JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Perspective

Artificial Intelligence for Retrosynthetic Planning Needs Both Data and Expert Knowledge

Felix Strieth-Kalthoff,[▽] Sara Szymkuć,[▽] Karol Molga,[▽] Alán Aspuru-Guzik, Frank Glorius,[*] and Bartosz A. Grzybowski[*]

Cite This: https://doi.org/10.1021/jacs.4c00338

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Rapid advancements in artificial intelligence (AI) have enabled breakthroughs across many scientific disciplines. In organic chemistry, the challenge of planning complex multistep chemical syntheses should conceptually be well-suited for AI. Yet, the development of AI synthesis planners trained solely on reaction-example-data has stagnated and is not on par with the performance of "hybrid" algorithms combining AI with expert knowledge. This Perspective examines possible causes of these shortcomings, extending beyond the established reasoning of insufficient quantities of reaction data. Drawing attention to the intricacies and data biases that are specific to the domain of synthetic chemistry, we advocate augmenting the unique capabilities of AI with the knowledge base and the reasoning strategies of domain experts. By actively involving synthetic chemists, who are the end users of any synthesis planning software, into the development process, we envision to bridge the gap between computer algorithms and the intricate nature of chemical synthesis.

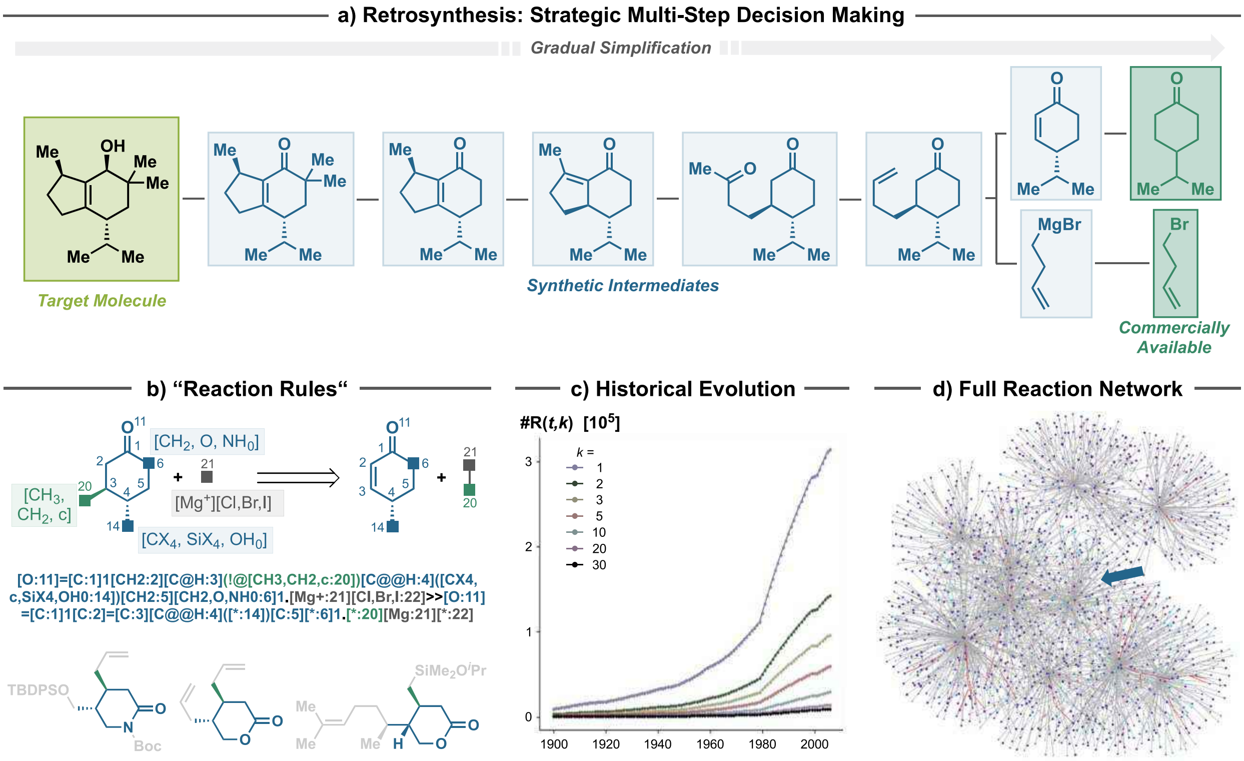
Artificial intelligence (AI) is having a transformative impact on fields as diverse as interaction and reasoning in natural language,[1,2] solving games of strategy,[3,4] computer vision,[5,6] recommender systems,[7] theory of computing,[8] and many more. In the molecular sciences, the success of AlphaFold2[9,10] in predicting the structures of proteins has generated widespread enthusiasm and has raised expectations for similar feats in chemistry. Many of these expectations have been met by recent successes in the AI-driven optimization of catalysts,[11−13] materials,[14−17] or reactions[18−20] but not yet in the classic problem of synthesis planning.[21,22] Recent efforts in this area[23−25] have given rise to a vision of "exploit[ing] artificial intelligence to automatically learn organic synthesis from reaction examples".[26] Unfortunately, as of today, purely data-driven algorithms have remained confined to the syntheses of simple targets and face severe challenges when confronted with, e.g., complex scaffolds with multiple stereocenters. These limitations have often been attributed to insufficient amounts of reaction examples,[27] especially those reporting failed reactions. However, this Perspective argues that the problem has deeper roots than just the quantities of reaction examples—from noise and biases in the reaction data sets to complex reaction-condition relationships hidden "beneath" the published data—and collecting a sufficient number of examples will not be feasible in the near term. Instead, we argue the field can make significant advances by (i) embracing the key elements of domain knowledge and "classical" theoretical tools developed by chemists over the past centuries (but not contained in reaction databases) and (ii) improving the AI's algorithmic basis to better match the thinking of human experts. Early examples of such AI-knowledge "hybrids" have already been demonstrated to attack demanding natural product targets[28−30] and will define further development of "algorithmic chemistry" in synthesis. We believe that key to these improvements will be closer cross-talk between algorithm specialists and synthetic chemists—whose expertise cannot be easily replaced by more reaction examples.

THE CHALLENGE OF SYNTHESIS PLANNING

Synthetic routes are typically designed following the logic of retrosynthesis, originally introduced by E. J. Corey in the 1960s,[21,22] in which a desired target molecule is iteratively disconnected into progressively simpler intermediates, ultimately reaching commercially available and ideally inexpensive starting materials (Figure 1a). From a chemical viewpoint, these disconnections are defined by known reaction types (e.g., nucleophilic substitution, Diels−Alder cycloaddition, etc.) and can be considered as basic "moves" from which complete synthesis "games" are made (Figure 1b). From an algorithmic standpoint, the design of retrosynthetic routes is certainly not a simple problem, as there are as many as $\sim 10^5$ different types of reaction moves (Figure 1c and previous analyses[31]) and because the networks of synthetic possibilities for any given target are very large (Figure 1d), scaling as $\sim 100^n$, where n is the number of reactions needed to complete the synthesis.[32] For complex natural products, n is usually measured in tens, and within such enormous networks, only a few sequences of moves may exist that define a viable synthesis.

The Promise and Limitations of Data-Driven Synthesis. However, these very large numbers by themselves are

Received: January 9, 2024
Revised: March 10, 2024
Accepted: March 11, 2024



```
[O:11]=[C:1]1[CH2:2][C@H:3](!@[CH3,CH2,c:20])[C@@H:4]([CX4,
c,SiX4,OH0:14])[CH2:5][CH2,O,NH0:6]1.[Mg+:21][Cl,Br,I:22]>>[O:11]
=[C:1]1[C:2]=[C:3][C@@H:4]([*:14])[C:5][*:6]1.[*:20][Mg:21][*:22]
```

Figure 1. Synthesis planning, reaction rules, and retrosynthetic networks. (a) Retrosynthesis by strategically disconnecting a target molecule (here, natural product (+)-brasilenol[33]) into progressively simpler intermediates. (b) Example of a "reaction rule" describing a substrate-controlled 1,4-addition of an organometallic reagent, used in the synthesis from (a). Such rules are often represented in the alphanumeric SMARTS notation,[34] specifying the identities of atoms in the chemical synthesis and mapping atoms across the reaction, and can be accompanied by further information (conditions, incompatible groups, etc.). Importantly, the rules are broader than any specific literature example and can be applicable to other targets—the three structures shown at the bottom illustrate the application of the 1,4-addition rule to some intermediates used in the syntheses of other natural products (Strychnofoline, Yohimbane, and Crenulatan diterpene). (c) The number of reaction rules, *#R*, governing organic syntheses is very large but finite. The graph depicts *#R* as a function of time, *t*. Each curve represents rules above a specific "popularity", *k*, defined by the number of literature examples in which a given rule was used, as discussed in detail in our previous study.[31] Rules with $k = 1$ (i.e., reaction types reported only once) are mostly database entry errors. For the synthetically more useful rules, i.e., $k > 2$, the numbers are large but finite (∼50,000−100,000). Contrary to what has often been claimed,[35] the rate at which new reaction types are discovered is not increasing exponentially but has decreased in recent decades (i.e., *#R*(*t*) curves flatten out). (d) With so many rules, each intermediate ("retron") considered during retrosynthetic planning can be expanded into multiple "synthons" (on average, ca. 100; see our previous work for a systematic analysis[32]), and the networks of possibilities can rapidly become extremely large. The image is an excerpt from a network originating from Valsartan (node marked by the blue arrow) and expanding just ten progeny/"synthon" nodes. Realistic retrosynthetic design of complex targets requires tens of thousands of such expansions.[28] Image in (c) is adapted with permission from ref 31. Copyright 2021 Wiley-VCH GmbH. Network in (d) is a screenshot of the Synthia program.[28,32,36]

not necessarily an unsurmountable obstacle to developing algorithms for synthesis planning, given that modern AI tools have mastered problems such as chess or GO, in which the networks of possibilities to consider are even larger.[32] Moreover, millions of published reaction examples have nowadays been digitized and are stored in reaction databases such as publicly available USPTO[37] or proprietary *Reaxys*[38] or *SciFinder*;[39] it can be argued that they represent all existing synthesis knowledge, from which the reaction rules could easily be extracted explicitly (in the form of subgraph edits characterizing different reaction types) or learned implicitly from the data. With these components in place, various neural network architectures[40−48] could be used to associate specific rules/"moves" with specific types of molecules to which they have been applied. In this way, when AI is confronted with a new target, it could recognize the pertinent structural features of the molecule and suggest the most appropriate reaction rule/"move" to guide the exploration of retrosynthetic networks. What really fuels the imagination in this vision is that this processing of and learning from the *entire* synthesis knowledge could be realized in one afternoon!

Unfortunately, it appears that this vision is affected by certain overoptimistic assumptions and inaccurate analogies. To begin with, in games like chess, the rules are very well-defined in terms of both what is allowed (e.g., a bishop moves along diagonals) and what is not (e.g., a bishop cannot jump over other pieces). In contrast, reaction data alone is not only littered by erroneous and incomplete entries but, as we will discuss, allows neither for the chemically unambiguous definition of reaction rules nor for their extrapolation to the "impossible cases", without which the synthesis planning programs tend to invent impossible reactions. We detail key challenges below, arguing that at least some of them can be remedied by broadening the scope of purely reaction-example-driven AI: In fact, the mastery of synthesis experts extends well beyond the knowledge of reaction examples and encompasses symbolic and implicit knowledge such as reaction mechanisms and reactive intermediates, 3D structures, steric, electronic, and stereoelectronic considerations, and often "fuzzy" insights

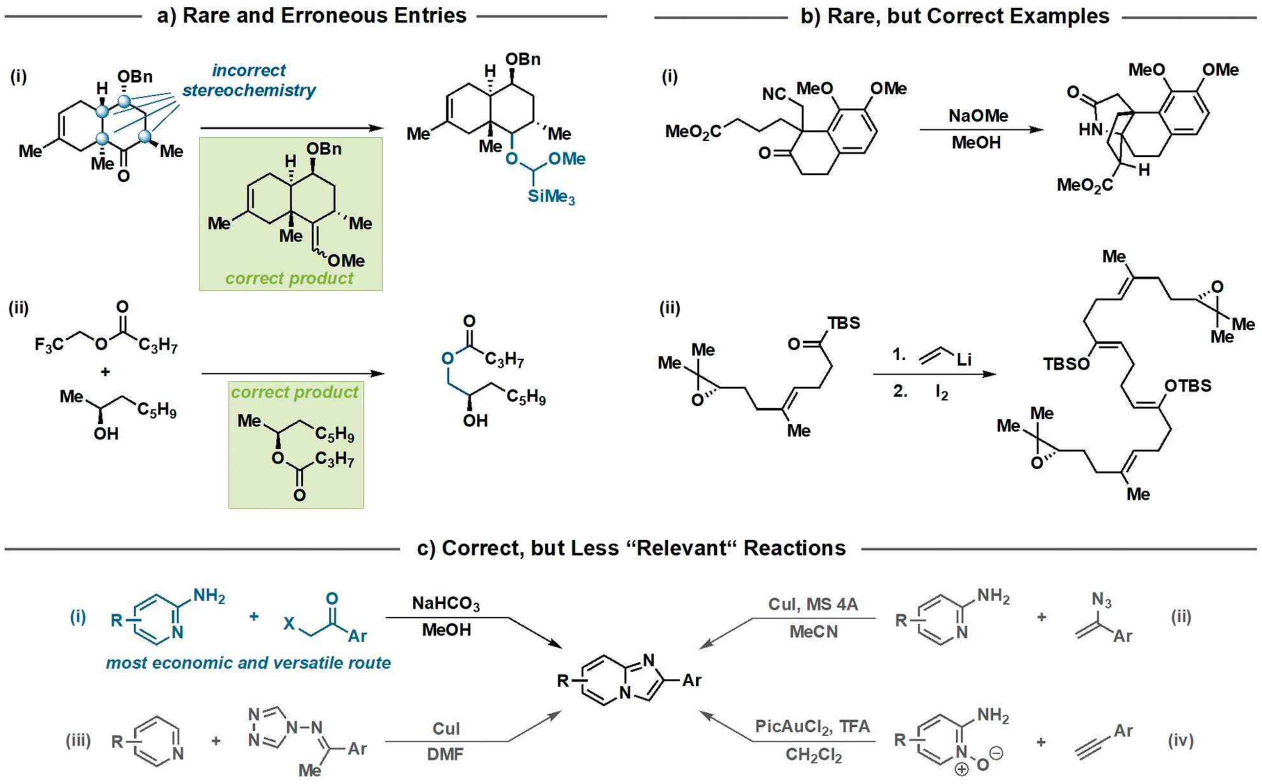

a) Rare and Erroneous Entries
(i)
incorrect stereochemistry
correct product
(ii)
correct product
b) Rare, but Correct Examples
(i)
NaOMe
MeOH
(ii)
1. Li
2. I2
c) Correct, but Less "Relevant" Reactions
(i)
most economic and versatile route
NaHCO3
MeOH
CuI, MS 4A
MeCN
(ii)
(iii)
CuI
DMF
PicAuCl2, TFA
CH2Cl2
(iv)

Figure 2. Examples of erroneous, rare, and biased reaction examples. (a) Examples of erroneous reaction types that have only one occurrence in the literature ("popularity" $k = 1$[31]). Reactions (i) and (ii) are both database entry errors with incorrect stereochemistry of a substrate (on highlighted atoms) and/or incorrect products. (b) Examples of correct reaction types that have only one occurrence in the literature ($k = 1$[31]) and have been applied to specialized scaffolds only ((i) stephadiamine synthesis by Stoltz, Trauner, and co-workers;[51] (ii) (+)-α-onocerin synthesis by Corey and co-workers[52]). (c) Imidazopyridines can be prepared via different methods reported in the literature, including: (i) condensation of 2-aminopyridines with a-haloketones; (ii) copper-mediated condensation of 2-aminopyridines with vinyl azides;[53] (iii) dehydrogenative coupling of pyridines with hydrazones;[54] or (iv) gold-catalyzed cycloaddition of 2-aminopyridine *N*-oxides and alkynes.[55] The most economical and versatile methodology is (i). Methods (ii)−)iv) have >50 citations and ∼20 scope examples (i.e., popularity $k \sim 20$), but even a decade after their publication, they have not been used other than by the original authors (presumably owing to the low accessibility of reactants (ii)−(iv), safety concerns regarding azides (ii), or economic considerations (iv) with respect to gold catalysts).

derived from practical laboratory experience. While we do not dismiss the potential of inferring this knowledge from data against the limit of infinite reaction examples, we posit that, given the current data landscape (and the foreseeable data quantities in the coming decades), computer-aided synthesis planning can be markedly enhanced by incorporating these additional modalities of synthesis expert knowledge.

In the first part of this Perspective, we will discuss the above-mentioned challenges of learning from reaction examples and will outline how to enhance the prediction quality for single reaction steps. Subsequently, we will discuss algorithms that combine these single-step predictions into searches for multistep synthetic routes.

■ PREDICTING THE OUTCOMES OF CHEMICAL REACTIONS

The Challenges of Data-Driven Reaction Prediction. *i. Errors and Biases in Chemical Reaction Data.* To begin with, reaction repositories contain surprisingly large numbers, up to tens of percent,[31,49] of erroneous entries missing key atoms, lacking entire substrates or products, or listing common solvents or reagents in their lieu (particularly in the USPTO data set, see ref 31). This causes significant problems, as such erroneous entries prompt any data-driven algorithm to learn incorrect reaction rules. The simplest remedy has been to eliminate entries whose reaction templates (i.e., patterns of atoms defining the reaction type) appear only once or twice in the database—such extremely rare reactions are, indeed, often erroneous[31] (Figure 2a), although it should be remembered that they can also be simply unique (yet useful in the synthesis of rare scaffolds; see Figure 2b). Unsupervised, purely data-reliant AI methods have been developed to address this problem, but they sometimes rely on rather dangerous—especially for more advanced chemistries—assumptions, e.g., that "the most difficult examples to learn while training reaction prediction models are probably examples of wrong chemistry".[50] Such assumptions lead to the rejection of perfectly viable reactions, mostly on account of an algorithm's inability to detect trivial missing reactants/reagents or because it cannot recognize reactions that are reported in shorthand, with multiple steps concatenated into one (see specific examples in Supporting Information, SI Section 1). Our experience has shown that many of such problems could be avoided by straightforward knowledge-based analyses—e.g., matching the reaction entries with lists of common solvents and reagents (to detect their mis-assignments as reactants or products), detecting reactions in which masses of substrates and products differ significantly (suggesting missing substrates), etc.

On the flipside of the coin, a less widely appreciated complication is that not all correct literature examples, even

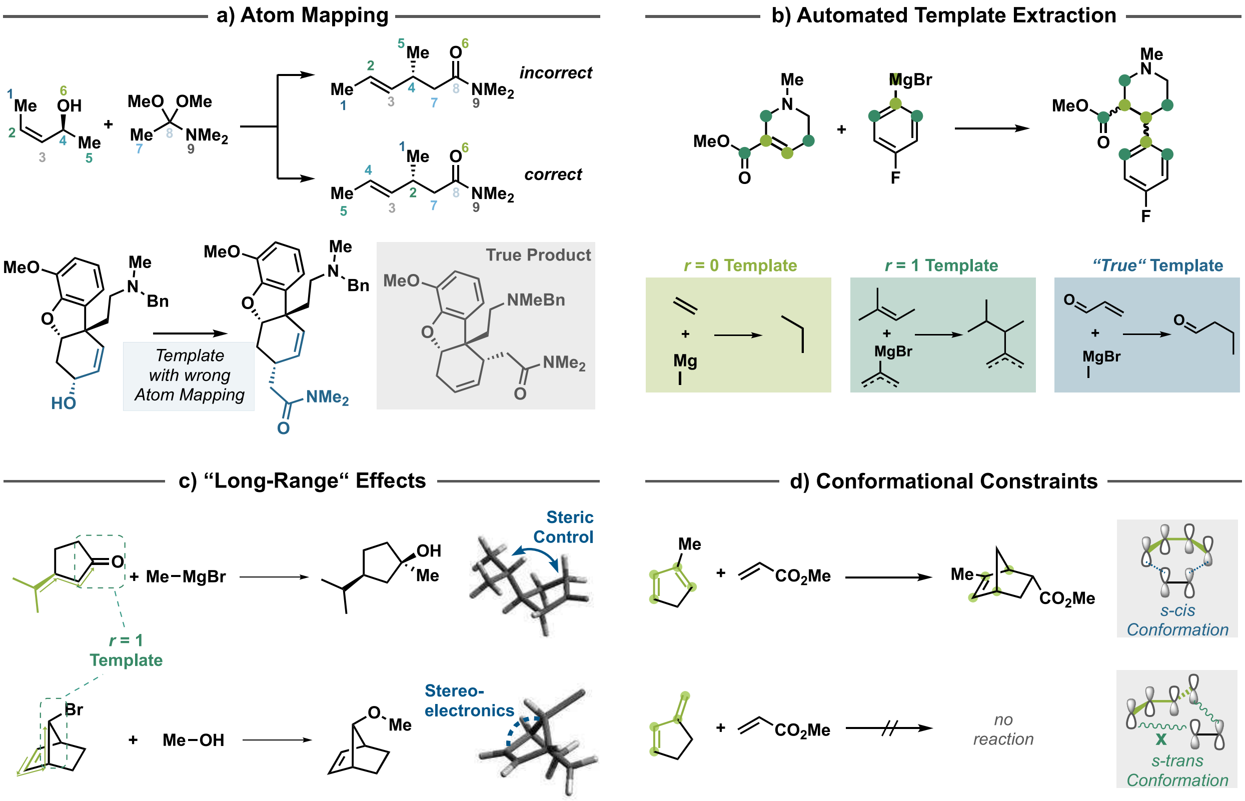


Figure 3. Challenges of automated atom mapping and reaction-template extraction. (a) Example of atom mapping in an Eschenmoser-Claisen rearrangement. With incorrect mapping, the template applied to the substrate shown in the bottom row will predict formation of an incorrect product (one with relevant parts colored in blue). (b) Definition of bond "radii" around the reactive center. (c) Examples of longer-range interactions "through space" that determine reactivity. The $r = 1$ environments are in dashed-line boxes. However, reaction outcomes are dictated by the more distant parts colored in green. (d) An example of conformational constraints (here, imposed by the ring structure) dictating the reactivity of simple dienes in Diels–Alder reactions.

those representing relatively popular methodologies, may be actually desirable to include in the AI's training data, at least not until some bounds are imposed on their applicability. As an example, defunctionalization reactions (e.g., dehalogenations, deoxygenations, decarboxylations, alkene reductions) without any constraints may prompt AI to later use them promiscuously, generating synthons featuring halogens, hydroxy groups, carboxyl groups, or C=C bonds in all possible positions (see example in SI Section 2). This, in turn, can dramatically increase the number of synthetic options to evaluate, which may be negligible for shorter syntheses, but very problematic for complex targets for which the sizes of reaction networks test the limits of computing power even without such "decoys". In earlier works, we have shown the use of such constraints, implemented with caution and only upon expert scrutiny of *pros* (productively limiting the search space) and *cons* (occasionally missing unconventional pathways): For instance, "converting" an aromatic C–H bond into an amine (in the retrosynthetic direction) is allowed only if it leads to *para* and/or *ortho* substituted arenes. In that way, deamination (in the forward direction) becomes part of a synthetic strategy in which the amino group activates the aromatic ring and facilitates electrophilic aromatic substitution(s) at earlier retrosynthetic steps. When the desired substitution pattern is reached, the amino group can be removed. Similarly, retrosynthetic carboxylations can be constrained to cases in which they lead to 1,3-dicarbonyl synthons, which may, during further retroanalysis, enable facile alkylations or other base-catalyzed reactions with various electrophiles.

A further bias to consider is when the literature provides similar numbers of examples of different methodologies producing the same scaffold type–some of these methods may be truly useful (e.g., offering a different scope of incompatibilities) but for some, high database counts merely document the scope published in one, original publication. There are no formal errors in such reactions but, when taught as generalizable rules, they may produce syntheses that chemists might find impractical (see examples in Figure 2c). In our experience, such corner cases require careful, expert analysis.

ii. Atom Mapping. Assuming a reaction data repository is properly precleaned to some satisfactory degree, all algorithms that rely on explicit reaction rules or "templates" require the identification of the corresponding atoms in products/"retrons" and substrates/"synthons". If this so-called atom mapping is off, AI will learn incorrect changes in the bonding patterns, resulting in nonsensical reaction predictions (Figure 3a). While atom mapping can be considered a "vintage" problem and may be trivial for simple reaction types, there are currently no tools to ensure its correctness in complex reactions. Automated atom mapping tools have been developed both by knowledge-aided[56] and fully data-driven algorithms.[57] In the context of our discussion, it is worth noting that the best-in-class AI reaction mapper[57] (based on attention weights in a transformer neural network trained on

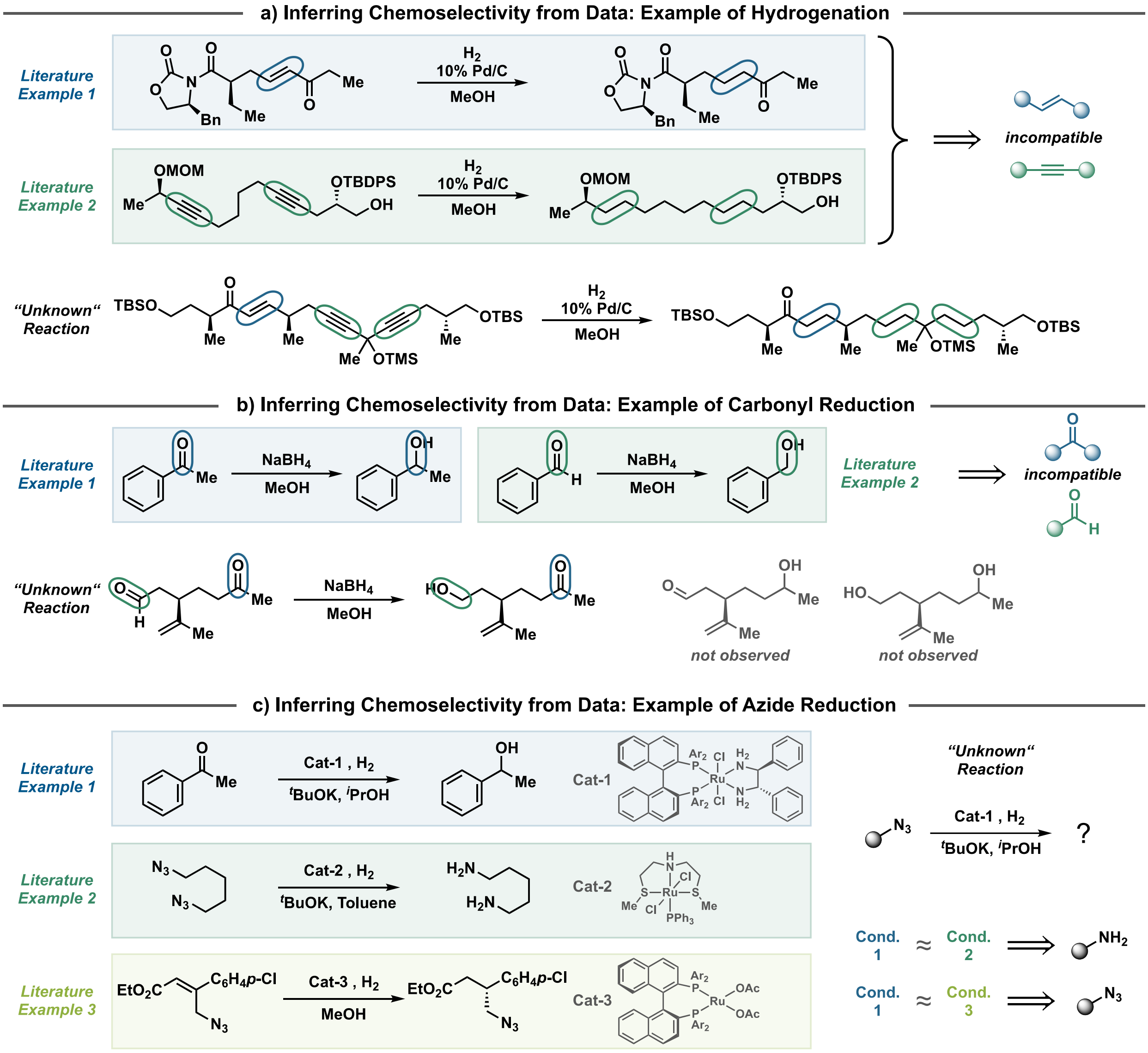


Figure 4. The hurdles of learning group (in)compatibilities from data alone. (a) Published reactions 1 and 2 are alkene and alkyne hydrogenations, respectively, under the same reaction conditions (H_2, Pd/C in MeOH). It is deduced that these groups, if present in the same molecule, would be incompatible in the hydrogenation reaction under these conditions. This is, in fact, the case in the example in the bottom row, where hydrogenation is not chemoselective for an intermediate toward a *Phytophthora α1* mating hormone containing both an alkene and an alkyne.[70] (b) Selective outcomes can be achieved even if conditions of examples 1 and 2 are (formally) identical but reaction rates differ, as for the aldehyde and ketone reduction with $NaBH_4$. Selective outcome in the bottom reaction was, indeed, observed in the synthesis of *trans-β*-elemene.[71] (c) Prediction complexity is further increased when the reported conditions differ, e.g., when predicting whether an azide would remain intact in (i.e., be compatible with) reduction of a ketone under Noyori's conditions. Existing literature covers ketone reduction under these conditions (example 1, blue) but no examples of azide reduction. Literature examples under "similar" conditions do not provide a definitive answer: There are examples in which apparently similar conditions resulted in reduction (example 2, dark green[72]) but also those under which the azide stayed intact (example 3, bright green[73]).

millions of reactions) is significantly faster but less accurate than a simple isomorphic mapping algorithm supplemented with 20 knowledge-based heuristics formulated by synthesis experts to guide the algorithm's intermediate atom assignment[56] (78% vs 84% accuracy). Clearly, the number of 20 expert-assigned rules is small but illustrates the importance of including additional knowledge modalities, and further improvements can be expected with additional expert input or through hybrid AI−expert systems.

iii. Reaction Templates. In the rule-based approaches to retrosynthesis, chemical reactions are treated as "templates", by extracting parts of the molecules at and near the reaction center and subsequently using them as "operators" to synthesize new molecules (Figure 1b). These template-based algorithms are of particular interest because, assuming correct mapping, they capture known reactivity patterns and outperform template-free methods[43−48] (see Molga et al.[58] and additional examples in SI Section 3), especially for more complex chemistries. Automated extraction procedures first identify atoms that change their local bonding patterns and constitute the reaction "core," as well as some flanking atoms defining the "environment" to some radius *r* measured in bond distance (Figure 3b).[58−60] The *r* = 0 "core-only" templates often miss the important structural motifs proximal to the reaction center; consequently, they generate many false positive predictions when applied to new molecules. As r increases, the templates become more accurate but also applicable only to molecules increasingly similar to those already in the data set (for very large r, the rules would only be able to reproduce the original reaction data set).[59] To balance these two tendencies, $r = 1$ or $r = 2$ values are typically applied to extract templates from reaction repositories. A major challenge, however, is that the value of r may need to be adjusted for different reaction types or molecular environments. Coley and co-workers made valuable contributions to mitigate some of the resulting issues by adding knowledge-based heuristics preventing, e.g., disconnection of common functional groups at the template's extremities.[60] Yet, existing template extraction solutions are far from being universal, and stereochemistry in particular remains a challenge: As of today, it is unclear how to ensure that the templates retain proper stereochemistry (see examples in SI Section 4) or how to recognize larger, stereochemically complex motifs that may impact reactivity (e.g., shielding or conformation-determining groups). Moreover, such motifs may be distant in a two-dimensional graph but proximal in space, owing to the molecule's three-dimensional structure (Figure 3c). Augmenting template extraction with traditional conformational analysis and metrics of conformational flexibility or steric hindrance could represent an attractive direction in this regard. Additionally, some domain knowledge—desirably, including information on the underlying reaction mechanisms—seems inevitable to define *which* groups and spatial arrangements are relevant for specific reactions (Figure 3d and study by Moskal et al.[61]).

iv. Nonselectivities and Incompatibilities. Finally, detecting structural motifs that may engage in undesired side reactions is a major problem plaguing all AI synthesis design programs, even at the level of very simple targets and routes (cf. examples in Molga et al.[58]). This problem actually consists of two parts. On one hand, it is relatively easy to ensure that the *same* reaction type cannot be applied in places of the reactant other than the desired one; this is checked by applying the reaction rule in the forward direction and verifying that it does not give more than one product (see previous works for more nuanced discussion of chemo- and regioselectivity modeling[13,61−69]). On the other hand, it is much harder to ascertain that synthons cannot engage in side reactions of *other* types. Given very limited availability of data sets reporting side reactions or "failed" reaction outcomes, auxiliary models to detect such reactivity conflicts have been built. In most cases, these have been trained on negative data generated *in silico*, e.g., by assuming that, if molecules A and B react to product C, they do not react to products D, E, F, etc. or by shuffling the associated pairs of products and corresponding "correct" reactions.[35] However, these schemes are chemically speaking rather crude, and it has been recognized that, to properly predict conflicting reactivity, one needs to consider reaction conditions—this task was anticipated to be a relatively straightforward feature to add, requiring "additional search in condition space", limited only by "time constraints".[35] Unfortunately, this is not the case.

To see why, assume that along with the reactions, we managed to extract the corresponding conditions—which, in itself, is a bold assumption, as the inconsistencies, gaps, and errors in terms of reporting reaction conditions are far more pronounced than in the case of reactants and products (vide infra). Notwithstanding, let us consider a situation in which some reactions R_1 and R_2 were reported under the same solvent and reagent conditions, C_1. For instance, in the example in Figure 4a, R_1 hydrogenates a double bond, while R_2 hydrogenates a triple bond, both under formally identical conditions C_1 = {10% Pd/C, H_2, MeOH}. In this case, R_1 and R_2 can be recognized as "competing" and, consequently, if both a double and a triple bond are present in the same molecule, neither can be selectively hydrogenated under these conditions (see Figure 4a). However, even such identical-condition cases may be more nuanced if the rates differ significantly and selective outcomes can still be achieved; for instance, reactions R_1 and R_2 in Figure 4b may seem competing but, in reality, the rate differences can be harnessed to perform a reaction in the bottom row selectively. Interestingly, this aldehyde/ketone differentiation is well-known to chemists but can be hard to learn automatically from published reaction examples alone, as R_1 and R_2 have been reported to proceed on similar time scales[74−76] (in general, reaction times are often not indicative of actual reaction kinetics but rather reflect operational convenience[77]). Learning becomes even more problematic when different conditions are at play. Say, if R_1 was performed under conditions C_1 and R_2 *could* also work under these conditions, but in reality, was executed and reported under some other conditions C_2, then we would not find R_2 as potentially competing with R_1. A first-line approach to avoid such problems would be to curate "dictionaries" of "synonymous" conditions, e.g., in our particular example, informing the algorithm that conditions C_1 could be used in lieu of C_2. However, such condition matching would require an extensive study by expert chemists and would likely only be applicable for extremely "simple" reaction types (see example in Figure 4c in which such reasoning by condition similarity is inconclusive).

The Need for Additional and More Representative Data. Whereas the problems of atom mapping or template extraction are largely independent of the available data quantities, there is little doubt that additional data *will* improve the performance of AI models for reaction outcomes, especially if this data also describes "failed" reactions[78] which are rare in existing data sets.[77−80] Given that reaction data is much more costly than image or speech data and is associated with significant generation of waste, open-science initiatives like the Open Reaction Database (ORD)[81,82] and the goal to systematically capture metadata (including metadata that is often absent or wrong in existing databases) are of high value. This said, we feel that the ways in which reactions are reported and harnessed require certain improvements.

For instance, chemists in method development, natural product synthesis, or materials discovery perform large numbers (often thousands) of experiments to optimize synthetic routes or conditions; yet, these experiments either are not published at all or are relegated to the supporting materials, meaning that they are "lost" to chemical databases and AI algorithms. Collecting these experiments in initiatives like ORD will require not only simplified and standardized upload protocols (e.g., via electronic lab notebooks[83−85]) but also external incentives to the authors, from both the publishers and funding agencies. In addition to such ongoing standardization efforts for newly added reactions, recent advances in NLP could provide an attractive strategy for digitizing older, heterogeneously reported data. It should be remembered, however, that these efforts will require time and,

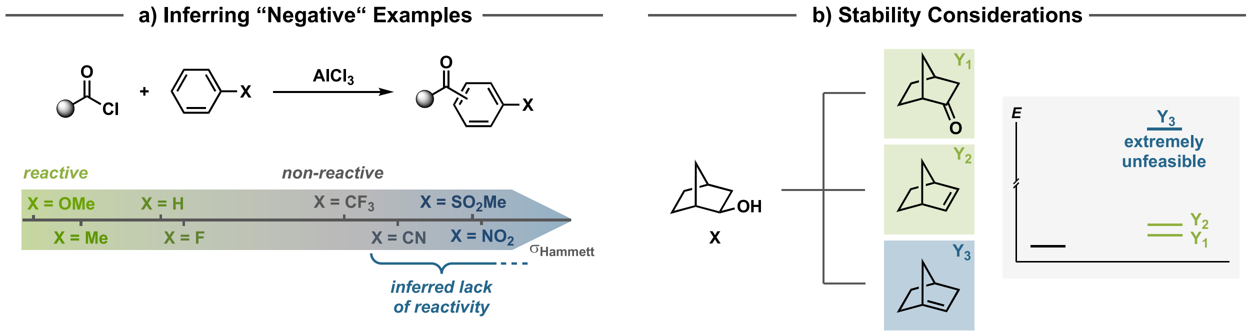


Figure 5. Knowledge-guided heuristics for augmenting synthesis-planning algorithms. (a) Example of a "reactivity series" for a Friedel–Crafts acylation. As no reaction is observed with a CF_3 substituent, experiments with more electron-withdrawing substituents are unlikely to be successful (indicated on the right part of the axis). Since electron-withdrawing/donating propensities are well-known (and can be approximated by, e.g., Hammett constants or more advanced calculations; see the supporting information of ref 36), similar series could be considered for other, less explored ring systems or reaction types (e.g., Prins cyclizations suffering from competitive Oxonia-Cope rearrangements). In this way, a single or few experimental data points would allow data augmentation with "negative" examples that are likely to give no reactivity. (b) Structurally viable (green) and impossible (blue) synthons to the 2-norbornanol retron, as governed by thermodynamic constraints.

based on the experiences of the past several years, we expect that open reaction data sets with satisfactory quantity and quantity of reaction examples are at least a few years away.

A directly related issue is the question of *which* reaction data is needed most to improve AI synthesis planning. In fact, if we continue to amass reactions in the same manner as up until now, we will make the popular reaction types even more popular (by the mechanism of preferential attachment which has guided the evolution of organic chemistry for over 200 years; see systematic analyses[31,86,87]) and will thus generate more data points in the already well-explored regions of synthetic space. Arguably, an additional thousands of nucleophilic substitutions, amide couplings, or Suzuki or Buchwald–Hartwig reactions may not be of prime importance to AI models—especially if applied to simple substrates—since the existing data sets already give us a decent understanding of which groups are compatible/incompatible in these popular reactions.

Also, given the aforementioned cost of performing reactions, exhaustive campaigns with the sole purpose of populating reaction databases are not feasible from both economic and ecological standpoints. As a rough estimate, meaningful group compatibility/incompatibility and conditions screens would involve hundreds[20,28,36,88] if not thousands[89] of variants of each reaction type, scaling to billions of experiments if implemented for a representative subset of reactions.[86] Even in the age of high-throughput experimentation, such exhaustive efforts are not implementable. Again, we think it is time to go beyond the calls for just *more* data[90,91] and, instead, carefully consider *which* data is actually needed to improve the AI algorithms the most. In this regard, we believe that, even a few-days expert panel could be helpful in prioritizing the areas in which more data would matter: Examples could include established reactions on complex scaffolds or new desirable reactivity classes ("green" methods, multicomponent reactions, etc.).

One other point we wish to comment on is that, with the admixture of physical-organic knowledge, a significant portion of the otherwise required experimental data gathering could be avoided. As a textbook example, if an S_N2 reaction on a given scaffold is shown to proceed in poor yield with Br as a leaving group, then there is a low probability of improved reactivity for the Cl-containing substrate instead. Similarly, if a Friedel–Crafts-type acylation (Figure 5a) does not proceed on a trifluoromethylated arene, it is unlikely to proceed with even more electron-withdrawing groups. With such physical-organic considerations, even few experimental data points—delineating reactivity "thresholds" in terms of (stereo)electronic or steric factors—can help determine unlikely reactions and augment the data without spurious experimentation. We have used this kind of reasoning (supplemented by calculations detailed in the supporting information of ref 36) in our own expert-coding of reaction rules[58] and highly recommend its inclusion in AI driven efforts.

The Knowledge of Impossible Molecules. Let us now assume a scenario in which we have overcome the aforementioned problems and have finally curated a set of high-quality reaction rules. In Figure 5b, such rules, *R*, are applied in the retrosynthetic direction to 2-norbornanol, producing three synthons. Whereas ketone reduction from 2-norbornanone (**Y_1**) and oxidative hydroboration from 2-norbornene (**Y_2**) are perfectly viable transformations, the application of oxidative hydroboration to yield the bottom intermediate **Y_3** is problematic, as this synthon features a double bond at a bridgehead atom, creating excessive ring strain in the molecule and violating "Bredt's rule".[92] Naturally, such nonexisting molecules are not featured in reaction data sets and are not recognized as problematic by AI (see SI Section 5 for actual AI-generated examples). On the other hand, they could be readily recognized and eliminated by well-established molecular mechanics simulations plugged into the retrosynthetic algorithms. If such simulations are too time-consuming to accompany complex retrosynthetic analyses (in which millions of synthons need to be evaluated), the most prominent unstable motifs can be precalculated, tabulated based on existing knowledge (e.g., in evident cases of small rings with triple bonds, allenes, or *trans* ring-fused epoxides), or used to train AI-based stability classifiers. In previous works,[28,32,36] we used a knowledge-based tabulation to assemble a list of 500–1000 scaffolds spanning not only strained motifs but also those that readily decompose (e.g., geminal diols or α-haloalcohols). In the end, a relatively small curation effort translated into marked improvements in the quality of the computer-planned syntheses.

■ ALGORITHMS FOR SEARCHING SYNTHETIC NETWORKS

The eventual goal of synthesis planning is not only to correctly predict individual reaction steps but also to search the networks of synthetic possibilities—which can, again, significantly benefit from "anthropomorphic" improvements. At present, the most popular search algorithm in the field[35,93,94] is the Monte Carlo Tree Search (MCTS), which was introduced in 2006 as "bandit-based Monte Carlo planning", outperformed alternatives strategies like asynchronous dynamic programming,[95] and has since been used successfully in programs such as AlphaGO.[3] In the context of retrosynthesis, however, there has been little justification as why MCTS should be the algorithm of choice. In fact, a recent analysis showed[96] that MCTS using neural networks trained on reaction data showed comparable performance to a "vintage" A*-type algorithm[32] using a simplistic scoring that promotes disconnections of retrons into similar-sized synthons. As a matter of fact, there has not been a single example of MCTS successfully planning a longer route to a complex target (possible reasons have been discussed elsewhere[97]). What should make us revisit the network search problem is that there have been demonstrations of algorithms which are capable of planning such syntheses, but these are based on anthropomorphic rather than AI scoring functions, reflecting human heuristics such as key disconnections, disconnections of rings, or creation of sterocenters.[28,97] We do not claim that these algorithms are yet optimal, but their satisfactory performance in complex, natural-product-level tasks[28] certainly points to the benefits of mimicking human reasoning.

As a remarkable parallel, in the context of NLP, the success of ChatGPT has demonstrated the immense value of human feedback (reinforcement learning with human feedback, RLHF)[98] for enabling human-like reasoning. We envision that a related approach with expert feedback—to compare and rank the "quality" of different syntheses proposed by the machine—could significantly enhance the capabilities of retrosynthetic search algorithms. It should be noted, however, that many of these approaches have focused on RLHF tasks involving only dozens of options to choose from. Synthetic pathways, on the other hand, would involve multiple thousands of options (in the form of different reaction types and retrons), making this approach fascinating yet practically challenging in terms of the amount of sustained expert feedback to be harnessed. In fact, some primitive tools for providing feedback were incorporated in the early versions of our *Chematica* platform ("thumbs-up" voting and "envelope" comment buttons in Figures S18−S25 in ref 32), but they were not popular with the users and were ultimately removed, perhaps prematurely.

Complementary to this, we see enormous potential in large-scale, systematic surveys of how domain experts approach the synthesis problem and then develop advanced network search algorithms based on such guidance. Inspiring questions could include: How often do chemists choose disconnections based on Corey's rules?[21] How often are they inspired by similarity to other scaffolds? When considering a particular disconnection, what are the features that allow one to better anticipate downstream problems? How many steps ahead do experts think? Do they always work "just" in the retrosynthetic direction or have some key substrates or intermediates in mind?[99,100]

Searching for Synthetic Routes Like an Expert. One of these "design logic" aspects we wish to single out regards the depth to which synthesis planning should be performed. The readers may find it curious that existing retrosynthesis tree search algorithms lack the capacity to systematically plan multiple steps ahead, scoring all synthons after each disconnection "move". This may be sufficient in syntheses of simple targets, but it does not capture the farsighted, strategic thinking of human experts constructing routes to complex molecules. These experts plan many steps ahead and are trained to consider—sometimes almost intuitively—certain multistep *sequences* within which individual steps offer little immediate gain but set the scene for efficient downstream disconnection(s). As a case in point, the popular functional group interconversions (FGIs)[101] entail seemingly unproductive "moves" but, taken together, they serve to convert stability into more reactive groups or to adjust oxidation states. In a similar genre, two-step tactical combinations, TCs,[29] may prefer to initially complexify the synthon's structure (compared to its retron) but, by doing so, enable a subsequent disconnection offering a high degree of structural simplification. In previous work,[28] we showed that it is neither the sheer number and quality of reaction rules nor the nature of the scoring function choosing the "next move" (including scoring schemes incorporating Corey's rules[21]) but the inclusion of such sequences that gave the retrosynthetic searches the greatest performance boost, effectively allowing them to think up to five steps ahead and plan syntheses of complex targets.

Important for our discussion here is that these sequences may not be readily derived from the available reaction data sets, as reaction repositories are dominated by single-step reactions or short, simple sequences. As such, most of the possible TCs are not reported therein[29] and, for FGIs, the individual reaction types are not strongly correlated. One data-oriented solution would be to assemble and learn from a smaller data set focusing on longer "classics" of total syntheses; there are a few expert-curated repositories out there (Chemistry by Design,[102] Hans Reich's Collection,[103] SynArchive,[104] Organic Chemistry Portal[105]) but, unfortunately, these only provide images rather than computer-readable data. Alternatively, the problem could be addressed by a consultation with synthesis experts (and a human analysis of the synthesis literature); in fact, the above-mentioned performance boost[28] was due to only ∼100 FGIs and ∼1000 TCs carefully selected by human experts (the total number of important TCs is likely to be significantly larger, but 100−1000 already enabled major improvements).

Scoring Synthetic Options for Real-World Applicability. Finally, we want to touch on an issue that may not be essential for all synthetic plans but is of growing importance in planning syntheses that meet the demands of economy and green chemistry. In this context, reaction repositories may be informative of certain relevant parameters (e.g., unfavorable conditions such as cryogenic or very high temperatures), which should be deprioritized if other options are available; but for additional information, one needs to tap into the expertise of process chemistry or consider publicly available domain knowledge. As examples, the EPA List of Extremely Hazardous Substances[106] or the REACH regulation List of Substances of Very High Concern[107] can be connected to the network-search algorithms to flag and eliminate hazardous and toxic reagents;[88] published guidelines (e.g., the GSK criteria) can be helpful in suggesting "greener" replacements for reagents or

https://doi.org/10.1021/jacs.4c00338

solvents,[108−110] and process variables such as cumulative process mass intensity, cPMI, can help estimate reactions' economics under different purification methods. Some of this data may need to be preprocessed, and others may require careful categorization (e.g., cPMI values are different for different reaction classes, which require manual assignments to reaction rules[88]). Such efforts and the obtained process criteria, however, can provide significant added value into computational synthesis design. The algorithms guiding the retrosynthetic searches and scoring the solutions found should be customizable to prioritize the above "variables" to user-specified degrees, while also being aware of other aspects described in previous works (e.g., prices of substrates[28,32,36] and of entire routes,[111] diversity of solutions presented to the user,[111] the use of common intermediates when planning the syntheses of entire libraries of targets,[112] and possibly more).

HOW MUCH ADDITIONAL KNOWLEDGE IS NEEDED?

Taken together, the above considerations bring us to the following recommendations regarding the development of AI−knowledge hybrids for computer-aided synthesis planning in the coming decade(s).

Complex Syntheses. If the algorithm is concerned with the problem of designing tens-of-steps-long syntheses of complex, stereochemically defined natural product targets, our recommendation is to integrate as many knowledge-derived improvements as possible. In our own works, after some initial but unsatisfactory exploration of data-driven template extraction in the early 2010s, we embarked on a campaign to expert-curate a comprehensive collection of high-quality reaction rules, considering, for each individual reaction class, the template span, stereochemical consequences, the scope of admissible substituents, effects of distant substituents for specific scaffolds, and hundreds of potentially incompatible groups (400−600 per rule; for a detailed discussion, including the consideration of reaction mechanisms, see our previous study[58]). Often, it has been criticized that such an effort is not scalable since chemistry is expanding exponentially, and expert coding cannot match the pace at which new reactions are reported.[35] We ourselves had identified this exponential growth of organic chemistry,[86] which, however, only pertains to molecules and reaction examples rather than to reaction *types*, whose number has plateaued in recent years at ∼50,000 to 100,000 rules (cf. Figure 1c).[31] We acknowledge that the effort to encode such a collection is certainly significant (in our case, it took a decade or work for a dedicated team of computer-adept organic chemists) and goes along with further challenges: As an example, the encoded rule set inherently represents the knowledge at the time of encoding, inevitably introducing human biases. In addition, especially for "modern" reaction classes, their increasing complexity (in terms of altering the atom connectivity patterns from substrates to products[31]) and the inherently lower level of mechanistic understanding pose additional barriers to the unambiguous encoding of reaction rules. Yet, as of today, such an effort appears to be inevitable for obtaining usable and applicable synthetic routes to complex targets.[28−30,36] While we believe that, at their core, data-driven algorithms have the potential to contend with these expert encodings, the necessary data advancements, in terms of selection, quality, and quantity, are presumably decades down the road (see discussion above).

Beyond the expert encoding of reaction templates, we found that extensive collections of unfeasible motifs, anthropomorphic scoring functions, and search algorithms (as discussed earlier) are highly beneficial for obtaining realistic synthesis routes. This said, we expect data-driven and knowledge-enhanced approaches to be highly synergistic—e.g., for reaction classes with large numbers of examples or for reactions with substituent combinations too numerous to be enumerated within rule templates (e.g., combinations of multiple substituents dictating regio- or stereoselectivity).[61−64,68,69]

Synthesis of Drug-Like Molecules. If the algorithm is intended to provide more concise syntheses of smaller, drug-like molecules, we advocate a two-tier approach: The few hundred most prominent reaction types for medicinal chemistry[113] should be expert-curated (including substrate scope, incompatibilities, and importantly, stereochemistry), while the remaining reaction types can be automatically extracted but fine-tuned, especially for incompatibilities, using the approach of equivalent conditions. Nonsensical molecules and intermediates should be filtered rigorously by simulations or lists attached to the search routines. For shorter routes, the nature of the search algorithm is perhaps of lesser importance,[96] but for industrial relevance, the scoring functions should include process criteria.

Simple Target Molecules. For simple targets, the pharmaceutical criteria could be further relaxed, though the programs should still be improved compared to current AI engines. At minimum, if the objective is to generate ideas rather than realistic pathways, the emphasis should be on avoiding some critical blunders, and routines for eliminating impossible intermediates and for detecting at least the most striking incompatibilities should be in place.

CONCLUSION AND OUTLOOK

In summary, we have argued that the view of computerized retrosynthesis as aiming to "automatically learn organic synthesis from reaction databases" may be too narrow and not realistic in the near future, as existing reaction examples do not capture all nuances of organic syntheses. In fact, computer-aided retrosynthetic planning is not about the speed and novelty of algorithms but the ability to deliver efficient solutions to the target audience, that is, to synthetic organic chemists. In this respect, incorporation of diverse modalities of chemical knowledge is necessary to increase the quality of predictions and provide more appeal to the chemist users. It should be borne in mind that these chemists might simply not need computerized synthesis if it only provides plausible disconnection plans for simple molecules. In contrast, if the AI methods could rapidly supply multiple, chemically correct, and diverse pathways to more complex targets and, at the same time, could also rank them for additional process criteria, then such solutions could gain widespread popularity, as no single chemist can memorize the rules of synthesis planning, the tables of hazardous reagents and solvent, "greener" replacements for such chemicals, and many other parts of chemical knowledge. The knowledge base for such improvements is out there,[114] but not only in the reaction repositories, and often in the heads of synthesis experts with whom algorithm designers should form closer ties. Such joint efforts should benefit not only the designers but also the synthetic chemists; even if they do not commit to multiyear campaigns to deploy all-encompassing retrosynthesis platforms, they can engage their theory colleagues in developing AI models for some synthetic

https://doi.org/10.1021/jacs.4c00338

subproblems of particular interest (see examples in refs 13, 61−64, and 66−69), in the process also transplanting the AI know-how into the synthetic community.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c00338.

Further examples of inconsistencies in chemical reaction databases (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Frank Glorius − *Universität Münster, Organisch-Chemisches Institut, 48149 Münster, Germany;* orcid.org/0000-0002-0648-956X; Email: glorius@uni-muenster.de

Bartosz A. Grzybowski − *Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland; IBS Center for Algorithmic and Robotized Synthesis, CARS, Ulju-gun, Ulsan 689-798, South Korea; Department of Chemistry, UNIST, Ulju-gun, Ulsan 689-798, South Korea;* orcid.org/0000-0001-6613-4261; Email: nanogrzybowski@gmail.com

Authors

Felix Strieth-Kalthoff − *University of Toronto, Department of Chemistry and Department of Computer Science, Toronto, Ontario M5S 3H6, Canada; University of Toronto, Department of Computer Science, Toronto, Ontario M5S 3G4, Canada*

Sara Szymkuć − *Allchemy, Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland*

Karol Molga − *Allchemy, Highland, Indiana 46322, United States; Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw 01-224, Poland*

Alán Aspuru-Guzik − *University of Toronto, Department of Chemistry and Department of Computer Science, Toronto, Ontario M5S 3H6, Canada; University of Toronto, Department of Computer Science, Toronto, Ontario M5S 3G4, Canada; Vector Institute for Artificial Intelligence, Toronto, Ontario M5G 1M1, Canada; University of Toronto, Department of Chemical Engineering and Applied Chemistry, Toronto, Ontario M5S 3E5, Canada; University of Toronto, Department of Materials Science and Engineering, Toronto, Ontario M5S 3E4, Canada;* orcid.org/0000-0002-8277-4434

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c00338

Author Contributions

$^{\nabla}$F.S.-K., S.S., and K.M. contributed equally to this work.

Funding

F.S.-K. is a postdoctoral fellow in the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship Program, a program of Schmidt Futures. S.S., K.M., and A.A.-G. gratefully acknowledge funding from the Defense Advanced Research Projects Agency under the Accelerated Molecular Discovery Program (cooperative agreement no. HR00111920027 dated 1 August 2019). The content of the information presented in this work does not necessarily reflect the position or the policy of the US government. A.A.-G. acknowledges support from the Canada 150 Research Chairs Program as well as Anders G. Fröseth. F.G. acknowledges generous financial support from the Deutsche Forschungsgemeinschaft (SPP 2363, "Molecular Machine Learning"). Writing of the article by B.A.G. was supported by the Institute for Basic Science, Korea (project code IBS-R020-D1).

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS

AI, artificial intelligence; cPMI, cumulative process mass intensity; EPA, Environmental Protection Agency; MCTS, Monte Carlo tree search; NLP, natural language processing; ORD, open reaction database; REACH, registration, evaluation, authorization, and restriction of chemicals; SMARTS, SMILES arbitrary target specification; SMILES, simplified molecular input line entry system; TC, tactical combination; USPTO, United States patent and trademark office

■ REFERENCES

(1) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models Are Few-Shot Learners. *arXiv* **2020**, 2005.14165v4 (accessed 2023−06−09).

(2) OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S. et al. GPT-4 Technical Report. *arXiv* **2023**, 2303.08774 (accessed 2023−06−09).

(3) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529* (7587), 484−489.

(4) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play. *Science* **2018**, *362* (6419), 1140−1144.

(5) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60* (6), 84−90.

(6) Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, No. 7068349.

(7) Portugal, I.; Alencar, P.; Cowan, D. The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review. *Expert Syst. Appl.* **2018**, *97*, 205−227.

(8) Fawzi, A.; Balog, M.; Huang, A.; Hubert, T.; Romera-Paredes, B.; Barekatain, M.; Novikov, A.; Ruiz, F. J. R.; Schrittwieser, J.; Swirszcz, G.; Silver, D.; Hassabis, D.; Kohli, P. Discovering Faster Matrix Multiplication Algorithms with Reinforcement Learning. *Nature* **2022**, *610* (7930), 47−53.

(9) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(10) Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; Velankar, S.; Kleywegt, G. J.; Bateman, A.; Evans, R.; Pritzel, A.; Figurnov, M.; Ronneberger, O.; Bates, R.; Kohl, S. A. A.; Potapenko, A.; Ballard, A. J.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Clancy, E.; Reiman, D.; Petersen, S.; Senior, A. W.; Kavukcuoglu, K.; Birney, E.; Kohli, P.; Jumper, J.; Hassabis, D. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596* (7873), 590−596.

(11) Zhao, S.; Gensch, T.; Murray, B.; Niemeyer, Z. L.; Sigman, M. S.; Biscoe, M. R. Enantiodivergent Pd-Catalyzed C-C Bond Formation Enabled through Ligand Parameterization. *Science* **2018**, *362* (6415), 670−674.

(12) Ferreira, M. A. B.; De Jesus Silva, J.; Grosslight, S.; Fedorov, A.; Sigman, M. S.; Copéret, C. Noncovalent Interactions Drive the Efficiency of Molybdenum Imido Alkylidene Catalysts for Olefin Metathesis. *J. Am. Chem. Soc.* **2019**, *141* (27), 10788−10800.

(13) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of Higher-Selectivity Catalysts by Computer-Driven Workflow and Machine Learning. *Science* **2019**, *363* (6424), No. eaau5631.

(14) Dave, A.; Mitchell, J.; Kandasamy, K.; Wang, H.; Burke, S.; Paria, B.; Póczos, B.; Whitacre, J.; Viswanathan, V. Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning. *Cell Rep. Phys. Sci.* **2020**, *1* (12), 100264.

(15) S. V, S. S.; Law, J. N.; Tripp, C. E.; Duplyakin, D.; Skordilis, E.; Biagioni, D.; Paton, R. S.; St. John, P. C. Multi-Objective Goal-Directed Optimization of de Novo Stable Organic Radicals for Aqueous Redox Flow Batteries. *Nat. Mach. Intell.* **2022**, *4* (8), 720−730.

(16) Kim, S. C.; Oyakhire, S. T.; Athanitis, C.; Wang, J.; Zhang, Z.; Zhang, W.; Boyle, D. T.; Kim, M. S.; Yu, Z.; Gao, X.; Sogade, T.; Wu, E.; Qin, J.; Bao, Z.; Bent, S. F.; Cui, Y. Data-Driven Electrolyte Design for Lithium Metal Anodes. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (10), No. e2214357120.

(17) Moon, J.; Beker, W.; Siek, M.; Kim, J.; Lee, H. S.; Hyeon, T.; Grzybowski, B. A. Active Learning Guides Discovery of a Champion Four-Metal Perovskite Oxide for Oxygen Evolution Electrocatalysis. *Nat. Mater.* **2024**, *23*, 108−115.

(18) Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. A Brief Introduction to Chemical Reaction Optimization. *Chem. Rev.* **2023**, *123* (6), 3089−3126.

(19) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590* (7844), 89−96.

(20) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-Loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling. *Science* **2022**, *378* (6618), 399−405.

(21) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley, 1989.

(22) Corey, E. J. Robert Robinson Lecture. Retrosynthetic Thinking—Essentials and Examples. *Chem. Soc. Rev.* **1988**, *17*, 111−133.

(23) de Almeida, A. F.; Moreira, R.; Rodrigues, T. Synthetic Organic Chemistry Driven by Artificial Intelligence. *Nat. Rev. Chem.* **2019**, *3* (10), 589−604.

(24) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154−6168.

(25) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine Intelligence for Chemical Reaction Space. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), No. e1604.

(26) Kreutter, D.; Reymond, J.-L. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *ChemRxiv* **2022** DOI: 10.26434/chemrxiv-2022-8khth (accessed 2023−06−09).

(27) Lemonick, S. A New Database for Machine-Learning Research. *Chem. Eng. News* 2021; https://cen.acs.org/physical-chemistry/computational-chemistry/new-database-machine-learning-research/99/web/2021/11.

(28) Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational Planning of the Synthesis of Complex Natural Products. *Nature* **2020**, *588* (7836), 83−88.

(29) Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem.* **2020**, *6* (1), 280−293.

(30) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-Aided Key Step Generation in Alkaloid Total Synthesis. *Science* **2023**, *379* (6631), 453−457.

(31) Szymkuć, S.; Badowski, T.; Grzybowski, B. A. Is Organic Chemistry Really Growing Exponentially? *Angew. Chem., Int. Ed.* **2021**, *60* (50), 26226−26232.

(32) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904−5937.

(33) Greene, A. E.; Serra, A. A.; Barreiro, E. J.; Costa, P. R. R. Expeditious, Stereocontrolled Syntheses of Racemic and Natural Brasilenol through Intramolecular Asymmetry Transfer. Absolute Stereochemistry of Brasilenol. *J. Org. Chem.* **1987**, *52* (6), 1169−1170.

(34) *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*; https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

(35) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604−610.

(36) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4* (3), 522−532.

(37) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Thesis, University of Cambridge, 2012; DOI: 10.17863/CAM.16293.

(38) *Reaxys*; August 2022 release; Elsevier B.V.: Amsterdam (NL), 2022; accessed at https://www.reaxys.com.

(39) *SciFinder*; August 23, 2022 release; Chemical Abstracts Service: Columbus, OH, USA, 2022; accessed at https://scifinder.cas.org.

(40) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725−732.

(41) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem.—Eur. J.* **2017**, *23* (25), 5966−5971.

(42) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434−443.

(43) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3* (10), 1103−1113.

(44) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9* (28), 6091−6098.

(45) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572−1583.

(46) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *arXiv* **2020**, 2003.12725 (accessed 2023−06−09).

(47) Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; Huang, J. RetroXpert: Decompose Retrosynthesis Prediction Like A Chemist. *arXiv* **2020**, 2011.02893 (accessed 2023−06−09).

(48) Somnath, V. R.; Bunne, C.; Coley, C.; Krause, A.; Barzilay, R. Learning Graph Models for Retrosynthesis Prediction. *arXiv* **2020**, 2006.07038 (accessed 2023−06−09).

(49) Gimadiev, T. R.; Lin, A.; Afonina, V. A.; Batyrshin, D.; Nugmanov, R. I.; Akhmetshin, T.; Sidorov, P.; Duybankova, N.; Verhoeven, J.; Wegner, J.; Ceulemans, H.; Gedich, A.; Madzhidov, T. I.; Varnek, A. Reaction Data Curation I: Chemical Structures and Transformations Standardization. *Mol. Inf.* **2021**, *40* (12), 2100119.

(50) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted Noise Reduction of Chemical Reaction Datasets. *Nat. Mach. Intell.* **2021**, *3* (6), 485−494.

(51) Hartrampf, N.; Winter, N.; Pupo, G.; Stoltz, B. M.; Trauner, D. Total Synthesis of the Norhasubanan Alkaloid Stephadiamine. *J. Am. Chem. Soc.* **2018**, *140* (28), 8675−8680.

(52) Mi, Y.; Schreiber, J. V.; Corey, E. J. Total Synthesis of (+)-*α*-Onocerin in Four Steps via Four-Component Coupling and Tetracyclization Steps. *J. Am. Chem. Soc.* **2002**, *124* (38), 11290−11291.

(53) Donthiri, R. R.; Pappula, V.; Reddy, N. N. K.; Bairagi, D.; Adimurthy, S. Copper-Catalyzed C-H Functionalization of Pyridines and Isoquinolines with Vinyl Azides: Synthesis of Imidazo Heterocycles. *J. Org. Chem.* **2014**, *79* (22), 11277−11284.

(54) Yu, J.; Jin, Y.; Zhang, H.; Yang, X.; Fu, H. Copper-Catalyzed Aerobic Oxidative C-H Functionalization of Substituted Pyridines: Synthesis of Imidazopyridine Derivatives. *Chem.—Eur. J.* **2013**, *19* (49), 16804−16808.

(55) Talbot, E. P. A.; Richardson, M.; McKenna, J. M.; Toste, F. D. Gold-Catalyzed Redox Synthesis of Imidazo[1,2-a]Pyridines Using Pyridine N-Oxide and Alkynes. *Adv. Synth. Catal.* **2014**, *356* (4), 687−691.

(56) Jaworski, W.; Szymkuć, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kaźmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic Mapping of Atoms across Both Simple and Complex Chemical Reactions. *Nat. Commun.* **2019**, *10* (1), 1434.

(57) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7* (15), No. eabe4166.

(58) Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski, B. A. The Logic of Translating Chemical Knowledge into Machine-Processable Forms: A Modern Playground for Physical-Organic Chemistry. *React. Chem. Eng.* **2019**, *4* (9), 1506−1521.

(59) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chem. Sci.* **2020**, *11* (1), 154−168.

(60) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59* (6), 2529−2537.

(61) Moskal, M.; Beker, W.; Szymkuć, S.; Grzybowski, B. A. Scaffold-Directed Face Selectivity Machine-Learned from Vectors of Non-Covalent Interactions. *Angew. Chem., Int. Ed.* **2021**, *60* (28), 15230−15235.

(62) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019**, *58* (14), 4515−4519.

(63) Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. Predicting Regioselectivity in Radical C-H Functionalization of Heterocycles through Machine Learning. *Angew. Chem., Int. Ed.* **2020**, *59* (32), 13253−13259.

(64) Ree, N.; Göller, A. H.; Jensen, J. H. RegioML: Predicting the Regioselectivity of Electrophilic Aromatic Substitution Reactions Using Machine Learning. *Digit. Discovery* **2022**, *1* (2), 108−114.

(65) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186−190.

(66) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature* **2018**, *559* (7714), 377−381.

(67) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem* **2020**, *6* (6), 1379−1390.

(68) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and on-the-Fly Quantum Mechanical Descriptors. *Chem. Sci.* **2021**, *12* (6), 2198−2208.

(69) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and Accurate Prediction of the Regioselectivity of Electrophilic Aromatic Substitution Reactions. *Chem. Sci.* **2018**, *9* (3), 660−665.

(70) Bajpai, R.; Yang, F.; Curran, D. P. On the Structure of the Phytophthora A1 Mating Hormone: Synthesis and Comparison of Four Candidate Stereoisomers. *Tetrahedron Lett.* **2007**, *48* (45), 7965−7968.

(71) Benito Iglesias, D.; Herrero Teijón, P.; Rubio González, R.; Fernández-Mateos, A. Synthesis of trans-*β*-Elemene. *Eur. J. Org. Chem.* **2018**, *2018* (35), 4926−4932.

(72) Schörgenhumer, J.; Zimmermann, A.; Waser, M. SNS-Ligands for Ru-Catalyzed Homogeneous Hydrogenation and Dehydrogenation Reactions. *Org. Process Res. Dev.* **2018**, *22* (7), 862−870.

(73) Thakur, V. V.; Nikalje, M. D.; Sudalai, A. Enantioselective Synthesis of (R)-(−)-Baclofen via Ru(II)-BINAP Catalyzed Asymmetric Hydrogenation. *Tetrahedron Asymmetry* **2003**, *14* (5), 581−586.

(74) Schnapperelle, I.; Hummel, W.; Gröger, H. Formal Asymmetric Hydration of Non-Activated Alkenes in Aqueous Medium through a "Chemoenzymatic Catalytic System. *Chem.—Eur. J.* **2012**, *18* (4), 1073−1076.

(75) He, C.; Zhang, X.; Huang, R.; Pan, J.; Li, J.; Ling, X.; Xiong, Y.; Zhu, X. Synthesis of Structurally Diverse Diarylketones through the Diarylmethyl Sp3 CH Oxidation. *Tetrahedron Lett.* **2014**, *55* (32), 4458−4462.

(76) Piovan, L.; Wu, L.; Zhang, Z.-Y.; Andrade, L. H. Hypervalent Organochalcogenanes as Inhibitors of Protein Tyrosine Phosphatases. *Org. Biomol. Chem.* **2011**, *9* (5), 1347−1351.

(77) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E. P.; Grzybowski, B. A.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7* (1), 3582.

(78) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angew. Chem., Int. Ed.* **2022**, *61* (29), No. e202204647.

(79) Pitzer, L.; Schäfers, F.; Glorius, F. Rapid Assessment of the Reaction-Condition-Based Sensitivity of Chemical Transformations. *Angew. Chem., Int. Ed.* **2019**, *58* (25), 8572−8576.

(80) Collins, K. D.; Glorius, F. A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nat. Chem.* **2013**, *5* (7), 597−601.

(81) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143* (45), 18820−18826.

(82) *Open Reaction Database*; https://open-reaction-database.org/client/browse.

(83) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52* (7), 1745−1756.

(84) Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E. A.; Webster, J.; Gillet, V. J. Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature. *J. Chem. Inf. Model.* **2019**, *59* (10), 4167−4187.

(85) Jablonka, K. M.; Patiny, L.; Smit, B. Making the Collective Knowledge of Chemistry Open and Machine Actionable. *Nat. Chem.* **2022**, *14* (4), 365−376.

(86) Fialkowski, M.; Bishop, K. J. M.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and Evolution of Organic Chemistry. *Angew. Chem., Int. Ed.* **2005**, *44* (44), 7263−7269.

(87) Bishop, K. J. M.; Klajn, R.; Grzybowski, B. A. The Core and Most Useful Molecules in Organic Chemistry. *Angew. Chem., Int. Ed.* **2006**, *45* (32), 5348−5354.

(88) Wołos, A.; Koszelewski, D.; Roszak, R.; Szymkuć, S.; Moskal, M.; Ostaszewski, R.; Herrera, B. T.; Maier, J. M.; Brezicki, G.; Samuel, J.; Lummiss, J. A. M.; McQuade, D. T.; Rogers, L.; Grzybowski, B. A. Computer-Designed Repurposing of Chemical Wastes into Drugs. *Nature* **2022**, *604* (7907), 668−676.

(89) Cernijenko, A.; Risgaard, R.; Baran, P. S. 11-Step Total Synthesis of (−)-Maoecrystal V. *J. Am. Chem. Soc.* **2016**, *138* (30), 9425−9428.

(90) Gomolión-Bel, F. Chemists Debate Machine Learning's Future in Synthesis Planning and Ask for Open Data. *Chem. Eng. News* **2022**, *100* (18), 1.

(91) Baldi, P. Call for a Public Open Database of All Chemical Reactions. *J. Chem. Inf. Model.* **2022**, *62* (9), 2011−2014.

(92) Bredt, J. Über Sterische Hinderung in Brückenringen (Bredtsche Regel) Und Über Die Meso-Trans-Stellung in Kondensierten Ringsystemen Des Hexamethylens. *Liebigs Ann. Chem.* **1924**, *437* (1), 1−13.

(93) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), No. eaax1566.

(94) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning. *J. Chemoinf.* **2020**, *12* (1), 70.

(95) Barto, A. G.; Bradtke, S. J.; Singh, S. S. *Real-Time Learning and Control Using Asynchronous Dynamic Programming*; COINS technical report; University of Massachusetts at Amherst, Department of Computer and Information Science, 1991.

(96) Genheden, S.; Bjerrum, E. PaRoutes: Towards a Framework for Benchmarking Retrosynthesis Route Predictions. *Digit. Discovery* **2022**, *1* (4), 527−539.

(97) Grzybowski, B. A.; Badowski, T.; Molga, K.; Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-Level Computerized Synthesis Planning. *WIREs Comput. Mol. Sci.* **2023**, *13* (1), No. e1630.

(98) Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; Amodei, D. Deep Reinforcement Learning from Human Preferences. *arXiv* **2017**, 1706.03741 (accessed 2023−06−09).

(99) Hanessian, S.; Franco, J.; Larouche, B. The Psychobiological Basis of Heuristic Synthesis Planning - Man, Machine and the Chiron Approach. *Pure Appl. Chem.* **1990**, *62* (10), 1887−1910.

(100) Hanessian, S. Man, Machine and Visual Imagery in Strategic Synthesis Planning: Computer-Perceived Precursors for Drug Candidates. *Curr. Opin. Drug Discovery Devel.* **2005**, *8* (6), 798−819.

(101) Serratosa, F.; Xicart, J. *Organic Chemistry in Action: The Design of Organic Synthesis*; Materials Science Monographs; Elsevier, 1990.

(102) *Chemistry By Design*; https://chemistrybydesign.oia.arizona.edu/ (accessed 2023−06−09).

(103) *Hans Reich Collection of Total Syntheses*; https://organicchemistrydata.org/hansreich/resources/syntheses/ (accessed 2023−06−09).

(104) *SynArchive - The Organic Synthesis Database*; https://synarchive.com/ (accessed 2023−06−09).

(105) *Organic Chemistry Portal*; https://www.organic-chemistry.org/ (accessed 2023−06−09).

(106) *Code of Federal Regulations (eCFR)*; https://www.ecfr.gov/current/title-40/chapter-I/subchapter-J/part-355 (accessed 2023−06−09).

(107) *Candidate List of substances of very high concern for Authorisation - ECHA*; https://echa.europa.eu/candidate-list-table (accessed 2023−06−09).

(108) Adams, J. P.; Alder, C. M.; Andrews, I.; Bullion, A. M.; Campbell-Crawford, M.; Darcy, M. G.; Hayler, J. D.; Henderson, R. K.; Oare, C. A.; Pendrak, I.; Redman, A. M.; Shuster, L. E.; Sneddon, H. F.; Walker, M. D. Development of GSK's Reagent Guides - Embedding Sustainability into Reagent Selection. *Green Chem.* **2013**, *15* (6), 1542−1549.

(109) Henderson, R. K.; Hill, A. P.; Redman, A. M.; Sneddon, H. F. Development of GSK's Acid and Base Selection Guides. *Green Chem.* **2015**, *17* (2), 945−949.

(110) Henderson, R. K.; Jiménez-González, C.; Constable, D. J. C.; Alston, S. R.; Inglis, G. G. A.; Fisher, G.; Sherwood, J.; Binks, S. P.; Curzons, A. D. Expanding GSK's Solvent Selection Guide - Embedding Sustainability into Solvent Selection Starting at Medicinal Chemistry. *Green Chem.* **2011**, *13* (4), 854−862.

(111) Badowski, T.; Molga, K.; Grzybowski, B. A. Selection of Cost-Effective yet Chemically Diverse Pathways from the Networks of Computer-Generated Retrosynthetic Plans. *Chem. Sci.* **2019**, *10* (17), 4640−4651.

(112) Molga, K.; Dittwald, P.; Grzybowski, B. A. Computational Design of Syntheses Leading to Compound Libraries or Isotopically Labelled Targets. *Chem. Sci.* **2019**, *10* (40), 9219−9232.

(113) Brown, D. G.; Boström, J. Analysis of Past and Present Synthetic Methodologies on Medicinal Chemistry: Where Have All the New Reactions Gone? *J. Med. Chem.* **2016**, *59* (10), 4443−4458.

(114) Schrader, M. L.; Schäfer, F. R.; Schäfers, F.; Glorius, F. Bridging the information gap in organic chemical reactions. *Nat. Chem.* **2024**, *16*, 491−498.