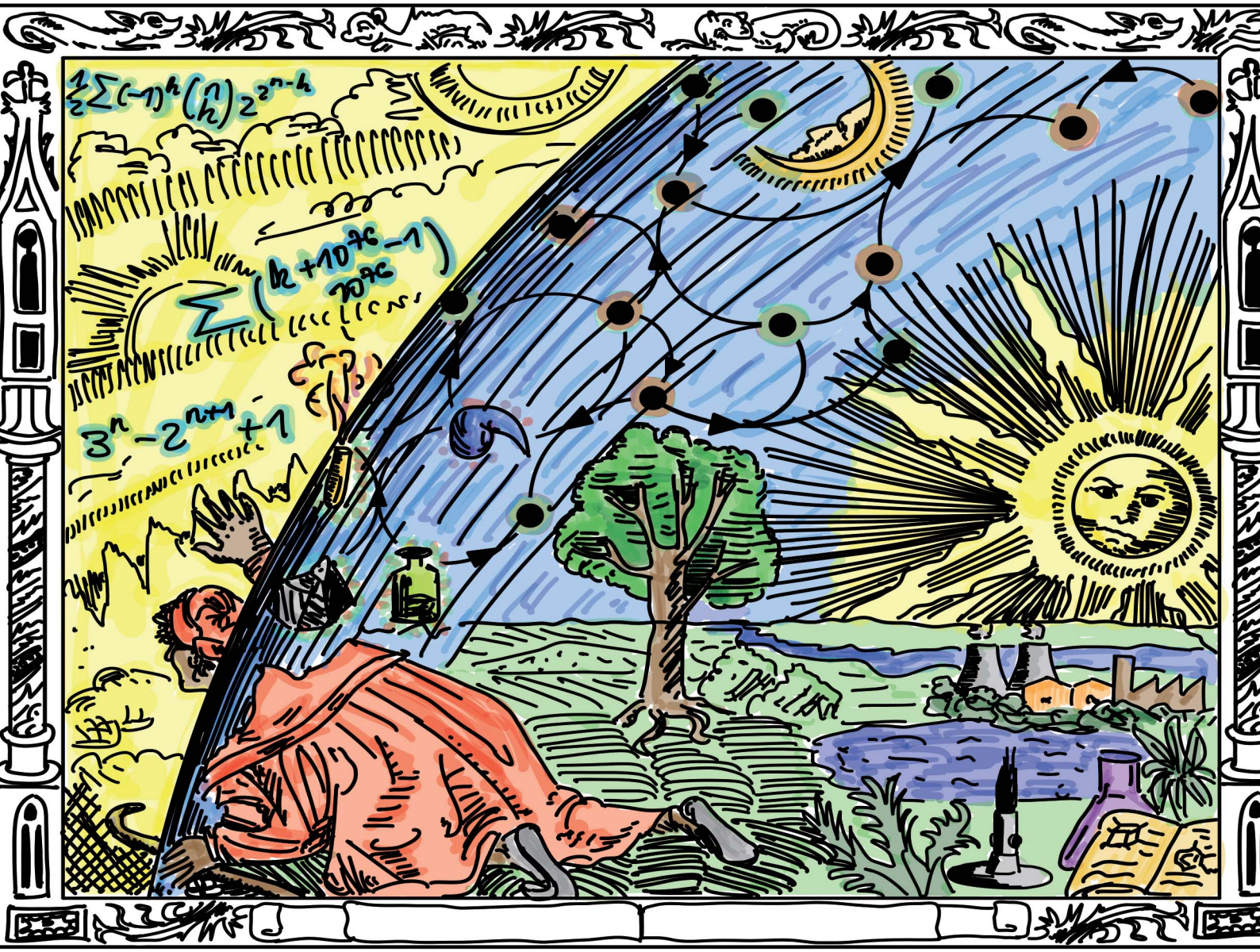
Volume 1
Number 5
October 2022
Pages 543-746

Digital Discovery

rsc.li/digitaldiscovery



ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE
Guillermo Restrepo
Chemical space: limits, evolution and modelling of an object bigger than our universal library

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2022, **1**, 568

Chemical space: limits, evolution and modelling of an object bigger than our universal library†‡

Guillermo Restrepo *

Chemical space entails substances endowed with a notion of nearness that comes in two flavours: similarity and synthetic reachability. What is the maximum size for the chemical space? Is there an upper bound for its classes of similar substances? How many substances and reactions can it house? Can we store these features of the chemical space? Here I address these questions and show that the physical universe does not suffice to store the chemical one embodied in the chemical space. By analysing the historical evolution of the space as recorded by chemists over the centuries, I show that it has been mainly expanded by synthesis of organic compounds and unfolds at an exponential rate doubling its substances each 16 years. At the turn of the 20th century it left behind an expansion period driven by reactions and entered the current era ruled by substance discovery, which often relies on some few starting materials and reaction classes. Extrapolating from these historical trends, synthesising a large set of affordable chemicals in the foreseeable future would require trebling the historical stable speed rate of discovery of new chemicals. Likewise, creating a database of failed reactions accounting for 25% of the known chemical space to assist the artificial intelligence expansion of the space could be afforded if the synthetic efforts of the coming five years are entirely dedicated to this task. Finally, I discuss hypergraph reaction models to estimate the future shape of the network underlying the chemical space.

Received 7th April 2022
Accepted 4th July 2022

DOI: 10.1039/d2dd00030j

rsc.li/digitaldiscovery

1 Introduction

Chemistry is about producing new substances and innovative methods to procure them. It is about documenting such a material enterprise for the sake of reproducibility and, above all, of expanding chemical knowledge.[2] This perspective is about those records of substances and reactions, which constitute the *chemical space*, and about the possibilities of storing the space for the generations to come.§

In the first half of this document I address the following questions: what is the maximum size the chemical space may reach? How many substances and reactions are mathematically possible? What is the upper bound for the number of classes of similar substances? Can we store the information of the chemical space?

The second half of the document is devoted to the evolution of the chemical space, based on the digitised historical record of its expansion. I address the following questions: are there historical trends on the report of chemicals and reactions? How has been the interplay of substances and reactions along the unfolding of the chemical space? Are there other than chemical driving forces guiding the evolution of the chemical space? Can we model the evolution of the chemical space?

This perspective is about the past, present and future of the material core of chemistry, which sheds light on its history and on the possible reaches of the discipline.

1.1 From substances to reactions: the two approaches to the chemical space

Why is the chemical space a space? One might say that it is just a trendy expression for the chemical output of centuries of research. This has some truth as easily checked, for instance, by typing "chemical space" in Ngrams of Google.[6] This expression began its popularity in the 2000s and today is much more used than "computational chemistry" and even than "Nobel prize." However, it is not much more popular than "organic chemistry", "quantum chemistry" or "inorganic chemistry". Chemical space is trendy because these times of rapid computational advances, mainly led by artificial intelligence (AI) successes upon big data, bring to the fore the colossal amount of chemical information, which not only grows rapidly but that enjoys

Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany. E-mail: restrepo@mis.mpg.de; Fax: +49 341 9959 658; Tel: +49 341 9959 601

†Electronic supplementary information (ESI) available. See https://doi.org/10.1039/d2dd00030j

‡The universe (which others call the Library) Jorge Luis Borges, 1941.[1]

§Chemical space encompasses different concepts, for example the collection of substances of pharmacological interest,[3,4] or the collection of molecular properties.[5] In this perspective, the former concept is a subset of what I call the chemical space and the latter is a mapping of (or from) the chemical space here defined to the space of molecular properties. See main text and note ¶¶¶.

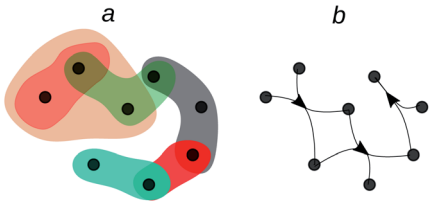


Fig. 1 Two approaches to the chemical space, where substances (black dots) are endowed with nearness relationships by (a) similarity or by (b) synthetic steps. In a, similarity classes are represented by subsets (coverings or hypergraphs) of different colours. In (b), chemical reactions, represented by arrows, lead to a chemical reaction network. Educts of a reaction lie at the tail of the arrow, while reaction products emerge from the arrow head.

a tradition of more than two centuries of curation and annotation. But besides the fashionability of the chemical space, it carries an interesting formal load, namely that of "space", which is a concept philosophers and mathematicians have tirelessly developed and elaborated upon over centuries.[7,8] Let us analyse in some detail the idea of "space" encoded in the chemical space.

A *space* entails a set of objects and a notion of nearness among them.[8,9] A chemical space can, therefore, be thought of as a set of chemicals endowed with a notion of nearness.[9,10] One may think of an Euclidean space, where substances are somehow located in a coordinate system allowing for measuring distances among them. Based on that nearness one may classify substances (Fig. 1a).[9]¶ Likewise, one can imagine the space as a set of substances related by synthetic paths, which leads to a network (Fig. 1b). In this setting, as often in chemistry, we can talk about distant substances if several synthetic steps separate one from each other. This nearness notion may also be used to classify substances.[9]

In the 1970s the nearness among substances was addressed from a molecular similarity stance.[11] Several mathematical and computational methods were developed to quantify such a similarity, which eventually led to the paradigm of Quantitative Structure–Activity Relationships models (QSAR), of widespread use today in medicinal and environmental chemistry.[12] These approaches are today applied to other branches of chemistry, which span substances such as polymers and materials.[13–16]

Currently, there is a surge of reports addressing the chemical space from a network perspective,[10,17,18] which has been motivated by the digitisation of reaction information that grew in the 1980s, 1990s and which is today analysed using machine learning tools,[17] for example for the optimal design of synthesis plans.[19–21] Another factor contributing to the current network studies of the chemical space is the maturity network theory has attained by important contributions in the 1990s and in subsequent years.[22]

In the following section I ponder on the size of the chemical space, from its similarity and network perspectives.

¶All over the text I talk about classes, which are to be understood as sets or subsets. This implies that classes in this text may overlap, as in Fig. 1a.

2 So huge, we cannot even name it

Analysing the *size of the chemical space* requires considering the number of substances and the number of similarity relationships among them, as well as the number of reactions among them. Some of these questions have entertained chemists, physicists, mathematicians and other scientists. Here I address those questions from a theoretical stance, aiming at determining upper bounds for the number of substances, classes of similar substances and possible number of reactions.‖

2.1 More than a universal library of substances

It has been estimated that the number of particles in the universe is about 10^{80},[28–30]** which amounts to 7×10^{76} atoms.†† A first approach to estimating the possible number of substances is determining the theoretical number of collections of atoms held together by chemical bonds. The number of such possible *atomic ensembles* is given by $\mathscr{C} = \sum_{k=1}^{10^{76}} \binom{k+10^{76}-1}{10^{76}}$, where $\binom{k+10^{76}-1}{10^{76}}$ is the number of ways of selecting k

‖Lower bounds are also interesting. As chemistry is built upon chemical elements, a natural smallest set of substances able to lead to the entire possible substances is the set of chemical elements. Although it is hypothesised that the maximum number of elements is 172,[23] chemical space requires elements able to form compounds, which leads to elements with lifetimes greater than 10^{-14} s.[24] This is enough time for most nuclei to reach their ground state and to gather their electrons at ambient conditions.[24] As nuclei stabilised by electrons are not enough to form compounds, a further requirement is that those atoms can actually form chemical bonds, which takes about 10^{-11} s.[25] At any rate, it seems the chemical space under ambient conditions cannot include more than 118 elements.[24] Further discussion on the lifetimes of nuclei and atoms is found in ref. 26. If chemical elements constitute the lower bound of the chemical space, the respective bound for the number of similarity classes may be provided by the smallest number of similarity relationships among chemical elements. A trivial lower bound is that no similarities exist at all, that is, that each element is only self-similar, providing n similarity classes for n elements. But there are actual resemblances among chemical elements. If we require at least a similarity relationship for each element, then for n elements there will be at least $\lceil n/2 \rceil$ similarities. An account of the 19th-century evolution of similarities among chemical elements is Ssection 4.2 found in ref. 27. Likewise, if we regard the lower bound of the number of chemical reactions as the case where every chemical element reacts with at least another one, then $\lceil n/2 \rceil$ is the size of the smallest set of chemical reactions for n elements. Clearly, I am not counting nuclear reactions in this approximation.

**Initial calculations were reported by Eddington in 1931 and refer to the number of hydrogen atoms accounting for the mass of the observable universe.[28] The calculation was based on hydrogen given that about 75% (Table S1, ESI†) of the mass of the universe is provided by this element. Although Eddington's number (2.36×10^{79}) can be obtained by dividing the mass of the universe (1.45×10^{53} kg) by the mass of the a hydrogen atom (1.67×10^{-27} kg), the number has been refined to include the number of baryons and electrons in the universe, which amounts to 1.93×10^{80} particles.[30]

††By considering the abundances of elements in the universe[30,31] and their atomic weights, the number of atoms per element can be calculated, which leads to the total number of atoms spanning the universe (Table S1, ESI†).

atoms from a collection of 10^{76} atoms, such that order is not important and repetitions are allowed.[32] So, here we are counting mono-, di, tri-, …, *n*-atomic ensembles up to the ultimate largest compound made of all 10^{76} atoms in the universe.‡‡ factors determining whether an atomic ensemble is chemically feasible or not. This requires determining the suitable conditions of pressure and temperature holding together the given atoms by electrostatic interactions. Although the chemical space has been traditionally regarded at ambient conditions (see Section 4), there is uncharted land at extreme conditions.[33] As usual in chemistry, we do not require the simultaneous "existence" of those substances, but the mere theoretical possibility of their existence and, importantly, of recording it.§§

Clearly, we are counting here ensembles far from the experimental possibilities we currently know, which also challenge the concept of chemical substance. Nevertheless, a piece of information that must be included corresponds to the further combinatorial possibilities arising from the manifold structures those ensembles may take. This can be addressed, as a zeroth-order approximation, by multiplying each ensemble by the possible number of graphs.¶¶ As these structures are based on binary relations of objects, in this case of atoms, graphs are perfectly suited for atomic ensembles made of bonds relating two elements.|||| Nevertheless, as there are substances such as boranes, which do not always hold classical 2-centre 2-electron bonds, a more general setting is needed, which is provided by hypergraphs.*** In this case, for instance, a 3-centre 2-electron bond as in B–H–B, can be modelled as a hyperedge made of three atoms, that is {B,H,B}.††† Likewise, aromatic systems constitute a hyperedge, where equivalent aromatic atoms become part of the hyperedge. Hence, a more accurate approximation to the number of substances in the chemical space is given by multiplying each atomic ensemble by a constrained number of possible hypergraphs associated to the given atomic ensemble.

At any rate, a higher order approximation to the upper bound of the number of substances in the chemical space requires chemical and mathematical knowledge, which may be attained by interdisciplinary collaboration. Scientists early in the 19th century recognised these possibilities when, for example, mathematician Rothe, chemist Bischof and botanist Nees von Esenbeck undertook the combinatorial challenge posed by chemical isomerism,[35] a subject, decades latter, continued by Cayley[36] and in the 20th century by Blair[37] and Pólya,[38] who counted the number of acyclic molecular structures.[39] Further interest in the subject arose by the advent of spectroscopic methods in the 1950s requiring determining the number of theoretical substances under particular chemical structural constraints of valence to come up with possible candidates for the different signals in spectra of different provenances.[40]‡‡‡ More recent approaches involve the collaboration of computer scientists, mathematicians and chemists, with outcomes such as MOLGEN,[42,43] a software package that, among several other features, provides the number of isomers of a given chemical formula based on a blending of group and graph theories, along with group algebra.

In the 1990s Weininger hypothesised that the number of possible substances is about 10^{200}, which is known as the "Weininger number" $\mathscr{W}$.[44,45] According to Gorse, $\mathscr{W}$ is "a lower limit of the number of different (chiral) molecular graphs possible given known chemistry (*i.e.*, bond types), restricted elements (C, N, O, P, S, halogens) and a molecular weight of less than <1000 dalton. Of these, it was further estimated that only about 1 in 10^{20} compounds could possibly be physically and chemically stable, giving 10^{180} compounds".[45] Although $\mathscr{W} \ll \mathscr{C}$, $\mathscr{W}$ is anyhow huge. Assuming that $\mathscr{W}$ provides a more realistic upper bound for the size of the chemical space, the question that arises is whether the information of those $\mathscr{W}$ substances can actually be stored. This would secure the expansion of chemical knowledge and would continue a strong disciplinary annotation tradition which began with the 13th-century encyclopedists such as Angelicus and Beauvais[2] and reached us through colossal handbooks such as those by Gmelin and Beilstein.[2] Annotating $\mathscr{W}$ substances would lead us to continue the current joy of having the whole corpus of chemical experimentation at our fingertips through electronic databases such as Reaxys and SciFinder.

Can we annotate $\mathscr{W}$? Unfortunately, no, we cannot! The entire universe does not suffice to store the most simple labels characterising those substances! The universe is able to store no more than 10^{123} bits.§§§ So, our universal library, to put it in terms of Borges,[1] is too small to accommodate the most simple

‡‡This material upper bound requires further adjustments to touch physical and, above all, chemical reality. It requires taking some few atoms out of the 10^{76} to account for the synthesiser of the largest compound, which may be either a human or a robot. Besides the constraints discussed in note ‖, energetic conditions constitute the key.

§§An instance of one of the largest atomic ensembles already synthesised is $C_{934893}H_{1495830}O_{49203}Si_{49203}Co_{19683}P_{19683}F_{118098}$, corresponding to a giant cobalticinium dendrimer accounting for 2 637 390 atoms.[34] So, we have achieved species accounting for 10^6 atoms, which although big for traditional standards, are very far from the theoretical upper bound of 10^{76} atoms.

¶¶A *graph* $G = (V, E)$ is made of a set of *vertices* V and a set of *edges* E. An edge is a set of two vertices. Thus, E is a collection of pairs of vertices. So, if $V = \{a, b, c\}$, a possible graph is $G = \{\{a, b, c\}, \{\{a, b\}, \{b, c\}\}\}$.

||||In this setting a single bond corresponds to a graph edge, while a double bond to the repetition of the edge. In general, any bond of order n requires a graph with n repeated edges.

***In a *hypergraph* $H = (V, E)$, V is a set of vertices and E is a collection of hyperedges, that is of sets of vertices of any size. So, for instance, for $V = \{a, b, c\}$ a possible hypergraph is $H = \{\{a, b, c\}, \{\{a, b\}, \{a, b, c\}\}\}$, as well as $H' = \{\{a, b, c\}, \{\{a, b\}, \{b, c\}\}\}$. Note that H' is the graph of previous note. In fact, graphs are a particular case of hypergraphs.

†††Assuming we can distinguish atoms of the same element, as typical in molecular representations, the hypergraph model for B_2H_6 is {{B,B,H,H,H,H,H,H},{{B,H},{B,H},{B,H},{B,H},{B,H,B},{B,H,B}}}, while that for H_2O is {{H,H,O},{{H,O},{H,O}}}.

‡‡‡A more detailed list of references on counting of substances or subregions of the chemical space can be found in ref. 41.

§§§This comes from Bekenstein bound, which sets up an upper limit on the thermodynamic entropy, or Shannon entropy $\mathscr{H}$, for a given finite region of physical space with a finite amount of energy. It also sets up an upper limit for the amount of information required to perfectly describe a physical system down to the quantum level. Bekenstein bound states that $\mathscr{H} \leq \frac{2\pi RE}{c \ln 2} = \frac{2\pi RMc}{\ln 2}$, which when using the mass of the universe, 10^{53} kg, and its diameter, 8.8×10^{26} m, yields $\mathscr{H} \leq 2.268 \times 10^{123}$ bits. A more refined calculation is found in ref. 46.

versions of all possible Gmelin and Beilstein handbooks that at least mention each substance of the chemical space.

Having discussed the size of the number of substances, I proceed to analyse the role of relations among substances, which constitute the "glue" holding together the chemical space and which actually turn the set of chemicals into a formal space endowed with a notion of nearness.

2.2 Covering the space

As stated in Section 1.1, there are two main approaches to nearness in the chemical space: similarity and synthetic separation. Let us first consider the case of similarity.

By experimentation and theoretical work, chemists have defined and determined several substance properties, ranging from the source of the substances, their chemical and physical properties, to molecular models assigned to each substance.[47] Substance properties are then used to classify chemicals into physiological classes such as medicines and poisons; or into chemical classes such as alcohols, amines, and several other groupings. It is based on those classifications that the complexity of individual chemicals with individual properties may be reduced and which allows for estimating new compounds and their properties.[2,48]¶¶¶.

Chemical classification leads to sets of chemicals that may be regarded as hyperedges of a hypergraph spanning the whole chemical space. These classes, or hyperedges, also correspond to *coverings* of the set of substances.‖‖‖‖‖‖ That is, they endow the set of substances with subsets of similar compounds, which are not necessarily disjoint (Fig. 1a) and that when taken together do not leave out any substance. This is seen, for instance, with the classification of amino acids, which belong to the class of amines and also to that of carboxylic acids. Interestingly, amino acids also constitute a class of compounds at the intersection of amines and carboxylic acids.

Such a rich collection of classes of similar chemicals turns the chemical space into a *topological space*.**** These spaces generalise the idea of *metric spaces*, which are made by sets of objects, for which it is possible to measure distances. In a metric space, the distance defines *open sets*, which contain objects close to a reference object; instances of open sets are open intervals on the real line.[8] Topological spaces generalise the notion of nearness by using open sets, which may or not come from a distance. In this setting, classes of similar compounds may be taken as open sets allowing for studying further properties of the chemical space and of mappings of the chemical space into other spaces, for instance relating chemical compounds with their properties. Further topological concepts can be applied to those mappings, for example continuity, which is central for QSAR studies, as they allow, among other possibilities, to study similarity cliffs[52] in a formal way.††††

The question that arises is: if the chemical space is so big, what can we say about its coverings, that is about the possible similarity classes we can define upon substances? Ultimately, what can we say about the open sets of its topology? If we gather together all possible substances of the chemical space in the set $\mathbb{C}$, the number of similarity classes must not exceed the possible number of coverings of $\mathbb{C}$. That is, it cannot be larger than $\frac{1}{2}\sum_{k=0}^{n}(-1)^k\binom{n}{k}2^{2^{n-k}}$, with n indicating the number of chemicals in $\mathbb{C}$.[50] We can decide whether n is either $\mathscr{C}$ or $\mathscr{U}$, as discussed before. This latter case would lead us to $C(\mathscr{U}) = \frac{1}{2}\sum_{k=0}^{10^{200}}(-1)^k\binom{10^{200}}{k}2^{2^{10^{200}-k}}$ possible classes of similar substances.‡‡‡‡ Again, this number is by far much bigger than the possibilities of our universal library. Therefore, any approach to systematically pinpoint the most relevant similarity classes of the chemical space is welcomed given the high likelihood of obtaining non-interesting ones by random selection.§§§§

In topological terms, as coverings have associated topologies¶¶¶¶ if the subsets in the former meet the conditions mentioned in note **** to warranty the notion of nearness, the question that arises is about the number of topologies. For a given

¶¶¶As discussed later, this is a case where topological concepts such as *continuity* become important. The general idea here is that there is a set of substances that may become a space (chemical space) because of resemblance of substance properties. Substance properties, in turn, constitute a space, as property values may be embedded in a metric space (a space where the notion of nearness among its objects is given by a metric, or distance). Hence, if we call the property space $\mathbb{P}$ and the chemical space $\mathbb{C}$, then the classification of chemicals by properties corresponds to a mapping f from $\mathbb{P}$ to $\mathbb{C}$. The prediction of new compounds and of their properties is associated to the reverse mapping, that is from $\mathbb{C}$ to $\mathbb{P}$. In formal terms, we can say that f is continuous at a property $p \in \mathbb{P}$ if and only if for each neighbourhood M of $f(p)$, $f^{-1}(M)$ is a neighbourhood of p. Here the notion of *neighbourhood* of an object is to be thought of as containing all the objects of a set that are sufficiently close to the object in question. More on these topological ideas in ref. 49.

‖‖‖‖‖‖A *covering*, or a *cover*, of a set X corresponds to a collection of subsets of X whose union is X.[50,51] If $X = \{a, b\}$, a covering of X is $\{\{a, b\}, \{a\}\}$. There are, actually, five coverings for this X. Besides the already mentioned, the other four are $\{\{a\}, \{b\}\}$, $\{\{a, b\}\}$, $\{\{a, b\}, \{b\}\}$ and $\{\{a, b\}, \{a\}, \{b\}\}$. See main text for an expression to determine the number of coverings of a given set.

****A *topological space* is a set X endowed with a collection τ of subsets of X satisfying: (1) $\varnothing$, $X \in \tau$, (2) for every $S \subseteq \tau$, the union of subsets in S is in τ and (3) for every finite $S \subseteq \tau$, the intersection of the subsets in S is in τ. τ is called a *topology* on X and the sets in τ are called *open sets* of the topology.[51]

††††Similarity cliffs turn out to be cases of lack of continuity between the space of properties and that of substances (chemical space). Interesting topological ideas of straightforward application to the chemical space can be found in the works by Stadler[53,54] and in some early studies of my research group.[55,56]

‡‡‡‡To have an idea of the rapid growth of the number of coverings for a set X of n elements, that is $C(n)$, I list the values of $C(n)$ for $n = 1$ to 7: 1; 5; 109; 32; 297; 2, 147, 321, 017; $\sim 9.223 \times 10^{18}$ and $\sim 1.701 \times 10^{39}$.[50]

§§§§The number of coverings with chemical meaning can be reduced by noting that relevant coverings in chemistry must not involve the whole set of chemicals. Coverings containing the whole set of chemicals are trivial, as the presence of this set in the covering indicates that the whole set of chemicals is similar. Therefore, if we leave out those coverings that include $\mathbb{C}$, the number of *proper coverings* for $\mathbb{C}$ is given by $C'(n) = \left(\frac{1}{2}\sum_{k=0}^{n}(-1)^k\binom{n}{k}2^{2^{n-k}}\right) - \frac{2^{2^n}}{4}$, with n being the number of elements in $\mathbb{C}$.[50]

¶¶¶¶See next note, where it is observed that the topologies there listed correspond to coverings, which also include $\varnothing$. I note in passing that the number of topologies is lower than the number of coverings, as here observed by counting the four topologies for $X = \{a, b\}$ and the five coverings for the same set (see note ‖‖‖‖‖‖). See further discussion in main text.

set, the number of topologies corresponds to the number of *preorders* defined over the set.[57]‖‖‖‖‖‖‖ For large sets of n objects, this has been found to be proportional to $n!/2(\log 2)^{n+1}$,[58] which for $n = \mathscr{W}$ amounts to have more than $10^{200}!/2(\log 2)^{10^{200}+1}$ topologies. Despite the computational complexity associated to determining the number of topologies, there are studies characterising the kinds of topologies arising from similarity studies,[59] which drastically reduce the complexity of the calculation of possible topologies of relevance for the chemical space.

Although approaching the chemical space from a similarity stand is compelling, aiming at its complete topological description turns difficult for practical reasons. For example, not all known substances of the space hold the same sort of properties, or for some of them the property cannot be measured, which poses difficulties for the application of the mapping between substance properties and substances discussed in note xiv. Examples include the molecular weight of polymers or the crisp composition of alloys. This occurs thanks to the great diversity of compounds constituting the chemical space. Therefore, instead of trying to explore the structure of the whole chemical space, studies are restricted to subregions of the chemical space, for instance of new materials, binary compounds, or oral-drug-like substances, where particular properties of the involved compounds lead to meaningful coverings, which may be taken to define local topologies and to use them. What is the topology of the oral-drug-like space? How can it be used to shed light on the limits of this subspace and on the mappings to other spaces of chemical interest, such as those of protein-receptor interactions?[60]

2.3 The seams of the chemical space

A further relation connecting substances and building up a chemical space is provided by chemical reactions (Fig. 1b). In this setting, substances that are connected by few synthetic steps are regarded as closer than all others either not connected by chemical reactions or requiring many synthetic steps to be connected. The question that arises is about the theoretical number of reactions connecting a given set of substances. The answer to this question depends on the model we have for treating reactions.

Over the years, different mathematical models have been proposed to encode chemical reactions. Some of them model the dynamic behaviour of substance concentrations during the chemical transformation using mathematical settings ranging from graph and hypergraph theories to Petri nets.[54,61,62] These approaches aim at finding, for instance, the kind and amount of substances produced after certain time, given particular amounts of starting materials. Solving these questions requires knowing reaction rates, as well as the connectivity patterns of educts and products participating in the reactions. The basis for any model of reactions is the underlying network of chemical reactions connecting substances. It is upon this network that information on reaction kinetics is added to solve the above mentioned dynamical questions on the amount of educts and products. Given the fundamental role of the underlying network of reactions, let us focus on it for the purposes of building up an alternative chemical space to the one discussed in the previous section based on similarity.

Reaction networks have been modelled using graphs, Petri nets and directed hypergraphs.[54,61,63] An example of the former is shown in Fig. 2a and the hypergraph model is shown in Fig. 2b. Petri nets may be considered as refinements of the hypergraph model. Therefore, I will discuss the graph and hypergraph models.

The graph model is the most simple approach to the network. Here I discuss the *educt–product model* that encodes the directionality of the chemical transformation. In this setting, if there is a reaction in which substance y can be obtained from another x, this is encoded as $x \rightarrow y$ (Fig. 2a).

The organic chemistry part of the network of chemical reactions, for the period 1850–2004, was studied using the educt–product model.[18] By analysing the statistics of this network it was found that chemists have had preferences in the way they relate substances through reactions. This was evident in the fact that only a few substances are involved in a high number of educt–product connections; which contrasts sharply with the vast majority of substances which have very few connections to any others. These results were confirmed by my research group when analysing the entire network of the chemical space from 1800 up to 2015.[10] Examples of frequently used educts (*toolkit compounds*) are: acetic anhydride, methyl iodide and formaldehyde.[10] For instance, acetic anhydride has been the most used educt ever since 1940 and it has been part of the top-10 of most used educts since 1880 (about 30 years after its synthesis[64]).

Substances have been also classified according to their degrees of connectivity in the network into *core compounds*, corresponding to toolkit substances; *peripheral substances*, obtained within no more than seven synthetic steps from core substances; and *island substances*, corresponding to compounds neither synthesised from core nor from peripheral substances.[65] Other studies of parts of the chemical reaction network include further statistics, such as assortativity, average shortest paths, clustering coefficients and betweenness centrality.[66] These statistics coincide with noteworthy results indicating a small set of frequently used chemicals that connect, in few synthetic steps, a large proportion of the remaining network.[10,18,65]

From the perspective of the educt–product model, the theoretical number of chemical reactions is given by the densest possible reaction network, which entails connecting every single substance of the chemical space with the other substances. Thus, if we have $\mathscr{C}$ substances, we cannot produce more than $\mathscr{C}(\mathscr{C}-1)$ reactions, which is, of course, a chemical space we cannot store in our universal library.

But the educt–product model disregards an essential piece of chemical information encoded in every reaction, namely that chemical reactions relate two sets of substances in a directed fashion, rather than couples of single substances. These two sets are the set of educts and the set of products, which are related by the temporality of the chemical transformation of the former into the latter. Thus, the educt–product model does not include the important AND relation among educts indicating that they react

‖‖‖‖‖‖‖For instance, for the set $X = \{a, b\}$ there are four possible topologies on X: (1) $\{\varnothing, X\}$, (2) $\{\varnothing, \{a\}, X\}$, (3) $\{\varnothing, \{b\}, X\}$, (4) $\{\varnothing, \{a\}, \{b\}, X\}$. Note that (4) corresponds to the *discrete topology* on X. Topology 1 is often called the *trivial topology* or *indiscrete topology* on X.[8]

together, nor it informs about the AND relationship among the products, which hold them together in the reaction vessel once the reaction has taken place.***** For example, in the reaction A + B → C, the educt–product model indicates that C can be produced from A or from B, but it does not inform whether C is produced by a rearrangement of A, or by the reaction of A with B (A AND B in the set of educts) (Fig. 2a).

A suitable model encoding the AND relation among products and among educts is the *directed hypergraph model*. Fig. 2b shows the application of this model to the previous reactions analysed with the educt–product model. The hypergraph model shows that to obtain C, A requires to react with B and that there is no direct reaction from A to B. Thus, the hypergraph model actually depicts the asymmetric relation between educts and products. Educts lead to products but not the converse, and if so, as in reversible reactions, the reverse reaction is modelled as a new reaction where products and educts of the former interchange roles.

Just recently, hypergraph network statistics have been reported, such as clustering coefficients, spectral properties and curvatures,[67–69] which provide local and global information of the network. These statistics have not been applied yet to the network of chemical reactions, but they will inform about the presence of central reactions using educts that are obtained by different reactions to produce substances in turn used as starting materials of many other reactions. Likewise, those statistics will shed light on the global structure of the reaction network and its evolution.

In the hypergraph setting, the possible number of chemical reactions is therefore given by the number of directed hypergraphs one can build up on a given set of chemicals. This corresponds to the number of ordered pairs of subsets of chemicals one can form that are disjoint.††††† Hence, the possible number of directed hypergraphs over $\mathscr{C}$ substances is given by $3^{\mathscr{C}} - 2^{\mathscr{C}+1} + 1$.‡‡‡‡‡ Again, our universal library cannot afford storing this information.

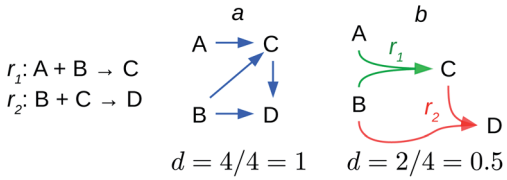


Fig. 2 Chemical reaction models. Two chemical reactions r_1 and r_2 (left) giving place to (a) a network modelled as an educt–product graph and as (b) a directed hypergraph. Blue arrows are called *arcs*, while green and red ones *directed hyperedges*. The density (d) of each network is shown. If either the number of arcs or of hyperedges is given by E and the number of substances by n, then d is calculated as the ratio E/n.

A more realistic account of the possible number of reactions must consider that actual reactions take place by the interaction of quasi-molecular species. Therefore, expecting to have a single collision among all $\mathscr{C}$ species is very unlikely. In fact, chemists have traditionally combined no more than a handful of substances in their chemical reactions.§§§§§ We recently found that the most traditional combination of educts involves no more than two.[10] The question is, how many possible reactions involving no more than two educts are theoretically possible? It is actually $\mathscr{C}(\mathscr{C} - 1)/2$, which is the number of couples of substances one could bring together to the reaction vessel. Hence, even if chemists keep performing binary combinations, and even if the number of possibilities is much lower than the possible number of reactions, still all binary combinations of educts cannot be annotated in our universal library.¶¶¶¶¶.

In Section 2.2 I discussed how similarities lead to topologies for the chemical space, which provides a formal approach to treat nearness on the space. I note in passing that directed hypergraphs also lead to topologies, where open sets for each chemical are defined in terms of the hyperedges of different order the substance belongs to. By hyperedges of different order I mean the different sets of substances associated to a substance in reactions of n steps, with n indicating how far (number of reactions) we want to go to define the open sets of the substance.|||||||| Interesting results on the application of topological ideas to chemical reaction networks are found in the works by Stadler and collaborators,[54,72–74] which not only provide suitable tools for the analysis of the chemical network but for studies in the origin of life.

So far, the message I have tried to convey is that the chemical space is actually huge, so huge that the universe does not suffice to store the labels of its substances let alone those of their

*****Part of the abstraction of reaction models entails focusing the attention on the starting materials and the final products, which disregards the appearance of non reacting educts in the final reaction mixture.

†††††Here we are assuming the simple case of reactions whose educts are not part of the products. That is, autocatalytic reactions such as A + B → 2B or 2B → A + B are not considered. I note in passing that both the educt–product and the directed hypergraph model may incorporate stoichiometric details of the reactions by weighting their arcs or hyperedges with stoichiometric coefficients.

‡‡‡‡‡A directed hypergraph on a set X consists of an ordered pair of disjoint subsets of X. If X has n elements, each subset of size $k \leq n$ is disjoint to the remaining 2^{n-k} subsets in X. As the empty set is considered in this counting, but hypergraphs connecting sets of chemicals with the empty set are meaningless, then the ordered pair between the set and the empty set is disregarded. Therefore, each subset of size k (excluding the empty set) is disjoint to the remaining $2^{n-k} - 1$ subsets in X. As each subset of size k is chosen from n elements, then the number of disjoint pairs of subsets for sets of size k is given by $\binom{n}{k}\left(2^{n-k} - 1\right)$. Finally, the total number of disjoint pairs of subsets of any size $k \leq n$ is given by $\sum_{k=1}^{n}\binom{n}{k}\left(2^{n-k} - 1\right)$, which can be expressed as $\sum_{k=1}^{n}\binom{n}{k}2^{n-k} - \sum_{k=1}^{n}\binom{n}{k}$. Each of these addends, by using the binomial theorem, can be expressed as $3^n - 2^n$ and $2^n - 1$, respectively, which finally leads to $3^n - 2^{n+1} + 1$.

§§§§§It is actually a chemical challenge to go beyond the four educts in a chemical reaction in the so-called multicomponent reactions.[70] There are, nevertheless, famous examples of reactions involving more than four components or substances, namely Dömling and Ugi 7-component reaction.[71]

¶¶¶¶¶The temporal discovery of n-ary compounds between 1800 and 1869 is reported in ref. 27.

||||||||Thus, for the substance A, given the chemical space made of reactions A + B → C + D and C + E → F, we can define a 1-order open set as {A, B, C, D} and a 2-order open set as {A, B, C, D, E, F}. Further refinements can be added by incorporating the directed nature (educts → products) of chemical reactions.

topologies and reactions. In the second part of this perspective, instead of asking for the limits of the chemical space, I turn to what has been achieved on the exploration of the chemical space, from a network perspective.

3 The evolving chemical space

A further aspect of the chemical space is its temporality, which leads to questions about its dynamics. That is about the temporal change of number of known substances and of reactions. Unlike social networks where actors enter and leave the network, for instance driven by life-spans, chemical "actors," that is substances, can only enter the network. Even if we do not have an actual physical repository storing a sample of each reported substance in the history of chemistry, whenever the isolation or the synthesis of a substance is reported in a publication, any chemist equipped with the technology reported in the publication may "bring to life" the substance in question.****** Therefore, the number of substances of the chemical space can either increase or remain constant.†††††† The latter case involves no chemical activity at all. As we will see, this has never happened in the history of chemistry.‡‡‡‡‡‡

If the very likely scenario for the chemical substances is to increase in number, what about the seams of the chemical space? That is, what about chemical reactions? As well as substances reported in the literature, chemical reactions can be brought to life with the suitable starting materials and technologies allowing for reproducing the experimental settings of the literature. What are the possibilities for chemical reactions? If the number of chemicals were expanded mainly by extractions, the number of new reactions would drop. If synthesis were the driving force of chemistry, the number of reactions would grow.§§§§§§.

These dynamical aspects of substances, reactions and their interplay in the temporal unfolding of the chemical space are the subject of this section.

3.1 Exponential discovery of substances

The first study analysing the growth of the cumulative number of reported chemicals over the history was published by Schummer in 1997 and spanned the period 1800–1995.[78] He manually analysed the indices of printed sources, including handbooks of organic and inorganic chemistry. An exponential growth with an annual rate $r = 5.5\%$ was found, which amounts to doubling the number of new substances every 13 years. The digitisation of several of the sources used by Schummer led to the second study on the growth of the cumulative number of reported substances. In such a computational study, tailored to the organic chemistry chemical space between 1850 and 2004, Grzybowski and collaborators analysed the Beilstein database (now part of Reaxys). As Schummer, they found an exponential growth, which was actually divided into two growth periods. From 1850 to 1900, with $r = 8.3\%$ and from 1900 until 2004 with $r = 4.4\%$.[18] Hence, organic chemistry, in the second half of the 19th century, doubled the number of known substances every 8.3 years; while in the 20th century this prolific production of chemicals dropped to half 19th-century growth rates and led to doubling times of 16 years.

A more recent study of the chemical space was reported by my research group. It spans the period 1800–2015 and analyses all substances involved in single step reactions and published in academic journals, as retrieved from Reaxys database. Such a study was not devoted to organic chemistry alone as in Grzybowski' study,[18] it rather aimed at a holistic depiction of the chemical space, as it was also Schummer's aim in his 1997 study.[78] We found a rather stable growth rate of $r = 4.3\%$ (Fig. 3) and no signs of an early period growing very fast and another slowing down as in the account by Grzybowski.¶¶¶¶¶¶.

By contrasting Schummer, Grzybowski and our results, the rapid growth of organic chemistry before 1900 was observed by Grzybowski because the important contribution of inorganic chemistry before 1900, and especially before 1860,[10,79] was not taken into account.||||||||||| After 1900, as the chemical space was mainly populated by organic compounds,[10,27] Grzybowski's results agree with those of Schummer and my group.

The rapid and constant expansion speed of the number of new substances of the chemical space ($r = 4.3\%$) indicates that all over the history about each 16 years chemists have doubled the number of new substances reported. This speed can be expressed as that the number of new chemicals reported by the chemical community in 2015 roughly amounts to all substances reported between 1800 and 1992. "That is, in a single year of contemporary chemistry, chemists produced the same number of new substances as reported in 192 years of the history of chemistry. This is the dramatic speed at which the chemical space grows"![80]

******To make justice, I shall acknowledge the several experimental difficulties historians of science face when reproducing experiments originally conducted centuries ago.[75–77]

††††††An interesting question is determining the actual number of active chemical substances participating in chemical reactions in different periods of the history of chemistry.

‡‡‡‡‡‡Not even World-War (WW) periods have prevented chemical production. See Fig. 3, where it is observed that these social setbacks caused drops in the production of new chemicals, but they never reached zero production of new chemicals. The effects of WWs upon the chemical space are discussed in Section 4.2.

§§§§§§The role of synthesis for expanding the chemical space is discussed in Section 4.

¶¶¶¶¶¶Our study was based on the annual report of new substances, rather than in the cumulative number of these substances, as in Schummer' and Grzybowski' studies. For the sake of comparison, I calculated the cumulative values based upon the fitting equation of our study (see equation in Fig. 3). To avoid problems with the initial values of the cumulative distribution, caused by the lack of figures before 1800, I estimated those figures by extrapolating backwards using the fitting equation of our study. These values were appended to our 1800–2015 values of annual number of new chemicals and the cumulative distribution was calculated. The corresponding fitting equation for this distribution led to a growth rate of $r = 4.3\%$. This result is not surprising as exponential growths enjoy the particularity of being correlative with their exponential cumulative values.

|||||||||||Quantitative evidences of the important role of inorganic chemistry in the 19th century are found in ref. 10 and 27. This is seen, for example, by analysing the distribution of platinum metal compounds (Fig. 4), as well as those of alkali metals over the history.[27]

3.2 Exponential wiring through reactions

Are the seams of the chemical space also growing exponentially? This question was partially solved by Grzybowski and collaborators in 2005 when analysing the network of organic chemistry between 1850 and 2004.[18] The same trends for the growth of cumulative number of new substances were observed for reactions, that is a rapid wiring of the network between 1850 and 1900 (r = 8.7%) and a subsequent slowing down (r = 3.8%). Nevertheless, this exponential growth of single reactions does not necessarily lead to an exponential growth of classes of chemical reactions.[81]******

By analysing reactions gathered in Reaxys database, Grzybowski and his team found that new classes of chemical reactions have grown at most linearly from 1900 up to 2016.[81] They also found that the number of new classes of reactions becoming popular, that is frequently used to wire the chemical space, has been very small for the period studied. Their results show a core of popular reactions classes. The top 10 of these preferred reactions, arranged in decreasing order is:[81]

1. Amide synthesis from carboxylic acid and amine.
2. Alkylation of alcohols or phenols with primary or secondary halides/*O*-sulfonyls.
3. Hydrolysis of esters (carboxylic acid as the main product).
4. Acylation of amines.
5. Reduction of carbonyl to alcohols.
6. Esterification.
7. Alkylation of amines with primary or secondary halides/*O*-sulfonyls.
8. Oxidation of alcohols to aldehydes/ketones.
9. Acylation of alcohols/phenols.
10. Buchwald–Hartwig coupling/nucleophilic aromatic substitution with amines.

Similar preferences for some few classes of chemical reactions have been found in medicinal chemistry.[82] For instance, there is a strong preference toward *para* substitution in phenyl rings within drug discovery research.[82]††††††

If Grzybowski's results on the selection of some few classes of reactions to expand the organic chemistry space actually span the whole chemical space, one may wonder about the shape of the space driven by an exponential growth of substances and reactions and confined to the repeated use of some few reaction classes. This is discussed in Section 5. Before going in this direction, let us explore in detail the interplay of substances and reactions expanding the space.

3.3 Ups and downs of the interplay between substances and reactions

One of the first questions one may ask about the interplay of substances and reactions is whether new substances are actually used in chemical reactions. We addressed this question for the whole chemical space between 1800 and 2015 (ref. 10) and found that once substances are synthesised, they are very seldom used in new reactions.[10] There is, nevertheless, a small fraction of substances (toolkit compounds) that are often used to expand the chemical space. Some of these heavily used substances for the period 2000–2015 are acetic anhydride, methyl iodide, benzaldehyde, formaldehyde, trifluoroacetic acid, phenylacetylene and benzyl bromide.[10] There are also frequent reaction products, which include traditional side products such as carbon dioxide and water, but also targets such as metallic oxides (CuO, ZnO, NiO and CoO), often synthesised ever since the 1980s with an important surge between 2000 and 2015.[10]‡‡‡‡‡‡ There is however much more the use of

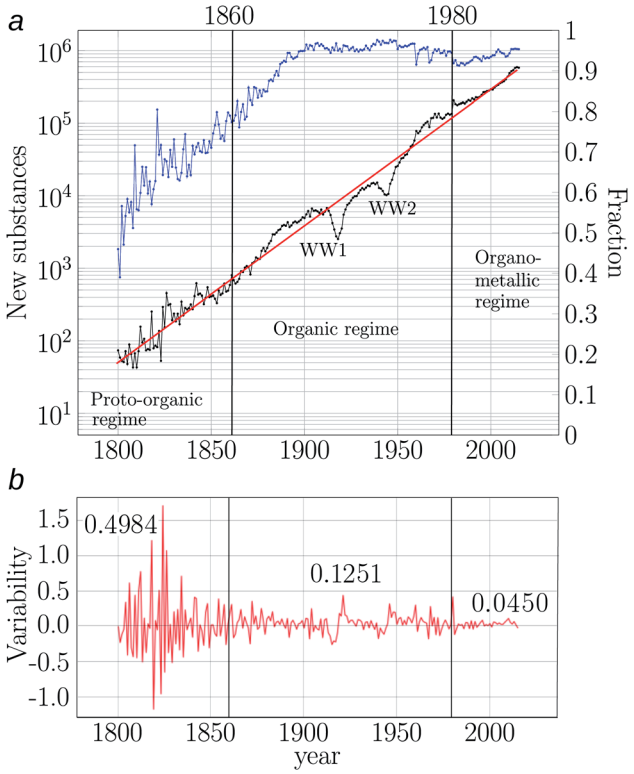


Fig. 3 The expanding chemical space. (a) Middle-black curve: annual growth of number of new substances between 1800–2015 (left axis). The exponential equation fitting the growth is indicated as a red straight line with equation: $s_t = 51.85 \times 10^{0.04324(t-1800)}$ (R^2 = 0.9829, residual standard error = 0.3575). Upper-blue curve: fraction of new synthesised compounds to the total of new ones (right axis). (b) Variability of the annual output of new substances, calculated as ln s_{t+1} − ln s_t. Figures close to the red curve in (b) correspond to the average variabilities for the periods (regimes) 1800–1860, 1861–1980 and 1981–2015, demarcated by vertical lines. The three *statistical regimes* resulting from the annual variability of the number of new substances are indicated (proto-organic, organic and organometallic regimes), with transitions occurring in 1860 and in 1980 (vertical lines). The effects of the World Wars (WWs) are indicated close to the black curve. Plots adapted from Fig. 1a and S1 in ref. 10.

******By a class of chemical reactions I mean, for instance, Diels–Alder reaction.

††††††This *p*-preference over *meta* and *ortho* phenyls is said to be caused by historical models of medicinal chemistry where *p*-substituted compounds were more easily accessed, and further reinforced by Topliss in 1972, who argued that if a phenyl was active, the *p*-Cl phenyl should be made because of ease of synthesis and hydrophobicity driven potency effects.[83]

‡‡‡‡‡‡Other often used and reported products over the history are found in ref. 10.

preferred educts than the synthesis of heavily produced substances. The most used educts have participated in about 70 000 reactions, while the most synthesised targets have been produced by no more than 2000 reactions.[10]

A further proxy indicating the interplay between substances and reactions is the density of the network of the chemical space, defined as the number of edges per node.§§§§§§§ Hence, in the educt–product model used by Grzybowski and collaborators, network density is calculated as the ratio between the number of arcs and the number of substances (Fig. 2a). Grzybowski's results on the density of the organic chemistry network between 1850 and 2004 show an initial "wiring" period between 1850 and 1885, where chemists reported more arcs (roughly speaking reactions) than substances, increasing network density. The wiring period was followed by a period where chemists reported more substances than arcs leading to low density values.[18] Grzybowski's results indicate that current density values, of around two arcs per substance, are far from 1885 ones, where the network achieved about four arcs per substance. Hence, since about the turn of the 20th century, chemists have been busier adding substances to the space than wiring them.

These results open several questions. If we consider the most appropriate model for chemical reactions of directed hypergraphs along with the whole chemical space, rather than its organic part, what will the density values look like? Lower values of density are expected because directed hypergraphs count actual reactions rather than arcs (Fig. 2). In the educt–product model, the number of arcs results from multiplying the number of educts and of products of each reaction (Fig. 2a). This difference is observed in Fig. 2b, where the two reactions (two directed hyperedges) are expanded in the educt–product model to four arcs.¶¶¶¶¶¶ The advantage of the network density according to the directed hypergraph model is that it has a direct chemical interpretation as it accounts for the actual number of reactions per substance.

Further questions based on the density are of historical and philosophical nature. Which conditions facilitated the wiring of the network before the turn to the 20th century and which ones triggered the subsequent emphasis on substances? Is there a preferred mode (wiring or substance based) of expanding the chemical space that is optimal for speeding up chemical knowledge?

Density values of the network of chemical reactions seem to indicate that historical events play a role in the expansion of the network. The role of past events and of human intervention in the unfolding of the network is analysed in the next section.

4 Anthropogenic effects upon the expansion of the chemical space

The chemical space, besides being an object of chemical interest, is a historical object driven by the interplay of social and semiotic factors.[2] The former are related to the social structures and forms of organisation giving place to the chemical space and the associated technologies for its expansion. Semiotic factors include the theoretical structures of chemistry, as well as the language of the discipline and the communication channels used by chemists.[2] These factors are tied together by the human factor. The chemical space is driven by chemists, which act on the thermodynamic, kinetic and quantum chemistry possibilities of the chemical transformations.

An instance of anthropogenic aspects driving the chemical space is the finding that since the early years of the 19th century, more than half of the reported substances have been synthesised by chemists (Fig. 3).[10] Actually, by the turn of the 20th century, almost all reported substances were synthesised and this trend has mainly remained so ever since (Fig. 3). It is traditionally accepted that synthesis in organic chemistry began after Wöhler' synthesis of urea in 1828.[64,84] Nevertheless, by calculating the fraction of new chemicals synthesised and extracted over the history, we found that more than half of the new substances have come from synthesis ever since the dawn of the 19th century (Fig. 3). In particular, already 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 before that. Hence, the kicking-off landmark event of 1828 must be considered a myth. Today about 95% of the reported substances come from synthesis (Fig. 3). Thus, the chemical space and its expansion is the product of chemists' ingenuity.||||||||||||.

The collection of reaction conditions used to expand the chemical space constitute a further example of the anthropogenic driving force guiding the unfolding of the chemical space. Almost all reactions have been performed at 1 atm and 25 C,[89,90]

§§§§§§A more traditional network density measure is the ratio between the number of actual edges or arcs in a network and the theoretical number of possible edges or arcs. Hence, if the network houses n vertices, its density is given by $d = 2N/(n(n-1))$, considering vertices linked by edges. Here N stands for the actual number of edges in the network. If the network is modelled as a directed graph (educt–product model), its density is given by $d = N/(n(n-1))$, with N indicating the actual number of arcs. For a network modelled through directed hypergraphs, $d = N'/(3^n - 2^{n+1} + 1)$, with N' representing the actual number of directed hyperedges (reactions). See Section 2.3.

¶¶¶¶¶¶As each reaction in Fig. 2 is of the form $x + y \rightarrow z$, then the number of arcs is 2×1. Hence, the two reactions (directed hyperarcs) of the figure amount to 4 arcs. In the directed hypergraph model, the density of the chemical Plots adapted from Fig. 1a network by year t (d_t), consisting of s_t substances and r_t reactions, is given by r_t/s_t, where r_t corresponds to the amount of directed hyperedges. In the educt–product model, $d_t = \sum_k (e_i \times e_j)_{k,t}/s_t$, with $(e_i \times e_j)_{k,t}$ indicating the number or arcs provided by reaction k, which is known by year t. Here e_i and e_j stand for the number of educts and of products, respectively, in reaction k.

||||||||||||As a human construct, the size of the chemical community is an important factor to consider when analysing the chemical space.[78] Several studies coincide in reporting a historical exponential growth of the chemical community. See for instance the chapters in ref. 85. The question that arises is about the conditions allowing for the exponential growth of the community. They involve the social interest for chemistry, which is attached to the public image of this science and its ability to trigger innovations able to pull further resources to accommodate more chemists. The public image of chemistry is also regulated by ideologies and economic factors.[86] They range from the changing roles in public acceptance of alchemy in the antiquity[87] and middle ages to the pro-scientific ideologies of the 19th century and to the present antiscientific attitudes.[88]

that is at the ambient conditions at which we have evolved. The influence of our circadian clocks upon the unfolding of the space is also evident, as the duration of chemical reactions corresponds to simple fractions of multiple integers of our day–night cycles.[89,90]

Human decisions or preferences have also left its mark on the expansion of the space. A consequence of the use of some few reaction classes and of a selected set of toolkit compounds is the uneven span of the chemical space at the level of molecular structures. In a 2019 study, Lipkus and coworkers analysed part of the organic chemistry space between the years 2009 and 2018 and found that there is an uneven distribution of molecular frameworks (molecular backbones consisting of all ring systems and all chain fragments connecting them).[91] A consequence of this uneven distribution is that the likelihood of reporting a framework in the future becomes proportional to the number of times the framework has been already reported.*******

Anthropogenic factors play also a major role in the present synthesis plannings and actual syntheses based on the artificial intelligence (AI) exploration of the chemical space. Although it is true that AI approaches, when coupled to robotic synthesisers, speed up chemical discovery, they do it in a rather conservative manner, as the training set of AI algorithms is based on past synthesis, which perpetuates the anthropogenic biases of the known chemical space.[92] As suggested by Grzybowski and collaborators, the same ongoing automatisation and robotisation of chemical synthesis may become instrumental to free us from the learnt manner in which we have expanded the chemical space. The idea is to rapidly repeat synthetic protocols under different conditions to enlarge the training set with a richer dataset that includes a large number of failed reactions. In the best scenario, this may lead to actual discoveries of synthesis plans and of chemicals with novel properties. Although the proposal of "playing the tape of the chemical space again" under different conditions sounds interesting, we need not to forget the limits of the exponential growth of the space. Let us suppose that we suffice with playing only 25% of the current chemical space. That is, that we need a training set of 25% the number of current substances, which is about 18, 500, 000.††††††† So, re-playing 25% of the tape of the chemical space requires synthesising about 4 600 000 substances, which implies entirely devoting the next five years to the production of the training set. How much of the annual outcome of chemistry are we willing to spend in re-playing the tape of the chemical space to free us from our anthropogenic bias to achieve more reliable AI results fostering innovative chemistry?

Even if AI algorithms coupled to automatised and robotised devices can take us out of the path-dependent box created by the expansion of the chemical space, can we really afford the supply of starting materials to re-play the tape? At least for the moment this is not affordable, as for instance our capacities to reuse chemicals such as metals are very poor. A further point of re-playing the tape is the storage of the new play. I have discussed the impossibility of annotating the chemical space in its totality. This poses serious difficulties for annotating an enlarged chemical space containing multiple repetitions of the current one under different conditions.‡‡‡‡‡‡‡ This further supports the important need of methods to annotate what is chemically interesting, not only for us, but rather for the future generations. The challenge is to determine the requirements of the future.

4.1 Regularisation of the expansion

I have discussed so far how regular the growth of the number of substances and reactions has been across history. These results depict the cumulative effect of a community of chemists working mostly independently in different corners of the globe at different periods of the history. In quantitative terms, the result is clear. There is a historical and rather stable growth trend for new substances and reactions, which has not been affected by theories, schools of thought and wars.[2] These conclusions are based on the historical growth rates of the number of substances and reactions. But these temporal signals are much richer in quantitative information, which provides further insight on the temporal unfolding of the chemical space.

Elaborating in this direction, we analysed the variability of the annual output of new chemicals between 1800 and 2015 (ref. 10) and found that chemical production has undergone two big transitions demarcating three statistical regimes (Fig. 3a). That is, we found three stationary processes (regimes) in the historical production of new chemicals (Fig. 3b). The first regime covers the period 1800–1860, corresponding to the highest variability in the annual production of new substances. This may be caused by the small size of the chemical community, where local setbacks in production of particular research groups could affect the global production to a large extent. This hypothesis needs to be further explored by contrasting Fig. 3a with annual data of number of active chemists.§§§§§§§ While this was the period with the highest percentage of metal compounds reported, C and H compounds nevertheless dominated during the entire period (Fig. 4). In fact, the second half of the regime was mainly characterised by C, H, N, O and halogen based compounds (Fig. 4). According to historians of chemistry, this period witnessed the rise of organic chemistry and especially

*******Although this seems to indicate a Barabási–Albert (or preferential attachment) dynamics[22] on the report of frameworks, Lipkus and coworkers also noted that the model for estimating the annual output of frameworks may be driven by stochastic processes, presumably arising from economic and other factors of the chemical practice.[91]

†††††††Estimated using S_t in Section 5.

‡‡‡‡‡‡‡Strictly speaking, these repeated spaces do not hold the same substances of the original space, as we need to discount reaction products, as failed reactions do not lead to them. Nevertheless, this reduction only refers to a hypothetical bound because variations of reaction conditions may also lead to new products. This has been shown, for instance, through the synthesis of a huge set of diverse chemicals by varying the reaction conditions under which amines and carboxylic acids react.[93]

§§§§§§§Interesting ideas and hypotheses in this direction were reported by Schummer in 1997.[78]

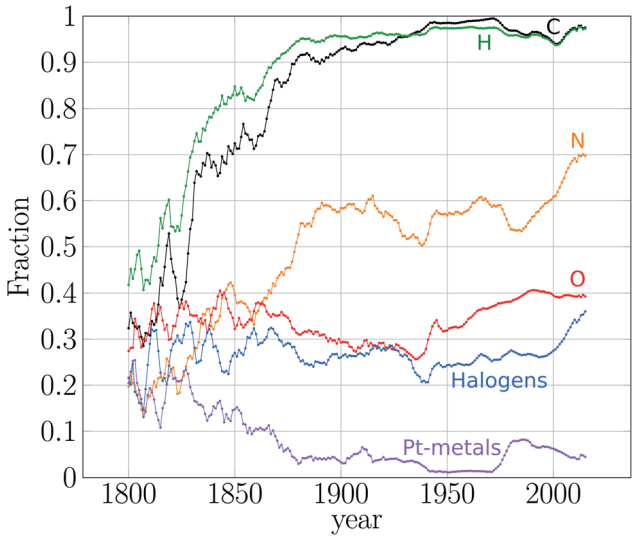


Fig. 4 Temporal behaviour of particular regions of the chemical space. Annual fraction of new compounds containing C, H, N, O, halogens, and platinum metals. These latter correspond to Fe, Co, Ni, Ru, Rh, Pd, Os, It and Pt. Distributions are convoluted using the moving average method with a five-year window. Plot adapted from Fig. 1c in ref. 10.

the changing role of this kind of chemistry, from an analytic approach to a markedly synthetic one (Fig. 3a).[94,95] Because of these features, we called this period the *proto-organic regime*.[10]

After 1860 the second regime of chemical production began, which is evident by a drastic reduction of the variability of the annual output of new chemicals (Fig. 3b). This regime was strongly driven by organic chemistry synthesis. The role of organic chemicals is evident, for instance in the large percentage of C and H compounds spanning the space in this period – by 1880 C and H compounds constituted 90% of the new substances (Fig. 4). This predominance of organic substances has remained so ever since. In fact, as early as 1870 most of the compounds were made of CHNO and the same composition is the most populated still today.[10] The rise of organic chemistry contrasts with the reduction of the percentage of compounds containing metals (Fig. 4). We called this period the *organic regime*.[10]¶¶¶¶¶¶

Historians agree that by 1860 molecular structural theory changed the practice of chemistry, which we observe in the chemical space.[79,94,96] This theory became a powerful tool used by chemists to explore in a more controlled fashion the chemical space.[2] Structural theory is to chemistry as a tourist guide is to a newcomer willing to explore a city. As well as the newcomer, chemists could explore the space in a random way, following different streets and discovering from time to time interesting spots. However, a tourist guide offers the possibility of getting directly to the most interesting sights of the city. This metaphor, nevertheless, presupposes a given space to be discovered. Chemical space, as shown here, is "invented" and the question that arises is whether awe of structural theory is not hindering the possibility of expanding the chemical space in less traditional directions.

The third regime of the expansion of the chemical space, evident by a further reduction in the annual variability of production of new chemicals, started about 1980 (Fig. 3) and, in contrast to the transition occurred by 1860, the event(s) triggering this transition are still unknown and open to discussion and common research between chemists and historians. Some possible causes could be the computerisation of chemistry. Another possible explanation is the expansion of instrumental techniques to cope with heavier substances such as macromolecules and even solid materials. This period, which is our current regime, has been dominated by organic compounds, some of those including metals. In this regime platinum metal compounds surged (Fig. 4) as well as silicon ones.[10] Here the variability of the annual production of new chemicals is the lowest of the three regimes, which indicates that more than ever chemists have regularised the year-to-year output of new compounds. We call this regime the *organometallic regime*.[10]

The historical reduction of the variability in the production of new chemicals indicates a regularisation of the expansion of the chemical space, which is reinforced by a growing community of chemists. The fact that this regularisation has occurred through drastic transitions, rather than in a continuous fashion, indicates that there are historical events affecting the unfolding of the chemical space. If historical events regularise the annual output of the chemical space, can common work among historians and chemists shed light on the driving forces leading to those crucial events? To which extent this regularisation affects the future reaches of the chemical space? I will address those questions in Section 5. In the next section I discuss further aspects of the influence of historical events upon the unfolding of the chemical space.

4.2 The toll of World Wars upon production of chemicals

In Fig. 3 two important dips are observed and coincide with World War (WW) periods. Although the literature is rich on the role of chemistry in those devastating events,[97] the same cannot be said about the role of wars upon chemical production. Recently, we quantified this later relation.[10]

We found a devastating effect of WW1 and a mild effect of WW2 upon the annual output of new chemicals. WW1 sent chemistry back 37 years and WW2 16 years (Fig. 3a). WW1 also caused a drop in the rate of chemical production three times more dramatic than the rate of WW2.[10] The reason underlying the devastating effect of WW1 is found in the social system of chemistry, which concentrated the chemical industry and research around Germany in pre-WW1 times.[98] After WW1, chemistry decentralised from Germany and other nations accommodated their research and production infrastructures to this new scheme, for instance the USA.[98]

Interestingly, WWs have not permanently affected the expansion of the chemical space, as after these events, chemical production recovered and returned to its 4.3% annual growth

¶¶¶¶¶¶The important role of organic chemistry in the evolution of the chemical space indicates that Grzybowski's results[18,65,81] on the organic chemistry side of the chemical space are likely representative of the whole chemical space.

rate (Fig. 3). This catching-up recovery phenomenon contrasts with typical production delays of other sorts, such as publication of abstracts in other disciplines.[99] Some early analysis on the possible reasons behind these phenomenon are discussed by Schummer.[78]

Although chemical production has not been affected in the long run by WWs, these events do have motivated changes in chemical research. For instance, during WW1, there was a surge in the number of As, Sb, and Bi compounds, while Al, Ga, In, and Tl decreased. N and alkali metals dropped during WW2 but S, B, P and Si benefited.[10] The surge of As compounds may be the result of the arsenic warfare agents developed during WW1.[100] P compounds began to be often reported after WW2 when P biological role was established and when P compounds started to be used in daily-life applications and as novel insecticides and other industrial materials.[101]‖‖‖‖‖‖‖‖‖‖‖‖‖

A worthwhile research subject on the chemical space involves incorporating the different facets of the social, semiotic and material dimensions driving the evolution of the chemical space into a mathematical model. Jürgen Jost and I have recently sketched the main aspects of this model, actually a complex dynamical system, and we have discussed the different sources of data to feed the model.[2] While results of this research are obtained, a less complex approximation to the modelling of the chemical space involves modelling the dynamics of the network of chemical reactions. This is the subject of the next section.

5 Modelling and estimating the evolution of the chemical space

In Section 2.3 we discussed the advantages of modelling the chemical space as an expanding network of substances related by chemical reactions and we mentioned the advantages of the directed hypergraph model over the graph model. Hence, if we are interested in estimating the future of the chemical space, a straightforward path involves modelling the dynamics of the network by incorporating probabilistic models for the appearance of substances and reactions in the network. Before discussing these models, I shall discuss the modelling of the growth of the number of substances. As available statistics exist only for the growth of substances, I will restrict the discussion to the estimation of new chemicals. The approach can be used for the estimation of new reactions as well, and in turn for the density of the network of chemical reactions.

5.1 Modelling the appearance of new substances in the chemical space

The fitting equation for the annual output of chemicals between 1800 and 2015 is $s_t = 51.85 \times 10^{0.04324(t-1800)}$ ($R^2 = 0.9829$, residual standard error = 0.3575) (Fig. 3a),[10] where s_t indicates the number of new substances reported the year t. Following the procedure discussed in note ¶¶¶¶¶¶, this equation leads to $S_t = 1310.29 \times 10^{0.04305(t-1800)}$ ($R^2 = 0.9977$, residual standard error = 0.1299), where S_t corresponds to the cumulative number of chemicals by year t. When will we reach $\mathscr{W}$ substances? That is, which t satisfies $10^{200} = 1310.29 \times 10^{0.04305(t-1800)}$? The answer is $t = 12,330.52$, which means that to attain $\mathscr{W}$ substances we would need to keep working, as we have done it for more than 200 years, some further 10 309 years, from 2022 on. If we accept Hawking's estimations that humans will turn Earth inhabitable by 2600 (ref. 103) and we suppose we are not able to keep expanding the chemical space, a further estimation indicates that by 2600 we will have reached 1.187×10^{18} substances, which when compared with 10^{200} shows the negligible fraction of the chemical space we could actually afford.

Being more optimistic, one may suppose that we are able, somehow, to increase the historical rate of expansion of the number of substances. Then, we can ask ourselves for the needed growth rate to be able to synthesise $\mathscr{W}$ substances by, say, 2050. This prompts us to increase the growth rate of discovery of new substances to $r = 12.69\%$. That is, we would need to double the size of the new substances every 5.46 years. This is indeed a very fast growth. It may bring some relief to note that even if it is a very rapid growth, it is not much faster than the growth of bacteria cultures.[104] What is the technology and the social and semiotic infrastructure to, at least, increase the growth rate of new chemicals beyond the stable rate of the last more than 200 years?

5.2 Dynamical models for directed hypergraphs

A further approach to model the evolution of the chemical space entails devising simple rules for the appearance of new substances and reactions of the chemical space. If we model the network of chemical reactions with the directed hypergraph model, this falls in the realm of network dynamics.[105]

There are different models for dynamical networks, mostly developed for graph-theoretical settings, which include the Erdős–Rényi, Barabási–Albert and small world models, among others.[22] Although these models have been studied for hypergraphs,[106–116] there are only few accounts of dynamical models for directed hypergraphs.[117–119] At any rate, a general setting for dynamical directed hypergraphs requires probabilistic rules to include new substances (vertices) in the network and also to wire substances by chemical reactions, that is rules to include new hyperedges in the hypergraph. These rules may come from two different sources. They may result from well-established probability distributions such as normal, power law, Poisson and other distributions, or they may be obtained from the historical records of the unfolding of the chemical space. The first approach entails developing a theory for random directed hypergraphs, which is a vibrant field of research for chemists, physicists and mathematicians. The second approach is an empirical one designed for the specific purpose of modelling the network of chemical reactions.

Models based on normal distributions may shed light, for instance, on directed hypergraphs where substances do not exhibit any particular preference to be incorporated in new reactions. In contrast, if the probability of belonging to

‖‖‖‖‖‖‖‖‖‖‖‖‖The toll of wars is also evident in the current He shortage exacerbated by the war in Ukraine.[102]

a reaction follows a power-law distribution, then biased directed hypergraphs are obtained, where a substance, or even a set of them, may become a hub for the expansion of the network.******** Although we have actually found that the above mentioned preferences for some few toolkit compounds are in line with the presence of these hubs in the network,[10] our statistical tests reject the hypothesis of a power-law distribution. This shows that the dynamics of the underlying network of the chemical space does not follow a Barabási–Albert model or any other generative process of power-law behaviours, as suggested in ref. 18. Thus, the question on the suitable model for the dynamic behaviour of the chemical space remains open. In general, dynamical hypergraph models are important as they set a background to compare the actual network of chemical reactions with its theoretical extremes.

The contrasting of the experimental chemical reaction network with the several theoretical models allows for determining whether there have been periods in history where the chemical network has been closer to normal random distributions or much depending on the role of some few substances (power-law distributions). Several of the results discussed in this perspective seem to indicate that the early days of the 19th century depicted a network with no chief role of few substances, which would indicate an exploratory period of expansion of the chemical space. However, the rise of organic chemistry seems to have changed the structure of the network by highlighting the importance of some few substances and even of some few kinds of reactions.

A further approach to endow new reactions and substances with probabilities entail extracting those probabilities from the historical records of new substances and new reactions. Hence, if the participation of substances in chemical reactions follows a heavy-tailed distribution, as actually observed between 1800 and 2015,[10] then probabilities of participation of substances in reactions may be assigned based on this sort of distribution. This, combined with the current values of variability of the annual output of chemicals may be used to estimate the future expansion of the chemical space.

Regarding reactions, the often reliance on few classes of chemical reactions[81] may be used to assign reaction probabilities among sets of chemicals. As every reaction class is characterised by a reaction centre, these centres may be weighed by the frequency of use of the reaction class. The probability of application of a chemical reaction class over a given set of substances is then given by the frequency of use of the reaction centre. A reaction centre is made by the atoms and bonds undergoing changes in a chemical reaction.[81,121] This model would shed light on the future shape of the chemical space under the current preferences for amide formation reactions, alkylation of alcohols or phenols and the other classes of reactions constituting the current preferred toolkit of chemical transformations.

6 Conclusions

I present upper bounds for different facets of the chemical space, which is constituted by substances endowed with a notion of nearness. The bounds range from the possible number of substances and reactions to its similarity classes and topologies. Based on the number of atoms in the universe, 10^{76}, the upper bound for the number of substances is $\mathscr{C} = \sum_{k=1}^{10^{76}} \binom{k+10^{76}-1}{10^{76}}$, which by chemical constraints can be reduced to about $\mathscr{W} = 10^{200}$ substances. If n is either $\mathscr{C}$ or $\mathscr{W}$, the upper bound for chemical reactions is $3^n - 2^{n+1} + 1$. The possible number of similarity classes on the chemical space is $\frac{1}{2}\sum_{k=0}^{n}(-1)^k\binom{n}{k}2^{2^{n-k}}$, which is related to the $n!/(2(\log 2)^{n+1})$ possible topologies. As these upper bounds imply storing by far more than 10^{123} bits, which is the information our physical universe can house, the ultimate record of substances, reactions, similarity classes and topologies associated to the complete chemical space is simply impossible.

Beyond analysing upper bounds of the chemical space, I discussed recent data-driven results on the evolution of the chemical space based on the historical records of thousands of chemistry publications from 1800 up to date, where substances, reactions and properties have been reported. The historical trends of expansion of the chemical space, doubling the number of substances each 16 years, show that if we keep expanding the space as usual, the target of producing at least $\mathscr{W}$ substances would be only attained in 10 300 further years. By setting 2050 as an arbitrary deadline for obtaining $\mathscr{W}$ substances, we would need to double the number of known substances every 5.46 years.

Before analysing the implications of the above quantitative results, I discuss some features of the relationships that endow substances with a notion of nearness and which turn the set of chemicals into an actual space. I analysed two main relations among substances, namely chemical similarity and synthetic separation.

Chemical similarity is currently used in QSAR studies and is based on the molecular structural resemblance among substances. But the chemical space is not only populated by molecular substances, for which QSAR approaches have been developed. Chemical space also includes inorganic substances, glasses and alloys, composites and nanostructures. Methods relating substance "structure" with their properties, to which QSAR approaches belong, need to be further developed to treat non-molecular substances. Interesting advances in this direction are discussed in ref. 13 and current innovative methods combining experimental data with Machine Learning approaches are being reported.[14–16] These approaches allow for endowing largest regions of the chemical space with similarity classes to estimate new compounds. They also allow for actually extending our understanding of the chemical space beyond its

********An instance of a power-law distribution is Pareto's law or 80–20 rule, which for the case of the distribution of wealth, as originally studied by Pareto, states that about 80% of wealth is held by 20% of the population.[120] In the case of chemical reactions a Pareto-law-like distribution would mean that about 20% of the substances would be involved as educts of about 80% of the reactions.

molecular limits. Further challenges for the chemical space defined through its similarity classes involve devising substance descriptors able to encode the salient features of the chemicals studied, which are not always related to their composition and 3D arrangement, but also with the processes involved to end up with those chemicals. This entails incorporating in the description of the substance also information about its environment and its production, which may include behaviours under different conditions as well as reaction conditions. Hence, it is important to widen the concept of substance structure beyond its traditional niche of "balls and sticks". These are challenges that are to be tackled by the wise interplay of large volumes of very accurate experimental information with novel computational methods such as those provided by machine learning as well as with modern mathematical settings.

The second approach to "glue" chemicals in order to build up a chemical space is through the reactions connecting substances, which leads to the network of chemical reactions. I discussed two models for this network, one based on graphs and the other, more general, on hypergraphs. As well as the concept of substance "structure" needs to be widened, the concept of chemical reaction network structure needs to be widened. These networks are not only an extension of the "balls and sticks" model for molecules where atoms are linked. A chemical network is much more than replacing atoms for substances and chemical bonds for substance co-occurrence in chemical reactions. Chemical reactions encode the relationship between two sets of substances: educts and products. I discussed how directed hypergraphs constitute a suitable model encoding these directed relationships among sets. Notwithstanding the importance of hypergraph models for chemical reactions, there are still challenges for this model. They include developing directed hypergraph statistics able to capture local and global properties of the network. As chemists have experience solving similar challenges, for instance when developing thousands of global and local descriptors for molecular structures, what are the descriptors for directed hypergraphs conveying information about the structure of the chemical space? Which are the Wiener and topological indices of the chemical network? Network descriptors need also to involve the temporality of the network, which connects the field with the mathematical subject of dynamical systems and of complexity studies. Expected advances in the field of chemical directed hypergraphs include the development of approaches to treat the interaction of the directed hypergraph with other structures such as spaces of reaction conditions or of substance properties, whose mathematical settings seem to be continuous rather than discrete. This also becomes a fruitful field of research for novel topological approaches to chemistry.

I emphasised that topological treatments of the chemical space may shed light on interesting mathematical properties such as continuity and connectivity, which turn crucial to better understand the similarity among chemicals. They may become central for understanding, for instance, similarity cliffs and, in general, may shed light on emerging corners of the chemical space as well as on the possibility of finding similar substances to a target one, which is pivotal for material sciences or medicinal chemistry. In turn, if topologies are obtained from the chemical network, they may become powerful tools to navigate the chemical space, which, if coupled to AI approaches, may contribute to further improve algorithms for the search of optimal synthesis plans.

I mentioned before that the physical universe cannot store the information of the chemical space. If we cannot store the complete story of material achievements, which is in the end what the chemical space is, what is the fate of chemistry? Is chemistry ultimately stamp collection, as it is said Rutherford once claimed? I do not believe so. The challenge posed by the vast chemical space is not to synthesise every possible single substance. It is not a combinatorial realisation of the space by extension. This could be, in the end, a routine task for a "Penelope-robot" knitting atoms into new substances and ripping them apart to begin new synthesis until reaching $\mathscr{C}$ substances. The challenge is not a combinatorial realisation of the space by extension, which not even the robot can remember. The challenge posed by the chemical space is gauging it by intension. The charm of chemistry lies in finding a minimal set of features characterising the extension of the space and its diversity. I dream of defining the chemical space as the intertwine of substances and reactions in a chemical network meeting the conditions $p_1(t)$, $p_2(t)$, …, $p_n(t)$, where, hopefully, n is not that large. With t in $p_i(t)$ I stress the temporal facet of these conditions, as what is important today may be irrelevant in the future. Therefore, all efforts to sharpen our theoretical and experimental tools to detect those relevant features constitute the real and reachable challenge posed by the chemical space. Every method aiming at finding relevant similarity classes, coverings and topologies on the space, as well as better models for the evolving chemical network, is paramount for the future of chemistry. In a more formal claim, there is an absolute chemical space $\mathbb{C}$, which I have extensively discussed in this document, but there is also the "interesting chemical space" at time t ($\mathbb{I}(t)$). The charm of chemistry is finding the function $f(t)$ mapping $\mathbb{C}$ to (t). That is $\mathbb{C} \xrightarrow{f(t)} \mathbb{I}(t)$. By definition $\mathbb{I}(t) \subseteq \mathbb{C}$, which I hope to be $\mathbb{I}(t) \subset \mathbb{C}$. Possible constituents of $f(t)$ include that $\mathbb{I}(t)$ can be afforded with the technology and the knowledge at t, as well as that $f(t)$ improves known preparative methods of chemistry, that is by detecting representative substances and reactions triggering new classification in the chemical space. $f(t)$ must also involve criteria to sharpen, challenge or debunk chemical theories, to find new applications of the substances in $\mathbb{C}$ and to pick up substances in $\mathbb{C}$ with novel structures, either at the bounding or the microscopic level.

Besides wondering about the limits of the chemical space and the challenges they pose, I put forward that an additional level of understanding of the space is attained by looking back at its evolution, which allows for exploring the historical interplay of substances and reactions and the possible influence of past events upon the unfolding of the space. By analysing several studies on the historical unfolding of the chemical space I showed that both substances and reactions have been discovered at an exponential rate.

There is no account of the growth rate of the number of similarity classes discovered, as it requires setting up a universal similarity criterion, which is difficult to attain, as similarity is often tailored to particular realms of the chemical space or to particular properties of the substances. Nevertheless, there is a report on the similarity of chemical reactions, which are classified by their reaction centres.[81] The report shows that classes of chemical reactions, unlike single reactions, do not grow exponentially, but at most linearly. This indicates that chemists often use a small set of chemical transformations to expand the chemical space, which when combined with the participation of substances in reactions, shows that there is also an often used subset of substances acting as educts in many chemical reactions. Moreover, these results on preferences chemists have regarding kinds of transformations and regarding starting materials, coupled with the exponential growth of substances and of reactions, indicate that most of the substances once discovered are very seldom used in further reactions.

Chemistry is today definitely driven by the production of new chemicals, as shown by the historical density values of the chemical network. Since the turn of the 20th century, density has dropped, indicating an emphasis on substances rather than on wiring them through chemical reactions. The report of new chemicals, which could, in principle, come from a balance of extractions from natural products and synthesis, is definitely tilted towards synthesis. Chemical synthesis is the driving force expanding the chemical space since the early years of the 19th century and not, as traditionally claimed, a trend triggered by Wöhler' synthesis of urea in 1828. Chemical space is therefore a human construction, resulting from the intervention of chemists on matter. This intervention is not only guided by the thermodynamics, kinetics and quantum chemistry of chemical species but by the social, semiotic, technological, economic and historical conditions allowing for such an intervention. Several of these conditions have been studied when analysing the progress of science and it has been discussed how they may level off the growth of science.[99,122,123] Nevertheless, a more nuanced analysis for the particular case of chemistry, especially including recent chemical, social and economic data, is still to be done. For instance, in the 1960s Solla Price claimed that science was reaching its saturation point in terms of growth. This statement has been disproved by different accounts.[2,78] Holton, still in the 1960s, proposed models for scientific growth and its dependence on innovations and man-power.[122] Later on, Rescher took the subject over and further explored it from an economic perspective.[123] Schummer, in the 1990s, was the first to analyse the case of the growth of chemistry in terms of the chemical space[78] and he found no sign of saturation in the production of substances. This condition has also been confirmed by my research group.[10] Nevertheless, further studies on the conditions for levelling off the growth of the chemical space are crucial to try to avoid those conditions, if possible, and to better understand the complex dynamics expanding the chemical space.

By analysing the evolution of the annual output of new chemicals over the history, going beyond the exponential trend, and focusing on the variability of this production, the chemical space exhibits a general regularisation trend. That is, with the passage of time, chemists produce, every time, a number of substances closer to the number of substances of recent years. This regularisation has not, nevertheless, occurred in a continuous way. It has had two major transitions. The first one was marked by the incorporation of the molecular structural theory, which became a guide to explore new corners of the space. The second transition occurred around 1980 and it is a still and open question to determine its causes.

Although the aim of chemistry cannot be reaching completeness of the chemical space, more knowledge on it, that is of its substances, reactions, similarity classes, topologies and the structure of the reaction network, would be important additions to advance in the detection of the relevant aspects of the space. I already mentioned the 10 300 further years we would require to attain $\mathscr{W}$ substances if we keep expanding the chemical space as we have done in the last 200 years. This means that by expanding the chemical space as usual, by the year 12 322 we would have reached 10^{200} chemicals. As this is not an affordable time for our current species and for its interaction with our planet, I calculated the required speed to attain $\mathscr{W}$ chemicals by a closer year, actually 2050. As discussed above, this would require doubling the number of discovered chemicals each 5.46 years, which implies leaving behind more than 200 years of doubling the number of discovered substances each 16 years. Can we afford these unprecedented speeds? Can we find the suitable technologies, theories and social structures of chemistry to speed up the discovery of the chemical space? I also discussed the possibility of dedicating some synthetic efforts to create databases of failed reactions, which may feed AI algorithms to take the chemical space to new arenas. It was found that, in order to create a training set involving 25% of the current chemicals, we would need to dedicate all our synthetic activities of the coming five years to complete the task. Are we willing to reduce our syntheses "as usual" for the sake of a rapid expansion of chemical diversity?

I posit that a deep understanding of the dynamics of the expanding reaction network may shed light on the inner workings of the chemical space, which we may eventually tune to speed up the discovery process. Therefore, modelling the network is of central importance for chemistry and I highlighted some advances and open questions in this direction. I stress that such modelling cannot be restricted to the simple weaving of substances in chemical reactions, it must incorporate social and semiotic aspects of the chemical practice, which when considered together may lead to simple rules required to speed up chemical discovery. I believe the ongoing digitisation of relevant social and semiotic sources along the history of chemistry, coupled to the well-established electronic annotation of substances and reactions, make this moment ripe for devising models for the evolution of the chemical space.

Chemical space is a thriving interdisciplinary scientific research area assisted by mathematical and computational methods, which relies on the exponential amount of information left behind by chemists of all times. Such a well-structured corpus of chemical information, which exceeds the possibilities

of our universal library, can only motivate data-driven approaches to gauge the essence and interesting features of the chemical space. The moment is ripe for these digital discoveries.

Data availability statement

As this is a Perspective article, no primary research results, data, software or code have been included.

Author contributions

The author wrote the document.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

I am grateful to Jürgen Jost, Peter F. Stadler and Joachim Schummer for interesting discussions over the years about the chemical space. I owe a debt of gratitude to Wilmer Leal for dumping and processing Reaxys data leading to our publications on the chemical space and for interesting discussions on the chemical space. I thank Duc H. Luu for his accurate mathematical analyses of the time series signals of our research and Eugenio J. Llanos for analysing part of the Reaxys data used in our work. I am indebted to Rainer Brüggemann for his critical comments on early versions of this document and for suggesting additions to it. I also thank Heber Mesa and Angel Garcia-Chung for his mathematical comments, as well as Jeff Seeman for his expert opinion on some of our results. Our findings on the chemical space could not be possible without the access to Reaxys database. Therefore, I am grateful to Elsevier for providing it.

Notes and references

1 J. L. Borges, *The Library of Babel*, ed. David R. Godine, Boston, 2000.
2 G. Restrepo and J. Jost, *The Evolution of Chemical Knowledge: A Formal Setting for its Analysis*, Springer, 2022.
3 T. Fink, H. Bruggesser and J.-L. Reymond, *Angew. Chem., Int. Ed.*, 2005, **44**, 1504–1508.
4 K. L. M. Drew, H. Baiman, P. Khwaounjoo, B. Yu and J. Reynisson, *J. Pharm. Pharmacol.*, 2011, **64**, 490–495.
5 M. P. Krein and N. Sukumar, *J. Phys. Chem. A*, 2011, **115**, 12905–12918.
6 Ngrams, **https://www.books.google.com/ngrams**, accessed 2022-02-03.
7 V. De Risi, *Mathematizing Space: The Objects of Geometry from Antiquity to the Early Modern Age*, Springer International Publishing, 2015.
8 J. Jost, *Mathematical Concepts*, Springer, Cham, 2015, p. 312.
9 G. Restrepo, *Nachr. Chem.*, 2020, **68**, 12–15.
10 E. J. Llanos, W. Leal, D. H. Luu, J. Jost, P. F. Stadler and G. Restrepo, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 12660–12665.
11 *Concepts and Applications of Molecular Similarity*, ed. M. A. Johnson and G. M. Maggiora, Wiley, 1990.
12 *Advances in QSAR Modeling*, ed. K. Roy, Springer, 2017.
13 K. Wu, B. Natarajan, L. Morkowchuk, M. Krein and C. M. Breneman, *Informatics for Materials Science and Engineering*, Butterworth-Heinemann, Oxford, 2013, pp. 385–422.
14 J. Kerner, A. Dogan and H. von Recum, *Acta Biomater.*, 2021, **130**, 54–65.
15 G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *Journal of Physics: Materials*, 2019, **2**, 032001.
16 A. Pedone and M. C. Menziani, in *Computational Modeling of Silicate Glasses: A Quantitative Structure–Property Relationship Perspective*, ed. C. Massobrio, J. Du, M. Bernasconi and P. S. Salmon, Springer International Publishing, Cham, 2015, pp. 113–135.
17 G. Grethe, G. Blanke, H. Kraut and J. M. Goodman, *J. Cheminf.*, 2018, **10**, 22.
18 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2005, **44**, 7263–7269.
19 K. Molga, S. Szymkuć and B. A. Grzybowski, *Acc. Chem. Res.*, 2021, **54**, 1094–1106.
20 B. A. Grzybowski, *Synthesis Planning, Reaction Discovery, and Design of Chemical Systems Using Computers*, 2021, pp. 15–20.
21 B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich and B. A. Grzybowski, *Nature*, 2020, **588**, 83–88.
22 M. Newman, A. Barabási and D. J. Watts, *The Structure and Dynamics of Networks*, Princeton University Press, 2006.
23 P. Pyykkö, *Phys. Chem. Chem. Phys.*, 2011, **13**, 161–168.
24 C. Cao, R. E. Vernon, W. H. E. Schwarz and J. Li, *Front. Chem.*, 2021, **8**, 813.
25 G. Restrepo, *Chem.–Eur. J.*, 2019, **25**, 15430–15440.
26 P. Schwerdtfeger, O. R. Smits and P. Pyykkö, *Nat. Rev. Chem.*, 2020, **4**, 359–380.
27 W. Leal, E. J. Llanos, A. Bernal, P. F. Stadler, J. Jost and G. Restrepo, *ChemRxiv*, 2021, DOI: **10.26434/chemrxiv-2021-2frpz**.
28 E. Whittaker, *Math. Gaz.*, 1945, **29**, 137–144.
29 K. M. Guggenheimer, *Nature*, 1962, **193**, 664–665.
30 M. M. Vopson, *AIP Adv.*, 2021, **11**, 105317.
31 N. Greenwood and A. Earnshaw, *Chemistry of the Elements*, Elsevier Science, 2012.
32 A. Benjamin and J. Quinn, *Proofs that Really Count: The Art of Combinatorial Proof*, Mathematical Association of America, 2003.
33 C.-S. Yoo, *Matter Radiat. Extremes*, 2020, **5**, 018202.
34 C. Ornelas, J. Ruiz and D. Astruc, *Organometallics*, 2009, **28**, 2716–2723.

35 A. J. Rocke, *Chemical Atomism in the Nineteenth Century*, Ohio State University Press, 1984.
36 E. Cayley, *Ber. Dtsch. Chem. Ges.*, 1875, **8**, 1056–1059.
37 H. R. Henze and C. M. Blair, *J. Am. Chem. Soc.*, 1931, **53**, 3077–3085.
38 G. Pólya, *Comptes rendus de l'Académie des Sciences*, 1935, **201**, 1167–1169.
39 S. J. Weininger, *Hyle*, 1998, **4**, 3–27.
40 A. M. Duffield, A. V. Robertson, C. Djerassi, B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum and J. Lederberg, *J. Am. Chem. Soc.*, 1969, **91**, 2977–2981.
41 Bibliographie Kombinatorische Abzählungstheorie von Redfield, Pólya, de Bruijn, **http://www.scitron.de/bibliographie_abzaehlung.htm**, accessed: 19-01-2022.
42 R. Gugisch, A. Kerber, A. Kohnert, R. Laue, M. Meringer, C. Rücker and A. Wassermann, in *MOLGEN 5.0, a Molecular Structure Generator*, ed. S. Basak, G. Restrepo and J. Villaveces, Bentham Science Publishers B. V., Netherlands, 2014, vol. 1, pp. 113–138.
43 A. Kerber, *MATCH Communications in Mathematical and in Computer Chemistry*, 2018, **80**, 733–744.
44 *Twenty Five Years of Progress in Cheminformatics*, **https://www.warr.com/25years.html**, accessed: 2022-03-03.
45 A.-D. Gorse, *Curr. Top. Med. Chem.*, 2006, **6**, 3–18.
46 I. Haranas and I. Gkigkityis, *Mod. Phys. Lett. A*, 2013, **28**, 1350077.
47 J. Schummer, *The Impact of Instrumentation on Chemical Species Identity: From Chemical Substances to Molecular Species*, Royal Society of Chemistry, 2002, pp. 188–211.
48 J. Schummer, *Hyle*, 1998, **4**, 129–162.
49 B. Mendelson, *Introduction to Topology*, Dover Publications, 3rd edn, 2012.
50 A. J. Macula, *Math. Mag.*, 1994, **67**, 141–144.
51 J. Baum, *Elements of Point Set Topology*, Dover Publications, 1991.
52 M. Vogt, *Expert Opin. Drug Discovery*, 2018, **13**, 605–615.
53 B. M. R. Stadler and P. F. Stadler, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 577–585.
54 C. Flamm, B. M. R. Stadler and P. F. Stadler, *Generalized Topologies: Hypergraphs, Chemical Reactions, and Biological Evolution*, Bentham-Elsevier, 2015, ch. 2, pp. 300–328.
55 G. Restrepo, H. Mesa, E. J. Llanos and J. L. Villaveces, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 68–75.
56 G. Restrepo and H. Mesa, *Curr. Comput.-Aided Drug Des.*, 2011, **7**, 90–97.
57 M. Grandis, *Elementary Overview Of Mathematical Structures, An: Algebra, Topology And Categories*, World Scientific Publishing Company, 2020.
58 J. P. Barthelemy, *Discrete Math.*, 1980, **29**, 311–313.
59 H. Mesa and G. Restrepo, *MATCH Communications in Mathematical and in Computer Chemistry*, 2008, **60**, 371–384.
60 G. V. Paolini, R. H. B. Shapland, W. P. van Hoorn, J. S. Mason and A. L. Hopkins, *Nat. Biotechnol.*, 2006, **24**, 805–815.
61 J. C. Baez and B. S. Pollard, *Rev. Math. Phys.*, 2017, **29**, 1750028.
62 W. Bort, I. I. Baskin, T. Gimadiev, A. Mukanov, R. Nugmanov, P. Sidorov, G. Marcou, D. Horvath, O. Klimchuk, T. Madzhidov and A. Varnek, *Sci. Rep.*, 2021, **11**, 3178.
63 A. Bernal, E. Llanos, W. Leal and G. Restrepo, in *Similarity in Chemical Reaction Networks: Categories, Concepts and Closure*, Bentham-Elsevier, 2015, ch. 2, pp. 24–54.
64 J. R. Partington, *A History of Chemistry*, Macmillan, 1964.
65 K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2006, **45**, 5348–5354.
66 P.-M. Jacob and A. Lapkin, *React. Chem. Eng.*, 2018, **3**, 102–118.
67 J. Jost and R. Mulas, *Adv. Math.*, 2019, **351**, 870–896.
68 E. Estrada and J. A. Rodríguez-Velázquez, *Phys. A*, 2006, **364**, 581–594.
69 D. Zhou, J. Huang and B. Schölkopf, *NIPS*, 2006, 1–8.
70 S. Brauch, S. S. van Berkel and B. Westermann, *Chem. Soc. Rev.*, 2013, **42**, 4948–4962.
71 A. Dömling and I. Ugi, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 563–564.
72 B. M. R. Stadler and P. F. Stadler, *MATCH Communications in Mathematical and in Computer Chemistry*, 2018, **80**, 639–659.
73 R. Fagerberg, C. Flamm, R. Kianian, D. Merkle and P. F. Stadler, *J. Cheminf.*, 2018, **10**, 19.
74 J. L. Andersen, C. Flamm, D. Merkle and P. F. Stadler, *Chemical Transformation Motifs – Modelling Pathways as Integer Hyperflows*, 2017.
75 H. Fors, L. M. Principe and H. O. Sibum, *Ambix*, 2016, **63**, 85–97.
76 P.-O. Eggen, L. Kvittingen, A. Lykknes and R. Wittje, *Sci. Educ.*, 2012, **21**, 179–189.
77 M. M. A. Hendriksen and R. E. Verwaal, *Ber. Wissenschaftsgesch.*, 2020, **43**, 385–411.
78 J. Schummer, *Scientometrics*, 1997, **39**, 107–123.
79 A. Rocke, *The Theory of Chemical Structure and its Applications*, ed. M. J. Nye, Cambridge University Press, 2002, vol. 5, pp. 255–271.
80 G. Restrepo, *Bull. Hist. Chem.*, 2022, **47**, 91–106.
81 S. Szymkuć, T. Badowski and B. A. Grzybowski, *Angew. Chem., Int. Ed.*, 2021, **60**, 26226–26232.
82 D. G. Brown, M. M. Gagnon and J. Boström, *J. Med. Chem.*, 2015, **58**, 2390–2405.
83 J. G. Topliss, *J. Med. Chem.*, 1972, **15**, 1006–1011.
84 K. C. Nicolaou, *Angew. Chem., Int. Ed.*, 2013, **52**, 131–146.
85 A. Kildebæk Nielsen and S. Strbanova, *Creating Networks in Chemistry*, The Royal Society of Chemistry, 2008.
86 J. Schummer, B. Bensaude-Vincent and B. Van Tiggelen, *The Public Image of Chemistry*, World Scientific, 2007.
87 H. Leicester, *The Historical Background of Chemistry*, Dover Publications, 1971.
88 F. Azevedo and J. T. Jost, *Group Process. Intergr. Relat.*, 2021, **24**, 518–549.
89 G. M. Keserü, T. Soos and C. O. Kappe, *Chem. Soc. Rev.*, 2014, **43**, 5387–5399.
90 E. J. Llanos, W. Leal, G. Restrepo and P. Stadler, *Book of Abstracts, 254th American Chemical Society National Meeting*

& Exposition, Washington, D. C., August 20–24, 2017, American Chemical Society, Washington, DC, 2017, CINF-14.
91 A. H. Lipkus, S. P. Watkins, K. Gengras, M. J. McBride and T. J. Wills, *J. Org. Chem.*, 2019, **84**, 13948–13956.
92 W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke and B. A. Grzybowski, *J. Am. Chem. Soc.*, 2022, **144**, 4819–4827.
93 B. Mahjour, Y. Shen, W. Liu and T. Cernak, *Nature*, 2020, **580**, 71–75.
94 U. Klein, *Experiments, Models, Paper Tools: Cultures of Organic Chemistry in the Nineteenth Century*, Stanford University Press, 2003.
95 W. H. Brock, *The Norton History of Chemistry*, W. W. Norton & Company, 1993.
96 A. Rocke, *Image and Reality: Kekulé, Kopp, and the Scientific Imagination*, University of Chicago Press, 2010.
97 B. Friedrich, D. Hoffmann, J. Renn, F. Schmaltz and M. Wolf, *One Hundred Years of Chemical Warfare: Research, Deployment, Consequences*, Springer, 2017.
98 R. M. Friedman, *The Politics of Excellence*, Times Books, 2001.
99 D. J. de Solla Price, *Little Science, Big Science*, Columbia University Press, 1963.
100 B. Radke, L. Jewell, S. Piketh and J. Namieśnik, *Crit. Rev. Environ. Sci. Technol.*, 2014, **44**, 1525–1576.
101 D. E. C. Corbridge, *Phosphorus: Chemistry, Biochemistry and Technology*, CRC Press, 2013.
102 C. Bettenhausen, *War in Ukraine makes helium shortage more dire*, Chemical & Engineering News, 2022.
103 E. Mack, Forbes, 2017, **https://www.forbes.com/sites/ericmack/2017/11/07/stephen-hawking-apocalypse-2600-fireball-earth-breakthrough-starshot/**.
104 B. Gibson, D. J. Wilson, E. Feil and A. Eyre-Walker, *Proc. R. Soc. B*, 2018, **285**, 20180789.
105 J. Jost, *Dynamical Networks*, Springer London, London, 2007, pp. 35–62.
106 M. Dewar, J. Healy, X. Pérez-Giménez, P. Prałat, J. Proos, B. Reiniger and K. Ternovsky, *Subhypergraphs in Non-Uniform Random Hypergraphs*, 2017, **https://www.arxiv.org/abs/1703.07686**.
107 B. Kamiński, V. Poulin, P. Prałat, P. Szufel and F. Théberge, *PLoS One*, 2019, **14**, 1–15.
108 O. Parczyk and Y. Person, *Electron. Notes Discrete Math.*, 2015, **49**, 611–619.
109 P. Chodrow and A. Mellor, *Applied Network Science*, 2020, **5**, 9.
110 C. Cooper, *Random Struct. Algorithm*, 2004, **25**, 353–375.
111 M. Karoński and T. Łuczak, *J. Comput. Appl. Math.*, 2002, **142**, 125–135.
112 T. Carletti, D. Fanelli and S. Nicoletti, *Journal of Physics: Complexity*, 2020, **1**, 035006.
113 G. Ferraz de Arruda, M. Tizzani and Y. Moreno, *Commun. Phys.*, 2021, **4**, 24.
114 L. Neuhäuser, R. Lambiotte and M. T. Schaub, *Phys. Rev. E*, 2021, **104**, 064305.
115 R. Mulas, C. Kuehn and J. Jost, *Phys. Rev. E*, 2020, **101**, 062313.
116 N. W. Landry and J. G. Restrepo, *Hypergraph Assortativity: A Dynamical Systems Perspective*, 2022.
117 R. Mulas and D. Zhang, *Discrete Math.*, 2021, **344**, 112372.
118 J. Jost, R. Mulas and D. Zhang, *Vietnam J. Math.*, 2022, 323–358.
119 W. Leal, G. Restrepo, P. F. Stadler and J. Jost, *Adv. Complex Syst.*, 2021, **24**, 2150003.
120 A. Barabási and M. PÃ3sfai, *Network Science*, Cambridge University Press, 2016.
121 W. L. Chen, D. Z. Chen and K. T. Taylor, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2013, **3**, 560–593.
122 G. Holton, *Models for Understanding the Growth and Excellence of Scientific Research*, ed. S. R. Graubard and G. Holton, Columbia University Press, 1962, pp. 94–131.
123 N. Rescher, *Scientific Progress*, Basil Blackwell, 1978.