Exploration of the chemical space and its three historical regimes

Eugenio J. Llanos[a,b,c,d,1], Wilmer Leal[a,b,1], Duc H. Luu[b,e], Jürgen Jost[b,f], Peter F. Stadler[a,b,f,g,h], and Guillermo Restrepo[b,h,2]

[a]Bioinformatics Group, Department of Computer Science, Universität Leipzig, 04107 Leipzig, Germany; [b]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany; [c]Grupo de Biomatemáticas, Fundación Instituto de Inmunología de Colombia, 111321 Bogota, Colombia; [d]Grupo Chima, Corporación SCIO, 111321 Bogota, Colombia; [e]Institute of Mathematics, Vietnam Academy of Science and Technology, 10307 Hanoi, Vietnam; [f]The Santa Fe Institute, Santa Fe, NM 87501; [g]Institute for Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria; and [h]Interdisciplinary Center for Bioinformatics, Universität Leipzig, 04107 Leipzig, Germany

Edited by Peter Schuster, University of Vienna, Vienna, Austria, and approved May 17, 2019 (received for review September 17, 2018)

Chemical research unveils the structure of chemical space, spanned by all chemical species, as documented in more than 200 y of scientific literature, now available in electronic databases. Very little is known, however, about the large-scale patterns of this exploration. Here we show, by analyzing millions of reactions stored in the Reaxys database, that chemists have reported new compounds in an exponential fashion from 1800 to 2015 with a stable 4.4% annual growth rate, in the long run neither affected by World Wars nor affected by the introduction of new theories. Contrary to general belief, synthesis has been the means to provide new compounds since the early 19th century, well before Wöhler's synthesis of urea. The exploration of chemical space has followed three statistically distinguishable regimes. The first one included uncertain year-to-year output of organic and inorganic compounds and ended about 1860, when structural theory gave way to a century of more regular and guided production, the organic regime. The current organometallic regime is the most regular one. Analyzing the details of the synthesis process, we found that chemists have had preferences in the selection of substrates and we identified the workings of such a selection. Regarding reaction products, the discovery of new compounds has been dominated by very few elemental compositions. We anticipate that the present work serves as a starting point for more sophisticated and detailed studies of the history of chemistry.

history of chemistry | chemical space | chemical reactions | World War | structural theory

Chemical space (1, 2) comprises the possible set of chemical species. Most of them do not occur naturally, but have to be synthesized in chemical laboratories. Therefore, chemical space has been explored either by extraction from natural sources or by chemical reactions throughout the history of chemical research. We may still be rather ignorant about the full extension of this space, but at least we can trace how what is known today gradually emerged from more than 200 y of chemical research.

Chemical reactions have been systematically documented in databases, already since the 19th century up to the present, and currently this chemical corpus comprises millions of reactions.

The first study of the growth of chemical substances for the period 1800–1995 manually analyzed the indexes of eight printed sources, including handbooks of organic and inorganic chemistry (3). An exponential growth with an annual rate $r = 5.5\%$ was found. A second study considered the growth of organic substances by computationally treating the Beilstein database (now part of Reaxys) for the period 1850–2004 and an exponential growth was also found, with $r = 8.3\%$ before 1900 and $r = 4.4\%$ afterward (4).

In the present study we use mathematical tools to computationally analyze the chemical space explored up to the present. Besides studying the growth of production of chemicals, we also address its variability. Some other open questions on the exploration of the space are, e.g., Has the exploration been affected and to which extent by social and scientific events? Is chemical synthesis that central for the exploration as generally accepted? As there are more and more substances, therefore available substrates, can we identify the workings of substrate selection to explore the chemical space? Are chemists actually reaching new regions of the space? In this paper we answer these questions by analyzing millions of chemical reactions spanning more than 200 y of history.

The Three Regimes: Growth of Production of Compounds and the Explored Chemical Space

Analysis was performed on data stored in Reaxys*, a large repository of chemical information built from the Beilstein and Gmelin Handbooks and the Patent Chemistry Database, which covers data from 16,400 journals and patents (5). After a filtering procedure (*Materials and Methods*), our dataset comprises 14,341,955 compounds associated with 16,356,012 reactions from 1800 to 2015 reported in scientific journals.

Considering the publication year of a compound as its earliest report in a reaction of the database, the annual number of new compounds grows exponentially (Fig. 1*A*), following a heteroskedasticity model (6) (*Materials and Methods*, Eq. **1**) that distinguishes three historical regimes. The statistics of these regimes, plus some periods later discussed, are shown in

Significance

We found that the number of new chemical compounds has grown exponentially with a 4.4% annual production rate from 1800 to 2015 not even affected by World Wars. There are three distinct growth regimes: proto-organic, organic, and organometallic, with decreasing variability in the production of compounds over time. Contrary to the belief that organic synthesis developed only after 1828, synthesis had been a key provider of new compounds already at the beginning of the 19th century. By 1900, it became the established tool to report new compounds. We found that chemists are conservative when selecting starting materials and that despite the growing production of new compounds, most of them belong to a restricted set of chemical compositions.

Author contributions: P.F.S. and G.R. designed research; E.J.L., W.L., and D.H.L. analyzed data; and J.J. and G.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] E.J.L. and W.L. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: restrepo@mis.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1816039116/-/DCSupplemental.

Published online June 11, 2019.

*Copyright 2019 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited. Reaxys data were made accessible to our research project via the Elsevier R&D Collaboration Network.

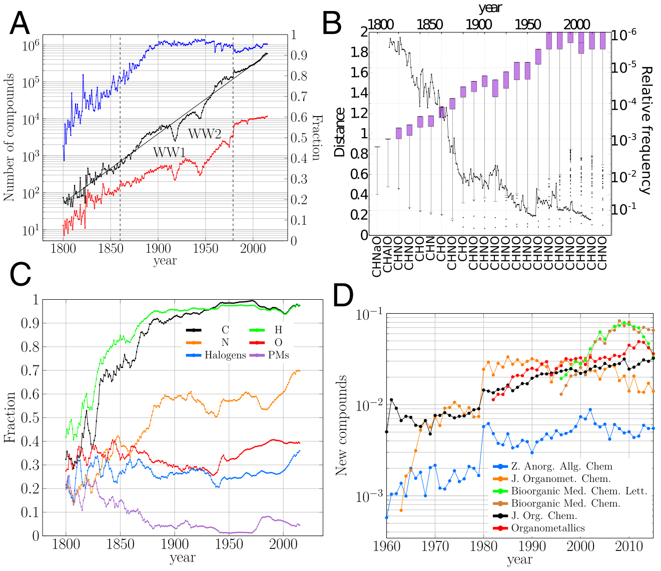

Fig. 1. Growth of compounds. (*A*) Annual number of new compounds (black) and plot of the fitted equation (left axis [l.a.]; *Materials and Methods* and Eq. **2**). WW1 and WW2 indicate the World War periods and the vertical dotted lines the change of production regime. Annual number of new compositions (l.a., red) and fraction of new synthesized compounds to the total of new ones (right axis [r.a.], blue) are shown. (*B*) Distance among compositions of successive years (l.a.; *Materials and Methods*; and *SI Appendix*, Fig. S3), box plots of the compositions every 10 y (*SI Appendix*, Fig. S4) with interwhisker distance accounting for 99.9% of the data, and most popular combinations of elements (compositions) reported in new compounds. Relative frequency of compositions is shown in the r.a. (*C*) Annual fraction of new compounds containing C, H, N, O, halogens, and platinum metals (PMs). Distributions are convoluted using the moving-average method with a 5-y window. (*D*) Annual number of new compounds reported in some specialized journals.

Table 1 and *SI Appendix*, Table S1. They are characterized by their growth rate and variability in the number of compounds reported (*Materials and Methods*; Eq. **1** and *SI Appendix*, Fig. S1). Remarkably, as our statistical analysis reveals, the transitions between regimes were quite sharp. In particular, while the general decrease of variability may simply be a consequence of the growth of the chemical research community (3), the sudden changes that emerged from our analysis cannot be so easily explained by such general factors.

In the first regime chemists reported compounds with an annual growth rate μ of 4.04% and with the highest standard variation σ (Table 1), indicating a very random period in the year-to-year output of new compounds. While this was the period with the highest percentage of metal compounds reported (Fig. 1*C* and *SI Appendix*, Table S2), C and H compounds nevertheless dominated during the entire period: In the first two decades of the regime they were present in about half of the compounds. At the end of this regime, 70% of the compounds contained C and H, but no metals (Fig. 1*C*). In fact, the second half of the regime was mainly characterized by C-, H-, N-, O-, and halogen-based compounds. During this proto-organic regime, the chemical space was explored mainly through the extraction and analysis of animal and plant products (7) with inorganic compounds (Tables 2 and 3 and *SI Appendix*, Tables S2 and S5). However, as we will show, synthesis also played a more active role than generally conceded.

In the second regime, compounds were reported somewhat faster and in a less variable fashion (Table 1). Remarkably, by 1880 C and H compounds constituted 90% of the new substances, and this has remained so ever since. In contrast, the proportion of metals decreased (Fig. 1*C*; Tables 2 and 3; and *SI Appendix*, Table S3). Organic chemistry became dominant, and compositions (combinations of elements associated to their compounds) such as CHNO became omnipresent (*Materials and Methods*; Fig. 1*B*; and *SI Appendix*, Table S3); more details on the role of organic substrates, products, and targets can be found in Tables 2 and 3 and *SI Appendix*, Table S5. The low variance of this regime indicates more regularity in the year-to-year report of new compounds. This goes hand in hand with the normalization brought to chemistry by the framework of valence and structural theory, which transformed research in organic chemistry around 1860 (8, 9). We call this period the organic regime.

The third period started around 1980. The percentage of C and H compounds dropped. N compounds also initially fell, but later rose to a historical maximum (Fig. 1*C*). There was a revival of metal compounds, both as substrates and as products (Tables 2 and 3 and *SI Appendix*, Table S5); e.g., 10% of the new compounds were platinum metals based. Silicon compounds, almost negligible over the history of chemistry, actually led off in this regime (*SI Appendix*, Table S4), CHOSi and CHNOSi being some of the most popular Si compositions in the period. This shift toward organometallic chemistry can also be seen in the journal landscape, for instance by looking at the compounds reported in specialized journals in the early 1980s, such as the *Zeitschrift für Anorganische und Allgemeine Chemie*, the *Journal of Organometallic Chemistry*, and *Organometallics*, launched in 1892, 1963, and 1982, respectively (Fig. 1*D*). We call this period the organometallic regime and its low variance (Table 1) indicates that more than ever chemists have regularized the year-to-year output of new chemical compounds. The growth rate is lower in this regime than in the previous ones, but a closer look shows that the regime can be split into two subregimes of different growth rates (Table 1): 1981–1994 had slow production, while during 1995–2015 the growth rate returned to values similar to those of the previous regimes. The first subregime started with the rise of organometallic chemistry in the early 1980s and ended with their drop around 1995, as seen in Fig. 1*D* through the decline and subsequent stagnation of reported substances in the two leading journals on the subject: *Organometallics* and *Journal of Organometallic Chemistry*. The second subregime began after 1995, when the report of bioorganic compounds took off, as evident in the surge of the journal *Bioorganic & Medicinal Chemistry Letters*, starting in 1991 as the 150th journal for reporting new substances (Fig. 1*D*), but already 8 y later surpassing the

Table 1. Statistics of chemical production of new compounds for different periods and historical regimes as modeled by Eq. 1, where μ and σ are the mean and the SD of the logarithm growth rate, respectively; P^{ShW} and P^{KS} stand for *P* values of the Shapiro–Wilk and Kolmogorov–Smirnov normality tests; and Orgmet is an abbreviation for organometallic

Regimes and periods	Period	μ, %	σ	P^{ShW}	P^{KS}
Proto-organic	Before 1861	4.04	0.4984	0.05267	0.5133
Organic	1861–1980	4.57	0.1251	0.07010	0.3391
Orgmet	1981–2015	2.96	0.0450	0.08297	0.6242
Orgmet-a	1981–1994	0.079	0.05356	0.4180	0.9040
Orgmet-b	1995–2015	4.40	0.03209	0.9770	0.9885
Pre-WW1	1861–1913	4.45	0.1229	0.5456	0.5069
WW1	1914–1918	−17.95	0.0682	0.7074	0.8721
Post-WW1a	1919–1924	18.98	0.1321	0.05726	0.3857
Post-WW1b	1925–1939	4.38	0.0487	0.05098	0.6151
WW2	1940–1945	−6.00	0.0745	0.9534	0.9666
Post-WW2a	1946–1959	12.11	0.0826	0.171	0.5765
Post-WW2b	1960–1979	4.25	0.1217	0.4731	0.7991

Table 2. Top 10 substrates over the history of chemistry

Top	Before 1860	1860–1879	1880–1889	1900–1919	1920–1939	1940–1959	1960–1979	1980–1999	2000–2015
1	H_2O	H_2O	HCl	EtOH	EtOH	Ac_2O	Ac_2O	Ac_2O	Ac_2O
2	NH_3	HCl	EtOH	HCl	AcOH	EtOH	MeOH	MeOH	MeOH
3	HNO_3	EtOH	H_2O	AcOH	HCl	AcOH	CH_2N_2	MeI	H_2O
4	HCl	H_2SO_4	AcOH	H_2O	Ac_2O	H_2O	MeI	CH_2N_2	MeI
5	H_2SO_4	HNO_3	Ac_2O	Ac_2O	H_2O	MeOH	CH_2O	CH_2O	PhCHO
6	EtOH	Br_2	H_2SO_4	H_2SO_4	Et_2O	HCl	EtOH	PhCHO	CH_2O
7	Cl_2	AcOH	MeI	MeI	H_2SO_4	C_6H_6	Morph	CuO	CO
8	Na_2CO_3	NH_3	$PhNH_2$	Et_2O	C_6H_6	Et_2O	$PhNH_2$	EtOH	TFA
9	KOH	$PhNH_2$	HNO_3	$PhNH_2$	MeOH	CH_2O	DMA	BzCl	PhAcet
10	I_2	MeI	Br_2	MeOH	Br_2	MeI	PhCHO	CO	BnBr

Abbreviations are defined in *Materials and Methods*.

Journal of Organic Chemistry that had been at the top for more than 30 y.

Despite the more certain year-to-year number of new compounds reported, an open question is whether chemists have focused or not on particular compositions. We found that chemists report not only more and more substances but also more compositions (Fig. 1*A*, black and red plots, respectively). However, they have preferences for some few of them, as found in the drop of year-to-year distance among compositions (*Materials and Methods* and *SI Appendix*, Fig. S3) and in the historical concentration of their distributions (Fig. 1*B*, box plots). In *SI Appendix*, Tables S2–S4 show the preferred compositions over time and the most popular one of each decade is found in Fig. 1*B*, where CHNO became the most explored since 1890.

Effects of Wars upon Exploration of the Space

As expected, the two World Wars (WW) led to dramatic reduction in the numbers of new chemical substances reported, as can be seen from the two dips in Fig. 1*A*. By taking the lowest war time number of new compounds, we found that WW1 sent chemistry back by 37 y and WW2 by 16 y. WW1 also caused a drop in the rate of chemical production three times more dramatic than the rate in WW2 (Table 1). The effect of WW1 was so devastating because chemical industry and research had been concentrated around Germany in pre-WW1 times (10). In fact, WW1 led to the rapid strengthening of chemistry in the United States and other non-German countries (10), which is possibly the reason that the decay of production during WW2 was not as dramatic.

After each war, however, there was a period of rapid recovery and then another one of slower surge, until production was back to prewar rates (Table 1). For post-WW1 (1919–1939), the rapid period corresponds to 1919–1924, with $\mu = 18.98\%$, the highest value in the entire history of chemistry, which somehow compensated the rapid decay preceding it. The rapid period after WW2 had only $\mu = 12.11\%$, but lasted almost three times as long (1946–1959). Some mechanisms behind these postwar stages have been suggested in ref. 3.

In summary, wars had a temporary effect on chemical production, but eventually chemistry returned to its prewar growth rates (Table 1). Such a phenomenon had already been observed for the growth of physical abstracts (11). But the wars also caused a shift of chemical research. During WW1, the importance of As, Sb, and Bi compounds, among others, increased, while that of Al, Ga, In, and Tl decreased (*SI Appendix*, Figs. S5–S8). N and alkali metals dropped during WW2 but S, B, P, and Si benefited. The interest in As compounds is presumably explained by the different arsenic warfare agents developed during WW1 (12). Phosphorus compounds started to be more often reported after WW2 when the important biological role of the element was established, as well as the use of its compounds in daily-life applications and as novel insecticides and other industrial materials (13).

Exploring Through Synthesis

Chemical substances can be extracted from plants, animals, minerals, or other sources, or synthesized, or both. Traditionally, it is claimed that organic synthesis began after Wöhler's famous synthesis of urea in 1828 (7, 14). To check this and, more generally, to assess the role of synthesis in chemical research (3, 15), we adopted the following rule. If $A + B \to C + D$ is the reaction where A and/or C is for the first time reported, we say that A is extracted and C is synthesized. The resulting annual ratio of synthesized compounds to all reported ones (Fig. 1*A*, blue) shows that, over history, more than half of the new substances have come from synthesis, except for the first 4 y of the 19th century where the percentage was slightly lower than 50% (7). In particular, already at the time of Wöhler's synthesis, new substances containing C, H, N, and O were about 50%, so organic synthesis

Table 3. Top 10 targets over the history of chemistry

Top	Before 1860	1860–1879	1880–1889	1900–1919	1920–1939	1940–1959	1960–1979	1980–1999	2000–2015
1	I_2	MAC	MAC	BZA	MAC	MAC	H_2S	TPPO	Glc
2	OA	BZA	BZA	CO_2	BZA	CO_2	TBTC	CO_2	CuO
3	CO_2	I_2	PhA	MAC	CO_2	BZA	TBTB	BZA	ZnO
4	Hg	OA	NH_3	NH_3	I_2	HCl	Ag	PhCHO	NiO
5	Cl_2	AcOH	OA	I_2	OA	Acetone	FC	Ph_2CO	CO_2
6	H_2	DHBZA	H_2O	OA	AcOH	CH_2O	$B(OH)_3$	PhAc	CoO
7	MAC	NH_3	CO_2	$PhNH_2$	NH_3	MeCHO	Ag_2S	NPhOH	MBPh
8	BZA	PhA	$PhNH_2$	PhCHO	HCl	AcOH	H_2O	Ph_2S_2	BZA
9	HgO	H_2S	I_2	H_2O	MeCHO	H_2O	TMTC	EDBB	Pd
10	NH_3	HCl	AcOH	AcOH	CH_2O	H_2S	UF_6	CuO	Ph_2

Abbreviations are defined in *Materials and Methods*.

was already well established before that (Fig. 1*A*). The percentage of synthesized compounds increased to 90% at the turn of the 20th century and has remained at such levels ever since.

Historical Trends for Substrates and Products in the Exploration of the Chemical Space

We now take a closer look at the distribution of substrates and products in chemical reactions, to address such questions as whether chemists prefer to work with those substrates that they know well, and whether those preferred substrates change over time, and analogously for products. We might expect a higher diversity of products than of substrates; that is, chemists might typically try to recombine known substrates or combine known ones with rarer ones, to produce a variety of products. Alternatively, they might search for even more efficient reactions to produce a limited number of targets.

The Workings of Selection of Substrates. By counting how often substrates are used in reactions, we obtained Fig. 2*A*. The distributions of Fig. 2*A* are heavy tailed, indicating that chemists have historically preferred some particular compounds as starting materials and that most chemicals are used as substrates only once (details of these preferences are given below).

In the analysis of the organic chemistry space a similar distribution was obtained and it was claimed that the distribution is of power-law type (4). Our analysis rejects the power-law hypothesis, which indicates that the mechanism underlying the selection of substrates cannot be modeled by multiplicative or Yule processes only or any other mechanisms generating power laws (16). Other distributions were likewise statistically rejected (*SI Appendix*, Tables S6 and S7). Therefore, it seems that the use of substrates results from a more complex process, whose study is beyond the scope of this paper.

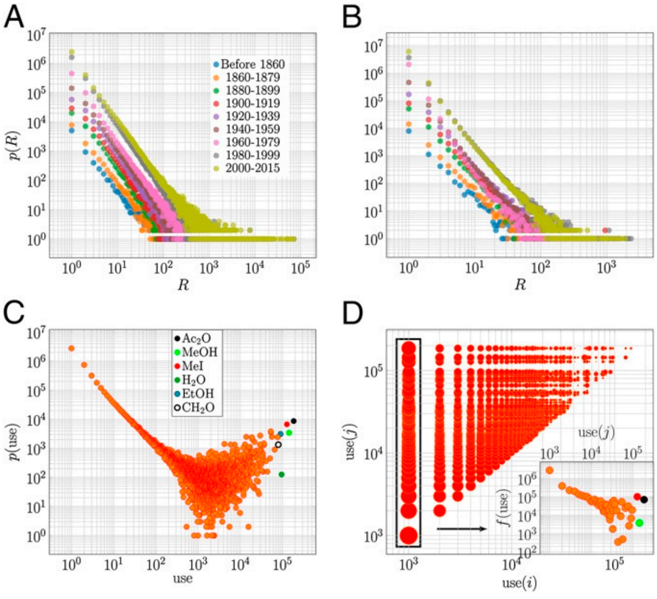

Fig. 2. Use and production of compounds. (*A* and *B*) Frequency distributions of participation of compounds in *R* different reactions as (*A*) substrates and (*B*) products. The left-hand side of the distributions corresponds to the many compounds appearing in few reactions, whereas the right-hand side corresponds to the few compounds appearing in many reactions. (*C*) Frequency distribution of uses of substrates that have participated in a one-substrate reaction. The following frequently used substrates are pinpointed: acetic anhydride (Ac_2O), methyl iodide (MeI), methanol (MeOH), ethanol (EtOH), water (H_2O), and formaldehyde (CH_2O). (*D*) Distribution of uses of substrates *i* and *j* that have participated in a two-substrate reaction. The size of each point is proportional to the frequency of use of the pair $\{i, j\}$ in reactions. (*D*, *Inset*) Frequency distribution of use of *j* in two-substrate reactions whose $use(i) \leq 10^3$ is fixed, where some frequently used substrates are shown (*C*).

Fig. 2*A* shows two jumps in the use of substrates, between 1860–1879 and 1880–1899 and between 1960–1979 and 1980–1999 (*SI Appendix*, Fig. S9). The first one marks the transition to a prolific period of production of new compounds from about 1800 to 1990, shown in Fig. 1*A* (black) as one of the regions where the number of new compounds is systematically higher than the long-term trend suggests. This transition goes hand in hand with a less focused period on particular compositions, observed as a stagnation of the distance among compositions (Fig. 1*B*).

The second jump coincides almost with the second prolific period of new compounds (Fig. 1*A*, black) and marks the transition from a period where compounds are concentrated on few compositions (1960–1979) to an even more concentrated period (1980–1999) (Fig. 1*B*).

The 10 most used substrates, i.e., max *R* in Fig. 2*A*, are shown in Table 2. The importance of strong acids and bases is notorious at the beginning of the 19th century, which gradually winds down, giving place to more organic chemistry-oriented substrates. A remarkable substrate is acetic anhydride; predicted theoretically in 1851, synthesized the next year (7), becoming the fifth most used substrate by 1880–1889 and the top substrate since 1940–1959, it is mainly used in acetylation reactions (17). Methyl iodide, which often appears among the top 10, is an important methylating agent (18).

To analyze the patterns in the use of substrates in more detail, we looked at reactions with one to three substrates, which account for 87.4% of the reactions and involve 6,081,963 compounds. One-third of the reactions report a single substrate, while half of them (48.7%) report two, and only 5.7% report three substrates.

When studying how often substrates that have participated in one-substrate reactions are used in other reactions, we see a distribution (Fig. 2*C*), where most of the reactions (about 87%) follow a log-log decay, but the tail has a smile shape (initial drop in frequency followed by its final surge). About half of the reactions have a substrate that has been used only once. At about 30 uses the remaining 13% of the one-substrate reactions spread in the mentioned smile fashion over higher uses. It is in this region, especially in its right uppermost part that some few extremely used substrates are observed (Fig. 1*C*).

Two-substrate reactions account for about half of the reported reactions (Fig. 2*D*), 94.3% of which combine a little-used substrate (participating in no more than 1,000 reactions) with another one of a wider range of uses (points in Fig. 2*D* with $use(i) \leq 10^3$ fixed). The frequency of such a combination decreases monotonically with the use of the latter, except for some particular substrates that chemists often prefer to use (Fig. 2*D*, *Inset*), e.g., acetic anhydride, methanol, and methyl iodide. These recurrent substrates are part of the chemical toolkit for synthesis and for unveiling the chemical properties of new substances. Thus, chemists like to let less-explored substances react with better-known substrates. We call this the fixed-substrate approach to the exploration of the chemical space.

For three-substrate reactions we found that reactions including a substrate of low use (less than 1,000 uses) and two substrates of any use account for 82.7% of the three-substrate reactions. Among those, reactions that have two substrates of low use account for only 36.5%. Thus, as for the two-substrate reactions, frequent substrates of these reactions are part of the chemical toolkit aforementioned.

Producing Compounds of the Chemical Space. We now turn to the products of reactions. By counting how often products are obtained in reactions, we plotted Fig. 2*B*. As for substrates, the distributions are heavy tailed, indicating that chemists have often synthesized some few products, while the majority of products are synthesized only once. As for substrates, the distributions of Fig. 2*B* do not follow a power law (*SI Appendix*, Tables S6 and S8), and they likely result from a complex process not explored in this paper.

The distributions of the number of syntheses for products overlap for the periods 1980–1999 and 2000–2015 (Fig. 2*B* and *SI Appendix*, Table S9), despite the exponential growth of new compounds. The stagnation in the report of new products in 2000–2015 was compensated by a higher fraction of new products than in 1980–1999. This is confirmed in Fig. 1*A*, where for 1980–1999 the percentage of synthetic products averaged 91.1%, which rose to 93.5% for 2000–2015. However, Fig. 1*B* indicates that these compounds were mainly of known compositions.

Fig. 2*B* shows two jumps, one between 1860–1879 and 1880–1899 and the other between 1940–1959 and 1960–1979. They indicate that products in those transitions were more often obtained than the average trend (*SI Appendix*, Fig. S10). The first jump coincides with the discussed first jump for substrates. The second one marks the transition from the WW2 recovery to the prolific period in production of new compounds, from about 1960 to 2000 (Fig. 1*A*, black), and coincides with a surge in the distance among compositions explored (Fig. 1*B*).

The 10 most synthesized compounds over history are shown in *SI Appendix*, Table S5. To actually assess whether chemists have synthetic targets, we looked at the numbers of products in their reports. A total of 81.6% of the reactions report no more than two products, and, in fact, 74.2% report only a single product. It seems that chemists aim at synthesizing complex compounds, which are typically the heaviest products of reactions; i.e., the probability of picking up the right target then is 74.2%. The distribution of appearances of targets in reactions is shown in *SI Appendix*, Fig. S11 and it follows the same trend as in Fig. 2*B* (*SI Appendix*, Tables S6 and S10).

The most synthesized targets per period are shown in Table 3. Organic acids such as oxalic and benzoic ones occur frequently; they are often used as synthetic intermediates (19). Hydrogen sulfide and uranium hexafluoride are examples of targets motivated by nuclear research in the post-WW2 period (20, 21). Likewise, organotin compounds and ferrocene evidenced the interest in organometallics, especially in materials science and homogeneous catalysis (22). We also draw attention to the presence of metallic oxides in the last period; they are often used in the synthesis of catalysts and nanomaterials (23).

Conclusion

We have found that in the exploration of the chemical space from 1800 to 2015, chemists have reported new compounds at an exponential rate. The year-to-year variability of the report of new compounds has two historical drastic reductions that distinguish three production regimes: proto-organic, organic, and organometallic. The proto-organic regime lasted until about 1860 and is characterized by the highest variance in the report of new extracted and synthesized C- and H-based compounds and by the highest use of typical inorganic substrates. The organic regime is witnessed by a sharp decrease in variance of the year-to-year output of mostly new C compounds. It began about 1860 when valence and structural theory were incorporated as a guide to explore the chemical space. The ensuing regime started about 1980 with the second drop of variance in the report of new compounds. It is a period where C-metal compounds increased, which were by about 1995 surpassed by the rapid production of compounds of biological interest.

In Rescher's (24) classification scheme, patterns of scientific growth with annual production rates (r) are identified as first-rate topics, when the growth is linear, and as very important ($r \sim 1.25\%$), important ($r \sim 2.5\%$), significant ($r \sim 3.75\%$), and routine research topics ($r > 5\%$), when the growth is exponential. r is related to the logarithm growth rate μ used here by $r = e^{\mu} - 1$; i.e., essentially $r \sim \mu$. Using our estimates for the growth rate combined with the results of the variance of the number of compounds reported for each regime, Rescher's scheme would suggest that the proto-organic regime was close to "significant" and the most uncertain of the regimes in chemistry in terms of year-to-year production, while the organic regime was roughly "routine research" with a much more certain character. A closer look at the organometallic regime shows that its first decade was "very important" and that subsequent years have been close to "routine research."

As expected, we see that production of compounds has been heavily affected by two external events: the World Wars, which temporarily reduced the production. However, after each war, chemistry recovered from these setbacks and returned to its long-term growth curve of about 4.4% annual growth. A similar trend also applies to scientific events such as the introduction of structural theory and the rise of organometallic chemistry, which marked the transition of regimes. At the transition, growth rates were somewhat perturbed, but again, chemistry quickly returned to the historical growth trend of 4.4%. This leads to the question of why chemistry maintains such a stable growth rate of 4.4% across different regimes despite major external perturbations. We speculate that this derives from the intrinsic structure of the underlying network of chemical reactions, and we are devising formal models to analyze this.

We found that the exploration of the space has been ruled since the early 19th century by synthesis, that is, even before Wöhler's synthesis of urea in 1828, which is traditionally considered the beginning of organic synthesis. Nevertheless, for a long period, extraction was similarly important to synthesis, and the latter became the established tool to report new compounds only around 1900, i.e., 70 y after Wöhler's synthesis and 40 y after the introduction of the structural theory. This time lag for a systematic shift in the practice of chemistry is remarkable.

In terms of the use of substrates and the production of compounds, chemists have been conservative in the selection of their starting materials, presumably as a disciplinary consequence of starting from substances that are readily available or as a way to develop valid and reliable expert intuition to explore the chemical space (2). Perhaps, as Mendeleev pointed out, "conservatism in science is completely inevitable, because science in essence is a legacy, unthinkable except as the wisdom of centuries past, and thus cannot be passed on without conservatism" (ref. 25, p. 146). Nevertheless, chemists have succeeded in synthesizing new compounds in an exponential manner. The exploration of chemical space, however, seems to have been rather uneven, with only a handful of compositions extensively explored, CHNO being foremost among them since 1890.

The set of explored combinations is narrow in the sense that a fixed-substrate approach is preferred. In fact, reported reactions typically include two substrates: one less known and the other part of the synthetic toolkit of preferred substrates, acetic anhydride leading since 1940.

The work reported here constitutes a computational, entirely data-driven approach to the history of chemistry. Using chemical compounds and reactions as the only data source goes to the very heart of chemistry as a science (26). This is in contrast to text-based approaches to the history of science (27) that focus on themes and topics of the scientific discourse.

Materials and Methods

Data. Up to January 2017 Reaxys reported 42,782,394 chemical reactions and 20,669,217 associated substances, with the first entry dating back to 1771. Given that single-step reactions are often contained in multistep reactions and that there are few entries for the 18th century, and to avoid those of the later months of 2016, which were still under curation and annotation, we restricted our analysis to the period from 1800 to 2015. Moreover, as we found that 70% of the new compounds over the history have been reported in journals, rather than in patents, we analyzed the former. Hence, 16,356,012 reactions and 14,341,955 compounds were the basis for the present study. By compounds we mean the reported chemical species, e.g., bulk substances, molecular clusters, van der Waals complexes, etc. (1).

Variance Analysis of the Number of New Compounds. For a continuous t, the exponential growth of new compounds is described by $s_t = s_0 e^{kt}$, from which $r = e^k - 1$. In a discrete case, data variance is an issue; therefore the normality of r values has to be tested. In this setting, $r_t = (s_{t+1}/s_t) - 1$, which for the chemical reactions data failed normality tests for 1804–2015 (*SI Appendix*, Fig. S1*A* and Table S1) and in particular for the period 1804–1860 that shows high fluctuation (*SI Appendix*, Table S1). Therefore we explored the distribution of $Y_t := \ln s_{t+1} - \ln s_t$ (*SI Appendix*, Fig. S1*B*), which we found to be normally distributed for three periods, but not for the whole period 1804–2015 (*SI Appendix*, Table S1). We observed that the variance is time dependent and Y_t follows an autoregressive conditional heteroskedasticity (ARCH) model. In the simplest form, Y_t is a heteroskedasticity model consisting of three periods with the general form $Y_t = \mu_i + \sigma_i Z_t$, i being an index for the respective period and the residuals $\{Z_t\}_{t\geq 1}$ a source of identically distributed Gaussian noises. This is equivalent to

$$s_{t+1} = s_t e^{\mu_i + \sigma_i Z_t}. \quad [1]$$

The values of μ_i and σ_i for the three periods are shown in Table 1. In all cases the residual $\{Z_t\}_{t_0}^{t_f}$ passes the Shapiro–Wilk and the Kolmogorov–Smirnov normality tests $\mathcal{N}(0, 1)$ (*SI Appendix*, Table S1), for t_0 and t_f the respective initial and final year of period i. The combination of all three subsequences of residuals passes the Kolmogorov–Smirnov normality test, but fails the Shapiro–Wilk one (*SI Appendix*, Table S1 and Fig. S2).

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

$$s_t = 51.85\, e^{0.04324(t-1800)}, \quad [2]$$

where t is a year between 1800 and 2015.

Compositions and Their Distance. The composition of a compound is its list of chemical elements arranged in lexicographic order, e.g., HOS for H_2SO_4. The distance among compositions of successive years is calculated as the Manhattan distance of the relative frequency of compositions for the 2 y (*SI Appendix*, Fig. S3).

Abbreviations for Compounds of Tables 2 and 3. Abbreviations shown in Tables 2 and 3 are defined in parentheses as follows: AcOH (acetic acid), Ac_2O (acetic anhydride), Ag (silver), Ag_2S (silver sulfide), BnBr (benzyl bromide), $B(OH)_3$ (boric acid), Br_2 (bromine), BZA (benzoic acid), BzCl (benzoyl chloride), C_6H_6 (benzene), CH_2N_2 (diazomethane), CH_2O (formaldehyde), Cl_2 (chlorine), CO (carbon monoxide), CO_2 (carbon dioxide), CoO (cobalt(II) oxide), CuO (copper(II) oxide), DHBZA (3,4-dihydroxybenzoic acid), DMA (dimethylamine), EDBB (1,1′-(1,2-ethanediyl)bisbenzene), Et_2O (diethyl ether), EtOH (ethanol), FC (ferrocene), Glc (glucose), H_2 (hydrogen), H_2O (water), HCl (hydrochloric acid), Hg (mercury), HgO (mercury(II) oxide), HNO_3 (nitric acid), H_2S (hydrogen sulfide), H_2SO_4 (sulfuric acid), I_2 (iodine), KOH (potassium hydroxide), MAC (methylammonium carbonate), MBPh (4-methoxylbiphenyl), MeCHO (acetaldehyde), MeI (methyl iodide), MeOH (methanol), Morph (morpholine), Na_2CO_3 (sodium carbonate), NH_3 (ammonia), NiO (nickel(II) oxide), NPhOH (4-nitro-phenol), OA (oxalic acid), Ph_2 (biphenyl), PhA (phthalic acid), PhAc (acetophenone), PhAcet (phenyl acetylene), PhCHO (benzaldehyde), Ph_2CO (benzophenone), $PhNH_2$ (aniline), Ph_2S_2 (diphenyl disulfide), TBTB (tributyltin bromide), TBTC (tributyltin chloride), TFA (trifluoroacetic acid), TMTC (trimethyltin chloride), TPPO (triphenylphosphine oxide), UF_6 (uranium hexafluoride), and ZnO (zinc oxide).

ACKNOWLEDGMENTS. We thank RELX Intellectual Properties SA for the reaction dataset; and Joachim Schummer, Rainer Brüggemann, and Jeffrey Seeman for their critical comments. E.J.L. was supported by the Fundación para la Promoción de la Investigación y la Tecnología from Banco de la República (Colombia), Project 201828. W.L. was supported by a PhD scholarship from the German Academic Exchange Service: Forschungsstipendien-Promotionen in Deutschland, 2017/2018 (Bewerbung 57299294). G.R.'s research was partly supported by the Alexander von Humboldt Foundation.

1. J. van Brakel, *Substances: The Ontology of Chemistry* (North-Holland-Elsevier, 2012), pp. 171–209.
2. G. M. Keserü, T. Soos, C. O. Kappe, Anthropogenic reaction parameters - the missing link between chemical intuition and the available chemical space. *Chem. Soc. Rev.* **43**, 5387–5399 (2014).
3. J. Schummer, Scientometric studies on chemistry I: The exponential growth of chemical substances, 1800–1995. *Scientometrics* **39**, 107–123 (1997).
4. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell, B. A. Grzybowski, Architecture and evolution of organic chemistry. *Angew. Chem. Int. Ed. Eng.* **44**, 7263–7269 (2005).
5. A. J. Lawson, J. Swienty-Busch, T. Géoui, D. Evans, *The Making of Reaxys - Towards Unobstructed Access to Relevant Chemistry Information* (American Chemical Society, 2014), chap. 8, pp. 127–148.
6. J. N. K. Rao, On the estimation of heteroscedastic variances. *Biometrics* **29**, 11–24 (1973).
7. J. R. Partington, *A History of Chemistry* (Macmillan, 1964).
8. A. Rocke, What did "theory" mean to nineteenth-century chemists? *Found. Chem.* **15**, 145–156 (2013).
9. U. Klein, *Experiments, Models, Paper Tools* (Stanford University Press, 2002).
10. R. M. Friedman, *The Politics of Excellence* (Times Books, 2001).
11. D. J. de Solla Price, *Little Science, Big Science* (Columbia University Press, 1963).
12. B. Radke, L. Jewell, S. Piketh, J. Namieśnik, Arsenic-based warfare agents: Production, use, and destruction. *Crit. Rev. Environ. Sci. Technol.* **44**, 1525–1576 (2014).
13. D. E. C. Corbridge, *Phosphorus: Chemistry, Biochemistry and Technology* (CRC Press, 2013).
14. K. C. Nicolaou, The emergence of the structure of the molecule and the art of its synthesis. *Angew. Chem. Int. Ed. Eng.* **52**, 131–146 (2013).
15. K. C. Nicolaou, The emergence and evolution of organic synthesis and why it is important to sustain it as an advancing art and science for its own sake. *Isr. J. Chem.* **58**, 104–113 (2018).
16. M. E. J. Newman, Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**, 323–351 (2005).
17. H. Held, A. Rengstl, D. Mayer, *Acetic Anhydride and Mixed Fatty Acid Anhydrides* (American Cancer Society, 2000).
18. P. A. Lyday, T. Kaiho, *Iodine and Iodine Compounds* (American Cancer Society, 2015), pp. 1–13.
19. W. Riemenschneider, M. Tanifuji, *Oxalic Acid* (American Cancer Society, 2011).
20. A. M. Rozen, The first plant in the world for the production of heavy water by the method of two-temperature water-hydrogen sulfide isotopic exchange. *At. Energy* **78**, 218–223 (1995).
21. T. A. O'Donnell, D. F. Stewart, P. Wilson, Reactivity of transition metal fluorides. II. Uranium hexafluoride. *Inorg. Chem.* **5**, 1438–1441 (1966).
22. A. Togni, R. L. Halterman, *Metallocenes: Synthesis Reactivity Applications* (Wiley, 2008).
23. J. Zhang, H. W. Richardson, *Copper Compounds* (American Cancer Society, 2016), pp. 1–31.
24. N. Rescher, *Scientific Progress* (Basil Blackwell, 1978).
25. M. D. Gordin, *A Well-Ordered Thing: Dmitrii Mendeleev and the Shadow of the Periodic Table* (Basic Books, 2004).
26. J. Schummer, The chemical core of chemistry I: A conceptual approach. *Hyle* **4**, 129–162 (1998).
27. M. D. Laubichler, J. Maienschein, J. Renn, Computational perspectives in the history of science: To the memory of Peter Damerow. *Isis* **104**, 119–130 (2013).