dramatic drop in density that spans seven decades.

Figure 6 shows that the rate at which wiring reactions were reported decreased after 1930, while the rate at which new substances reactions were reported maintained a stable growth. We wonder if this change in the purpose of chemists to develop new reactions, now focusing on the use and discovery of new substances, brought about a change of emphasis in the type of substances used or produced. We wonder, for example, if chemists focused on more or less complex substances during this period. For this, we use the number of atoms and the molecular weight associated with each substance as a proxy for molecular complexity. Figure 7 shows the evolution of the average number of atoms per substance (black, left axis) and the evolution of the average molecular weight (red, right axis). Figure 7 shows that the average number of atoms per substance has grown steadily throughout the history of chemistry, with one exception: starting in 1930, the trend shifted toward the use and production of substances with more atoms, compared to the trend of the previous three decades; This was followed by a drop in the average number of atoms that lasted until 1950. On the other hand, the average molecular weight fluctuated during the first regime of space exploration, i.e., 1800-1860. This was followed by an increase in molecular weight lasting seven decades. Around 1930, the molecular weight gain plateaued and by 1950, it declined.

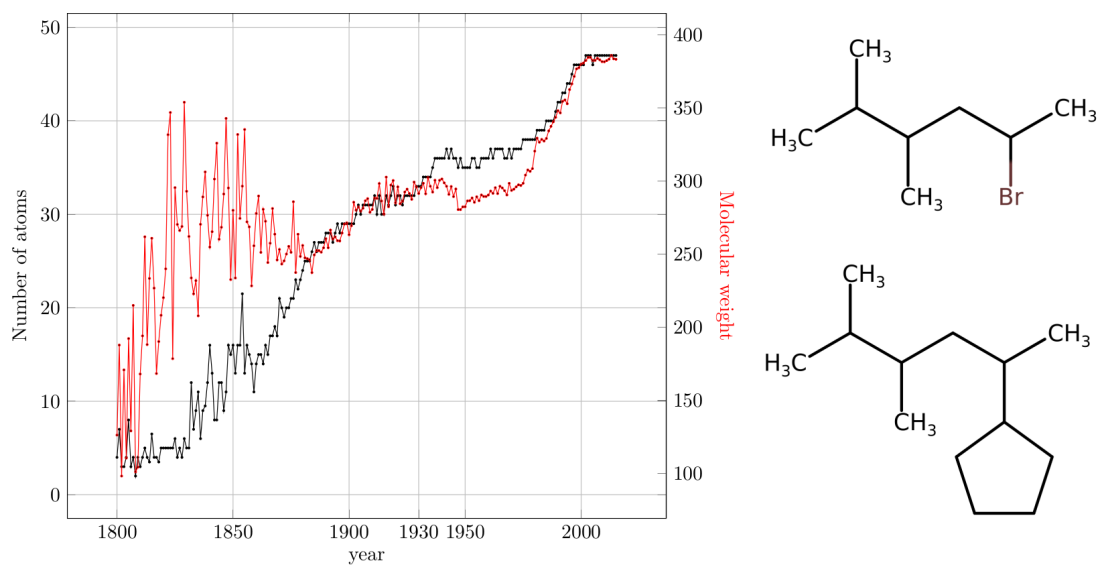
The above observations show that the drop in density around 1930 was due to a decrease in the number of wiring reactions, a steady growth in the number of new substance reactions, and the emphasis that chemists placed on using and producing new substances with more but lighter atoms.



**Figure 5:** Growth of the chemical space: Annual number of new compounds (red, left axis). Annual number of new reactions (black, left axis). Density of the chemical space, defined as:  $\text{Density} = \text{New reactions}(y) / \text{New substances}(y)$ ; (blue, right axis)



**Figure 6:** Chemists focused on producing new substances rather than discovering new ways to produce existing ones.



**Figure 7:** Chemists produced molecules with more but lighter atoms.

## CHAPTER 3

# Chemical space and similarity in the system of chemical elements

**Contents**

3.1	Similarity, atomic weights, and the expanding chemical space . . . . .	26
3.2	Evolution of the chemical space (1800-1868) . . . . .	26
3.3	Evolution of the system of chemical elements (1800-1868) . . . . .	30
3.3.1	Contemporary approach to the evolution of the system . . . . .	31
3.3.2	Historical approach to the evolution of the periodic system . . . . .	34

### 3.1 Similarity, atomic weights, and the expanding chemical space

A system of chemical elements is a structure depicting similarity and order relations among elements [76]. *Chemical similarity* arises from patterns of affinity of elements for other substances in chemical reactions [77]. For instance, alkali metals react with cold water to produce hydroxides such as LiOH, and NaOH; also, they react with oxygen to produce oxides like Li<sub>2</sub>O and Na<sub>2</sub>O. Chemical similarity can therefore be addressed by comparing the empirical and molecular formulas of compounds of two elements and using Mendeleev hypothesis that "[...] the elements, which are most chemically analogous, are characterized by the fact of their giving compounds of similar form [...]" [77, 97]. Coming back to the alkali metals, they are chemically analogous for many of their given compounds are of the same form, as illustrated by the examples XOH and X<sub>2</sub>O. On the other hand, *order* comes from atomic weights, which are estimated by finding the smallest common combining weight of a large set of molecular formulas containing a reference element [124]. These observations tell us that similarity between elements can be quantified by computing the number of forms they share [77], and therefore the outcome system strongly depends on the known chemical space [83], i.e., known substances and their molecular formulas.

UNESCO declared 2019 the "International Year of the Periodic Table", celebrating the 150th anniversary of the formulation of a periodic system of chemical elements around 1869, almost a decade after the Karlsruhe congress (1860) brought a standardized table of atomic weights and, in consequence, a normalization of molecular formulas. But substances were being reported at an exponential rate, from the beginning of the nineteenth century to 1869, and different tables of atomic weights were reported within that period, what was then the impact of the changing chemical space in the formulation of the system? Was the chemical space rich enough ever before the 1860's for chemists to formulate a periodic system that included most known elements? In this chapter, we aim to investigate these questions [74]. First, we report on the changing diversity of the chemical space between 1800 and 1869, then, we explore how the periodic system evolved from 1800 until its formulation, and finally, we investigate the impact of different nineteenth-century tables of atomic weights upon their systems.

### 3.2 Evolution of the chemical space (1800-1868)

Reaxys database is a corpus of empirical data on reactions, substances, and associated bibliographic sources, which has shown great potential for historical studies of chemistry [84, 119]. On January 2017, we retrieved records of reactions and substances reported from 1771 up to 1868 (two months before the publication of the first Mendeleev's system of elements [47]), accounting for 11,356 substances involved in 21,521 single-step reactions. Most records are credited to Gmelin's Handbook and were compiled from leading nineteenth-century journals [74]. Our data includes substances, their contemporary molecular formulas, and their earliest publication year in a chemical reaction. We manually curated 245 formulas coming from crystallisation species reporting non-integer stoichiometric coefficients (for instance, CdCO<sub>3</sub>\*0.5H<sub>2</sub>O was curated as CCdHO<sub>3.5</sub>). Furthermore, we discarded substances with intervals as stoichiometric coefficients (for example Ta<sub>1.15-1.35</sub>S<sub>2</sub>). Also, we discarded substances containing Er, Yt, and Di, that were problematic by 1868 (Appendix B). Our proxy for the chemical space by 1868 amounts to 11,356 substances composed of the 60 elements depicted in Figure 8.

Chemists reported new substances at an exponential rate [84], from just a few in 1800 up to

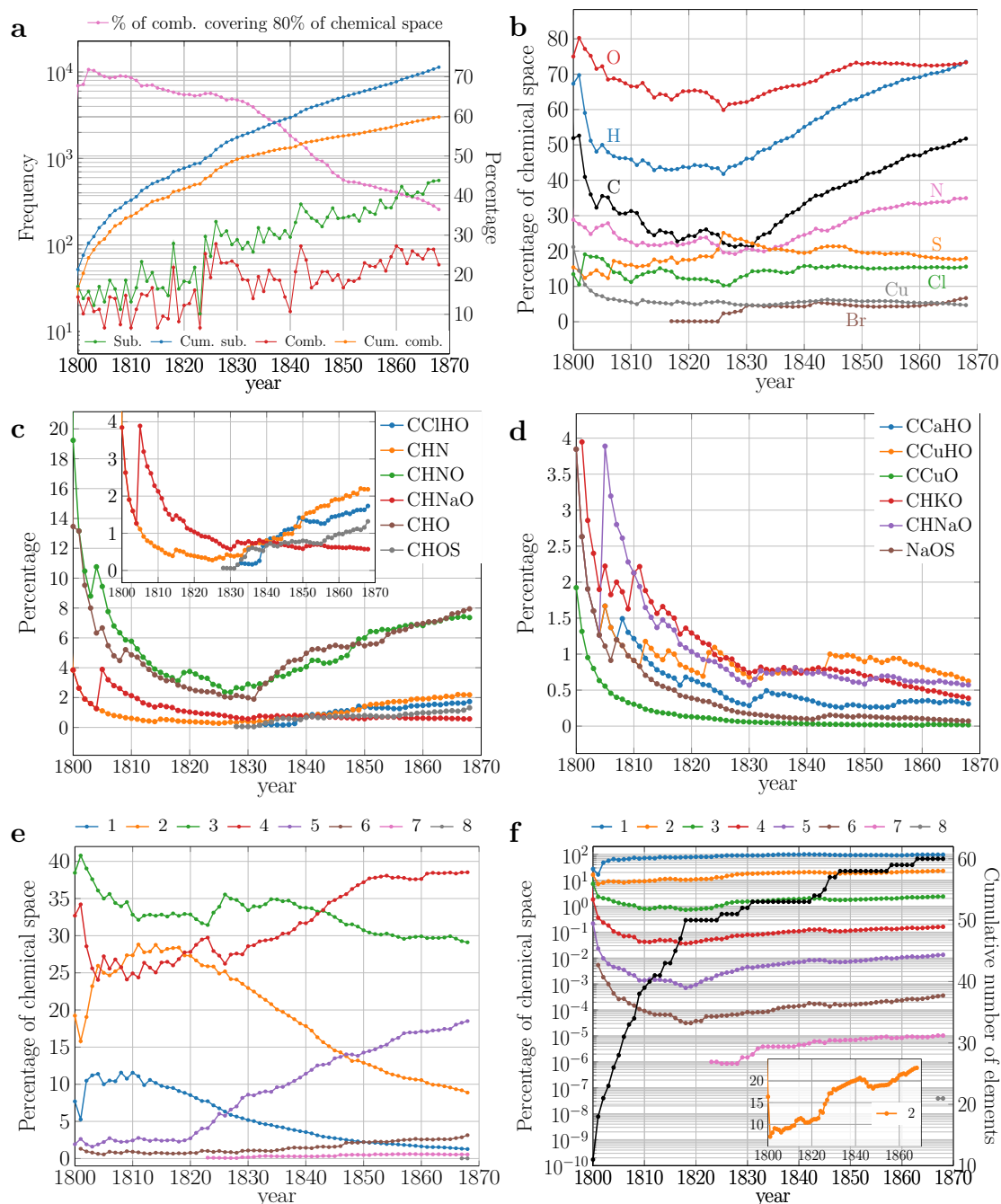
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba		Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra		Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg	Cn	Nh	Fl	Mc	Lv	Ts	Og
		La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	
Di	Er																
		Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	

**Figure 8:** Current system of chemical elements: elements known by 1869 (black), undiscovered elements (grey), and mixtures that were thought to be elements (red). Elements in black were considered in this study.

around 11 thousand by 1868 (Figure 9a). As a proxy for diversity changes, we studied how the number of new combinations of elements in molecular formulas changed over time (for instance, HOS is the combination associated with  $\text{H}_2\text{SO}_4$ ). New combinations were reported exponentially from 1800-1868 [84] albeit not at the same rate as new substances as shown in 9a. The pink curve shows that the percentage of combinations accounting for 80% of the space dropped from 70% around 1800 to 36% in 1868. In contrast to new substances, the number of new combinations was not increasing during 1830-1850, which translated into a very rapid decline in diversity, as shown by the pink curve. After 1850, the growth of the number of new combinations resumed and slowed down the dropping diversity.

Figure 9b provides further details about the turning point of 1830: around 1800, the chemical space was mainly populated by O, H, and C compounds; then, during the first quarter of the century their presence was reduced, which indicates that chemists explored new combinations of other elements (Figures 9c and 10). During this period the number of new substances and of new combinations grew jointly (Figure 9a). As shown in Figure 9c, CHO and CHNO reached their minimum around 1830 and then followed a steady growth; this emphasis on CHO and CHNO lasted until 1868 (Figure 9c). This pattern and the rapid increase in the number of C, H, and N compounds, is a consequence of the organic revolution [65, 124], which left two periods: before 1830 when most new combinations were metallic, and afterward when most were organic (Figures 9c, and 9d; Table 9). Figure 10 shows the growing occupancy of organic chemistry after 1830, where substances containing typical organic molecular fragments exploded, in contrast with those containing inorganic ones.

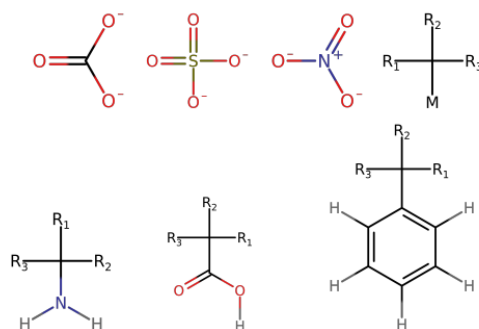
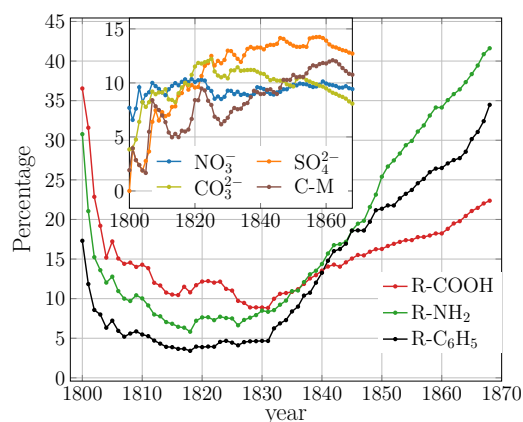
To continue assessing changes in diversity within the period 1800-1868, we turn our attention to the number of elements present in combinations, which we will call their size (Figure 9e). Substances of size 3 were the most abundant during the first two-thirds of this period, accounting for 30% of the space, surpassed only in 1842, by compounds of 4 elements, which came to cover 38% of the space by 1868. By then, compounds of size 5 were the third most abundant set in the space: despite a slow



**Figure 9:** Diversity of the chemical space up to 1869. a) Absolute (left axis [l.a.]) and cumulative values (right axis [r.a.]) of new substances and combinations. b) Percentage of chemical space spanned by some elements. These percentages are non-additive, e.g.  $\text{H}_2\text{O}$  contributes to both H and O counts. c) Percentage of chemical space spanned by different combinations. d) Percentage of chemical space spanned by some combinations containing metals. e) Percentage of chemical space spanned by substances made of  $n$  elements. After 1811 the number of uncombined forms (unary substances) in which elements appeared exceeded the number of known elements as a consequence of the allotropic forms and polymorphs of elements. For instance, by 1868 sulfur had nine uncombined forms. f) Cumulative number of elements (r.a.) and percentage of theoretical combinations of different sizes actually observed (l.a.).

**Table 9:** Most frequent combinations. Combinations including metals are in red and those made of metals and carbon are highlighted in gray.

1800	1805	1810	1815	1820	1825	1830	1835	1840	1845	1850	1855	1860	1865	1868
CHNO	CHNO	CHNO	CHNO	CHNO	CHNO	CHNO	CHO	CHO	CHO	CHNO	CHNO	CHO	CHO	CHO
CHO	CHO	CHO	CHO	CHO	CHO	CHO	CHNO	CHNO	CHNO	CHO	CHO	CHNO	CHNO	CHNO
CCuHO	CHNaOCHNaO	CHKO	CHKO	CHKO	CCuHO	CHKO	CHKO	CHN	CCIHO	CHN	CHN	CHN	CHN	CHN
CHN	OS	CHKO	CHNaO	CHNaO	CHKO	CCuHO	CHNaO	CCIHO	CHOPb	CCIHO	CCIHO	CCIHO	CCIHO	CCIHO
CHNaO	CHKO	HNO	CHOPb	CHOPb	CHNaO	CHNaO	CCuHO	CHKO	CHN	CCuHO	CCuHO	CHOS	CHOS	CHOS
Cu	CCuHO	OS	HOSb	HOSb	CHOPb	HNOP	CHN	CCuHO	CCuHO	CHOPb	CHOPb	CCuHO	BrCHO	BrCHO
NaOS	CCaHO	CCaHO	HNO	CCuHO	NaOW	CHOPb	CHOS	CCIH	CHKO	CHOS	CCIHNO	CCIHNO	CCIHNO	CCIHNO
OS	CIHg	CoHO	OS	HNOS	HOSb	NaOW	HNOP	CHNaO	CHOS	CCIHNO	CHOS	CHNOS	CHNOS	CHNOS
CCuO	CHN	CCuHO	CCuHO	HNOSe	HNOS	HNOS	CHOPb	CHOS	CCIHNO	CHKO	CHNaO	CHOPb	CCIH	CCIH
CH	Cu	CIHg	HNOS	HNO	KS	HOSb	HNOS	CHOPb	CCIH	CH	CH	CCIH	CCuHO	CH
CHgNO	NaOS	NaOS	OSb	OS	HNOSe	KS	CHNS	HNOS	CHNaO	CCIH	CHKO	CH	CH	BrCH
CNaO	CuO	AsO	CCaHO	OSb	OS	OS	CCaHO	CH	HNOS	CHNaO	CCIH	CHNaO	CHOPb	BrCHNO
ClCu	HNO	CoO	CoHO	CCaHO	CrO	CrO	NaOW	CCIHNO	CH	CHNS	CHNS	CHNS	CHNaO	CCuHO
ClCuHO	AsHO	NO	CIHg	NO	CFeN	AsHNaO	KS	CHNOS	BrCHO	HNOS	BrCHO	BrCHO	BrCH	CHNaO
CIH	AsO	Fe	AsO	CrO	CHNS	CHN	OS	FeHKOS	CHNOS	BrCHO	CHNOS	CHKO	CHNS	CHNS
CIHN	CO	AsHNiOHNOPbHHgOSe	SSb	FeHKOS	CrO	HNOP	CHNS	CHNOS	HNOS	AgCHNO	CHKO	CHOPb		
CIHNOPtCIHNTiHOSSn	NaOS	CoHO	HNO	HNOSe	FeHKOS	CCaHO	FeHOS	CCuHNO	CHMgO	CHIN	CHIN	CCIH		
CINa	ClO	CHN	CoO	CIHg	OSb	CFeN	NO	HOS	CIHNPT	FeHOS	CCuHNO	HNOS	CHS	CHIN
ClSn	CoHO	Cu	NO	AsO	CCaHO	CHNS	HOS	CHHgO	HOS	CIHNPT	AgCHNO	CHMgO	BrCHNO	CHS
CuHO	CoO	CuO	Fe	HNOPb	NO	SSb	CHNOS	CHNS	CHHgO	HOS	FeHOS	CCaHO	AgCHNO	AgCHNO

**Figure 10:** Typical organic, inorganic and organometallic molecular fragments. Left: distribution of tC-M means that at least one bond between C and M (see below) is reported on the table. It does not necessarily mean that C and M are bonded by a covalent bond. The structures of the analysed molecular fragments are shown on the right. In the inset, M stands for a metal, with  $M=\{\text{Li, Be, Al, Si, Fe, Co, Zn, As, Rh, Sb, Pt, Hg, Tl, Pb, Bi}\}$ . Right: Molecular fragments used to explore the evolution of the chemical space.

start (accounting for less than 3% for almost two decades), their presence began to increase steadily after 1818. This coincides with a drop in the percentage of the space covered by binary compounds, from 27% in 1818 to 8% in 1868. The biggest compounds for this period had eight elements (Figure 9e).

Now we want to investigate how many of the theoretical combinations of a given size were actually realised during 1800-1868. This number depends on the cumulative number of elements observed in substances that were involved in reactions up to a given year. This number went from 11 elements in 1800 to 60 in 1868 (Figure 9f). Since the theoretical number of combinations of size  $s$  is  $\text{theo}(n, s) = \binom{n}{s}$ , we define the theoretical number of combinations of (up to)  $n$  elements as  $\text{theo}(n) = \sum_{s=2}^n \binom{n}{s}$ . For instance, the number of theoretical combinations on 11 elements is 2,036 while there are  $1.15 \times 10^{18}$  on 60 elements. This is a rough upper bound disregarding valency and compound stability. The percentage of theoretical combinations actually observed (Figure 9f) corresponds to  $\exp(n, s)/\text{theo}(n, s)$ , where  $\exp(n, s)$  is the number of substances whose combination contains  $s$  of the  $n$  elements available.

The nineteenth century began with a rapid growth in the cumulative number of elements and the corresponding explosion of theoretical combinations (Figure 9f), which lasted until 1818. This explosion brought a rapid drop in the fraction of realised combinations. Once it slowed down, more combinations were observed increasing the proportion of realised theoretical combinations. In the mid-1840s a few more new elements came, reducing again the proportion of realised combinations, which coincides with a strong drop in the number of new substances, from 300 in 1842 to 163 in 1846 (Figure 9a). As expected, binary combinations were always closer to their theoretical possibilities than combinations of more elements. By 1825, after the stabilisation of the number of new elements in substances, about 10% of the theoretical number of binary compounds was reported (Figure 9f, inset), a growing percentage not even affected by the emphasis on compounds of three, four, and five elements (Figure 9e). In fact, by 1868 about 23% of the possible binary compounds (with 60 available elements) were already known (Figure 9f, inset).

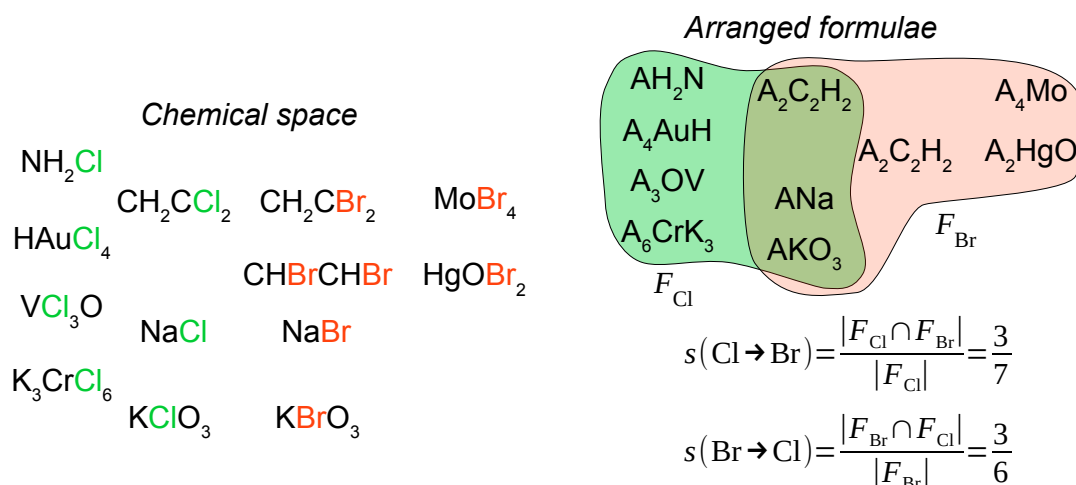
### 3.3 Evolution of the system of chemical elements (1800-1868)

In this section, we investigate how the evolution of the chemical space affected the system of chemical elements using a data-driven approach. Historians have concluded that the ripe moment for formulating the system came in the 1860s [10, 134], due largely to the standardised set of atomic weights and the associated normalisation of molecular formulas as a consequence of the Karlsruhe conference in 1860. But up to now, there is no account of how the evolution of the chemical space affected the system. We wondered whether the chemical space became rich enough for the formulation of the system only in the 1860s; or whether it could have been formulated earlier. We also assessed the impact of different nineteenth-century tables of atomic weights and their corresponding formulas upon the resulting systems. We analysed the interplay between the chemical space and the system of chemical elements from two perspectives, one using contemporary formulas and another one using formulas reconstructed from the atomic weight tables of nineteenth-century chemists. The *contemporary approach* “sees” the chemical space of the nineteenth-century through the eyes of twenty-first-century chemistry. For instance, the nineteenth-century formulas for water is  $\text{H}_2\text{O}$ , while Dalton’s was  $\text{OH}$ . Reaxys provides this data. The *nineteenth-century-tables approach* considers the evolution of the chemical space using formulas reconstructed from a given nineteenth-century table using an algorithm described below; in this approach, for instance, we use Dalton’s table

to reconstruct his formulas, which yields OH, Dalton's known formula for water. Therefore, the nineteenth-century-tables approach attempts to use the formulas of the leading chemists of that time.

### 3.3.1 Contemporary approach to the evolution of the system

Our approach to quantifying similarity among chemical elements is illustrated in Figure 11 (see Appendix B for a detailed description). This measure of similarity computes the number of forms shared by two elements [77], which follows Mendeleev's idea of similarity between elements [97].

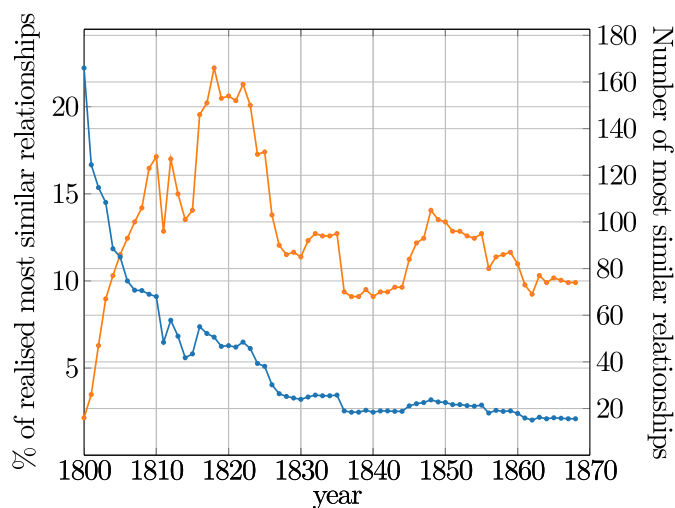


**Figure 11:** Similarity among chemical elements. Toy chemical space of 13 substances. Each compound provides an arranged formula for an element in the given formula when Cl or Br is replaced by A and the elements are lexicographically ordered. Arranged formulas of element A are gathered in  $F_A$ , which is a multiset as elements may appear more than once, e.g.  $\text{A}_2\text{C}_2\text{H}_2$  appears twice in  $F_{\text{Br}}$  (B). The similarity of an element  $x$  to the element  $y$  is given by  $s(x \rightarrow y)$ , which is the probability of  $x$  having a common arranged formula with  $y$ . In chemical terms, it is a measure of substitutability. This similarity is an asymmetric relation [99], e.g.  $s(\text{Br} \rightarrow \text{Cl}) > s(\text{Cl} \rightarrow \text{Br})$ , and generalizes that used in [77]. For instance, by 1869 we have  $s(\text{Br} \rightarrow \text{Cl}) = 344/659 = 0.52$ , while  $s(\text{Cl} \rightarrow \text{Br}) = 349/1,556 = 0.22$ . This means that Br could be substituted by Cl to obtain a known compound in roughly half of Br combinations, whereas Cl could be substituted by Br in about one-fourth of those of Cl.

We display only maximal similarities for each element since it is customary to present a system of chemical elements as tables, where similarities between neighbouring elements are the largest. Non-maximal but important similarities can be recovered from sequences of maximal similarity relationships, for instance, Li being most similar to Na and Na to K means that likely Li is quite similar to K as well (Figure 13). Families (groups) of elements on periodic tables will therefore appear as elements related by maximal similarities. After determining the maximal similarities of chemical elements, all that remains is to retrieve the system of elements to arrange them according to atomic weight. We depict these systems as similarity networks. Figure 13a-c presents three of them. All 69 networks can be found in the Interactive Information.

Despite the increasing number of elements (Figure 9f), the number of maximal similarities decreased over time, from a maximum of 166 in 1818 to 69 in 1862 (Figure 12). This indicates

that as chemists expanded the chemical space, they were converging towards a core set of similarity relationships. To assess this convergence, we calculated the similarity between systems resulting from each year in the period 1800-1868 (Figure 14a). The reddish region around the diagonal in Figure 14a indicates high similarity among adjacent years, in particular, the most similar periodic system of any year is always one of an adjacent year. The dark blue regions around the early years indicate that those systems transformed completely within the span of a few years and are essentially different from the system at the end of the period. Similarities in this early quarter of the century were mainly related to forms such as chlorides, oxides, hydroxides, sulfates, and other typical inorganic compounds (Interactive Information). By 1826, there was a sharp transition, as revealed by the light blue-yellow square in Figure 14a, which indicates that around 40% of the similarities found in 1826's system are also present until 1868. Some of these early known similarities were  $\text{Ag} \rightarrow \text{K}$  and  $\text{Pt} \rightarrow \text{Pd}$ , caused mainly by their inorganic compounds. Other similarities showing up a decade later were  $\text{K} \rightarrow \text{Na}$ ,  $\text{Hg} \rightarrow \text{Cu}$ ,  $\text{Si} \rightarrow \text{Ti}$ ,  $\text{Fe} \rightarrow \text{Co}$  and  $\text{Ni} \rightarrow \text{Co}$ , which were also mainly caused by inorganic compounds. Interestingly, famous nineteenth-century systems as those published by Meyer and Mendeleev in the 1860s, depict Cu and Ag as similar elements (Figure 40), mainly due to the similar low oxygen content of their oxides [95]. However, our results show that, ever since 40 years earlier, Cu had been more related to Zn group, mainly for its dominating +2 valence, and Ag to K, for its dominating +1 valence (Figure 13). By 1826, known chemical families such as halogens and Fe, Co, Ni were already formed (Figure 13a-c).



**Figure 12:** Number (orange) and percentage (blue) of realised maximal similarities. The orange curve shows the number of maximal similarities among elements for the system of year  $y$  ( $s_y$ ), that is the number of arrows in networks of Figure 13a-c. The percentage of realised maximal similarities among chemical elements over time (blue curve) is calculated as  $(s_y/u_y) \times 100$ , where  $u_y$  is the possible number of relationships for the year  $y$ , corresponding to  $u_y = n_y \times n_y$ , with  $n_y$  the number of known elements in year  $y$ .

The reddish region between columns 1826-1860 and rows 1835-1845 (Figure 14a) shows that about 80% of the similarities of the systems observed between 1835 and 1845 were present since 1826 and lasted until 1860. During the 1860s, similarity dropped down to about 60%. The period after 1845 shows that most similarities observed after this year lasted, and some others were just

transient, for instance, those of Nb, Ta, Rb and Cs.

Element discovery and tie-breaking drove the convergence to the 1868's system. In the early years of the century, only a few compounds were known as to unveil those similarities by 1868. For example, up to 1825, no combinations of Ce were known, which caused this element to be similar to almost any other element (Figure 13a). The first compound of Ce,  $\text{Ce}_2\text{S}_3$ , was reported in 1826, which brought Ce closer to other elements bearing the form (sulfide)  $\text{X}_2\text{S}_3$  and pruned similarity networks for the years to come, reflected in the transition on 1826. This mechanism was triggered by several newcomer elements, eventually leading to the systems of 1837-1845, which reproduced 80% of the similarities observed up to the 1860s. Therefore, a fairly accurate system could have been proposed as early as the 1840s. But atomic weights were still not standardised by then and different formulas were probably presumed for several compounds. We will estimate the effect of different atomic weight tables in the next section.

A system of elements with the most frequent maximal similarities of elements is depicted in Figure 13d, which we regard as the backbone of the periodic system from 1800 up to 1868. It shows known families of elements including alkali metals, halogens, chalcogens, pnictogens (without N), and {Fe, Co, Ni}, plus well-known families of transition metals such as {Pd, Pt, Ir} and {Mo, W, Ta}.

The values of maximal similarities are actually very small: more than 80% of the similarities had values lower than 0.1. The lowest similarity values ever recorded corresponded to those of organogenic elements. This fits Mendeleev's concept of "typical elements" [94, 96], today called singularity principle or uniqueness of second-period elements [117], which indicates that these elements hold weak similarities with elements of their families [48].

If the similarities were so small, how could they become so noticeable to chemists? This is even more surprising if we consider that each nineteenth-century chemist may not have had the complete knowledge provided to us by the database. We believe it has to do with ubiquity: these similarities span over the whole spread of the chemical space so that they are equally visible in any reasonably-sized portion of the chemical space. To test this hypothesis, we took random samples of different sizes of the space, for every year, and analysed how often the maximal similarities among elements were present in the samples. We found that most of the similarities observed in the first quarter of the nineteenth-century required more than 50% of the chemical space to be detected, indicating that in this period similarities of different elements were focused on different regions of the chemical space, hindering the discovery of the patterns of the system (Figure 14b). As time went by, especially after 1830, similarities became more ubiquitous and easier to detect. This effect is particularly intense for similarities among organogenic elements, as they were spread among the increasingly large number of organic compounds, which promptly became the majority of the chemical space. For instance, the similarity  $\text{S} \rightarrow \text{O}$ , detected as early as 1800, required at least 65% of the 1800 space to be observed, while by 1840 this fraction plummeted to 10% and dropped to 5% by 1868 (Figure 14b). In contrast,  $\text{Ba} \rightarrow \text{Ca}$  required in 1800 three quarters of the space, by 1840 60% and by 1868 a quarter, that is five times more chemical space than  $\text{S} \rightarrow \text{O}$  to be detected (Figure 14b). Therefore, the redundancy of the space on organogenic compounds facilitated the detection of similarities among organogenic elements, which contrasts with those of the transition metals, whose detection demanded the examination of a larger fraction of the chemical space.

This explains why nineteenth-century chemists, such as Meyer and Mendeleev, struggled with similarities among transition metals [46, 54, 79, 94, 96, 100–102, 109] (Figure 40, Table 12 Appendix B). Mendeleev also faced problems with the similarities of In and the rare earths he included in his system (Table 12 Appendix B) [148]. Remarkably, detecting  $\text{In} \rightarrow \text{Al}$  by 1869, as Meyer did, required more than 75% of the chemical space (Figure 14d). Examples of other similarities requiring

large amounts of chemical space to be detected were  $\text{Zn} \rightarrow \text{Mg}$ ,  $\text{Nb} \rightarrow \text{P,1}$  and  $\text{Nb} \rightarrow \text{Sb}$ . The first of these is explicit in Mendeleev and Meyer's 1869/70 systems (Figure 40), and the other two are explicit in Meyer's system and discussed as similarities by Mendeleev [94] (Figure 40, Table 12 Appendix B). Overall, we found that about 53% of the similarities among chemical elements arising from the chemical space were recovered by Meyer in his 1864 and 1868 systems (true positives, Table 13 Appendix B). Almost a quarter of nonsimilarities of the 1864 chemical space were observed as similarities by Meyer (false positives, Table 13 Appendix B). This fraction plummeted in 1868 to about 7%. At any rate, the best agreement between Meyer's systems and the system allowed by the chemical space was achieved in 1869/70, when 62% of the similarities of the space were gauged by his system, while there were only 6% nonsimilarities observed as similarities. Mendeleev, in turn, attained 58% and 10% of true and false positives, respectively (Table 13 Appendix B). Note that the (dis)agreements here discussed are based on the similarities reported by the two chemists in their systems, which were abundant and detailed in Mendeleev's case and very seldom discussed by Meyer, in which case similarities needed to be interpreted from his periodic tables. Also, the greater detail of Mendeleev's discussions on similarity is expected to yield a higher rate of false positives, due to our methodology being based on maximum similarities.

### 3.3.2 Historical approach to the evolution of the periodic system

The contemporary approach to the evolution of the system of chemical elements is based on the current table of atomic weights and molecular formulas. Nevertheless, analysing the historical evolution of the chemical space and its influence on the system requires considering the history of the atomic theory. That is, it requires considering the various nineteenth-century competing sets of atomic weights associated with different theoretical and experimental settings [123, 124], which led to chaos of formulas [12]. Since different atomic weights produce different orderings of the elements and different formulas, then chemists working with different systems of atomic weights may find distinct different systems, even if they worked with the same experimental data. Here we analyse the possible systems resulting from the different nineteenth-century chemical spaces that result from different atomic weight tables proposed over the period 1800-1868.

In the nineteenth-century, empirical data on composition came in the form of mass percentages for each element. For instance, Dalton knew that water was made of 88% and 12% by weight of oxygen and hydrogen, respectively. From Dalton on, chemists assumed formulas for key compounds such as water, ammonia, and oxides. Thus, chemists selected an element and assigned a reference atomic weight to it, upon which atomic weights of other elements were relative. The initial assumptions thus propagate through all the calculations, therefore creating a different chemical space for each chemist (Figure 15a). For example, Dalton's reference was an atomic weight of 1 for hydrogen. He assumed HO as the formula of water, therefore yielding an atomic weight of 7 for oxygen. This leads to molecular formulas of oxides whose coefficients are around half of those we know today. The determinations were made even more difficult by the varying quality of the experimental data [123, 124].

We gathered 13 sets of atomic weights (Table 10), corresponding to data published by Dalton (1810) [27], Thomson (1813) [147], Berzelius (1819 [17] and 1826 [14-16]), Gmelin (1843) [46], Lenßen (1857) [79], Meyer (1864 [101], 1868 [120] and 1869/70 [102]), Odling (1864) [109], Hinrichs (1867) [54] and Mendeleev (1869) [96], plus the current accepted atomic weights. Starting with Gmelin, these systems of atomic weights were proposed by authors who actually devised systems [134]. Although neither Dalton, Thomson nor Berzelius aimed at devising systems, they were some of

the key figures in the development of the atomic theory [123, 124], which is why we also explored the effects of their atomic weights upon the systems that could have been obtained from their respective chemical spaces. Figure 43 (Appendix B) shows the elements comprised by each system of atomic weights, which range from 30 for Dalton to 60 for Mendeleev. Information on the selection of these elements is found in Table 14 (Appendix B).

As any system of chemical elements is based on ordering and similarity of its chemical elements [76], we analysed the different orderings of elements associated to each set of weights. In all cases, they agreed in more than 80%, even with the current atomic weights (Table 15 Appendix B). This indicates that the ordering relationships among elements were rather stable since the beginning of the nineteenth-century. To determine element similarities it is necessary to reconstruct the formulas spanned by each system of atomic weights (Figure 15a). As there is no systematic record of the chemical formulas corresponding to the assumptions of each chemist, we devised an algorithm to obtain approximate formulas meeting the assumptions of the chemists here analysed (B). This entails, for instance approximating the current  $\text{Fe}_2\text{O}_3$  to  $\text{FeO}_3$  according to Dalton (Figure 15a). Our procedure takes all Reaxys formulas known by the time of publication of each chemist's atomic weights and rescales the modern formulas with 20 levels of *tolerance* (B). Often, the higher the tolerance, the lower the perturbation of Reaxys formulas.

For each system associated to each chemist-tolerance, we calculated the fraction of 1868 similarities it contains. This quantifies how close could have been the similarities known by 1868 of being detected by each chemist (Figure 41). For the sake of comparison, this fraction was normalised by the actual fraction of 1868 similarities that could have been observed by the time of chemist's publication (Figure 15b, red). That is, for instance, normalising Berzelius 1819 similarities with those provided by the actual 1819 chemical space. As a chemist's space could lead to several transient similarities not standing the test of the time, we also quantified chemist's fraction of similarities of this sort (Figures 15b, blue and 42).

By inspecting Figure 15b, we observe how, as the century progressed, chemists' systems could have contained more and more 1868 similarities and how transient similarities were further reduced. There is a remarkable leap with Gmelin, who becomes a turning point in the trends, separating systems with many transient similarities and few standing the test of time (on Gmelin's left in Figure 15b) from systems rich in 1868 similarities and with very few transient similarities (on Gmelin's right in Figure 15b). Gmelin's system contains about 78% of the 1868 similarities and about 40% of his similarities are transient. This is actually an improvement when contrasted with the systems of his predecessors. For instance, Dalton's system would have had about 10% of 1868 similarities and 93% of transient ones, and Berzelius (1826) 60% and 73%, respectively. The lack of accuracy of pre-Gmelin systems is caused by the many changes the chemical space underwent. Nevertheless, in those times, Berzelius' 1819 system stands out. In spite of its 75% of transient similarities, Berzelius' atomic weights would have led to a system with 63% of 1868 similarities.

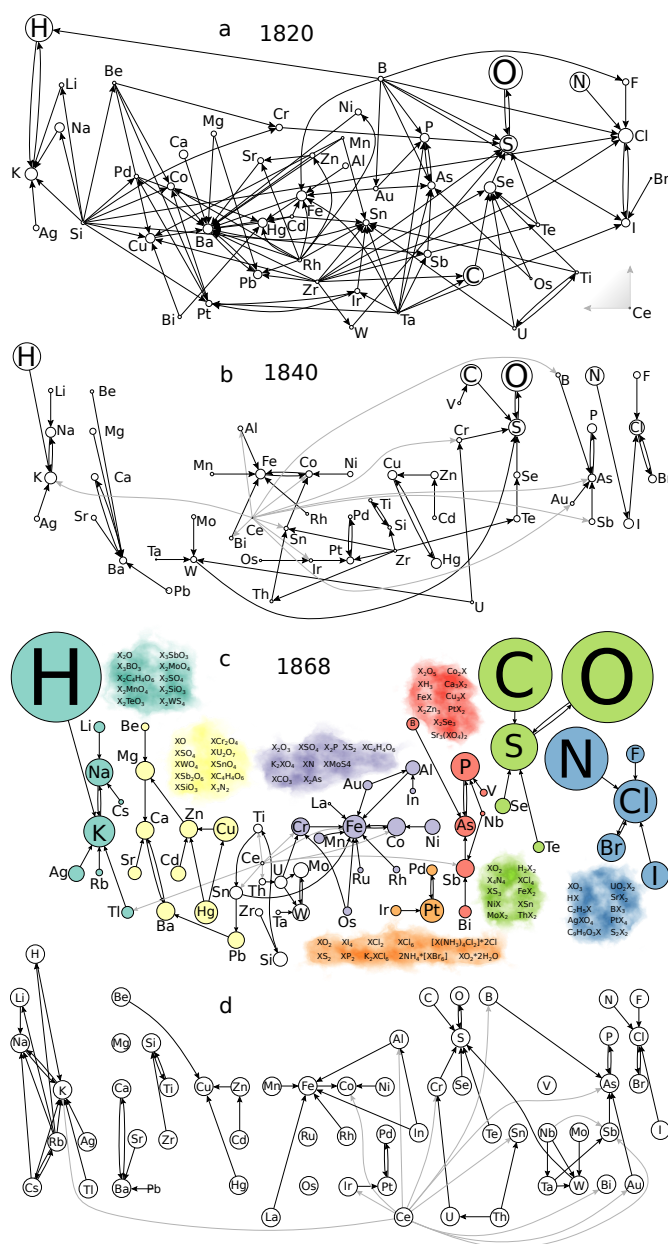
The remarkable separation of the two plots (Figure 15b) after Gmelin shows the strong relationship between the theoretical and experimental advances the atomic theory brought about and the raise of the backbone of the periodic system. Interestingly, this is particularly evident in the 1840s and not in the 1860s as traditionally accepted, which agrees with our contemporary approach results.

By analysing the systems by Meyer and Mendeleev, we found that each new Meyer's system would have achieved more 1868 similarities and reduced the number of transient similarities. His last set of atomic weights would have led to a system with no transient similarities matching 82% of the 1868 similarities. In turn, Mendeleev's system would have shown 83% of 1868 similarities and would have contained 6% of transient similarities. These results coincide with the different stances the

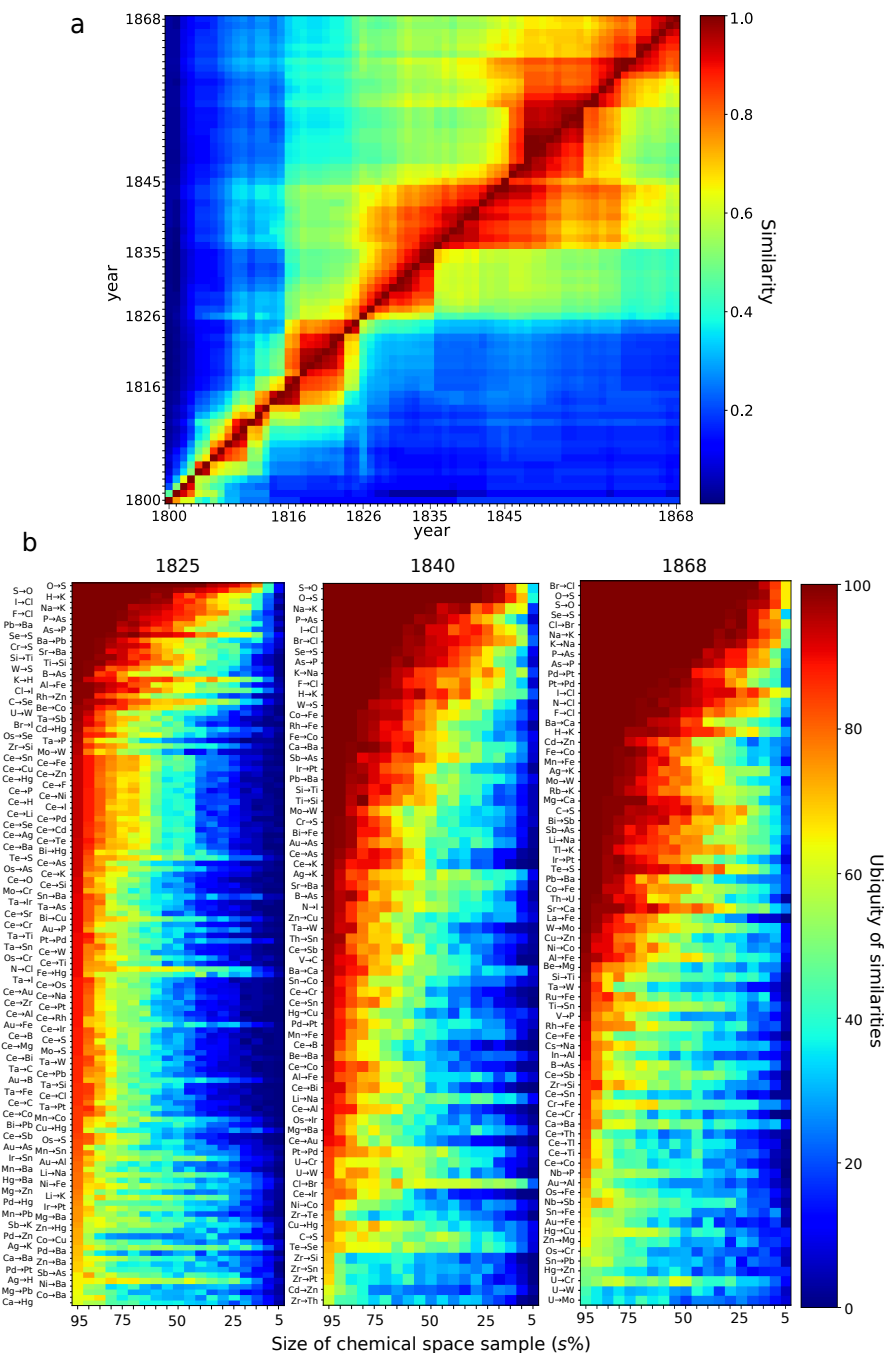
**Table 10:** Atomic weights as reported by nine nineteenth-century chemists. The primary sources for these figures are: Dalton [27], pages 546, 547; Thomson [147]; Berzelius 1819 [17]; Berzelius 1826 [14–16]; Gmelin [46], pages 50, 51; Lenßen [79], pages 122, 123 (we assume 1 for H); Meyer 1864 [101], we assume 1 for H; Odling [109], page 642; Hinrichs [54], pages 13, 14; Meyer 1868 [120, 122]; Mendeleev [96]; Meyer 1869/70 [102], current [93].

	Dalton 1810	Thomson 1813	Berzelius 1819	Berzelius 1826	Gmelin 1843	Lenßen 1857	Meyer 1864	Odling 1864	Hinrichs 1867	Meyer 1868	Mendeleev 1869	Meyer 1869/70	Current
H	1	0.132	6.2175	6.244	1	1	1	1	1	1	1	1	1.008
O	7	1	100	100	8	8	16	16	16	16	16	15.96	15.999
C	5.4	0.751	75.33	75	6	6	12	12	12	12	12	11.97	12.011
N	5	0.878	77.26	88.518	14	7	14.04	14	14	14.04	14	14.01	14.007
S	13	2	201.165	207.58	16	16	32.07	32	32	32.07	32	31.98	32.06
P	9	1.32	392.3	196.15	31.4	16	31	31	31	31	31	30.9	30.973
Cl	33	4.498	142.65	220	35.4	17.7	35.46	35.5	35.5	35.46	35.5	35.38	35.45
Na	28	5.882	581.84	290.92	23.2	23	23.05	23	23	23.05	23	22.99	22.989
K	42	5	979.83	487.915	39.2	39.11	39.13	39	39.1	39.13	39	39.04	39.0983
Ca	24	2.62	512.06	256.03	20.5	20	40	40	40	40	40	39.9	40.078
Mg	17	1.368	316.72	158.36	12.7	12	24	24	24	24	24	23.9	24.305
Sr	46	5.9	1094.6	547.3	44	43.67	87.6	87.5	87.6	87.6	87.6	87	87.62
Ba	68	8.731	1713.86	856.93	68.6	68.59	137.1	137	137	137.1	137	136.8	137.327
Fe	50	6.666	678.43	339.215	27.2	28	56	56	56	56	56	55.9	55.845
Cu	56	8	791.39	395.695	31.8	31.7	63.5	63.5	63.4	63.5	63.4	63.3	63.546
Zn	56	4.315	806.45	403.225	32.2	59	65	65	65.2	65	65.2	64.9	65.38
Ag	100	12.618	2703.21	1351.605	108.1	108	107.94	108	108	107.94	108	107.66	107.8682
Hg	167	25	2531.6	1265.8	101.4	100	200.2	200	200	200.2	200	199.8	200.592
Pb	95	25.974	2589	1294.5	103.8	103.6	207	207	207	207	207	206.4	207.2
Li			255.63	127.8	68.6	6.95	7.03	7	7	7.03	7	7.01	6.94
Be	30	3.6	662.56	331.28	17.7	7	9.3	9	9.3	9.3	9.4	9.3	9.012
B			69.655	135.98	10.8	11		11	11		11	11	10.81
F			75.03	116.9	18.7	9.5	19	19	19	19	19	19.1	18.998
Al	15	2.136	342.333	171.667	13.7	13.7		27.5	27.4	27.3	27.4	27.3	26.981
Si	45	4.066	296.42	277.8	14.8	15	28.5	28	28	28.5	28	28	28.085
Ti	40			389.1	24.5	25	48	50	50	48	50	48	47.867
Cr			703.638	351.86	28.1	26.8		52.5	52.2	52.6	52	52.4	51.9961
Mn	40	7.13	711.575	355.787	27.6	27.5	55.1	55	55	55.1	55	54.8	54.938
Co	55	7.326	738	369	29.6	29.5	58.7	59	60	58.7	59	58.6	58.933
Ni	50	7.305	739.51	369.755	29.6	29.6	58.7	59	58	58.7	59	58.6	58.6934
As	42	6	940.77	470.385	75.2	37.5	75	75	75	75	75	74.9	74.921
Se			495.91	494.59	40	39.7	78.8	79.5	79.4	78.8	79.4	78	78.971
Br					78.4	40	79.97	80	80	79.97	80	79.75	79.904
Rb							85.4	85	85.4	85.4	85.4	85.2	85.4678
Zr	45	5.656		420.21	22.4	33.6	90	89.5	89.6	90	90	89.7	91.224
Ce	45	11.494	1149.44	574.72	46.3	47.3		92			92		140.116
Mo		5.882	596.8	598.56	48	46	92	96	92	92	96	95.6	95.95
Rh			1500.1	750.65	52.1	51.2	104.3	104	104.4	104.3	104.4	104.1	102.905
Ru						52.1	104.3	104		104.3	104.4	103.5	101.07
Pd		14.204	1407.5	714.6	53.4	53.2	106	106.5	106.6	106	106.6	106.2	106.42
Cd			1393.54	696.77	55.8	20	111.9	112	112	111.9	112	111.6	112.414
U	60	12	3146.86	2711.36	217	60		120	120		116		238.028
Sn	50	14.705	1470.58	735.29	59		117.6	118	118	117.6	118	117.8	118.71
Sb	40		1612.9	806.45	129	60	120.6	122	122	120.6	122	122.1	121.76
Te		4.107	806.45	403.225	64	64.2	128.3	129	128	128.3	128	128	127.6
I				783.35	126	63.5	126.8	127	127	126.8	127	126.5	126.904
Cs							133	133	133	133	133	132.7	132.905
Tl							204	203	204	204	204	202.7	204.38
Bi	68	11.111	1773.8	1330.4	106.4	104	208	210	210	208	210	207.5	208.98
Th					59.6	59.5		231.5	231		118		232.0377
In									71		75.6	113.4	114.818
V					68.6	68.5	137	137		137	51	51.2	50.9415
La					36.1	47		92			94		138.905
Nb											94	93.7	92.906
Ta			1823.15	1152.87	185	92.3	137.6	138	137.6	137.6	182	182.2	180.947
W	56	8	1207.689	1183.2	95	92	184	184	184	184	186	183.5	183.84
Pt	100	12.161	1215.226	1215.23	98.7	99	197.1	197	198	197.1	197.4	196.7	195.084
Ir					98.7		197.1	197	198	197.1	198	196.7	192.217
Os					99.6	99.4	199	199		199	199	198.6	190.23
Au	140		2486	1243	199	98.4	196.7	196.5	197	196.7	197	196.2	196.966

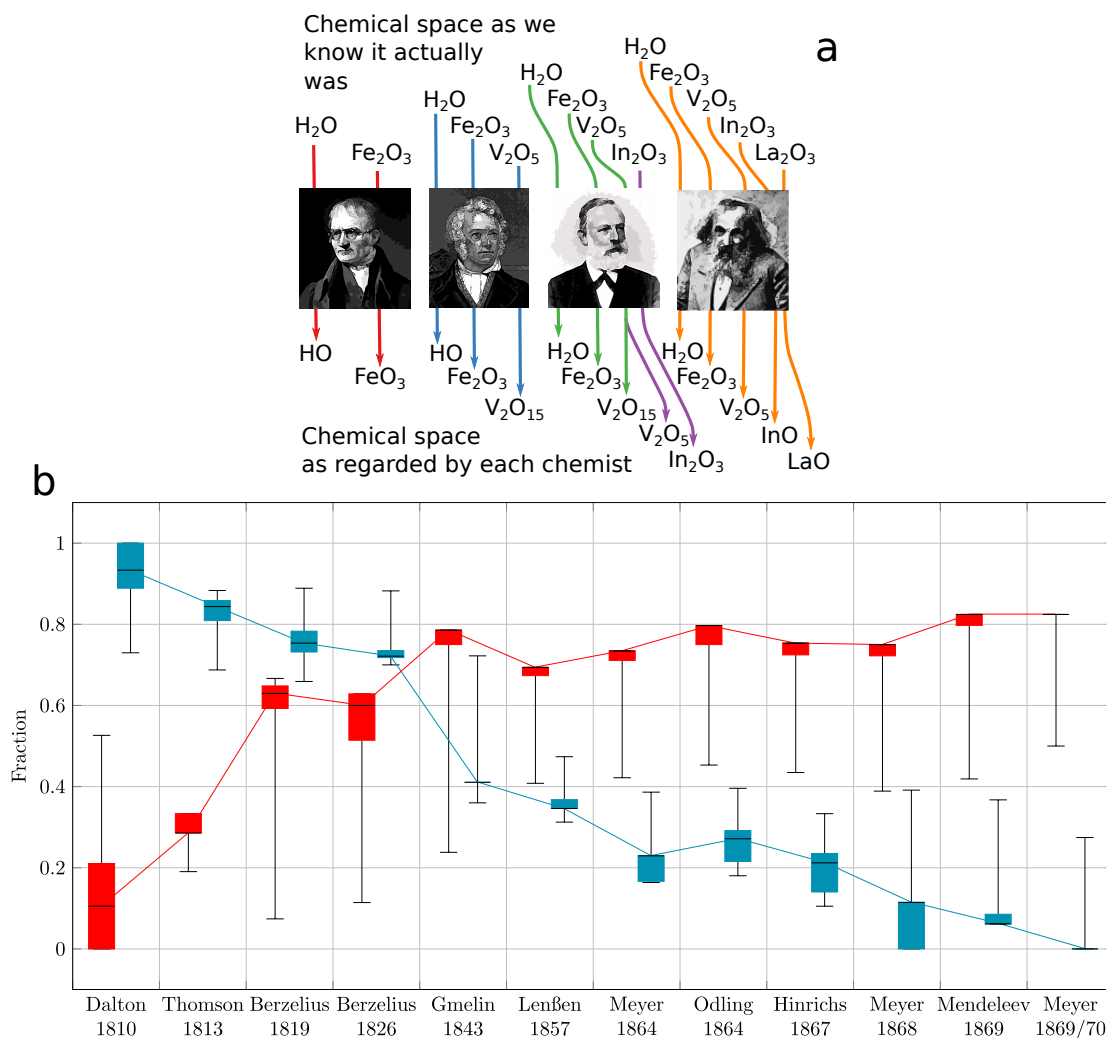
two chemists had regarding the system. Meyer favoured accurate atomic weights and experimental information and Mendeleev completeness [49, 122], as noted in the several elements left aside by Meyer, which were included by Mendeleev.



**Figure 13:** Evolution of the system of chemical elements. a-c) systems of three different years. Arrows  $x \rightarrow y$  indicate maximal similarities, i.e., that  $x$  is most similar to  $y$ . Node (element) size is proportional to the number of substances composed by the element. Similarities of Ce are coloured in light grey for the sake of readability. In a all Ce similarities are collapsed for the sake of simplicity. Some of the formulas shared by elements with the same colour are shown in c, where same-colour-elements correspond to X. Readers can also select in the Interactive Information any set of elements to retrieve the formulas making similar the elements in any particular year. d) Backbone of the system depicting the recurrent pairs of most similar chemical elements (Appendix B).



**Figure 14:** Similarity among systems of chemical elements and ubiquity of element resemblances. a) Resemblance between systems. The heatmap depicts similarity from the system of the column to the system of the row (Appendix B). Any row  $y$  indicates how similar the systems are, year after year, to the system of year  $y$ . Any column  $x$  shows which fraction of the system, year after year, is similar to the system of year  $x$ . b) Ubiquity of the similarities of the systems of 1825, 1840, and 1868. The ubiquity of each similarity corresponds to the percentage of appearance of such similarity in the sampled space of size  $s\%$  (Appendix B). Plots for every year from 1800 to 1868 are found in [74]. Appendix B contains heatmaps for every five years.



**Figure 15:** Contrast between systems of chemical elements by nineteenth-century chemists and the system of 1868. a) Examples of modified formulae according to the atomic weights of (left to right) Dalton (1810), Thomson (1813), Berzelius (1819), Berzelius (1826), Gmelin (1843), Lenßen (1857), Meyer (1864, green; 1869/70, purple), and Mendeleev (1869). For every chemist publishing a set of atomic weights in year  $y$ , known Reaxys substances ( $S_{y-1}$ ) up to year  $y - 1$  (inclusive) were retrieved, and the corresponding system  $P_{y-1}$  was obtained. Formulae of substances  $S_{y-1}$  were approximated with 20 different tolerance values ( $\tau$ ), each  $\tau$  yielding a system with similarities gathered in  $P_{y-1}^\tau$  (Materials and Methods, Supporting Information). The fraction of similarities allowed by the actual chemical space of  $y - 1$  is given by  $Z = |P_{y-1} \cap P_{1868}| / |P_{1868}|$ . b) Fraction of similarities observed in 1868 and available by the chemical space by  $y - 1$  that appear in chemists' systems (red). They are calculated as  $Z - |P_{y-1}^\tau \cap P_{1868}| / |P_{1868}|$  and depicted as box-plots (median (black), maximum and minimum values as whiskers). c) Fraction of transient similarities for each chemist's system, depicted as box-plots (blue, same settings as in b) and calculated as  $1 - |P_{y-1}^\tau \cap P_{1868}| / |P_{y-1}^\tau|$ .

**Part III**

**Geometry and Compositionality in  
Chemical Space**



## CHAPTER 4

# Curvature of hypergraphs and chemical space

## Contents

4.1	Introduction and outline . . . . .	44
4.2	A very brief story of Forman-Ricci curvature . . . . .	45
4.3	Forman-Ricci curvature of graphs . . . . .	46
4.3.1	Undirected graphs . . . . .	46
4.3.2	Directed graphs . . . . .	46
4.4	Forman-Ricci curvature of undirected hypergraphs . . . . .	48
4.4.1	Definition and bounds . . . . .	48
4.4.2	Curvature of the Wikipedia voting network . . . . .	50
4.5	Curvature of directed hypergraphs . . . . .	51
4.5.1	Definition and bounds . . . . .	51
4.5.2	Comparison with Ollivier-Ricci curvature . . . . .	53
4.5.3	Hyperloops and their curvature . . . . .	55
4.5.4	Curvature of metabolic networks . . . . .	57
4.6	Curvature of chemical space . . . . .	63
4.6.1	Indegree, outdegree, and the role of substances . . . . .	63
4.6.2	Forman-Ricci curvature and assortativity of chemical reactions . . . . .	65

## 4.1 Introduction and outline

Graphs are mathematical structures designed to encode binary relationships. Graph theory is a well developed area of mathematics devoted to these structures, but the networks that result from empirical data modeling are much less regular and often much larger than the graphs that are best understood. Therefore, various quantities have been introduced to characterize large-scale behavior or to identify the particularly important vertices in empirical networks, such as vertex degree and their distribution, clustering coefficients, betweenness centrality, assortativity, and, more recently, Forman-Ricci curvature [127].

Undirected and directed graphs are models for binary relations only (definition 1), but many empirical data sets encode higher order relations [140]. Authorship, for instance, is an undirected relation possibly among several individuals, in fact, by 2015 a physics paper set a record with more than 5,000 authors [24]. Chemical reactions, on the other hand, are evocative examples of directed relations between sets (reactants and products), and at least  $x\%$  of the reactions reported throughout history use or produce more than two substances. These systems are not only found in scientometrics and chemistry but also in physics, biology, computer science, combinatorial optimization, and several other fields<sup>1</sup>. Hypergraphs generalize graphs to encode simultaneous interactions between any number of entities [13, 18, 34, 130]. Several of the graph statistics have consequently been extended to hypergraphs, including vertex and hyperedge degrees, clustering coefficients [34, 64], and spectral properties [135]. Most of the commonly used quantities focus on vertices. As graphs and hypergraphs are models of relations, represented by edges, we shall systematically define and evaluate quantities assigned to the edges to complement vertex-centered measures.

In this chapter we present a brief review of graph curvature and its limits, then develop the Forman-Ricci curvature for undirected and directed hypergraphs and use it to explore social and biological networks, and finish the chapter using the Forman-Ricci curvature to probe the local structure of chemical space. This chapter is based on the following papers [31, 72, 73, 78].

Before moving to the next section, let us spell out the formal definitions of graphs [51] and hypergraphs [11, 18].

**Definition 1.** *Let  $V$  be a set, called vertices.*

- i) An undirected graph is a pair  $G = (V, E)$ , where  $E$  is a collection of two-vertex sets  $\{i, j\}$  called edges.*
- ii) A directed graph is a pair  $G = (V, E)$ , where  $E$  is a collection of ordered pairs  $e = (i, j)$  of vertices called directed edges. Moreover,  $i$  and  $j$  are called the tail and head of the directed edge  $e$ .*
- iii) A hypergraph is a pair  $G = (V, E)$ , where  $E$  is a collection of subsets of  $V$  called hyperedges.*
- iv) A directed hypergraph is a pair  $G = (V, E)$ , where  $E$  is a collection of pairs  $e = (e_1, e_2)$  of subsets of  $V$ . Moreover,  $e_1 \subset V$  is the tail and  $e_2 \subset V$  is the head of the directed hyperedge  $e$ .*

*If  $E$  is a set, the structures above are called simple; if  $E$  is a list, they are called multi.*

<sup>1</sup>See for instance [3, 8, 19, 43, 64, 67, 76, 103, 150, 158]

## 4.2 A very brief story of Forman-Ricci curvature

Recently various notions of “curvature” have been proposed for graphs and other, more general, discrete structures and applied to detect various local or global properties of such structures<sup>2</sup>. In 2003, Forman defined his notion of Ricci curvature for cell complexes [41]. While Forman’s definition applies to general CW-complexes, for our purposes it suffices to explain it for simplicial complexes. A simplicial complex  $\Sigma$  on a vertex set  $V = \{v_1, \dots, v_n\}$  consists of a collection of simplices, that is, subsets of  $V$ . When such a subset  $\alpha$  contains  $p + 1$ -vertices, it is called a  $p$ -simplex, because we can think of this combinatorial object also as a  $p$ -dimensional geometric simplex. A 0-simplex is simply a vertex, a 1-simplex is also called an edge, and a 2-simplex a triangle. We write  $\beta < \alpha$ , and say that  $\beta$  is a boundary-simplex of  $\alpha$ , when the  $(p - 1)$ -simplex  $\beta$  is a subset of the  $p$ -simplex  $\alpha$ . For the definition of a simplicial complex, it is required that whenever the  $p$ -simplex  $\alpha$  belongs to  $\Sigma$ , then so do all its boundary simplices  $\beta < \alpha$ . Thus, for instance, when our simplicial complex contains the triangle  $\{u, v, w\}$  with vertices  $u, v, w$ , it has to contain also the three edges  $\{u, v\}$ ,  $\{v, w\}$ , and  $\{u, w\}$ , as well as inductively also the 0-simplices  $\{u\}$ ,  $\{v\}$  and  $\{w\}$ . For the definition of a hypergraph, that is, a formal object of the type considered in this work, this condition will not be required (see definition 1). Thus, hypergraphs are more general objects than simplicial complexes, and correspondingly more difficult to treat mathematically, but this generality will be needed to adequately model metabolic networks, for instance, [73].

Returning to simplicial complexes for the moment, Forman defines functions

$$F_p : \{p\text{-simplices}\} \longrightarrow \mathbb{R} \quad (4.1)$$

by putting, for a  $p$ -simplex  $\alpha$ ,

$$\begin{aligned} F_p(\alpha) := & \#\{(p + 1)\text{-cells } \beta > \alpha\} \\ & + \#\{(p + 1)\text{-cells } \gamma < \alpha\} \\ & - \#\{\text{parallel neighbors of } \alpha\}, \end{aligned} \quad (4.2)$$

where, as mentioned,  $\beta > \alpha$  means that the  $p$ -simplex  $\alpha$  is contained in the boundary of the  $(p + 1)$ -simplex  $\beta$ , and analogously,  $\gamma < \alpha$  means that the  $(p - 1)$ -simplex  $\gamma$  is contained in the boundary of  $\alpha$ . And a parallel neighbor of the  $p$ -simplex  $\alpha$  is another simplex  $\alpha'$  of the same dimension that is disjoint from  $\alpha$ , but either contained in the boundary of some  $(p + 1)$ -simplex that also contains  $\alpha$  in its boundary or contains some  $(p - 1)$ -simplex in its boundary that is also contained in the boundary of  $\alpha$ , but not both. (As mentioned, the definition applies more generally to cell complexes, but that is not needed for our present purposes.)

The concept is perhaps most easily understood when we consider graphs. In fact, an undirected graph is a simplicial complex that has only 0-simplices (the vertices or nodes) and 1-simplices (the edges). The Forman curvature of an edge  $e$  with nodes  $i, j$  is simply given by  $F(e) = 4 - \deg(i) - \deg(j)$ . Here, the edge is not contained in any 2-simplex, because there are none in a graph, and hence the first term in (4.2) is 0. It has 2 0-simplices contained in it, its vertices  $i, j$ , and hence the second term is 2. Finally, it has  $\deg(i) + \deg(j) - 2$  parallel neighbors, the other edges emanating from the vertices  $i$  and  $j$ .

Edges connecting nodes with large degrees have very negative Forman-Ricci curvature values, and

<sup>2</sup>See [53, 86, 127–131, 141–143, 152, 154, 155]

these are typically edges that play a key role in the cohesion of a network. The Ricci curvature of a graph therefore can extract important information about the global structure of the graph from local quantities [73].

### 4.3 Forman-Ricci curvature of graphs

#### 4.3.1 Undirected graphs

Let  $G = (V, E)$  be a (multi)graph with vertex set  $V$  and multiset of edges  $E$ . The Forman-Ricci curvature of an edge  $e = \{i, j\} \in E$ , as introduced in [127], is given by:

$$F(e) = w_e \left( \frac{w_i}{w_e} + \frac{w_j}{w_e} - \sum_{e_l \sim i} \frac{w_i}{\sqrt{w_e w_{e_l}}} - \sum_{e_l \sim j} \frac{w_j}{\sqrt{w_e w_{e_l}}} \right) \quad (4.3)$$

where  $w_e$  denotes the weight of the edge  $e$ ,  $w_i$  and  $w_j$  are the weights of vertices  $i$  and  $j$ , respectively. The sums over  $e_l \sim k$  run over all edges  $e_l$  incident on the vertex  $k$  excluding  $e$ . The curvature for the unweighted multigraph, with vertex and edge weights set to 1, is given by [130]

$$F(e) = 4 - d_i - d_j \quad (4.4)$$

where  $d_k$  is the number of edges that are incident to  $k$ , called the degree of  $k$ . Defining  $D = \sum_{k \in e} d_k$  we have

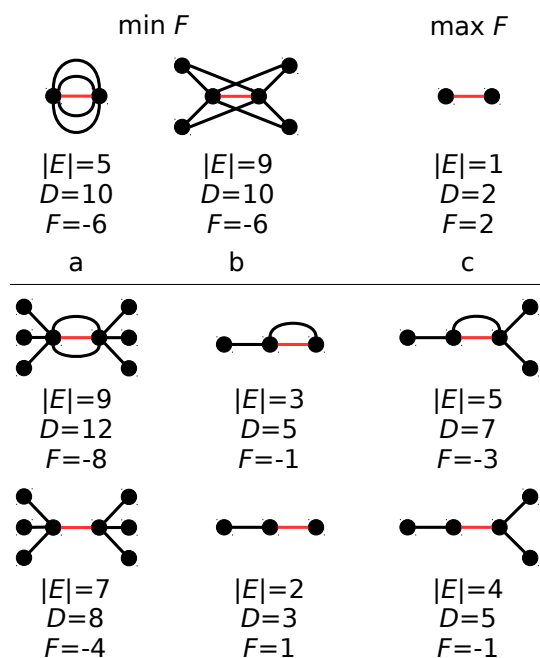
$$F(e) = 4 - D \quad (4.5)$$

As a multigraph may have repeated edges, whose number is independent of the number of vertices, the bounds for  $F(e)$  shall be expressed as a function of the known number of edges, namely,  $|E|$ . Therefore,  $2(2 - |E|) \leq F(e) \leq 2$ . The lower bound is attained when  $d_k = |E|$  for every  $k \in e$ , therefore  $D = 2|E|$  (Figure 16a). In turn  $F(e) = 2$ , for an isolated edge  $e$  (Figure 16c). In contrast to the multigraph case, for simple unweighted graphs, the lower bound can be expressed as a function of the number of vertices:  $2(3 - |V|) \leq F(e)$ , which is obtained for  $d_k = |V| - 1$  for every  $k \in e$ , i.e.,  $D = 2(|V| - 1)$  (Figure 16b). As for multigraphs,  $F(e)$  reaches its maximum value ( $F(e) = 2$ ) for an isolated edge (Figure 16c).

As shown in Figure 16, Forman-Ricci curvature quantifies the degree of spread of the vertices in  $e$ , from maximum spread (corresponding to  $\min F(e)$ ) to minimum spread (attained when  $\max F(e)$ ).

#### 4.3.2 Directed graphs

Here we are interested in an unweighted directed multigraph  $G = (V, E)$ , where  $e = (i, j) \in E$  is an *arc* (directed edge), and  $i, j \in V$ . Equation 4.4 indicates that the curvature of an edge depends on the degree of its vertices. As in a simple directed graph the degree can be split into indegree and outdegree, the curvature of  $e = (i, j)$  is defined in terms of indegree and outdegree as well [141]. There are different possibilities for the realization of the curvature, depending on the meaning one assigns to it. Here we emphasize the directed spread or *flow through*  $e$ , i.e., following the direction of the arc. Therefore, we consider the incoming arcs on  $i$  (indegree of  $i$ ,  $\text{in}(i)$ ) and the outgoing arcs from  $j$  (outdegree of  $j$ ,  $\text{out}(j)$ ).



**Figure 16:** Forman-Ricci curvatures  $F(e)$  calculated for the red edge  $e$  of undirected graphs.

The definition of the curvature in (4.4) can be split into separate contributions  $2 - d_i$  and  $2 - d_j$  for  $i$  and  $j$ , respectively. Accounting for the fact that the directed edge  $e$  does not contribute to the indegree of  $i$  or the outdegree of  $j$ , we define the curvature contributions for the in-flow at  $i$  ( $F(\rightarrow e)$ ) and for the out-flow at  $j$  ( $F(e \rightarrow)$ ), respectively, as

$$\begin{aligned} F(\rightarrow e) &= 1 - \text{in}(i) \\ F(e \rightarrow) &= 1 - \text{out}(j) \end{aligned} \quad (4.6)$$

Both are bounded below by  $2 - |E|$  for  $\text{in}(i) = \text{out}(j) = |E| - 1$ , and bounded above by 1 when  $\text{in}(i) = \text{out}(j) = 0$  (Figure 17a). For the simple directed graph, the lower bound for both, in- and out-flow, is  $2 - |V|$ , for  $\text{in}(i) = \text{out}(j) = |V| - 1$  (Figure 17b). The upper bound is reached, in both cases, when  $\text{in}(i) = \text{out}(j) = 0$  (Figure 17c). The curvature accounting for the flow through  $e = (i, j)$  is then given by

$$F(\rightarrow e \rightarrow) = F(\rightarrow e) + F(e \rightarrow) = 2 - \text{in}(i) - \text{out}(j) \quad (4.7)$$

where  $2(2 - |E|) \leq F(\rightarrow e \rightarrow) \leq 2$  for the multigraph case and  $2(2 - |V|) \leq F(\rightarrow e \rightarrow) \leq 2$  in the simple graph case. Figure 17c shows the case where  $F(\rightarrow e \rightarrow) = 2$ . Some further examples of calculations of curvatures  $F(\rightarrow e \rightarrow)$  are shown in Figure 17.

If the *flow-loss* along  $e$  is to be considered, two additional components are calculated that account

for the flow loss at  $i$  ( $F(\leftarrow e)$ ) and at  $j$  ( $F(e \leftarrow)$ ). Thus

$$\begin{aligned} F(\leftarrow e) &= 1 - \text{out}(i) \\ F(e \leftarrow) &= 1 - \text{in}(j) \end{aligned} \quad (4.8)$$

both bounded below by  $1 - |E|$ , for  $\text{out}(i) = \text{in}(j) = |E|$ , and bounded above by 0 for  $\text{out}(i) = \text{in}(j) = 1$  (Figure 17d). For the simple directed graph, we have  $2 - |V| \leq F(\leftarrow e) \leq 0$  and  $2 - |V| \leq F(e \leftarrow) \leq 0$ . Hence, the curvature for the flow-loss along  $e = (i, j)$  is

$$F(\leftarrow e \leftarrow) = F(\leftarrow e) + F(e \leftarrow) = 2 - \text{out}(i) - \text{in}(j) \quad (4.9)$$

where  $2(1 - |E|) \leq F(\leftarrow e \leftarrow) \leq 0$  (Figures 17a-e) holds in the multigraph case and  $2(2 - |V|) \leq F(\rightarrow e \rightarrow) \leq 0$  in the simple graph case. Some further examples are shown in Figure 17.

A curvature accounting for the *total flow over*  $e$  is then computed as

$$F(e) = F(\rightarrow e \rightarrow) + F(\leftarrow e \leftarrow) \quad (4.10)$$

In the following section, we extend the Forman-Ricci curvature to hypergraphs.

## 4.4 Forman-Ricci curvature of undirected hypergraphs

### 4.4.1 Definition and bounds

In the (unweighted) graph case, Forman-Ricci curvature is given by  $F(e) = 4 - d_i - d_j$ . An edge  $e$  is therefore flat, that is,  $F(e) = 0$ , if each of its endpoints stands in two relations; for instance, the curvature is zero for every edge of a cycle. Moreover, isolated edges are positively curved ( $F(e) > 0$ ). Also,  $e$  is negatively curved for edges containing nodes of a high degree. A generalization of the Forman-Ricci curvature ought to preserve these properties. In general, it has to quantify the difference between the hyperedge size and the number of connections of its nodes. In the hypergraph case, this translates into considering the trade-off  $\frac{w_k}{w_e} - \sum_{e_l \sim k} \frac{w_k}{\sqrt{w_e w_{e_l}}}$  in Equation 4.3 for all  $k \in e$ . That is, if  $e$  is a hyperedge (an arbitrarily large subset of vertices), then we shall compute

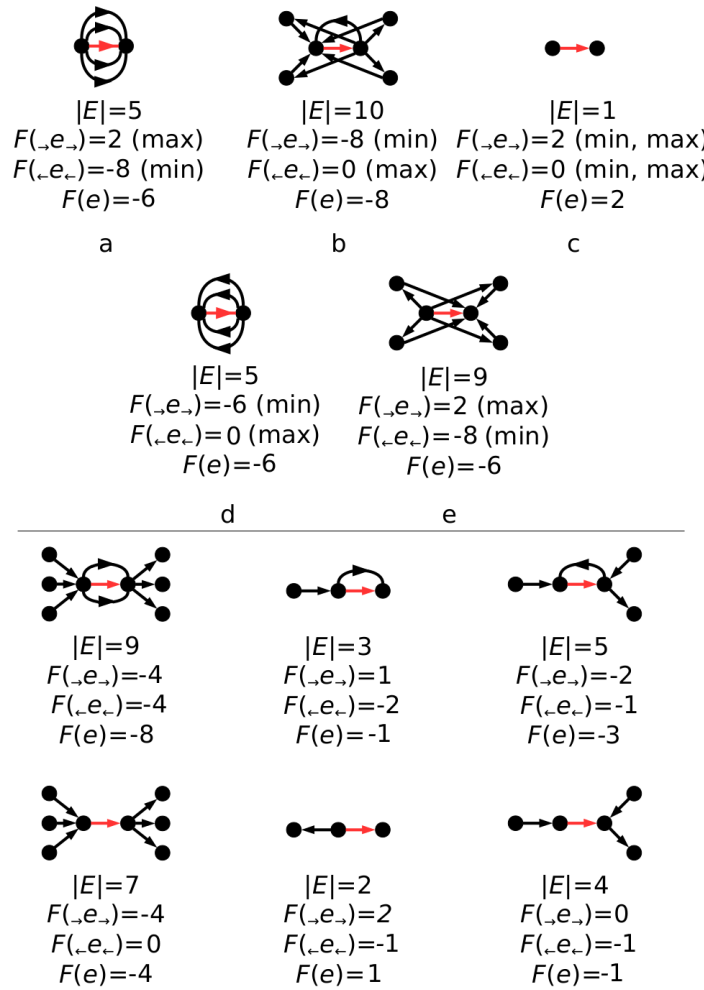
$$\sum_{k \in e} \left( \frac{w_k}{w_e} - \sum_{e_l \sim k} \frac{w_k}{\sqrt{w_e w_{e_l}}} \right).$$

Weighting this quantity by  $w_e$ , the weight of the hyperedge, we obtain the Forman-Ricci curvature of the hyperedge  $e$ :

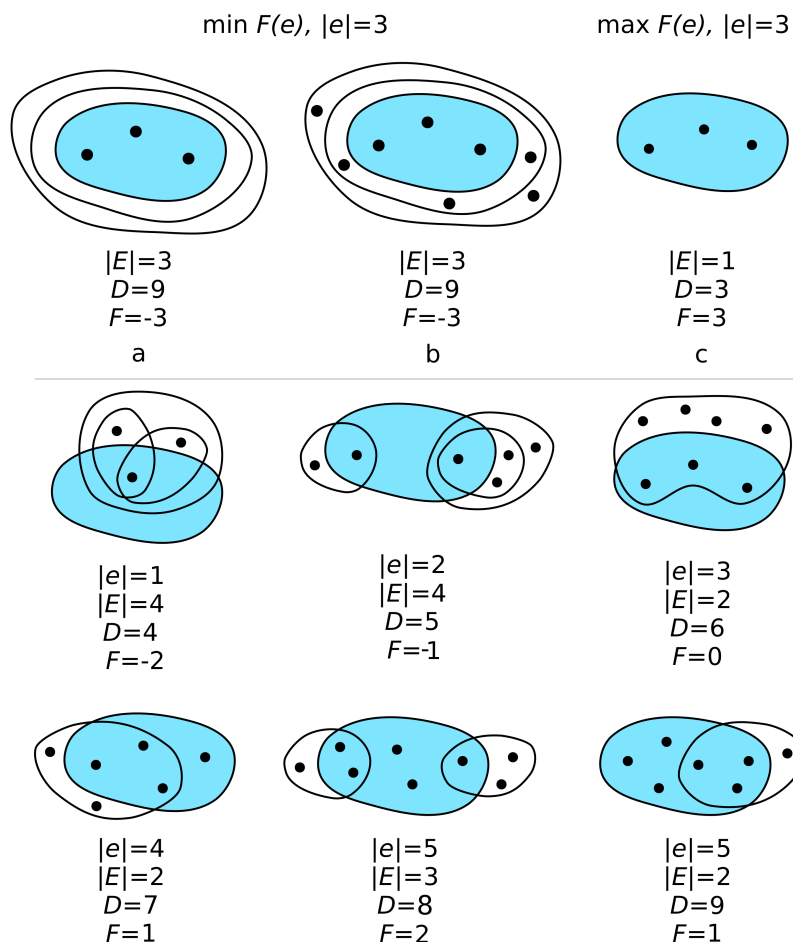
$$F(e) = w_e \left[ \sum_{k \in e} \left( \frac{w_k}{w_e} - \sum_{e_l \sim k} \frac{w_k}{\sqrt{w_e w_{e_l}}} \right) \right] \quad (4.11)$$

For the unweighted hypergraph, where all vertex weights are equal to 1, this expression simplifies to

$$F(e) = \sum_{k \in e} \left( 2 - d_k \right) = 2|e| - \sum_{k \in e} d_k = 2|e| - D \quad (4.12)$$



**Figure 17:** Forman-Ricci curvatures  $F(\rightarrow e \rightarrow)$ ,  $F(\leftarrow e \leftarrow)$ , and  $F(e)$  calculated for the red arc  $e$  of directed graphs.



**Figure 18:** Forman-Ricci curvatures  $F(e)$  calculated for the blue hyperedge  $e$  of hypergraphs.

which is bounded below by  $|e|(2 - |E|)$  when  $d_k = |E|$  for every  $k \in e$ , and bounded above by 1 when  $D = |e|$ . In other words, the minimum curvature occurs when every vertex in  $e$  belongs to each hyperedge (Figures 18a,b); the maximum is attained for an isolated hyperedge (Figure 18c).

For the particular case of simple hypergraphs, we, therefore, have the lower bound  $2|e|(1 - 2^{|V|-2})$  when  $d_k = |\mathcal{P}(V \setminus \{k\})|$  for every  $k \in e$ , and the upper bound  $|V|$ , when  $E = \{V\}$ . Note that in hypergraphs  $|e| \leq |V|$ , therefore the minimum value  $|e|$  may reach 1, unlike graphs. In such a case,  $2(1 - 2^{|V|-2}) \leq F(e) \leq |V|$ . Some further examples of curvature for hypergraphs are shown in Figure 18.

#### 4.4.2 Curvature of the Wikipedia voting network

Wikipedia is an encyclopedia written by volunteers. A small part of these users are administrators, who besides being active, regular long-term Wikipedia contributors, have gained the general trust of

the community and have taken on technical maintenance duties. A user becomes an administrator when a request for adminship is issued and the Wikipedia community via a public vote decides who to promote to administrator. Users can either submit their own requests for adminship or be nominated by other users. Using the January 3 2008 dump of Wikipedia page edit history [80], Leskovec *et al.* [81] extracted 2,794 elections (hyperedges in our setting) and 7,066 users (vertices) participating in the elections (either casting a vote or being voted on). We calculated the curvature for the resulting undirected hypergraph.

Figure 19 shows the distribution of hyperedge size (sometimes called "hyperedge degree") and of vertex degree. The data show that many of the elections involve a single user, although elections with 2-20 users are also common. There are few elections with more than 100 users, the largest one including 370 users (Figure 19a). The participation in elections is heavy-tailed distributed (Figure 19b), with most of the users participating in a single election and very few taking part in about a thousand elections. The curvature values are mostly negative (Figure 19c), indicating (i) the absence of elections with totally inexperienced users ( $\max F(e) \neq |e|$ ), i.e., all elections include at least a user that takes part in at least one other election; and (ii) for most elections the number of elections in which users take part is greater than their number of voting users ( $D > |e|$  in Equation 4.5). The minimum curvature value ( $-3,112$ ) is far from the lower bound ( $-19,728,272$ , calculated with  $|e| = 7,066$ ), since most users are experts in a limited number of fields only.

The magnitude of curvature could be understood in this context as the joint experience of the participants in a vote (hyperedge). A vote with few participants could have a high curvature if they are long-experienced voters. But according to Figures 19d)-f) the joint experience is large only when the group of voters is also large.

## 4.5 Curvature of directed hypergraphs

### 4.5.1 Definition and bounds

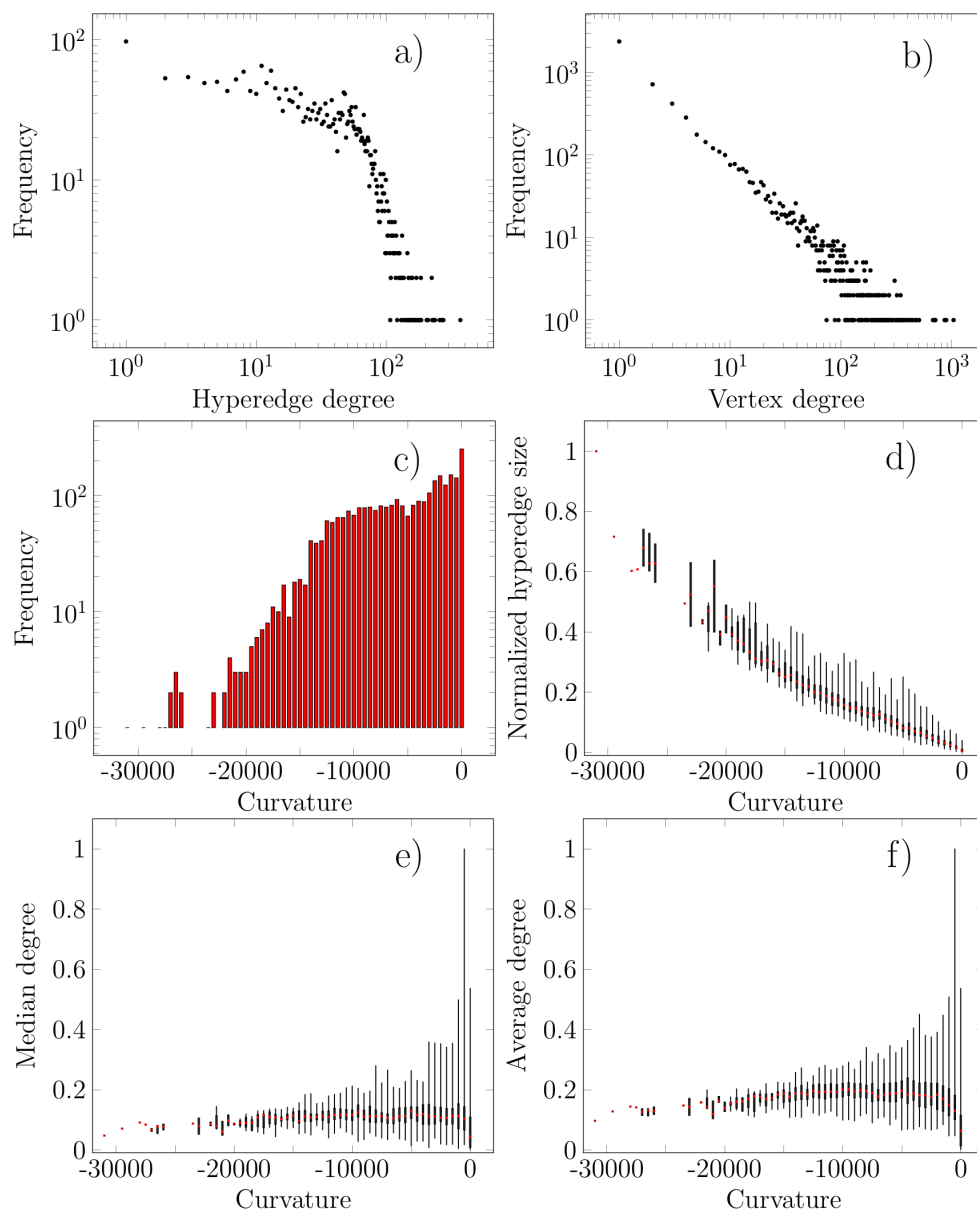
In a directed hypergraph, each hyperedge is composed of two subsets of vertices: the *tail* and the *head* of the hyperedge. Formally, we say that a *directed hypergraph*  $H$  is the couple  $(V, E)$  with  $V$  a set of vertices and  $E$  a multiset of hyperarcs. A *hyperarc* is a pair  $e = (e_i, e_j)$ , where  $e_i \subseteq V$  and  $e_j \subseteq V$  are called its *tail* and its *head*, respectively. Figure 20 depicts some examples of directed hypergraphs, where the sets  $e_i$  and  $e_j$  are highlighted.

Extending (Equation 4.6) to directed hyperedges, following the line of thought that lead to Equation 4.11 we introduce the components of the curvature  $F(\rightarrow e)$  and  $F(e \rightarrow)$  for a hyperarc as

$$\begin{aligned} F(\rightarrow e) &= |e_i| - \sum_{i \in e_i} \text{in}(i) \\ F(e \rightarrow) &= |e_j| - \sum_{j \in e_j} \text{out}(j) \end{aligned} \quad (4.13)$$

with bounds  $|e_i|(1 - |E|) \leq F(\rightarrow e) \leq |e_i|$  and  $|e_j|(1 - |E|) \leq F(e \rightarrow) \leq |e_j|$ . For the simple directed hypergraphs, we have  $|e_i|(1 - 2^{|V|-1}) \leq F(\rightarrow e) \leq |e_i|$  and  $|e_j|(1 - 2^{|V|-1}) \leq F(e \rightarrow) \leq |e_j|$ . With  $F(\rightarrow e)$  and  $F(e \rightarrow)$  at hand, we define the curvature for the flow through  $e = (e_i, e_j)$  as:

$$F(\rightarrow e \rightarrow) = F(\rightarrow e) + F(e \rightarrow) = |e_i| + |e_j| - \sum_{i \in e_i} \text{in}(i) - \sum_{j \in e_j} \text{out}(j) \quad (4.14)$$



**Figure 19:** Voting Wikipedia: Distribution of a) hyperedge size (size of elections) and b) vertex degree (participation of users in elections). c) Histogram of curvature values with bins of 500 units. Boxplots where red points are averages, and whiskers indicate minimum and maximum values of d) normalized hyperedge sizes, e) median, and f) average hyperedge degrees. For a given boxplot in d) - f), a single box, for instance, that at  $x = [-500, 0]$ , represents the distribution of values (e.g. of normalized hyperedge size) for hyperedges with curvature within  $[-500, 0]$ .

with bounds  $(1-|E|)(|e_i|+|e_j|) \leq F(\rightarrow e \rightarrow) \leq |e_i|+|e_j|$  in the general case and  $(1-2^{|V|})(|e_i|+|e_j|) \leq F(\rightarrow e \rightarrow) \leq |e_i|+|e_j|$  for the simple directed hypergraph (Figure 20). Note that if  $|e|$  is allowed to have its minimum value of 1, then  $|e_k| = 1$  and  $2(1-|E|) \leq F(\rightarrow e \rightarrow) \leq 2$ . Some examples of curvature values for directed hypergraphs are shown in Figure 20.

The respective flow-loss components are:

$$\begin{aligned} F(\leftarrow e) &= |e_i| - \sum_{i \in e_i} \text{out}(i) \\ F(e \leftarrow) &= |e_j| - \sum_{j \in e_j} \text{in}(j) \end{aligned} \quad (4.15)$$

with bounds  $|e_i|(1-|E|) \leq F(\leftarrow e) \leq 0$  and  $|e_j|(1-|E|) \leq F(e \leftarrow) \leq 0$  in the general case and  $|e_i|(1-2^{|V|}) \leq F(\leftarrow e) \leq 0$  and  $|e_j|(1-2^{|V|}) \leq F(e \leftarrow) \leq 0$  for the simple directed hypergraphs.

Equation 4.15 yields the flow-loss curvature

$$F(\leftarrow e \leftarrow) = F(\leftarrow e) + F(e \leftarrow) = |e_i| + |e_j| - \sum_{i \in e_i} \text{out}(i) - \sum_{j \in e_j} \text{in}(j) \quad (4.16)$$

with bounds  $(1-|E|)(|e_i|+|e_j|) \leq F(\leftarrow e \leftarrow) \leq 0$ , which becomes  $(1-2^{|V|-1})(|e_i|+|e_j|) \leq F(\leftarrow e \leftarrow) \leq 0$  for simple directed hypergraphs.

The Forman-Ricci curvature of directed hypergraphs is given by:

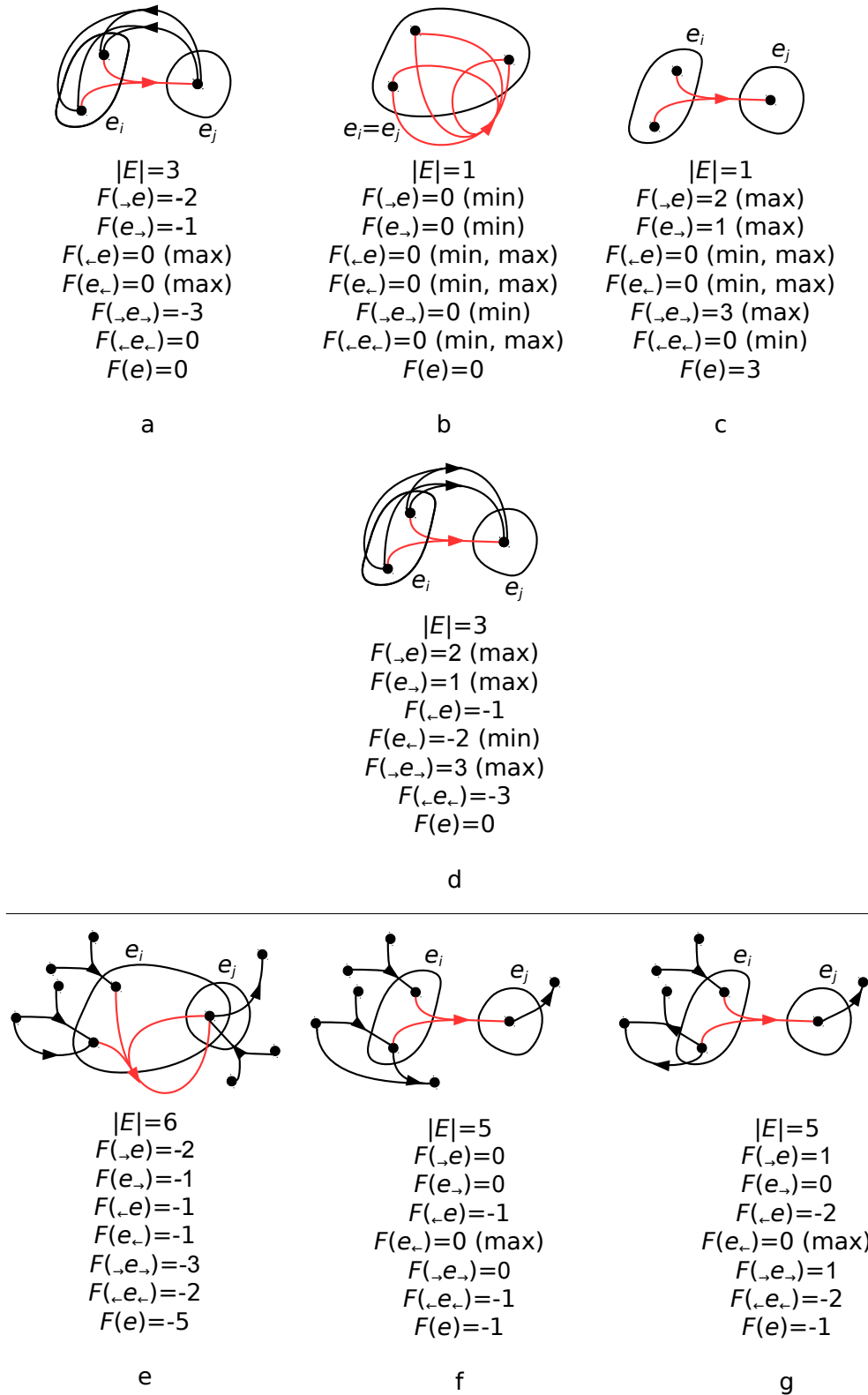
$$F(e) = F(\rightarrow e) + F(e \rightarrow) + F(\leftarrow e) + F(e \leftarrow) \quad (4.17)$$

#### 4.5.2 Comparison with Ollivier-Ricci curvature

The Ollivier-Ricci curvature of directed hypergraphs introduced by [32] follows a different idea of formalizing Ricci curvature for hypergraphs. There, to each directed hyperedge  $e = (e_i, e_j)$ , a probability measure  $\mu_{in}$  is assigned to a neighborhood of the tail (called masses and defined as the set  $\mathcal{M}$  of incoming neighbors and sources of  $e_i$ , the last are nodes in  $e_i$  without incoming neighbors), and a second probability measure  $\mu_{out}$  is assigned to a neighborhood of the head (called holes and defined as the set  $\mathcal{H}$  of outgoing neighbors and sinks of  $e_j$ , the last are nodes in  $e_j$  without outgoing neighbors). The Ollivier-Ricci curvature of  $e$  is then given by  $O(e) = 1 - W_1(\mu_{in}, \mu_{out})$ , where  $W_1(\mu_{in}, \mu_{out})$  is the 1-Wasserstein distance between the two probability measures  $\mu_{in}$  and  $\mu_{out}$ . As shown in Eidi and Jost [32], it can also be written as  $O(e) = \mu_0 - \mu_2 - 2\mu_3$ , where  $\mu_i$  is the amount of mass moved at distance  $i$  from a vertex in the set of masses to a vertex in the set of holes, in an optimal transport plan attained when computing  $W_1(\mu_{in}, \mu_{out})$ . Since  $\sum_i \mu_i = 1$ , then we have the bounds  $-2 \leq O(e) \leq 1$  [78].

Unlike  $O(e)$ , which is bounded by two constants, the different components of the Forman-Ricci curvature depend on the size of the hyperedge and the number of its incoming and outgoing arrows. In an isolated hyperedge, for instance,  $F(\rightarrow e \rightarrow)$  is equal to the size of the hyperedge, which can be an arbitrarily large integer, while  $O(e) = 0$ , since every mass is at distance 1 from any hole, i.e.,  $\mu_1 = 1$ . Curvature  $O(e) = 1$  is attained if a small hyperedge  $e$  has a large number of incoming arrows to  $e_i$  and a large number of outgoing arrows from  $e_j$ , but the set of masses is equal to the set of holes. Such a hyperedge, flat according to  $O(e)$ , is very negatively curved according to  $F(\rightarrow e \rightarrow)$  [78].

In general, since the two curvatures capture different aspects of the local connectivity of hyperedges, the sign of their values can be the same or opposite for the same hyperedge. In figure 22 we discuss



**Figure 20:** Forman-Ricci curvatures  $F(\rightarrow e\rightarrow)$ ,  $F(\leftarrow e\leftarrow)$ , and  $F(e)$  calculated for the red hyperarc  $e$ , connecting vertices in  $e_i$  with those in  $e_j$ , of hypergraphs.

signs and values of  $F(\rightarrow e \rightarrow)$  and  $O$  for the red hyperedges of nine directed hypergraphs, based upon the connections of their tails and heads. From left to right we can detect changes in the sign of  $O$  while the sign of  $F(\rightarrow e \rightarrow)$  is fixed. On the other hand, when we move vertically in the plot, the  $F(\rightarrow e \rightarrow)$  sign changes while the  $O$  sign is fixed. In the diagonal, directed hyperedges have the same sign for both curvatures. In the first column, each red hyperedge  $e = (e_i, e_j)$  has  $O(e) > 0$ . The reason is that the set of masses and the set of holes associated with  $e$  are equal, namely  $\mathcal{M} = \mathcal{H}$ . This implies that the distance between any mass and any hole is zero. Therefore,  $m_0 = 1$  while  $m_2 = m_3 = 0$ . Thus, since  $O(e) = \mu_0 - \mu_2 - 2\mu_3$ , then  $O(e) = 1$ . In contrast,  $F$  decreases when we move vertically in the same column. In particular,  $F(e)$  is positive for the uppermost hypergraph because the size of the tail of  $e$  is greater than the sum of indegree values of the vertices in  $e_i$ , namely,  $|e_i| - \sum_{i \in e_i} \text{in-deg}(i) = 1$  and, at the same time, the sole vertex in the head of  $e$  has only one outgoing arrow, then  $|e_j| - \sum_{j \in e_j} \text{out-deg}(j) = 0$ . When moving down in the same column,  $\sum_{i \in e_i} \text{in-deg}(i)$  and  $\sum_{j \in e_j} \text{out-deg}(j)$  increase by one in the second and third hypergraphs, respectively (with values  $F(e) = 0$  and  $F(e) = -1$ ) [73].

The arguments above explain the signs of  $F$  in the second and third columns. On the other hand,  $O(e)$  values decrease from left to right in rows, due to the increase in the distance between each mass and each hole of  $e$ . Distances in the second column increase by 1 compared to the first column, which means that  $m_0 = m_2 = m_3 = 0$  and  $O(e) = 0$ . Similarly, for any hypergraph in the third column, the distance from masses to holes becomes 3, and therefore,  $m_0 = m_2 = 0$ ,  $m_3 = 1$  and  $O(e) = -2$  [73].

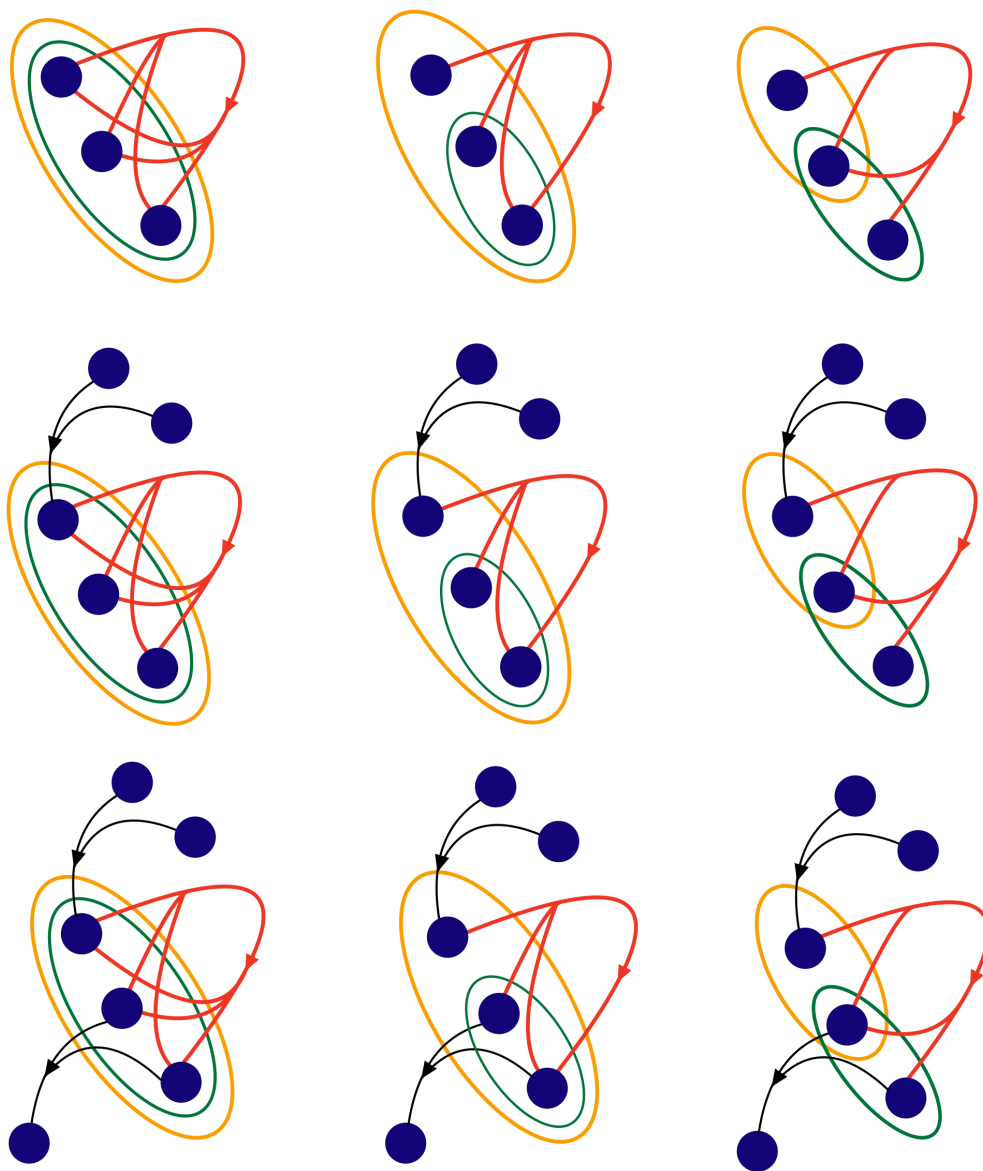
Forman-Ricci curvature detects the difference of the flow in the direction of the hyperedge under consideration and its size. On the other hand, Ollivier curvature informs about the existence of shorter alternative paths from incoming to outgoing neighbors of a given hyperedge, and in particular, measures the overlap of these two sets [73].

### 4.5.3 Hyperloops and their curvature

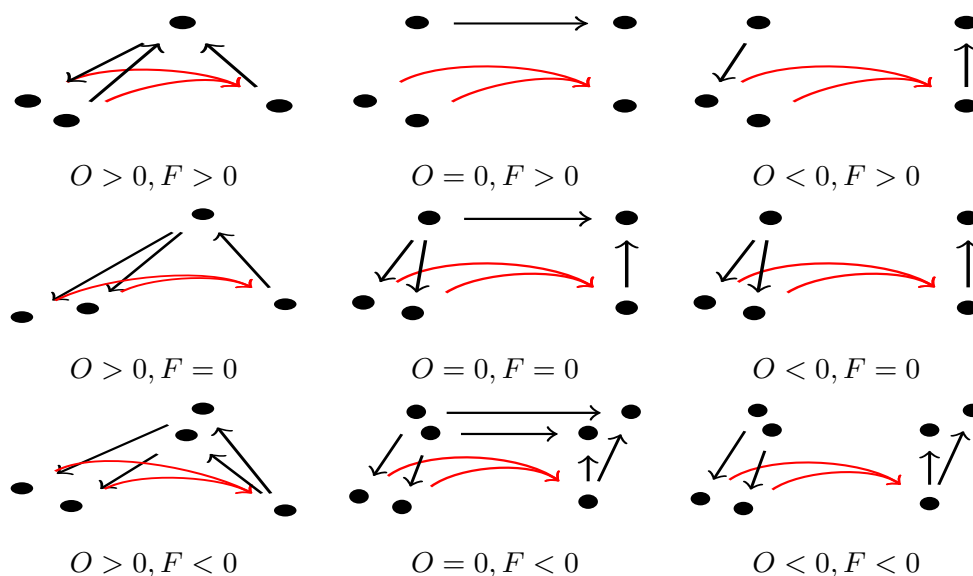
The definition of a loop in a hypergraph depends on the definition of a path. The strictest version requires that the head and tail of the hyperedge  $e = (e_i, e_j)$  coincide, i.e.  $e_i = e_j$ . If, moreover, the hyperedge is isolated, then  $F(e) = 0$  while  $O(e) = 1$  (see the top right hyperedge of Figure 21). If the hyperedge is not isolated, both curvature notions might change. As an example, if we move down in column one of Figure 21, we see  $F(e)$  and  $O(e)$  decrease, as the result of the incoming neighbors added to the tail of  $e$  and the outgoing neighbors added to its head.

A second, more flexible scenario, requires one of the following two cases to occur,  $e_i \subset e_j$  or  $e_j \subset e_i$ . In this case, if the hyperedge is isolated, then  $F(e) > 0$  and  $O(e) < 1$ . Again, both might change if the nodes of  $e$  have further connections. For instance, if we move down in column two of Figure 21,  $F(e)$  and  $O(e)$  decrease due to the addition of hyperedges incident to  $e$ .

Finally, the most flexible version of a hyperloop simply requires the tail and head of  $e$  to have non-empty intersection, namely,  $e_i \cap e_j \neq \emptyset$ . Naturally, any hyperedge of the two types described before represents a particular case of this version. All hyperloops in Figure 21 are instances of this definition, showing that they can be flat, positively or negatively curved, for both  $F$  and  $O$ .



**Figure 21:** Hyperloops and their curvature. The following are the values from left to right in rows. First row:  $F(e) = 0, 1, 2$ , while  $O(e) = 1.0, 0.4444, 0.25$ . Second row:  $F(e) = -1, 0, 1$ , and  $O(e) = 0.7222, 0.1111, -0.25$ . Third row:  $F(e) = -3, -2, -1$ , and  $O(e) = 0.2777, -0.3888, -1.0$ .



**Figure 22:** Local structure of directed hypergraphs with positive, negative, and zero values for both Ricci curvatures. For the given red directed hyperedge  $e$ ,  $O(e)$  and  $F(\rightarrow e \rightarrow)$  correspond to Ollivier and Forman curvatures respectively. From left to right we can detect changes in the signs of  $O(e)$  while the sign of  $F(\rightarrow e \rightarrow)$  is fixed. On the other hand, when we move vertically in the plot, Forman's sign changes while Ollivier's sign remains fixed. In the diagonal, directed hyperedges have the same sign for both curvatures.

#### 4.5.4 Curvature of metabolic networks

Metabolic networks have been extensively studied and thus offer an ideal setting to learn more about the Forman-Ricci curvature as a tool to elucidate the local geometry of empirical hypernetworks<sup>3</sup>.

This section is divided into two parts. First, we illustrate how to use the different components of the Forman-Ricci curvature for a deep structural analysis of a metabolic network. In the second part, we argue that these components carry information to evaluate the assortativity of directed hypergraphs, taking three metabolic networks as a case study.

##### Structure of metabolic networks

The metabolism of *Escherichia coli* is one of the most studied and best characterized among bacteria. Here we model the metabolism K-12 (iJR904 GSM/GPR) [118] of this bacterium as a directed hypergraph whose vertices are the metabolites (chemical species). Each chemical reaction is represented as a hyperarc  $e$ , whose educts (starting materials) correspond to  $e_i$  and products to  $e_j$ . There are  $|V| = 625$  metabolites and  $|E| = 1,176$  reactions accounting for 686 non-reversible and 245 reversible ones. These latter reactions, denoted by  $e_i \leftrightarrow e_j$  have been included as “forward”

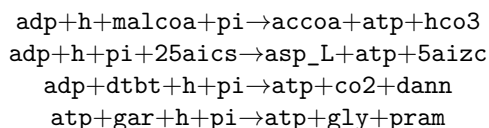
<sup>3</sup>The following references are great sources for readers interested in applications of Forman-Ricci curvature to systems modeled as graphs [86, 127, 129, 130, 141–143, 154, 155]

( $e_i \rightarrow e_j$ ) and “backward” ( $e_j \rightarrow e_i$ ) reactions. Extended data on the calculations of the curvature and its components (Equations 4.13 to 4.16) can be found in the Supplemental Material of [78].

As expected for chemical reactions, typically there are no more than three educts and three products (Figure 23a). The curvature values, therefore, vary little in response to hyperarc size, but rather depend more on the degree of vertices in  $e_i$  and  $e_j$ . Note that these degrees result, respectively, from the summation over vertex degrees of educts and of products (Equations 4.13 to 4.16). The distributions of educt and product sizes and their degrees are shown in Figure 23. The participation of educts and products in reactions does not yield a smooth distribution, as indicated by the gaps present in Figure 23b,c. The production of educts (Figure 23b) shows a large group of reactions whose educts are synthesized by less than 200 reactions and another group where they are obtained by more than 450 reactions. Likewise, there are two groups of reactions with different levels of use of their products (Figure 23c); one group has reactions whose products are used in less than 100 reactions and another with more than 300 reactions taking their products as starting materials.

The synthesis of products and the use of educts (Figures 23d and e), shows also a discontinuous participation of substrates in reactions. There are two groups of reactions according to the number of reactions synthesizing their products: one with reactions whose products are obtained by less than 200 reactions and another by more than 450 reactions (Figure 23d). Likewise, there are various groups of reactions according to the use of their educts, from some which are seldom used to some others with about 170, 230, and more than 330 uses (Figure 23e).

The extent to which the educts of the reaction  $e$  are produced from other reactions is measured by  $F(\rightarrow e)$ . The more reactions lead to the educts of  $e$ , the more negative  $F(\rightarrow e)$  becomes (Figure 24a). The theoretical bounds of  $F(\rightarrow e)$ , assuming  $\max |e_i| = 625$  are  $-734,375 \leq F(\rightarrow e) \leq 625$ . However, more realistic bounds are  $-7,050 \leq F(\rightarrow e) \leq 6$ , which result from taking the actual  $\max |e_i| = 6$  (Figure 23a). We found that  $\min F(\rightarrow e) = -735$ , which is attained by four reactions, with four educts (all substrate abbreviations are included in Appendix ?? (Table 16)):



These reactions are those whose educts are the most synthesized of all the metabolic reactions of *E. coli* (63% of the reactions produce their educts). In three of them atp is synthesized from adp, which shows the well-known central metabolic role of atp [59, 151].

$\max F(\rightarrow e) = 1$  corresponds to a single reaction:  $\text{cyan} + \text{tsul} \rightarrow \text{h} + \text{so3} + \text{tcynt}$ , where only one of its two educts is a product of a single reaction:

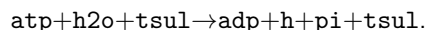
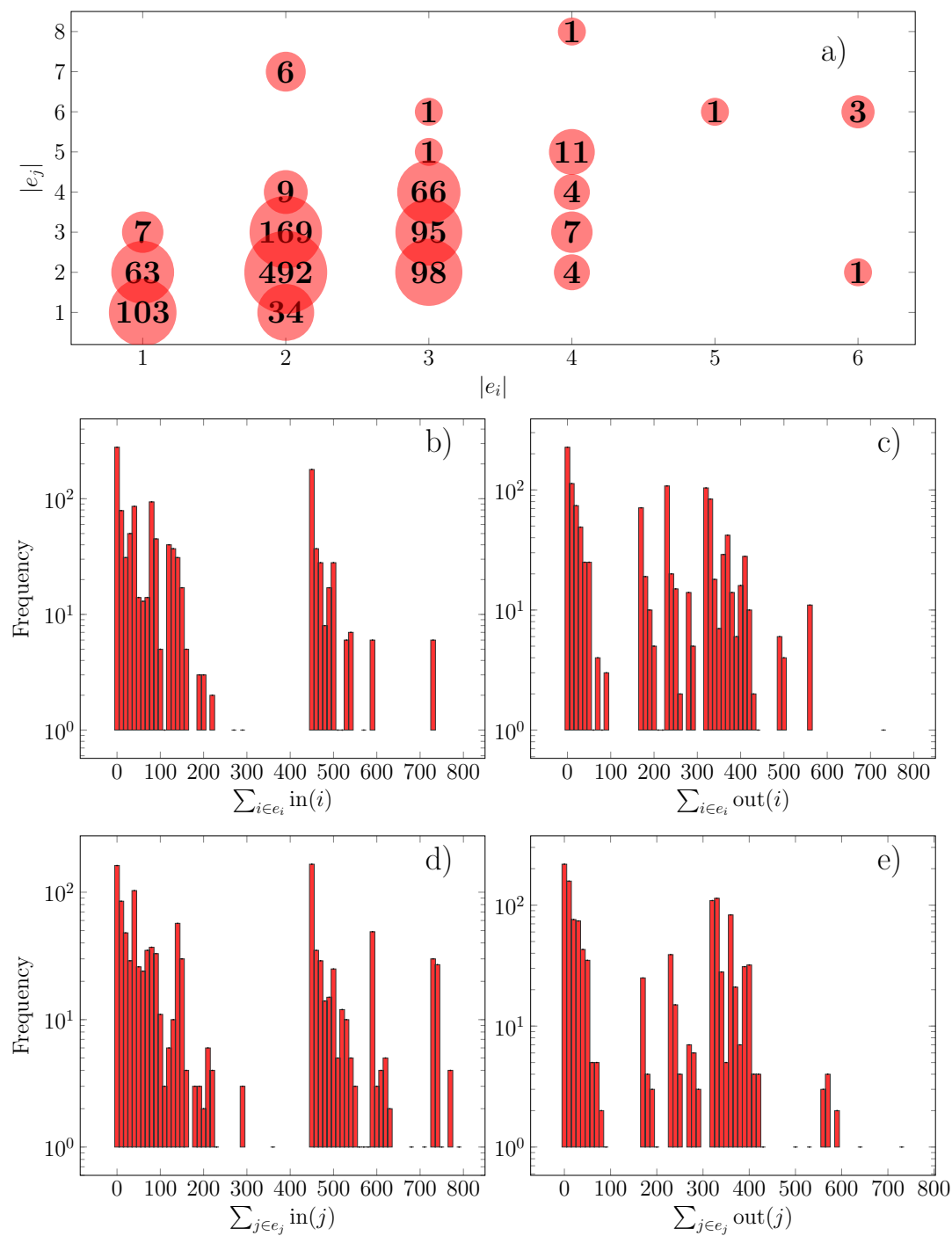
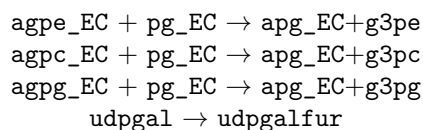


Figure 24a shows that the most frequent curvature value is 0 (for 73 reactions), i.e. 6% of the reactions have a trade-off between the number of educts and the number of reactions producing them; most of the remaining reactions have more ways to produce their educts than the number of educts. It is also found that there are almost no reactions with curvatures between -200 and -450, indicating that educts of reactions are mainly obtained either by less than 200 reactions (less than 17% of the reactions) or by 450 to 600 reactions (38 to 51% of the reactions). This is a consequence of the heavy-tailed indegree distribution of substrates [59].



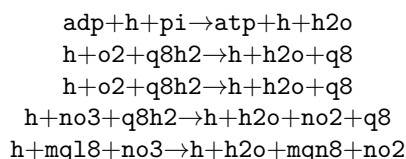
**Figure 23:** Metabolic network: a) Scatter plot of sizes of educts ( $|e_i|$ ) and products ( $|e_j|$ ), where circle radii correspond to  $\log f / \log 100$ , where  $f$  is the frequency of appearance of the couple  $(|e_i|, |e_j|)$  in reactions. Figures inside circles correspond to  $f$ . Distribution of b)  $\sum_{i \in e_i} \text{in}(i)$ , c)  $\sum_{i \in e_i} \text{out}(i)$ , d)  $\sum_{j \in e_j} \text{in}(j)$ , and e)  $\sum_{j \in e_j} \text{out}(j)$ .

Figure 24b shows the curvature values  $F(e_{\rightarrow})$ , which quantify the extent to which products of reactions are used in further reactions as educts. By taking  $\max |e_j| = 8$  (Figure 23a) this curvature takes values  $-9,400 \leq F(e_{\rightarrow}) \leq 8$ . The actual  $\min F(e_{\rightarrow}) = -729$ , for  $\text{adp}+\text{h}+\text{pi} \rightarrow \text{atp}+\text{h}+\text{h}_2\text{o}$ , i.e., this is the reaction whose three products are most used in other reactions as starting materials (used in 62% of the reactions). In contrast, there are four reactions with  $\max F(e_{\rightarrow}) = 1$ :



Hence, for those three reactions with two products, these substrates are only used in a further reaction as educts, while  $\text{udpgalfur}$  is not further used, i.e. it is a metabolic “dead-end” [118]. As most of the reactions (96%) have negative values of  $F(e_{\rightarrow})$ , this indicates the efficient use of reaction products [151], which can be divided into two regimes. For about half of the reactions, their products are used in no more than 9% of the reactions and about 40% of the reactions have products that are used in more than a quarter of the reactions. This is a consequence of the heavy-tailed distribution, this time, of the outdegrees of the substrates [59].

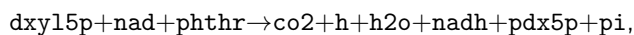
$F(\rightarrow e)$  showed that for most of the reactions their educts are produced by other reactions and  $F(e_{\rightarrow})$  that the products are used in several other reactions. The question that arises is whether those popular educts are connected through reactions with the popular products is positively answered by  $F(\rightarrow e_{\rightarrow})$ , which takes negative values for most of the reactions. The  $\min F(\rightarrow e_{\rightarrow}) = -1,463$  corresponds to  $\text{adp}+\text{h}+\text{pi} \rightarrow \text{atp}+\text{h}+\text{h}_2\text{o}$ . Hence, this is the reaction whose educts are most synthesized by other reactions and whose products are the most used as educts in other reactions. It is the bottleneck of the *E. coli* metabolism. Other reactions of this sort, with  $F(\rightarrow e_{\rightarrow}) < -1,000$  (Figure 24e), are:



Having analyzed the metabolism following the direction of educts to products in reactions, we now proceed to study the curvature in the backward direction, which quantifies to which extent a reaction is just one of the many connecting popular educts with popular products. We start by analyzing  $F(\leftarrow e)$  that shows to which extent educts of a reaction participate in other reactions. The theoretical bounds are  $-734,375 \leq F(\leftarrow e) \leq 0$  and we found that  $F(\leftarrow e)$  takes values in between  $-729$  and  $0$ ; the minimum is attained by  $\text{atp}+\text{h}+\text{h}_2\text{o} \rightarrow \text{adp}+\text{h}+\text{pi}$ , indicating that  $\text{atp}$  in an acidic aqueous medium is the most often used starting material.  $\max F(\leftarrow e)$  occurs for 51 reactions, whose involved 56 educts are only used in those 51 reactions, i.e. they are very specialized educts for very particular metabolic reactions. The distribution of  $F(\leftarrow e)$  values shows that for half of the reactions, their educts participate in less than 9% of the reactions, while for the rest, their educts take part in more than 15% of the reactions.

$F(e_{\leftarrow})$  shows to which extent products of a reaction are synthesized by other reactions. The theoretical bounds are given by  $\max |e_j| = 8$ , leading to  $-9,400 \leq F(e_{\leftarrow}) \leq 0$ . The actual values

range from -788 to 0. The minimum is reached by reaction:

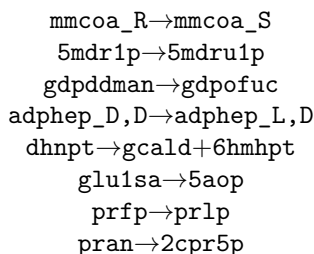


i.e. this set of products is the most synthesized by *E. coli* metabolism, which is expected, for the likelihood of a set of substances to be synthesized scales with the size of the set. This reaction with six products is one of the few where more than the frequent one to four products are synthesized (Figure 23a). Moreover, among the products, co2, h, h2o, nadh, and pi are often products of other reactions.

$\max F(e_{\leftarrow}) = 0$  is attained by 29 reactions, all of them leading to a single product, except for three reactions, each one with two products. Thus, those 32 products are of little synthetic relevance for the metabolism. The distribution of curvature values shows that there are three kinds of reactions whose products are synthesized by a different number of reactions. For 60% of the reactions, their products are synthesized by less than 200 reactions (17% of the reactions), and for the rest of the reactions by more than 450 reactions (38% of the reactions).

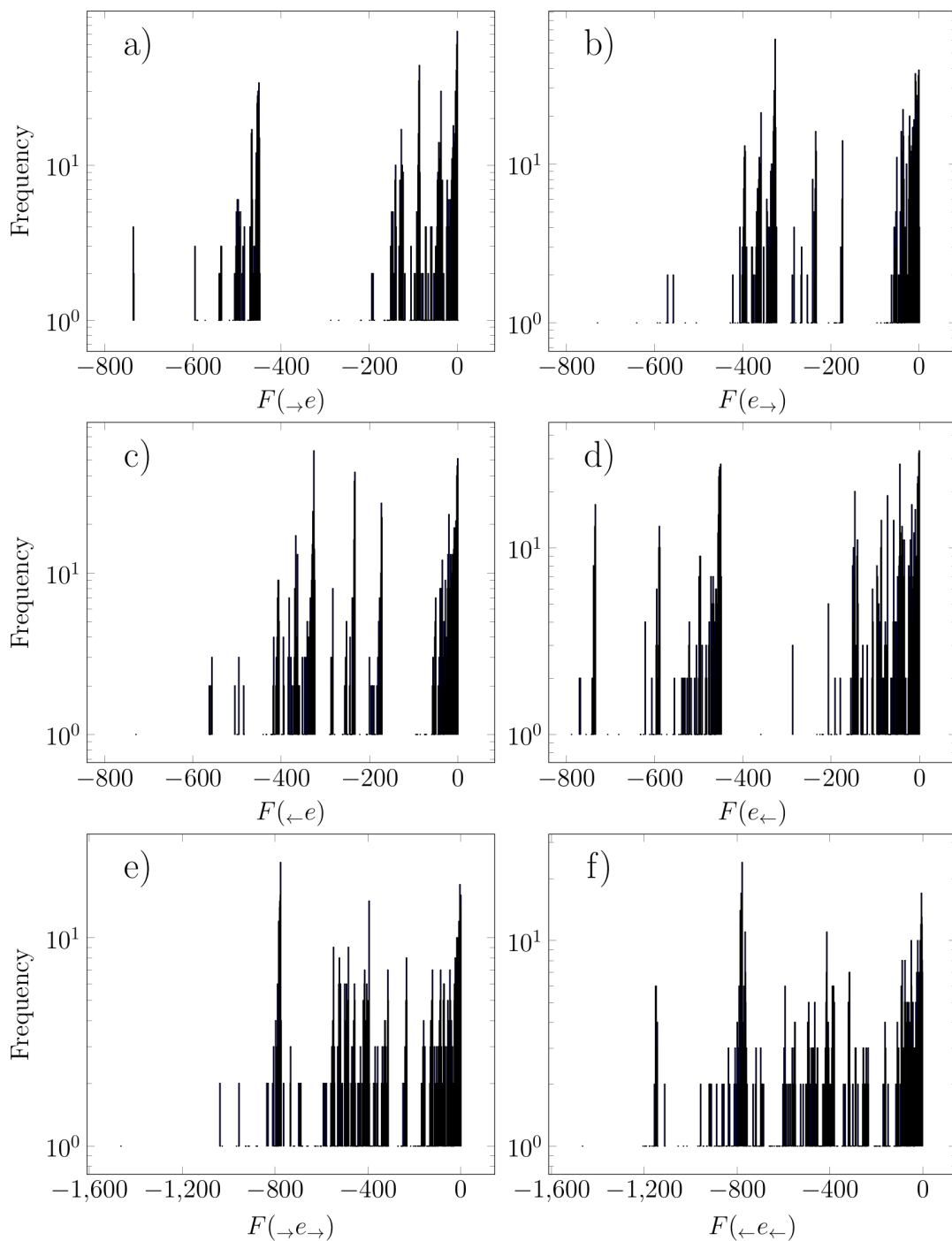
Curvatures  $F(\leftarrow e)$  and  $F(e_{\leftarrow})$  showed that half of the educts are often used and 40% of the products are often synthesized, which indicates that it is very likely to find alternative ways to link educts with products of existing reactions, as found in [7, 59, 151]. A measure of this degree of redundancy of a reaction or of its replaceability is given by  $F(\leftarrow e_{\leftarrow})$ , which indicates to which extent a reaction connects popular educts with popular products. The more negative the curvature, the more redundant or likely replaceable the reaction is.

By analyzing  $F(\leftarrow e_{\leftarrow})$  distribution (Figure 24f) we can see an ample spectrum of curvatures, with almost no gaps, indicating different degrees of redundancy for the metabolic reactions.  $\min F(\leftarrow e_{\leftarrow}) = -1,463$  corresponds to the hydrolysis of ATP, i.e.,  $\text{atp} + \text{h} + \text{h2o} \rightarrow \text{adp} + \text{h} + \text{pi}$ , indicating, e.g., that the dephosphorylation of atp to adp can be achieved by many other reactions (12% of the reactions).  $\max F(\leftarrow e_{\leftarrow}) = 0$  occurs for the following eight reactions, which are unique as they are the only way to connect their educts with their products:



### Forman-Ricci curvature and assortativity of directed hypergraphs

In network theory, assortativity is understood as the preference of nodes to link to others of similar degree. In undirected graphs, assortativity can be evaluated by computing the Pearson correlation coefficient,  $r$ , of the degree of nodes connected by an edge, i.e., correlation in the set  $\{(\deg(i), \deg(j))\}_{\{i,j\} \in E}$ . For directed graphs, it is possible to evaluate the correlation of indegree ( $r(in, in)$ ), outdegree ( $r(out, out)$ ), or any other combination (e.g.  $r(in, out)$ ). The question is how to assess assortativity in directed hypergraphs, where there might be more than one node at



**Figure 24:** Metabolic network: Histograms of curvature values for a)  $F(\rightarrow e)$ , b)  $F(e \rightarrow)$ , c)  $F(\leftarrow e)$ , d)  $F(e \leftarrow)$ , e)  $F(\rightarrow e \rightarrow)$ , and f)  $F(\leftarrow e \leftarrow)$ .

each end of a hyperedge,  $e = (e_i, e_j)$ . Suppose that we want to know if the number of incoming arrows to tails is correlated to the number of outgoing arrows from heads. This can be easily assessed by using Forman-Ricci curvature, since two of its components carry such information (normalized by the size of the tail and head, respectively), namely,  $F(\rightarrow e) = |e_i| - \sum_{i \in e_i} \text{in}(i)$ , and  $F(e \rightarrow) = |e_j| - \sum_{j \in e_j} \text{out}(j)$ . To illustrate this idea, let us consider three metabolic networks modeled as directed hypergraphs: *Escherichia coli*, *Mycobacterium tuberculosis* [58], and *Helicobacter pylori* [146].

Figures 25a, c, and e, show the distribution of  $(F(\rightarrow e), F(e \rightarrow))$  values for *Escherichia coli*, *Mycobacterium tuberculosis*, and *Helicobacter pylori*, respectively, where the size of a point is proportional to the logarithm of its frequency. Figures 25b, d, and f, represent the corresponding distributions of  $F(\rightarrow e) - F(e \rightarrow)$ <sup>4</sup> values as percentages. Figure 25 shows that none of the metabolic networks is assortative (disassortative), since the (normalized) number of incoming arrows to the tails and of outgoing arrows from the heads of the hyperedges are not correlated (anticorrelated). However, the three metabolisms show strong similarities. The biggest ball in Figures 25 a, c, and e, is at  $(0, 0)$ , accounting for 6.4%, 6.4%, and 6.1% of the metabolic reactions. The biggest humps in Figures 25b, d, and f, are around zero, accounting for more than 25% of the reactions in each network. These are sets of assortative reactions. Each network has four clusters in Figures 25a, c, and e, that determine four humps in Figures 8b, d, and f. There is a hump around 300 in Figures 25a and b (accounting for 8 and 10% of the reactions, respectively), and around 160 in Figures 25c (accounting for 9.3% of the reactions). A third hump, around  $-440$  in Figure 25a (7% of reactions), around  $-380$  in Figure 25 b (6% of reactions), and around  $-200$  in Figure 25c (8% of reactions). The two previous humps correspond to sets of disassortative reactions. Finally, the fourth hump appears around  $-120$ ,  $-100$ ,  $-40$ , in Figure 8b, d, and f, grouping 14.4%, 9.3% and 13.5% of the reactions, respectively.

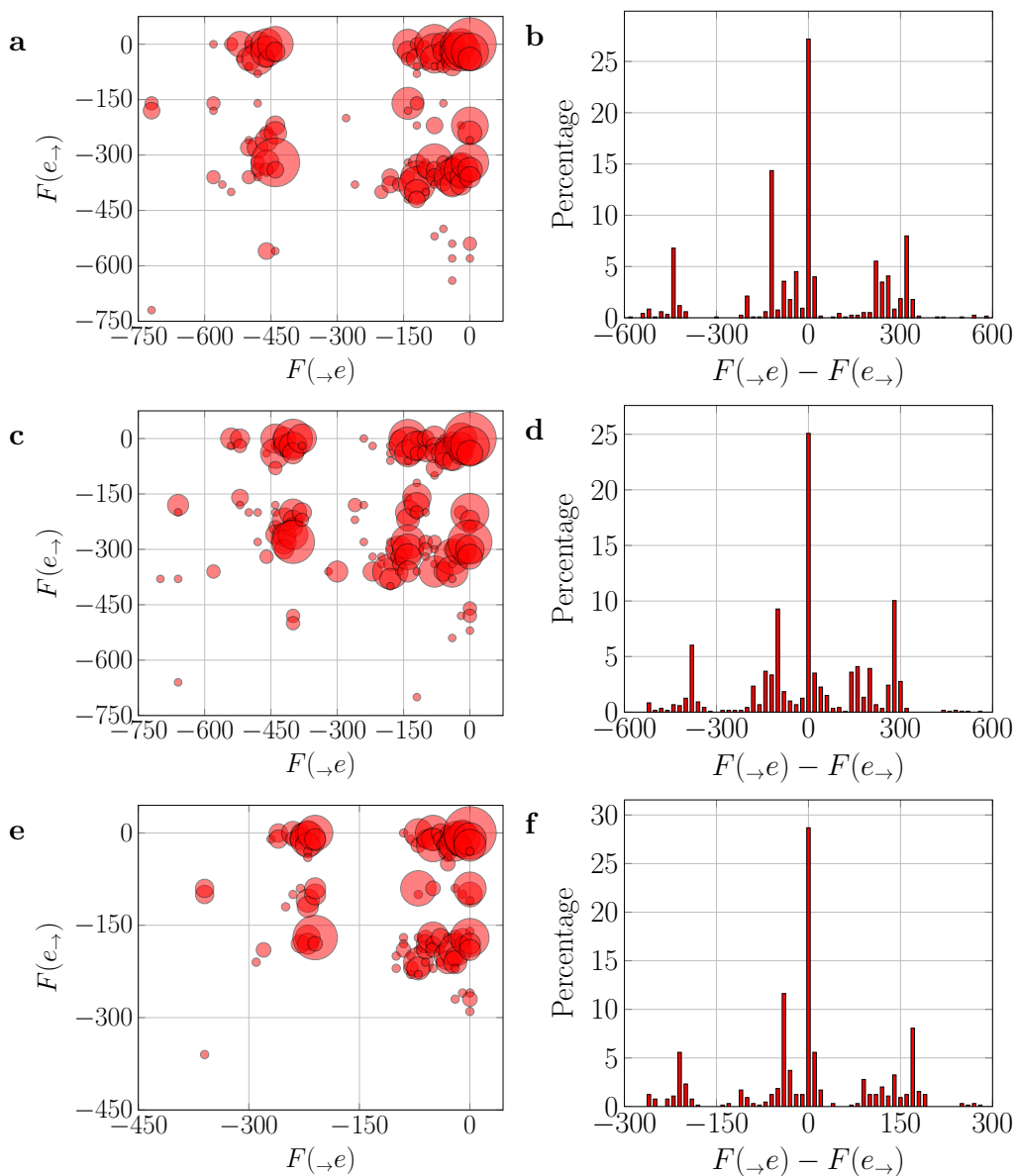
Other components of the Forman-Ricci curvature can be used to assess assortativity of all the possible combinations of degrees in a directed hypergraph.

## 4.6 Curvature of chemical space

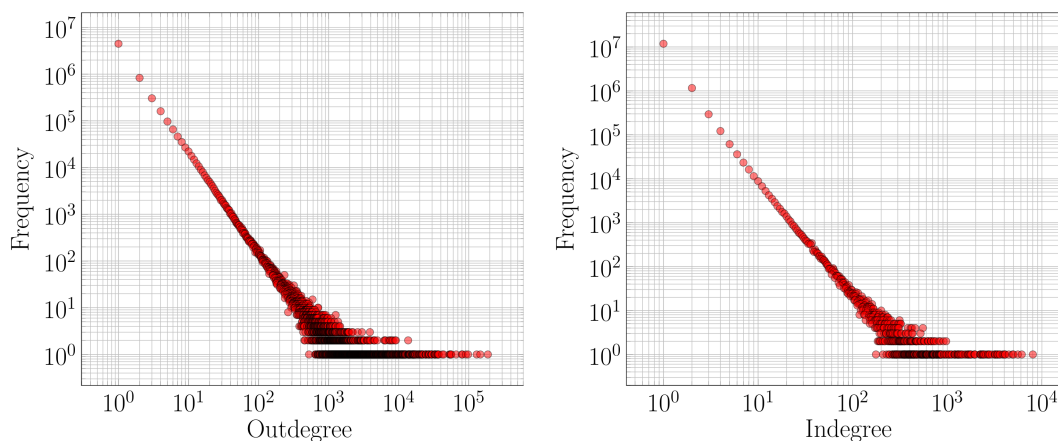
### 4.6.1 Indegree, outdegree, and the role of substances

We turn now to the distribution of outdegree and indegree. Outdegree of a substance is defined as the number of reactions that use it as a starting material. Indegree is the number of reactions that produce a substance. Figure 26 shows that both distributions are heavy-tailed, which means that most substances are used in a few reactions (97.4% are substrates in less than 5 reactions), similarly, most substances have been reported as products of few reactions (99.0% in less than 5 reactions), as shown by the top left part of the distributions. The tail in the indegree distribution reveals frequently reported products (the top-10 is shown in Table 11). The tail of the outdegree distribution indicates that some substances are starting materials in thousand reactions and a few in tens of thousands (the top-10 substrates are shown in Table 11). This pattern suggests that, to a large extent, chemists explore chemical space by reacting chemicals with standard reagents. I call

<sup>4</sup>The reader will note that the difference of these curvature components compares the number of incoming arrows at the tails and the number of outgoing arrows from the heads taking into account the size of the tails and heads. This is an important normalization in hypergraphs, where heads and tails can be arbitrarily large. A non-normalized version of the difference in degrees was introduced in [31], which is a generalization of the degree difference of graphs proposed in [38].



**Figure 25:** Using  $F(\rightarrow e) = |e_i| - \sum_{i \in e_i} \text{in}(i)$ , and  $F(e \rightarrow) = |e_j| - \sum_{j \in e_j} \text{out}(j)$  to assess assortativity in three metabolic networks. *Escherichia coli* (bin width of 20 units): Plot a represents the distribution of  $(F(\rightarrow e), F(e \rightarrow))$  values, where the size of a point is proportional to the logarithm of its frequency. Plot b, represents the distribution of  $F(\rightarrow e) - F(e \rightarrow)$  values as percentages. The second and third rows are the respective distributions for *Mycobacterium tuberculosis* (bin width of 20 units), and *Helicobacter pylori* (bin width of 10 units).



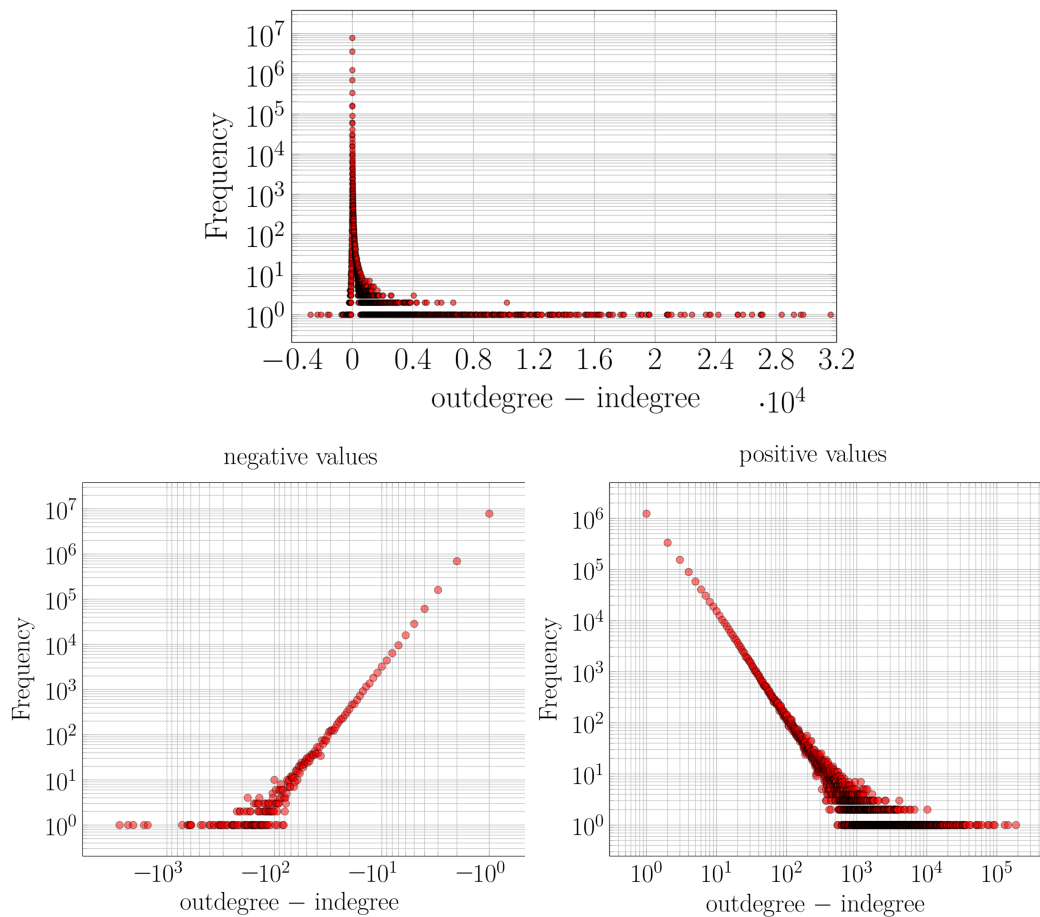
**Figure 26:** Outdegree (number of outgoing arrows) distribution of substances (left). Indegree (number of incoming arrows) distribution of substances (right).

this methodological feature of chemistry *the fixed substrate approach*.

The above discussion prompts the question of whether substances that act as prolific reactants (high outdegree) are also frequent products of reactions (high indegree). More generally, how correlated are the use and production of chemicals. This can be quantified by computing the difference between outdegree and indegree of each substance. The distribution of outdeg–indeg is presented in Figure 27 (top). The rightmost part of the positive tail was left out of this figure and presented in Table 17. Substances in Table 17 are the most prolific starting materials that are not frequent products of reactions, that is, the best representatives of the "precursors/reactants" role that do not perform the "targets/products" role equally well. Substances on the top of this list are acetic anhydride, methanol, and methyl iodide. Interestingly, these substances belong to the tool kit that chemists use for their *fixed substrate approach* to the exploration of chemical space. According to Figure 27, around 25% of substances have the same performance in the two roles (outdeg–indeg = 0). Moreover, most substances have an outdeg–indeg value close to zero: 97.8% hold  $-5 \leq \text{outdeg} - \text{indeg} \leq 5$ . The distribution of negative values of outdeg–indeg is presented in Figure 27 (bottom left), and accounts only for 14.7% of chemicals, which indicates that most substances have positive values of outdeg–indeg (60.6%) following the distribution shown in Figure 27 (bottom right). Values of outdeg–indeg increase (decrease), towards positive (negative) values, following a heavy-tailed distribution.

#### 4.6.2 Forman-Ricci curvature and assortativity of chemical reactions

We computed the Forman-Ricci curvature of chemical reactions according to Equation 4.17. The resulting distribution is shown in Figure 28. The most frequent curvature value is 1, accounting for 11.6% of reactions, followed by 0 (8.6%), and then by -1 (3.8%). So, nearly a quarter of the space is almost flat, in fact, 37.1% of the space has curvature values within  $-19 \leq F(e) \leq 1$  (see the list of



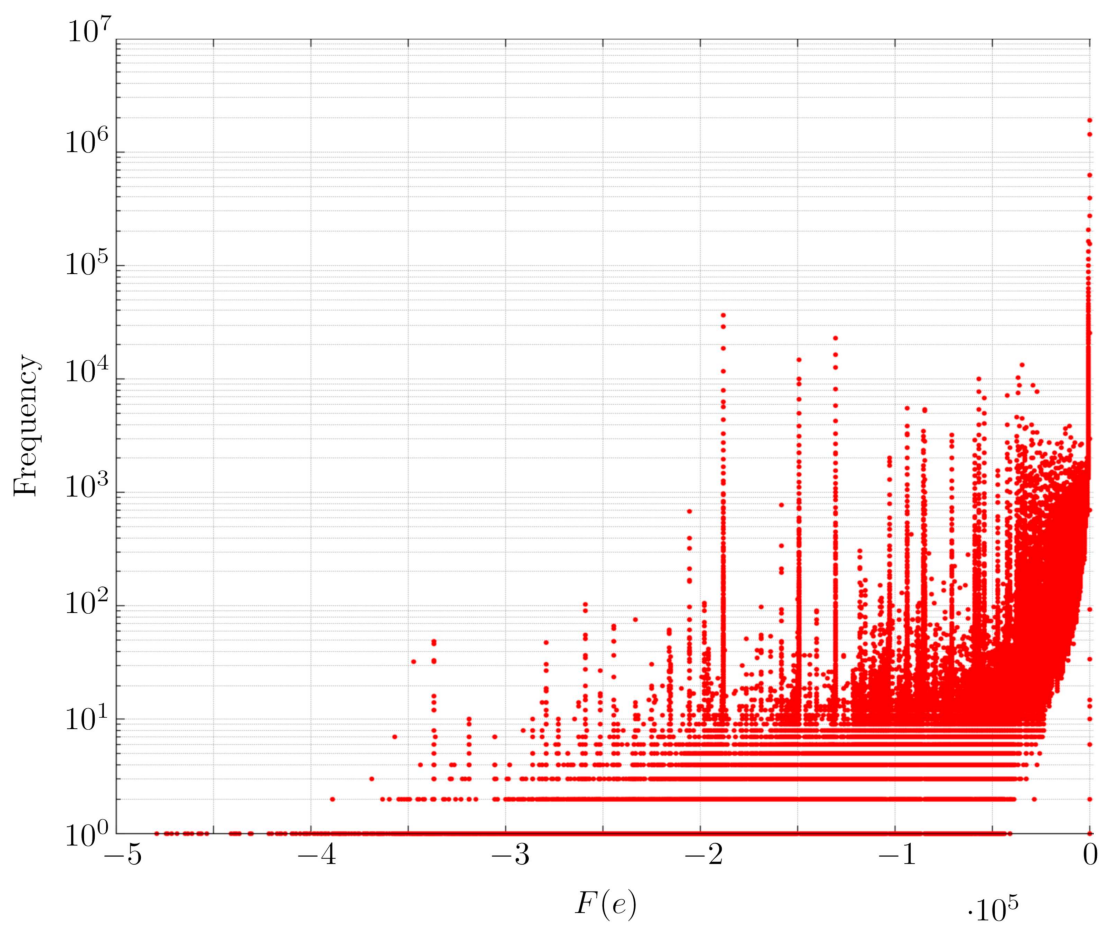
**Figure 27:** (top) Distribution of  $\text{outdeg} - \text{indeg}$  values using linear-log scales; some points of the positive tail were left out and presented in Table 17. (bottom left) The distribution of negative values of  $\text{outdeg} - \text{indeg}$  is presented in log-log scales; it was rendered by using first log-log scales on the collection of points  $(-(\text{outdeg} - \text{indeg})(v), \text{freq})$ , and then reversing the axis and relabeling the x-axis tick values. (bottom right) The distribution of positive values of  $\text{outdeg} - \text{indeg}$ .

Substance	outdegree	Substance	indegree
acetic anhydride	183480	carbon dioxide	8041
methanol	140632	benzoic acid	6338
benzaldehyde	129293	benzaldehyde	5929
methyl iodide	128409	ammonia	5031
benzyl bromide	100558	water	4926
water	85257	acetic acid	4462
benzoyl chloride	83912	methane	4051
formaldehyd	80456	ethene	4037
aniline	78365	biphenyl	3965
ethanol	72732	hydrogen (H <sub>2</sub> )	3962
di-tert-butyl dicarbonate	68030		

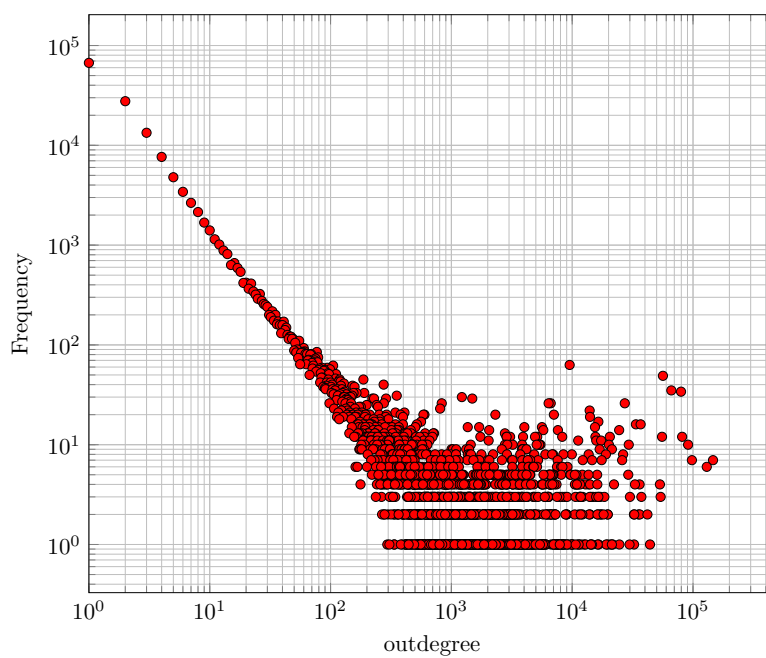
**Table 11:** Top-10 most popular reactants (outdegree) products (indegree).

more frequent values in Table 18). Interestingly, as we move to the left in Figure 28, we observe some prominent “humps”, each comprised of a series of points whose highest frequency is located at a certain curvature value, and the rest follow sliding to the left in descending order of frequency, like steps. These humps occur at curvature values near substances from the tool kit, for instance, the hump around -187966 is very close to the outdegree of acetic anhydride. This indicated that the hump is left by reactions using acetic anhydride as a substrate. The steps seem to indicate a pattern in the outdegree of substances reacting with acetic anhydride. Since 84% (Table 19) of reactions of acetic anhydride use it against exactly one other substance, we computed the outdegree distribution of those substances. This distribution explains the step-like pattern in the hump (Figure 29): the hump itself is a fingerprint of the chemistry of acetic anhydride; moreover, this substrate typically reacts with another substance (84%), and the outdegree of this substances follows (at least until outdegree around 200) a log-log decay. Similarly, other notable humps are due to the chemistry of methanol, methyl iodide, and other substrates in the toolkit, as well as to the outdegree pattern of their partner reactants. This raises the question of whether the pattern observed for the second substrate is due to its “age”, that is if the fact that  $\text{outdeg}(\text{acetic anhydride}) + \text{deg}(B) = 1$  is the most frequent value, followed by  $\text{outdeg}(\text{acetic anhydride}) + \text{deg}(B) = 2$ , etc., is a consequence of the exponential growth and the fixed-substrate approach: chemists use substances from the toolkit to characterise newly reported compounds, and this last collection is growing exponentially<sup>5</sup>. As we claimed before, the different components of curvature (Equation 4.17) carry information that can be used to quantify assortativity in directed hypergraphs. We have used  $F(\rightarrow e) - F(e \rightarrow)$  to quantify how aligned are the number of incoming arrows to tails and the number of outgoing arrows from heads, in other words, the number of available reactions yielding the substrates of  $e$  and the number of reactions using its products). Figure 30 shows that 12% of reactions are assortative and 40% have values neat to zero:  $-3 \leq F(\rightarrow e) - F(e \rightarrow) \leq 3$ . More details on reaction assortativity can be obtained from Figure 30.

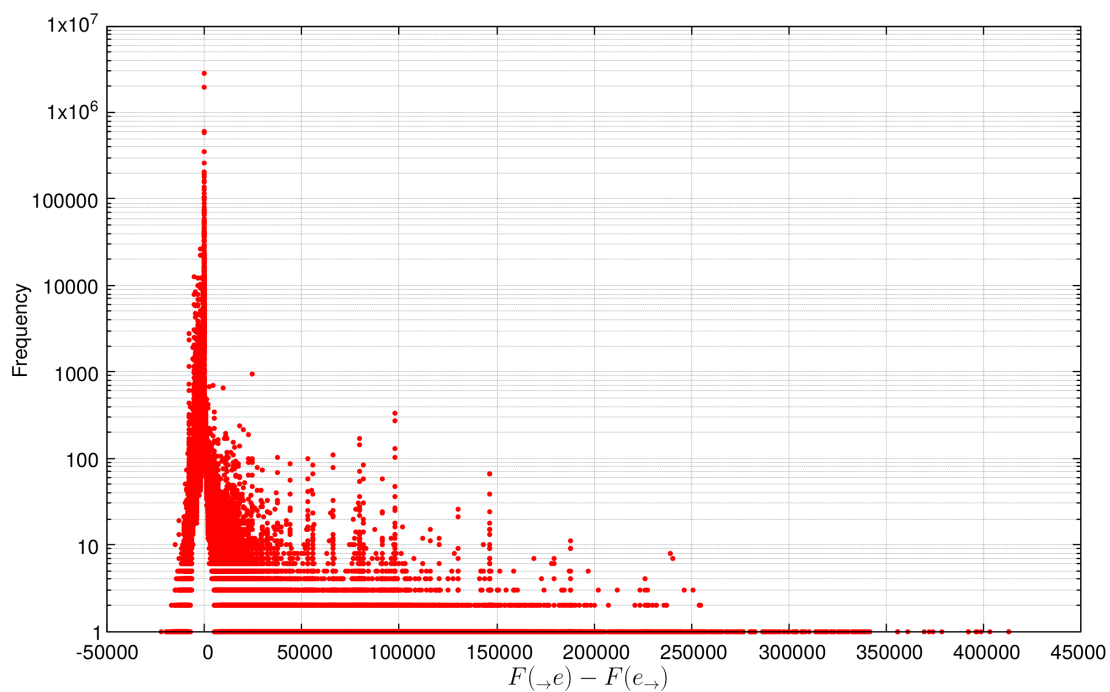
<sup>5</sup>Of course, we have to check the growth pattern of newly reported substances that are affine to the fix-substrate



**Figure 28:** Distribution of Forman-Ricci curvature for the chemical space. The curvature for a chemical reaction, i.e., for a directed hyperedge  $e$ , is given by  $F(e) = F(\rightarrow e) + F(e \rightarrow) + F(\leftarrow e) + F(e \leftarrow)$ , where  $F(\rightarrow e) = |e_i| - \sum_{i \in e_i} \text{in}(i)$ ;  $F(e \rightarrow) = |e_j| - \sum_{j \in e_j} \text{out}(j)$ ;  $F(\leftarrow e) = |e_i| - \sum_{i \in e_i} \text{out}(i)$ ; and  $F(e \leftarrow) = |e_j| - \sum_{j \in e_j} \text{in}(j)$ .



**Figure 29:** Outdegree distribution of substances participating in two-substrate reactions with acetic anhydride.



**Figure 30:** Distribution of  $F(\rightarrow e \rightarrow) - F(\leftarrow e \leftarrow)$  values. These two components of the curvature are used to assess the assortativity of the chemical space, where  $F(\rightarrow e) = |e_i| - \sum_{i \in e_i} \text{in}(i)$ , and  $F(e \rightarrow) = |e_j| - \sum_{j \in e_j} \text{out}(j)$ .

## CHAPTER 5

# A categorical model of chemical reaction networks

## Contents

5.1	Dialectica Petri nets . . . . .	<b>72</b>
5.1.1	Petri nets and their transitions . . . . .	73
5.2	Petri nets via Dialectica Categories . . . . .	<b>73</b>
5.3	The category $M_L\text{Set}$ and its structure . . . . .	<b>74</b>
5.3.1	Lineales . . . . .	75
5.3.2	The category $M_L\text{Set}$ . . . . .	75
5.4	A category of Petri nets . . . . .	<b>77</b>
5.5	Different lineales . . . . .	<b>79</b>
5.5.1	Original Dialectica $L = 2$ . . . . .	79
5.5.2	Kleene Dialectica $L = 3$ . . . . .	79
5.5.3	Multirelation Dialectica $L = \mathbb{N}$ . . . . .	80
5.5.4	Integers Dialectica $L = \mathbb{Z}$ . . . . .	81
5.5.5	Probabilistic Dialectica $L = [0, 1]$ . . . . .	82
5.5.6	Product of lineales . . . . .	82

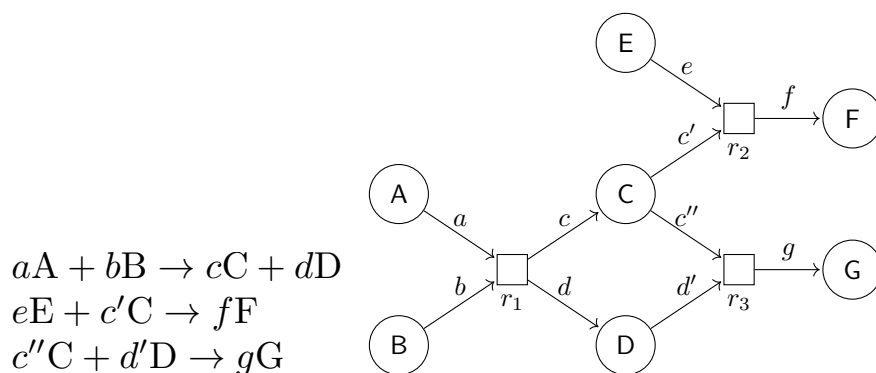
The network model of chemistry is inspired and motivated by the compositional nature of chemical reactions, which appear in several flavours. I wonder what categorical constructions underlie the structure of chemical reasoning and what are their formal connections. As the first step to answering this question is taken in the present chapter, where we investigate a categorical (compositional) network model of chemical reaction networks modeled as Petri nets (bipartite directed graphs) and its structure. Proofs for all propositions and the main theorem appear in Appendix D.

## 5.1 Dialectica Petri nets

Petri nets exert endless fascination over category theorists. Petri nets are only one of the many modeling languages for the description of distributed systems used by computer scientists, but they enjoy the distinction of being the one category theorists most write and talk about. Maybe category theorists see Petri nets as a gauntlet thrown at them because the definition of a morphism of Petri nets is not obvious and does lead to different categories. Maybe the bipartite graphs that usually depict Petri nets look too similar to automata ones, and these are the initial sources of good categorical examples in computing. In any case, many different categorical models of Petri nets do exist and some are fundamentally different from others. One fundamental difference is whether one concentrates in the token game and the behaviour of a given Petri net or on the graphs underlying different nets. Another difference is which possible combinators relating different Petri nets one considers. The aspect we concentrate on in this work is the kinds of labels that can be used in a Petri net.

A *Petri net* is simply a directed bipartite graph that has two types of nodes: *conditions* and *events* (also called places and transitions). Over this fixed structure of possible events and conditions, a causal dependency (or flow) relation between sets of events and conditions is described via pre- and post-relations, and it is this structure that determines the possible dynamic behaviour of the net. A transition in this causal dependency relation is enabled if all places connected to it as inputs contain at least one token. Finally given an initial position of tokens in a net (an initial marking) the "token game" (which we are not modeling) can be started and the system will evolve.

We explore the model originally introduced by Winskel [156, 157], but use it with morphisms, as in the work of Brown [21] and others, that relate Petri nets to constructors in Linear Logic [45].



**Figure 31:** Representation of chemical reaction data: as a list (a) and as a network (b). A, B, ..., H are substances and  $a, b, \dots, h$  are stoichiometric coefficients that indicate the proportion in which they combine.

### 5.1.1 Petri nets and their transitions

Networked systems are determined by their connections [30]. Perhaps the most basic type of relationship in any network is one that only allows us to express either presence or absence, that is, where the relationship connecting nodes uses the set  $\{0,1\}$  as a ruler or label set. In real-world applications, this is, though, not sufficient. In this section we explore frequent and rich applications of Petri nets, from chemical reaction networks to metabolic networks, searching for the kind of labels used on pre- and post-conditions.

*Chemical reaction networks.* Chemical combination is compositional in nature. Although data on substance reactivity are typically annotated as a list of chemical equations (see Figure 31), chemists reason on the network structure (see Figure 31) that emerges when the reactions are connected to make their concurrency explicit [137]. Synthesis planning is a prominent example where the synthesis of a substance results from composing reactions a sequence of reactions.

Directed hypergraphs and their enhancements, such as Petri nets, are used to model chemical reaction networks for they are models of concurrency of directed relations. At the level of abstraction described above, transitions of chemical reaction networks are discrete in nature, and pre- and post-conditions correspond to presence/absence or to stoichiometric coefficients, which can be modeled by the rulers  $\{0,1\}$  and  $\mathbb{N}$ , respectively.

*Metabolic networks.* These networks are comprised of the metabolic pathways (network of chemical reactions) and the gene interactions that regulate them. A key aspect of the former is kinetic modeling. There, Petri nets model reaction rates. For elementary reactions, which take place in a single step, the Law of Mass Action states that reaction rates are proportional to the concentration of reactants. Both quantities, rate of reactions, and concentration of reactants, are usually taken as positive real numbers; therefore, in this application, Petri nets are challenged to handle continuous tokens and transitions, which requires the ruler  $\mathbb{R}^+$ . On the other hand, gene interactions are handled by implementing genetic switches that are modeled by discrete transitions. A Petri net model for a metabolic network, therefore, needs two different rulers on the same net.

When applied to concrete metabolisms, a Petri net model will usually need to incorporate more than two rulers at the same time. For instance, [126] shows a hybrid Petri net representation of the gene regulatory network of *C. elegans* that is labeled with discrete and continuous transitions, but also with negative integers, real numbers, strings, and products of them.

Summarising, applications may need rulers such as  $\{0,1\}$ ,  $\{-1,0,-1\}$  (for data uncertainty, which is common in complex network systems [107]),  $\mathbb{N}$ ,  $\mathbb{R}^+$ ,  $\mathbb{Z}$ ,  $\mathbb{R}$ , strings, and their finite products. The ability to choose from a vast pool of rulers to label pre- and post-conditions is one of the strengths of the categorical construction presented in this chapter.

## 5.2 Petri nets via Dialectica Categories

Petri nets were described categorically in many works (e.g. [21], [98]) and are still been discussed [6], [90]. Models need to capture the practitioner's imagination and make themselves useful, both for calculations and for insights. Categorical models can be useful for both insights and calculations, but we have not seen categorical models that encompass different kinds of transitions in a single net.

Petri nets were modeled using Dialectica categories [113] previously, but the original Brown and Gurr model [20] worked only for *elementary nets*, that is nets whose transitions are marked with  $\{0,1\}$  for presence or absence of a relationship. An extension of this modeling to deal with integers  $\mathbb{N}$  was planned [22, 111], but never published. In this work, we put together different kinds of transitions, in

a single categorical framework. This way the categorical modeling applies to the many kinds of newer applications [33] that already use different kinds of labels on the transitions.

The original dialectica construction [113] was given in two different styles called the categories  $DC$  [112] and the categories  $GC$  [110]. For both constructions,  $C$  is a cartesian closed category with some other structure. The first style is connected to Gödel's Dialectica Interpretation hence the 'D' in  $DC$  for Dialectica. The second style called  $GC$  ([110]) is based on a suggestion of Girard's (hence the 'G') on how to simplify the first construction if one wants a model of Linear Logic. These two constructions are connected, via monoidal comonads as described in [113]. Here we are mostly interested in the construction called  $GC$ , whose morphisms are simpler. This construction can be explained, when the category  $C$  is  $\mathbf{Set}$ , using two 'apparently' different descriptions. This is because a relation in  $\mathbf{Set}$  between  $U$  and  $X$  can be thought of as either a subset of the product,  $\alpha \subseteq U \times X$ , or as a map into 2,  $\alpha: U \times X \rightarrow 2$ .

The work here uses only the second kind of description, defining general relation maps into algebraic structures called *lineales*. This is because changing the lineale where our relations take 'values', gives us the possibility of modeling several different kinds of processes. The original dialectica construction deals only with the Heyting algebra-like lineale 2. Here we discuss several other lineales and dialectica categories built over these different lineales.

### Related work

Our work fits in the vast landscape of categorical approaches to Petri nets building on [22, 111]. Meseguer's seminal work [98] focused on reachability properties of Petri nets, defining a category of all possible executions of a net. This work adopted the collective tokens philosophy. Its ideas were extended to the individual tokens philosophy in [23]. Other categorical models of Petri nets focus on obtaining nets by composing smaller nets along some boundaries. One of the first compositional models doing this was [60] where nets are composed along common places. In [116], nets are composed along common transitions and compositionality is used to study reachability properties of Petri nets. The work of [21], [111] and [22] concentrate on combining Petri nets via different monoidal products that give the category of Petri nets a linear logic structure. More recently, there have been numerous works building on the ideas of [98] and adopting the formalism of [60]. In [6] and [91], the authors focus on studying the categorical properties of reachability. In [66] a more fine-grained categorical model is proposed, that allows Kock to encompass the individual and collective token philosophies in the same framework. Finally, [5] constructs a unifying framework for [98], [23] and [66] extending [91].

Our work extends the approach of [21] to allow different kinds of arcs, e.g. inhibitor, probabilistic, partially defined, natural/integer numbers valued, and the coexistence of them in the same net.

## 5.3 The category $\mathbf{M}_L\mathbf{Set}$ and its structure

In order to define our category of Petri nets, we need to explain what kind of structure, that of a *lineale*, is required on the set of truth values that we use as codomain for pre- and post-condition relations.

### 5.3.1 Lineales

A lineale is a monoid together with a partial order compatible with the monoidal product, and such that every pair of elements has an internal hom. The monoidal product, the partial order, and the internal hom are used to define the adjunction that characterizes a lineale, as this is just a poset version of a symmetric monoidal closed category.

**Definition 2** (Partially ordered monoid). *A partially ordered monoid  $(L, \sqsubseteq, \otimes, e)$  is a monoid  $(L, \otimes, e)$  equipped with a partial order  $\sqsubseteq$  that is compatible with the monoidal operation, i.e. if  $a \sqsubseteq b$  and  $a' \sqsubseteq b'$  then  $a \otimes a' \sqsubseteq b \otimes b'$ .*

In the setting of partially ordered monoids, internal homs are easier to define.

**Definition 3** (Internal hom in a monoid). *Let  $(L, \sqsubseteq, \otimes, e)$  be a partially ordered monoid. A binary operation  $\multimap: L^{op} \times L \rightarrow L$  is said to be an internal hom when it is right adjoint to the monoidal product  $\otimes$ , i.e.  $\forall a, b, c \in L, b \otimes c \sqsubseteq a \Leftrightarrow b \sqsubseteq c \multimap a$ . The internal hom is also required to respect the ordering, contravariantly in the first coordinate and covariantly in the second, i.e. if  $b \sqsubseteq a$  and  $a' \sqsubseteq b'$  then  $a \multimap a' \sqsubseteq b \multimap b'$ .*

We can now define the central notion of this section, that of a lineale.

**Definition 4** (Lineale). *A lineale is a tuple  $(L, \sqsubseteq, \otimes, e, \multimap)$  such that  $(L, \sqsubseteq, \otimes, e)$  is a partially ordered monoid and  $\multimap$  is an internal hom for  $(L, \sqsubseteq, \otimes, e)$ .*

**Example 1.** *Examples of lineales are  $(\mathbb{N}, \geq, +, 0, \multimap)$  and  $(\mathbb{R}^+, \geq, +, 0, \multimap)$ , where*

$$a \multimap b = \begin{cases} b - a & a \leq b \\ 0 & a > b \end{cases}.$$

Any partially ordered group (see Definition 5) is a lineale with  $a \multimap b = b \otimes a^{-1}$  (see Proposition 1). Some of our examples will fall into this case.

**Definition 5** (Partially ordered group). *A partially ordered group  $(G, \sqsubseteq, \otimes, e, (-)^{-1})$  is a partially ordered monoid  $(G, \sqsubseteq, \otimes, e)$  together with an inverse operation  $(-)^{-1}$  that makes  $(G, \otimes, e, (-)^{-1})$  a group and respects the ordering contravariantly, i.e. if  $a \sqsubseteq b$  then  $b^{-1} \sqsubseteq a^{-1}$ .*

**Proposition 1.** *A partially ordered group  $(G, \sqsubseteq, \otimes, e, (-)^{-1})$  can be endowed with the structure of a lineale with  $a \multimap b := b \otimes a^{-1}$ .*

**Example 2.** *Examples of lineales obtained from partially ordered groups are  $(\mathbb{Z}, \geq, +, 0, -)$  and  $(\mathbb{R}, \geq, +, 0, -)$ .*

### 5.3.2 The category $M_L\text{Set}$

Having defined a lineale, we proceed to construct the intermediate category  $M_L\text{Set}$  over which our category of Petri nets  $\text{Net}_L$  is built.

**Definition 6** (Category  $M_L\text{Set}$ ). *Given a lineale  $(L, \sqsubseteq, \otimes, e, \multimap)$ , the category  $M_L\text{Set}$  is defined by the following data.*

- An object is a triple  $(U, X, \alpha)$ , denoted by  $U \xleftarrow{\alpha} X$ , where  $U, X$  are sets and  $\alpha: U \times X \rightarrow L$  is a function in  $\text{Set}$ .
- A morphism  $(f, F): (U, X, \alpha) \rightarrow (V, Y, \beta)$  is a pair of morphisms,  $f: U \rightarrow V$  and  $F: Y \rightarrow X$  in  $\text{Set}$ , such that  $\forall u \in U \forall y \in Y \alpha(u, Fy) \sqsubseteq \beta(fu, y)$ .

$$\begin{array}{ccc}
 U \times Y & \xrightarrow{f \times \mathbb{1}_Y} & V \times Y \\
 \downarrow \mathbb{1}_U \times F & \sqsubseteq & \downarrow \beta \\
 U \times X & \xrightarrow{\alpha} & L
 \end{array}$$

The category  $M_L\text{Set}$  allows us to have  $L$ -valued relations, including multirelations ( $L = \mathbb{N}$ ) and any other label set that can be seen as a lineale.

**Proposition 2.**  $M_L\text{Set}$  is a category.

We now proceed to define products and co-products in  $M_L\text{Set}$  and to endow it with a symmetric monoidal closed structure. We will define maps up to symmetries in  $\text{Set}$  to avoid distracting the reader with details.

**Definition 7** (Product and coproduct in  $M_L\text{Set}$ ). Given two objects  $A = (U \xleftarrow{\alpha} X)$  and  $B = (V \xleftarrow{\beta} Y)$  in  $M_L\text{Set}$ , we define their cartesian product  $A \& B$  as the following object.

$$A \& B = (U \times V \xleftarrow{\alpha \& \beta} X + Y)$$

The function  $\alpha \& \beta$  is  $U \times V \times (X + Y) \xrightarrow{[\alpha \times \epsilon_V, \beta \times \epsilon_U]} L$ , where  $\epsilon_U$  is the function that discards  $U$  in  $\text{Set}$ .

Similarly, we define their coproduct  $A \oplus B$  as the following object.

$$A \oplus B = (U + V \xleftarrow{\alpha \oplus \beta} X \times Y)$$

The function  $\alpha \oplus \beta$  is  $(U + V) \times X \times Y \xrightarrow{[\alpha \times \epsilon_Y, \beta \times \epsilon_X]} L$ .

Now, we use the monoidal operation of  $L$  to define a tensor product in  $M_L\text{Set}$ .

**Definition 8** (Monoidal product in  $M_L\text{Set}$ ). Given two objects  $A = (U \xleftarrow{\alpha} X)$  and  $B = (V \xleftarrow{\beta} Y)$  in  $M_L\text{Set}$ , we define their monoidal product  $A \otimes B$  as the following object.

$$A \otimes B = (U \times V \xleftarrow{\alpha \otimes \beta} X^V \times Y^U)$$

Where  $X^V$  and  $Y^U$  are internal hom objects in  $\text{Set}$  and the function  $\alpha \otimes \beta$  is defined by the following composition.

$$\begin{aligned}
 U \times V \times X^V \times Y^U & \xrightarrow{\Delta_U \times \Delta_V \times \mathbb{1}_{X^V \times Y^U}} U \times U \times V \times V \times X^V \times Y^U \\
 & \xrightarrow{\mathbb{1}_U \times \text{eval} \times \mathbb{1}_V \times \text{eval}} U \times X \times V \times Y \xrightarrow{\alpha \times \beta} L \times L \xrightarrow{\otimes} L
 \end{aligned}$$

where  $\Delta_U$  is the diagonal map on  $U$  and  $\text{eval}$  is the evaluation map in  $\text{Set}$ . Spelling out this definition elementwise, we obtain  $(\alpha \otimes \beta)(u, v, f, g) = \alpha(u, fv) \otimes \beta(v, gu)$ .

**Proposition 3.** *The construction above induces a functor  $\otimes: M_L\text{Set} \times M_L\text{Set} \rightarrow M_L\text{Set}$ , which is a monoidal product on  $M_L\text{Set}$ .*

**Definition 9** (Internal hom in  $M_L\text{Set}$ ). *Given two objects  $A = (U \xleftarrow{\alpha} X)$  and  $B = (V \xleftarrow{\beta} Y)$  in  $M_L\text{Set}$  we define their internal hom,  $[A, B]$ , as follows:*

$$[A, B] = V^U \times X^Y \xleftarrow{[\alpha, \beta]} U \times Y$$

The function  $[\alpha, \beta]$  is defined by the following composition.

$$\begin{aligned} V^U \times X^Y \times U \times Y &\xrightarrow{\mathbb{1}_V \times \mathbb{1}_X \times \Delta_U \times \Delta_Y} V^U \times X^Y \times U \times U \times Y \times Y \\ &\xrightarrow{\mathbb{1}_U \times \text{eval} \times \text{eval} \times \mathbb{1}_Y} U \times X \times V \times Y \xrightarrow{\alpha \times \beta} L \times L \xrightarrow{\circ} L \end{aligned}$$

Spelling out this definition elementwise, we obtain  $[\alpha, \beta](f, F, u, y) = \alpha(u, Fy) \circ \beta(fu, y)$ .

**Proposition 4.** *The construction above induces a functor  $[-, -]: M_L\text{Set}^{op} \times M_L\text{Set} \rightarrow M_L\text{Set}$ .*

The category  $M_L\text{Set}$  has products and coproducts and is a symmetric monoidal closed category with the structure defined so far.

**Theorem 1.** *The category  $M_L\text{Set}$  has products and coproducts as in Definition 7 and is a symmetric monoidal closed category with monoidal product as in Definition 8 and internal hom as in Definition 9.*

## 5.4 A category of Petri nets

A Petri net is given by a set of places  $X$ , a set of transitions  $U$ , and has two relations between these two sets that specify the precondition relation  $\blacktriangleright_\alpha$  and the postcondition relation  $\alpha_\blacktriangleright$ . In our case, these relations will be valued in a generic lineale  $L$  and the pre- and post-conditions will be objects in  $M_L\text{Set}$ . The category of Petri nets that we consider has Petri nets as objects and is obtained by putting together two copies of  $M_L\text{Set}$  (by taking a pullback in  $\text{Cat}$ ), one representing preconditions  $\blacktriangleright_\alpha: U \times X \rightarrow L$  and the other one representing postconditions  $\alpha_\blacktriangleright: U \times X \rightarrow L$ .

**Definition 10** (Category  $\text{Net}_L$ ). *Given a lineale  $(L, \sqsubseteq, \otimes, e, \circ)$ , the category  $\text{Net}_L$  is defined by the following data.*

- An object is a pair  $A = (\blacktriangleright A, A_\blacktriangleright)$  of objects  $U \xleftarrow{\blacktriangleright_\alpha} X$  and  $U \xleftarrow{\alpha_\blacktriangleright} X$  in  $M_L\text{Set}$ .
- A morphism  $(f, F): (\blacktriangleright A, A_\blacktriangleright) \rightarrow (\blacktriangleright B, B_\blacktriangleright)$  is a morphism both  $(f, F): \blacktriangleright A \rightarrow \blacktriangleright B$  and  $(f, F): A_\blacktriangleright \rightarrow B_\blacktriangleright$  in  $M_L\text{Set}$ .

**Example 3.** *Let's consider some examples of morphisms in  $M_L\mathbb{N}$ .*

- *Simulations regarding  $\alpha$  and  $\beta$ :* a morphism  $(f, F): A \rightarrow B$  in this category may represent the fact that the Petri net  $A$  is a simulation of the net  $B$  as the conditions on the morphisms in  $M_L\mathbb{N}$  ensure that the preconditions and the postconditions of  $B$  are ‘smaller or equal’ than those of  $A$ :  $\forall u \in U \ \forall y \in Y \ \alpha(u, Fy) \geq \beta(fu, y) \wedge \alpha(u, Fy) \geq \beta(fu, y)$ . Figure 32 (left) illustrates this type of morphism.
- *Refinements regarding  $F$ :* in Figure 32 (right), the Petri net  $B$  represents a refinement of the net  $A$  obtained by increasing its number of components (by adding a place).

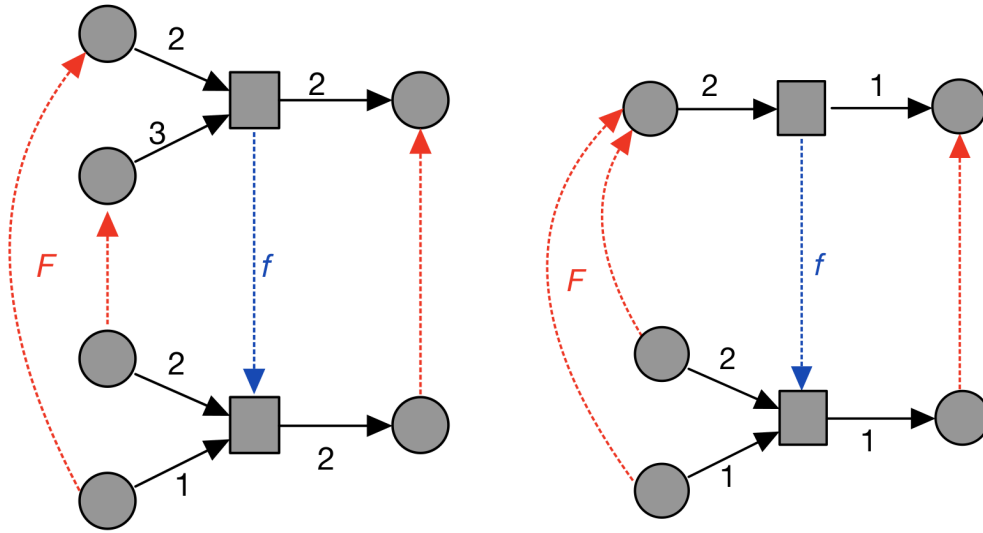


Figure 32: Examples of morphisms in  $M_L\mathbb{N}$ : simulations (left) and refinements (right).

The structure of  $M_L\text{Set}$  defines analogous structure in  $\text{Net}_L$ .

**Definition 11** (Structure of  $\text{Net}_L$ ). *The category  $\text{Net}_L$  inherits the structure of  $M_L\text{Set}$ . All the connectives are defined componentwise:*

- $A \otimes B = (\blacktriangleright A \otimes \blacktriangleright B, A\blacktriangleright \otimes B\blacktriangleright)$ .
- $[A, B] = ([\blacktriangleright A, \blacktriangleright B], [A\blacktriangleright, B\blacktriangleright])$ .
- $A \& B = (\blacktriangleright A \& \blacktriangleright B, A\blacktriangleright \& B\blacktriangleright)$ .
- $A \oplus B = (\blacktriangleright A \oplus \blacktriangleright B, A\blacktriangleright \oplus B\blacktriangleright)$ .

Examples of Petri nets modeled in this category are in the next section, where we will show how, with the possibility of changing the lineale, we can encompass different kinds of nets.

## 5.5 Different lineales

While the lineale 2 is associated with Boolean and Heyting algebras (traditional algebraic models for classical and intuitionistic propositional logic) other lineales are associated with different non-classical systems. We describe a 3-valued propositional logic where the undefined truth-value, the “unknown” state, can be thought of as neither true nor false. The adjunction determines the structure we consider for this lineale 3. We should also mention the lineale 4, associated with Belnap-Dunn’s four-valued logic. (These four values also correspond to the algebraic identities for the two conjunctions and two disjunctions of Linear Logic.)

We then consider the lineale built out of natural numbers, but with the opposite order from the natural order in  $\mathbb{N}$ . Using this lineale we build the dialectica category of multirelations  $M_L \mathbb{N}$  [111] based on Lawvere’s ideas about generalizing metric spaces [71].

Further we consider integers  $\mathbb{Z}$  with their usual order as a lineale. We believe this style of dialectica category can be profitably used to model systems where some transitions can cancel other transitions.

Next we consider the reals, in the form of the closed interval  $[0, 1]$ . These have long been considered for fuzzy sets, as the real number associated with a pair  $(u, x)$  can be thought of as the probability of the association between  $u$  and  $x$ . Finally, we consider the coexistence of several lineales in a single net by taking finite products of them. In possession of this collection of lineale structures, we define Petri nets using pre- and post-conditions between sets of events and conditions.

### 5.5.1 Original Dialectica $L = 2$

The original work on the categorical version of the dialectica interpretation has concentrated on relations that take values into 2, considered as a lineale. By considering relations with values on this lineale, which correspond to ordinary relations, we obtain the elementary Petri nets, those where pre- and post-conditions only say whether or not a place is a pre- or post-condition for a transition.

### 5.5.2 Kleene Dialectica $L = 3$

In this section, we consider a version of the Dialectica construction where the lineale is defined on the three elements set  $3 = \{-1, 0, 1\}$ . We can think of  $-1$  and  $1$  as *false* and *true* respectively. The additional truth value  $0$  can be interpreted as *undefined*. We can equip this set with the following structure.

- The monoid structure is given by the minimum,  $a \otimes b := \min\{a, b\}$ , whose unit is  $1$ .
- The ordering is given by the usual ordering  $-1 < 0 < 1$ .
- The implication is defined to yield an adjunction with the conjunction:  

$$a \multimap b := \max\{x : x \otimes a \leq b\}.$$

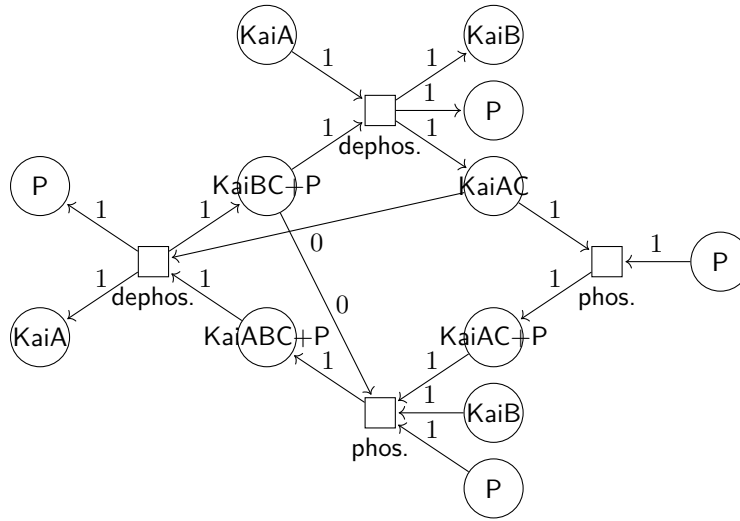
Spelling out the condition of the internal hom,  $a \multimap b = \begin{cases} 1 & \text{if } a \leq b \\ b & \text{if } a > b \end{cases}$ .

We can check that this structure is indeed that of a lineale.

**Lemma 1.** *The three-elements set is a lineale with the structure defined above.*

Thus, we can define the dialectica construction over  $(3, \leq, \otimes, 1, \multimap)$  and Petri nets with weights in 3 accordingly.

**Example 4.** We take as a motivating example the model of the chemical reactions regulating the circadian clock of *Synechococcus Elongatus* [55] that is composed of two successive phosphorylations and two successive dephosphorylations (which are the transitions labeled with *phos.* and *dephos.* in Figure 33). There is experimental evidence [4] for the existence of further feedback loops in this model. However, the precise underlying mechanism is still unknown. We can take into account these unknowns in our model for Petri nets by adding arcs with 0 weight (presence and absence are represented by 1 and -1 respectively). The Petri net in Figure 33 shows the values of the pre- and post- conditions relations as weights on the arcs.



**Figure 33:** Petri net representing the chemical reaction network regulating the circadian clock of *Synechococcus Elongatus*. Present and undefined relations are labeled by 1 and 0, respectively.

### 5.5.3 Multirelation Dialectica $L = \mathbb{N}$

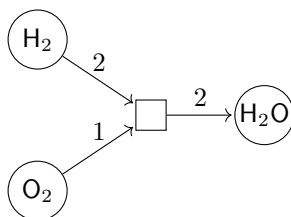
In this section, we consider a version of the Dialectica construction where the lineale is defined on the natural numbers  $L = \mathbb{N}$ . While we can think of the classical truth values as indicating whether or not a certain substance is present, we can think of the natural numbered truth values as indicating how much of a certain substance is present in a chemical reaction. We can equip this set with the following structure.

- The monoid structure is given by the sum of natural numbers,  $a \otimes b := a + b$ , whose unit is 0.
- The partial ordering is given by the opposite of the usual order on natural numbers.
- The implication is given by truncated subtraction:  $a \multimap b := \max\{b - a, 0\}$ .

The basic suggestion of this structure on the natural numbers comes from [71]. We can check that this structure is indeed that of a lineale.

**Lemma 2.** *The set of natural numbers is a lineale with the structure defined above.*

**Example 5.** *As with every chemical reaction, the one to obtain water from oxygen and hydrogen needs stoichiometric coefficients to be represented properly. We can use multirelations to take these into account, as shown in Figure 34.*



**Figure 34:** Petri net representing the chemical reaction  $2\text{H}_2 + \text{O}_2 \rightarrow 2\text{H}_2\text{O}$ .

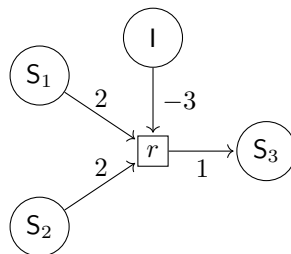
#### 5.5.4 Integers Dialectica $L = \mathbb{Z}$

Similarly to the Multirelation Dialectica, we can consider the Dialectica construction for the particular case of  $L = \mathbb{Z}$ . We endow the set of integers with a lineale structure in the following way.

- The monoid structure is given by the sum of integers,  $a \otimes b := a + b$ , whose unit is 0.
- The partial order is given by the usual ordering on the integers.
- The internal hom is given by subtraction:  $a \multimap b := b - a$ .

As  $(\mathbb{Z}, \leq, +, 0, -)$  is a partially ordered group, it is automatically a lineale with the internal hom defined above by Proposition 1.

**Example 6.** *Empirical systems often need to locally reverse the logic of preconditions to express that the presence of tokens in a given place “disables” a transition. Several different concepts of inhibitor arcs can be modeled by Petri nets including the “threshold inhibitor arc”. Reaction inhibitors in chemistry illustrate the situation: in Figure 35 chemical reaction  $r$  will not take place if the amount of substance  $I$  exceeds 3, a condition that is expressed by its inverse  $-3$ .*



**Figure 35:** Petri net representation of the chemical reaction  $S_1 + S_2 \rightarrow S_3$ . The inhibitor arc is labeled by  $-3$ , expressing that 3 is the minimum amount of substance  $I$  that prevents  $r$  from taking place.

### 5.5.5 Probabilistic Dialectica $L = [0, 1]$

So far we considered only (topologically) discrete lineales. Now we want to discuss real numbers, in particular, the closed interval  $[0, 1]$ . We first show that the closed interval  $[0, 1]$  admits a lineale structure.

- The monoid structure is given by the product of real numbers,  $a \otimes b := a \cdot b$ , whose unit is 1.
- The partial order is given by the usual ordering on the reals.
- The internal hom is given by a 'truncated division'  $a \multimap b := \begin{cases} \frac{b}{a} & a \neq 0 \wedge a \geq b \\ 1 & a = 0 \vee a < b \end{cases}$ .

**Lemma 3.** *The closed interval  $[0, 1]$  is a lineale with the structure defined above.*

**Example 7.** *The SIR (Susceptible, Infectious, Recovered) model is a simple compartmental model for infectious diseases. A susceptible individual has contact with an infectious individual with probability  $p_c$  and, after the contact, it can be infected with probability  $p_I$ , or remain susceptible with probability  $1 - p_I$ . On the other hand, an infectious individual can recover with probability  $p_R$  or remain infectious with probability  $1 - p_R$ . This setting can be represented with a Petri net where the relations between places and transitions are valued in  $[0, 1]$  (Figure 36).*

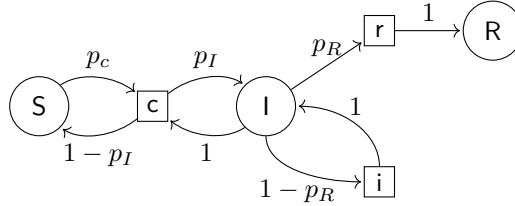


Figure 36: Petri net representing the SIR model.

### 5.5.6 Product of lineales

We have produced a pool of lineales, each of them suitable for transitions taking values in certain data types ( $\{0, 1\}$ ,  $\{-1, 0, -1\}$ ,  $\mathbb{N}$ ,  $\mathbb{Z}$  or  $\mathbb{R}^+$ ). As discussed in section 2, in empirical data analysis, a transition often carries data on more than one variable simultaneously. In this section, we show that any finite combination of lineales can be endowed with the structure of a lineale by taking finite products of them.

We remark that, because lineales are just the poset-version of symmetric monoidal closed categories, the next proposition is simply an instantiation of the idea that symmetric monoidal closed categories form a cartesian category **SymClosedCat** whose objects are symmetric monoidal closed categories, and morphisms are functors that preserve the adjunction. This category is Cartesian. This seems folklore, but we could not find it in any of the usual references (e.g. [61]).

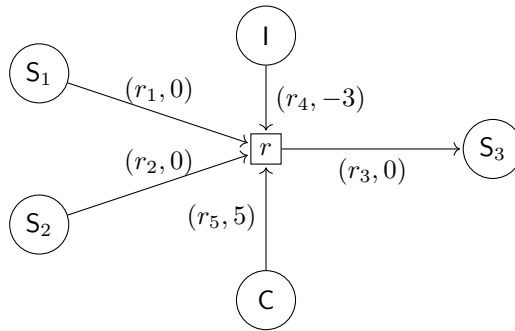
We only need the poset version here and we state it as such.

**Proposition 5.** *If  $(L_1, \leq_1, \otimes_1, e_1, \multimap_1)$  and  $(L_2, \leq_2, \otimes_2, e_2, \multimap_2)$  are lineales, then  $(L_1 \times L_2, \leq, \otimes, e, \multimap)$  is a lineale with the following structure.*

- $l \otimes l' = (l_1 \otimes_1 l'_1, l_2 \otimes_2 l'_2)$
- $e = (e_1, e_2)$
- $l \leq l'$  if and only if  $l_1 \leq_1 l'_1$  and  $l_2 \leq_2 l'_2$
- $l \multimap l' = (l_1 \multimap_1 l'_1, l_2 \multimap_2 l'_2)$

for  $l = (l_1, l_2), l' = (l'_1, l'_2) \in L_1 \times L_2$ .

**Example 8.** There is a dual situation to inhibition in chemistry, namely, catalysis. A catalyst is a substance that increases the reaction rate without being consumed by the reaction. The presence of a substance  $S$  in a chemical reaction might then play one of three roles: reactant/product, inhibitor, or catalyst. We claim that the product of the lineales  $\mathbb{R}^+$  and  $\mathbb{Z}$  has enough expressive power to model reaction rates in the presence of both inhibitors and catalysts. In Figure 37 pairs of the form  $(r, 0)$ , state that those substances are not inhibitors nor catalysts, and  $r$  is the rate at which a substance is consumed or produced. The negative number in the label  $(r_4, -3)$  expresses that  $I$  is an inhibitor of reaction  $r$ , and  $-3$  is the minimum amount of  $I$  required to slow down the reaction by the rate  $r_4$ . Finally, the label  $(r_5, 5)$  indicates that  $C$  is a catalyst and  $5$  is the minimum amount of  $C$  required to increase the reaction rate by  $r_5$ .



**Figure 37:** Petri net representation of reaction rates for the chemical reaction  $S_1 + S_2 \rightarrow S_3$  in the presence of an inhibitor  $I$  and a catalyst  $C$ . Labels are pairs  $(r, z)$  where  $z$  states the role of the substance as reactant/product (zero), inhibitor (negative integers), and catalysts (positive integers); and  $r$  the rate at which the substance is consumed/produced (if  $z = 0$ ), or at which the reaction rate increases ( $z > 0$ ) or is slowed down (if  $z < 0$ ).



## **Part IV**

### **Final remarks**



## CHAPTER 6

# Conclusions and further work

**Contents**

6.1	Discussion and conclusions . . . . .	88
6.1.1	On computational history of chemical space . . . . .	88
6.1.2	Local geometry of chemical space . . . . .	89
6.1.3	Compositionality in chemical space . . . . .	90
6.2	Further work . . . . .	91

## 6.1 Discussion and conclusions

Starting from the observation that substances and reactions are the central entities of chemistry, we have structured these data in a formal space called a directed hypergraph, which arises when substances are connected by their reactions. We call this hypernet chemical space. We have explored different levels of description of this space, namely chemical, historical, and formal aspects. The discussion and conclusions reached are presented in this section.

### 6.1.1 On computational history of chemical space

Our investigation of the chemical space from 1800 up to 2015 allowed us to quantify patterns in its growing mechanism. First, we report that substance discovery has been ruled by synthesis since the early 19th century and became the established method to report new compounds about 1900, that is, 72 years after Wöhler's synthesis of urea, providing quantitative evidence against Wöhler's myth. The results also show that it took 40 years from the introduction of chemical structure theory for chemists to adequate their discipline to the new resolution on substance structure and its power for the exploration of the space through synthesis.

The number of new substances is growing exponentially. Moreover, the production of compounds was heavily reduced by two external events: the World Wars. After each war, substance production went back to pre-war growth rates. This war invariance of growth is also observed for two scientific events: the introduction of structural theory and the raise of organometallic chemistry changed the rates, but the space went back to its historical growth trend very quickly.

We found patterns in the selection of substrates; chemists have been conservative in the selection of their starting materials, presumably as a disciplinary consequence of starting from substances that are readily available or as a way to develop valid and reliable expert intuition to explore the chemical space [62]. Or perhaps, it is as simple as Mendeleev put it: "conservatism in science is completely inevitable, because science, in essence, is a legacy, unthinkable except as the wisdom of centuries past, and thus cannot be passed on without conservatism." [47]. In contrast to conservatism in the selection of substrates, chemists showed to be motivated to discover new compounds of the space in an exponential manner. In spite of this, only a handful of compositions have been extensively explored, CHNO being the most popular one since 1890.

Most of the reported reactions over history use two substrates. Often, one of them is little-known and the other is part of the synthetic toolkit of preferred substrates; this trend is followed, for instance, by reactions of acetic anhydride, the most used substrate since 1940. We called this trend the *fixed-substrate approach*.

We have investigated the effect of changes in chemical space on systems of chemical elements for the period 1800-1868, in particular, in the identification of similarity patterns, which we found to play a major role in shaping the system. We found that the space, before 1840, was not ready for the formulation of a system of elements close to that of 1869.

We found that before 1830 the expansion of the space involved discovering a wide range of new combinations, which allowed observing several of the similarities of 1868 as early as 1826. The rise of organic chemistry after 1830 directed the expansion of the chemical space towards compounds of organogenic elements, mainly O, H, C, N, and S. This brought two effects: facilitated the detection of similarities among organogenic elements even with low fractions of the chemical space, and hindered some similarities among metals, which required large fractions of the chemical space to be observed.

This explains why chemists in the nineteenth-century struggled to detect similarities among metals but readily observed those among organogenic elements.

By 1837 a large number of similarity relations arose and remained in the systems of every year until 1845, several of them were to remain until Meyer and Mendeleev's formulation of their systems. This "golden period" was perhaps a ripe moment to have devised a system similar to theirs, which challenges the traditional claim that placed it in the 1860's [10, 134].

Our data-driven approach to element similarity could shed some light on the ongoing discussions on the conformation of families of similar elements [117]. For instance, whether Si is more similar to the carbon group or to the Ti group; or whether V resembles more the pnictogens or the current group 5. Our results show that by 1869 Si was actually closer to the Ti group and that V, as well as Nb (constituting the whole group 5 known at that time), was more similar to P. The output data of this study, including the chemical spaces computed from the atomic weight tables for several nineteenth-century chemists, is available here.

This work constitutes a computational, entirely data-driven approach to the history of chemistry. Using chemical compounds and reactions as the only data source places the study very closely to the core of chemistry [137]. Reaxys data relieved us almost entirely from the need to focus on themes and topics of the scientific discourse, which forms the main difficulty in text-based approaches to the history of science [68]. We hope that the approach presented here complements conventional approaches to the history of science.

### 6.1.2 Local geometry of chemical space

Drawing from Forman's definition of curvature for simplicial complexes (with roots in Riemannian geometry), and from its graph version, we generalized the Forman-Ricci curvature to hypergraphs, both directed and undirected (Equations 4.10 and 4.12). Graph curvatures used in previous studies thus become particular cases [127, 129, 130, 141–143, 152–155]. This curvature notion complements vertex-centered tools, providing an edge-based approach for the analysis of networks.

Also, after a brief account of the Forman-Ricci curvature for (un)directed graphs, we determined the upper and lower bounds for Forman-Ricci curvature for (multi)graphs and (multi)hypergraphs, which so far had not been investigated. In an undirected graph,  $F(e)$  is positive for isolated edges, zero for edges whose endpoints are connected each to two nodes, and negative if the nodes of  $e$  stand in more relations. Our generalization of Forman-Ricci curvature for undirected hypergraphs preserves these properties. In general, it quantifies the difference between the size of a hyperedge and the combined number of connections of its nodes (Equation 4.12).

Directed hypergraphs have a richer structure and so its curvature has four components:  $F(\rightarrow e)$ ,  $F(e \leftarrow)$ ,  $F(\leftarrow e)$ , and  $F(e \rightarrow)$ . Each of these components is the difference in tail/head size and the number hyperarcs either to or from its vertices. Since each component carries a part of the hyperedge structure, different combinations can be tailored to address different aspects of its structure, for instance,  $F(\rightarrow e \rightarrow) = F(\rightarrow e) + F(e \rightarrow)$  may be interpreted as the flow through  $e$ , while  $F(\rightarrow e) - F(e \rightarrow)$  was shown to be a simple yet powerful tool to quantify assortativity (for those directions). Moreover, our definition of Forman-Ricci curvature focuses entirely on hyperedge structure, that is, its size, and the degree of its vertices; it does not re-interpret hyperedges as "higher-dimensional" objects nor does it implicitly introduce additional structures, like boundaries of simplices, that are not part of the original data, which differentiates it from a version proposed in [132].

We applied these curvatures to the analysis of two large networks, one of social and the other of chemical interactions. The analysis of Wikipedia vote network exemplified the Forman-Ricci curvature

of undirected hypergraphs, where elections constituted hyperedges and users/voters vertices. Since our motivation, the chemical space, is a directed hypergraph, we characterized the directed Forman-Ricci curvature by studying its distribution for well-known metabolic networks (traditionally modeled as graphs): *Escherichia coli*, *Mycobacterium tuberculosis*, and *Helicobacter pylori*. We identified bottleneck reactions (defined as those reactions whose educts can be obtained from several reactions and whose products are often used as educts), which are characterized by having very negative  $F(\rightarrow e \rightarrow)$  values. For *E. coli* this reactions is:  $\text{adp} + \text{h} + \text{pi} \rightarrow \text{atp} + \text{h} + \text{h}_2\text{o}$ . Bottleneck reactions can be considered as assortative ones, for they transform popular products into popular educts.

The definition of Forman-Ricci curvature presented here can be weighted (Equation 4.11) based on meta information of the network, e.g. user's seniority in the Wikipedia example or stoichiometric coefficients in the metabolic network. The effects of using this feature of our definition on these and other weighting schemes need to be explored in future studies on the curvature of hypergraphs.

### 6.1.3 Compositionality in chemical space

We produced a categorical model of bipartite directed graphs, which are equivalent to directed hyperegraphs, called Petri nets. Most of the recent work on Petri nets focuses on the unfoldings of a single net. We have presented a categorical model for Petri nets that focuses on the diverse nature of network relations. This is a fundamentally different approach to Petri nets since it allows the use of different kinds of transitions (different kinds of labels in their graphs), while maintaining their compositionality.

Our model can handle different kinds of transition whose labels can be represented as a lineale (a poset version of a symmetric monoidal closed category). Several sets of labels, from those often used in empirical data modeling, can be endowed with the structure of a lineale, including: stoichiometric coefficients in chemical reaction networks ( $L = \mathbb{N}$ ), reaction rates ( $L = \mathbb{R}^+$ ), inhibitor arcs ( $L = \mathbb{Z}$ ), gene interactions ( $L = \{0, 1\}$ ), unknown or incomplete data ( $L = \{-1, 0, 1\}$ ), and probabilities ( $L = [0, 1]$ ).

The structure of the lineale is lifted to the category  $M_L\text{Set}$  from which the category  $\text{Net}_L$  of Petri nets is built. Both  $M_L\text{Set}$  and  $\text{Net}_L$  are symmetric monoidal closed categories with finite products and coproducts, providing a compositional way to put together smaller nets into bigger ones, making sure that algebraic properties of the components are preserved in the resulting net.

The category  $\text{Net}_L$  is a model for weighted and directed bipartite relations and therefore we anticipate applications of the compositionality of  $\text{Net}_L$  in the broader context of directed bipartite graphs [50, 57], in particular, for their applications to real-world networks. For instance, the labeled wiring of these graphs is key to the empirical analysis of metabolic networks, where the metabolism of an organism is studied in terms of the concurrence of smaller functional subnets called modules. We wonder whether our formal connectives may assist in the reconstruction and understanding of whole metabolisms in terms of the concurrence of their modules.

There is much more work to be done still. Both in the applications we are pursuing and in the theory of Dialectica Petri nets. On the theory side, notions of behavior (token game) should be investigated and on the practical side, we have still to investigate how the implemented systems can be modified to deal with our nets.

Dialectica Petri nets share some of the pros and cons of other Linear Logic based nets. As far as we know no one has investigated Differential Linear Logic Petri nets, yet (see [29] for relating differential interaction nets to the  $\pi$ -calculus). We wonder if this would make the exchange of information with modellers somewhat easier. Finally, we would like to investigate whether we could code our nets

using Catlab <https://github.com/AlgebraicJulia/Catlab.jl>, a framework for computational category theory, written in the Julia language and already used for other styles of Petri nets.

## 6.2 Further work

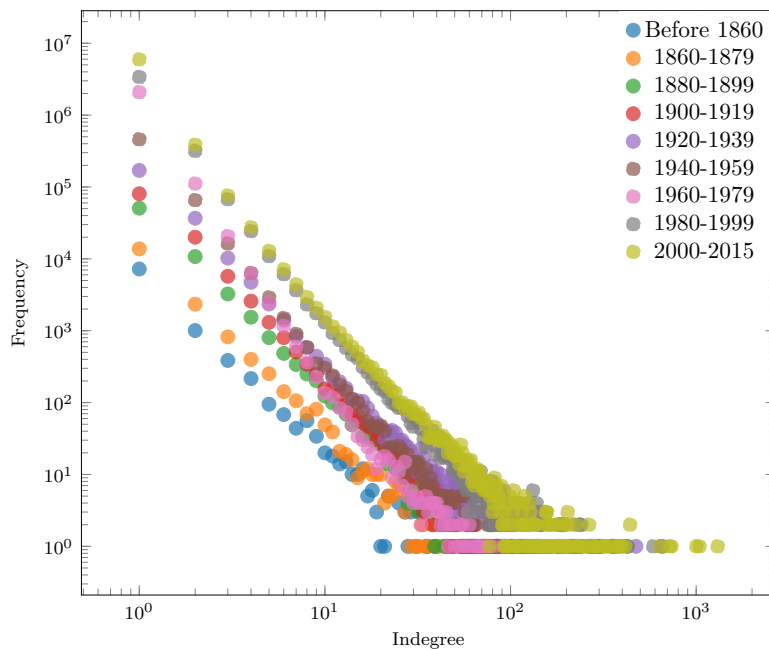
We have unveiled features of the growth mechanism and the role that the chemical space has played in shaping similarity relations among elements. We have also shown that the chemical space can be faithfully modeled as a growing directed hypergraph. These findings open up the following research questions:

- Could hypergraph random models be used to explain the growing mechanism of the space? If so, what are the generating rules that yield the patterns reported here of substance use, production, and assortativity, as well as the curvature and assortativity fingerprints of reactions?
- Can we extend our approach to assess similarity among elements to substances? If similarity among substances is given by functional molecular fragments, and if "similarity between elements comes from the forms they share", can we generalize our formal tools to deal with the more general claim that "similarity between substances comes from the functional molecular fragments they share?" If that is the case, what are the topological structures that underlie the chemical space?
- As argued in this work, compositionality is a key feature of chemical reasoning. What are then the appropriate categorical models of hypergraphs that may bring a compositional theory for the level of abstraction described in this work? And what are the formal connections between the hypergraph level and the "intermediate level of abstraction for chemistry" proposed by Stadler and coworkers [2, 9]?



## Appendices



APPENDIX **A****Supplementary data of Chapter 2**

**Figure 38:** Frequency distributions of number of reactions  $R$  producing the same target.

### Abbreviations for compounds of Tables 4 and 8:

AcOH (Acetic acid), Ac<sub>2</sub>O (Acetic anhydride), Ag (Silver), Ag<sub>2</sub>S (Silver sulfide), BnBr (Benzyl bromide), B(OH)<sub>3</sub> (Boric acid), Br<sub>2</sub> (Bromine), BZA (Benzoic acid), BzCl (Benzoyl chloride), C<sub>6</sub>H<sub>6</sub> (Benzene), CH<sub>2</sub>N<sub>2</sub> (Diazomethane), CH<sub>2</sub>O (Formaldehyde), Cl<sub>2</sub> (Chlorine), CO (Carbon monoxide), CO<sub>2</sub> (Carbon dioxide), CoO (Cobalt(II) oxide), CuO (Copper(II) oxide), DHBZA (3,4-dihydroxybenzoic acid), DMA (Dimethylamine), EDBB (1,1'-(1,2-ethanediyl)bisbenzene), Et<sub>2</sub>O (Diethyl ether), EtOH (Ethanol), FC (Ferrocene), Glc (Glucose), H<sub>2</sub> (Hydrogen), H<sub>2</sub>O (Water), HCl (Hydrochloric acid), Hg (Mercury), HgO (Mercury(II) oxide), HNO<sub>3</sub> (Nitric acid), H<sub>2</sub>S (Hydrogen sulfide), H<sub>2</sub>SO<sub>4</sub> (Sulfuric acid), I<sub>2</sub> (Iodine), KOH (Potassium hydroxide), MAC (Methylammonium carbonate), MBPh (4-methoxybiphenyl), MeCHO (Acetaldehyde), MeI (Methyl iodide), MeOH (Methanol), Morph (Morpholine), Na<sub>2</sub>CO<sub>3</sub> (Sodium carbonate), NH<sub>3</sub> (Ammonia), NiO (Nickel(II) oxide), NPhOH (4-nitro-phenol), OA (Oxalic acid), Ph<sub>2</sub> (Biphenyl), PhA (Phthalic acid), PhAc (Acetophenone), PhAcet (Phenyl acetylene), PhCHO (Benzaldehyde), Ph<sub>2</sub>CO (Benzophenone), PhNH<sub>2</sub> (Aniline), Ph<sub>2</sub>S<sub>2</sub> (Diphenyl disulfide), TBTB (Tributyltin bromide), TBTC (Tributyltin chloride), TFA (Trifluoroacetic acid), TMTC (Trimethyltin chloride), TPPO (Triphenylphosphine oxide), UF<sub>6</sub> (Uranium hexafluoride), ZnO (Zinc oxide).

### Analysis of jumps for products

The distribution of participation of compounds in  $R$  different reactions as products for the following periods is shown in Figure 2b (main text): Period 1 (Before 1860), Period 2 (1860-1879), Period 3 (1880-1899), Period 4 (1900-1919), Period 5 (1920-1939), Period 6 (1940-1959), Period 7 (1960-1979), Period 8 (1980-1999), Period 9 (2000-2015).

We found that the participation of products in less than 21 reactions ( $R \leq 20$ ) accounts for 97% of the whole variance of the distribution (Figure 2b, main text). Therefore, we plotted (Figure 39) the logarithm growth of participation of compounds in  $R \leq 20$  different reactions as products in the nine periods of Figure 2b. The least squares regression method shows that the logarithm growth of participation of compounds  $\log \text{Comp}_i$ , with  $i = 1, \dots, 9$  follows the equation

$$\log \text{Comp}_i = 0.8447904i + 8.3520785 + \text{Residual}_i,$$

where  $\text{Residual}_1 = 0.04787297$ ,  $\text{Residual}_2 = -0.17878533$ ,  $\text{Residual}_3 = 0.28512578$ ,  $\text{Residual}_4 = -0.07595590$ ,  $\text{Residual}_5 = -0.20859164$ ,  $\text{Residual}_6 = -0.16903558$ ,  $\text{Residual}_7 = 0.35814213$ ,  $\text{Residual}_8 = 0.16272740$ , and  $\text{Residual}_9 = -0.22149983$ .

These residuals pass the Shapiro-Wilk and Kolmogorov-Smirnov normality tests for mean 0 and standard deviation  $sd = 0.2228266$ , with the corresponding probabilities  $p^{ShW} = 0.1398$  and  $p^{KS} = 0.6966$ . There are two historical jumps, namely between 1860-1879 and 1880-1899, and between 1940-1959 and 1960-1979, corresponding to  $\text{Residual}_3 = 0.28512578$  and  $\text{Residual}_7 = 0.35814213$ , which are higher than the standard deviation  $sd = 0.2228266$  (and also higher than the other residuals), and still in the range  $(-2sd, 2sd)$ . As a result, the logarithm difference  $\log \text{Comp}_i - \log \text{Comp}_{i-1}$ , with  $i = 2, \dots, 9$  reaches the highest and the second highest at these two periods.

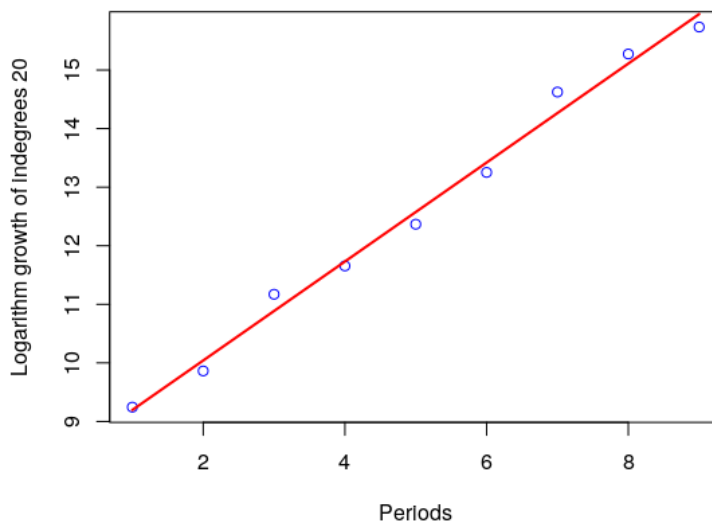
### Variance analysis of the number of new compounds

For a continuous  $t$ , the exponential growth is described by  $s_t = s_0 e^{kt}$ , from which  $r = 100(e^k - 1)$ . In a discrete case, data variability needs to be regarded, therefore the normality of  $r$  values has to be

tested. In this setting,  $r_t = (s_{t+1}/s_t) - 1$ , which for the chemical reactions data failed normality tests for 1804-2015 and for the period 1804-1860 that is of high fluctuation. Therefore we explored the distribution of  $Y_t := \ln s_{t+1} - \ln s_t$ , which we found to be normally distributed for three periods, but not for the whole period 1804-2015. We observed that the variance is time dependent and  $Y_t$  follows an ARCH (autoregressive conditional heteroskedasticity) model. In the simplest form,  $Y_t$  is a switching model consisting of three periods with the general form  $Y_t = \mu_i + \sigma_i Z_t$ , being  $i$  an index for the respective period and the residuals  $\{Z_t\}_{t \geq 1}$  a source of identically distributed Gaussian noises. This is equivalent to:

$$s_{t+1} = s_t e^{\mu_i + \sigma_i Z_t} \quad (\text{A.1})$$

In all cases the residuals  $\{Z_t\}_{t_0}^{t_f}$  passes the Shapiro-Wilk and the Kolmogorov-Smirnov normality tests  $\mathcal{N}(0, 1)$ , for  $t_0$  and  $t_f$  the respective initial and final year of period  $i$ . The combination of all three subsequences of residuals passes the Kolmogorov-Smirnov normality test, although fails the Shapiro-Wilk one.



**Figure 39:** Logarithm growth of participation of compounds in  $R \leq 20$  different reactions as products in the nine periods.

### Fitting the growth model

Calculations for the growth fitting were conducted by linear regression methods and the equation for the annual number of new compounds is

$$s_t = 8.18 \times 10^{-33} e^{0.04324t} \quad (\text{A.2})$$

where  $t$  is a year between 1800 and 2015.

## Supplementary data of Chapter 3

### Similarity between systems of chemical elements

We quantified the similarity of element  $x$  regarding element  $y$  as the fraction of substances of  $x$  in whose formulae  $x$  can be replaced by  $y$  yielding a formula that is part of the chemical space. Hence, for an element  $x$  having  $s_x$  substances in the chemical space, which are gathered as  $\{s_x^1, s_x^2, \dots, s_x^{s_x}\}$ , there is an associated multiset of formulae  $F_x = \{f(s_x^1), f(s_x^2), \dots, f(s_x^{s_x})\}$ , where  $f(s_x^i)$  is the arranged formula of substance  $i$  containing element  $x$ . Arranged formulae are assigned to a reference element, whose similarity regarding other elements is to be calculated. These formulae are found by replacing the reference element  $x$  in their formulae by the symbol  $A$ . The resulting formula is lexicographically ordered for the sake of comparison with other formulae. For example, the arranged formula of  $\text{NH}_2\text{Cl}$ , with  $\text{Cl}$  as reference element, is  $\text{AH}_2\text{N}$ . Figure 2 (main text) shows a toy-example for the similarity between  $\text{Cl}$  and  $\text{Br}$  ( $s(\text{Cl} \rightarrow \text{Br})$ ) and  $\text{Br}$  and  $\text{Cl}$  ( $s(\text{Br} \rightarrow \text{Cl})$ ). There,  $\text{Cl}$  has eight substances in the chemical space, namely  $\{\text{NH}_2\text{Cl}, \text{HAuCl}_4, \text{VCl}_3\text{O}, \text{K}_3\text{CrCl}_6, \text{CH}_2\text{CCl}_2, \text{NaCl}, \text{KClO}_3\}$ . In turn,  $\text{Br}$  has seven substances:  $\{\text{CH}_2\text{CBr}_2, \text{CHBrCHBr}, \text{NaBr}, \text{KBrO}_3, \text{MoBr}_4, \text{HgOBr}_2\}$ . The multiset of arranged formulae for  $\text{Cl}$  is  $F_{\text{Cl}} = \{\text{AH}_2\text{N}, \text{A}_4\text{AuH}, \text{A}_3\text{OV}, \text{A}_6\text{CrK}_3, \text{A}_2\text{C}_2\text{H}_2, \text{ANa}, \text{AKO}_3\}$  and for  $\text{Br}$  it is  $F_{\text{Br}} = \{\text{A}_2\text{C}_2\text{H}_2, \text{A}_2\text{C}_2\text{H}_2, \text{ANa}, \text{AKO}_3, \text{A}_4\text{Mo}, \text{A}_2\text{HgO}\}$ . Note that in both cases all formulae have multiplicity one, except  $\text{A}_2\text{C}_2\text{H}_2$ , which has multiplicity two in  $F_{\text{Br}}$ . This shows that  $\text{Br}$  has two substances with arranged formulae  $\text{A}_2\text{C}_2\text{H}_2$ . They are  $\text{CH}_2\text{CBr}_2$  and  $\text{CHBrCHBr}$ , which are isomers participating in single-step reactions before 1869. Note that  $\text{A}_2\text{C}_2\text{H}_2$  has multiplicity one for  $\text{Cl}$ , as only  $\text{CH}_2\text{CCl}_2$  had participated in single-step reactions by 1869. Thence,  $F_{\text{Br}}$  can be rewritten as  $F_{\text{Br}} = \{\text{A}_2\text{C}_2\text{H}_2^2, \text{ANA}^1, \text{AKO}_3^1, \text{A}_4\text{Mo}^1, \text{A}_2\text{HgO}^1\}$ , which, in general, has the form  $F_x = \{f_i^{m_x(i)}\}$ , where  $f_i$  is the  $i$ -th arranged formulae of element  $x$  whose formula multiplicity is  $m_x(i)$ . By multiplicity of a formula is meant the number of times the formula shows up in the multiset, that is the number of times the formula is found in the chemical space of element  $x$ .

With the list of arranged formulae for elements  $x$  and  $y$ , we can calculate  $s(x \rightarrow y)$  as:

$$s(x \rightarrow y) = \frac{|F_x \cap F_y|}{|F_x|}$$

As  $|F_x|$  amounts to counting the multiplicities of arranged formulae of  $x$ , then  $|F_x| = \sum m_x(i)$ . Likewise, finding the common arranged formulae between  $F_x$  and  $F_y$  amounts to counting the minimum multiplicity of formulae that appear in both multisets  $F_x$  and  $F_y$ :  $|F_x \cap F_y| = \sum \min(m_x(i), m_y(i))$ , which for  $|F_{\text{Cl}} \cap F_{\text{Br}}| = \min(m_{\text{Cl}}(\text{A}_2\text{C}_2\text{H}_2), m_{\text{Br}}(\text{A}_2\text{C}_2\text{H}_2)) + \min(m_{\text{Cl}}(\text{ANa}), m_{\text{Br}}(\text{ANa})) + \min(m_{\text{Cl}}(\text{AKO}_3), m_{\text{Br}}(\text{AKO}_3)) = \min(1, 2) + \min(1, 1) + \min(1, 1) = 3$ . Therefore, the similarities  $s(\text{Cl} \rightarrow \text{Br})$  and  $s(\text{Br} \rightarrow \text{Cl})$  are calculated as  $s(\text{Cl} \rightarrow \text{Br}) = 3/7$  and  $s(\text{Br} \rightarrow \text{Cl}) = 3/6$ .

## Stability of similarities regarding chemical space size

### Sampling the chemical space

For each year we randomly took  $s\%$  of the space and determined the most similar element(s) for each element. This experiment was carried out 100 times. For each similarity  $x \rightarrow y$  resulting for the whole space of that year, we counted in how many of the 100 experiments appeared  $x \rightarrow y$ . The higher the frequency of appearance of the similarity  $x \rightarrow y$  in the 100 experiments, the more stable the similarity. We carried out this analysis for the 19 sample sizes 95%, 90%, 85%, ..., 5%. The higher the frequency for different values of  $s$ , the less that similarity depends on the sample size.

### Contrasting Meyer and Mendeleev' systems of chemical elements with those of the chemical space (contemporary approach)

We took the three systems by Meyer, which were formulated in 1864 [101], 1868 [120] and 1869/70 [102]; and the first Mendeleev' system published in 1869 [96]. We extracted the similarities among the elements out of these systems and contrasted them with the "most similar" relationships of the systems of elements of the respective years 1863, 1867 and 1868. The time difference of one year between the system of elements of each author and the system of elements of the chemical space is to regard the time required for a chemist to be updated with the literature in the nineteenth-century.

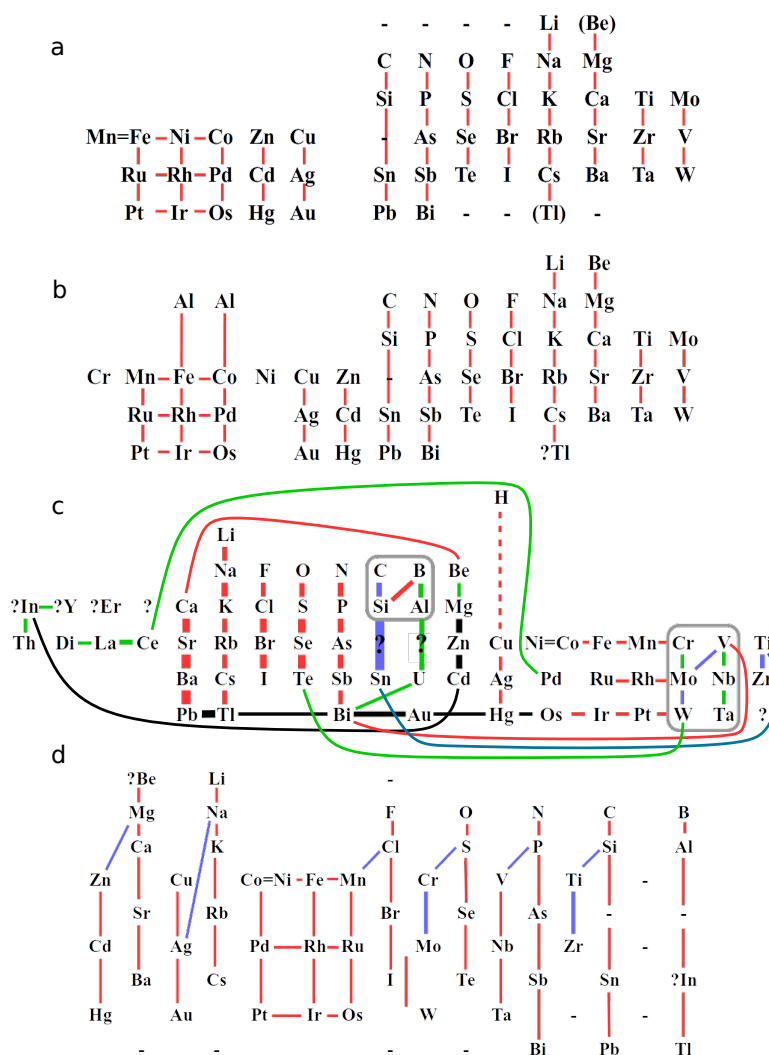
As Meyer did not explicitly discuss similarities among elements, we followed his principle of regarding as similar those elements belonging to a column of his 1864 and 1868 tables. In addition, we regarded elements with the same valency as similar for his 1864 table, plus the well-known similarities among transition metals  $\text{Mn}=\text{Fe}$ ,  $\text{Ni}$ ,  $\text{Co}$ ,  $\text{Ru}$ ,  $\text{Rh}$ ,  $\text{Pd}$  and  $\text{Pt}$ ,  $\text{Ir}$ ,  $\text{Os}$ . Hence, for instance, we regarded  $\text{Li}$ ,  $\text{Na}$ ,  $\text{K}$ ,  $\text{Rb}$ ,  $\text{Cs}$ ,  $\text{Tl}$  and  $\text{Cu}$ ,  $\text{Ag}$ ,  $\text{Au}$  as similar, as they hold valence 1. Note that Meyer did not write the valency of  $\text{Cu}$ ,  $\text{Ag}$  and  $\text{Au}$  in 1864, but one can infer it was 1 from the groups on the left of this group of elements [101]. For Meyer's periodic system of 1868, besides elements belonging in the same column, we consider the following transition metal similarities observed in his table:  $\text{Mn}$ ,  $\text{Fe}$ ,  $\text{Co}$ ,  $\text{Ru}$ ,  $\text{Rh}$ ,  $\text{Pd}$  and  $\text{Pt}$ ,  $\text{Ir}$ ,  $\text{Os}$ . In this case, we did not regard same valency as a similarity criterion, because valency is not highlighted in Meyer's table and because of his shifting and reshifting of  $\text{Al}$  [120, 122]. Similarities in his 1869 table are further discussed in the paper where they were published. Therefore, besides the usual vertical similarities (in the 1869 published representation corresponding to rows), we also included those similarities mentioned by Meyer [102], plus the transition metal ones:  $\text{Mn}$ ,  $\text{Ru}$ ,  $\text{Os}$ ,  $\text{Fe}$ ,  $\text{Rh}$ ,  $\text{Ir}$  and  $\text{Co}=\text{Ni}$ ,  $\text{Pd}$ ,  $\text{Pt}$  (Figure 40). Mendeleev discussed thoroughly the similarities and even some lack of similarities [94], both of them listed in Table 12.

By contrasting the similarities reported by each author with those corresponding to most similar relationships among chemical elements allowed by the chemical space, we calculated the true positive

and false negative rates of each system of elements (Table 13). True positives ( $TP$ ) correspond to similarities observed by year  $y$  and extracted from chemist's system of elements. True negatives ( $TN$ ) correspond to non most similar relationships observed in year  $y$  that are also non similarities according to chemist's system of elements. False positives ( $FP$ ) are non similarities in year  $y$  corresponding to similarities in chemist's system of elements. False negatives ( $FN$ ) are similarities in year  $y$  not retrieved from chemist's system of elements. The true positive rate ( $TPR$ ) is given by  $TP/(TP + FN)$ , false positive rate ( $FPR$ ) by  $FP/(FP + TN)$ .

**Table 12:** Similarities as mentioned by Mendeleev in his extended 1869 publication [94]. Red entries correspond to similarities Mendeleev thought did not exist and the blue one to “not so well studied.”

Bi,V,Sb,As	P,N	Te,Se,S
I,F	Mg,Zn,Cd	Cu,Ag
Pb,Ba,Sr,Ca	Pd,Rh,Ru	Os,Ir,Pt
Ag,Cu,Hg	Sb,Bi	Tl,Cs
Alkaline-earth metals	N group	S group (in part)
Ce companions	Li,K,Na	Ca,Sr,Ba
O,S,Se,Te	N,P,As,Sb	Mg,Zn,Cd
P,As,Sb	Os,Pt,Ir	W,V,Mo
Si,B,F	O,N,C	Cr,Ni,Cu
N,P,As	Na,K,Rb,Cs	F,Cl,Br,I
V,Nb,Ta	Cr,Mo	Sr,Ba,Pb
Li,Na,K,Cu,Rb,Ag,Cs,Tl	Be,Mg,Ca,Zn,Sr,Cd,Ba,Pb	B,Al,U,Bi
N,P,V,As,Nb,Sb,Ta	O,S,Se,Te,W	Be,Ca,Sr,Ba,Pb
Ti,Si,Sn	Cr,S,Te	Mn,Cl,Br
P,V,As	S,Cr,Se	Cl,Mn,Br
Nb,V,Sb	In,Mg,Zn,Cd	Zr,C,Sn
Y,Th,In	Co,Ni,Cr,Mn,Fe	Ce,La,Di,Pd,Rh,Ru
Mg,{Ca,Sr,Ba}	Pb,Tl,Bi,Au,Hg,Pt,Ir,Os,W	Pb,Tl
Hg,Pt	H	H,Cu,Ag,Hg
Ba,Pb,Tl	V,Cr,Nb,Mo,Ta,W	B,Al,U
Si,Ti,Zr,Sn	Tl,alkali metals	Si,B
Halogens	Pt companions	Cl,I,Br
Ag,Pb,Hg	Ta,Sn,Ti	Be,Zr,U
C,Si,...,Sn	Li,Na,K,Rb,Cs,Tl	C,Si,Ti,Zr,Sn
V,P,Sb,As	Si,Ti,?	As,Sb,Nb
{Mg,Zn,Cd},{Ca,Sr,Ba}	Pt,Ir,Os	Bi,Au
C,B,Si,Al		



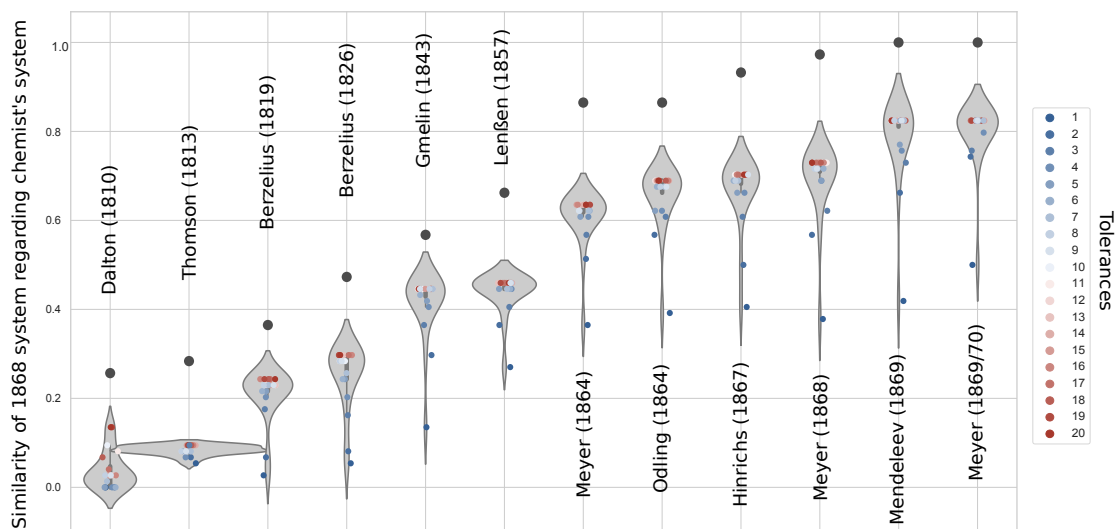
**Figure 40:** Systems of chemical elements by Meyer (a: 1864, gathering together his three separate tables; b: 1868; d: 1869/70) [101, 102, 120] and Mendeleev (c: 1869, rotated and reflected for the sake of comparison with the other systems) [96]. Element symbols are updated to current notation. Lines and boxes indicate similarities. The complete list of similarities for Mendeleev is found in Table 12. Line widths are proportional to the number of times the similarity is discussed by each author. Line colours are used only for the sake of clarity

## Chemical spaces from atomic weights

As contemporary atomic weights are related by simple fractions with atomic weights of different chemists (Table 10), we adjusted a contemporary chemical formula  $F = X_x Y_y \dots Z_z$  to

$$F_A = X_{x f_{A(X)}} Y_{y f_{A(Y)}} \cdots Z_{z f_{A(Z)}} \quad (\text{B.1})$$

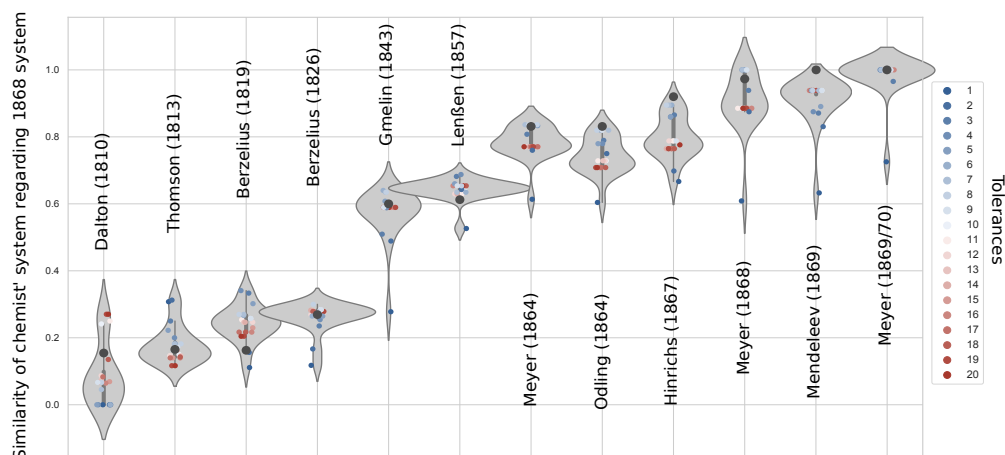
Here,  $X, Y, \dots, Z$  are chemical elements and  $x, y, \dots, z$  their stoichiometric coefficients in  $F$ ;  $f_{A(X)}, f_{A(Y)}, \dots, f_{A(Z)}$  are the respective coefficients modifying  $x, y, \dots, z$  to yield the formula  $F_A$ , as an approximation to that regarded by chemist  $A$ . Coefficients  $f_A$  are calculated as follows: knowing the current ( $W(e)$ ) and chemist's ( $A(e)$ ) atomic weights of element  $e$  (Table 10), as well as the respective values for hydrogen ( $W(H)$  and  $A(H)$ ), we calculate the ratios  $(W/A)(e) = (W(e)/W(H))/(A(e)/A(H))$  and  $(A/W)(e) = (A(e)/A(H))/(W(e)/W(H))$ . Our aim is determining the simplest fraction  $f$  approximating either  $(W/A)(e)$  or  $(A/W)(e)$ . As these ratios either fall in the real interval  $(0, 1]$  or correspond to figures of the form  $\alpha + \beta$ , where  $\alpha$  is an integer and  $\beta$  is a real number in the interval  $(0, 1]$ , we need to find a  $0 < f \leq 1$  that best approximates either  $\beta$  or the ratio falling in the interval  $(0, 1]$ . The best  $f$  corresponds to a fraction of a Farey sequence [37] (Supporting Information) minimising the relative error of the approximation ( $\text{error}(r, f) = |r - f|/r$ , with  $r$  either  $(W/A)(e)$  or  $(A/W)(e)$ ). We allowed 20 different error tolerances  $\tau$  for the approximation, from 1 to 20% of relative error, in such a manner that for each  $\tau$ , the selected fraction  $f$  always approximates  $r$  with an  $\text{error} \leq \tau$ . Hence, for a given  $\tau$  a fraction  $f$  is found, which corresponds to the coefficient  $f_{A(e)}$  in equation B.1. By applying this algorithm to each element of the contemporary formula  $F$ , the respective fractions are found and the adjusted formula  $F_A$  of chemist  $A$  is found (Further details in the Supporting Information). By applying this method it is found, for instance, that contemporary  $\text{Fe}_2\text{O}_3$  corresponds to  $\text{FeO}_3$  according to Berzelius' table of atomic weights of 1819 (Table 10).



**Figure 41:** Fraction of 1868 similarities observed by chemist space with tolerance  $\tau$ , calculated as  $|P_{y-1}^{\tau} \cap P_{1868}| / |P_{1868}|$ , with  $P_{1868}$  collecting the actual similarities by 1868. The 20 similarity values (coloured dots) for each chemist are gathered together in a violin plot. For the sake of comparison the similarity  $|P_{y-1} \cap P_{1868}| / |P_{1868}|$  is depicted as a black dot

**Table 13:** True positives and negatives ( $TP$ ,  $TN$ ), false positives and negatives ( $FP$ ,  $FN$ ), true positive and false positive rates ( $TPR$ ,  $FPR$ ) and the data used for their calculation.

	Meyer 1864	Meyer 1868	Meyer 1869/70	Mendeleev 1869
Number of elements ( $n$ ) in chemist's system of elements	50	52	55	60
Possible number of similarities for $n$ elements: $n \times (n - 1)$	2450	2652	2970	3540
Number of most similar relationships observed from the chemical space in year $y - 1$ , with $y$ the publication year of chemist's system of elements	77	74	74	74
Number of similarities retrieved from chemist's system of elements	586	204	222	403
Number of most similar relationships observed from the chemical space and not retrieved from chemist's system of elements	36	35	28	31
Number of most similar relationships observed from the chemical space and from chemist's system of elements	41	39	46	43
Number of similarities retrieved from chemist's system of elements but not corresponding to most similar relationships from the chemical space in year $y - 1$	545	165	176	360
True positives ( $TP$ )	0.532	0.527	0.622	0.581
True negatives ( $TN$ )	0.77	0.907	0.939	0.896
False positives ( $FP$ )	0.23	0.064	0.061	0.104
False negatives ( $FN$ )	0.468	0.473	0.378	0.419
True positive rate ( $TPR$ )	0.532	0.527	0.622	0.581
False positive rate ( $FPR$ )	0.23	0.066	0.061	0.104



**Figure 42:** Fraction of similarities observed by chemist' space with tolerance  $\tau$  in year  $y - 1$  that are observed in 1868, calculated as  $|P_{y-1}^\tau \cap P_{1868}| / |P_{y-1}^\tau|$ . The 20 similarity values (coloured dots) for each chemist are gathered together in a violin plot. For the sake of comparison the similarity  $|P_{y-1} \cap P_{1868}| / |P_{y-1}|$  is depicted as a black dot.

## Backbone of the system of chemical elements

Figure 13d depicts similarities  $e_i \rightarrow e_j$  that appeared in more than 60% of the systems where the elements  $e_i$  and  $e_j$  are present. The percentage of appearance is computed by  $100 \times f(e_i \rightarrow e_j) / (1868 - y)$ , where  $f(e_i \rightarrow e_j)$  is the number of systems where  $e_i \rightarrow e_j$  is observed, and  $y$  is the first year in which  $e_i$  and  $e_j$  appeared in a system. The normalization factor,  $1868 - y$ , represents the time window where such similarity could have been observed.

## Selection of elements for each chemist

The most complete system was formulated by Mendeleev [49], including the 60 elements of Figure 8 plus Er, Yt and Di [96]. Nonetheless, we excluded these three elements because of their unreliable information by 1869: Yt (currently Y [26]) first reported reaction dates back to 1872. Er and Di were later found to be mixtures of other elements [35].

**Table 14:** Selection of elements for each chemist. The number of initial elements refers to those we used to build up the systems of elements of each chemist (Figure 43). Disregarded elements correspond to those not having substances in the database participating in single step chemical reactions. Notes: *a* We assumed 1 as the atomic weight for H. *b* Meyer reported 50 elements (Figure 43), but it is clear that H made part of his system [122]. *c* Meyer reported 52 elements, here we included H (see note b). *d* Meyer reported 55 elements, here we included H (see note b). *e* Bi, Ce, Mn, Si, W, Y, Zr. *f* Ce, Mn, Mo, Y, Zr. *g* Mo, Y. *h* Y. *i* La

Chemist	Number of initial elements	Number of disregarded elements
Dalton (1810)	37	7 <sup>e</sup>
Thomson (1813)	37	5 <sup>f</sup>
Berzelius (1819)	46	2 <sup>g</sup>
Berzelius (1826)	49	1 <sup>h</sup>
Gmelin (1843)	54	1 <sup>i</sup>
Lenßen (1857) <sup>a</sup>	53	0
Meyer (1864) <sup>a</sup>	51 <sup>b</sup>	0
Odling (1864)	58	0
Hinrichs (1867)	55	0
Meyer (1868)	53 <sup>c</sup>	0
Mendeleev (1869)	60	0
Meyer (1869/70)	56 <sup>d</sup>	0

**Table 15:** Spearman rank correlation for the ordering of elements according to atomic weights by each of the nine nineteenth-century chemists.

	Dalton 1810	Thomson 1813	Berzelius 1819	Berzelius 1826	Gmelin 1843	Lenßen 1857	Meyer 1864	Odling 1864	Hinrichs 1867	Meyer 1868	Mendeleev 1869	Meyer 1869/70	Current
Dalton 1810	1	0.9050	0.8750	0.8907	0.8037	0.9008	0.8779	0.8786	0.8849	0.8817	0.8771	0.8817	0.8711
Thomson 1813	0.9050	1	0.9037	0.8932	0.8322	0.8516	0.8389	0.8430	0.8444	0.8447	0.8440	0.8447	0.8518
Berzelius 1826	0.8750	0.9037	1	0.9561	0.8747	0.9033	0.8903	0.8968	0.8929	0.8938	0.8954	0.8975	0.9047
Berzelius 1826	0.8907	0.8932	0.9561	1	0.8886	0.9436	0.9409	0.9393	0.9370	0.9434	0.9378	0.9437	0.9440
Gmelin 1843	0.8037	0.8322	0.8747	0.8886	1	0.8626	0.8545	0.8512	0.8439	0.8577	0.8469	0.8488	0.8321
Lenßen 1857	0.9008	0.8516	0.9033	0.9436	0.8626	1	0.9454	0.9437	0.9414	0.9478	0.9277	0.9280	0.9056
Meyer 1864	0.8779	0.8389	0.8903	0.9409	0.8545	0.9454	1	1	0.9997	1	0.9750	0.9750	0.9741
Odling 1864	0.8786	0.8430	0.8968	0.9393	0.8512	0.9437	1	1	0.9998	1	0.9676	0.9762	0.9552
Hinrichs 1867	0.8849	0.8444	0.8929	0.9370	0.8439	0.9414	0.9997	0.9998	1	0.9997	0.9639	0.9664	0.9584
Meyer 1868	0.8817	0.8447	0.8938	0.9434	0.8577	0.9478	1	1	0.9997	1	0.9758	0.9758	0.9750
Mendeleev 1869	0.8771	0.8440	0.8954	0.9378	0.8469	0.9277	0.9750	0.9676	0.9639	0.9758	1	0.9946	0.9669
Meyer 1869/70	0.8817	0.8447	0.8975	0.9437	0.8488	0.9280	0.9750	0.9762	0.9664	0.9758	0.9946	1	0.9994
Current	0.8711	0.8518	0.9047	0.9440	0.8321	0.9056	0.9741	0.9552	0.9584	0.9750	0.9669	0.9994	1

H																	
Li	Be										B	C	N	O	F		
Na	Mg										Al	Si	P	S	Cl		
K	Ca		Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br
Rb	Sr		Zr	Nb	Mo		Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	
Cs	Ba			Ta	W		Os	Ir	Pt	Au	Hg	Tl	Pb	Bi			
		La	Ce														
			Th		U												

H																			
Li	Be										B					C	N	O	F
Na	Mg										Al					Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br		
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			
Cs	Ba			Ta	W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi					
		La	Ce																
		Th			U														

H																			
Li	Be										B					C	N	O	F
Na	Mg										Al					Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br		
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			
Cs	Ba			Ta	W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi					
		La	Ce																
		Th			U														

H																	
Li	Be											B	C	N	O	F	
Na	Mg											Al	Si	P	S	Cl	
K	Ca		Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br
Rb	Sr		Zr	Nb	Mo		Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	
Cs	Ba			Ta	W		Os	Ir	Pt	Au	Hg	Tl	Pb	Bi			
		La	Ce														
			Th		U												

H																			
Li	Be										B					C	N	O	F
Na	Mg										Al					Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br		
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			
Cs	Ba			Ta	W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi					
		La	Ce																
		Th		U															

H																	
Li	Be										B	C	N	O	F		
Na	Mg										Al	Si	P	S	Cl		
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	
Cs	Ba			Ta	W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi			
La		Ce															
		Th	U														

Meyer 1864

H																
Li	Be										B		C	N	O	F
Na	Mg										Al		Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn	As		Se	Br	
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I
Cs	Ba	Ta		W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi			
La		Ce														
Th		U														

Odling 1864

H																
Li	Be											B	C	N	O	F
Na	Mg											Al	Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn	As		Se	Br	
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I
Cs	Ba	Ta		W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi			
La		Ce														
				Th		U										

Hinrichs 1867

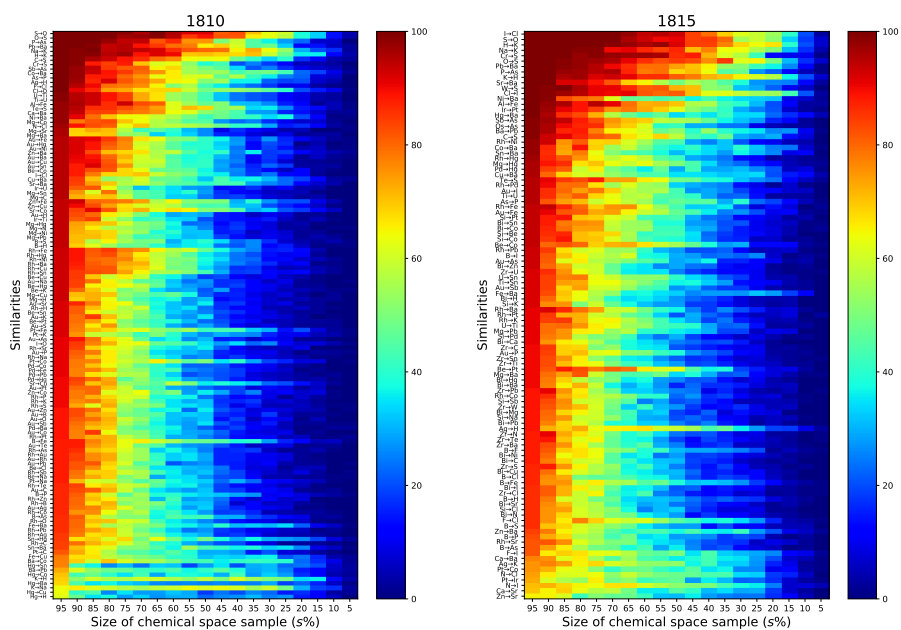
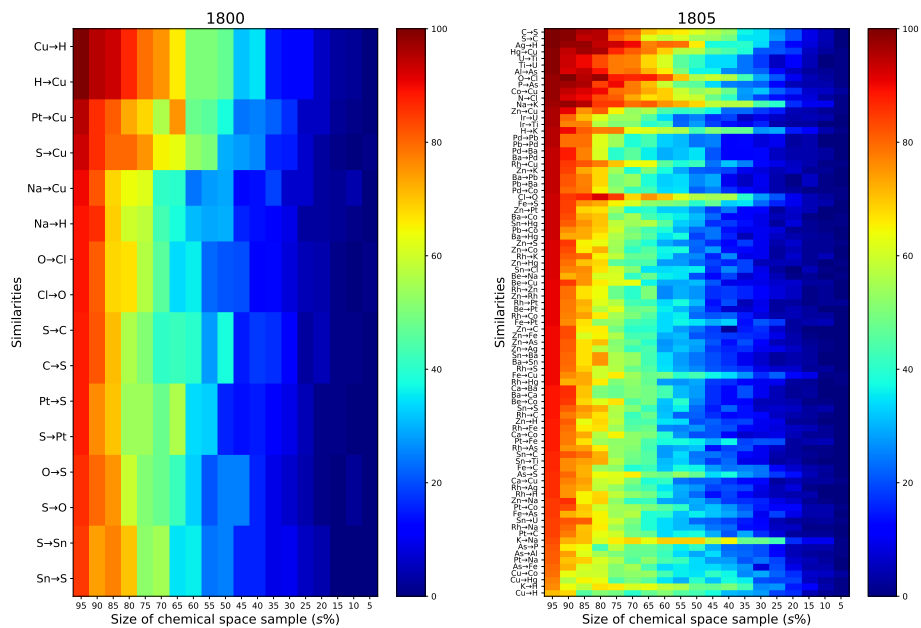
H																
Li	Be											B	C	N	O	F
Na	Mg											Al	Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn	As		Se	Br	
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I
Cs	Ba	Ta		W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi			
		La	Ce													
		Th		U												

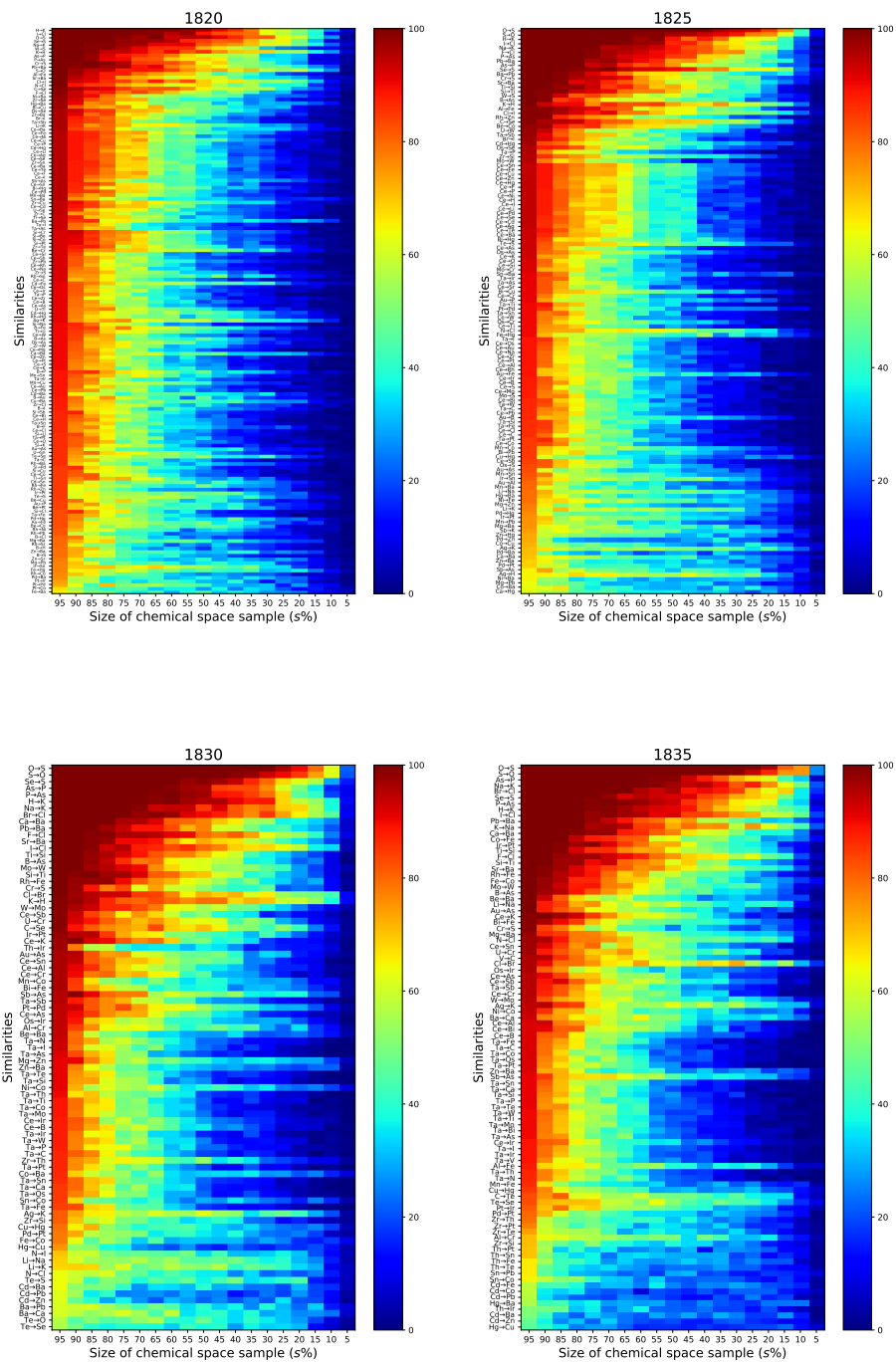
H																			
Li	Be										B					C	N	O	F
Na	Mg										Al					Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br		
Rb	Sr	Zr	Nb	Mo			Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			
Cs	Ba			Ta	W			Os	Ir	Pt	Au	Hg	Tl	Pb	Bi				
		La	Ce																
					Th	U													

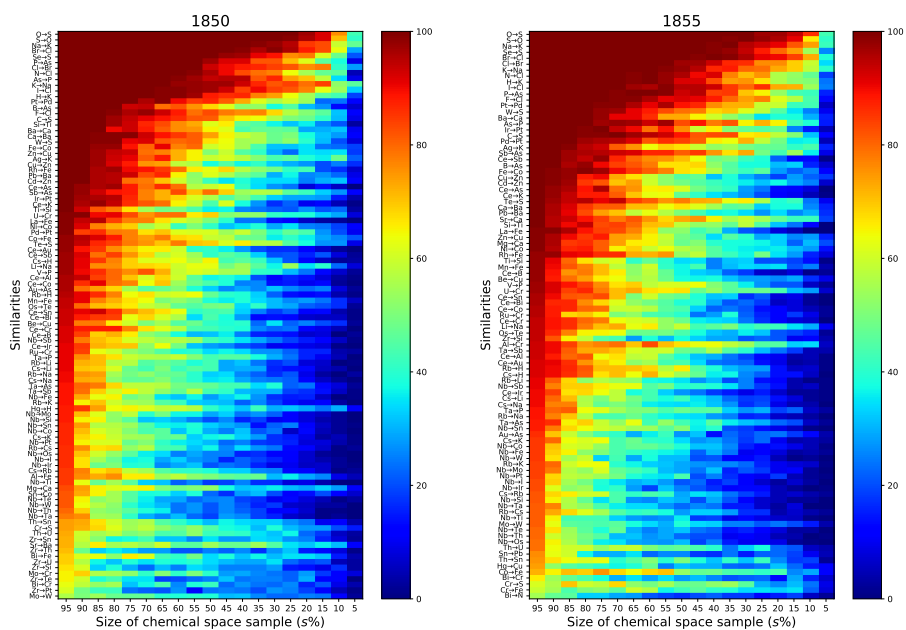
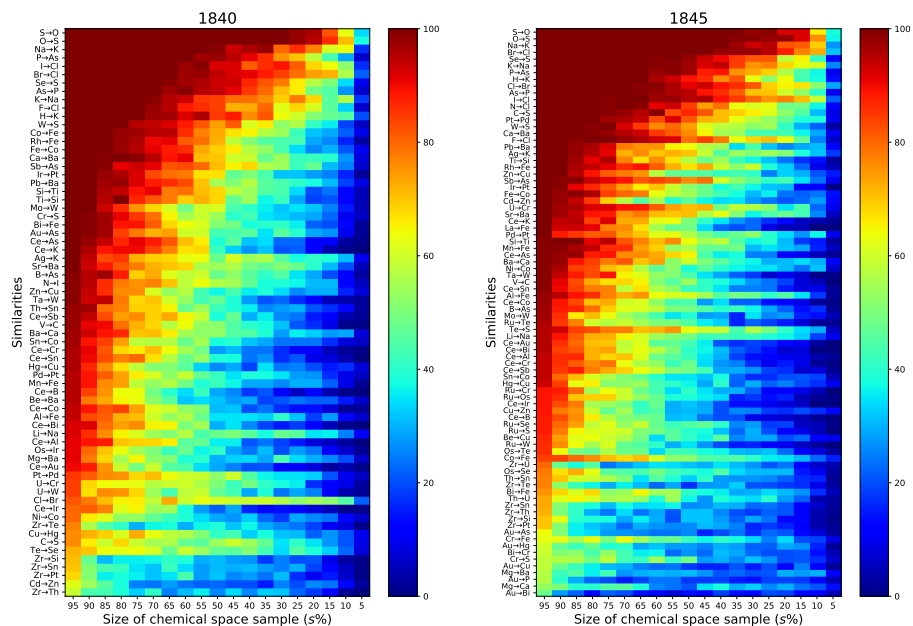
H																
Li	Be										B		C	N	O	F
Na	Mg										Al		Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn	As		Se	Br	
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	I	
Cs	Ba	Ta		W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi			
La		Ce														
Th				U												

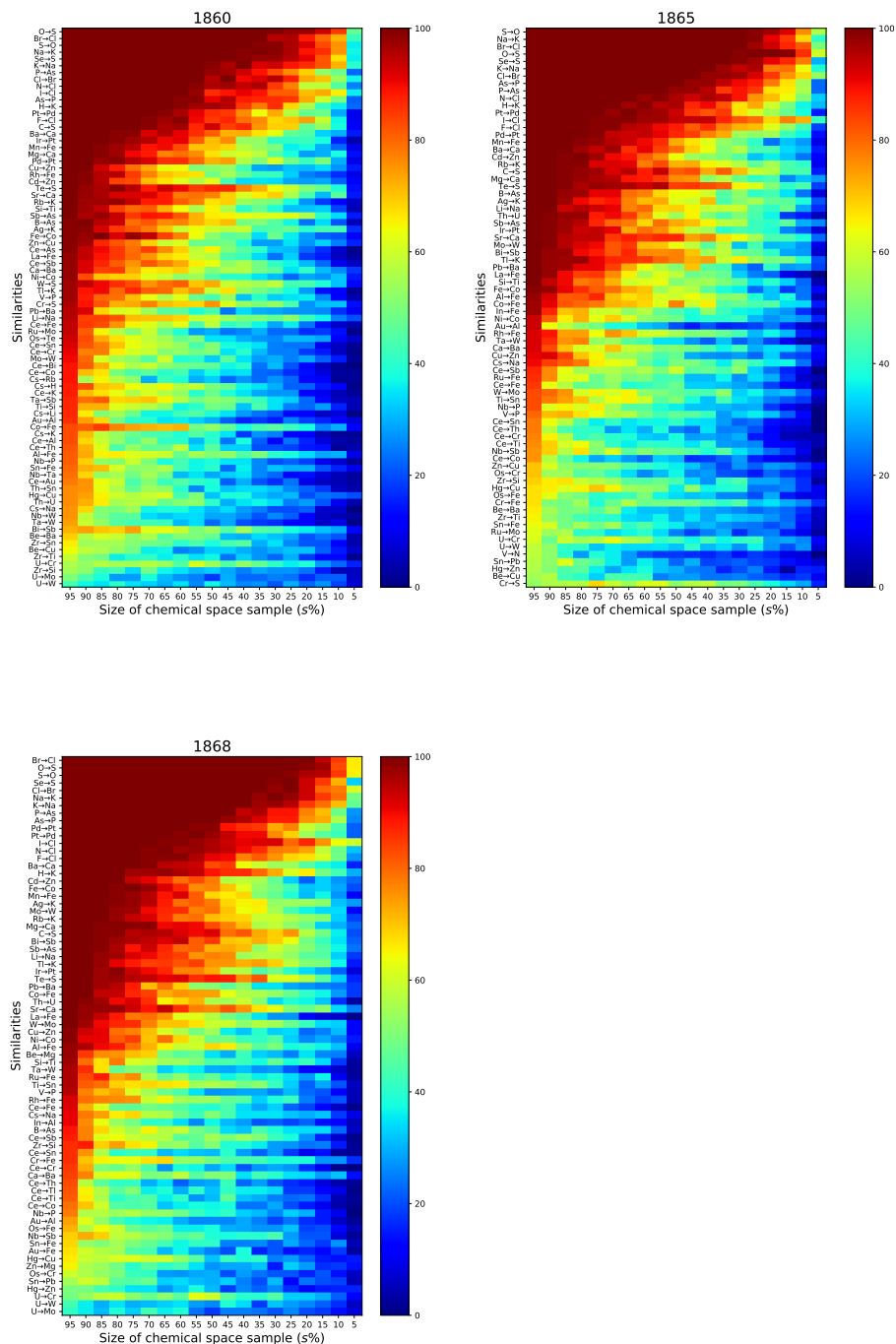
H																			
Li	Be										B					C	N	O	F
Na	Mg										Al					Si	P	S	Cl
K	Ca	Ti		V	Cr	Mn	Fe	Co	Ni	Cu	Zn				As	Se	Br		
Rb	Sr	Zr		Nb	Mo	Ru		Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I			
Cs	Ba			Ta	W	Os		Ir	Pt	Au	Hg	Tl	Pb	Bi					
		La	Ce																
		Th			U														

**Figure 43:** Chemical elements used to build up the systems of elements of the nine nineteenth-century chemists. Elements known by the year discussed in each table are shown in black, while undiscovered elements in grey and in red mixtures that were thought to be elements.









**Figure 44:** Similarity stability over random samples of the chemical spaces within 1800-1868. Samples of 95%, 90%, ..., 5% of the chemical space were considered. Only the results of every five years are shown (the 69 heatplots can be found in [75]). Each row contains a given similarity observed by considering the chemical space in year  $y$ . The stability of each similarity corresponds to the percentage of appearance of such similarity in the sampled space of size  $s\%$ . Colours associated to this percentage are shown on the right bar.



APPENDIX

C



## Supplementary data of Chapter 4

**Table 16:** List of substrates abbreviations appearing in the metabolic network of *Escherichia coli*.

Abbreviation	Chemical name
25aics	(S)-2-[5-Amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxamido]succinate
5aizc	5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate
5aop	5-Amino-4-oxopentanoate
2cpr5p	1-(2-Carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate
6hmhpt	6-hydroxymethyl dihydropterin
5mdr1p	5-Methylthio-5-deoxy-D-ribose 1-phosphate
5mdru1p	5-Methylthio-5-deoxy-D-ribulose 1-phosphate
accoa	Acetyl-CoA
adp	Adenosine diphosphate
adphep_D,D	ADP-D-glycero-D-manno-heptose
adphep_L,D	ADP-L-glycero-D-manno-heptose
asp_L	L-Aspartate
atp	Adenosine triphosphate
co2	Carbon dioxide
cyan	Hydrogen cyanide
dann	7,8-Diaminononanoate
dhnpt	Dihydroneopterin
dtbt	Dethiobiotin
dxy15p	1-deoxy-D-xylulose 5-phosphate
tsul	Thiosulfate
gar	N1-(5-Phospho-D-ribosyl)glycinamide
gcald	Glycolaldehyde
gdiddman	GDP-4-dehydro-6-deoxy-D-mannose
gdpoauc	GDP-4-oxo-L-fucose
glu1sa	L-Glutamate 1-semialdehyde
gly	Glycine
h	H <sup>+</sup>
h2o	Water
hco3	Bicarbonate
malcoa	Malonyl CoA C <sub>24</sub> H <sub>33</sub> N <sub>7</sub> O <sub>19</sub> P <sub>3</sub> S
mmcoa_R	(R)-Methylmalonyl-CoA
mmcoa_S	(S)-Methylmalonyl-CoA
mql8	Menaquinol 8
mqn8	Menaquinone 8
nad	Nicotinamide adenine dinucleotide
nadh	Nicotinamide adenine dinucleotide - reduced
no2	Nitrite
no3	Nitrate
o2	Molecular oxygen
pdx5p	Pyridoxine 5'-phosphate
phthr	O-Phospho-4-hydroxy-L-threonine
pi	Phosphate
pram	5-Phospho-beta-D-ribosylamine
pran	N-(5-Phospho-D-ribosyl)anthranilate
prfp	1-(5-Phosphoribosyl)-5-[(5-phosphoribosylamino)methylideneamino]imidazole-4-carboxamide
prlp	5-[(5-phospho-1-deoxyribulos-1-ylamino)methylideneamino]-1-(5-phosphoribosyl)imidazole-4-carboxamide
q8	Ubiquinone-8
q8h2	Ubiquinol-8
so3	Sulfite
tcynt	Thiocyanate
tsul	Thiosulfate

outdegree—indegree	Substance	Outdegree	Indegree
187324	acetic anhydride	187646	322
144000	methanol	146367	2367
129529	methyl iodide	129937	408
93006	water	97892	4886
88766	ethanol	91036	2270
77857	formaldehyd	81254	3397
74004	benzaldehyde	79578	5574
61658	acetic acid	66120	4462
56275	diazomethane	56365	90
53756	benzoyl chloride	53942	186
51976	aniline	55398	3422
50027	carbon monoxide	53200	3173
41652	benzyl bromide	41935	283
41461	hydrogenchloride	44034	2573
36628	acetyl chloride	36842	214
36022	trifluoroacetic acid	36212	190
35564	benzylamine	36116	552
35283	p-toluenesulfonyl chloride	35392	109
34127	methanesulfonyl chloride	34161	34
33783	acetone	37374	3591
33269	phenylacetylene	33676	407
32917	N,N-dimethyl-formamide	33064	147
32458	allyl bromide	32510	52
32433	morpholine	32659	226
31608	chloro-trimethyl-silane	32450	842

**Table 17:** Substances with the largest positive outdegree—indegree values. These substances belong to the rightmost part of the positive tail of Figure 27. These substances are the most prolific starting materials that are not frequent products of reactions, that is, the best representatives of the "precursors/reactants" role that do not perform the "targets/products" role equally well.

$F(e)$	Frequency (#rxns)	% of reactions	Cummulative %
1	1910725	11.5596	11.5596
0	1426797	8.63191	20.1915
-1	626782	3.79194	23.9834
-2	395965	2.39553	26.379
-3	273380	1.65391	28.0329
-4	206346	1.24836	29.2812
-5	162988	0.986053	30.2673
2	154715	0.936003	31.2033
-6	133341	0.806693	32.01
-7	114459	0.69246	32.7025
-8	98752	0.597435	33.2999
-9	86763	0.524903	33.8248
-10	77768	0.470485	34.2953
-11	69762	0.42205	34.7173
-12	63088	0.381673	35.099
-13	58112	0.351569	35.4506
-14	54023	0.326831	35.7774
-15	49813	0.301361	36.0788
-16	46476	0.281173	36.3599
-17	43792	0.264935	36.6249
-18	41079	0.248522	36.8734
-19	39015	0.236035	37.1094
-187966	36714	0.222114	37.3315
-20	36703	0.222048	37.5536
-21	34902	0.211152	37.7647
-22	33462	0.20244	37.9672
-23	31693	0.191738	38.1589
-24	30234	0.182911	38.3418
-25	29287	0.177182	38.519
-187965	28563	0.172802	38.6918

Table 18: Most frequent values of curvature.

Number of reactants ( $ e_i $ )	Number of reactions	Percentage	cumulative percentage
1	8651	4.61035	4.61035
2	157517	83.945	88.5554
3	16927	9.02085	97.5762
4	2326	1.23959	98.8158
5	1208	0.643776	99.4596
6	231	0.123106	99.5827
7	181	0.0964598	99.6792
8	213	0.113513	99.7927
9	152	0.0810049	99.8737
10	77	0.0410354	99.9147
11	72	0.0383707	99.9531
12	19	0.0101256	99.9632
13	12	0.00639512	99.9696
14	29	0.0154549	99.9851
15	12	0.00639512	99.9915
16	7	0.00373049	99.9952
17	1	0.000532927	99.9957
18	1	0.000532927	99.9963
19	2	0.00106585	99.9973
20	5	0.00266463	100

**Table 19:** Distribution of number of reactants in reactions using acetic anhydride as a substrate.



APPENDIX

D

# Proofs for Chapter 5

*Proof of Proposition 1.* A group is a monoid, thus we only need to check that  $a \multimap b$  actually defines an internal hom and that it respects the ordering.

$$\begin{aligned} b \otimes c &\sqsubseteq a \\ \Leftrightarrow b \otimes c \otimes c^{-1} &\sqsubseteq a \otimes c^{-1} \\ \Leftrightarrow b &\sqsubseteq c \multimap a \end{aligned}$$

$$\begin{aligned} b &\sqsubseteq a \wedge a' \sqsubseteq b' \\ \Rightarrow a^{-1} &\sqsubseteq b^{-1} \wedge a' \sqsubseteq b' \\ \Rightarrow a' \otimes a^{-1} &\sqsubseteq b' \otimes b^{-1} \\ \Rightarrow a \multimap a' &\sqsubseteq b \multimap b' \end{aligned}$$

□

*Proof of Proposition 2.* The identity arrow of an object  $U \xleftarrow{\alpha} X$  in  $M_L\text{Set}$  is given by the pair  $(1_U, 1_X)$  of identities in  $\text{Set}$ . Moreover, given objects  $A = (U \xleftarrow{\alpha} X)$ ,  $B = (V \xleftarrow{\beta} Y)$ , and  $C = (W \xleftarrow{\gamma} Z)$ , and morphisms  $(f, F): A \rightarrow B$  and  $(g, G): B \rightarrow C$ , their composition is computed componentwise as  $(g, G) \circ (f, F) = (g \circ f, F \circ G): A \rightarrow C$ . Notice that  $(g, G) \circ (f, F)$  is a morphism in  $M_L\text{Set}$ : given  $u \in U$  and  $z \in Z$ , we have  $\alpha(u, FGz) \sqsubseteq \beta(fu, Gz) \sqsubseteq \gamma(gfu, z)$ . Associativity and unitality come from associativity and unitality in  $\text{Set}$ . □

*Proof of Proposition 3.* The object  $A \otimes B = (U \times V \xleftarrow{\alpha \otimes \beta} X^V \times Y^U)$  is clearly an object of  $M_L\text{Set}$ . The unit is the object  $I = (1 \xleftarrow{e} 1)$ , which assigns to  $1 \times 1$  the monoidal unit  $e$  of  $L$ . On morphisms  $(f, F): A \rightarrow A'$  and  $(g, G): B \rightarrow B'$ , the monoidal product can be defined as

$$(f, F) \otimes (g, G) = (f \times g, F(-)g \times G(-)f): A \otimes B \rightarrow A' \otimes B'$$

where  $f \times g: U \times V \rightarrow U' \times V'$  and  $F(-)g \times G(-)f: X^{V'} \times Y^{U'} \rightarrow X^V \times Y^U$ . We need to check that the monoidal product is well defined, which means that  $(f, F) \otimes (g, G)$  satisfies the condition on morphisms.

$$\begin{aligned}
 & \alpha \otimes \beta(u, v, (F(-)g \times G(-)f)(f', g')) \\
 &= \alpha \otimes \beta(u, v, Ff'g, Gg'f) \\
 &= \alpha(u, Ff'Gv) \otimes \beta(v, Gg'fu) \\
 &\sqsubseteq \alpha'(fu, f'Gv) \otimes \beta'(gv, g'fu) \\
 &= \alpha' \otimes \beta'(fu, gv, f', g') \\
 &= \alpha' \otimes \beta'((f \times g)(u, v), f', g')
 \end{aligned}$$

The monoidal product is a functor as it preserves composition

$$\begin{aligned}
 & ((f', F') \circ (f, F)) \otimes ((g', G') \circ (g, G)) \\
 &= (f' \circ f, F' \circ F) \otimes (g' \circ g, G' \circ G) \\
 &= ((f' \circ f) \times (g' \circ g), FF'(-)g'g \times GG'(-)f'f) \\
 &= ((f' \times g') \circ (f \times g), (F(-)g \times G(-)f) \circ (F'(-)g' \times G'(-)f')) \\
 &= (f' \times g', F'(-)g' \times G'(-)f') \circ (f \times g, F(-)g \times G(-)f) \\
 &= ((f', F') \otimes (g', G')) \circ ((f, F) \otimes (g, G))
 \end{aligned}$$

and identities

$$\begin{aligned}
 & (1_U, 1_X) \otimes (1_V, 1_Y) \\
 &= (1_U \times 1_V, 1_X(-)1_V \times 1_Y(-)1_U) \\
 &= (1_{U \times V}, 1_{X^V \times Y^U})
 \end{aligned}$$

The associator is defined by the following isomorphisms in Set

$$\alpha_{A,B,C} = (\alpha_{U,V,W}, A_{X,Y,Z}): (A \otimes B) \otimes C \rightarrow A \otimes (B \otimes C)$$

where  $\alpha_{U,V,W}: (U \times V) \times W \rightarrow U \times (V \times W)$  is the associator in Set and  $A_{X,Y,Z}: X^{V \times W} \times (Y^W \times Z^V)^U \rightarrow (X^V \times Y^U)^W \times Z^{U \times V}$  is the composition of isomorphisms in Set given by

$$\begin{aligned}
 X^{V \times W} \times (Y^W \times Z^V)^U &\xrightarrow{\cong} X^{V \times W} \times (Y^{U \times W} \times Z^{U \times V}) \\
 &\xrightarrow{\cong} (X^{V \times W} \times Y^{U \times W}) \times Z^{U \times V} \xrightarrow{\cong} (X^V \times Y^U)^W \times Z^{U \times V}
 \end{aligned}$$

The unitors are defined by the following isomorphisms in Set

$$\lambda_A = (\lambda_U, L_X): I \otimes A \rightarrow A \quad \rho_A = (\rho_U, R_X): A \otimes I \rightarrow A$$

where  $\lambda_U: 1 \times U \rightarrow U$  and  $\rho_U: U \times 1 \rightarrow U$  are the unitors in Set, and  $L_X: X \rightarrow 1^U \times X^1$  and  $R_X: X \rightarrow X^1 \times 1^U$  are the compositions of isomorphisms in Set given by

$$X \xrightarrow{\cong} 1 \times X \xrightarrow{\cong} 1^U \times X^1 \quad X \xrightarrow{\cong} X \times 1 \xrightarrow{\cong} X^1 \times 1^U$$

We are left to prove that the above are actually morphisms in  $M_L\text{Set}$ , that they are natural isomorphisms and that they satisfy the pentagon and triangle equations [89]. The associator is a morphism because for all  $((u, v), w) \in (U \times V) \times W$  and all  $(f, (g, h)) \in X^{V \times W} \times (Y^W \times Z^V)^U$

$$\begin{aligned}
 & ((\alpha \otimes \beta) \otimes \gamma)((u, v), w), A_{X,Y,Z}(f, (g, h))) \\
 &= ((\alpha \otimes \beta) \otimes \gamma)((u, v), w), ((f, g), h)) \\
 &= (\alpha(u, f(v, w)) \otimes \beta(v, g(u, w))) \otimes \gamma(w, h(u, v)) \\
 &= \alpha(u, f(v, w)) \otimes (\beta(v, g(u, w)) \otimes \gamma(w, h(u, v))) \\
 &= (\alpha \otimes (\beta \otimes \gamma))((u, (v, w)), (f, (g, h))) \\
 &= (\alpha \otimes (\beta \otimes \gamma))(\alpha_{U,V,W}((u, v), w), (f, (g, h)))
 \end{aligned}$$

The unitors are morphisms because for all  $u \in U$  and all  $x \in X$

$$\begin{aligned}
 (I \otimes \alpha)((*, u), L_X(x)) & & (\alpha \otimes I)((u, *), R_X(x)) \\
 = (I \otimes \alpha)((*, u), (*, x)) & & = (\alpha \otimes I)((u, *), (x, *)) \\
 = I(*, *) \otimes \alpha(u, x) & & = \alpha(u, x) \otimes I(*, *) \\
 = e \otimes \alpha(u, x) & & = \alpha(u, x) \otimes e \\
 = \alpha(u, x) & & = \alpha(u, x)
 \end{aligned}$$

The associator and the unitors are natural isomorphisms because they are natural isomorphisms component-wise. Finally, the triangle and pentagon equations hold because they hold in  $\text{Set}$ .  $\square$

*Proof of Proposition 4.* The object  $[A, B] = V^U \times X^Y \overset{[\alpha, \beta]}{\dashv} U \times Y$  is clearly an object of  $M_L\text{Set}$ . On morphisms  $(f, F): A' \rightarrow A$  and  $(g, G): B \rightarrow B'$  in  $M_L\text{Set}$ , the internal hom can be defined as

$$[(f, F), (g, G)] = (g(-)f \times F(-)G, f \times G): [A, B] \rightarrow [A', B']$$

where  $g(-)f \times F(-)G: V^U \times X^Y \rightarrow V'^{U'} \times X'^{Y'}$  and  $f \times G: U' \times Y' \rightarrow U \times Y$ . We need to check that the internal hom is well defined, which means that  $[(f, F), (g, G)]$  needs to satisfy the condition on morphisms. For all  $(h, H) \in V^U \times X^Y$  and all  $(u', y') \in U' \times Y'$

$$\begin{aligned}
 & [\alpha, \beta](h, H, (f \times G)(u', y')) \\
 &= [\alpha, \beta](h, H, fu', Gy') \\
 &= \alpha(fu', HGY') \multimap \beta(hfu', Gy') \\
 &\sqsubseteq \alpha'(u', FHGY') \multimap \beta'(ghfu', y') \\
 &= [\alpha', \beta'](ghf, FHG, u', y') \\
 &= [\alpha', \beta']((g(-)f \times F(-)G)(h, H), u', y')
 \end{aligned}$$

because  $\alpha'(u', FHGY') \sqsubseteq \alpha(fu', HGY')$  and  $\beta(hfu', Gy') \sqsubseteq \beta'(ghfu', y')$  as  $(f, F)$  and  $(g, G)$  are

morphisms. The internal hom is a functor as it preserves composition

$$\begin{aligned}
& [(f', F') \circ (f, F), (g', G') \circ (g, G)] \\
&= [(ff', F'F), (g'g, GG')] \\
&= (g'g(-)ff' \times F'F(-)GG', ff' \times GG') \\
&= ((g'(-)f' \times F'(-)G') \circ (g(-)f \times F(-)G), (f \times G) \circ (f' \times G')) \\
&= (g'(-)f' \times F'(-)G', f' \times G') \circ (g(-)f \times F(-)G, f \times G) \\
&= [(f', F'), (g', G')] \circ [(f, F), (g, G)]
\end{aligned}$$

and identities

$$\begin{aligned}
& [\mathbb{1}_A, \mathbb{1}_B] \\
&= [(\mathbb{1}_U, \mathbb{1}_X), (\mathbb{1}_V, \mathbb{1}_Y)] \\
&= (\mathbb{1}_V(-)\mathbb{1}_U \times \mathbb{1}_X(-)\mathbb{1}_Y, \mathbb{1}_U \times \mathbb{1}_Y) \\
&= (\mathbb{1}_{V^U \times X^Y}, \mathbb{1}_{U \times Y}) \\
&= \mathbb{1}_{[A, B]}
\end{aligned}$$

□

*Proof of Theorem 1.* To prove the adjunction  $- \otimes B \dashv [B, -]$  we have to show that, for every objects  $A$  and  $C$  in  $\mathbf{M}_L\mathbf{Set}$ , there is a bijection  $\psi_{A,C}: \mathbf{Hom}_{\mathbf{M}_L\mathbf{Set}}(A \otimes B, C) \cong \mathbf{Hom}_{\mathbf{M}_L\mathbf{Set}}(A, [B, C])$  that is natural in  $A$  and  $C$ .

Let  $\phi_{U,Z}: \mathbf{Hom}_{\mathbf{Set}}(U \times V, Z) \rightarrow \mathbf{Hom}_{\mathbf{Set}}(U, Z^V)$  be the natural isomorphism witnessing the adjunction between the cartesian product and the internal hom in  $\mathbf{Set}$  and let  $\sigma_{U,V}: U \times V \rightarrow V \times U$  be the symmetry of the cartesian product in  $\mathbf{Set}$ . Define the maps

$$\begin{aligned}
\psi_{A,C}(f, F) &= (\langle \phi(f), \phi(\phi^{-1}(F_2) \circ \sigma_{U,Z}) \rangle, \phi^{-1}(F_1) \circ \sigma_{V,Z}) \\
\psi_{A,C}^{-1}(g, G) &= (\phi^{-1}(g_1), \langle \phi(G \circ \sigma_{Z,V}), \phi(\phi^{-1}(g_2) \circ \sigma_{Z,U}) \rangle)
\end{aligned}$$

We can check that they are well defined. An element of  $\mathbf{Hom}_{\mathbf{M}_L\mathbf{Set}}(A \otimes B, C)$  is a pair  $(f, \langle F_1, F_2 \rangle)$  with  $f: U \times V \rightarrow W$  and  $F = \langle F_1, F_2 \rangle: Z \rightarrow X^V \times Y^U$  such that  $\forall (u, v) \in U \times V \forall z \in Z (\alpha \otimes \beta)(u, v, Fz) \sqsubseteq \gamma(f(u, v), z)$ , which is equivalent to  $\alpha(u, (F_1(z))(v)) \otimes \beta(v, (F_2(z))(u)) \sqsubseteq \gamma(f(u, v), z)$ . On the other hand, an element of  $\mathbf{Hom}_{\mathbf{M}_L\mathbf{Set}}(A, [B, C])$  is a pair  $(\langle g_1, g_2 \rangle, G)$  with  $g = \langle g_1, g_2 \rangle: U \rightarrow W^V \times Y^Z$  and  $G: V \times Z \rightarrow X$  such that  $\forall u \in U \forall (v, z) \in V \times Z \alpha(u, G(v, z)) \sqsubseteq [\beta, \gamma](g(u), v, z)$ , which is equivalent to  $\alpha(u, G(v, z)) \sqsubseteq \beta(v, (g_2(u))(z)) \multimap \gamma((g_1(u))(v), z)$ .

They are morphisms because the inequality condition for morphisms in  $\mathbf{M}_L\mathbf{Set}$  holds with equality. We check that they are inverses to each other.

$$\begin{aligned}
& \psi_{A,C} \circ \psi_{A,C}^{-1}(g, G) \\
&= (\langle \phi(\phi^{-1}(g_1)), \phi(\phi^{-1}(\phi(\phi^{-1}(g_2) \circ \sigma_{Z,U})) \circ \sigma_{U,Z}) \rangle, \phi^{-1}(\phi(G \circ \sigma_{Z,V})) \circ \sigma_{V,Z}) \\
&= (\langle g_1, g_2 \rangle, G)
\end{aligned}$$

$$\begin{aligned}
& \psi_{A,C}^{-1} \circ \psi_{A,C}(f, F) \\
&= (\phi^{-1}(\phi(f)), \langle \phi(\phi^{-1}(F_1) \circ \sigma_{V,Z} \circ \sigma_{Z,V}), \phi(\phi^{-1}(\phi(\phi^{-1}(F_2) \circ \sigma_{U,Z})) \circ \sigma_{Z,U}) \rangle) \\
&= (f, \langle F_1, F_2 \rangle)
\end{aligned}$$

We check that they are natural. Naturality comes from naturality of  $\phi$  in Set.  $\square$

*Proof of Lemma 1.* The ordering of the poset  $\sqsubseteq$  is given by the usual ordering on the integers,  $a \sqsubseteq b \Leftrightarrow a \leq b$ . The conjunction is associative and unital, with 1 as unit, so it is a monoid operation on  $\mathbb{3}$ . Moreover, it respects the ordering and this makes  $\mathbb{3}$  a partially ordered monoid. We are left to check whether the internal hom is right adjoint to the conjunction, namely, whether  $b \otimes c \leq a \Leftrightarrow b \leq c \multimap a$ . By the definition of internal hom,

$$\begin{aligned} b &\leq c \multimap a \\ \Leftrightarrow b &\leq \max\{x : x \otimes c \leq a\} \\ \Leftrightarrow b \otimes c &\leq a \end{aligned}$$

$\square$

*Proof of Lemma 2.* The ordering of the poset  $\sqsubseteq$  is given by the opposite of the usual ordering on the natural numbers,  $a \sqsubseteq b \Leftrightarrow a \geq b$ . The conjunction is associative and unital, with 0 as unit, so it is a monoid operation on  $\mathbb{N}$ . Moreover, it respects the ordering and this makes  $\mathbb{N}$  a partially ordered monoid. We are left to check whether the internal hom is right adjoint to the conjunction, namely, whether  $b \otimes c \sqsubseteq a \Leftrightarrow b \sqsubseteq c \multimap a$ . By the definition of internal hom,

$$\begin{aligned} b &\sqsubseteq c \multimap a \\ \Leftrightarrow b &\geq \max\{a - c, 0\} \\ \Leftrightarrow b &\geq a - c \wedge b \geq 0 \\ \Leftrightarrow b + c &\geq a \\ \Leftrightarrow b \otimes c &\sqsubseteq a \end{aligned}$$

$\square$

*Proof of Lemma 3.* The image of the product of real numbers in the closed interval  $[0, 1]$  is the closed interval itself. This product is associative and unital, thus it gives a monoid structure to  $[0, 1]$ . Moreover, it preserves the ordering and this makes  $[0, 1]$  a partially ordered monoid. We need to check that truncated division as defined above gives an internal hom.

If $c \neq 0 \wedge c \geq a$ $b \sqsubseteq c \multimap a$ $\Leftrightarrow b \leq \frac{a}{c}$ $\Leftrightarrow b \cdot c \leq a$ $\Leftrightarrow b \otimes c \sqsubseteq a$	If $c = 0$ $b \sqsubseteq c \multimap a$ $\Leftrightarrow b \leq 1$ $\Leftrightarrow 0 \leq a$ $\Leftrightarrow b \cdot 0 \leq a$ $\Leftrightarrow b \otimes c \sqsubseteq a$	If $c < a$ $b \sqsubseteq c \multimap a$ $\Leftrightarrow b \leq 1$ $\Leftrightarrow b \leq 1 < \frac{a}{c}$ $\Leftrightarrow b \cdot c \leq a$ $\Leftrightarrow b \otimes c \sqsubseteq a$
---	--	--

$\square$

*Proof of Proposition 5.*  $(L_1 \times L_2, \otimes, e)$  is the cartesian product of two monoids and therefore it is a monoid.  $(L_1 \times L_2, \leq)$  is a partial ordered set with the ordering defined above. Since  $l_i \leq l'_i$  implies both  $l_i * k_i \leq l'_i * k_i$  and  $k_i * l_i \leq k_i * l'_i$  for each  $k_i \in L_i$  and  $i \in 1, 2$ ; then  $l \leq l'$  implies  $l \otimes k \leq l' \otimes k$  and  $k \otimes l \leq k \otimes l'$  for every  $k = (k_1, k_2) \in L_1 \times L_2$ . This proves that  $(L_1 \times L_2, \leq, \otimes, e, \multimap)$  is a partially ordered monoid. We need to prove that the internal hom defined above is right adjoint to the monoidal product.

$$\begin{aligned}
& b \otimes c \leq a \\
& \Leftrightarrow (b_1 \otimes_1 c_1, b_2 \otimes_2 c_2) \leq (a_1, a_2) \\
& \Leftrightarrow b_1 \otimes_1 c_1 \leq_1 a_1 \wedge b_2 \otimes_2 c_2 \leq_2 a_2 \\
& \Leftrightarrow b_1 \leq_1 c_1 \multimap_1 a_1 \wedge b_2 \leq_2 c_2 \multimap_2 a_2 \\
& \Leftrightarrow (b_1, b_2) \leq (c_1 \multimap_1 a_1, c_2 \multimap_2 a_2) \\
& \Leftrightarrow b \leq c \multimap a
\end{aligned}$$

This proves that the product of lineales is again a lineale. □





## Bibliography

- [1] Albert, R., H. Jeong, and A.-L. Barabási. "Diameter of the World-Wide Web". In: *Nature* 401.6749 (Sept. 1999), pp. 130–131. *ISSN*: 1476-4687. *DOI*: 10.1038/43601.
- [2] Andersen, J. L., C. Flamm, D. Merkle, and P. F. Stadler. "An intermediate level of abstraction for computational systems chemistry". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 375.2109 (2017), p. 20160354. *DOI*: 10.1098/rsta.2016.0354. *URL*: <https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2016.0354>.
- [3] Andersen, J. L., C. Flamm, D. Merkle, and P. F. Stadler. "Maximizing output and recognizing autocatalysis in chemical reaction networks is NP-complete". In: *J. Syst. Chem.* 3 (2012), p. 1. *DOI*: 10.1186/1759-2208-3-1.
- [4] Axmann, I. M., S. Legewie, and H. Herzel. "A minimal circadian clock model". In: *Genome Informatics 2007: Genome Informatics Series Vol. 18*. World Scientific, 2007, pp. 54–64.
- [5] Baez, J. C., F. Genovese, J. Master, and M. Shulman. "Categories of Nets". In: *arXiv preprint arXiv:2101.04238* (2021).
- [6] Baez, J. C. and J. Master. "Open Petri nets". In: *Mathematical Structures in Computer Science* 30.3 (2020), pp. 314–341.
- [7] Barabási, A.-L. and R. Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (1999), pp. 509–512. *ISSN*: 0036-8075. *DOI*: 10.1126/science.286.5439.509.
- [8] Barbosa, V. C., R. Donangelo, and S. R. Souza. "Coevolution of the mitotic and meiotic modes of eukaryotic cellular division". In: *Phys. Rev. E* 98 (3 2018), p. 032409. *DOI*: 10.1103/PhysRevE.98.032409. *URL*: <https://link.aps.org/doi/10.1103/PhysRevE.98.032409>.
- [9] Benkő, G., C. Flamm, and P. F. Stadler. "A Graph-Based Toy Model of Chemistry". In: *Journal of Chemical Information and Computer Sciences* 43.4 (July 2003), pp. 1085–1093. *ISSN*: 0095-2338. *DOI*: 10.1021/ci0200570.
- [10] Bensaude-Vincent, B. "Mendeleev's periodic system of chemical elements". In: *The British Journal for the History of Science* 19.1 (1986), pp. 3–17. *DOI*: 10.1017/S000708740002272X.
- [11] Berge, C. *Hypergraphs : combinatorics of finite sets*. eng. Amsterdam ; New York : North Holland : Distributors for the U.S.A. and Canada, 1989. *ISBN*: 0444874895.
- [12] Bernadette, B.-V. "Languages in chemistry". In: *The Cambridge History of Science*. Ed. by N. Mary Jo. Vol. 5. Cambridge: Cambridge University Press, 2003, pp. 174–190.

- [13] Bernal, A. and E. Daza. "Metabolic networks: beyond the graph". In: *Curr. Comput. Aided Drug Des.* 7 (2011), pp. 122–132. DOI: 10.2174/157340911795677611.
- [14] Berzelius, J. J. "Ueber die Bestimmung der relativen Anzahl von einfachen Atomen in chemischen Verbindungen". In: *Annalen der Physik* 83.8 (1826), pp. 397–416. DOI: <https://doi.org/10.1002/andp.18260830802>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.18260830802>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18260830802>.
- [15] Berzelius, J. J. "Ueber die Bestimmung der relativen Anzahl von einfachen Atomen in chemischen Verbindungen". In: *Annalen der Physik* 84.9 (1826), pp. 1–24. DOI: <https://doi.org/10.1002/andp.18260840902>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.18260840902>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18260840902>.
- [16] Berzelius, J. J. "Ueber die Bestimmung der relativen Anzahl von einfachen Atomen in chemischen Verbindungen". In: *Annalen der Physik* 84.10 (1826), pp. 177–190. DOI: <https://doi.org/10.1002/andp.18260841006>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/andp.18260841006>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/andp.18260841006>.
- [17] Berzelius, J. J. *Essai sur la théorie des proportions chimiques et sur l'influence chimique de l'électricité*. Paris, 1819.
- [18] Bretto, A. "Dirhypergraphs: Basic Concepts". In: *Hypergraph Theory: An Introduction*. Heidelberg: Springer International Publishing, 2013, pp. 95–110. ISBN: 978-3-319-00080-0. DOI: 10.1007/978-3-319-00080-0\_6.
- [19] Bretto, A. *Hypergraph Theory: An Introduction*. Springer Publishing Company, Incorporated, 2013. ISBN: 3319000799, 9783319000794.
- [20] Brown, C. and D. Gurr. "A Categorical Linear Framework for Petri Nets". In: *Information and Computation* 122.2 (1995), pp. 268–285. ISSN: 0890-5401. DOI: <https://doi.org/10.1006/inco.1995.1150>.
- [21] Brown, C. and D. Gurr. "A categorical linear framework for Petri nets". In: *[1990] Proceedings. Fifth Annual IEEE Symposium on Logic in Computer Science*. IEEE. 1990, pp. 208–218.
- [22] Brown, C., D. Gurr, and V. de Paiva. *A Linear Specification Language for Petri Nets (Aarhus Technical Report)*. 1991.
- [23] Bruni, R., J. Meseguer, U. Montanari, and V. Sassone. "Functorial models for Petri nets". In: *Information and Computation* 170.2 (2001), pp. 207–236.
- [24] Castelvechi, D. "Physics paper sets record with more than 5,000 authors". In: *Nature* (May 2015). ISSN: 1476-4687. DOI: 10.1038/nature.2015.17567.
- [25] Catanese, S., P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti. "Extraction and Analysis of Facebook Friendship Relations". In: *Computational Social Networks: Mining and Visualization*. Ed. by A. Abraham. London: Springer London, 2012, pp. 291–324. ISBN: 978-1-4471-4054-2. DOI: 10.1007/978-1-4471-4054-2\_12.

- [26] Coplen, T. B. and H. S. Peiser. "History of the recommended atomic-weight values from 1882 to 1997: A comparison of differences from current values to the estimated uncertainties of earlier values (Technical Report)". In: *Pure and Applied Chemistry* 70 (1 2019), pp. 237–257. DOI: 10.1351/pac199870010237.
- [27] Dalton, J. *A new system of chemical philosophy, part 2*. Manchester, 1810.
- [28] Dobson, C. M. "Chemical space and biology". In: *Nature* 432.7019 (Dec. 2004), pp. 824–828. ISSN: 1476-4687. DOI: 10.1038/nature03192.
- [29] Ehrhard, T. and O. Laurent. "Interpreting a finitary pi-calculus in differential interaction nets". In: *Information and Computation* 208.6 (2010). Special Issue: 18th International Conference on Concurrency Theory (CONCUR 2007), pp. 606–633. ISSN: 0890-5401. DOI: <https://doi.org/10.1016/j.ic.2009.06.005>.
- [30] Eidi, M., A. Farzam, W. Leal, A. Samal, and J. Jost. "Edge-based analysis of networks: curvatures of graphs and hypergraphs". In: *Theory in Biosciences* 139.4 (2020), pp. 337–348. ISSN: 1611-7530. DOI: 10.1007/s12064-020-00328-0. URL: <https://doi.org/10.1007/s12064-020-00328-0>.
- [31] Eidi, M., A. Farzam, W. Leal, A. Samal, and J. Jost. "Edge-based analysis of networks : curvatures of graphs and hypergraphs". In: *Theory in biosciences* 139.4 (2020), pp. 337–348. ISSN: 1431-7613. DOI: 10.1007/s12064-020-00328-0.
- [32] Eidi, M. and J. Jost. "Ollivier Ricci curvature of directed hypergraphs". In: *Sci. Rep.* 10.1 (July 2020), p. 12466. ISSN: 2045-2322. DOI: 10.1038/s41598-020-68619-6. URL: <https://doi.org/10.1038/s41598-020-68619-6>.
- [33] Emzivat, Y., B. Delahaye, D. Lime, and O. H. Roux. "Probabilistic time Petri nets". In: *International Conference on Application and Theory of Petri Nets and Concurrency*. Springer. 2016, pp. 261–280.
- [34] Estrada, E. and J. A. Rodríguez-Velázquez. "Subgraph centrality and clustering in complex hyper-networks". In: *Physica A* 364 (2006), pp. 581–594. ISSN: 0378-4371. DOI: 10.1016/j.physa.2005.12.002.
- [35] Evans, C. *Episodes from the history of the rare earth elements*. Boston : Kluwer Academic Publishers, 1996. ISBN: 978-94-009-0287-9. DOI: 10.1007/978-94-009-0287-9\_3.
- [36] Fagerberg, R., C. Flamm, D. Merkle, P. Peters, and P. F. Stadler. "On the Complexity of Reconstructing Chemical Reaction Networks". In: *Mathematics in Computer Science* 7.3 (Sept. 2013), pp. 275–292. ISSN: 1661-8289. DOI: 10.1007/s11786-013-0160-y.
- [37] Farey, J. "On a curious Property of vulgar Fractions". In: *The Philosophical Magazine and Journal* 47 (1816), pp. 385–386.
- [38] Farzam, A., A. Samal, and J. Jost. "Degree difference: a simple measure to characterize structural heterogeneity in complex networks". In: *Sci. Rep.* 10.1 (Dec. 2020), p. 21348. ISSN: 2045-2322. DOI: 10.1038/s41598-020-78336-9.
- [39] Fialkowski, M., K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, and B. A. Grzybowski. "Architecture and Evolution of Organic Chemistry". In: *Angew Chem Int Ed Eng* 44.44 (2005), pp. 7263–7269. DOI: 10.1002/anie.200502272. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200502272>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.200502272>.

- [40] Flamm, C., B. M. R. Stadler, and P. F. Stadler. "Generalized Topologies: Hypergraphs, Chemical Reactions, and Biological Evolution". In: *Advances in Mathematical Chemistry and Applications*. Ed. by S. C. Basak, G. Restrepo, and J. L. Villaveces. Bentham-Elsevier, 2015, pp. 300–328. ISBN: 978-1-68108-053-6.
- [41] Forman, R. "Bochner's Method for Cell Complexes and Combinatorial Ricci Curvature". In: *Discrete Comput. Geom.* 29.3 (2003), pp. 323–374. ISSN: 1432-0444. DOI: 10.1007/s00454-002-0743-x. URL: <https://doi.org/10.1007/s00454-002-0743-x>.
- [42] Friedman, R. M. *The politics of excellence. Behind the Nobel prize in science*. Times Books, 2001.
- [43] Gallo, G., G. Longo, S. Pallottino, and S. Nguyen. "Directed hypergraphs and applications". In: *Discrete Appl. Math.* 42.2 (1993), pp. 177–201. ISSN: 0166-218X. DOI: [https://doi.org/10.1016/0166-218X\(93\)90045-P](https://doi.org/10.1016/0166-218X(93)90045-P). URL: <http://www.sciencedirect.com/science/article/pii/0166218X9390045P>.
- [44] Gautschi, W. "Error function and Fresnel integrals". In: *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Ed. by M. Abramowitz and I. A. Stegun. Dover, 1972, p. 299.
- [45] Girard, J.-Y. "Linear logic". In: *Theoretical computer science* 50.1 (1987), pp. 1–101.
- [46] Gmelin, L. *Handbuch der Chemie*. Vol. 7, in 2 pts. Vol.8 published in 1858. Heidelberg: Karl Winter, 1843, 8 v. in 9. URL: [http://hdl.handle.net/2027/uc1.b4059118%20\(v.3\)](http://hdl.handle.net/2027/uc1.b4059118%20(v.3)).
- [47] Gordin, M. D. *A well-ordered thing: Dmitrii Mendeleev and the shadow of the periodic table*. New York: Basic Books, 2004. ISBN: 0-465-02775-X.
- [48] Gordin, M. D. "The organic roots of Mendeleev's periodic law". In: *Historical Studies in the Physical and Biological Sciences* 32 (2 2002), pp. 263–290. DOI: 10.1525/hsps.2002.32.2.263.
- [49] Gordin, M. D. "The textbook case of a priority dispute: D. I. Mendeleev, Lothar Meyer, and the periodic system". In: *Nature Engaged: Science in Practice from the Renaissance to the Present*. Ed. by M. Biagioli and J. Riskin. New York: Palgrave Macmillan US, 2012, pp. 59–82. ISBN: 978-0-230-33802-9. DOI: 10.1057/9780230338029\_4. URL: [https://doi.org/10.1057/9780230338029\\_4](https://doi.org/10.1057/9780230338029_4).
- [50] Hammack, R. H. "Proof of a conjecture concerning the direct product of bipartite graphs". In: *European Journal of Combinatorics* 30.5 (2009). Part Special Issue on Metric Graph Theory, pp. 1114–1118. ISSN: 0195-6698. DOI: <https://doi.org/10.1016/j.ejc.2008.09.015>.
- [51] Harary, F. *Graph Theory*. Reading, Massachusetts: Addison-Wesley, 1969.
- [52] Held, H., A. Rengstl, and D. Mayer. "Acetic Anhydride and Mixed Fatty Acid Anhydrides". In: *Ullmann's Encyclopedia of Industrial Chemistry*. American Cancer Society, 2000. ISBN: 9783527306732. DOI: 10.1002/14356007.a01\_065.
- [53] Higuchi, Y. "Combinatorial curvature for planar graphs". In: *J. Graph Theory* 38.4 (2001), pp. 220–229. DOI: <https://doi.org/10.1002/jgt.10004>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jgt.10004>.
- [54] Hinrichs, G. *Atomechanik*. Iowa, 1867.

- [55] Hinze, T., T. Lenser, G. Escuela, I. Heiland, and S. Schuster. "Modelling signalling networks with incomplete information about protein activation states: a p system framework of the KaiABC oscillator". In: *International Workshop on Membrane Computing*. Springer. 2009, pp. 316–334.
- [56] Hoffmann, R. W. "Introduction". In: *Elements of Synthesis Planning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-79220-8. DOI: 10.1007/978-3-540-79220-8\_1. URL: [https://doi.org/10.1007/978-3-540-79220-8\\_1](https://doi.org/10.1007/978-3-540-79220-8_1).
- [57] Imrich, W. and S. Klavžar. *Product graphs, structure and recognition*. New York: Wiley, 2000.
- [58] Jamshidi, N. and B. Ø. Palsson. "Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ 661 and proposing alternative drug targets". In: *BMC Syst. Biol* 1.1 (June 2007), p. 26. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-26.
- [59] Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási. "The large-scale organization of metabolic networks". In: *Nature* 407 (2000), pp. 651–654.
- [60] Katis, P., N. Sabadini, and R. F. Walters. "Representing place transition nets in Span(Graph)". In: *International Conference on Algebraic Methodology and Software Technology*. Springer. 1997, pp. 322–336.
- [61] Kelly, M. *Basic Concepts of Enriched Category Theory*, vol. 64. Cambridge University Press, Lecture Notes in Mathematics. Republished in Reprints in Theory and Applications of Categories, No. 10 (2005) pp. 1-136., 1982.
- [62] Keserű, G. M., T. Soós, and C. O. Kappe. "Anthropogenic reaction parameters – the missing link between chemical intuition and the available chemical space". In: *Chem. Soc. Rev.* 43 (15 2014), pp. 5387–5399. DOI: 10.1039/C3CS60423C. URL: <http://dx.doi.org/10.1039/C3CS60423C>.
- [63] Kirkpatrick, P. and C. Ellis. "Chemical space". In: *Nature* 432.7019 (Dec. 2004), pp. 823–823. ISSN: 1476-4687. DOI: 10.1038/432823a.
- [64] Klamt, S., U.-U. Haus, and F. Theis. "Hypergraphs and Cellular Networks". In: *PLOS Comput. Biol.* 5.5 (2009), pp. 1–6. DOI: 10.1371/journal.pcbi.1000385. URL: <https://doi.org/10.1371/journal.pcbi.1000385>.
- [65] Klein, U. *Experiments, Models, Paper Tools: Cultures of Organic Chemistry in the Nineteenth Century*. Stanford: Stanford University Press, 2003, p. 320. URL: <http://www.sup.org/books/title/?id=1917>.
- [66] Kock, J. "Elements of Petri nets and processes". In: *arXiv preprint arXiv:2005.05108* (2020).
- [67] Lambiotte, R., M. Rosvall, and I. Scholtes. "From networks to optimal higher-order models of complex systems". In: *Nature Physics* 15.4 (Apr. 2019), pp. 313–320. ISSN: 1745-2481. DOI: 10.1038/s41567-019-0459-y. URL: <https://doi.org/10.1038/s41567-019-0459-y>.
- [68] Laubichler, M. D., J. Maienschein, and J. Renn. "Computational Perspectives in the History of Science: To the Memory of Peter Damerow". In: *Isis* 104.1 (2013), pp. 119–130. DOI: 10.1086/669891. eprint: <https://doi.org/10.1086/669891>. URL: <https://doi.org/10.1086/669891>.
- [69] Lavore, E. D., W. Leal, and V. de Paiva. "Dialectica Petri nets". ArXiv. (Submitted). 2021.

- [70] Lawson, A. J., J. Swienty-Busch, T. Géoui, and D. Evans. "The making of Reaxys - Towards unobstructed access to relevant chemistry information". In: *The future of the history of chemical information*. 2014. Chap. 8, pp. 127–148. DOI: 10.1021/bk-2014-1164.ch008. eprint: <https://pubs.acs.org/doi/pdf/10.1021/bk-2014-1164.ch008>. URL: <https://pubs.acs.org/doi/abs/10.1021/bk-2014-1164.ch008>.
- [71] Lawvere, F. W. "Metric spaces, generalized logic, and closed categories". In: *Rendiconti del seminario matematico e fisico di Milano* 43.1 (1973), pp. 135–166.
- [72] Leal, W., M. Eidi, and J. Jost. "Curvature-based analysis of directed hypernetworks". In: *Complex networks 2019: the 8th international conference on complex networks and their applications; December 10 - 12, 2019 Lisbon, Portugal; book of abstracts*. Ed. by H. Cherifi. [s.l.]: International conference on complex networks and their applications, 2019, pp. 32–34. ISBN: 978-2-9557050-3-2.
- [73] Leal, W., M. Eidi, and J. Jost. "Ricci curvature of random and empirical directed hypernetworks". In: *Applied network science* 5.1 (2020), p. 65. ISSN: 2364-8228. DOI: 10.1007/s41109-020-00309-8.
- [74] Leal, W., E. J. Llanos, A. Bernal, P. F. Stadler, J. Jost, and G. Restrepo. "The expansion of chemical space in 1826 and in the 1840s prompted the convergence to the periodic system". (Accepted in *Proceedings of the National Academy of Sciences of the United States of America*). 2022.
- [75] Leal, W., E. J. Llanos, P. F. Stadler, J. Jost, and G. Restrepo. "The chemical space from which the periodic system arose". In: (Aug. 2019). DOI: 10.26434/chemrxiv.9698888.v1. URL: [https://chemrxiv.org/articles/preprint/The\\_Chemical\\_Space\\_from\\_Which\\_the\\_Periodic\\_System\\_Arose/9698888](https://chemrxiv.org/articles/preprint/The_Chemical_Space_from_Which_the_Periodic_System_Arose/9698888).
- [76] Leal, W. and G. Restrepo. "Formal structure of periodic system of elements". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475.2224 (2019), p. 20180581. ISSN: 1364-5021. DOI: 10.1098/rspa.2018.0581.
- [77] Leal, W., G. Restrepo, and A. Bernal. "A network study of chemical elements: from binary compounds to chemical trends". In: *MATCH communications in mathematical and in computer chemistry* 68.1 (2012), pp. 417–442.
- [78] Leal, W., G. Restrepo, P. F. Stadler, and J. Jost. "Forman-Ricci curvature for hypergraphs". In: *Advances in Complex Systems* 24.01 (2021), p. 2150003. ISSN: 0219-5259. DOI: 10.1142/S021952592150003X.
- [79] Lenßen, E. "Ueber die Gruppierung der Elemente nach ihrem chemisch - physikalischen Character". In: *Justus Liebigs Annalen der Chemie* 103.2 (1857), pp. 121–131. DOI: <https://doi.org/10.1002/jlac.18571030202>. eprint: <https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/jlac.18571030202>. URL: <https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/10.1002/jlac.18571030202>.
- [80] Leskovec, J. *Wikipedia vote network*. 2017. URL: <https://snap.stanford.edu/data/wiki-Vote.html>.
- [81] Leskovec, J., D. Huttenlocher, and J. Kleinberg. "Governance in Social Media: A Case Study of the Wikipedia Promotion Process". In: *Proc. Int. Conf. on Weblogs and Social Media*. Menlo Park: AAAI, 2010, arXiv: 1004.3547.

- [82] Lilienfeld, O. A. von, K.-R. Müller, and A. Tkatchenko. "Exploring chemical compound space with quantum-based machine learning". In: *Nature Reviews Chemistry* 4.7 (July 2020), pp. 347–358. *ISSN*: 2397-3358. *DOI*: 10.1038/s41570-020-0189-9.
- [83] Llanos, E. J., W. Leal, A. Bernal, G. Restrepo, J. Jost, and P. F. Stadler. "A Network model of the Chemical Space provides similarity structure to the system of chemical elements". In: *Complex networks 2019: the 8th international conference on complex networks and their applications; December 10 - 12, 2019 Lisbon, Portugal; book of abstracts*. Ed. by H. Cherifi. [s.l.]: International conference on complex networks and their applications, 2019, pp. 308–310. *ISBN*: 978-2-9557050-3-2.
- [84] Llanos, E. J., W. Leal, D. H. Luu, J. Jost, P. F. Stadler, and G. Restrepo. "Exploration of the chemical space and its three historical regimes". In: *Proc. Natl. Acad. Sci. U.S.A.* 116.26 (2019), pp. 12660–12665. *ISSN*: 0027-8424. *DOI*: 10.1073/pnas.1816039116.
- [85] Llanos, E. J., W. Leal, G. Restrepo, and P. F. Stadler. "Computational approach to the history of chemical reactivity: Exploring Reaxys database". In: *Abstracts of papers of the American Chemical Society* 254 (2017).
- [86] Lohmann, G., E. Lacosse, T. Ethofer, V. J. Kumar, K. Scheffler, and J. Jürgen. *Predicting intelligence from fMRI data of the human brain in a few minutes of scan time*. Tech. rep. 2021.03.18.435935. bioRxiv, 2021. *DOI*: 10.1101/2021.03.18.435935.
- [87] Lu, Z. Q. and L. M. Berliner. "Markov switching time series models with application to a daily runoff series". In: *Water Resour Res* 35.2 (1999), pp. 523–534. *DOI*: 10.1029/98WR02686. eprint: <https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/98WR02686>. *URL*: <https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98WR02686>.
- [88] Lyday, P. A. and T. Kaiho. "Iodine and Iodine Compounds". In: *Ullmann's Encyclopedia of Industrial Chemistry*. American Cancer Society, 2015, pp. 1–13. *ISBN*: 9783527306732. *DOI*: 10.1002/14356007.a14\_381.pub2.
- [89] MacLane, S. *Categories for the Working Mathematician*. Graduate Texts in Mathematics, Vol. 5. New York: Springer-Verlag, 1971, pp. ix+262.
- [90] Master, J. "Generalized Petri Nets". In: *arXiv preprint arXiv:1904.09091* (2019).
- [91] Master, J. "Petri nets based on Lawvere theories". In: *Mathematical Structures in Computer Science* 30.7 (2020), pp. 833–864.
- [92] McKie, D. "Wöhler's 'Synthetic' Urea and the Rejection of Vitalism: A Chemical Legend". In: *Nature* 153.3890 (May 1944), pp. 608–610. *ISSN*: 1476-4687. *DOI*: 10.1038/153608a0. *URL*: <https://doi.org/10.1038/153608a0>.
- [93] Meija, J. et al. "Atomic weights of the elements 2013 (IUPAC Technical Report)". In: *Pure and Applied Chemistry* 88.3 (1Mar. 2016), pp. 265–291. *DOI*: <https://doi.org/10.1515/pac-2015-0305>. *URL*: <https://www.degruyter.com/view/journals/pac/88/3/article-p265.xml>.
- [94] Mendeleev, D. "On the Correlation Between the Properties of the Elements and their Atomic Weights". In: *Mendeleev on the periodic law: Selected writings, 1869-1905*. Ed. by W. B. Jensen. New York: Dover, 2002. Chap. 2, pp. 18–37.

- [95] Mendeleev, D. "On the periodic regularity of the chemical elements". In: *Mendeleev on the periodic law: Selected writings, 1869-1905*. Ed. by W. B. Jensen. New York: Dover, 2002. Chap. 3, pp. 38–109.
- [96] Mendeleev, D. "On the Relation of the Properties to the Atomic Weights of the Elements". In: *Mendeleev on the periodic law: Selected writings, 1869-1905*. Ed. by W. B. Jensen. New York: Dover, 2002. Chap. 1, pp. 16–17.
- [97] Mendeleev, D. "The grouping of the elements and the periodic law". In: *Mendeleev on the periodic law: Selected writings, 1869-1905*. Ed. by W. B. Jensen. New York: Dover, 2002. Chap. 13, pp. 253–314.
- [98] Meseguer, J. and U. Montanari. "Petri nets are monoids". In: *Information and computation* 88.2 (1990), pp. 105–155.
- [99] Mestres, J. and G. M. Maggiora. "Putting molecular similarity into context: Asymmetric indices for field-based similarity measures". In: *Journal of Mathematical Chemistry* 39.1 (Jan. 2006), pp. 107–118. ISSN: 1572-8897. DOI: 10.1007/s10910-005-9007-3. URL: <https://doi.org/10.1007/s10910-005-9007-3>.
- [100] Meyer, L. "Kritik einer Abhandlung von A. Kekulé". In: *Zeitschrift für Chemie* 8 (1865), pp. 250–254.
- [101] Meyer, L. *Die modernen Theorien der Chemie und ihre Bedeutung für die chemische Statik*. Breslau: Verlag von Maruschke & Berendt, 1864.
- [102] Meyer, L. "Die Natur der chemischen Elemente als Function ihrer Atomgewichte". In: *Ann. Chem. Pharm.* VII Supplementband (1870), pp. 354–364.
- [103] Michoel, T. and B. Nachtergaele. "Alignment and integration of complex networks by hypergraph-based spectral clustering". In: *Phys. Rev. E* 86 (5 2012), p. 056111. DOI: 10.1103/PhysRevE.86.056111. URL: <https://link.aps.org/doi/10.1103/PhysRevE.86.056111>.
- [104] Newman, M. E. J. and M. Girvan. "Finding and evaluating community structure in networks". In: *Physical Review E* 69.2 (Feb. 2004). ISSN: 1550-2376. DOI: 10.1103/physreve.69.026113.
- [105] Newman, M. "Power laws, Pareto distributions and Zipf's law". In: *Contemporary Physics* 46.5 (2005), pp. 323–351. DOI: 10.1080/00107510500052444.
- [106] Nicolaou, K. C. "The Emergence of the Structure of the Molecule and the Art of Its Synthesis". In: *Angew Chem Int Ed Eng* 52.1 (2013), pp. 131–146. DOI: 10.1002/anie.201207081.
- [107] Nicolaou, Z. G. and A. E. Motter. "Missing links as a source of seemingly variable constants in complex reaction networks". In: *Physical Review Research* 2 (4 2020), p. 043135. DOI: 10.1103/PhysRevResearch.2.043135.
- [108] O'Donnell, T. A., D. F. Stewart, and P. Wilson. "Reactivity of Transition Metal Fluorides. II. Uranium Hexafluoride". In: *Inorg Chem* 5.8 (1966), pp. 1438–1441. DOI: 10.1021/ic50042a035.
- [109] Odling, W. "On the proportional numbers of the elements". In: *Quarterly Journal of Science* 1 (1864), pp. 642–648.

- [110] Paiva, V. de. "A Dialectica-like model of linear logic". In: *Category Theory and Computer Science*. Springer Berlin Heidelberg. 1989, pp. 341–356.
- [111] Paiva, V. de. *Categorical Multirelations, Linear Logic and Petri Nets, Technical Report*. 1991.
- [112] Paiva, V. de. "The Dialectica Categories". In: *Categories in Computer Science and Logic: Proceedings of the AMS-IMS-SIAM Joint Summer Research Conference Held June 14–20, 1987 with Support from the National Science Foundation*. Vol. 92. American Mathematical Society. 1989, p. 47.
- [113] Paiva, V. de. "The Dialectica Categories". In: University of Cambridge, Computer Lab Technical Report, PhD thesis. 1991.
- [114] Partington, J. R. *A history of chemistry*. London: Macmillan, 1964.
- [115] Ramberg, P. J. "Myth 7. That Friedrich Wöhler's Synthesis of Urea in 1828 Destroyed Vitalism and Gave Rise to Organic Chemistry". In: *Newton's Apple and Other Myths about Science*. Ed. by R. L. Numbers and K. Kampourakis. Harvard University Press, 2015, pp. 59–66. DOI: doi:10.4159/9780674089167-009. URL: <https://doi.org/10.4159/9780674089167-009>.
- [116] Rathke, J., P. Sobociński, and O. Stephens. "Compositional reachability in Petri nets". In: *International Workshop on Reachability Problems*. Springer. 2014, pp. 230–243.
- [117] Rayner-Canham, G. *The Periodic Table*. World Scientific, 2020. DOI: 10.1142/11775. eprint: <https://worldscientific.com/doi/pdf/10.1142/11775>. URL: <https://worldscientific.com/doi/abs/10.1142/11775>.
- [118] Reed, J. L., T. D. Vo, C. H. Schilling, and B. Ø. Palsson. "An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR)". In: *Genome Biol.* 4 (9 2003), R54.1–R54.12.
- [119] Restrepo, G. and J. Jost. *A formal setting for the evolution of chemical knowledge*. Preprint. 2020. URL: <https://www.mis.mpg.de/publications/preprints/2020/prepr2020-77.html>.
- [120] Rex, F. "Zur Erinnerungen an Felix Hoppe-Seyler, Lothar Meyer und Walter Hückel". In: *Bausteine zur Tübinger Universitätsgeschichte* 8 (1997), pp. 103–130.
- [121] Riemenschneider, W. and M. Tanifuji. "Oxalic Acid". In: *Ullmann's Encyclopedia of Industrial Chemistry*. American Cancer Society, 2011. ISBN: 9783527306732. DOI: 10.1002/14356007.a18\_247.pub2.
- [122] Rocke, A. J. "Lothar Meyer's pathway to periodicity". In: *Ambix* 66.4 (2019), pp. 265–302. DOI: 10.1080/00026980.2019.1677976. eprint: <https://doi.org/10.1080/00026980.2019.1677976>. URL: <https://doi.org/10.1080/00026980.2019.1677976>.
- [123] Rocke, A. J. "Atoms and Equivalents: The Early Development of the Chemical Atomic Theory". In: *Historical Studies in the Physical Sciences* 9 (1978), pp. 225–263. ISSN: 00732672. URL: <http://www.jstor.org/stable/27757379>.
- [124] Rocke, A. J. *Chemical atomism in the nineteenth century*. Columbus: Ohio University Press, 1984.
- [125] Rozen, A. M. "The first plant in the world for the production of heavy water by the method of two-temperature water-hydrogen sulfide isotopic exchange". In: *Atomic Energy* 78.3 (Mar. 1995), pp. 218–223. ISSN: 1573-8205. DOI: 10.1007/BF02407494.

- [126] Saito, A., M. Nagasaki, H. Matsuno, and S. Miyano. "Hybrid Functional Petri Net with Extension for Dynamic Pathway Modeling". In: *Modeling in Systems Biology: The Petri Net Approach*. Ed. by I. Koch, W. Reisig, and F. Schreiber. London: Springer London, 2011, pp. 101–120. ISBN: 978-1-84996-474-6. DOI: 10.1007/978-1-84996-474-6\_6. URL: [https://doi.org/10.1007/978-1-84996-474-6\\_6](https://doi.org/10.1007/978-1-84996-474-6_6).
- [127] Samal, A., R. P. Sreejith, J. Gu, S. Liu, E. Saucan, and J. Jost. "Comparative analysis of two discretizations of Ricci curvature for complex networks". In: *Scientific Reports* 8 (2018). DOI: 10.1038/s41598-018-27001-3.
- [128] Sandhu, R., T. Georgiou, E. Reznik, L. Zhu, I. Kolesov, Y. Senbabaoglu, and A. Tannenbaum. "Graph Curvature for Differentiating Cancer Networks". In: *Sci. Rep.* 5.1 (July 2015), p. 12323. ISSN: 2045-2322. DOI: 10.1038/srep12323. URL: <https://doi.org/10.1038/srep12323>.
- [129] Saucan, E. and J. Jost. *Network topology vs. geometry: from persistent homology to curvature*. MIS-Preprint 5/2017. Max Planck Institute for Mathematics in the Sciences, 2017.
- [130] Saucan, E., A. Samal, M. Weber, and J. Jost. "Discrete curvatures and network analysis". In: *MATCH Communications in Mathematical and in Computer Chemistry* 80.3 (2018), pp. 605–622.
- [131] Saucan, E., R. P. Sreejith, R. P. Vivek-Ananth, J. Jost, and A. Samal. "Discrete Ricci curvatures for directed networks". In: *Chaos Soliton. Fract.* 118 (2019), pp. 347–360. DOI: 10.1016/j.chaos.2018.11.031.
- [132] Saucan, E. and M. Weber. *Forman's Ricci curvature – From networks to hypernetworks*. Tech. rep. 1810.07749v1. ArXiv, 2018.
- [133] Savić, M., M. Ivanović, and L. C. Jain. "Co-authorship Networks: An Introduction". In: *Complex Networks in Software, Knowledge, and Social Systems*. Cham: Springer International Publishing, 2019, pp. 179–192. ISBN: 978-3-319-91196-0. DOI: 10.1007/978-3-319-91196-0\_5.
- [134] Scerri, R. E. *The periodic table: Its story and its significance*. New York: Oxford University Press, 2019. ISBN: 9780195305739.
- [135] Schölkopf, B., J. Platt, and T. Hofmann. "Learning with Hypergraphs: Clustering, Classification, and Embedding". In: *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*. MIT Press, 2007, pp. 1601–1608. ISBN: 9780262256919. URL: <https://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6287378>.
- [136] Schummer, J. "Scientometric studies on chemistry I: The exponential growth of chemical substances, 1800–1995". In: *Scientometrics* 39.1 (May 1997), pp. 107–123. ISSN: 1588-2861. DOI: 10.1007/BF02457433. URL: <https://doi.org/10.1007/BF02457433>.
- [137] Schummer, J. "The chemical core of chemistry I: A Conceptual approach". In: *Hyle* 4.2 (1998), pp. 129–162.
- [138] Sloan, R. H., D. Stasi, and G. Turan. *Hydras: Directed Hypergraphs and Horn Formulas*. 2015. arXiv: 1504.07753 [cs.DM].
- [139] Solla Price, D. J. de. "Networks of Scientific Papers". In: *Science* 149.3683 (1965), pp. 510–515. ISSN: 0036-8075. DOI: 10.1126/science.149.3683.510.
- [140] Spivak, D. I. *Higher-dimensional models of networks*. 2009. arXiv: 0909.4314 [cs.NI].

- [141] Sreejith, R. P., J. Jost, E. Saucan, and A. Samal. "Forman curvature for directed networks". In: *ArXiv e-prints* (2016). arXiv: 1605.04662.
- [142] Sreejith, R. P., J. Jost, E. Saucan, and A. Samal. "Systematic evaluation of a new combinatorial curvature for complex networks". In: *Chaos Solitons and Fractals* 101 (2017), pp. 50–67. DOI: 10.1016/j.chaos.2017.05.021. arXiv: 1610.01507.
- [143] Sreejith, R. P., K. Mohanraj, J. Jost, E. Saucan, and A. Samal. "Forman curvature for complex networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 6 (2016), p. 063206. DOI: 10.1088/1742-5468/2016/06/063206. eprint: 1603.00386.
- [144] Stadler, B. M. R. and P. Stadler. "Reachability, Connectivity, and Proximity in Chemical Spaces". In: *MATCH Communications in Mathematical and in Computer Chemistry* 80.3 (2018), pp. 639–659. ISSN: 0340-6253.
- [145] Stadler, B. M. R. and P. F. Stadler. "Generalized Topological Spaces in Evolutionary Theory and Combinatorial Chemistry". In: *Journal of Chemical Information and Computer Sciences* 42.3 (May 2002), pp. 577–585. ISSN: 0095-2338. DOI: 10.1021/ci0100898.
- [146] Thiele, I., T. D. Vo, N. D. Price, and B. Ø. Palsson. "Expanded Metabolic Reconstruction of *Helicobacter pylori* (iT341 GSM/GPR): an In Silico Genome-Scale Characterization of Single- and Double-Deletion Mutants". In: *J. Bacteriol.* 187.16 (2005), pp. 5818–5830. DOI: 10.1128/JB.187.16.5818-5830.2005. eprint: <https://journals.asm.org/doi/pdf/10.1128/JB.187.16.5818-5830.2005>. URL: <https://journals.asm.org/doi/abs/10.1128/JB.187.16.5818-5830.2005>.
- [147] Thomson, T. "On the Daltonian theory of definite proportions in chemical combinations". In: *Annals of Philosophy* 2 (1813), pp. 32–52.
- [148] Thyssen, P. and K. Binnemans. "Mendeleev and the rare-earth crisis". In: *Philosophy of Chemistry: Growth of a New Discipline*. Ed. by E. Scerri and L. McIntyre. Dordrecht: Springer Netherlands, 2015, pp. 155–182. ISBN: 978-94-017-9364-3. DOI: 10.1007/978-94-017-9364-3\_11. URL: [https://doi.org/10.1007/978-94-017-9364-3\\_11](https://doi.org/10.1007/978-94-017-9364-3_11).
- [149] Togni, A. and R. L. Halterman. *Metallocenes: Synthesis Reactivity Applications*. Wiley, 2008.
- [150] Vazquez, A. "Population stratification using a statistical model on hypergraphs". In: *Phys. Rev. E* 77 (6 2008), p. 066106. DOI: 10.1103/PhysRevE.77.066106. URL: <https://link.aps.org/doi/10.1103/PhysRevE.77.066106>.
- [151] Wagner, A. and D. A. Fell. "The small world inside large metabolic networks". In: *Proc. R. Soc. B.* 268.1478 (2001), pp. 1803–1810. ISSN: 0962-8452. DOI: 10.1098/rspb.2001.1711. eprint: <http://rspb.royalsocietypublishing.org/content/268/1478/1803.full.pdf>. URL: <http://rspb.royalsocietypublishing.org/content/268/1478/1803>.
- [152] Weber, M., E. Saucan, and J. Jost. *Can one see the shape of a network?* Tech. rep. 1608.07838. arXiv, 2016.
- [153] Weber, M., J. Stelzer, E. Saucan, A. Naitsat, G. Lohmann, and J. Jost. *Curvature-based Methods for Brain Network Analysis*. Tech. rep. 1707.00180. arXiv, 2017.
- [154] Weber, M., J. Jost, and E. Saucan. "Forman-Ricci Flow for Change Detection in Large Dynamic Data Sets". In: *Axioms* 5 (2016), p. 26. ISSN: 2075-1680. DOI: 10.3390/axioms5040026.

- [155] Weber, M., E. Saucan, and J. Jost. "Characterizing complex networks with Forman-Ricci curvature and associated geometric flows". In: *Journal of Complex Networks* 5.4 (2017), pp. 527–550. DOI: 10.1093/comnet/cnw030. eprint: /oup/backfile/content\_public/journal/comnet/5/4/10.1093\_comnet\_cnw030/2/cnw030.pdf.
- [156] Winskel, G. "A category of labelled Petri nets and compositional proof system". In: *[1988] Proceedings. Third Annual Symposium on Logic in Computer Science*. IEEE, 1988, pp. 142–154. DOI: 10.1109/LICS.1988.5113.
- [157] Winskel, G. "Petri nets, algebras, morphisms, and compositionality". In: *Information and Computation* 72.3 (1987), pp. 197–238. ISSN: 0890-5401. DOI: [https://doi.org/10.1016/0890-5401\(87\)90032-0](https://doi.org/10.1016/0890-5401(87)90032-0).
- [158] Xiong, F.-L., Y.-Z. Zhen, W.-F. Cao, K. Chen, and Z.-B. Chen. "Qudit hypergraph states and their properties". In: *Phys. Rev. A* 97 (1 2018), p. 012323. DOI: 10.1103/PhysRevA.97.012323. URL: <https://link.aps.org/doi/10.1103/PhysRevA.97.012323>.
- [159] Zhang, J. and H. W. Richardson. "Copper Compounds". In: *Ullmann's Encyclopedia of Industrial Chemistry*. American Cancer Society, 2016, pp. 1–31. ISBN: 9783527306732. DOI: 10.1002/14356007.a07\_567.pub2.

# Curriculum Scientiae

## Personal Information

Name: Wilmer Leal  
Birth: May 10, 1989  
Birthplace: Pamplona, Colombia  
Home address: Paul-List-Straße 26, 04103 Leipzig, Germany  
Work address: Inselstraße 22, 04103 Leipzig, Germany  
Nationality: Colombian  
Marital status: Married

## Education

2017 - 2022 PhD Student in Computer Science, University of Leipzig & Max Planck Institute for Mathematics in the Sciences  
*Summa cum laude*  
2009-2013 Bachelor of Sciences in Mathematics, University of Pamplona  
*Laureate thesis*  
2005-2010 Bachelor of Science in Chemistry, University of Pamplona  
*Excellence Thesis*

## Languages

Spanish: Native speaker  
English: Fluent  
German: Basic knowledge

## Marks of recognition

2022	Graduated summa cum laude
2022	Selected participant of the 71st Lindau Nobel Laureate Meeting (dedicated to Chemistry) – from 26 June to 1 July 2022
2021	Nominated to attend the 71st Lindau Nobel Laureate Meeting (Chemistry)
2021	Selected by Max Planck Institute director Prof. Dr. Jost to give a lecture for the evaluation of the institute at the 2021's MPG Scientific Advisory Board meeting
2020	Selected to attend the Applied Category Theory Adjoint School 2020 at MIT, Boston, USA (16 students/researchers selected from 144 applicants)
2019 April	Cover of the journal <i>Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences</i>
2018 Dec	Visiting Scientist, University of Primorska
2018 -	Guest Scientist, Max Planck Institute for the Mathematics in the Sciences
2017	PhD grant, Deutscher Akademischer Austauschdienst (DAAD)
2017	Guest Scientist, Max Planck Institute for the Mathematics in the Sciences
2017	Guest Scientist, University of Leipzig
2016	Guest Scientist, University of Leipzig
2016	Featured Article, <i>Journal of Cheminformatics</i>
2013	Excellence Thesis
2010	Laureate Thesis

## Articles

2022	Leal, W.; Llanos, E. J.; Bernal, A.; Jost, J.; Stadler, P. F.; Restrepo, G. The expansion of chemical space in 1826 and in the 1840s prompted the convergence to the periodic system. <i>Proceedings of the National Academy of Sciences USA</i> 119:30, e2119083119. Llanos, E. J.; Leal, W.; Bernal, A.; Jost, J.; Stadler, P. F. Are the chemical families still there? exploration of similarity among elements. <i>ChemRxiv</i> Teich, M.; Leal, W.; Jost, J. Corpus-based Metaphor Analysis through Graph Theoretical Methods. <i>arXiv:2209.12234</i> .
2021	Leal, W.; Restrepo, G.; Stadler, P. F.; Jost, J. Forman-Ricci curvature for hypergraphs. <i>Advances in Complex Systems</i> 24:01, 2150003. <a href="https://doi.org/10.1142/S021952592150003X">https://doi.org/10.1142/S021952592150003X</a> . Lavore, E. D.; Leal, W.; de Paiva, V. Dialectica Petri nets. <i>ArXiv</i> . <a href="https://arxiv.org/abs/2105.12801">https://arxiv.org/abs/2105.12801</a> .
2020	Eidi, M.; Farzam, A.; Leal, W.; Samal, A.; Jost, J. Edge-based analysis of networks : curvatures of graphs and hypergraphs. <i>Theory in Biosciences</i> 139:4, 337-348. <a href="https://doi.org/10.1007/s12064-020-00328-0">https://doi.org/10.1007/s12064-020-00328-0</a> . Leal, W.; Eidi, M.; Jost, J. Ricci curvature of random and empirical directed hypernetworks. <i>Applied Network Science</i> 5:65. <a href="https://doi.org/10.1007/s41109-020-00309-8">https://doi.org/10.1007/s41109-020-00309-8</a> (invited).

- 2019 Leal, W.; Llanos, E. J.; Jost, J.; Stadler, P. F.; Restrepo, G. The chemical space from which the periodic system arose. *ChemRxiv. Preprint* <https://doi.org/10.26434/chemrxiv.9698888.v1>.  
Llanos, E. J.\*; Leal, W.\*; Luu, D. H.; Jost, J.; Stadler, P. F.; Restrepo, G. Exploration of the chemical space and its three historical regimes. *Proceedings of the National Academy of Sciences USA* 116:26, 12660-12665. <https://doi.org/10.1073/pnas.1816039116> (\* identifies co-first authors).  
Leal, W.; Restrepo, G. Formal structure of periodic system of elements. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 475:20180581 <http://dx.doi.org/10.1098/rspa.2018.0581> (cover article).
- 2016 Leal, W.; Llanos, E. J.; Restrepo, G.; Suárez, C.; Patarroyo, M. E. How frequently do clusters occur in hierarchical clustering analysis? A graph theoretical approach to studying ties in proximity. *Journal of Cheminformatics* 8:4. <https://doi.org/10.1186/s13321-016-0114-x> (featured article).
- 2012 Leal, W.; Restrepo, G.; Bernal, A. A network study of chemical elements: from binary compounds to chemical trends. *MATCH Communications in Mathematical and in Computer Chemistry* 68, 417–442.

## Book Chapters

- 2015 Bernal, A.; Llanos, E. J.; Leal, W.; Restrepo, G. Similarity in chemical reaction networks: categories, concepts and closures. In: *Advances in Mathematical Chemistry and Applications, Vol. 2*, 24-54. Elsevier.

## Abstracts

- 2019 Leal, W.; Eidi, M.; Jost, J. Curvature-based analysis of directed hypernetworks. In: *Complex networks 2019: the 8th international conference on complex networks and their applications 2-34*; December 10-12, 2019 Lisbon, Portugal; book of abstract / Hocine Cherifi (ed.) [s.l.]: International conference on complex networks and their applications.  
Llanos, E. J.; Leal, W.; Bernal, A.; Restrepo, G.; Jost, J.; Stadler, P. F. A Network model of the chemical space provides similarity structure to the system of chemical elements. In: *Complex networks 2019: the 8th international conference on complex networks and their applications 308-310*; December 10-12, 2019 Lisbon, Portugal; book of abstract / Hocine Cherifi (ed.) [s.l.]: International conference on complex networks and their applications.
- 2017 Llanos, E. J.; Leal, W.; Restrepo, G.; Stadler P. F. Computational approach to the history of chemical reactivity: exploring Reaxys database. *Abstracts of Papers of the American Chemical Society* 254.
- 2016 Leal, W.; Llanos, E. J.; Restrepo, G.; Suárez, C.; Patarroyo, ME. How frequent are your clusters in hierarchical cluster analysis? Quantifying their frequencies considering ties in proximity. *Abstracts of Papers of the American Chemical Society* 252.
- 2009 Leal, W.; Restrepo, G. Chemical elements, their sociology. *J. Sci. Edu.*, 10, 126-127.

## Participation in scientific conferences (selected)

- 2021 Leal, W. *Exploration of the chemical space*. Leipzig, Germany. (Selected by Max Planck Institute Director Prof. Jost to be presented at the 2021's Scientific Advisory Board meeting, for the evaluation of the institute).  
Di Lavore, E.; Leal, W.; de Paiva, V. 4th International Conference on Applied Category Theory. *Dialectica Petri nets*. University of Cambridge, Cambridge, United Kingdom.
- 2020 Leal, W. Encuentro Nororiental de Matemáticas 2020. *Aspectos formales de la exploración del Espacio Químico [Formal aspects of chemical space exploration]*. Pamplona, Colombia. (Invited)  
Leal, W.; Llanos, E. J.; Bernal, A.; Restrepo, G.; Luu, D. H.; Jost, J.; Stadler, P. F. NetSci-X 2020 Tokyo: International School and Conference on Network Science. *Exploring the hypergraph structure underlying the Chemical Space*. Tokyo, Japan.  
Leal, W.; Llanos, E. J.; Restrepo, G.; Jost, J.; Stadler, P. F. Inorganic colloquium series at Faculty of Chemistry and Mineralogy, Leipzig, Germany
- 2019 Leal, W. Data Day. MPI for Mathematics in the Sciences. *Exploring the Chemical Space: a data driven approach*. Leipzig, Germany. (Invited)  
Leal, W. Max Planck Institute for Mathematics in the Sciences. *Mathematical approaches to exploring the Chemical Space*. Leipzig, Germany.  
Leal, W.; Eidi, M.; Jost, J. The 8th International Conference on Complex Networks & Their Applications. *Curvature-based analysis of directed hypernetworks*. Lisbon, Portugal.  
Llanos, E. J.; Leal, W.; Bernal, A.; Restrepo, G.; Jost, J.; Stadler, P. F. The 8th International Conference on Complex Networks & Their Applications. *A network model of the chemical space provides similarity structure to the system of chemical elements*. Lisbon, Portugal.  
Leal, W.; Llanos, E. J.; Restrepo, G.; Luu, D. H.; Jost, J.; Stadler, P. F. PHASE-Chimie lecture at Université Paris Diderot. *Chemical space: historical trends and changes over time*. Paris, France.  
Leal, W.; Llanos, E. J.; Restrepo, G.; Jost, J.; Stadler, P. F. XXI Mendeleev Congress on General and Applied Chemistry. *The expanding chemical space from which the periodic system arose*. St. Petersburg, Russia.  
Llanos, E. J.; Leal, W.; Restrepo, G.; Jost, J.; Stadler, P. F. The 12th International Conference on the History of Chemistry. *The chemical space and its three historical regimes*. Maastricht, The Netherlands.  
Leal, W.; Restrepo, G. Mendeleev 150: 4th International Conference on the Periodic Table. *Formal structure of periodic system of elements*. Saint Petersburg, Russian Federation.
- 2018 Leal, W.; Llanos, E. J.; Restrepo, G.; Jost, J.; Stadler, P. F. 4th Spring School Complex Networks: Theory, Methods, and Applications. *Growth patterns and graph transformation rules underlying the Network of Chemical Reactions*. Como, Italy.  
Leal, W.; Llanos, E. J.; Restrepo, G.; Jost, J.; Stadler, P. F. Max Planck Institute for Mathematics in the Sciences. *Growth patterns and graph transformation rules underlying the network of chemical reactions*. Leipzig, Germany.

- 2017 Llanos, E. J.; Leal, W.; Restrepo, G.; Stadler, P. F. 34th TBI Winterseminar in Bled. *News on the history of chemistry through chemical reactions*. Bled, Slovenia.  
 Llanos, E. J.; Leal, W.; Restrepo, G.; Stadler, P. F. 253rd American Chemical Society National Meeting & Exposition, Washington, USA. *Computational approach to the history of chemical reactivity: exploring Reaxys database*, Washington, USA.  
 Llanos, E. J.; Leal, W.; Restrepo, G.; Stadler, P. F. EON Workshop on Computational Chemistry, from Components to Systems and Back, *A computational approach to the Reaxys network of chemical reactions*. Tokyo, Japan.
- 2016 Leal, W.; Restrepo, G.; Stadler, P. F. Herbstseminar der Bioinformatik. *Graph transformation rules underlying patterns of organic chemical reactions*. Doubice, Czech Republic.
- 2014 Leal, W. Minería de Patrones, desde los Elementos Químicos Hasta la Vacuna Contra la Malaria. *Minería de patrones, desde los elementos químicos hasta la vacuna contra la malaria [Pattern mining, from chemical elements to Malaria vaccine]*. Pamplona, Colombia.
- 2010 Leal, W.; Esquivel, L.; Restrepo, G. Universidad de Pamplona: 50th anniversary celebration with Mathematical Chemistry. *On the topology of Formal Concept Analysis*. Pamplona, Colombia.
- 2009 Leal, W.; Restrepo, G.; Bernal, A. VII Simposio nororiental de Matemáticas. *Teoría de redes y los elementos químicos [Network theory and chemical elements]*. Bucaramanga, Colombia.  
 Leal, W.; Restrepo, G.; Bernal, A. International Congress of Science Education. *Chemical elements, their sociology*. Cartagena, Colombia.  
 Leal, W.; Restrepo, G.; Bernal, A. Symposium on the Philosophy of Chemistry. *Binary compounds, a network theory approach* Bogota, Colombia.  
 Leal, W.; Restrepo, G.; Bernal, A. First Mathematical Chemistry Workshop of the Americas. *From network theory to periodic trends*. Bogota, Colombia.

## Schools and Courses

- 2022 **School**  
*Mathematics Research Communities (MRC)*  
 American Mathematical Society, New York, USA  
 Mentor: Professor Dr. John Baez
- 2021 **School**  
*Graduate Student Combinatorics Conference (GSCC)*  
 University of Minnesota, Twin Cities, USA  
**School**  
*CSH Winter Lecture Series*  
 Complexity Science Hub Vienna, Vienna, Austria
- 2020 **School** (Feb-Jul)  
*Applied Category Theory Adjoint School 2020*  
 MIT, Boston, USA  
 Mentor: Professor Dr. Valeria de Paiva  
 (16 students/researchers selected from 144 applicants)
- 2019 **Course** (summer term)

*Schemes and Sheaves - An introduction to modern algebraic geometry*

Lecturer: Professor Dr. Jürgen Jost

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**Course** (summer term)

*Category theory and applications*

Lecturer: Dr. Paolo Perrone

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**Course** (summer term)

*Information and Complexity*

Lecturer: Professor Dr. Jürgen Jost

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**Summer School**

*First School on Reaction Systems*

Nicolaus Copernicus University, Torun, Poland

2018

**Course** (summer term)

*Geometric and topological methods of data analysis*

Lecturer: Professor Dr. Jürgen Jost

Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany

**Summer School**

*Network Science in the Humanities*

Max Planck Institutes for Mathematics in the Sciences & for the History of Science, Leipzig, Germany.

**Spring School**

*4th Spring School Complex Networks: Theory, Methods, and Applications*

Lake Como School of Advanced Studies, Como, Italy.

## Research Experience in Groups and Institutes

2018 -	Guest Scientist Max Planck Institute for the Mathematics in the Sciences Leipzig, Germany
Feb-Apr 2017	Guest Scientist Max Planck Institute for the Mathematics in the Sciences Leipzig, Germany
Aug-Oct 2016	Guest Scientist Bioinformatics Group University of Leipzig Leipzig, Germany
2012-2013	Researcher Biomathematics Group Fundación Instituto de Inmunología de Colombia - FIDIC Bogota, Colombia
2010-2018	Researcher Laboratorio de química teórica

2005-2006      Universidad de Pamplona  
                  Pamplona, Colombia  
                  Researcher  
                  Grupo de Investigación en Transporte Molecular - GITRAM  
                  Universidad de Pamplona.  
                  Pamplona, Colombia

## Teaching

2020-      Category Theory and its applications to chemistry (seminar)  
                  Mathematics Department, University of Pamplona  
 2015-I      Differential Equations  
                  Mathematics Department, University of Pamplona  
 2014-II      Differential Geometry, Functional Analysis, & Complex Variable  
                  *Mathematics Department, University of Pamplona*  
 2014-I      Real Analysis I, Linear Algebra, Differential Calculus  
                  *Mathematics Department, University of Pamplona*  
 2011-I      Mathematics I & Algebra and Geometry  
                  *Mathematics Department, University of Pamplona*

## Refereeing

2021      Reviewer for *Journal of Biosciences*  
 2021 -      Reviewer for *zbMATH*  
 2020      Reviewer for *Ingeniería y Ciencia*  
 2014      Chapter reviewer  
                  Book: *Advances in Mathematical Chemistry Vol. 2*  
                  Elsevier  
 2014      Referee of BSc thesis in Mathematics  
                  Mathematics Department, Pamplona University  
 2013      Referee BSc thesis in Mathematics  
                  Mathematics Department, Pamplona University

## IT-Knowledge

Operating      GNU-Linux, Windows.  
 systems:  
 Programming      Gawk, Bash, Python,  $\text{\LaTeX}$ .  
   & Markup  
 languages:

## Professional Affiliations

2020 -      Student member of the Society for Industrial and Applied Mathematics (SIAM)

2019 -	Member of the Colombian Association for the Advancement of Science (ACAC)
2019 -	Member of the Colombian Mathematical Society (SCM)
2018 -	Member of the Complex Systems Society (CSS)
2013-2014	Member of the Association for Symbolic Logic (ASL)

### Media coverage (selected)

2019	<p>By "Chemical &amp; Engineering News" (American Chemical Society). The following article features one of our papers: "Chemists discovered new compounds at an exponential rate over the past two centuries", by Sam Lemonick (C&amp;EN associate editor). By "Springer Professional".</p> <p>The following article features one of our papers: "Die verborgene Struktur des Periodensystems [The hidden structure of the periodic table ]", by Dieter Beste. By "Der Standard".</p> <p>The following article features one of our papers: "Neue chemische Substanzen werden mit überraschender Regelmäßigkeit entdeckt [New chemical substances are discovered with surprising regularity]". By "Max Planck Society".</p> <p>The following article features one of our papers: "The hidden structure of the periodic system". By "Chemistry World".</p> <p>The following article features one of our papers: "Hopes raised of a 'super-table' to end periodic table disputes", by Philip Ball (Former editor of <i>Nature</i>).</p>
------	--

**Selbständigkeitserklärung**

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

---

(Ort, Datum)

---

(Unterschrift)