Spaces of mathematical chemistry

Guillermo Restrepo[1,2,3*]

[1*]Max Planck Institute for Mathematics in the Sciences, Inselstr. 22, Leipzig, 04103, Saxony, Germany.
[2]Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstr. 16-18, Leipzig, 04107, Saxony, Germany.
[3]School of Applied Sciences and Engineering, EAFIT University, Carrera 49 N° 7 Sur-50, Medellin, 050022, Antioquia, Colombia.

Corresponding author(s). E-mail(s): restrepo@mis.mpg.de;

Abstract

In an effort to expand the domain of mathematical chemistry and inspire research beyond the realms of graph theory and quantum chemistry, I explore five mathematical chemistry spaces and their interconnectedness through mappings. These spaces are characterised by their elements and the concept of proximity that binds these elements within the space. These spaces comprise the chemical space, which encompasses substances and reactions; the space of reaction conditions, spanning the physical and chemical aspects involved in chemical reactions; the space of reaction grammars, which encapsulates the rules for creating and breaking chemical bonds; the space of substance properties, covering all documented measurements regarding substances; and the space of substance representations, composed of the various ontologies for characterising substances.

Keywords: Chemical space, space of reaction conditions, space of reaction grammars, space of substance properties, space of substance represenations

1 Introduction

Some years ago Willet and I analysed the evolution of the subject content of the two leading journals of mathematical chemistry, namely the *Journal of Mathematical Chemistry* (JMC) and *MATCH Communications in Mathematical and in Computer Chemistry* (MATCH). We found an initial emphasis on chemical graph theory in the former, followed by a shift towards quantum chemistry [1]; a shift that was not observed

in the second journal, which has specialised in graph theory [2]. By analysing the citations of papers published in these journals, we found that besides the high citations between the two journals, most of the citations to JMC come from physical and quantum chemistry journals, while citations for MATCH mainly come from mathematically oriented journals. By assuming that the community of mathematical chemists publish in these two journals, the above results indicate that the community has some connection with mathematics and with chemistry via theoretical chemistry. Nevertheless, as noted by Klein [3], there are multiple examples of mathematical chemistry achievements published in other journals, typically physics or chemistry oriented ones. This poses questions on the role of the JMC and MATCH to nucleate the community.[1]

From a more personal experience arising from attending and organising several meetings on mathematical chemistry over the years, I have the impression that the community is shrinking and that no young scientists enter into the field.[2] This is particularly worrisome, as several years ago I pushed to show that mathematical chemistry actually constitute a scientific discipline [5]. There must be social and epistemic reasons behind the dynamics I observe. Leaving aside the exploration of the social reasons for this trend, which may be well understood in terms of the philosophy of chemistry [6], I concentrate in this document on the epistemic side of mathematical chemistry.

By further exploring the connections of mathematical chemistry with mathematics and theoretical chemistry, Willett and I found that the connection with mathematics is mainly achieved through graph theory, while the connection with theoretical chemistry via the mathematics of quantum chemistry [1, 2]. If mathematical chemistry entails the formalisation of chemical concepts and the study of their mathematical properties [5], perhaps mathematical chemistry, as has been practised, has focused on a narrow but prolific domain that misses potential impact and communication with other branches of chemistry and of mathematics.

I strongly believe that even if graph theory suits very well to treat molecular structures, chemistry is much more than molecular structures, as well as mathematics is much more than graph theory. Likewise, I strongly believe that theoretical chemistry is much more than quantum chemistry. Therefore, other branches of mathematics, not necessarily focusing on quantum chemistry, have plenty of potential in helping to structure non-quantum chemical theories for chemistry. This latter claim, comes from the acceptance of the different ontological levels of chemistry, where bulk substances, molecules, clusters, atomic ensembles, nuclei, electrons and other species of chemical interest are regarded as the objects of chemistry. These species, when coupled with the object allowing for chemical transformation (chemical reactivity) complete the ontological panorama of chemistry.[3] It is by accepting the richness of that ontology that new theories can enter into the realm of theoretical chemistry. Thus, theories lying at the level of substances are valid epistemic objects of chemistry, as well as those based on molecules or on electrostatic interactions of nuclei and electrons. This is particular evident today, when Artificial Intelligence (AI) and Machine Learning

[1] In fact, having core journals, as noted by Nye [4], is central for the emergence of scientific disciplines.

[2] This can be easily assessed by checking, for instance, the names and ages of members of the International Academy of Mathematical Chemistry.

[3] From Aristotle to contemporary chemistry, chemical change has been central to chemical studies [7] and the whole edifice of theoretical chemistry revolves around ways of encoding and explaining chemical change.

(ML) approaches are estimating, with unprecedented precision, new synthetic routes and new materials with tailored properties [8]. As long as these approaches help to discover or build up new concepts in chemistry, advancing the epistemic structure of the discipline, there is plenty of room for theories in chemistry based on different ontological levels.[4] In this respect, Hoffmann and Malrieu are sceptical about the new knowledge brought up by AI and ML methods in chemistry [11–13]. This scepticism poses an interesting challenge for mathematical chemistry, namely the one of using AI and ML methods to discover new concepts in chemistry that can be taught to the new generations of chemists.[5]

It is my opinion that by broadening the scope of mathematical chemistry, that is by welcoming all branches of mathematics and by addressing questions and problems of all branches of chemistry, mathematical chemistry can thrive as a discipline. This perspective, in particular, aims at broadening the scope of action of mathematical chemistry by highlighting alternative fields of research where mathematical chemistry can contribute by producing new mathematics and by bringing back new insights into chemistry.[6] I present this scope of mathematical chemistry through the description and interplay of five mathematical spaces for chemistry.

A graphical depiction of the spaces of chemistry is illustrated in Figure 1. They correspond to mathematical spaces as they are constituted by sets endowed with a notion of nearness [14]. These spaces are: the *chemical space*, corresponding to substances and chemical reactions; the *space of reaction conditions*, spanning all pressures, temperatures, catalysts, solvents, pHs and other parameters driving chemical reactions; the *space of reaction grammars*, corresponding to the encoding of the reaction centres and its matching with educts and products for chemical reactions; the *space of substance properties*, containing chemical, physical, biological and ecological properties of substances; and the *space of substance representations*, which range from molecular graphs to geometrical structures and quantum chemical descriptions.

2 Five spaces of mathematical chemistry

This section presents the five spaces I regard as central for broadening the scope of mathematical chemistry. It also contains information on what it has been achieved in their study and poses new areas of research for each space and for its intertwining with the other spaces.

[4] Some of the recent AI and ML approaches use SMILES (molecular representations) as the input of their algorithms [8], as well as the composition of the substances, devoid of molecular structure [9], for instance. I argue that high-order structures in chemistry, such as the partially ordered sets of chemicals devised by Klein [10] or the chemical reaction networks, may constitute new avenues of research for AI and ML algorithms based on strong and well structured mathematical concepts.

[5] This idea comes from a private discussion with Hoffmann (26th November 2022).

[6] This is another subject of importance to the sustainability and growth of mathematical chemistry, namely caring for the discipline and for its founding pillars, that is chemistry and mathematics. I hardly imagine mathematical chemistry without interacting with mathematics and chemistry. It is therefore important to avoid going into the solely direction of mathematics or the solely direction of chemistry. The former path would lead to mathematical chemistry to be absorbed by mathematics and the second to be absorbed by chemistry. In such a tension between chemistry and mathematics, mathematical chemistry is called to contribute both to mathematics and to chemistry.

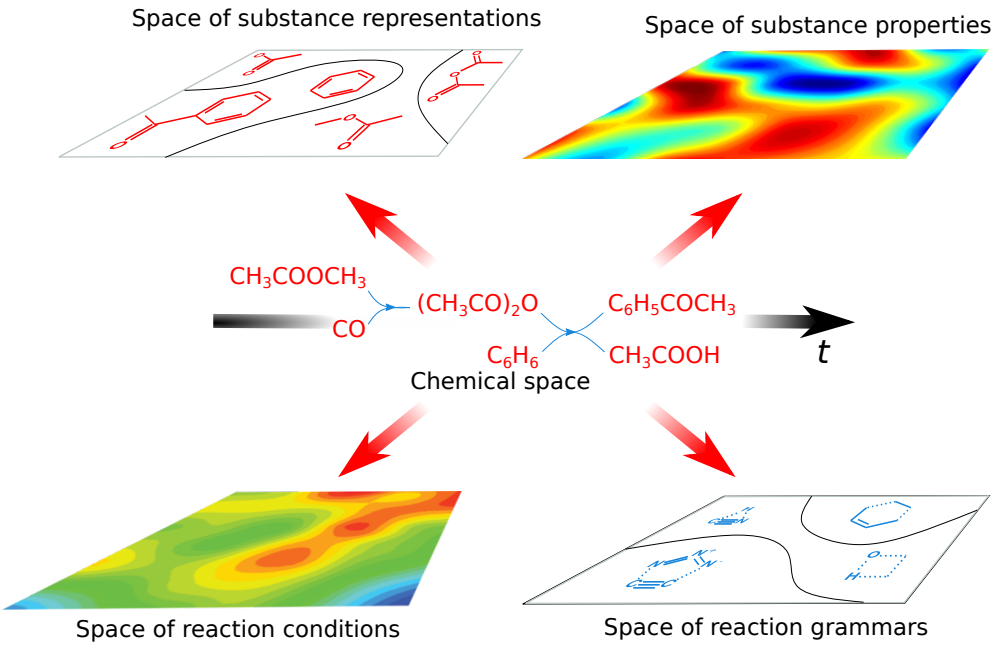

Fig. 1 Five spaces of mathematical chemistry. The central role of the chemical space is highlighted, as well as the temporal dimension of this and the remaining fours spaces. In the space of reaction grammars only reaction centres are depicted for the sake of simplicity.

2.1 Chemical space

Chemistry is all about substances and reactions, therefore the chemical space occupies a central stage in the description of the other spaces (Figure 1). Although there are different definitions of the "chemical space," all of them share the essence of a structure made of chemical species and of a relation among them. Thus, there are for instance definitions of the chemical space based on species that are molecular structures characterised as graphs, whose relationship is given by the nearness of their graphs, which entails all works on molecular similarity [15, 16]. Other definitions of chemical space look for a geometrical similarity based on the three dimensional resemblance of the arrangement of atoms in solids [17, 18]. There are other approaches to the chemical space that characterise molecules in terms of their quantum chemistry descriptions, often electronic densities. In this case the notion of nearness is given by the resemblance of those quantum chemical descriptions [19, 20]. A more recent account of the chemical space, although the most traditional in the history of chemistry, is that of a set of chemicals related by chemical reactions [21, 22].

The question that arises is how much has been accomplished in the characterisation of those different flavours of the chemical spaces. Answering this question triggers questions on the number of substances constituting the chemical space. After all, the chemical spaces just described have the commonality of referring to substances for which it is possible to talk about their nearness.

2.1.1 Chemicals, reactions and a model for the chemical space

In this section I discuss some approaches to determine the number of possible chemicals and of reactions. In the first case the notion of a chemical as the result of atoms holding a relationship is central, which is generalised to the possible number of sets of related atoms. In the second case, determining the theoretical number of reactions give a certain number of atoms is determined by assuming a mathematical model for chemical reaction.

Number of chemicals

As recently discussed [21], the possible number of chemicals in the chemical space depends on the number of atoms in the universe. The estimated number of particles in the universe is about 10^{80}[23–25],[7] which amounts to 7×10^{76} atoms.[8] As discussed in [21], 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 quasi-molecular species is given by $\mathcal{C} = \sum_{k=1}^{10^{76}} \binom{k+10^{76}-1}{k}$, where $\binom{k+10^{76}-1}{k}$ corresponds to the number of ways of selecting k atoms from a collection of 10^{76} atoms, such that order is not important and repetitions are allowed [27]. This means that we are counting mono-, di, tri-, ..., n-atomic quasi-molecular ensembles up to the ultimate largest compound made of all 10^{76} atoms in the universe.[9] As having packages of atoms is not enough to have actual chemical substances, energetic conditions turn central to determine 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 chemicals have been traditionally observed and treated at ambient conditions, there is uncharted land at extreme conditions of pressure and temperature [28]. It is important to note that in chemistry, as well as in mathematical chemistry, we are not interested in the simultaneous "existence" of the possible substances, but rather on their theoretical possibilities.

A further piece of information that needs to be added to the possible quasi-molecular ensembles entails the diverse structural arrangements these ensembles can adopt. To approximate this, we can take a basic approach by multiplying each ensemble by the potential number of graph-theoretical representations. Since graphs are based on binary relationships between objects (in this case, atoms), these structures prove to be a suitable representation for ensembles formed by bonds connecting two atoms. Nonetheless, when we encounter substances like boranes, which do not consistently adhere to the traditional 2-centre 2-electron bonding model, a more inclusive

[7] As reported in [21], early calculations were reported by Eddington in 1931 and account for the number of hydrogen atoms spanning the mass of the observable universe [23]. Eddington's calculation was based on hydrogen, taking into account that about 75% 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 a hydrogen atom (1.67×10^{-27} kg), the figure has been refined to include the number of baryons and electrons in the universe, which amounts to 1.93×10^{80} particles [25].

[8] By considering the abundances of elements in the universe [25, 26] and their atomic weights, the number of atoms per element can be calculated, which leads to the total number of atoms in the universe. See the Supplementary Information in [21] for a complete account of the number of atoms per element in the universe.

[9] As discussed in [21], this upper bound requires further adjustments to touch physical and chemical ground. 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.

framework is required. This broader context is provided by hypergraphs.[10] In the context of hypergraphs, an example of a 3-centre 2-electron bond, such as the one seen in B-H-B, is represented as the hyperedge involving three atoms: {B, H, B}.[11] Moreover, aromatic systems also translate into hyperedges, with equivalent aromatic atoms forming part of these hyperedges [29]. Consequently, a more precise estimation of the number of substances can be achieved by multiplying each quasi-molecular ensemble by a restricted set of potential hypergraphs associated with that specific ensemble [21].

At any rate, a higher order approximation to the upper bound of the number of substances requires chemical and mathematical knowledge, and, importantly, the mathematical knowledge need not be restricted to the realm of graphs. As shown, it requires an extension to high-order structures such as hypergraphs.[12] Therefore, traditional collaborations between chemists and mathematicians, for instance those leading to the MOLGEN package [31, 32], require further extension.[13]

When the above lines are presented to chemists, they often regard such approaches as mere mathematical diversion, far from touching chemistry reality. Moving in this direction, in the 1990s Weininger hypothesised that the number of possible substances under ambient conditions is about 10^{200}, which is known as the "Weininger number" $\mathcal{W}$ [33, 34]. According to Gorse, $\mathcal{W}$ 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" [34]. Although $\mathcal{W} << \mathcal{C}$, $\mathcal{W}$ is anyhow huge.

Given the high number of possible chemicals, all flavours of chemical spaces discussed above and in further sections of this document lead to the conclusion that all chemical spaces have been very little explored [21, 35].

Directed hypergraphs as a model for chemical reactions

Chemical space was initially modelled using directed graphs [36], where, for instance a space made of the reaction A + B → C + D was modelled as the directed graph whose arcs or directed edges corresponded to A → C, A → D, B → C and B → D. The problem of this representation, as discussed in [21, 35], is that it introduces artifacts in the description that hinder the recovery of the original reaction. Thus, the sequence of four arcs just described may lead to the following set of reactions: A → C, A → D, B → C, B → D, A + B → C, A + B → D, A → C + D, B → C + D and A + B → C + D. Clearly, only the last reaction is the one provided by the actual space, but the graph representation provides by far many more options not actually provided by the chemical space.

[10] 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' also corresponds to a graph. In fact, graphs are a particular case of hypergraphs.

[11] For instance, assuming we can distinguish between atoms of the same element, following the typical molecular representations, the hypergraph model for B_2H_6 would look like {{B, B, H, H, H, H, H, H}, {{B, H}, {B, H}, {B, H}, {B, H}, {B, H, B}, {B, H, B}}}. Similarly, the model for H_2O would be {{H, H, O}, {{H, O}, {H, O}}} [21].

[12] Other further high-order structures of potential use in mathematical chemistry include simplicial complexes [30].

[13] Among several other features, MOLGEN provides the number of isomers of a given chemical formula based on a blending of group and graph theories, along with group algebra.

The above shortcoming is solved by using directed hypergraphs that encode the essential feature of chemical reactions, namely that they relate sets of substances rather than individual substances [21]. A chemical reaction is the binary relation between sets of substances, namely the directed binary relationship between the set of educts and of products. The advantage of directed hypergraphs becomes apparent when modelling of A + B → C + D, which is now represented as the *directed hyperedge* {A, B} → {C, D}, leaving no room for alternative interpretations except for the fact that A and B react to yield C and D.[14] The directed hypergraph representation can also be framed in a graph-theoretical version as a bipartite directed hypergraph. In this case, there are two kinds of vertices, namely substances and reactions, which are related via binary relations. For the above example, the directed bipartite graph is given by the following set of arcs: A → r, B → r, r → C and r → D. In this case r corresponds to the vertex representing the chemical reaction.

The hypergraph description of chemical reactions opens a new field of research in mathematical chemistry, as the mathematics of these structures is still to be developed. As discussed in [35, 37], only a few mathematical properties of these structures have been studied, for example vertex and hyperedge degrees [35], clustering coefficients [38, 39], spectral properties [40], curvatures [37] and more recently the Erdős-Rényi model for the random hypergraph [35]. Nevertheless, other aspects, including different random models, measures of assortativity, and betweenness centrality, among others, remain unexplored, as well as further curvatures and network geometry notions as those pioneered by Jost and collaborators [30, 37, 41–44]. On top of this mathematics to develop, the connection with chemistry is central, that is the interpretation and implications of those mathematical properties for the study of the chemical space.

Number of reactions spanning the chemical space

Equipped with the directed hypergraph model, we can now turn to solve the question about the possible number of chemical reactions given a certain number n of chemicals. As shown in [21], this corresponds to $3^n - 2^{n+1} + 1$ reactions (directed hyperedges). This is an important result allowing for the calculation of densities of chemical spaces, as it provides the background to contrast the actual number of reactions realised in a chemical space. Hence, the density of a chemical space housing n substances is given by $d(n) = |R|/(3^n - 2^{n+1} + 1)$, with $|R|$ being the actual number of reactions spanning the given chemical space. The just discussed upper bound for the number of reactions also allows for devising random models for the chemical space, as recently reported when developing the Erdős-Rényi model [35].[15] But, as discussed above, several other mathematical properties of directed hypergraphs are to be explored, and the determination of the upper bound of the number of reactions constitutes a step forward in the mathematical formalisation of the properties of directed hypergraphs.

[14] Formally, a chemical reaction $r : E \to P$, with E being the set of educts and P the set of products, is modelled through the directed hyperedge $r = (E, V)$. In general, given a chemical space with substances gathered in X and reactions collected in R, the chemical space is modelled with the directed hypergraph $H = (X, R)$.

[15] Actually, the model developed in [35] was devised for oriented hypergraphs, which are directed hypergraphs, whose direction has been disregarded. Hence, the Erdős-Rényi model, as detailed in [35], aligns with the chemical space model, in which substances involved in reactions are segregated into products and reactants, but it is impossible to discern between the two sets.

2.1.2 Accomplishments and further open questions in the study of the chemical space

By exploring the records of reports of substances and reactions reported in the scientific literature since 1800 up to date, it has been found that the chemical space grows at an exponential rate [22]. In fact, the number of new substances double every 16 years and there are evidences that the number of reactions follows a similar growth pattern.[16] Such a stable growth pattern has been possible thanks to the reliance of chemists in chemical synthesis, where they often combine no more than three educts in their reactions [22]. Of those educts, very often one is a chemical whose chemistry has been well explored, while the other chemicals are new substances in the chemical space [22]. This approach to expand the chemical space has been coined as the *fixed substrate approach*, where the fixed part is given by the chemical whose chemistry is well understood [22]. Further investigation has shown that the set of fixed chemicals is very small and only a handful of compounds have been used as anchor points to trigger new reactions leading to new chemicals [22]. These often recurring chemicals have been dubbed *toolkit compounds*. Typical examples of toolkit compounds include acetic anhydride and methyl iodide [22].

Although the chemical space was dominated by inorganic compounds at the dawn of the 19th century, after the first quarter of that century organic chemistry began to take the lead in the expansion of the space, to the extent that today most of the chemical space is spanned by substances resulting from the combinatorial capacities of carbon, hydrogen, oxygen and nitrogen [22].

Open questions in the study of the chemical space include determining the underlying reasons of the exponential growth of the space. Is it the result of the growth of the chemical community? Is it related to the mechanisms chemists use to connect chemicals through chemical reactions? Early ideas on the underlying mechanism of growth of the chemical space were reported by Schummer in 1997 [45]. Today the readily available information on the substances and reactions reported over the history of chemistry, as well as the bibliographic information associated to chemical publications, allow for a data-driven analysis of the information, which when combined with computational and mathematical tools may shed light on the mechanisms expanding the chemical space. From a more mathematical perspective, the statistics of the directed hypergraphs are anticipated to provide insight into the geometric and topological characteristics of the chemical space and its evolution throughout the history of chemistry. Are there geometric or topological patterns in the historical expansion of the chemical space allowing for estimating the future of the chemical space?

2.2 Space of reaction conditions

Chemical reactions reported by chemists often occur in controlled experimental settings. These settings refer to the contextual conditions of pressure, temperature, pH, reaction time, nature of the atmosphere in which the reaction mixture is embedded, as well as to the kinds of catalysts and solvents used. These reaction conditions, key for

[16] This arises from results by Grzybowski, who determined the growth rate of reactions and substances for the subspace of organic chemistry [36].

affording reproducibility, are often disregarded when analysing molecular structures, substances properties or the structure of the network of chemical reactions.

As discussed before, the notion of space results central for mathematical chemistry and reaction conditions are not far from a space description. In fact, there is a *space of reaction conditions* made by the set of reaction parameters describing the reaction context, which are endowed with a nearness notion given by similarity among those parameters. Thus, similarity among real valued parameters is defined by the embedding of those parameters in a metric space, which allows, for instance, for measuring similarities among temperatures, pressures, pHs and reaction times. For chemical contextual parameters such as solvents, catalysts and atmospheric embeddings of the reaction, methods of chemical similarity such as those based on descriptor of the molecular graph or on physical properties of the substances can be applied to assess the nearness of these reaction parameters.

Reaction conditions are central for the chemical practice, as they often determine the final fate of a chemical transformation and are determining factors when competing reaction paths exist. It was just recently reported that a classic combination of amines and carboxylic acids not only leads to amides, as traditionally accepted, but to substituted amines, esters, alkanes, ketones and other families of compounds, under particular reaction conditions [46]. Likewise, extreme regimes of pressure and temperature have led to compounds with unconventional stoichiometries, where, for example NaCl is transformed into Na_3Cl, Na_2Cl, Na_3Cl_2, $NaCl_3$, and $NaCl_7$ [47].[17] The systematic exploitation of the reaction conditions is just a nascent field in chemistry, as most of the reactions in the history of the discipline have been performed under a narrow window of reaction conditions revolving around ambient conditions [48, 49].

Now that the space of reaction conditions is aimed to be expanded and further explored, it is central to understand how chemists have explored it over history and how large and diverse it is. This space becomes an important subject of research for mathematical chemistry as it offers opportunities for its mathematical characterisation. Likewise, it poses new questions. For instance, to what extent the diversity of the space of reaction conditions is related to the diversity of the chemical space? Could we estimate the effect of chemistry at extreme conditions upon the space of substance properties?[18]

An interesting mapping involving the space of reaction conditions is the one to the chemical space. In this setting, every chemical reaction of the chemical space, modelled as a directed hyperedge, is annotated by its associated reaction conditions. Hence, the directed hyperedge is associated to a tuple that gathers the information on solvents, catalysts, temperatures, reaction times and other parameters associated to the modelled chemical transformation. This reaction-information link is today used to optimise synthesis plans that meet green chemistry conditions or that facilitate one-pot reactions [50]. Several examples of the coupling of this mapping with ML and AI algorithms have been recently reported in the literature [8] and offer more opportunities for mathematical chemistry research.

[17] Extreme conditions chemistry is an emerging field of research supported by the recent interests in mining other planets and asteroids of our galaxy.

[18] This space is discussed in a latter section.

2.3 Space of reaction grammars

The expanding chemical space has brought up the need of systematising chemical knowledge at different levels, for instance by classifying chemical reactions [29, 51], or by producing abstract structures such as periodic systems encoding the salient features of chemical knowledge [52]. In terms of the network of chemical reactions spanning the chemical space, classifying reactions entails reducing the size and complexity of the network to an associated network of classes of substances [53]. Given the central role of reaction classification for chemical knowledge [29, 51], several approaches have been devised to meet this end, which range from the so-called "name reactions" such as Claisen condensation or Grignard reaction to more sophisticated settings involving physicochemical properties of educts and products or reaction grammars.

Name reactions rely on a manual indexing method based on the name of the chemist introducing the reaction, or on the identity of a reaction product, or on a relevant functional group involved in the transformation, or on the reagent used. At any rate, name reactions are far from being systematic [54], which poses not only a nomenclature problem for chemistry, but makes it difficult to process the deluge of new reactions appearing every day. Physicochemical classifications are based on the resemblance of physicochemical properties of educts and products [54]. The shortcoming of this approach is its lack of a set of physicochemical properties acting as the basis for the classification, which leads to subjectivities. The approach using reaction grammars regards a chemical reaction at its atomic level, by considering educts and products as molecular structures related through a rewrite rule acting upon the structures of educts and leading to the structure of products. It is considered a grammar as it contains the information, the rule, of which bonds to break and which ones to form. Stadler and his team have pioneered this approach [55, 56] by computationally encoding the grammars of certain reactions and by letting them to iteratively act upon an initial set of molecular graphs.

Figure 2 explains the concept of a reaction grammar. Given a chemical reaction, as the Diels-Alder reaction depicted in Figure 2, an atom mapping algorithm detects which atoms of the educts correspond to those of the products (coloured regions in Figure 2). This allows for detecting the context of the reactions, that is those atoms and bonds actually taking part in the reaction (Figure 2). In the example, as all bonds around the six atoms in brackets change (four of the diene and two of the dienophile), the context is given only by these carbon atoms. The context of any reaction encodes the essential information of the transformation and therefore can be used to classify reactions in such a manner that all those reactions with common context belong in the same reaction class.[19] The context is also called the reaction centre.

In this setting, every reaction class, such as the Diels-Alder reaction is expressed as a triple made of the left hand side graph (L), the context (K) and the right-hand side graph (R) (Figure 2). The former corresponds to the atoms (vertices) and bonds (edges) of the educts affected by the reaction. The context gathers together the atoms and bonds making it part of the reaction centre. In turn, the right-hand side corresponds to the atoms and bonds affected during the rewriting process. Based

[19] This approach was recently followed by Grzybowski and his team, when analysing the number of reactions classes of organic chemistry [57].

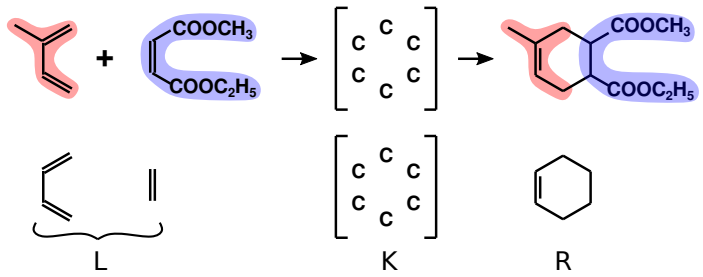

Fig. 2 Chemical reaction grammar. Top: Diels-Alder chemical reaction with educts on the left and product on the right. Coloured regions indicate the atom mapping and the structure in brackets corresponds to the context of the reactions, that is the atoms actually involved in the reaction. Bottom: Grammar rule for the Diels-Alder reaction. L constitutes the graph of educts, K the context and R the graph of products.

on this encoding of a reaction class, whenever a set of educts (molecular graphs) matches with the left hand side graph, the reaction context can be applied and the product is estimated by the rewriting rule encoded by the reaction grammar. This approach has been useful to estimate potential chemical spaces that could be obtained by the repetitive use of particular reaction classes (reaction grammars). Research in this direction is central for understanding questions, for instance, on the origin of life, where a few reaction grammars act upon a limited set of substances.

Equipped with the notion of a reaction grammar, then a *space of reaction grammars* can be defined as the collection of reaction grammars, whose nearness is given by their similarity. Such a similarity may be assessed in different forms, for instance by looking for the Maximal Common Subgraph (MCS) of the context of two grammars, or by including the MCS of the left-hand side of the two contrasted grammars, as well as the MCS of the right-hand side graphs.

Key for characterising reaction grammars is finding the atom-to-atom mapping of educts into products. Different mathematical and computational approaches to address the detection of atom mappings have been reported over the years and today the most versatile ones include the use of Large Language Models (LLM) [58]. In a recent study by Grzybowski it was found that despite the exponential growth of the chemical space, chemists discover new reaction centres, that is reaction classes, at a linear or sublinear pace [57]. This implies that despite the rapid growth of chemistry, materialised in its new substances and reactions, the innovative creation of new reaction classes is not that rapid. From a philosophical perspective, chemistry seems to be producing, in a rapid manner, new chemicals and reactions but through the same preparative methods, which makes it ponder on the actual growth of chemical knowledge [59]. According to Grzybowski's results, the most used reactions by chemists to expand the chemical space are:

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.

Some open questions regarding the space of reaction grammars involve, for instance determining the most meaningful context to characterise reaction classes. As noted by Grzyboswki and collaborators [57], the context of a reaction does not always distinguish different classes of reactions, and they exemplified it with the case of N-acylations and of substitutions of vinyl chlorides via addition-eliminations. In both cases, the context involves the fragments C-Cl and NH_2, but clearly both reactions cannot be classified as the same reaction. A further point, related to the space of reaction conditions, is that the fact of finding a match between the graph of educts and the left-hand side of a grammar does not necessarily lead to a particular set of reaction products. As seeing in the case of the mixture of amines and carboxylic acids [46], the selection of the grammar finally leading to amides, or substituted amines, or esters, or alkanes, or ketones, or other families of compounds is strongly attached to the reaction conditions. It is therefore important to explore the mapping or reaction grammars and reaction conditions. That is, determining the region in the space of reaction conditions associated to each grammar. Are regions in the space of reaction conditions only associated to a grammar? If not, which further criteria are needed to select the acting grammar? From a more mathematical perspective, further formalisation of the reaction grammars is needed, as noted by Stadler and his team, when approaching composition rules of grammars via category theory [60].

2.4 Space of substance properties

This space is made of reported properties such as melting and boiling points, densities and biological activities, among several others [51], whose nearness is quantified by the distance between actual property values. These distances result from embedding the properties in a metric space.

An early attempt to analyse the behaviour of a single property was carried out by Grzybowski and his team [36], who studied the molecular weight of all organic substances reported between 1850 and 2004. It was found that chemists use raw materials of about 150 g/mol to reach products weighing, on average, 100 g/mol more. Other studies have focused on the 21st century behaviour of the subspace of pharmacological substances, with emphasis on their molecular weight and logarithm of the octanol/water partition coefficient [61]. These properties are suitable proxies for the oral-drug like activity of those compounds [62].

Despite these tailored studies of the space of properties, there is no systematic study of the substance properties reported by chemists. Therefore, analysing the space of substance properties constitutes an interesting field of research for mathematical chemistry. And the moment is ripe to begin these studies as the large volume of chemical records today stored in electronic databases offers ample possibilities to analyse

the size, diversity and temporal expansion of the properties reported by chemists bout the substances they have procured either by extraction or by chemical synthesis.

Interesting questions for this space include, as for the other spaces, its mathematical characterisation and its mappings to other spaces. Can we unveil a function depending on reaction grammars and reaction conditions that leads to substance properties? This would entail that substance properties are the result of a process and not necessarily encoded in the substances. In the end, it would incorporate the idea that substance properties are somehow encoded in the related substances, that is on the educts leading to the substance of interest, as well as on its products.

2.5 Space of substance representations

There are different ontological levels for chemistry, which range from bulk substances to quasi-molecular species [63] and chemical annotation, over the history, has spanned all these levels.[20] Hence, before the acceptance of molecular structural theory, substances were labelled by their composition, then by the binary notation following the dualistic theory by Berzelius [59]. Today the ontology of chemistry revolves around molecular structures represented as labelled graphs and encoded via SMILES or InChI structures. When needed, this encodings are endowed with geometrical information, where the connectivity relationships among atoms belonging to the molecular graph are enriched by interatomic distances. This is the case of crystal structures or of substances whose solid state is of relevance. Likewise, the chemical ontology may be enriched by introducing information on the nanoscale arrangements of quasi-molecular species, which are of relevance for new materials chemistry and nanotechnology, for instance.

Therefore, the space of substance representations involves selecting the ontology in place for chemistry and the assessment of the similarity among those substance representations. For an ontology of compositions the similarity can be assessed in terms of lexicographic resemblance, but it can also be addressed by considering the role of the substances in the chemical space, as reported in [65]. For graph theoretical representations, the usual similarity is based on the determination of the MCS of the graphs compared. As in the case of the composition, similarity among molecular graphs can also be assessed by considering the role of the analysed graphs in the chemical space. Initial results in this direction use LLM to analyse the similarity among chemicals [8].

Early results analysing the space of molecular structures only consider the period 2009-2018 and are restricted to the organic chemistry part of the whole chemical space [66, 67]. It is found that chemists have actually reported very few *molecular frameworks*, that is general molecular templates depicting atom connectivity and disregarding bond order [66, 67] (Figure 3). Interestingly, a large fraction of the chemical space is spanned by a reduced set of these frameworks. In Figure 3 the three most populated frameworks are depicted. The finding that chemists do not populate the chemical space with a large diversity of molecular frameworks is presumably related to the conservatism in the selection of starting materials to undergo chemical synthesis

[20] Examples of quasi-molecular species include molecular clusters, van der Waals complexes and other chemical species [64].

[22, 68], a conservatism that has been also found for the subspace of pharmaceutical chemicals [69].

Despite these results, a wider analysis of the space of molecular structures is still to be done, where not only organic but inorganic, organometallic and biochemical substances are to be included. This poses serious challenges and opportunities for mathematical chemistry. For instance, the SMILES and InChI annotation of inorganic substances is still an open problem, as well as the annotation of large structures such as proteins and new materials. There is another aspect of the space of substance representations which has not been considered yet, namely side-groups. These are molecular fragments that remain after extracting molecular frameworks. In Figure 3 they are highlighted in blue. Although the molecular backbone, gauged through the molecular frameworks, provides information about the building blocks of the chemical space, all those side-groups making possible the production of new chemicals are an important part of the exploration of the chemical space.[21]

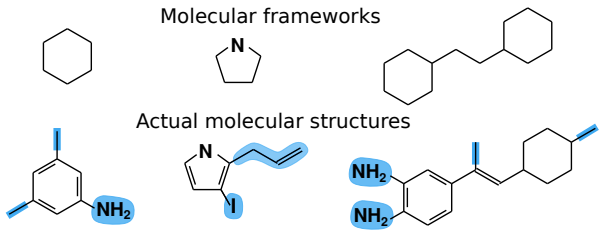

Fig. 3 Molecular structural patterns. Most produced molecular frameworks and some actual molecular structures containing them. Blue fragments correspond to side-groups.

Questions related to the space of substance representations arise from the mapping between that space and the space of reaction grammars. For instance, is it possible to find reaction grammars departing from the graph representation of substances? If, as noted in a previous section, hypergraphs are also suited to represent non 2-centre 2-electron bonds, can we generalise reaction grammars to operate upon those high-order structures? This would lead to study hypergraph isomorphism, which is a vibrant field of mathematical research [70]. A further mapping relates the space of substance representations with the space of substance properties. This mapping has been traditionally studied in QSAR studies. The challenge for mathematical chemistry is to generalise the accumulated QSAR knowledge, based on graph representations, to hypergraph representations. A further avenue of research entails representing substances, as discussed in a previous section, not by their internal structures but rather by their connectivity in the chemical space. In this setting, the identity of a substance would not be encoded in its molecular structure but rather in the number and diversity of the compounds it interacts with and produces. This diversity could be evaluated in relation to the constitution of these substances, as well as their properties.

[21] This is an aspect suggested by Stadler in a private communication (September 2021).

3 Conclusion and outlook

I have provided the description of five spaces that I consider relevant to trigger further research in mathematical chemistry. They are the chemical space, the space of reaction conditions, the space of reaction grammars, the space of substance properties and the space of substance representations. Besides describing them in terms of the elements that constitute them and the relationship that holds together those elements, I provided some examples of important mappings between couples of those spaces, which either have become subjects of research in mathematical chemistry or that constitute potential avenues of research. The presentation of these spaces has been nuanced by the incorporation of high-order structures in mathematical chemistry, especially of hypergraphs. This lifting of high-order structures aims at showing that mathematical chemistry, besides keep delving into the wonders of graph theory and the mathematics of quantum chemistry, may also explore new avenues of research in other mathematical structures, which may eventually contribute to bringing insight to chemistry as one of the outcomes of mathematical chemistry research.

Acknowledgments. G. R. is grateful to Jürgen Jost, Peter Stadler, Douglas Klein, Peter Willett, Rainer Brüggemann, Jos'e L. Villaveces, and Angel Garcia-Chung for generously sharing their insights and wisdom over the years, which have been distilled into the present manuscript.

References

[1] Restrepo, G., Willett, P.: The Journal of Mathematical Chemistry: a bibliometric profile. Journal of Mathematical Chemistry **55**(8), 1589–1596 (2017) https://doi.org/10.1007/s10910-017-0747-7

[2] Restrepo, G., Willett, P.: A bibliometric profile of MATCH Communications in Mathematical and in Computer Chemistry. MATCH Communications in Mathematical and in Computer Chemistry **77**, 235–242 (2017)

[3] Klein, D.J.: Mathematical chemistry! is it? and if so, what is it? HYLE–International Journal for Philosophy of Chemistry **19**(1), 35–85 (2013)

[4] Nye, M.J.: From Chemical Philosophy to Theoretical Chemistry: Dynamics of Matter and Dynamics of Disciplines, 1800-1950. University of California Press, Berkeley (1994). https://books.google.de/books?id=zOXB_AVT08kC

[5] Restrepo, G.: Mathematical chemistry, a new discipline. In: Scerri, E., Fisher, G. (eds.) Essays in the Philosophy of Chemistry, pp. 332–351. Oxford University Press, New York (2016)

[6] Schummer, J.: Why mathematical chemistry cannot copy mathematical physics and how to avoid the imminent epistemological pitfalls. HYLE–International Journal for Philosophy of Chemistry **18**(1), 71–89 (2012)

[7] Bensaude-Vincent, B., Simon, J.: Chemistry: The Impure Science. Imperial College Press, London (2012). https://www.worldscientific.com/worldscibooks/10.1142/p569

[8] Jablonka, K.M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J.D., Bran, A.M., Bringuier, S., Brinson, L.C., Choudhary, K., Circi, D., Cox, S., Jong, W.A., Evans, M.L., Gastellu, N., Genzling, J., Gil, M.V., Gupta, A.K., Hong, Z., Imran, A., Kruschwitz, S., Labarre, A., Lála, J., Liu, T., Ma, S., Majumdar, S., Merz, G.W., Moitessier, N., Moubarak, E., Mouriño, B., Pelkie, B., Pieler, M., Ramos, M.C., Ranković, B., Rodriques, S.G., Sanders, J.N., Schwaller, P., Schwarting, M., Shi, J., Smit, B., Smith, B.E., Van Herck, J., Völker, C., Ward, L., Warren, S., Weiser, B., Zhang, S., Zhang, X., Zia, G.A., Scourtas, A., Schmidt, K.J., Foster, I., White, A.D., Blaiszik, B.: 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. Digital Discovery **2**, 1233–1250 (2023) https://doi.org/10.1039/D3DD00113J

[9] Goodall, R.E.A., Lee, A.A.: Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nature Communications **11**(1), 6280 (2020) https://doi.org/10.1038/s41467-020-19964-7

[10] Klein, D.J.: Prolegomenon on partial orderings in chemistry. MATCH Communications in Mathematical and in Computer Chemistry **42**, 7–21 (2000)

[11] Hoffmann, R., Malrieu, J.-P.: Simulation vs. understanding: A tension, in quantum chemistry and beyond. part a. stage setting. Angewandte Chemie International Edition **59**(31), 12590–12610 (2020) https://doi.org/10.1002/anie.201902527 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201902527

[12] Hoffmann, R., Malrieu, J.-P.: Simulation vs. understanding: A tension, in quantum chemistry and beyond. part b. the march of simulation, for better or worse. Angewandte Chemie International Edition **59**(32), 13156–13178 (2020) https://doi.org/10.1002/anie.201910283 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201910283

[13] Hoffmann, R., Malrieu, J.-P.: Simulation vs. understanding: A tension, in quantum chemistry and beyond. part c. toward consilience. Angewandte Chemie International Edition **59**(33), 13694–13710 (2020) https://doi.org/10.1002/anie.201910285 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201910285

[14] Jost, J.: Mathematical Concepts, p. 312. Springer, Cham (2015)

[15] Johnson, M.A., Maggiora, G.M., Meeting, A.C.S.: Concepts and Applications of Molecular Similarity. A Wiley-Interscience publication. Wiley, New York (1990). https://books.google.de/books?id=iGbJaz1ziWkC

[16] Wang, Y., Wang, J., Cao, Z., Barati Farimani, A.: Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence **4**(3),

279–287 (2022) https://doi.org/10.1038/s42256-022-00447-x

[17] Pettifor, D.G.: A chemical scale for crystal-structure maps. Solid State Communications **51**(1), 31–34 (1984) https://doi.org/10.1016/0038-1098(84)90765-8

[18] Glawe, H., Sanna, A., Gross, E.K.U., Marques, M.A.L.: The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. New Journal of Physics **18**(9), 093011 (2016) https://doi.org/10.1088/1367-2630/18/9/093011

[19] Carbó-Dorca, R., Gironés, X., Mezey, P.G.: Fundamentals of Molecular Similarity. Springer, New York (2013). https://books.google.de/books?id=4v0JswEACAAJ

[20] Carbó-Dorca, R.: Quantum similarity and QSPR in Euclidean-, and Minkowskian–Banach spaces. Journal of Mathematical Chemistry **61**(5), 1016–1035 (2023) https://doi.org/10.1007/s10910-023-01454-y

[21] Restrepo, G.: Chemical space: limits, evolution and modelling of an object bigger than our universal library. Digital Discovery **1**, 568–585 (2022) https://doi.org/10.1039/D2DD00030J

[22] Llanos, E.J., Leal, W., Luu, D.H., Jost, J., Stadler, P.F., Restrepo, G.: Exploration of the chemical space and its three historical regimes. Proceedings of the National Academy of Sciences **116**(26), 12660–12665 (2019) https://www.pnas.org/content/116/26/12660.full.pdf

[23] Whittaker, E.: Eddington's theory of the constants of nature. The Mathematical Gazette **29**(286), 137–144 (1945) https://doi.org/10.2307/3609461

[24] Guggenheimer, K.M.: Fundamental length, fine structure constant and cosmological number. Nature **193**(4816), 664–665 (1962) https://doi.org/10.1038/193664a0

[25] Vopson, M.M.: Estimation of the information contained in the visible matter of the universe. AIP Advances **11**(10), 105317 (2021) https://doi.org/10.1063/5.0064475 https://doi.org/10.1063/5.0064475

[26] Greenwood, N.N., Earnshaw, A.: Chemistry of the Elements. Elsevier Science, New York (2012). https://books.google.de/books?id=EvTI-ouH3SsC

[27] Benjamin, A.T., Quinn, J.J.: Proofs that Really Count: The Art of Combinatorial Proof. Dolciani Mathematical Expositions. Mathematical Association of America, New York (2003). https://books.google.de/books?id=kGD0DwAAQBAJ

[28] Yoo, C.-S.: Chemistry under extreme conditions: Pressure evolution of chemical bonding and structure in dense solids. Matter and Radiation at Extremes **5**(1), 018202 (2020) https://doi.org/10.1063/1.5127897

https://doi.org/10.1063/1.5127897

[29] Restrepo, G., Jost, J.: The Evolution of Chemical Knowledge: a Formal Setting for Its Analysis. Springer, Cham (2022)

[30] Mulas, R., Horak, D., Jost, J.: Graphs, Simplicial Complexes and Hypergraphs: Spectral Theory and Topology, pp. 1–58. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-91374-8_1

[31] Gugisch, R., Kerber, A., Kohnert, A., Laue, R., Meringer, M., Rücker, C., Wassermann, A.: In: Basak, S., Restrepo, G., Villaveces, J. (eds.) MOLGEN 5.0, a Molecular Structure Generator. Advances in Mathematical Chemistry and Applications, vol. 1, pp. 113–138. Bentham Science Publishers B.V., Netherlands (2014). https://doi.org/10.2174/978160805928711401001O

[32] Kerber, A.: Molgen, a generator for structural formulas. MATCH Communications in Mathematical and in Computer Chemistry **80**(3), 733–744 (2018)

[33] Twenty Five Years of Progress in Cheminformatics. https://www.warr.com/25years.html. Accessed: 2022-03-03

[34] Gorse, A.-D.: Diversity in medicinal chemistry space. Current Topics in Medicinal Chemistry **6**(1), 3–18 (2006) https://doi.org/10.2174/156802606775193310

[35] Garcia-Chung, A., Bermúdez-Montaña, M., Stadler, P.F., Jost, J., Restrepo, G.: Chemically inspired Erdös-Rényi oriented hypergraphs (2023)

[36] Fialkowski, M., Bishop, K.J.M., Chubukov, V.A., Campbell, C.J., Grzybowski, B.A.: Architecture and evolution of organic chemistry. Angewandte Chemie International Edition **44**(44), 7263–7269 (2005) https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200502272

[37] Leal, W., Restrepo, G., Stadler, P.F., Jost, J.: Forman–Ricci curvature for hypergraphs. Advances in Complex Systems **24**(01), 2150003 (2021) https://doi.org/10.1142/S021952592150003X https://doi.org/10.1142/S021952592150003X

[38] Estrada, E., Rodríguez-Velázquez, J.A.: Subgraph centrality and clustering in complex hyper-networks. Physica A: Statistical Mechanics and its Applications **364**, 581–594 (2006) https://doi.org/10.1016/j.physa.2005.12.002

[39] Klamt, S., Haus, U.-U., Theis, F.: Hypergraphs and cellular networks. PLOS Computational Biology **5**(5), 1–6 (2009)

[40] Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. In: Schölkopf, B., Platt, J., Hoffman, T. (eds.) Advances in Neural Information Processing Systems, vol. 19. MIT Press, Cambridge (2006). https://proceedings.neurips.cc/paper_files/paper/2006/file/

dff8e9c2ac33381546d96deea9922999-Paper.pdf

[41] Jost, J., Mulas, R.: Hypergraph laplace operators for chemical reaction networks. Advances in Mathematics **351**, 870–896 (2019)

[42] Eidi, M., Jost, J.: Ollivier Ricci curvature of directed hypergraphs. Scientific Reports **10**(1), 12466 (2019)

[43] Eidi, M., Farzam, A., Leal, W., Samal, A., Jost, J.: Edge-based analysis of networks: Curvatures of graphs and hypergraphs. Theory in Biosciences **139**(4), 337–348 (2020)

[44] Mulas, R., Kuehn, C., Jost, J.: Coupled dynamics on hypergraphs: Master stability of steady states and synchronization. Physical Review E **101**, 062313 (2020) https://doi.org/10.1103/PhysRevE.101.062313

[45] Schummer, J.: Scientometric studies on chemistry i: The exponential growth of chemical substances, 1800-1995. Scientometrics **39**(1), 107–123 (1997)

[46] Mahjour, B., Shen, Y., Liu, W., Cernak, T.: A map of the amine-carboxylic acid coupling system. Nature **580**(7801), 71–75 (2020) https://doi.org/10.1038/s41586-020-2142-y

[47] Zhang, W., Oganov, A.R., Goncharov, A.F., Zhu, Q., Boulfelfel, S.E., Lyakhov, A.O., Stavrou, E., Somayazulu, M., Prakapenka, V.B., Konôpková, Z.: Unexpected stable stoichiometries of sodium chlorides. Science **342**(6165), 1502–1505 (2013) https://doi.org/10.1126/science.1244989 https://science.sciencemag.org/content/342/6165/1502.full.pdf

[48] Keserü, G.M., Soos, T., Kappe, C.O.: Anthropogenic reaction parameters - the missing link between chemical intuition and the available chemical space. Chemical Society Reviews **43**, 5387–5399 (2014)

[49] Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang'at, I., Milder, A., Ruby, A.E., Wang, H., Friedler, S.A., Norquist, A.J., Schrier, J.: Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nature **573**(7773), 251–255 (2019) https://doi.org/10.1038/s41586-019-1540-5

[50] Mikulak-Klucznik, B., Gołebiowska, P., Bayly, A.A., Popik, O., Klucznik, T., Szymkuć, S., Gajewska, E.P., Dittwald, P., Staszewska-Krajewska, O., Beker, W., Badowski, T., Scheidt, K.A., Molga, K., Mlynarski, J., Mrksich, M., Grzybowski, B.A.: Computational planning of the synthesis of complex natural products. Nature **588**(7836), 83–88 (2020) https://doi.org/10.1038/s41586-020-2855-y

[51] Schummer, J.: The chemical core of chemistry I: a conceptual approach. HYLE–International Journal for Philosophy of Chemistry **4**(2), 129–162 (1998)

[52] Leal, W., Restrepo, G.: Formal structure of periodic system of elements. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences **475**(2224), 20180581 (2019)

[53] Chapter 2 - similarity in chemical reaction networks: Categories, concepts and closures. In: Basak, S.C., Restrepo, G., Villaveces, J.L. (eds.) Advances in Mathematical Chemistry and Applications, pp. 24–54. Bentham Science Publishers (2015)

[54] Kraut, H., Eiblmaier, J., Grethe, G., Löw, P., Matuszczyk, H., Saller, H.: Algorithm for reaction classification. Journal of Chemical Information and Modeling **53**(11), 2884–2895 (2013)

[55] Andersen, J.L., Flamm, C., Merkle, D., Stadler, P.F.: A software package for chemically inspired graph transformation. In: Echahed, R., Minas, M. (eds.) Graph Transformation, pp. 73–88. Springer, Cham (2016). https://link.springer.com/chapter/10.1007/978-3-319-40530-8_5

[56] Andersen, J.L., Flamm, C., Merkle, D., Stadler, P.F.: Chemical transformation motifs - Modelling pathways as integer hyperflows (2017)

[57] Szymkuć, S., Badowski, T., Grzybowski, B.A.: Is organic chemistry really growing exponentially? Angewandte Chemie International Edition **60**(50), 26226–26232 (2021) https://doi.org/10.1002/anie.202111540 https://onlinelibrary-wiley-com.ezproxy.mis.mpg.de/doi/pdf/10.1002/anie.202111540

[58] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., Laino, T.: Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Science Advances **7**(15), 4166 (2021) https://doi.org/10.1126/sciadv.abe4166 https://www.science.org/doi/pdf/10.1126/sciadv.abe4166

[59] Jost, J., Restrepo, G.: Self-reinforcing Mechanisms Driving the Evolution of the Chemical Space. Perspectives on Science, 1–39 (2023) https://doi.org/10.1162/posc_a_00588 https://direct.mit.edu/posc/article-pdf/doi/10.1162/posc_a_00588/2150616/posc_a_00588.pdf

[60] Andersen, J.L., Flamm, C., Merkle, D., Stadler, P.F.: Inferring chemical reaction patterns using rule composition in graph grammars. Journal of Systems Chemistry **4**(1), 4 (2013) https://doi.org/10.1186/1759-2208-4-4

[61] Paolini, G.V., Shapland, R.H.B., Hoorn, W.P., Mason, J.S., Hopkins, A.L.: Global mapping of pharmacological space. Nature Biotechnology **24**(7), 805–815 (2006) https://doi.org/10.1038/nbt1228

[62] Hann, M.M.: Molecular obesity, potency and other addictions in drug discovery. MedChemComm **2**, 349–355 (2011) https://doi.org/10.1039/C1MD00017A

[63] Restrepo, G., Harré, R.: Mereology of quantitative structure-activity relationships models. HYLE–International Journal for Philosophy of Chemistry **21**(1), 19–38 (2015)

[64] Substances: The ontology of chemistry. In: Woody, A.I., Hendry, R.F., Needham, P. (eds.) Philosophy of Chemistry. Handbook of the Philosophy of Science, vol. 6, pp. 191–229. North-Holland, Amsterdam (2012)

[65] Leal, W., Llanos, E.J., Bernal, A., Stadler, P.F., Jost, J., Restrepo, G.: The expansion of chemical space in 1826 and in the 1840s prompted the convergence to the periodic system. Proceedings of the National Academy of Sciences **119**(30), 2119083119 (2022) https://doi.org/10.1073/pnas.2119083119 https://www.pnas.org/doi/pdf/10.1073/pnas.2119083119

[66] Lipkus, A.H., Watkins, S.P., Gengras, K., McBride, M.J., Wills, T.J.: Recent changes in the scaffold diversity of organic chemistry as seen in the cas registry. The Journal of Organic Chemistry **84**(21), 13948–13956 (2019)

[67] Lipkus, A.H., Yuan, Q., Lucas, K.A., Funk, S.A., Bartelt, W.F., Schenck, R.J., Trippe, A.J.: Structural diversity of organic chemistry. a scaffold analysis of the cas registry. The Journal of Organic Chemistry **73**(12), 4443–4451 (2008) https://doi.org/10.1021/jo8001276 . PMID: 18505297

[68] Bishop, K.J.M., Klajn, R., Grzybowski, B.A.: The core and most useful molecules in organic chemistry. Angewandte Chemie International Edition **45**(32), 5348–5354 (2006) https://doi.org/10.1002/anie.200600881 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200600881

[69] Brown, D.G., Boström, J.: Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? Journal of Medicinal Chemistry **59**(10), 4443–4458 (2016) https://doi.org/10.1021/acs.jmedchem.5b01409 https://doi.org/10.1021/acs.jmedchem.5b01409. PMID: 26571338

[70] Feng, Y., Han, J., Ying, S., Gao, Y.: Hypergraph Isomorphism Computation (2023)