COMPUTATIONAL HISTORY OF CHEMISTRY

Guillermo Restrepo, Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany; restrepo@mis.mpg.de

Abstract

In this essay it is shown how mathematical and computational approaches can be used to model the underlying mechanisms of historical processes, which transform the structure, dynamics and function of chemistry. By chemical knowledge, I refer to a complex dynamical system emerging from the interaction of the social, material and semiotic systems of chemistry. Besides instantiating some watershed events of the history of chemistry in this framework, the increasing availability of large datasets amenable to computational exploration is discussed, as well as the suitable mathematical theories to carry out these studies. I show how this framework allows for exploring possible alternative histories of chemistry by perturbing its past, leading to solving questions of the sort "what would have happened if." This not only sheds light on the past of chemistry, it rather allows modelling the future of the discipline, with its societal and pedagogical reaches. This approach complements conventional methodologies for the history of chemistry and becomes an interdisciplinary field of research for linguists, mathematicians, physicists, historians and chemists, to name but a few scholars and scientists.

"The search for regularities in human history is becoming a trifle more respectable than it was formerly. That could well portend some significant improvement in our ability to discuss the human future."

—Murray Gell-Mann, 2017 (1)

1 Introduction

The charm of the history of chemistry lies in its convincing, and often literary, narratives of past events whose temporal paths have influenced the evolution of chemistry. History of chemistry shows how the backbone of chemistry emerges from a multidimensional and noisy dynamics of contingencies and certainties. Wonderfully it teaches us about friendships, rivalries, cooperations, academic-industrial alliances, professionalization and technologies, which when embedded in changing social and scientific contexts lead us to the chemistry of the twenty first-century.

History elaborates on the past but it should not be forced to wait until events happen. Its niche lies in the past, but it also includes the present and future, as well as the possible pasts, presents and futures. History spans all tenses. History of chemistry, therefore, ought not beguile us only with the past of chemistry, but with its present and future reaches. Moreover, given the key societal role of chemistry, the history of chemistry endeavors with the past, present and future of our civilization (2).

Tracing back the conditions leading to the discovery of penicillin, or to the development of the Haber-Bosch process is paramount, especially if we want to disentangle the workings of innovation. Similarly, analyzing the conditions facilitating the production and commercialization of thalidomide or the formulation and use of Napalm is extremely relevant to avoid repeating these

disasters. These constitute a few examples of the enormous contributions and responsibilities of the history of chemistry. Addressing these questions entails that history of chemistry, at its core, concerns itself with the search for regularities leading to causal relationships (3).

However, most of the historical work, although illuminating, is far from the search for patterns. And this has its own history, as expected. Carl von Clausewitz (1780-1831) and Leo Tolstoy (1828-1910) believed that historical processes were driven by some sort of law, an idea supported by several nineteenth- and twentieth-century historians (4, 5). Nevertheless, the concept of history as a scientific enterprise in the search for patterns lost popularity in the second half of the twentieth-century (6); a movement epitomized by Karl Popper (1902-1994) with his critics to search for historical regularities to foretell the future. Currently, searching for regularities is assumed as a task of the natural sciences, which deal with far less complex systems than those of the humanities. All in all, particles, atoms and genes lack free will, which facilitates the detection of their patterns, while people and their organizations are unpredictable and prone to act upon contingencies. However, scientists are typically busy working within their specialties and very seldom venture beyond their disciplines to seek regularities. Nevertheless, one would expect that chemists, having a strong tradition of detecting patterns, could easily look into their discipline and detect the relevant threads that after being separated from noise would resolve into the driving forces shaping chemistry. The fact of the matter shows that historians are left alone in this central task.

In this essay I argue that Clausewitz and Tolstoy's belief in historical processes driven by patterns is not only a reality but that the moment is ripe to undertake these studies seizing upon present computational and mathematical capacity to delve into the colossal corpus of chemical information gathered to date.

2 History of Chemistry and the Search for Patterns

Historians are very good at devising narratives, often involving causal relationships, where they weave the different dimensions of the historical subject under study. It is gripping to find how nineteenth-century organic chemistry benefited from the strong academe-industry relationship in Germany (7); how professional rivalry hindered the recognition of watershed chemical constructions or theories, for example the frictions between Arrhenius and Mendeleev or between the former and Nernst (8). History contributes to understanding how the deluge of organic chemicals in the first quarter of the nineteenth-century led to devising the molecular structural theory (9), disciplinarily so rooted in our chemical minds. It is fascinating to find that Guyton de Morveau, Lavoisier, Berthollet, de Fourcroy, Hassenfratz and Adet's revolutionary nomenclature (10) was motivated by philosophy of language, which can be traced back to Leibniz (11). These and several other remarkable developments in the history of chemistry evidence how chemistry and its knowledge have been driven by material, social and semiotic aspects of the discipline (2).

Then, in its broadest sense, the history of chemistry entails the temporal analysis of events leading to the status of what has been regarded as chemistry in a given time, this latter discipline understood as the science devoted to transforming matter and theorizing upon it. The events of interest for the history of chemistry involve complex social, semiotic and material factors whose interaction is driven by contingencies and regularities. The question that arises is about the essence of these events. What is the currency of the history of chemistry? I claim it is chemical knowledge.

2.1 Chemical Knowledge, the Currency of History of Chemistry

Jürgen Renn has framed the history of science within a broader history of knowledge (12). This entails considering the history of science as the study of knowledge and of its evolution. In this setting the different dimensions of historical events are integrated into the new object of study: knowledge. Following Renn, Jürgen Jost and I posited that the history of chemistry entails analyzing the evolution of chemical knowledge (2). But, what is knowledge and what is specifically chemical knowledge?

Humans accumulate experiences, which they cognitively structure allowing for predictions of new experiences (13). *Knowledge* entails developing those cognitive structures and predicting, as well as the feedback resulting from predictions (12, 14). When predictions are realized, new experiences are added to the cognitive structures and the predictive method is strengthened. Otherwise, new experiences falling outside the scope of the initial experiences are used to tune the predictive model, while enlarging the cognitive structures. Therefore, knowledge is a dynamical process and it depends on social constraints, such as economic and political interest; as well as cognitive frameworks, which involve theories and meaning generation. Knowledge is stored, shared

and transmitted across generations. Thus, it requires a material system to be preserved and spread, for instance through signs and text (2).

Chemical knowledge involves the cognitive structures generated to make sense of the experimentation upon matter transformations. These structures are either used by chemists to estimate future outcomes of their experiments or to modify their cognitive structures in such a manner that they span the new experimental findings. As chemists are embedded in social frameworks, which vary over time and which determine particular ways of thinking (15), their cultures and semiotic systems influence chemical knowledge, which also depends on the materials and technologies available for exploring matter transformations (2).

In our account of the evolution of chemical knowledge we claim that it can be modelled as a *complex dynamical system* made of at least three interacting systems. These are the semiotic, material and social systems (2). As typical systems, they are made of objects and their relationships (16, 17). The objects of the social system include people, academic and scientific societies, committees, enterprises, industries and other forms of social organization, plus computational objects such as robots and artificial intelligence technologies. These social objects are held together by economic, political, cultural, academic and other relations (2). The semiotic system, following Peirce's distinction among objects, signs and interpretants (18, 19), involves substances, reactions and other concepts of chemistry along with their historical representations (signs) that chemists (interpretants) have associated to them (2). These semiotic objects are related by the ternary relation object-sign-interpretant of Peirce's semiotics and by the high-order relations resulting from their combinations (20). The material system is made of substances, reactions, technologies and apparatus, as well as the relationships they establish in the chemical practice (2, 21). Our setting is that chemical knowledge arises from the mutual interaction of the semiotic, social and material systems of chemistry. Chemical knowledge is an emerging property and history of chemistry involves the analysis of the dynamics of this complex object.

So far this does not bring anything new, it sounds just like the use of complex systems jargon to describe what historians of chemistry have been doing for over a century. Nonetheless, this setting brings new possibilities for the history of chemistry. The theory of complex systems regards the emergence of macroscopic phenomena as resulting from multiple relations caused by simple dynamical rules (22), which are often detected through the patterns they form (23). Hence, if by studying past events of chemistry, we detect rules or patterns leading to the rules, we will be in a good position to make predictions and even retrodictions, this latter understood as "predictions" about the past.

The most realistic case of the two mentioned is that of detecting patterns in a large corpus of chemical information about the social, material and semiotic systems of chemistry. If that happens, we have good reasons to think there is an underlying rule driving the historical process. The other case is a bit more difficult, as it involves the direct detection of the rule driving the complexity of the evolution of chemical knowledge. In any case the rule is accepted or rejected insofar as it reproduces the patterns of chemical knowledge.

The first case, that is from patterns to rules, leads to modeling. If we find a pattern in the evolution of chemical knowledge, we may devise a model and let it evolve over time. If the resulting pattern obtained from the model matches that of the historical evolution of chemical knowledge, we have good reasons to take the model as encoding the driving force of the pattern (24). The second case, from model to pattern, is a derivation of the first one. This requires the direct evaluation of the validity of the model by contrasting it with the historical pattern.

Thus, considering chemical knowledge as a complex dynamical system leads to a model, which brings the history of chemistry to new reaches. It allows for estimating the future outcomes of the observed pattern. Interestingly, it also allows for retrodictions by running the model on arbitrary pasts and letting it predict events occurring afterwards but still in the past (2). This is particularly suitable to solve questions of the sort, "what would have happened if." One could ask, for instance, what would have happened with the material and social systems of chemistry, and with chemical knowledge in general, if Berzelius had not devised its notation of empirical formulae. What would have happened if the pre World War I conditions of chemical knowledge had been maintained for longer (25).

In short, the history of chemistry involves the detection of historical patterns in the evolution of chemical knowledge, which we may model as a complex dynamical system arising from the mutual interaction of the semiotic, material and social systems of chemistry. The question that now arises is how to do it. I argue that these patterns are to be detected by analyzing large

corpora of historical chemical data through mathematical and computational tools.

3 The Necessity of Mathematical and Computational Methods for Studies in the History of Chemistry

Computational approaches to the history of science have been recently recognized as complementary to the practice of history and sociology of science (26-29). As recently noted by Abraham Gibson, Manfred D. Laubichler and Jane Maienschein (30), history, as a discipline, is currently undergoing a computational revolution. In recent years *Isis* has published some papers on "the computational turn," "the computational revolution" and "the electronic information revolution" (31-33) and in 2019 the journal dedicated a focus section to "Computational history and philosophy of science" (30). In 2017 the annual issue of *Osiris* was devoted to "Historicizing Big Data" (34). In turn, in the *American Historical Review* "the digital revolution" and "the digitized revolution" has been discussed (35). Despite the pros and cons of computation in the practice of history, addressed in the above references, it is clear that "digital sources and computational tools have transformed how we engage with the historical record, including the history of science" (30).

What does this computational turn offer for the history of chemistry and for the search for historical patterns? Computational approaches allow for processing large amounts of historical data and, when coupled with mathematical and statistical methods, for detecting the sought-after historical patterns, if they actually exist. It is important to note that these patterns are not observable by traditional history of science methods, which are often restricted to analyzing periods spanning decades and covering specific geographical regions. Furthermore, such studies typically rely on fewer than hundreds of primary and secondary sources, only some of which are digitized (36). In contrast, datasets for computational studies depend upon millions of digitized records spanning centuries and often the whole globe (37). This change of scale signals a change in the kinds of analysis offered by computational approaches to the history of chemistry.

Furthermore, I propose that mathematical approaches to the history of chemistry promote insights that are otherwise unsupported speculations, or simply unavailable because they are out of reach. For example, the proposal that chemical knowledge is a complex dynamical system involving social, material and semiotic components can be elaborated by employing a suite of mathematical theories for pattern formation, evolution and adaptation, nonlinear dynamics, as well as systems, network, game and collective behavior theories (38). I discuss some particular instances of these theories in section 5. But the central point of mathematical methods for the history of chemistry goes beyond the use of mathematical theories as "canned" tools of straightforward application (2). Rather, I posit that close interactions between chemists, historians, and mathematicians generates fruitful interdisciplinary work while also stimulating some surprising insights into the history of chemistry (39), a topic further discussed in section 6, below. Some of the aforementioned theories are instrumental for detecting patterns and analyzing statistical properties. The products of such combined mathematical methods allow us to glimpse the future of chemical knowledge.

A case study of the mathematical and computational approach to the history of chemistry follows.

4 The Evolution of the Chemical Space

Recent studies of the growth of the chemical space provide a case study of patterns in the history of chemistry (40). By chemical space I mean the substances reported over the history of chemistry, which have been extracted or synthesized by chemists, apothecaries, pharmacists, metallurgists and other chemistry practitioners (2, 41). These substances are endowed with a notion of nearness by chemical reactions and therefore constitute a mathematical space (42). Hence, one talks of substances that are closely related by very few synthetic steps, in contrast to the majority of other substances which are far apart in our current knowledge of possible synthesis plans connecting them. Note that the notion of nearness is arbitrary, as one could select other criteria of nearness. For example, substances can be characterized by their molecular structures and their nearness can be determined by the resemblance of their structures (43).

In any case, the chemical space is an important concept leading to new questions. For example: how does it grow? How rapidly? How are its dynamics affected by social perturbations such as wars or pandemics? Is it perturbed by semiotic changes? Moreover, can we model its evolution? Which are the rules driving its dynamics?

In 1963 de Solla Price briefly discussed the annual report of some chemical starting materials and the growth of chemical elements (44). The first complete account

of the growth of the chemical space was reported by Joachim Schummer in 1997, who analyzed the period 1800-1995 by manually screening the indexes of eight printed sources, including handbooks of organic and inorganic chemistry (45). He found an exponential growth with an annual growth rate $r = 5.5\%$, indicating a doubling time of about 13 years. A further study analyzed the growth of organic substances by computationally treating the Beilstein database for the period 1850-2004 and an exponential growth was also found, with $r = 8.3\%$ before 1900 and $r = 4.4\%$ afterwards (46).

In a more recent account, we analyzed 16,356,012 reactions and 14,341,955 substances published between 1800 and 2015 in chemical journals, gathered in the Reaxys electronic database. We found that the chemical space has historically grown at an exponential rate, with a stable growth rate ($r = 4.4\%$) (40). This indicates that about each 16 years chemists have doubled the number of new substances reported. The speed of this rapid chemical production can be expressed in these terms: 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 (47)!

In our model of chemical knowledge as a complex dynamical system, we claim that chemical knowledge is driven by the mutual interaction of the semiotic, social and material systems of chemistry. Evidence of these interactions, at least in their binary forms, have been already reported and discussed in (2).

We found that the expansion of the chemical space has been affected by social setbacks such as World Wars (WWs) (40). We observed two drops in chemical production around WWs and quantified the effect of these events in the annual output of new chemicals. It was found that WW I sent back chemical production 37 years, while WW II 16 years. The dramatic effect of WW I follows from the centralized structure of chemistry in the first quarter of the twentieth century, whose capital was Germany. WW I actually motivated a restructuring of chemical industrial and research production, prompting a decentralization in which the USA began to take the lead. By the time of WW II, the social system of chemistry had changed and production was sufficiently less centralized such that, to a large extent, WW II did not affect the annual output of new chemicals.

An interesting semiotic event leading to changes in the material system of chemistry was the introduction of the molecular structural theory (2). This theory solved a semiotic crisis, mainly driven by the deluge of organic substances brought about by the improvement of analytical methods of the early nineteenth-century (48, 49). Before chemists gained control over organic substances, their extractions and synthesis, chemical knowledge was mainly driven by inorganic chemistry (49, 50). Therefore, its semiotics and theoretical structure was tailored to these substances. In this period, for instance, the Berzelian dualistic theory became a paper tool to understand the chemical space and to keep expanding it (48).

At any rate, the 1830s and 1840s brought a growing number of organic substances, which challenged the dualistic formulae of Berzelius and the way of thinking these formulae incorporated in the chemistry of the first half of the nineteenth century (2, 48). Finding a growing number of very different substances with the same composition was totally unexpected, for instance. The introduction of the molecular structure as a new semiotic object of chemistry added a new dimension to the Berzelian algebra of formulae and to the Lavoisian concept of classes of substances (2). Chemists had a new and powerful paper tool at their hands, a topological one where elements hold relationships through chemical bonds. Molecular structures added to the visual character of chemistry and, as Berzelian formulae, became a way of thinking for expanding the chemical space (2).

Our data-driven study of the evolution of the chemical space, besides revealing a stable historical growth, one not halted by social setbacks such as WWs (51), showed other patterns and sharp transitions. This is a further instance of a result attainable only by applying mathematical and computational tools to vast corpora of chemical information. By analyzing the variability of the annual output of new chemicals, we found three clear statistical regimes of chemical production, the first one spanning the period 1800-1860, the second running between 1860 and 1980 and the third and present one beginning in 1980 (40).

Typically, growth studies involve analyzing growth as the slope of the growth curves. However, in our interdisciplinary work, we were fortunate enough to bring mathematicians to our study, for example experts in time series analyses, who indicated to us the importance of the variability of the signal. Duc H. Luu found three statistical regimes where the variability of the annual output of chemicals has been normally distributed (40). These are

the three statistical regimes just mentioned, which are characterized by an emphasis on inorganic chemistry before 1860, where the variability of the annual output of new chemicals was the highest in the history of chemistry. This indicates an exploratory regime of the chemical space, where some years brought several new substances and some others not so many. It is clear that the size of the chemical community also played a major role.

By 1860 there was a drastic reduction of the variability, which we argued was caused by the widespread introduction of the structural theory that regularized chemical production, assisted by an ever growing chemical community, which was benefiting from the positive public image of chemistry and of its important chemical industry (9). The interplay between a powerful chemical theory and the social conditions led to regularize the annual output of chemicals, which were mostly organic substances. This trend lasted more than a century—impressively, with two WWs in between! It is noteworthy that WWs did not delay chemical production as observed in other disciplines according to their bibliographic production (44), but rather caused a drop in production followed by a rapid recovery after WWs, leading chemical production to pre-WWs trends (40). This is what Schummer has dubbed as a *catching-up phenomenon* (45). Although he has explored the role of the size of the chemical community in this effect, further work is needed to gauge the dynamics of these postwar recoveries.

A second sudden reduction of the variability of the annual production of chemicals was observed around 1980, but it is still an open question as to the leading event or collection of events causing this further regularization of the variability. It could be caused by the increasing computerization of the chemical practice or a delayed effect of the widespread adoption of spectroscopic chemical instrumentation that took place in the 1950s. At any rate, this third regime of chemistry was characterized by a revival of compounds containing metals, specifically organometallics, followed by a surge of substances of biological interest (40).

Having discussed the mathematical and computational analysis of the evolution of the chemical space, I now turn to discuss the data and methods we have, and we expect to have, to further extend the computational history of chemistry.

5 Data and Methods

Chemistry is the science with the largest output of publications (2) associated to its material practice. Therefore, it is not short of data. Moreover, chemists have developed a strong tradition of data curation, annotation, storage and dissemination initiated by the encyclopedists of the thirteenth century such as Batholomaeus Anglicus (before 1203-1272), Vincent of Beauvais (c. 1190- c. 1264) and Albertus Magnus (1193-1280) (49) and continued by towering nineteenth-century figures such as Leopold Gmelin (1788-1853) and Friedrich Konrad Beilstein (1838-1906). The efforts of these pioneers, plus the colossal amount of new chemical data of the twentieth and twenty-first century are today at our fingertips through electronic databases, which besides being a source of chemical information, constitute an important corpus of historical information.

A well organized electronic database of proven use for history of chemistry is *Reaxys*, owned by Elsevier, and a further one is *SciFinder* hosted by the American Chemical Society. These sources offer important information about the material system of chemical knowledge, which includes substances, reactions, substance properties and reaction conditions. This data is also associated to authors and to their affiliations. Hence, these databases may be used as a source of data for the social system of chemistry as well. As most of the reactions stored in these sources contain details about how the reactions were performed, these databases contribute to the corpus needed to explore the semiotic system of chemistry (2).

Despite the advantages of the chemical databases, historians still lack a well organized and curated database of information about the semiotic and social systems of chemistry. In (2) we mention some reliable sources which could be used to build up such historical databases, which include other electronic databases such as the *ISI Web of Knowledge* and *Dimensions*. However, these databases do not convey a complete picture of the social and semiotic systems. The methods of historians are needed to collect, curate and digitize, for instance, membership in chemical societies, chemical industries and registration records of academic institutions, if the aim is building up a complete database for the semiotic and social systems (2).

Perhaps the most challenging system is the semiotic one, because gathering its data requires developing standards for selecting and curating information. For example, it is important to define what counts as a

diagram, table or reaction scheme in historical records. This entails going beyond a formal definition as it has to account for the natural evolution of these terms and concepts. A further question concerns the storage of these data: should they be saved as images? Or is it better to convert them into a machine readable format regardless of whether it is human readable (52)?

Here we must briefly examine the history of codification of molecular structures and reactions. In the 1950s chemists started to ponder the problem of computational encoding of molecular structures (53), which led to connection tables, later on to SMILES and today to InChIs (2, 54). These frameworks were mainly adjusted to encode organic molecules, therefore presenting difficulties for encoding organometallic or inorganic compounds (2). Moreover, they are based on the "chemistry" of the molecules, as encoded in their molecular structures, which disregards several dimensions of their semiotic load. For instance, the intention of the chemist who draws one particular shape, among many possible alternatives to represent a molecule, is lost in these encodings. Fortunately, the growing computational memory and processing capacity is making it possible to store and process molecular structures as images (55), which are coupled with machine learning algorithms to advance chemical knowledge, nevertheless mainly concentrated on the material system. We consider this an opportunity to use current computational power and algorithms to collect chemical information of relevance for the semiotic system of chemistry, and, in general, for the evolution of chemical knowledge.

In (2) we discuss different computational and mathematical methods we find appropriate for studies on the history of chemistry. I have mentioned how time series analysis becomes a powerful tool to analyze historical data. The quality of results obtained through these methods depends to a large extent on the temporal resolution of the data. In section 4 we show the application of these methods to data with annual resolution. Current publishing speeds and editorial policies are making it possible to think about continuous sources of data, which will sharpen the possibilities of time series analysis methods (56). A time series analysis study takes a random temporal signal allowing for studying temporal trends and forecasting (57), as well as detecting changes and evolving behaviors (58). This technique has been recently recognized by historians of science as a suitable method to assist their narratives and to find causal relationships (59).

A powerful collection of mathematical techniques is found in statistical physics that involves the dynamic behavior of mathematical structures, such as graphs and hypergraphs (60-62), which I further discuss in section 6. In general these structures are suitable models for chemical knowledge and its constitutive systems. The general idea is to define a set of objects of study and some relations among them of interest for historical study. Hence, for the material system one may think about analyzing how substances relate to each other through chemical reactions and how that structure of reactions has evolved over time. Likewise, one may consider the dynamics of the social structure of chemistry relating people and institutions as well as the temporal behavior of the connections between concepts and other semiotic tokens of chemical knowledge. Graphs, and in general hypergraphs, are perfectly suited to study the dynamics of the relationships between the systems of chemistry from which chemical knowledge emerges (2).

Another important tool is agent-based modelling, which is used to model simple local interaction patterns and to understand the emerging global complex dynamics. Thus, these models are especially appropriate for understanding the evolution of chemical knowledge. Besides allowing for estimations, agent based models become relevant to study retrodictions, for example, of competing narratives in the history of chemistry (2).

Text analysis tools as well as natural language techniques become important parts of the computational tools for the history of chemistry. These techniques allow for detecting concepts and topics of importance throughout the evolution of chemical knowledge which can be applied, for instance, to treat large corpora of chemical abstracts and reaction details. Therefore, these techniques are applicable to each one of the systems of chemistry (social, semiotic and material) as well as to our exploration of their mutual interplay (2).

A further tool historians can profit from is machine learning, which runs through large databases allowing for predictions, for example through regression techniques, or for classifications. These algorithms also involve reinforcement learning, where the algorithm make decisions based on rewards that it intends to maximize (2).

Bayesian networks constitute a powerful tool for the practice of history of chemistry as they provide a formal setting to explore causal relationships (63), of utter importance for weaving historical narratives. In this setting a time series signal is perturbed and the temporal propagation of the perturbation is analyzed (2).

Some other computational tools and mathematical settings are discussed in (2), but nothing restricts the use of others and the development of new ones motivated by questions from the history of chemistry.

6 Mathematics Triggered by Research on the History of Chemistry

Two questions that computational history of chemistry may address, which I have so far not discussed, are: can we model the evolution of chemical space? What are the rules regulating its dynamics? The answer to the first question is "yes, we can" and we are working on gauging the right model for the evolution of the chemical space. With information about the use of substances to expand the space we can propose statistical models based on the dynamics of chemical reaction networks, described as hypergraphs. In (40) we analyzed whether there are statistical patterns in the way chemists combine their substrates in chemical reactions. We found that although a large part of chemistry is exploratory, that is chemists combine new chemicals with other new chemicals to produce novel substances, there is also a high degree of conservatism in the expansion of the chemical space. We dubbed this trend as the *fix substrate approach* and it entails reactions of very well known substrates, such as acetic anhydride, with new substances (64). Thus, as we know the statistical patterns of the annual participation of substrates in reactions, we can model the expansion of the chemical space. The estimated space can therefore be contrasted with the actual chemical space as recorded, for instance in Reaxys. The mathematical setting to encode participation of substances in chemical reactions is that of hypergraphs, which may be regarded as belonging to the field of network science. What follows is a detailed discussion of how these mathematical structures afford this encoding and how history of chemistry questions have triggered mathematical research in this area.

The success of network theory as a suitable model for systems of an ample range of disciplines has not left chemistry untouched. Collaboration networks and relationships among disciplines are but a few examples of uses of networks to analyze cases of chemical interest. An interesting case at the core of chemistry is Schummer's claim that the logical structure of chemical knowledge corresponds to a network of substances related through chemical reactions (65). This is what today is dubbed as a chemical reaction network. Graph theory has become the mathematical setting of choice for modeling networks (66). In a graph (or network), the objects of the system under study constitute the vertices (or nodes) of the network and the different relationships among objects, the edges of the network. A familiar graph-like chemical concept is that of molecular structure. In this setting, atoms and bonds, respectively, correspond to vertices (V) and edges (E) of the graph, which embodies the molecular structure. A graph G is the couple (V, E). As in a typical molecule, where pairs of atoms are bonded, in a graph, edges connect pairs of vertices (67).

Not surprisingly, graphs have been used to model networks of chemical reactions (46). However, Klamt, Haus and Theis showed in 2009 that these structures miss an important piece of chemical information when used to model chemical reactions (68). As it is well known, reactions typically entail sets of substances. That is, a collection of substrates, solvents, catalysts, which are typically heated and stirred to end up with a mixture of products, which are later on diligently separated and analyzed (69). Klamt, Haus and Theis's argument is that the essential information of a chemical reaction is the AND connector for some substances. We say, A reacts with B to produce C and D, for example, as the result of an experiment (70). Here the important information is that A AND B react together to produce C AND D. If we model this reaction through a substrate-product graph, where substrates are related with products in a directed fashion, that is with arrows from every substrate to each product (71), then we obtain the following relations: A → C, A → D, B → C and B → D. Klamt, Haus and Theis show that this model does not allow for deducing that A AND B are the substrates of the reaction. This occurs because there are further possible interpretations of this model such as that A decomposes into C AND D. The same can be said for the decomposition of B. We can also infer that A AND B react together to produce C through an addition reaction and that under other reaction conditions they produce D. This lack of accuracy of the model, or of several interpretations, multiply as the network grows, that is as more and more reactions are concatenated in the reaction network.

Klamt, Haus and Theis provide the solution to this conundrum by highlighting the possibilities of hypergraphs to gauge the essence of the AND relation of chemical reactions. The basic idea is that reactions simply relate sets of substrates with sets of products. Hence, we talk of {A, B} → {C, D} as a suitable reaction model emphasizing the direction from the left hand side set of substrates to the right-hand side set of products. The main difference between graphs and hypergaphs is that while graphs depict only relations between pairs of ele-

ments, hypergraphs do it for sets of elements of any size. As seen, hypergraphs are a novel concept in chemistry and it is also of recent appreciation in mathematics (72), where the attention has been mostly on graphs and their mathematical and statistical properties (73). Nevertheless the further study of these properties for hypergraphs is gaining momentum, in part, motivated by efforts to model large networks of chemical reactions. Interestingly, hypergraphs are a generalization of graphs in such a manner that all that is known about graphs should be reduced to a single case of a general theory of hypergraphs (74). This is an example of mathematics triggered by the study of the history of chemistry. In fact, the study of directed hypergraphs, such as chemical reactions, is just beginning and constitutes a niche of vibrant mathematical research motivated by chemical questions (75).

These mathematical studies may shed light not only on the very concept of chemical space but on the whole of chemical knowledge, as we have argued that the most suitable model for studying the interaction of the three systems of chemical knowledge is a hypergraph. The dynamics of hypergraphs allows for exploring arbitrary moments in the temporal unfolding of chemical knowledge, that is, they lead to predictions and retrodictions, which I discuss in the next section.

7 Predictions and Retrodictions

Detecting statistical patterns in the history of chemistry enables the historical endeavor of the chemical practice to extend beyond the past by allowing extrapolations. It further allows perturbing the past to come up with alternative pasts, presents and futures of chemistry. All in all, statistical patterns in the history of chemistry allow for simulations, predictions and retrodictions.

In 2005 Grzybowski and his team modelled chemical prices based on the use the chemical community made of substances (46). They found that the price of a chemical rapidly decreases with both, the number of synthetic ways of producing it (k_{in}) and the number of uses the substance has as starting material of other chemicals (k_{out}). In fact, the cost follows a power-law distribution that is proportional to $k^{-\nu}$, with k being either k_{in} or k_{out} (46). This research was conducted with a fraction of the known chemical space, namely the organic chemical space between 1850 and 2004. Hence, these trends apply to prices of organic chemicals. The same team found that chemists have traditionally used organic substrates of 150 g mol^{-1}, on average, to produce substances with an average of 250 g mol^{-1} (in 2004) (46). This pattern may shed light on the large scale trends of organic chemistry. However, further statistical tests need to be conducted to analyze the reliability of possible estimations.

In section 4 we discussed an important pattern in the evolution of the chemical space, namely that for more than 200 years the growth rate of the chemical space has been stable, with doubling times of about 16 years and with variabilities that have been reducing over time. This trend allows us to predict an expansion of chemical space at a similar pace (76). This leads us to think that an extrapolation is possible and that eventually we could know when, for instance, a specific number of new chemicals would be afforded. However, even if the present variability of the annual output of new substances is the lowest, still the uncertainty of the estimations is high (40). This result shows two important things: i) that extrapolations must be supported by sound statistical analyses to avoid oversimplifications and ii) that estimations of the expansion of the chemical space are complex, presumably calling for information from the social and semiotic systems of chemical knowledge. This latter point highlights the relevance of building up strong social and semiotic databases to foresee the future of the discipline.

A further possibility for computational history of chemistry lies in the retrodictions it permits. This basically entails perturbing past historical data to observe the effects upon subsequent events of that arbitrary past. We conducted a retrodictive study when analyzing the influence of the chemical space upon the evolution of the periodic system (77). In that study we took the known chemical space between 1800 and 1869, with annual resolution, and by taking several systems of atomic weights in vogue in that period, we perturbed the chemical space to obtain the possible periodic systems that several leading chemists such as Dalton and Berzelius may have devised if they had attempted to do so with the chemical space at their disposal (78). We also analyzed how the chemical space by the time of Gmelin, Meyer, Odling, Hinrichs and Mendeleev influenced the systems they reported (79). This allowed us to calculate the false positives and true negatives rates for these chemists by contrasting the periodic systems obtained from the chemical space and those reported by them.

8 Conclusions and Outlook

The increasing amount of data and of computational power are turning computational approaches into an integral part of the historians' tools, which when combined

with mathematical insight open new possibilities for the practice of history of chemistry. Here I have shown how computational history of chemistry provides new ways to solve historical questions and allows for asking and solving novel questions related to large scale historical patterns.

Arguing that the currency of the history of chemistry is chemical knowledge, I show how a computational history of chemistry sheds light on the evolution of knowledge by regarding it as a complex dynamical system made of the mutual interaction of the social, semiotic and material systems of chemistry.

Different sorts of data available to undertake studies belonging to the setting here presented are analyzed, as well as several methods and theoretical frameworks assisting those studies. Likewise, I discuss the rich sources of information for the material system of chemistry, vastly documented in chemical databases, and I emphasize the lack of databases for the social and semiotic system. I argue that this is actually an opportunity for the history of chemistry, rather than a limitation, where historians can decide on the sorts of data relevant to curate and to preserve as well as their formats. Clearly, historians cannot create these databases alone; they will require the support of computer scientists and chemists, neither of which alone can come up with databases of relevant use for the history of chemistry.

Computational history of chemistry therefore requires above all interdisciplinary work and constitutes an interesting research niche at least for historians of chemistry, chemists, physicists, computer scientists, mathematicians, philosophers and semioticians. As discussed in (2), the success of these approaches depends, more than on data and methods, on the common and synchronized work of different disciplines, which requires being able to understand the language and jargon of other specialists and even other ways of thinking. All in all, the success of this setting for the history of chemistry is a matter of interdisciplinary respect and of interdisciplinary thinking.

Besides interdisciplinary research, the approaches here discussed also involve interdisciplinary research assessment. That is, when refereeing a publication or research proposal of this sort, a traditional historian, with little knowledge of computation, data analysis or the mathematical techniques discussed, will face problems, as she/he cannot go beyond the validity of the historical question and the kind of data suitable for the study. Likewise, computer scientists, mathematicians and chemists cannot go beyond judging the methods, tools and logic of the subject under study (80). A complete and holistic evaluation of research in this field requires the common work of different experts, with coordinated communication among them during the assessment process.

The success of novel outcomes discussed in this essay necessitate ready access to databases and adequate computing capacity. Traditional history of science departments do not count on computational resources and when these are available, they normally do not offer the minimal conditions of memory and processing capacity to handle large corpora of chemical information. Moreover, they typically do not count on system administrators in charge of machine and software updates. Therefore, alliances between computer science departments and history and chemistry ones are essential.

Based on an analysis by Daniel J. Cohen and Roy Rosenzweig of the pros and cons of computational history (81), in general, the approach here presented solves several issues of the traditional practices of history of chemistry. It brings almost unlimited storage capacity, as information is stored in electronic form. Traditional storage capacities are driven by library and archive capacities of printed records such as books and letters. In contrast, a colossal amount of information can be stored in tiny computer memories. Another advantage is the accessibility to information. This is already evident when consulting century-old printed records by searching, for instance in the Internet Archive. Several users can get direct and simultaneous access to the same records without leaving their desks (82). Computational history of chemistry also offers flexibility, as digital information may be converted into different formats, ranging from text and image to audio and video. One can, for instance, think of recorded interviews of leading chemists or even politicians, which are stored as sound but can easily be transformed into text. Graphical abstracts and videos of current publications also enlarge the possibilities of the data sources for the historical practices of the present and the future.

Disciplinary diversity is a further advantage of the mathematical and computational approaches. The very fact of accessibility to the information makes it possible that not only historians can access the archives, but also practicing chemists and other scholars and scientists with interests in the history of chemistry. Large repositories of information for historical studies contribute to the manipulability and searchability of information we have become accustomed to in the current information era. It is almost unimaginable to search for a particular string of characters in hundreds of handwritten records

of the Hellenistic, Chinese or Arabic alchemy without computational aid. Computation has turned this task into a customary process, where segmentation and optical character recognition techniques allow for rapid and efficient queries. The approaches here presented also offer the possibility of interactivity, where databases can be fed from different sources, by different specialists (and even laypeople as in the case of Wikipedia) and the different users can interact and discuss particular subjects, leaving digital records of their interactions, which constitute sources of further historical inquiry.

Despite the advantages of computational history of chemistry, there are challenges and problems to be faced. One of these is data quality. Building up large repositories of information relies on the common work of several annotators and curators. This practice may sound foreign for traditional historians, which often look for primary sources delving into the archives and rely basically on their own interpretations. The approaches here presented often rely on commercial databases which are regularly updated. These updates may modify records of previous entries, for example information dumped today covering the nineteenth century may vary from another dump from the same database and spanning the same period but performed months later. This occurs because annotators and curators may include new sources, for instance journals not considered before, or patents; or because annotation errors are found and corrected. Although this poses a problem for historical studies, as the historical source may vary, it is actually a minor problem as long as the bulk of the database is not affected. The reason is statistical. If updates and corrections do not change the historical trends, they may be considered as background noise of the signal; otherwise, the previous observed trends correspond to data artifacts, which could have been detected running statistical tests over the data. In any case, it is a good policy to perform regular dumps of the data and run statistical tests to determine whether variations are significant or not.

A further issue of these approaches is the inaccessibility to some of the relevant data. Computational history of chemistry requires accessing large databases, some of them owned by private companies or involving costly subscription licenses, which are normally granted to institutions rather than to individuals. Moreover, mathematical and computational approaches require special access to these databases, which go far beyond the limited number of queries allowed to typical users and which often imply dumping the entire database. This issue can be overcome by signing collaboration agreements with database owners. This has been my particular experience with Reaxys, for instance. Providers of other relevant databases such as Clarivate and Digital Science & Research Solutions Inc. offer a variety of similar opportunities for academic partnerships. As mentioned, providing access on individual basis is not the norm, rather, these companies generally grant access to research groups who already have the infrastructure to store and process the information. Typically, a further condition is that access is granted based on a guarantee of information security to avoid data leaks, which may affect the commercial interests of database owners.

A large part of this essay has been devoted to depicting how methods of the natural sciences, especially those devised to look for patterns, find application in the history of chemistry. However, this relationship is far from being unidirectional, and it should be actually taken as a symmetric relation, where historians have a central role in the advancement of the different sciences. As Robin George Collingwood (1889-1943) claimed (83):

> The methods of historical research have, no doubt, been developed in application to the history of human affairs; but is that the limit of their applicability? They have already before now undergone important extensions: for example, at one time historians had worked out their methods of critical interpretation only as applied to written sources containing narrative material, and it was a new thing when they learnt to apply them to the unwritten data provided by archaeology. Might not a similar but even more revolutionary extension sweep into the historian's net the entire world of nature? In other words, are not natural processes really historical processes, and is not the being of nature an historical being?

Acknowledgements

I thank Michael Gordin, Evan Hepler-Smith, Ernst Homburg, Jeffrey Johnson, Jürgen Jost, Ursula Klein, Manfred Laubichler, Farzad Mahootian, Carsten Reinhardt, Jürgen Renn, Camilo Restrepo, Alan Rocke, Jeff Seeman and Joachim Schummer, for interesting and motivating discussions, over the years, about some aspects discussed in this essay.

References and Notes

1. M. Gell-Mann, "Regularities in Human Affairs," in D. Krakauer, J., Gaddis and K. Pomeranz, Eds., *History, Big History, & Metahistory*, ch. 4, pp 63-90, SFI Press, Santa Fe, 2018.

2. G. Restrepo and J. Jost, *The Evolution of Chemical Knowledge, a Formal Setting for its Analysis*, Wissenschaft und Philosophie—Science and Philosophy, Springer, forthcoming.
3. Note that I am not advocating either for a teleological or for a whig history. I put forward a history of chemistry advancing hand-in-hand with mathematics and statistics, and as I will argue later, also with computer sciences.
4. P. Turchin, "Toward Cliodynamics—an Analytical, Predictive Science of History," *Cliodynamics*, **2011**, *2*, 167-186, https://doi.org/10.21237/C7clio21210.
5. W. McNeill, *The Rise of the West: A History of the Human Community,* University of Chicago Press, 2009. https://books.google.de/ books?id=_RsPrzrsAvoC (accessed 12 Oct. 2021).
6. As Turchin has remarked, there were still some supporters of the idea of history as the search for patterns such as S. H. Williamson, "The History of Cliometrics," *Research in Economic History*, Suppl. 6, **1991**, 15-31 and V. Bonnell and L. Hunt, "Introduction," in V. Bonnell and L. Hunt, Eds., *Beyond the Cultural Turn: New Directions in the Study of Society and Culture*, University of California Press, Berkeley, 1999, pp 1-32.
7. J. A. Johnson, "Germany: Discipline—Industry—Profession. German Chemical Organizations, 1867-1914," in A. K. Nielsen and S. Štrbáňová, *Creating Networks in Chemistry: The Founding and Early History of Chemical Societies in Europe*, Royal Society of Chemistry, Cambridge, UK, 2008, ch. 6, pp 113-138.
8. R. M. Friedman, *The Politics of Excellence: Behind the Nobel Prize in Science*, W. H. Freeman, New York, 2001.
9. A. Rocke, "The Theory of Chemical Structure and its Applications," in M. J. Nye, Ed., *The Cambridge History of Science,* vol. 5, *The Modern Physical and Mathematical Sciences*, Publisher: Cambridge University Press, Cambridge, UK, 2002, pp 255-271.
10. I am referring to Louis-Bernard Guyton de Morveau (1737-1816), Antoine-Laurent de Lavoisier (1743-1794), Claude Louis Berthollet (1748-1822), Antoine François, comte de Fourcroy (1755-1809), Jean-Henri Hassenfratz (1755-1827) and Pierre-Auguste Adet (1763-1834).
11. G. Restrepo and J. L. Villaveces, "Chemistry, a Lingua Philosophica," *Found. Chem.*, **2011**, *13*, 233-249.
12. J. Renn, *The Evolution of Knowledge: Rethinking Science in the Anthropocene*, Princeton Univ. Press, Princeton, NJ, 2020.
13. A recent study models how we operate upon experiences to build up such cognitive structures. See C. W. Lynn and D. S. Bassett, "How Humans Learn and Represent Networks," *Proc. Nat. Acad. Sci. USA*, **2020**, *117*, 29407-29415, https://doi.org/10.1073/pnas.1912328117.
14. J. Renn, "The Evolution of Knowledge: Rethinking Science in the Anthropocene," *HoST—Journal of History of Science and Technology*, **2018**, 12, 1-22, https://doi.org/10.2478/host-2018-0001.
15. M. Foucault, *Les mots et les choses*, Gallimard, Paris, 1966.
16. L. v. Bertalanffy, "An Outline of General System Theory," *Br. J. Philos. Sci.*, **1950**, *1*, 134-165, https://doi.org/10.1093/bjps/I.2.134.
17. A typical chemical system is a system of elements, where chemical elements hold order and similarity relationships (W. Leal and G. Restrepo, "Formal Structure of Periodic System of Elements," *Proc. R. Soc. (London) A*, **2019**, *475*, 20180581, http://dx.doi.org/10.1098/rspa.2018.0581). I note in passing that systems of elements are not set in stone, but rather historical objects emerging from the evolution of chemical knowledge. See G. Restrepo, "Compounds Bring Back Chemistry to the System of Chemical Elements," *Substantia*, **2019**, *3*, 115-124 (2019), https://doi.org/10.13128/Substantia-739 and G. Restrepo, "Das periodische System und die Evolution des chemischen Raums," *Nachrichten aus der Chemie*, **2020**, *68*, 12-15, https://doi.org/10.1002/nadc.20204094740.
18. C. S. Peirce, *The Essential Peirce, Volume 2*, Indiana University Press, Bloomington and Indianapolis, 1998.
19. T. L. Short, *Peirce's Theory of Signs,* Cambridge University Press, Cambridge, UK, 2007.
20. An example of this ternary relation is the association of the sign "water" to the colorless liquid wetting materials, quenching our thirst, flowing in rivers, boiling at 100 °C, dissolving table salt, plus several other features; an association mediated by the interpretant (2). By high-order relations I mean the further relations between ternary related semiotic objects. For instance, a chemical reaction such as

    $$A + B \rightarrow C + D$$

    corresponds to a high-order relation entailing the ternary semiotic objects A, B, C and D. The ternary relation in A is given by the sign the interpretant has given to this substance, likewise occurs with B, C and D. As we have shown elsewhere (2), a suitable high-order structure for chemical reactions is a hypergraph, which I discuss in section 6.
21. In Ref. 2, we have highlighted the strong relationship between the semiotic and the material systems, as there are semiotic objects which, through experimental evidence, have been introduced as objects of the material system. This is the case of atoms and molecules, first devised as chemical entities lacking physical reality used to advance chemical knowledge. Later experimental results led to the adoption of Avogadro's hypothesis and the kinetic theory of gases, which prompted the introduction of atoms and molecules as material entities of the chemical practice.

See A. J. Rocke, *Chemical Atomism in the Nineteenth Century*, Ohio State University Press, Columbus, 1984.

22. C. Hooker, "Introduction to Philosophy of Complex Systems: A: Part a: Towards a Framework for Complex Systems," in C. Hooker, Ed., *Philosophy of Complex Systems*, vol. 10 of *Handbook of the Philosophy of Science*, North-Holland, Amsterdam, 2011, pp 3-90, https://doi.org/10.1016/B978-0-444-52076-0.50001-8.

23. An example of simple rules leading to the emergence of a pattern is the *game of life* by John Horton Conway (1937-2020). The setting is as follows (A. Adamatzky, *Game of Life Cellular Automata*, Springer, London, 2010, https://books.google.de/books?id=5iz6C0zzWKcC): imagine a graph paper sheet, where every square can be either black (dead) or white (live). The color of each square is determined by the colors of an initial set of squares that one selects at one's will. Each square is surrounded by eight other squares. The rules read that any live square with two or three live neighboring squares survives. Any dead square with three live neighboring squares becomes a live square. All other live squares die in the next generation. Likewise, all other dead squares stay dead. The first generation of squares is created by simultaneously applying the aforementioned rules to every square. Births and deaths occur simultaneously. Each generation depends on the preceding one and the rules continue to be applied repeatedly to create further generations. The Wikipedia article, "Conway's Game of Life (https://en.wikipedia.org/wiki/Conway%27s_Game_of_Life, (accessed 12 Oct. 2021)) depicts interesting visualizations.

24. In practice, accepting a model requires some further steps, such as validating it. This entails, for example, perturbing (or deleting) part of the input data to observe the stability of the model.

25. This was a period with one of the most rapid annual outputs of new chemicals in the history of chemistry, as shown in E. J. Llanos, *et al*., "Exploration of the Chemical Space and its Three Historical Regimes," *Proc. Nat. Acad. Sci. USA*, **2019**, *116*, 12660-12665, https://doi.org/10.1073/pnas.1816039116.

26. A. Buyalskaya, M. Gallo and C. F. Camerer, "The Golden Age of Social Science," *Proc. Nat. Acad. Sci. USA*, **2021**, *118*, https://doi.org/10.1073/pnas.2002923118.

27. J. Li, Y. Yin, S. Fortunato, and D. Wang, "Scientific Elite Revisited: Patterns of Productivity, Collaboration, Authorship and Impact," *Journal of The Royal Society Interface*, **2020**, *17*, 20200135, https://doi.org/10.1098/rsif.2020.0135.

28. A. Edelmann, T. Wolff, D. Montagne and C. A. Bail, "Computational Social Science and Sociology," *Annu. Rev. Sociol.*, **2020,** *46*, 61-81, https://doi.org/10.1146/annurev-soc-121919-054621.

29. J. Schummer, "The Impact of Instrumentation on Chemical Species Identity," in P. Morris, Ed., *From Classical to Modern Chemistry: The Instrumental Revolution*, Royal Society of Chemistry, Cambridge, UK, 2002, pp 188-211.

30. A. Gibson, M. D. Laubichler and J. Maienschein, "Introduction," *Isis*, **2019**, *110*, 497-501, https://doi.org/10.1086/705542

31. A.-L. Post and A. Weber, "Notes on the Reviewing of Learned Websites, Digital Resources, and Tools," *Isis*, **2018**, *109*, 796-800, https://doi.org/10.1086/701651.

32. M. D. Laubichler, J. Maienschein and J. Renn, "Computational Perspectives in the History of Science: To the Memory of Peter Damerow, *Isis*, **2013**, *104*, 119-130, https://doi.org/10.1086/669891.

33. S. P. Weldon, "Introduction," *Isis*, **2013,** *104*, 537-539, https://doi.org/10.1086/673272.

34. E. Aronova, C. v. Oertzen, and D. Sepkoski, "Introduction: Historicizing Big Data," *Osiris,* **2017,** *32*, 1-17, https://doi.org/10.1086/693399.

35. L. Putnam, "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast," *American Historical Review*, **2016**, *121*, 377-402, https://doi.org/10.1093/ahr/121.2.377.

36. When I state that traditional historical studies concentrate on specific regions and on periods of decades, I am mainly referring to studies leading to journal publications. There are, nevertheless, important publications, mainly books, where historians present large analyses of longer periods and often of global scope. These studies typically rely on journal publications as well as on other secondary sources.

37. Here sizes are relative to the largest possible historical records. Hence, one can take as a reference for the material system all substances, reactions, instrumentation and technologies used in the history of chemistry. Likewise, the size of the records pertaining to the social system is determined by the number of names and personal data of all alchemists, apothecaries, metallurgists and chemists in the history of chemistry as well as their organizations and institutions. It would also require the inclusion of novel technologies such as robots directly interacting with humans in the practice of chemistry. The size of the semiotic system is given by the collection of all semiotic objects devised and used by practitioners of chemistry over the evolution of chemistry. It is in comparison with this scale that I argue that a traditional historical study relies on small datasets, which are nevertheless a subset of the datasets of computational studies.

38. Particular fields of mathematics in play include partial differential equations, cellular automata, artificial neural networks, evolutionary computation, genetic algorithms, machine learning, time series analysis, agent-based modelling, as well as bifurcation, information, (hyper)graph and complexity theories.

39. There exist examples of mathematics motivated by chemistry. For instance, nineteenth-century settings of

molecular structures as mathematical objects contributed to graph theory with the works of James Joseph Sylvester (1814-1897). Likewise, Arthur Cayley (1821-1895), motivated by the question on the number of isomers for some alkanes, contributed to enumerative mathematics, which were expanded by George Pólya (1887-1985) in the twentieth century. Further information on this subject is found in D. J. Klein, "Mathematical Chemistry! Is It? and if So, What Is It?" *Hyle*, **2013**, *19*, 35-85 (http://www.hyle.org/journal/issues/19-1/klein.pdf, accessed 12 Oct. 2021) and G. Restrepo, "Mathematical Chemistry, a New Discipline," in E. Scerri and G. Fisher, Eds., *Essays in the Philosophy of Chemistry*, Oxford University Press, Oxford and New York, 2016, ch. 15, pp 332-351, doi:10.1093/oso/9780190494599.003.0023.

40. Llanos, *et al*. (Ref. 25).
41. Restrepo, "Das periodische System..." (Ref. 17).
42. Although there is no unique definition of mathematical space, this concept encodes the idea of a set endowed with a notion of nearness. For example, if nearness is associated with distance, then we talk about metric spaces. If it is associated with relationships among elements, then this remits to topological spaces.
43. This is, for example, the basis of the Quantitative Structure-Activity Relationship paradigm.
44. D. J. de Solla Price, *Little Science, Big Science,* Columbia University Press, New York, 1963.
45. J. Schummer, "Scientometric Studies on Chemistry I: The Exponential Growth of Chemical Substances, 1800-1995," *Scientometrics*, **1997**, *39*, 107-123.
46. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, "Architecture and Evolution of Organic Chemistry," *Angew. Chem. Int. Ed.*, **2005**, *44*, 7263-7269, doi:10.1002/anie.200502272.
47. A natural question is how this speed compares to those of other spaces, for instance biological spaces. In this case, one has also to account for rates of extinction and of mutation. See J. W. Bull and M. Maron, "How Humans Drive Speciation as Well as Extinction," *Proc. R. Soc. (London) B*, **2016**, *283*, 20160600, https://doi.org/10.1098/rspb.2016.0600.
48. U. Klein, *Experiments, Models, Paper Tools: Cultures of Organic Chemistry in the Nineteenth Century*, Stanford University Press, Stanford, CA, 2003.
49. H. Leicester, *The Historical Background of Chemistry*, Dover Publications, 1971 (first published Wiley, 1956).
50. W. H. Brock, *The Norton History of Chemistry*, W. W. Norton and Company, New York, 1993.
51. See below in this section for a discussion on the recovery after WWs.
52. Discussing these and other related questions was one of the main reasons we organized the recent *Computational approaches to the history of chemistry* meeting (Max Planck Institute for Mathematics in the Sciences, Leipzig, March 2021), where historians, chemists, computer scientists and mathematicians analyzed different aspects of data driven approaches for the practice of history of chemistry. It was clear that more work is required to advance the construction of electronic databases for the social and semiotic systems of chemistry.
53. L. C. Ray and R. A. Kirsch, "Finding Chemical Records by Digital Computers," *Science*, **1957**, *126*, 814-819, doi:10.1126/science.126.3278.814.
54. Connection tables are lists of atoms and bonds belonging to a molecular structure with further information such as coordinates in a two- or three-dimensional space. SMILES (Simplified Molecular-Input Line-Entry System) are string representations of molecular structures. International Chemical Identifiers (InChIs) are strings used to encode as much as possible information about chemical substances, which include their associated molecular structures.
55. J. Staker, K. Marshall, R. Abel and C. M. McQuaw, "Molecular Structure Extraction from Documents Using Deep Learning," *J. Chem. Inf. Model.*, **2019**, *59*, 1017-1029, https://doi.org/10.1021/acs.jcim.8b00669.
56. Publications before the digital era had regular discrete outputs ranging from annual to weekly, a periodization that is becoming more continuous with the advent of preprint servers and with the online publication of papers right after acceptance, while proofs are edited (2).
57. Z. Wu, N. E. Huang, S. R. Long and C.-K. Peng, "On the Trend, Detrending, and Variability of Nonlinear and Nonstationary Time Series," *Proc. Nat. Acad. Sci. USA*, **2007**, *104*, 14889-14894, https://doi.org/10.1073/pnas.0701020104.
58. D. Brillinger, "Time Series: General," In N. J. Smelser and P. B. Baltes, Eds., *International Encyclopedia of the Social and Behavioral Sciences*, 15724-15731, Pergamon, Oxford, 2001, https://doi.org/10.1016/B0-08-043076-7/00519-2.
59. M. D. Laubichler, J. Maienschein and J. Renn, "Computational History of Knowledge: Challenges and Opportunities," *Isis*, **2019**, *110*, 502-512, https://doi.org/10.1086/705544.
60. A. Aleta and Y. Moreno, "Multilayer Networks in a Nutshell," *Annu. Rev. Condens. Matter Phys.*, **2019**, *10*, 45-62.
61. M. M. Danziger, I. Bonamassa, S. Boccaletti and S. Havlin, "Dynamic Interdependence and Competition in Multilayer Networks," *Nature Physics*, **2019**, *15*, 178-185.
62. P. Chodrow and A. Mellor, "Annotated Hypergraphs: Models and Applications," *Applied Network Science*, **2020**, *5*, 9.

63. J. Pearl, *Causality*, Cambridge University Press, Cambridge, UK, 2009.

64. This finding goes hand in hand with others indicating also preferences for some particular reactions to explore the chemical space, actually the space of substances of potential pharmacological interest. In this case, for example, Suzuki-Miyaura is one of the preferred reactions since 1980 (D. G. Brown and J. Boström, "Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone?" *J. Med. Chem.*, **2016**, *59*, 4443-4458, https://doi.org/10.1021/acs.jmedchem.5b01409.).

65. J. Schummer, "The Chemical Core of Chemistry I: A Conceptual Approach," *Hyle*, **1998**, *5*, 129-162, http://www.Hyle.org/journal/issues/4/schumm.htm (accessed 12 Oct. 2021).

66. Sociologists were one of the first to recognize the advantages of graph theory. The same community further extended the mathematical possibilities of these structures, which drew the attention of mathematicians. Graph theory is today a vibrant area of research in discrete mathematics. Although I highlight the important role of sociologists for the expansion of graph theory, I do not disregard the early works of mathematicians such as Leibniz and Euler in presenting instances of graphs and their mathematics.

67. Interesting discussions on the development of chemical graph theory are found in Klein (Ref. 39).

68. S. Klamt, U.-U. Haus and F. Theis, "Hypergraphs and Cellular Networks," *PLOS Computational Biology*, **2009**, *5*, 1-6.

69. This is a very broad depiction of a chemical reaction in solution. One could claim that there are also solid state reactions, or even one-pot reactions and others where the separation process is automated, or others where the evolution of the reaction is monitored in real time through the direct insertion of a probe into the reaction vessel (B. Wittkamp, In Situ Monitoring of Chemical Reactions—a Molecular Video, https://www.chemeurope.com/en/whitepapers/126365/in-situ-monitoring-of-chemical-reactions-a-molecular-video.html, (accessed 12 Oct. 2021)). In any case the general idea of reactions as relating sets of chemicals holds and it is an essential part of what a chemical reaction is.

70. The development of the concept of chemical reaction is of central historical and epistemological relevance for understanding the evolution of chemical knowledge. Much work in this direction is needed, where, for instance, semioticians and philosophers of science may contribute to a large extent. At any rate, Lavoisier initiated chemical semiotics by using chemical equations to describe material transformation in terms of substances undergoing changes (11). Further discussions on alternative ways to describe chemical transformations are found in (2)

71. In graph theory these arrows correspond to arcs.

72. Claude Berge (1926-2002) in the 1970s analyzed several of their properties and discussed their delayed recognition in mathematics: C. Berge, *Graphs and Hypergraphs*, North-Holland Mathematical Library, North-Holland, Amsterdam, 1973.

73. W. Leal, G. Restrepo, P. F. Stadler and J. Jost, "Forman-Ricci Curvature for Hypergraphs," *Advances in Complex Systems*, **2021**, *24*, 2150003, https://doi.org/10.1142/S021952592150003X.

74. In this setting a hypergraph H corresponds to the couple (V, E), where V are the objects and E contains the subsets of V that are related. For example, for the reaction A + B → C + D, its corresponding hypergraph is $H = (V, E)$, with $V = \{A, B, C, D\}$ and $E = \{(\{A, B\}, \{C, D\})\}$.

75. In (2) we discussed some of the mathematical properties already studied for hypergraphs and some others still to be explored.

76. Our results show no sign of saturation in this expansion, as de Solla Price wrongly anticipated for the whole science in the 1960s (44).

77. W. Leal, E. J. Llanos, A. Bernal, P. F. Stadler, J. Jost and G. Restrepo, "Computational Data Analysis Shows that Key Developments Towards the Periodic System Occurred in the 1840s," *ChemRxiv,* **2021**.

78. The underlying hypothesis of this study is that the chemical space directly influenced the ordering and similarities of the chemical elements, therefore the periodic system. See Restrepo "Compounds Bring Back…" (Ref. 17) and G. Restrepo, "Challenges for the Periodic Systems of Elements: Chemical, Historical and Mathematical Perspectives," *Chem. Eur. J.*, **2019**, *25*, 15430-15440, https://doi.org/10.1002/chem.201902802.

79. These are some of the formulators of periodic systems, ranging from Leopold Gmelin (1788-1853) in the 1840s to Julius Lothar Meyer (1830-1895), William Odling (1829-1921), Gustavus Detlef Hinrichs (1836-1923) and Dmitri Ivanovich Mendeleev (1834-1907) in the 1860s. See E. Scerri, *The Periodic Table: Its Story and Its Significance*, Oxford University Press, New York, 2nd ed., 2019, https://books.google.de/books?id=9x2yDwAAQBAJ.

80. Sometimes, even for computer scientists it is difficult to judge the value of a computational approach as reproducibility may become an issue.

81. D. Cohen and R. Rosenzweig, *Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web*, University of Pennsylvania Press, Philadelphia, 2006, available online at https://chnm.gmu.edu/digitalhistory/ available online at https://chnm.gmu.edu/digitalhistory/ (accessed 12 Oct. 2021).

82. One is tempted to argue that this practice contributes to lowering global warming emissions, as trips to consult archives and other sources are reduced to a large extent.

However, the environmental costs of computation are far from being disregarded, as shown in E. Strubell, A. Ganesh and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3645-3650*,* Association for Computational Linguistics, Florence, Italy, 2019, https://www.aclweb.org/anthology/P19-1355 (accessed 12 Oct. 2021). The question that arises concerns the net environmental costs of the approaches here presented and how they actually compare with the traditional approaches to the history of chemistry.

83. R. G. Collingwood and W. J. van der Dussen, *The Idea of History*, ACLS Humanities E-Book, Oxford University Press, Oxford, UK, 1994, https://hdl.handle.net/2027/heb.05489.

About the Author

Guillermo Restrepo is a chemist working at the Max Planck Institute for Mathematics in the Sciences (Germany). His scientific interests include the evolution of chemical knowledge, the relationship between chemistry and mathematics and the history and philosophy of chemistry.

Born in the Same Year as HIST (1922)

- Har Gobind Khorana: January 9, 1922, in Raipur, in the Punjab in (then British) India.
- Robert William Holley: January 28, 1922, in Urbana, Illinois, USA.
- George C. Pimentel: May 2, 1922, in Rolinda, California, USA.
- John Goodenough: July 25, 1922 in Jena, Germany.

Khorana and Holley, born in the same month in widely separated places, shared the 1968 Nobel Prize in Physiology or Medicine for their roles in deciphering the genetic code, the link between nucleic acid and protein sequences. Khorana and his co-workers at the University of Wisconsin-Madison synthesized short chains of RNA which were used to direct the synthesis of short protein fragments. Later Khorana and co-workers made what is considered the first artificial gene. Holley and his co-workers at Cornell University and the US Plant, Soil and Nutrition Laboratory on the Cornell campus, isolated and sequenced alanine transfer RNA, which directs incorporation of alanine into proteins.

Pimentel's research is well known to physical chemists and his service to chemical education and the chemical profession is widely recognized in the American Chemical Society. Matrix isolation methods and chemical lasers are his main legacies in physical chemistry. He received the highest award of the American Chemical Society, the Priestley medal, in 1989, the year of his death. He had served as ACS president in 1986. The ACS award in chemical education was renamed the George C. Pimentel Award in Chemical Education in his honor.

Goodenough shared the 2019 Nobel Prize in Chemistry for the development of lithium ion batteries, ubiquitous power-storage devices for mobile electronics. Goodenough and co-workers made their key contributions to this technology at the University of Oxford in the 1970s and 1980s. The Materials Chemistry division of the Royal Society of Chemistry have an award named for Goodenough.