Angewandte Chemie

GDCh

Eine Zeitschrift der Gesellschaft Deutscher Chemiker

www.angewandte.de

Titel: Is organic chemistry really growing exponentially?

Autoren: Sara Szymkuc, Tomasz Badowski, and Bartosz A. Grzybowski

Zitierweise: *Angew. Chem. Int. Ed.* 10.1002/anie.202111540

Link zur VoR: https://doi.org/10.1002/anie.202111540

WILEY-VCH

Angewandte Chemie
Angewandte Communications

10.1002/ange.202111540

Chemoinformatics

DOI: 10.1002/anie.201((will be filled in by the editorial staff))

Is organic chemistry really growing exponentially?**

Sara Szymkuć[+] *,Tomasz Badowski*[+] *Bartosz A. Grzybowski**

Abstract. *In terms of molecules and specific reaction examples reported in the literature, organic chemistry features an impressive, exponential growth. However, new reaction classes/types that fuel this growth are being discovered at a much slower and only linear (or even sublinear) rate. In effect, the proportion of newly discovered reaction types to all reactions being performed keeps decreasing, suggesting that synthetic chemistry becomes more reliant on reusing the well-known methods. On the brighter side, the newly discovered chemistries are more complex than decades ago, and allow for the rapid construction of complex scaffolds in fewer numbers of steps. In this paper, we study these and other trends in the function of time, reaction-type "popularity" and complexity based on the algorithm that extracts generalized reaction class templates. These analyses are useful in the context of computer-assisted synthesis, machine learning (to estimate the numbers of models with sufficient reaction statistics), and also for identifying erroneous entries in reaction databases.*

Nearly two decades ago we published the first paper[1a] analyzing the development of organic chemistry from a network perspective – as we showed, the literature-reported reactions form a giant, scale-free network ("Network of Organic Chemistry," NOC) that features relatively few but highly connected "hub" molecules and evolves according to statistical laws that have not changed from the times of Wöhler all the way to the present. Later, these trends were confirmed and extended by us[1b,c] and others[1d,e] to quantify other topological properties of the network and to relate its evolution to various economic aspects. On one hand, this early work was the cornerstone of our subsequent work on network-search algorithms[2] and computational synthesis planning (Chematica/Synthia), recently culminating in automated design and experimental execution of syntheses leading to medicinally-relevant targets[3a] as well as complex natural products[3b]. On the other hand, one aspect of the 2005 paper[1a] caused some rather unexpected outcomes and interpretations. Namely, we showed therein that the numbers of known molecules and connections within the network (i.e., reactions making these molecules) grow exponentially with time – this result has since been interpreted consistently but overoptimistically as the *knowledge base* of organic chemistry expanding equally rapidly (in particular, so rapidly as that only data-driven AI systems can keep up with the pace; see ref. [4] and footnote [5]). This interpretation, however, entails a key fallacy in that specific reactions reported in literature are not necessarily new reaction *types* but rather repeating manifestations of the same reaction classes (e.g., thousands of esterifications, amide bond formations, or Suzuki couplings reported each year for different types of substrates). Here, we show that when the reactions reported in the NOC or in the patent literature[6] are grouped according to commonly recognized types/classes[7] – reflecting similar properties of different functional groups and generally sharing common reaction mechanisms – the growth of our discipline is much less dynamic and, over the past few decades, has become linear or even sublinear. Only few thousands of new reaction types are reported each year and their proportion, compared to all reported reactions, keeps decreasing. On a more positive note, within these newly discovered reaction types there is a growing fraction of complex and multicomponent transformations that can generate complex scaffolds in fewer steps and offer improved atom and "pot-efficiencies"[8]. These results can be considered from several perspectives. More narrowly, in the context of computer-aided synthesis planning, the limited rate at which chemistry expands means that expert coding of reaction rules *is*, at least for the foreseeable future, scalable and can be kept *au courant*. Our analyses also provide clues which of the reported reactions are reliable and which ones are likely erroneous and should not be taught to the machines (and also be removed from reaction repositories). More broadly, one may take these results as a challenge and ponder whether it is possible to break away from the nearly constant reaction discovery rate. Does this constancy – and, in fact, relative decrease with respect to all performed reactions – reflect chemistry gradually exhausting the possibilities in which atoms of relatively few types (Csp3, Csp2, O, etc.) can form/break bonds? Or can the discovery process be revitalized and accelerated by new technologies such as modern robotized systems capable of analyzing multicomponent mixtures and looking for unexpected reactivity patterns[9]?

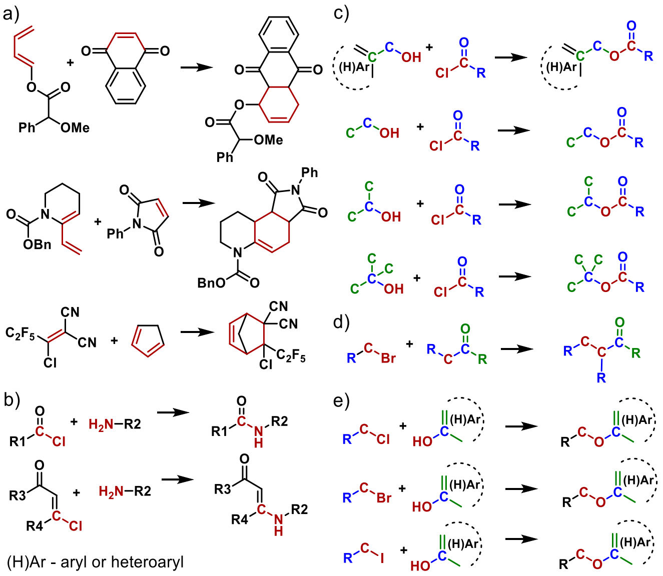


Figure 1. Reaction cores, core environments, and the challenge of automatically classifying reactions by type. Core atoms changing bonding patterns during the reaction are colored in *red*. One-bond

---

[*] Dr. Sara Szymkuć+, Dr. Tomasz Badowski+, Prof. B.A. Grzybowski
Institute of Organic Chemistry, Polish Academy of Sciences
Ul. Kasprzaka 44/52, 01-224 Warsaw, Poland &
Allchemy, Inc., Highland, IN, USA &
Prof. B.A. Grzybowski
IBS Center for Soft and Living Matter and
Department of Chemistry, UNIST,
50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, South Korea
E-mail: nanogrzybowski@gmail.com

[+] Authors contributed equally.

10.1002/ange.202111540

environments are, when considered, colored in *blue*. Environments two bonds away from the core are colored in *green*. **a)** Diels-Alder cycloaddition is uniquely and properly defined as a reaction type by the core atoms alone. Including various substituents on the diene or the dienophile would only serve to artificially increase the number of reaction templates. In contrast, in **b)**, the core itself cannot distinguish between, e.g., N-acylation and substitution of vinyl chlorides via addition-elimination. In **c)**, templates spanning environments two bonds away from the core are too wide and unnecessarily multiply reactions of the same type (here, acylation of alcohols). **d)** On the other hand, taking just one-atom environments may be too narrow. Not taking into account carbonyl/carboxyl functionality would produce a chemically nonsensical reaction template. This alkylation reaction can proceed because the EWG group helps increase the acidity of alpha protons. **e)** Irrespective of the width of the "environment," one must often consider the identity of specific functional groups/atoms. In the alkylation of phenols shown here, Cl, Br, I act merely as leaving groups and should be combined into one template. Hydrogens were omitted for clarity.

The cornerstone of our work is to automatically categorize literature-reported reactions into distinct types. To do so, it is necessary to extract parts of the molecules that define these types. In some cases, it is sufficient to consider only the "reaction core", i.e., atoms that change their bonding patterns and formal charges. For instance, reactions in **Figure 1a** can all be uniquely assigned as Diels-Alder cycloadditions based on the reaction cores (colored in *red*) spanning only the carbon atoms of the diene and the dienophile and not any substituents. In general, however, the "reaction cores" are too narrow – for instance, they are identical for and not able to distinguish between, e.g., N-acylation and substitution of vinyl chlorides via addition-elimination (**Figure 1b**). Such examples point to the need to consider wider environments – one or two bonds away from the core (in **Figures 1c-e** colored in *blue* and *green*, respectively). In this context, environments up to two bonds away from the core overspecify the reactions – this is illustrated in **Figure 1c** where this environment would incorrectly classify acylations of primary, secondary and tertiary alcohols as different reaction types. On the other hand, one-bond environments may not reflect all relevant mechanistic details – in the example in **Figure 1d**, the fact that the neighboring atom is not only "some carbon" but a carbon within an electron-withdrawing carbonyl/carboxyl group is essential for the acidity of the proton abstracted during the alkylation reaction. Yet another problem is illustrated in **Figure 1e** whereby one-atom environments of phenol alkylation reactions include the halides – in each individual case, this inclusion is proper to describe specific reactions, but in terms of a broader reaction type, all halides act merely as leaving groups. Such mechanistically identical reactions should be grouped into one reaction type. To account for these and many other nuances, we developed multiple heuristics grounded in chemical-mechanistic knowledge and specifying which environments should be applied and which groups should be assigned as equivalent. These heuristics are summarized below:

(1) For cycloadditions such as Diels-Alder or 1,3-dipolar, the reaction core atoms are kept. However, for cycloadditions such as [2+2] De Mayo, we retain the one-atom environment since the presence of a EWG group is required (and substitution patterns without EWGs suggest alternate reaction mechanism).

(2) For other reactions, one-atom environments are taken as default but with the following exceptions:

(2a) If an alkyl or an aryl atom of the substrate is connected to a leaving group (halogens Cl, Br, I, sulfonyls, or leaving groups containing boron, e.g., boronic acids or esters, or tin, e.g., tributyltin) then only the core atoms and a leaving group are extracted as a template. This being said, we keep track of whether the alkyl carbon atom is flanked by one or two vs. three carbon substituents (to subdivide such general templates based on whether S_N2 vs. S_N1 mechanism is more likely).

(2b) Alkyl and aromatic carbons connected to "core" heteroatoms are not extracted. In general, heteroatoms connected with alkyl vs. aryl carbons share similar chemical properties and undergo reactions with the same substrates and via the same mechanism (e.g., both alcohols and phenols undergo acylation; similarly, alkyl and aryl thiols can react with the same partners).

(2c) Likewise, carbons connected by a single bond to a "core" carbon are not extracted. An exception to this rule is to include carbons within electron-withdrawing groups (EWGs: ketone, ester, amide, imine, carboxylic acid, aldehyde, nitrile~~,~~ etc.; for a full list of groups please see SI, **Section S1**) that are connected to alkyl, alkenyl or alkynyl "core" carbons. The EWGs (i) increase the acidity of adjacent, α-alkyl protons and facilitate generation of carbanions that subsequently can react with various electrophiles, or (ii) make unsaturated bonds prone to attack of a nucleophile. This heuristic also allows to distinguish between mechanistically distinct reactions like Michael addition vs. hydroamination.

(2d) If a "core" nonaromatic heteroatom has a neighbor which is, in turn, part of a double or triple bond, this multiple bond is kept. This is to distinguish between, e.g., esterification vs. etherification, or alkylation of amines vs. amides. We note that without this heuristic, hydrolysis of amides to carboxylic acids would result in an erroneous template in which, e.g., a secondary amine substrate is converted into an alcohol – such a reaction is unknown.

(2e) Aromatic nitrogen and phosphorus are extracted without neighboring atoms if these neighbors do not participate in the transformation. This heuristic prevents multiplication of mechanistically identical transformations that differ only in the structure of the heterocycle (e.g., alkylation or acylation of pyrrole and pyrazole).

(2f) If a reaction involves opening of a three-membered ring, all atoms of the ring are retained in the template. This heuristic helps to correctly delineate core atoms for reactions of epoxides, aziridines or cyclopropanes.

Two additional heuristics were applied to group mechanistically equivalent transformations:

(3) Reaction differing in only a halogen atom were all grouped as one template. Likewise, if the sole difference was potassium vs. sodium cations, these reactions were also grouped.

(4) Amides, sulfonamides, phosphoramides, and their derivatives were grouped into one category if a reaction took place on the nitrogen atom (whose reactivity is similar within these groups). Similar grouping was applied if nitrogen of thioamides, thiophosphinic amides, or their derivatives reacted.

These template extraction rules were applied to organic reactions from the NOC collection and from the patent literature. We limited our analysis to database entries in which atoms changing their bonding patterns formed a graph with only one connected component (see SI, **Section S2.5**). In this way, we did not count towards the new reaction types those examples which, in reality, were "composed" of several known reactions performed simultaneously or in a multistep fashion at different loci (e.g., multiple protections or deprotections, simultaneous reduction of both a nitro group and a double bond). For reactions occurring at one locus, we additionally checked whether they could be represented as a sequence involving some most popular reaction types (Diels-Alder, click); such reactions were removed but other types of cascades were allowed. As described previously[1a,b], both sets were pruned for duplicate reactions as well as those that were missing substrates or products. In addition, reactions were deionized (to unify, say, COO^- and COOH in products and substrates), and those reporting as products unstable intermediates (e.g., carbocations and carbanions) were removed. In the end, the filtered NOC collection comprised ~3,72 million reactions and the patent one, 907 thousand reactions. Atoms across all these reactions were mapped using either our own MAPPET[10a] and/or IBM's

10.1002/ange.202111540

softwares[10b,11], and the extracted templates were in the SMARTS notation.

The templates derived as a result of these extraction (1,2a-f) and grouping (3,4) operations provided a satisfactory categorization of chemical knowledge into commonly recognized reaction classes according to similar properties of different functional groups. Naturally, there were some corner cases not distinguished by the methods we applied: for instance, S_NAr and Buchwald coupling with an amine cannot be distinguished without inspecting reaction conditions (though such a subdivision would only serve to duplicate one reaction template). Conversely, one could argue that some reactions could be further grouped to decrease the template count. For instance, one could imagine combining all nucleophiles in S_N2 reaction into one template. We do not do so, however, because, these nucleophiles may have different chemical properties and their S_N2 reactions may require different conditions (e.g., alkylation of thiols usually proceeds without addition of base, at room temperature, and in various solvents; in contrast, alkylation of unactivated esters requires strong base to generate carbanion acting as a nucleophile, proceeds at low or very low temperatures, and solvent scope is significantly narrower, typically THF or diethyl ether). The finer granularity we apply in defining reaction types can, if anything, overcount some reaction classes – but even then, the growth of chemistry is still slower than the previously purported exponential. We also note that our templates are qualitatively different than those extracted by, e.g., RDChiral[12] which aims to extract detailed information about a specific reaction and does not classify functional groups according to similar properties (for details, see SI, **Section S5**). We are also not pre-supposing any “key” hand-coded templates as in ref. [4d].

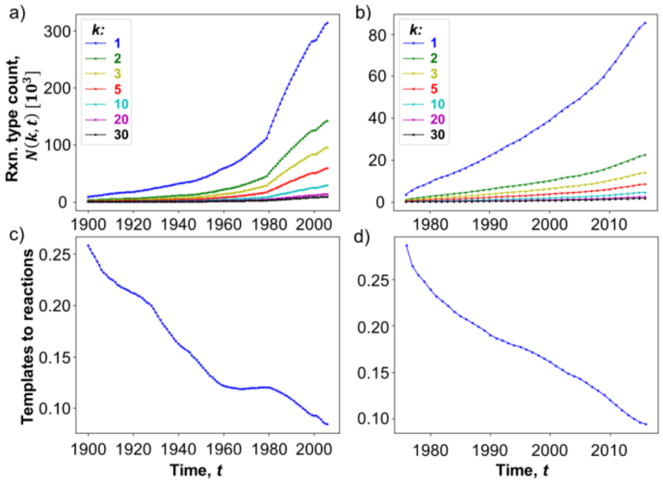


Figure 2. Discovery of reaction types in the function of time and for different popularities *k*. Each curve corresponds to different *k*, with specific values given in the legends. The trends are for **a)** NOC and **b)** patent datasets. **c)** and **d)** plot the number of reaction types divided by all reactions performed till a given year. For both NOC, c, and patents, d, this ratio decreases with time indicating the slow-down in the rate of new reaction discovery.

With these preliminaries, we extracted ca. 310,000 templates from the NOC and only ca. 85,000 from the patent literature (which is in line with the industrially relevant syntheses relying on fewer chemistries). The patent templates are available at https://github.com/badtom/expchem/blob/main/data/templates_USPTO.csv (the NOC templates are based on Reaxys’ proprietary data and cannot be publicly shared). We tracked the discovery of these reaction types as a function of time, t, and also of their synthetic usefulness/popularity measured as the minimum number of times, k, a given template has been used. The time dynamics is quantified in **Figures 2a**,**b** which plot how many templates, $N(k,t)$, were known by a given year. Different-color curves correspond to different values of k, for instance, *red* curve has the number of templates used in at least $k = 5$ reactions up to a given year. For the larger and more diverse NOC collection, the early growth was faster than linear but since ca. 1980 transitioned to linear (for higher k) or even distinctly sublinear (lower k) (see footnote [13]). The recent (1980-2006) slopes of these curves are, at most, few thousand new templates per year: ~7500/yr for $k = 1$, ~2500/yr for $k = 3$, 770/yr for $k = 10$, and only 246 for $k = 30$. Taking as practically useful reaction types that appeared $k \geq 3$ times (as in the expansion policy neural network in [4b]) translates into the modern-day corpus of ~95,000 reaction classes – interestingly, this is commensurate with the number of reaction transforms taught over the years to synthesis-planning programs such as Chematica/Synthia[2,3]. For the patents, the numbers of reaction types also increase approximately linearly with time, with slopes increasing slightly after 2009 (e.g.; ~3900/yr for $k = 1$, ~670/yr for $k = 3$, 240/yr for $k = 10$, and 99 for $k = 30$). The most pronounced increase in the slope is for $k = 1$ but, as we will discuss later, these reactions are often erroneous entries. We observe that for both the NOC and patents, the number of reaction types decreases steadily compared to all specific reactions executed (**Figures 2c,d**). We interpret this trend as yet another manifestation of the slowdown in the *discovery* of new reaction classes/methodologies relative to the *exploration* of synthesizable molecular space using already known reaction types.

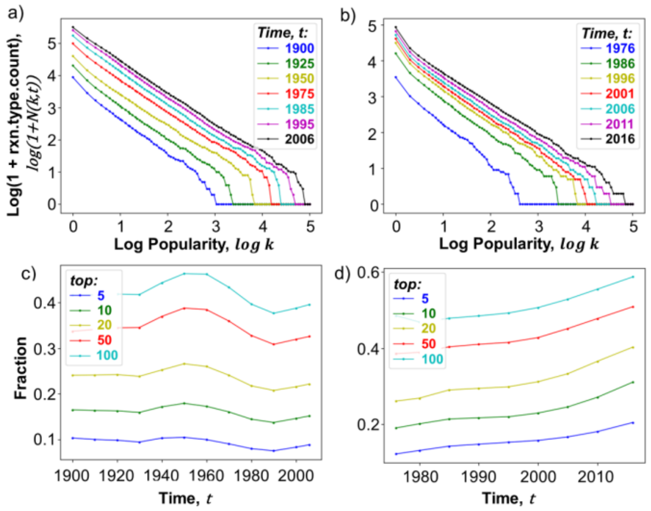


Figure 3. Distributions of reaction templates of different popularities. a,b) Doubly-logarithmic plots of template count, $N(k,t)$, as a function of popularity k. Different-color curves correspond to different years. **c,d)** Fractions of the most popular reaction types amongst all reactions for NOC (*left*) and patents (*right*) plotted as a function of time. Different curves give the fractions for the most popular 5,10, 20, 50 and 100 reaction types.

It is also interesting to consider the dependence of $N(k,t)$ on k and study the rate at which reaction classes gain popularity. In this context, the doubly logarithmic plots in **Figure 3a** and **3b** reveal that for different times, $N(k,t)$ scales with k in a power-law fashion, with exponent close to -1, $N(k,t) \propto 1/k$. It follows that the numbers of templates that become very popular is much lower than those that are less popular – in fact, $N(k+1,t) - N(k,t) \propto 1/k^2$. This means, for instance, that the number of reactions reaching, at some instance of time, popularity of $k = 50$ is roughly 100 times smaller than those that reach popularity $k = 5$. Leading the list of these most popular reactions we find for the NOC amide synthesis from carboxylic acids and amines ($k = 77,116$), followed by alkylation of alcohols or phenols with primary or secondary halides/sulphonates ($k = 70,863$), and hydrolysis of esters ($k = 64,813$); for the patents, the highest $k =$

10.1002/ange.202111540

63,434 corresponds to the said amide synthesis from carboxylic acids and amines, followed by hydrolysis of esters ($k = 38{,}088$) and Buchwald Hartwig/nucleophilic aromatic substitution with amines with $k = 32{,}694$ (for the list of the top fifty types, see SI **Section S4**). Interestingly, the fraction of the most popular reactions amongst all reactions remains roughly constant (**Figure 3c**) for the NOC but increases with time for the patents (**Figure 3d**) – the latter observation indicates that simple, "basic" transformations continue to offer robustness sought in industrial applications. In fact, for the patents, the top 100 most popular reaction types account for ca. 60% of all reactions reported in the patent collection (*light blue* line in **Figure 3d**). Also, the $N(k,t)$ vs. k dependencies allow us to estimate the number of reaction classes for which the number of specific reaction examples measured by k is large enough – say, from ~100 to, ideally, 1000 and above – to construct meaningful machine learning, ML, models[14]. For instance, based on the NOC, one could construct ML models based on at least 1,000 reaction examples for 297 reaction types; with the relaxed requirement of only 100 examples, there could be ML models for 2838 types (which is, still, only ~1% of all reaction classes).

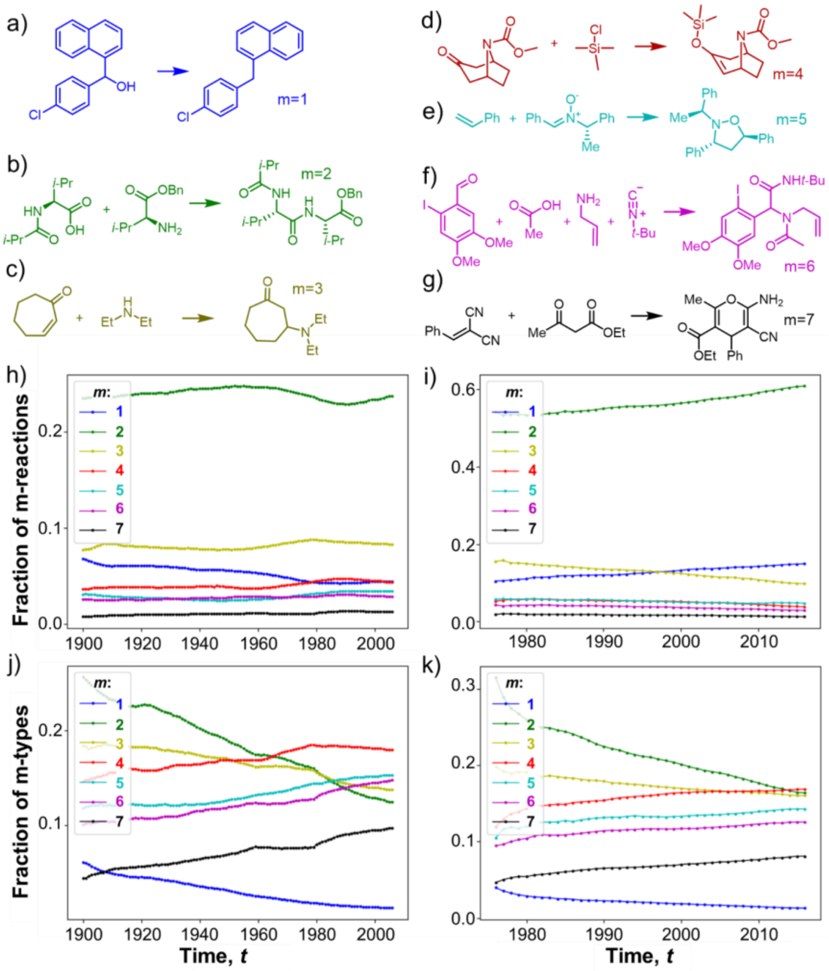


Figure 4. Examples and dynamics of reactions and reaction types characterized by different degrees of complexity. Complexity is defined here by the numbers, m, of non-H atoms that change bonding patterns compared to the substrates (according to the reaction mapping). For instance, a) Deoxygenation reaction is characterized by $m = 1$. b) Amide bond formation, $m = 2$. c) Michael addition, $m = 3$. d) Formation of silyl enol ethers, $m = 4$. e) Nitrone-olefin [3+2] cycloaddition, $m = 5$. f) Four-component Ugi reaction, $m = 6$. g) Formation of 4*H*-pyrans, $m = 7$. Plots below have fractions of reactions and reaction types characterized by different complexities m and known till a given year. The colors of the curves correspond to examples of different-complexity reactions given above (values of m are also given in the legends). h,i) Plot fractions of all reactions for, respectively NOC and patents. j,k) plot fractions of reaction *types* for NOC and patents.

Next, we consider the complexity of the reactions being discovered. As a measure of complexity, we use here the number of main product's non-hydrogen atoms m that change their bonding patterns from the substrates. For instance, $m = 1$ reactions can be deoxygenation, carbonyl to alkane reduction, or *O*-desililation; $m = 2$ corresponds to reactions such as amide bond formation or esterification, and $m = 7$ is seen in more specialized transformations such as, say, 4*H*-pyran ring formation (see examples in **Figure 4a-g**).

Figures 4h,i quantify – respectively, for the NOC and patent collections – the fractions of all reactions (not reaction types) characterized by different complexities m and reported up to a given year. These fractions remain roughly constant meaning that chemists execute both simple and complex transformations with frequencies that have remained approximately constant for the past century. In contrast, **Figures 4j,k** plot similar dependencies but for reaction *types*. The key trend here is that as chemistry progresses, newly discovered reaction classes become more complex. For instance, hundred years ago, chemistry was dominated by $m = 2$ and $m = 3$ reaction types (esterification, acylation, etherification, Michael addition, etc.). In the NOC collection (**Figure 4j**), between 1960 and 1980, a crossover took place and nowadays most newly discovered reaction classes are those characterized by $m = 4,5,6$ (e.g., 1,3-dipolar cycloaddition, Pauson-Khand reaction, Ugi reaction, Sakurai reaction, etc.). Interestingly, for the patent literature (**Figure 4k**), this cross-over is just taking place now, as $m = 2,3$ reaction types are being overtaken by $m = 4$ (but not yet by $m = 5,6$). This is yet another manifestation of patent literature relying on simpler types of chemistries.

One aspect of our analyses may have practical ramifications – namely, the k-dynamics from **Figure 3** can be a useful tool in automatic tracking of erroneous entries in reaction databases, which is important for avoiding false-positive reaction "rules" marring data-driven synthesis-design softwares. Recently, Toniato *et al*. presented an interesting deep-learning method[15] that assumed that erroneous entries are likely those that are most difficult to learn by the AI models and contain features rarely seen across the training set (though this definition also encompasses specialized but correct chemistries producing some rare scaffolds). Our analyses offer an alternative approach – namely, reaction types that retain $k = 1$ popularity for years (i.e., are never used again) are suspicious. Such reaction types may correspond to some highly specialized transformations but may also be database entry errors. In this context, we note that analysis based on the k popularity of reaction *types* is advantageous over analysis of specific reaction precedents – in fact, majority of specific reactions are nor repeated/reported multiple times and there is nothing wrong or suspicious in them retaining "unitary popularity" forever.

To verify our hypothesis, we analyzed 280 reaction templates retaining $k = 1$ for at least 5 years (140 examples from NOC and 140 from patents, 20 randomly chosen per each value of complexity $m = 1,\ldots,6$, and at least 7). Comparing them to source publications revealed that in the patent collection, ca. 60% were outright incorrect or highly suspicious (**Figure 5**); in the NOC collection, the percentage was lower but still significant, ca. 28%. With all of these examples tabulated and commented at https://github.com/badtom/expchem/blob/main/data/rare_not_new_reactions (files NOC_rare_ not_new.csv and USPTO_rare_not_new.csv), we observe that the main problems were incorrect structures of substrates or products, substrates entered as products (or vice versa), solvent or reagents given instead of substrates (especially in patents), and multistep reactions written as one transformation (unless performed one-pot, a combination of two known reactions cannot be classified as a new reaction type). Extrapolating the „error rates" to the entire NOC and patent collections leads to an estimate that some 36,000 and 29,000 k = 1 examples these repositories contain might be erroneous. Furthermore, for the patent collection, the k = 1 reactions contain unrealistically high proportion of multicomponent reactions, MCRs, some 30% of the total (compared to only 6% in the NOC). Inspection of a sample of these reactions (listed and commented on at https://github.com/badtom/expchem/blob/main/data/USPTO_rare_multicomponent_reactions.csv), revealed that 86% were erroneous and 94% were incorrectly classified as multicomponent reactions in which solvent, base or reagent was entered instead of substrate. Assuming similar rates hold across the entire patent collection, we estimate that the combined k = 1 @ 5 yr and MCR sets (not double-counting the conjunction of these two sets) contain ca. 36,000 erroneous entries – that is, some 40% of the entire patent collection. This helps us understand why, for instance, AI-based synthesis-design programs trained on USPTO often give chemically nonsensical suggestions[7] – these suggestions reflect the poor quality of the underlying reaction set from which the reaction templates/rules are automatically extracted. We advocate that at least for the USPTO set, k = 1 MCRs could all be removed automatically, as the likelihood of them being erroneous is very high; from the remaining subset, reactions retaining k = 1 popularity for 5 years should be extracted and carefully inspected by expert chemists (which, given the few tens of thousands of such entries, would be a tedious but realistic undertaking).

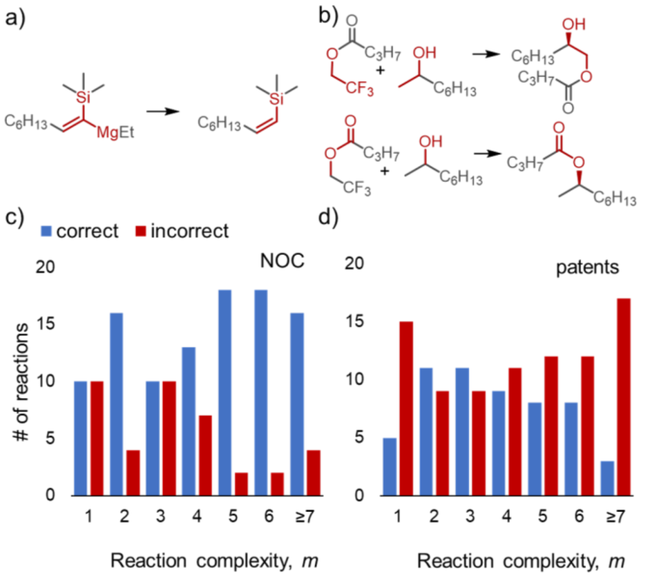


Figure 5. Examples and correctness statistics of low-popularity reaction types. a) An example of a correct but highly specialized reaction template that has maintained k =1 popularity for at least 5 years. The template itself is delineated by red bonds; gray bonds specify the remaining parts of the specific substrates/products as reported in [16a]. **b)** Another example but this time erroneous on account of the incorrect structure of the product. The correct reaction from the source publication[16b] is shown below. **c)** Histograms summarizing the counts of correct and incorrect k = 1 @ 5 yr templates based on 140 examples from the NOC. Different pairs are for different values of reaction complexity, m (20 examples for each and calculated based on atom-mapping) **d)** Analogous histogram but for 140 examples from patents.

To sum up, we showed that in terms of reaction discovery, organic chemistry expands slower than generally assumed and becomes more reliant on well-known reaction types. These trends are unlikely to be artifacts of a particular reaction set as they are seen both in the NOC and patent collections curated in different ways, by different organizations, and with different focus (academic for NOC vs. industrial for USPTO). One reinvigorating trend is that increasing fractions of newly discovered reactions become more and more complex (m > 4) – since such reaction are often multicomponent (cf. **Figure S2**) or cascade transformations, they fit well with the aim of modern chemistry to reduce reaction operations and increase pot economy. In the context of chemoinformatics, the reaction-type templates we defined here (1) can allow for more objective scrutiny of reaction collections to estimate their true diversity (i.e., numbers of distinct chemistries rather vs. multiple manifestations of the same reaction types) and also identify erroneous entries; and (2) can help assess which reaction classes encompass sufficient numbers of precedents (measured by popularity k) to allow for accurate machine learning models reflecting reaction mechanism (as in [14]).

Keywords: chemical reactions, organic synthesis, chemoinformatics

Acknowledgements: This research was developed with internal funding from Allchemy, Inc. B.A.G. acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1.

---

[1] a) M. Fialkowski, K.J.M. Bishop, V.A. Chubukov, C.J. Campbell, B.A. Grzybowski, *Angew. Chem. Int. Ed.* **2005**, *44*, 7263-7269; b) K.J.M. Bishop, R. Klajn, B.A. Grzybowski, *Angew. Chem., Int. Ed.* **2006**, 45, 5348-5354; c) B.A. Grzybowski, K.J.M. Bishop, B. Kowalczyk, C.E. Wilmer, *Nat. Chem.* **2009**, *1*, 31-36; P.M. Jacob, A. Lapkin, *React. Chem. Eng.* **2018**, *3*, 102-118; e) E.J. Llanos, L. Wilmer, D.L. Luu, J. Jost, P.F. Stadler, G. Restrepo, G. *Proc. Nat. Acad. Sci.* **2019**, *116*, 12660-12665.

[2] a) M. Kowalik, C.M. Gothard, A.M. Drews, N.A. Gothard, A. Weckiewicz, P.E. Fuller, B.A. Grzybowski, K.J.M. Bishop, *Angew. Chem. Int. Ed.* **2012**, *51*, 7928–7932; b) C.M. Gothard, S. Soh, N.A. Gothard, B. Kowalczyk, Y. Wei, B. Baytekin, B.A. Grzybowski, *Angew. Chem. Int. Ed.* **2012**, *51*, 7922–7927; c) S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937; d) K. Molga, P. Dittwald, B.A. Grzybowski, *Chem. Sci.* **2019**, *10*, 9219–9232.

[3] a) T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M.P. Startek, G.J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S.L.J.Trice, B.A. Grzybowski, *Chem* **2018**, *4*, 522–532; b) B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich, B. A. Grzybowski, *Nature* **2020**, *588*, 83–88.

[4] a) M.H.S. Segler, M.P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966-5971; b) M.H.S. Segler, M. Preuss, M.P. Waller, *Nature* **2018**, *555*, 604–610; c) F. Strieth-Kalthoff, S. Frederik, M.H.S. Segler, F. Glorius, *Chem. Soc. Rev.* **2020**, *49*, 6154-6168; d) P. Schwaller, D. Probst, A.C. Vaucher, V.H. Nair, D. Kreutter, T. Laino, J.L. Reymond, *Nat. Mach. Intell.* **2021**, *3*, 144-152; e) C.W. Coley, R. Barzilay, T.S. Jaakkola, W.H. Green, K.F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 5, 434–443; f) W. Jin, R. Barzilay, T. Jaakkola, C.W. Coley, *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*: **2017**, 2604– 2613; g) F. Feng, L. Lai, J. Pei, *Front. Chem.* **2018**, 6, #199; h) K. Lin, Y. Xu, J. Pei, *Chem. Sci.* **2020**, *11*, 3355-3364; i) P.P. Plehiers, G.B. Marin, C.V. Stevens, K.M. Van Geem, *J. Cheminf.* **2018**, *10*, #11; j) Ch. Yan. Q. Ding., P. Zhao, S. Zheng, J. Yang, Y. Yu, J. Huang, J. **2020**,

10.1002/ange.202111540

https://arxiv.org/abs/2011.02893; k) A. Thakkar, S. Johansson, K. Jorner, D. Buttar, J.L. Reymond, O. Engkvist, *React. Chem. Eng.* **2021**, *6*, 27-51; l) C.A. Nicolaou, I.A. Watson, M. LeMasters, T. Masquelin, J. Wang, *J. Chem. Inf. Model.* **2020**, 60, 2728–2738.

[5] This belief of chemistry expanding "exponentially" has been used to advocate automated extraction of reaction rules from databases rather than their expert coding. While expert curation of tens of thousands of reaction rules is, without a doubt, a laborious undertaking, it offers vastly superior quality and, ultimately, synthetic predictions that work in the laboratory not only for simple but also complex targets. For further discussion, see ref. [7].

[6]https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873

[7] K. Molga, E. P. Gajewska, S. Szymkuć, B. A. Grzybowski, *React. Chem. Eng.* **2019**, *4*, 1506-1521.

[8] Y. Hayashi, *Acc. Chem. Res.* **2021**, *54*, 1385-1398.

[9] a) L. Wilbraham, S.H.M. Mehr, L. Cronin, *Acc. Chem. Res.* **2021**, *54*, 253–262; b) L. Porwol, D.J. Kowalski, A. Henson, D.L. Long, N.L. Bell, L. Cronin, *Angew. Chem. Int. Ed.* **2020**, *59*, 11256-11261.

[10] a) W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin, B.A. Grzybowski, *Nat. Commun.* **2019**, *10*, #1434; b) P. Schwaller, B. Hoover, J.L. Reymond, H. Strobelt, T. Laino, *Sci. Adv.* **2021**, *7*, eabe4166.

[11] As MAPPET is more accurate for complex reactions, it was used mostly to map the NOC dataset; on the other hand, IBMs software can handle better patent reactions in which solvents and/or reagents are often incorrectly listed as substrates. This being said, the general trends described in the main text were similar when patent reactions were remapped using MAPPET and when NOC reaction were mapped using IBM's tool.

[12] C.W. Coley, W.H. Green, K.F. Jensen, *J. Chem. Inf. Model.* **2019**, *59*, 2529–2537.

[13] In ref. [1e], the authors observed a "kink" in the expansion of chemistry – measured in terms of specific reactions and molecules – around 1980 and attributed it to the emergence of organometallic and bioorganic chemistries. While it might be tempting to use similar reasoning for our NOC data from Fig. 2a, we note that the "kink" for patents is only around 2009, making such a chemical-oriented explanation somewhat doubtful. Instead, the change to linear trends – with increase of slope immediately after the kinks – might reflect maturation and automation of data input technologies into the Beilstein/Reaxys and USPTO repositories.

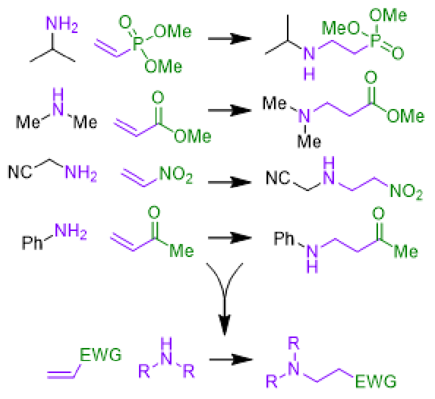
[14] W. Beker, E.P. Gajewska, T. Badowski, B.A. Grzybowski, *Angew. Chem. Int. Ed.* **2019**, *58*, 4515–4519; b) X. Li, S. Zhang, L. Xu, X. Hong, *Angew. Chem. Int. Ed.* **2020**, *59*, 13253–13259; c) S. Moon, S. Chatterjee, P.H. Seeberger, K. Gilmore, *Chem. Sci.* **2021**, *12*, 2931–2939; d) G. Pesciullesi, P. Schwaller, T. Laino, J.L. Reymond, *Nat. Commun.* **2020**, *11*, 3878; e) M. Moskal, W. Beker, S. Szymkuć, B.A. Grzybowski, *Angew. Chem. Int. Ed.* **2021**, *60*, 15230–15235.

[15] A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, T. Laino, *Nat. Mach. Intell.* **2021**, *3*, 485-494.

[16] a) E. Negishi, T. Takahashi, *J. Am. Chem. Soc.* **1986**, *108*, 3402–3408; b) L.T. Kanerva, J. Vihanto, M.H. Halme, J.M. Loponen, E.K. Euranto, *Acta Chem. Scand.* **1990**, *44*, 1032-1035.

10.1002/ange.202111540

Chemoinformatics



TITLE:
Is organic chemistry really growing exponentially?

TOC TEXT:
Although the number of published reaction examples continues to increase exponentially with time, the number of new reaction types being discovered grows only linearly or even sub-linearly. As a discipline, we rely on re-using the most familiar and popular reactions classes, although we also discover increasingly more complex transformations, especially in recent years. Analysis of reaction-type dynamics is a useful tool with which to estimate the numbers of meaningful machine-learning models that can be constructed and also for tracing suspicious/erroneous entries in reaction databases.