Reaction Chemistry
& Engineering

The logic of translating chemical knowledge into machine-processable forms: A modern playground for physical-organic chemistry

Journal:	*Reaction Chemistry & Engineering*
Manuscript ID	RE-PER-02-2019-000076.R2
Article Type:	Perspective
Date Submitted by the Author:	06-Jun-2019
Complete List of Authors:	Molga, Karol; Polish Academy of Sciences, Institute of Organic Chemistry Gajewska, Ewa; Polska Akademia Nauk Instytut Chemii Organicznej Szymkuc, Sara; Polska Akademia Nauk Instytut Chemii Organicznej Grzybowski, Bartosz; Ulsan National Instiutte of Science and Technology, IBS Center for Soft and Living Matter and Department of Chemistry; Polish Academy of Sciences, Institute of Organic Chemistry

The logic of translating chemical knowledge into machine-processable forms: A modern playground for physical-organic chemistry.

Karol Molga[1], *Ewa P. Gajewska*[1], *Sara Szymkuć*[1], *Bartosz A. Grzybowski*[1,2,]*

[1] Institute of Organic Chemistry, Polish Academy of Sciences, ul. Kasprzaka 44/52, Warsaw 01-224, Poland

[2] IBS Center for Soft and Living Matter and Department of Chemistry, UNIST, 50, UNIST-gil, Eonyang-eup, Ulju-gun, Ulsan, 689-798, South Korea

*Correspondence to: nanogrzybowski@gmail.com

Abstract.

Recent years have brought renewed interest – and tremendous progress – in computer-assisted synthetic planning. Although the vast majority of the proposed solutions rely on individual reaction rules that are subsequently combined into full synthetic sequences, surprisingly little attention has been paid in the literature to how these rules should be encoded to ensure chemical correctness and applicability to syntheses which organic-synthetic chemists would find of practical interest. This is a dangerous omission since any AI algorithms for synthetic design will be only as good as the basic synthetic "moves" underlying them. This Perspective aims to fill this gap and outline the logic that should be followed when translating organic-synthetic knowledge into reaction rules understandable to the machine. The process entails numerous

considerations ranging from careful study of reaction mechanism, to molecular and quantum mechanics, to AI routines. In this way, the machine is not only taught the reaction “cores” but is also able to account for various effects that, historically, have been studied and quantified by physical-organic chemists. While physical organic chemistry might no longer be at the forefront of modern chemical research, we suggest that it can find a new and useful embodiment though a junction with computerized synthetic planning and related AI methods.

Introduction.

The idea of computers designing synthetic routes and/or predicting the outcomes of chemical reactions dates back to 1960s[1]. The pioneering efforts of eminent chemists such as E.J Corey (LHASA program[2,3]), C. Djerassii[4], I. Ugi[5], and J.B Hendrickson[6] were, in many ways, ahead of their time though, for various reasons (narrated in [7]) did not become widely adopted by the synthetic community. Fortunately, recent years have witnesses a revival of interest in this interesting and potentially impactful area of chemical research and several platforms for organic-synthetic analyses have emerged. Our own effort in this area – starting with the 2005 paper on the analysis of large chemical networks[8] – has culminated in the development of Chematica retrosynthesis platform[7,9] that has recently been commercialized by Sigma-Aldrich (under the trade name of Synthia) and validated experimentally via execution of several synthetic routes designed by the machine[9]. Other notable efforts have been the ARChem engine[10] (to be incorporated into SciFinder), the ASKCOS tool[11] from MIT, or Segler and Waller's software based on Monte-Carlo searches and described in reference [12]. While the ways in which the algorithms underlying these engines come up with complete synthetic pathways differ in many substantial ways, the common component they all share are the chemical rules ("transforms") describing individual chemical reactions. In fact, the quality of these rules is absolutely crucial since synthetic pathways are very "unforgiving" to errors in individual steps – if our individual rules are, say, 80% correct, the chance that a n-step synthesis is going to be error-free and executable in the laboratory is only 0.8^n (i.e., only 10% for a ten-step synthesis). It is a straightforward but essential conclusion that any synthesis planning software will be only as good as the reaction rules it incorporates. Yet, despite such considerations, the articles on computer-aided synthesis focus mostly on the (often quite advanced) AI routines for concatenating individual steps into pathways[8,9,11-14] while, at the same time, spend little time on how the individual reaction rules are (or should be) coded. We think that given the progress and interest – or even hype[15] – surrounding this re-emerging field of chemical research, the time is ripe to systematize the approaches to and the logic of reaction rule coding. Accordingly, in this Perspective, we will strive to provide an overview of these aspects for reaction rules (i) machine-extracted from repositories of published reactions (e.g., Reaxys, SciFinder, InfoChem, USPTO databases[16-19]), or (ii) coded by expert chemists based on the underlying reaction mechanisms. One of the main conclusions of our survey is that while machine extraction approach is very rapid, the chemical correctness of the rules it yields is lower than those coded by experts – this difference becomes all the more pronounced as one becomes interested in the synthesis of more complex targets, for whose synthesis the human experts might actually benefit from computer's help. Another message this article is intended to convey is that translation of chemical knowledge into machine-readable rules is a very nuanced process, in which one has to consider not only the scope of admissible substituents and incompatible groups, but also a range of physical-organic effects including electron densities, steric bulk, molecular strain, and more. In calculating such quantities by quantum mechanical, QM, or molecular mechanics, MM, methods, an added challenge is to make them very fast and compatible with automated synthetic planning, during which the numbers of candidate intermediates considered can be very large (hundreds of thousands[7,9]). Finally, we also aim to illustrate which types of rules can be fine-tuned and improved by AI models, and what precautions should be taken for such models to be robust and meaningful. Altogether, we suggest that rule coding combining mechanistic considerations, elements of QM and MM modelling, and AI can be viewed as a modern embodiment of classic physical-organic chemistry[20-21], possibly reviving interest in this somewhat forgotten but essential field of organic-chemical research.

Defining reaction core and its environment.

To begin with, a reaction rule must specify atoms that are changing their environments and/or bonding patterns during a chemical reaction. For example, during addition of an amine to an alkene, a bond is formed between an incoming amine nucleophile and a carbon atom, the double C=C bond becomes a single one, and another C-H bond is formed (**Figure 1a**). In a Diels-Alder cycloaddition, the double bonds of the diene and dienophile rearrange to form a cyclohexene ring (**Figure 1b**). Of course, such a "local" description is chemically incomplete since the flanking atoms are generally also important or even all-important. In the first example, the reaction can proceed only when there is an electron withdrawing group (e.g., an ester or a ketone) activating the alkene for the nucleophilic attack (**Figure 1c**); for the Diels-Alder reaction, the electron-withdrawing/accepting ability of the substituents on the diene and dienophile will dictate regio-, site, or diastereoselectivity (**Figure 1d**). To account for such effects, the reaction rules need to be extended to include admissible, nearby substituents at various positions. Such extensions can be to different "radii" – to within atoms just flanking the reaction core (radius = 1), to within two atoms from reaction core (radius = 2), etc. (**Figure 1e**). In addition, chemical reactions might be prohibited by conflicting groups present anywhere in the molecule – in our example of double-bond amination, an iodide and a thiol present in reaction partners introduce an incompatibility because allylic iodide is a better electrophile than α,β-unsaturated alkene and thiol is a more reactive nucleophile than amine (**Figure 1f)**.

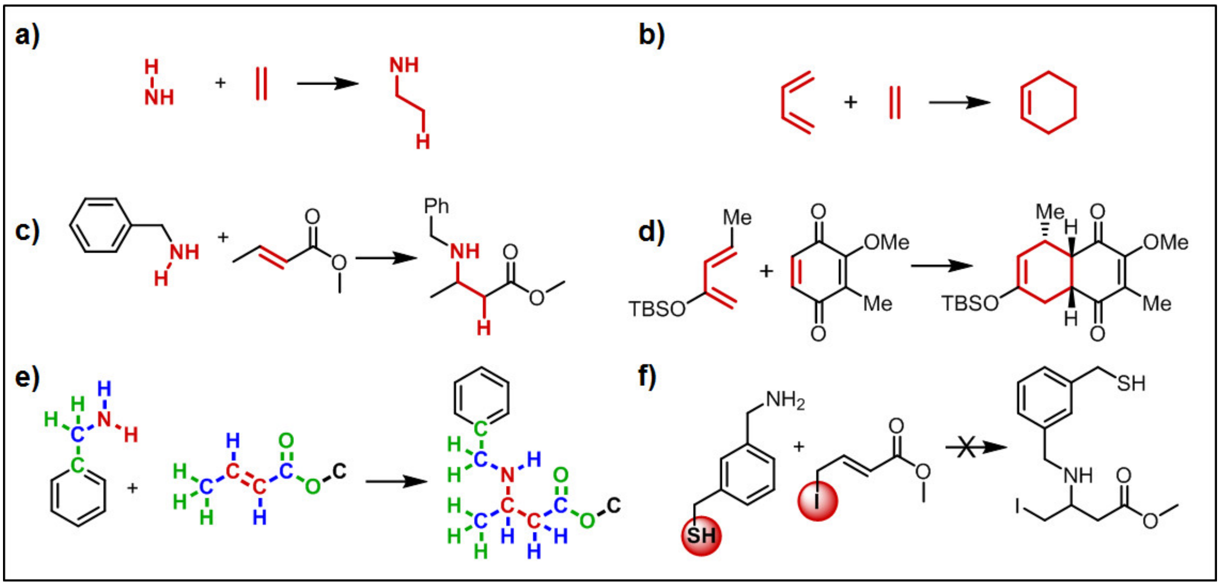


Figure 1. Defining reaction core. Atoms changing their environments and bonds modified during **a)** addition of an amine to alkene and **b)** Diels-Alder cycloaddition. Literature precedents[22,23] used for extraction of these cores are shown in **c)** and **d)**, respectively. **e)** Expansion of the reaction core. The core atoms changing their bonding patterns are coloured in red. Inclusion of the nearest-neighbour atoms (radius = 1, blue) and next-nearest-neighbour atoms (radius = 2; green) increases the accuracy of the reaction rule and begins to cover important substituents – here, the presence of an electron withdrawing group attached to the more distant end of C=C bond. **f**) Cross-reactive groups (here, allyl iodide and thiol) present anywhere in the substrates must also be considered as they will interfere with a desired reaction outcome.

There are few conceivable ways of capturing such intricacies of chemical reactivity into a machine readable format. Over the past years, two major approaches have emerged (1) Extraction of reaction cores from databases such as Reaxys[16] or USPTO[19] and fine-tuning their substituent scope/applicability based on the synthetic latitude of the examples found; or (2) Coding the rules manually, by chemists taking into account pertinent mechanistic considerations. In the following, we will focus on these two core-based approaches and will not discuss methods in which reactivity is predicted based on the AI-trained scores for atom pairs[24] or linguistic sequence-to-sequence models[25-26]. The performance of these approaches is described in the SI to ref [27] and also ref [28] – in addition, examples included in the next section and in the Supplementary **Figure S1** evidence that such models produce unacceptably large fraction of chemically problematic predictions.

Limitations of Automatic rule extraction.

The advantage of machine-extracting rules from reaction databases is the speed of the method – in fact, with adequate computational power, an entire Reaxys collection can be scripted within few days. In the end, one ends up with tens to hundreds of thousands of reaction rules with the specific number depending on the source database and the radius around the reaction core. For instance, Segler and Waller[12] extracted transformation rules from 12.4 million single-step reactions from the proprietary Reaxys repository. For the transformations having at least 50 literature precedents and encompassing the core atoms plus the radius =1 environments, they extracted ~17,000 rules; for the relaxed requirements of three literature precedents and only the core atoms, the number was ~300,000. In a recent study[29] by Watson *et al.*, the authors extracted from the publicly available United States Patent and Trade Office, USPTO, database (close to 2 million entries) a total of 83,942 unique transforms for radius 0, 180,862 for radius 1 and 325,873 for radius 2.

Of course, not all machine-extracted rules derive from similar numbers of literature precedents – for popular reaction classes, the number of such precedents can be up to hundreds of thousands per one extracted rule (e.g., for generalized cores of Wittig olefination, Suzuki-

Miyaura coupling, or formation of an amide from carboxylic acid and amine substrates); for more specialized chemistries, however, there might be just few examples in the literature (e.g., in Reaxys, there are only ~20 examples of stereospecific C-H insertions of carbenes yielding tertiary alcohols; **Figure 2a** and [30]). This is significant for our discussion since any machine-learning of reaction-rule applicability is possible only for the popular reaction classes. For the ones with just few examples, it is impossible to automatically extract meaningful statistics delineating the scope of the reaction, i.e., which substituents are admissible and which are not, which groups are conflicting with the reaction core, etc. This is, in fact, a serious flaw if one wishes to teach the machine some more advanced syntheses – for instance, the number of reported examples related to the stereospecific reduction of tertiary alcohols used in total syntheses of curcumene[31,32] and himachalene[32] is very limited and does not exceed 10. Importantly, all of these reactions were performed on simple substrates (**Figure 2b**) complicating prediction of the extracted transformations' applicability to more advanced intermediates, especially bearing potentially cross-reactive functional groups. In another example, anionic 4+2 cycloaddition in **Figure 2c** is represented by only 30 precedents but was the key step in the synthesis of murrayafoline-A[33], olivine[34], clausamine E[35] and claulansine D[36]. This annulation occurs via stepwise conjugate addition of a lithiated lactone, subsequent cyclisation forming [2.2.1] bicycloheptane, and elimination-tautomerisation (**Figure 2c**) leading to the fused aromatic system. Without mechanistic understanding of this complicated sequence, it is virtually impossible to properly define the necessary substituents, e.g., the presence of electron withdrawing group necessary for the conjugate addition to proceed.

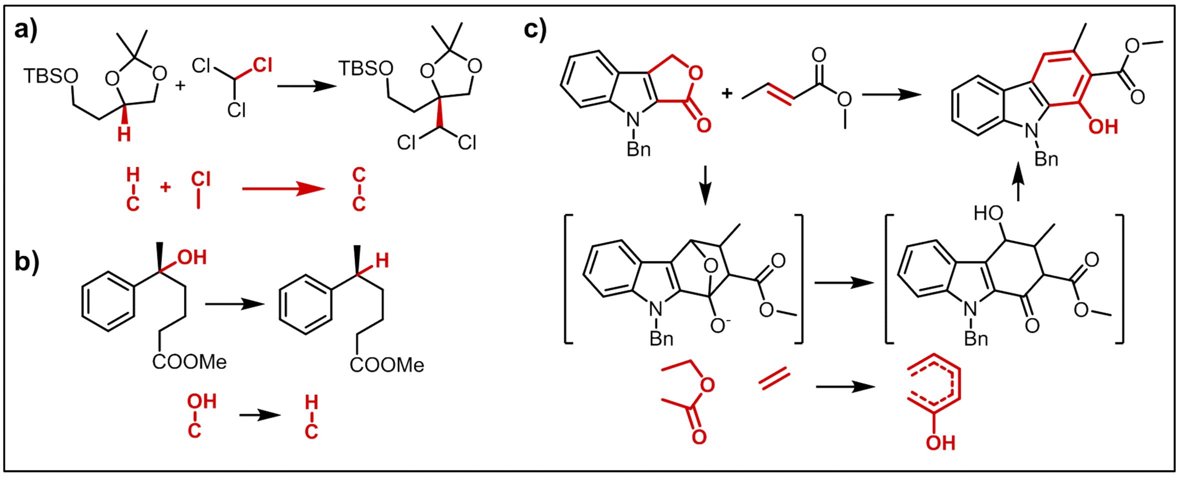


Figure 2. Examples of rare but useful reactions. **a)** Base-induced dichloromethylation of secondary alcohols (only 20 literature examples). **b)** Stereospecific deoxygenation of tertiary alcohols (only 10 examples). **c)** Anionic 4+2 cycloaddition forming a fused aromatic system (only 30 examples). Reaction cores without any environments are coloured in red.

Even for the popular reaction classes one has to be careful. Although widening the core to the radius 1 or 2 environments generally increases accuracy of the transforms, it also limits their scope, makes them very much case-specific and non-generalizable (cf. previous paragraph), and still does not treat adequately many chemical details. This point is corroborated by examples in **Figure 3** describing popular aldol condensation between an ester and an aldehyde. Limiting the reaction rule to only the core of changing bonds/atoms (b) or supplementing this core (c) with immediate, radius = 1 neighbouring atoms allows for erroneous results such as those shown in the rightmost column of **Figures 3b,c**. Even if radius = 2 is applied, the reaction template is incomplete as it allows for the presence of highly acidic H’s interfering with expected reaction outcome (**Figure 3d**).

The automated approach is also ill-suited to account for truly long-distance effects that might come from groups many bonds away – for instance, in **Figure 4a**, the reduction of a double bond is directed by a hydroxyl group[37] located four atoms away, while the regioselectivity of dihydroxylation of a triene in **Figure 4b** is controlled by an electron withdrawing group [38] five atoms away. We observe that extending the environments is not an answer here because in this way one would soon end up with the number of "rules" approaching the number of literature precedents in the database one is learning the rules from – the only way around this problem is to have an expert organic chemist determine in which cases narrower cores are adequate and in which they must be extended to capture distant substituent or scaffold effects.

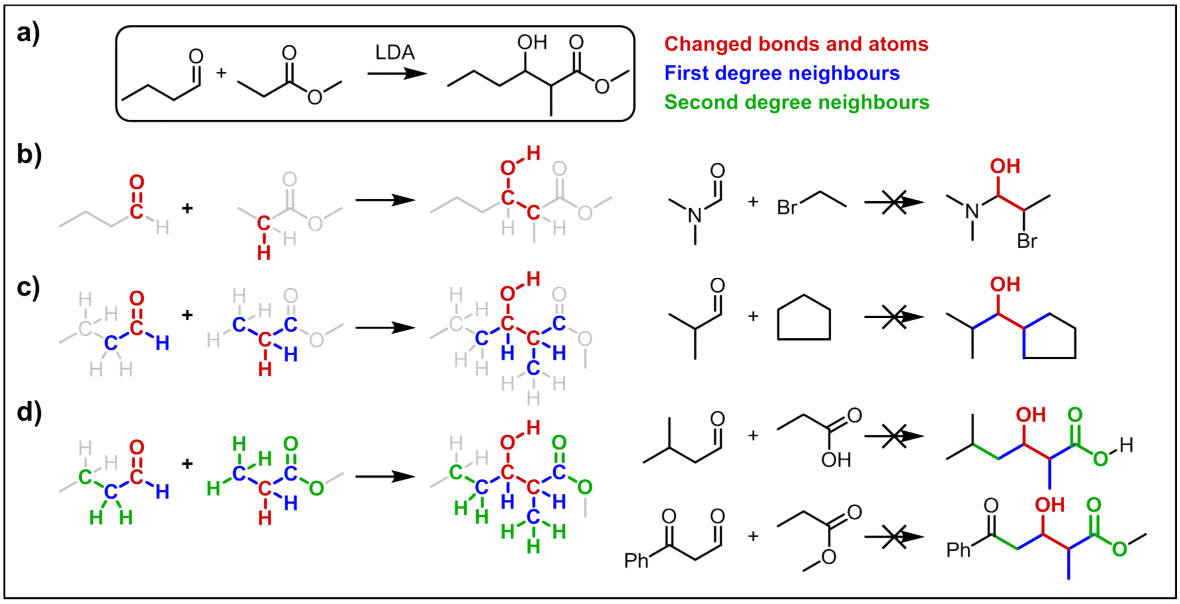


Figure 3. Defining proper span of reaction rules. **a)** Base-induced aldol condensation between an ester and an aldehyde. **b)** Limiting the reaction template to the bare reaction core (red) allows the rule to capture nonsensical results such as the one shown on the right. **c)** Inclusion of first-degree neighbours (blue) still does not eliminate faulty predictions – here, addition

of cyclopentane to aldehyde is still captured though it is chemically nonsensical for this reaction type. **d)** Extension to next-nearest-neighbour, radius = 2 environments (green) limits the number of nonsensical predictions but even this extended rule allows for the presence of acidic H's from carboxylic acid (top) or benzoyl acetaldehyde (bottom) interfering with expected reaction outcome.

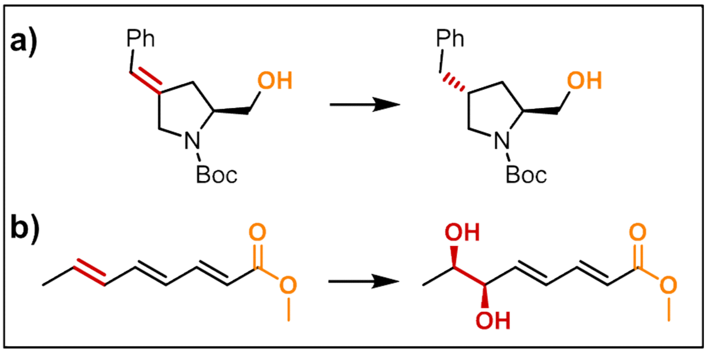


Figure 4. Long-range control of organic reactions. **a)** Stereoselective reduction of an alkene with Crabtree's catalyst[37] is controlled by a remote hydroxyl group; **b)** Sharpless dihydroxylation of polyene[38] is controlled by a remote electron withdrawing group. The reacting groups are coloured in red and the controlling, distant functionalities, in orange.

Another relevant issue is the treatment of incompatible groups – the generic automated approach devised so far[11,12,14,39] is to identify which groups are surviving a transformation of interest and deem them as "compatible"; those groups that are not seen in reactions corresponding to a given rule (or are "destroyed" in such transforms), are deemed incompatible. This heuristics is rather crude since the fact that there are no reactions in which a particular functional group is absent does not mean this group is generally prohibited in such a reaction –

perhaps no one just tried a particular group or groups' combination, which is all the more likely for more specialized transforms having less literature precedents. Second, there are numerous cases in which statistical approach fails because the same group might be incompatible with a given reaction rule in some targets, but compatible in others. As a case in point, consider examples shown in **Figure 5** for which automatic assignment of certain functional groups as compatible is chemically incorrect. Specifically, for methyl ester reduction performed *en route* to jujuboside saponin[40] (**Figure 5a**, top), we might "learn" that geminal methyl diester can survive the mono-ester's reduction; however, for the same reaction conditions applied to an intermediate in Plumisclerin A synthesis[41] (**Figure 5a**, bottom), we would learn the opposite – namely, that geminal methyl diester reacts while monoester survives. Naturally, these conclusions are chemically faulty: in general, methyl esters are incompatible during reduction of methyl diesters (and vice versa) and the examples shown are scaffold-specific exceptions for which one needs to separately code an additional, wider-core reaction rule. Similarly, during the synthesis[42] of Milbemycin β_9 (**Figure 5b**) one lactone is reduced in the presence of another – this, however, is an exceptional example and, in general, lactones should be qualified as incompatible during reduction of another lactone moiety. The same problem of falsely qualifying an alkene as compatible with ozonolytic cleavage of another alkene is illustrated for the specific case in **Figure 5c** taken from the synthesis[43] of Hippospongic Acid A.

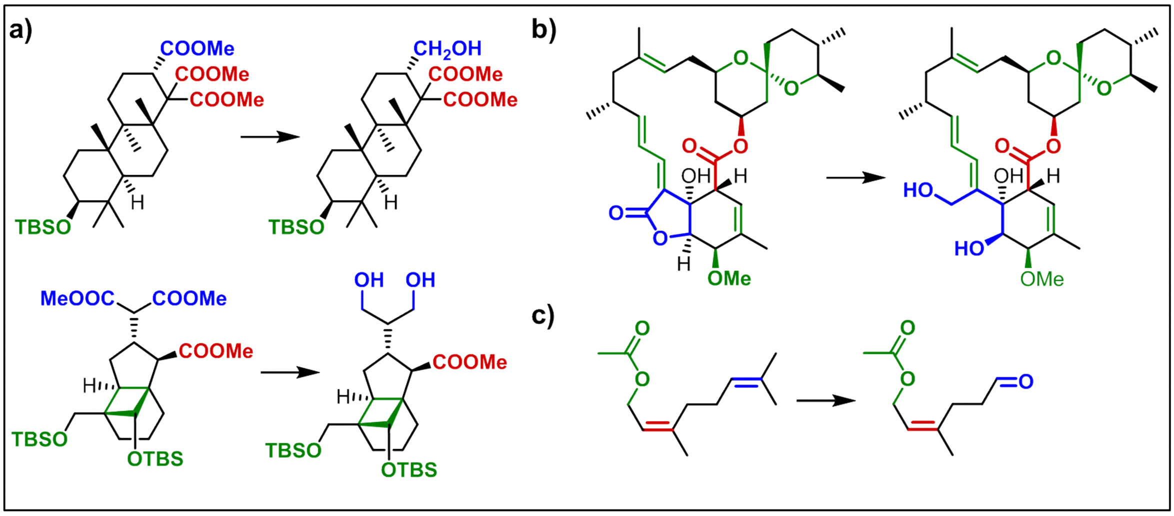


Figure 5. False-positive assignments of functional groups as compatible or incompatible. Blue – reacting functional group; green – appropriately assigned, stable functional groups; red – false-positive assignments of purportedly "compatible" groups. These groups are compatible only for certain reacting molecules while for others, under the same reaction conditions, they are not stable (i.e., incompatible). For detailed discussion, see main text.

Third, there are important classes of reactions for which the cores are identical yet the mechanisms differ and so the requirements of admissible vs. conflicting groups can be markedly different. As an example, consider Buchwald-Hartwig amination vs. nucleophilic aromatic substitution, where in both cases amine reacts with an aryl halide (**Figure 6**). Although the core of these reactions is identical, the mechanisms, scopes of admissible substituents and incompatible groups are substantially different. In aromatic substitution, the reacting aryl halide should be attached to an electron deficient ring (e.g., pyridine or nitroarene) while other aryl halides (including iodides) attached to neutral or electron-rich rings – even if present in the same molecule (**Figure 6b**) – remain unreactive. In the Buchwald-Hartwig amination, the electronic requirements are less important and reaction is not limited to electron deficient halides, though the reactivity of halogens usually follows the order I>Br>Cl and thus aryl iodides/bromides cannot be present while the chloride is supposed to react (**Figure 6c**).

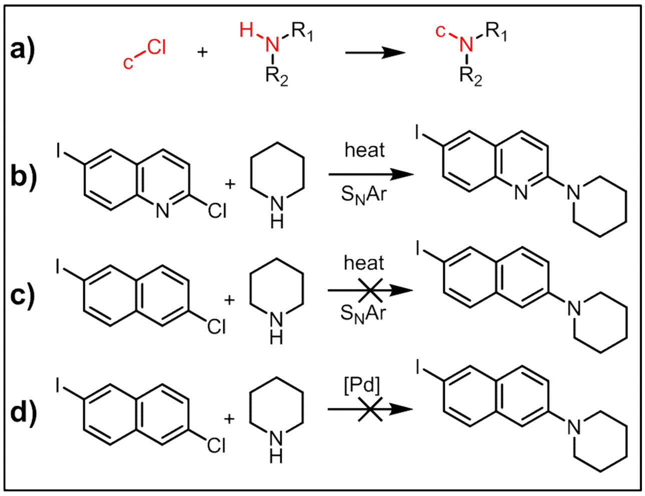


Figure 6. Reactions with identical cores but different mechanisms, scopes of substituents, and incompatible groups. **a)** Buchwald-Hartwig amination and nucleophilic aromatic substitution share the same reaction core and convert aryl chloride to arylamine. **b)** Nucleophilic aromatic substitution can occur in the presence of aryl iodide but requires electron deficient ring to proceed (compare with **c**). **d)** In contrast, Buchwald-Hartwig amination of aryl chloride cannot be performed in the presence of a more reactive aryl iodide.

Finally, machine-extracted rules may not properly handle the stereochemical information. From the very basic level, the proper handling of stereochemistry requires incorporation of so called canonical SMILES to ensure identical order of parsing atoms in the product and in the substrates. Even if this requirement is met, however, the tools like RDKit[44] still do not handle stereochemical notation properly and more complex solutions like Stereofix[7] or RDChiral[45] are needed. Moreover, the problem remains how to define the reaction transform to handle relevant stereochemical information not only in the substrate(s) but also chiral catalysts or reagents. In the latter case, properly defined reaction core should not allow for any nearby stereocenters in the substrate causing the so-called "mismatching effects" and lowering stereoselectivity or yield (**Figure 7**). For instance, Morken's hydroboration-oxidation[46] catalyzed by phosphonite ligand (denoted L4 in **Figure 7a**) should allow for unbiased substrate (**I**) and "harmless" nearby stereocenters (**II**, **III**) but preclude mismatched α oxygenated stereocenters (**IV**, red). Significantly, is it not easy to generalize such catalyst/reagent effects, and the scope of "harmless" vs. "harmful" environments may be quite different for different transformations – and thus not readily amenable to machine rule extraction. For instance, Roush's[47] *anti*-selective crotylation with *E*-boronate (**Figure 7b**, denoted *R*,*R*-1/*S*,*S*-1) affords products with acceptable diastereoselectivities in both matched (**I**) and mismatched (**II**) cases whereas *syn*-selective crotylation with *Z*-boronate denoted *R*,*R*-2/*S*,*S*-2 performs well only in the matched (**IV**) case (**Figure 7b**). Additionally, diastereoselectivity of Leighton's[48] *syn*- and *anti*- crotylations with *R*,*R*-3/*S*,*S*-3 (**Figure 7b**, *bottom-right structure*) is equally high in each case.

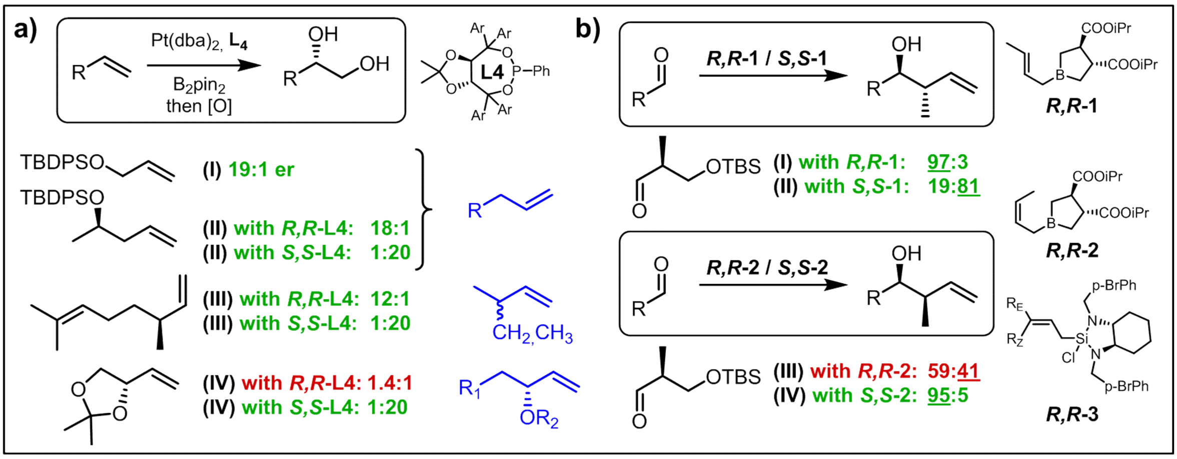


Figure 7. Defining reaction core to handle matched/mismatched cases in catalyst- or reagent-controlled stereoselective reactions. **a)** Morken's hydroboration-oxidation cannot be performed in the presence of mismatched nearby oxygenated stereocenters. Properly defined reaction cores are shown in blue. **b)** Lack of general rules for predicting mismatch effects. Scopes of admissible substituents for Roush's *syn-* and *anti*-selective crotylations are significantly different. In contrast, Leighton's crotylation with *R*,*R*-3/*S*,*S*-3 performs well in both matched and mismatched reagent/substrate pairs (**I-IV**).

On the other hand, when the observed stereochemical outcome of the reaction is substrate-controlled, one should carefully evaluate the mechanism responsible for the observed stereochemistry and ensure that reaction transform includes all necessary structural features. For example, reactions of alkenes are often controlled by the so-called allylic strain[49] (**Figure 8a**), intramolecular reactions and sigmatropic rearrangements (**Figure 8b**) often yield products deriving from the most energetically favourable, pseudoequatorial transition state, face selectivity of additions to cyclic systems is controlled by an entire set of substituents attached to the ring and occurs from the less hindered side (**Figures 8c**,**d**), whereas stereoselective 1,2-additions to carbonyl groups (**Figure 8e)** and alkylations of enolates (**Figure 8f**) proceed via chelated intermediates. In each case, the environments necessary to capture these stereoelectronic factors (green) are of different sizes and are significantly larger than just the cores of modified bonds or atoms (red). As in some other examples described before, there is – at least currently – no possibility to automate the extraction of such reaction rules.

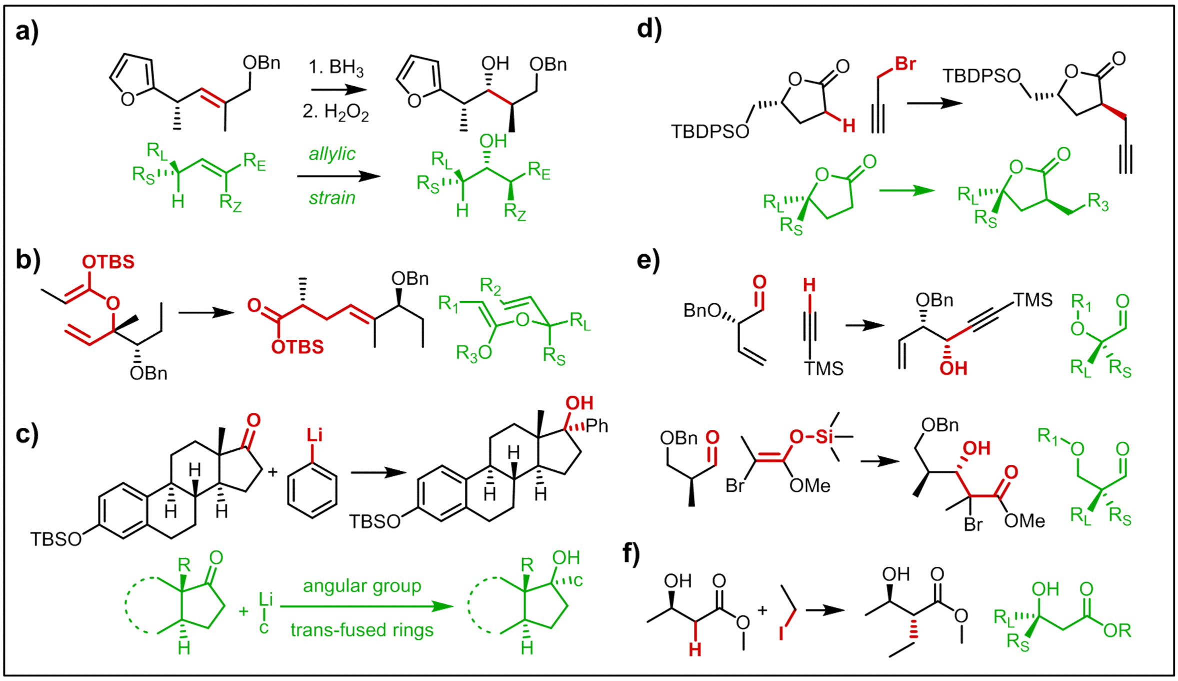


Figure 8. Defining reaction core for substrate-controlled stereoselective transformations. In each case, different model explains the observed stereoselectivity. **a)** Hydroboration-oxidation of alkenes is controlled by allylic strain and adequate reaction core should address relative size of substituents (R_L, R_S *L-large, S-small*). **b)** Stereochemistry of product obtained in Claisen rearrangement is dictated by relative size of substituents and occurs via a pseudo-chair transition state; **c)** Addition of a nucleophile to a cyclic ketone is controlled by an angular methyl group. Properly defined reaction core must address the presence of *trans*-fused bicyclic system and angular substituent of any size. **d)** Alkylation of lactone is controlled by a distant substituent and occurs from the less hindered face. **e)** Additions of nucleophiles to aldehydes and **f)** alkylations of enolates are commonly performed and controlled by chelated intermediates. Corresponding reaction rules should capture the presence of chelating groups (here, OBn or OH) and relative sizes of substituents. In all panels, red = reaction cores spanning changing atoms and

bonds, green = correct environments capturing substituents responsible for observed stereochemical outcomes. Examples are taken from [50-56].

Without spending much additional time on complications deriving from manual entry errors in reaction databases (**Figure 9a,b**) or serious problems with proper atom mapping across the reaction rules[28] (important to ensure that the rules "know" which atom of the substrate becomes which atom of the product; **Figure 9c**), our conclusion from this part is that the automatic-extraction approach entails serious chemical problems. In **Figure 10** and in **Figures S1-S25**, this conclusion is further corroborated by specific examples of erroneous, automatically extracted rules – in many cases, describing very basic reaction classes – underlying the operation of platforms such as Waller's MCTS[12] or MIT's ASKCOS[11,14]. We, of course, envision a possibility that the quantity *and* the quality of available literature examples will, one day, increase sufficiently to allow for more meaningful extraction and subsequent machine learning – though one should remember that although the "universe" of chemical reactions grows rapidly[8], the increase is mostly due to the proliferation of the popular reaction types whereas the statistics on the expert-level transformations, often scaffold-specific and applied to natural products, is growing much slower. How to ensure that more such expert-level data becomes available to the community is currently unclear to us given that total synthesis is, unfortunately, becoming less popular than few decades ago. In the meantime, at least some other problems related to automated rule extraction (e.g., more accurate treatment of incompatibilities) could be alleviated by publishing more negative results from which machine-learning approaches could benefit. Leaving such considerations to the community (and the funding agencies) to ponder, we reminisce that we ourselves were initially enticed to follow the path of automatic rule extraction – but only until the rules so derived proved inadequate when applied to retrosynthetic planning involving non-trivial targets. When one reaches this conclusion, one must reconsider the entire approach and face the enormity of the task ahead – that is, of coding the rules individually while taking into account reaction mechanism and several physical-organic considerations.

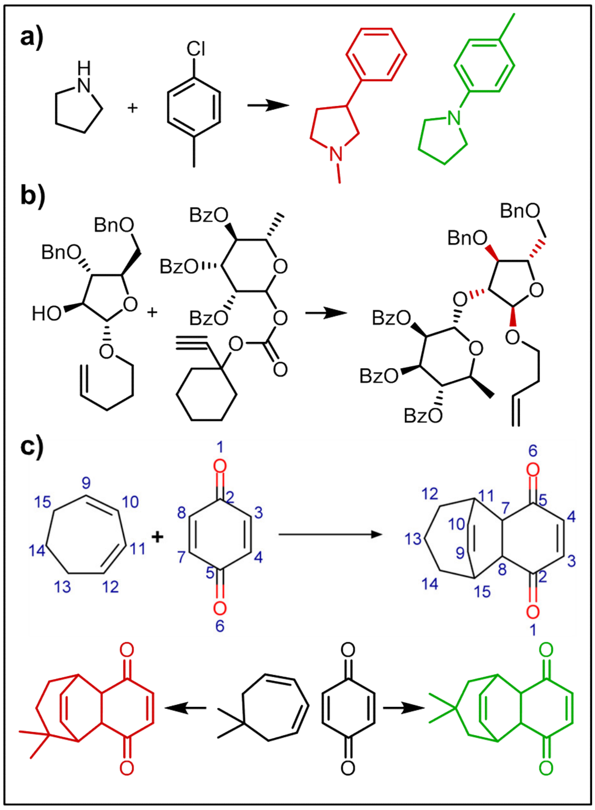
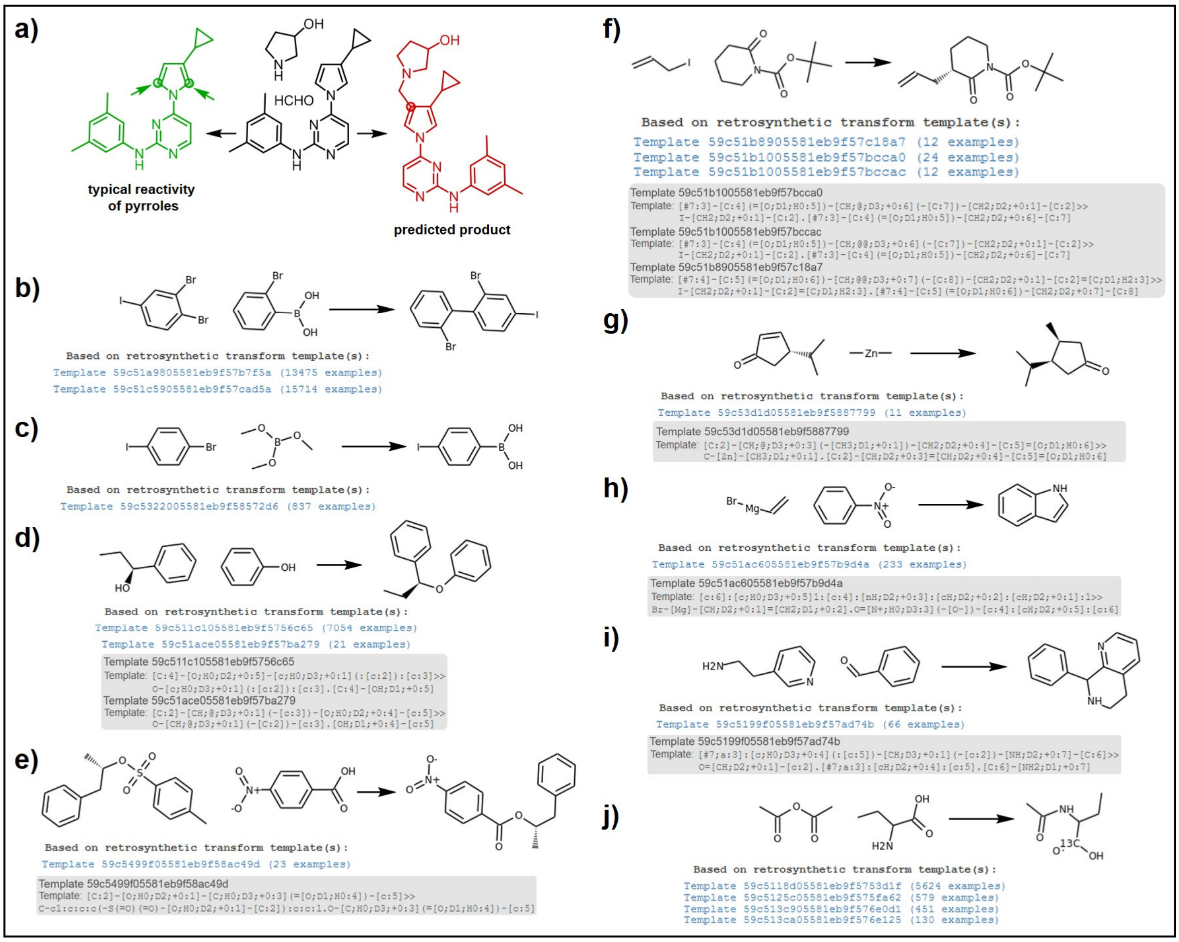


Figure 9. Errors in data sources translate into meaningless reaction rules. **a)** A product in data entry [57] in SciFinder database (red) does not match the one reported in the source publication [58] (green). The database entry corresponds to a tandem *N*-methylation/C3-arylation of pyrrolidine (with chlorotoluene being fragmented) while the original reported reaction is an ordinary *N*-arylation. **b)** Stereochemistry of several stereocenters is mis-assigned in Reaxys database entry 42850901 describing an instance of gold-catalyzed acetalisation[59]. Stereocenters marked red are

all incorrectly inverted. **c)** Reaction template derived from an incorrectly mapped Diels-Alder reaction from the USPTO dataset (top) applied to retrosynthetic planning generates faulty predictions. The reaction and substrates proposed for the synthesis of an adduct coloured red will, in fact, yield the product coloured green.



```
Based on retrosynthetic transform template(s):
Template 59c51a9805581eb9f57b7f5a (13475 examples)
Template 59c51c5905581eb9f57cad5a (15714 examples)
```

```
Based on retrosynthetic transform template(s):
Template 59c5322005581eb9f58572d6 (837 examples)
```

```
Based on retrosynthetic transform template(s):
Template 59c511c105581eb9f5756c65 (7054 examples)
Template 59c51ace05581eb9f57ba279 (21 examples)
Template 59c511c105581eb9f5756c65
Template: [C:4]-[O;H0;D2;+0:5]-[c;H0;D3;+0:1](:[c:2]):[c:3]>>
          O-[c;H0;D3;+0:1](:[c:2]):[c:3].[C:4]-[OH;D1;+0:5]
Template 59c51ace05581eb9f57ba279
Template: [C:2]-[CH;@;D3;+0:1](-[c:3])-[O;H0;D2;+0:4]-[c:5]>>
          O-[CH;@;D3;+0:1](-[C:2])-[c:3].[OH;D1;+0:4]-[c:5]
```

```
Based on retrosynthetic transform template(s):
Template 59c5499f05581eb9f58ac49d (23 examples)
Template 59c5499f05581eb9f58ac49d
Template: [C:2]-[O;H0;D2;+0:1]-[C;H0;D3;+0:3](=[O;D1;H0:4])-[c:5]>>
C-c1:c:c:c(-S(=O)(=O)-[O;H0;D2;+0:1]-[C:2]):c:c:1.O-[C;H0;D3;+0:3](=[O;D1;H0:4])-[c:5]
```

```
Based on retrosynthetic transform template(s):
Template 59c51b8905581eb9f57c18a7 (12 examples)
Template 59c51b1005581eb9f57bcca0 (24 examples)
Template 59c51b1005581eb9f57bccac (12 examples)
Template 59c51b1005581eb9f57bcca0
Template: [#7:3]-[C:4](=[O;D1;H0:5])-[CH;@;D3;+0:6](-[C:7])-[CH2;D2;+0:1]-[C:2]>>
          I-[CH2;D2;+0:1]-[C:2].[#7:3]-[C:4](=[O;D1;H0:5])-[CH2;D2;+0:6]-[C:7]
Template 59c51b1005581eb9f57bccac
Template: [#7:3]-[C:4](=[O;D1;H0:5])-[CH;@@;D3;+0:6](-[C:7])-[CH2;D2;+0:1]-[C:2]>>
          I-[CH2;D2;+0:1]-[C:2].[#7:3]-[C:4](=[O;D1;H0:5])-[CH2;D2;+0:6]-[C:7]
Template 59c51b8905581eb9f57c18a7
Template: [#7:4]-[C:5](=[O;D1;H0:6])-[CH;@@;D3;+0:7](-[C:8])-[CH2;D2;+0:1]-[C:2]=[C;D1;H2:3]>>
          I-[CH2;D2;+0:1]-[C:2]=[C;D1;H2:3].[#7:4]-[C:5](=[O;D1;H0:6])-[CH2;D2;+0:7]-[C:8]
```

```
Based on retrosynthetic transform template(s):
Template 59c53d1d05581eb9f5887799 (11 examples)
Template 59c53d1d05581eb9f5887799
Template: [C:2]-[CH;@;D3;+0:3](-[CH3;D1;+0:1])-[CH2;D2;+0:4]-[C:5]=[O;D1;H0:6]>>
          C-[Zn]-[CH3;D1;+0:1].[C:2]-[CH;D2;+0:3]=[CH;D2;+0:4]-[C:5]=[O;D1;H0:6]
```

```
Based on retrosynthetic transform template(s):
Template 59c51ac605581eb9f57b9d4a (233 examples)
Template 59c51ac605581eb9f57b9d4a
Template: [c:6]:[c;H0;D3;+0:5]1:[c:4]:[nH;D2;+0:3]:[cH;D2;+0:2]:[cH;D2;+0:1]:1>>
Br-[Mg]-[CH;D2;+0:1]=[CH2;D1;+0:2].O=[N+;H0;D3:3](-[O-])-[c:4]:[cH;D2;+0:5]:[c:6]
```

```
Based on retrosynthetic transform template(s):
Template 59c5199f05581eb9f57ad74b (66 examples)
Template 59c5199f05581eb9f57ad74b
Template: [#7;a:3]:[c;H0;D3;+0:4](:[c:5])-[CH;D3;+0:1](-[c:2])-[NH;D2;+0:7]-[C:6]>>
          O=[CH;D2;+0:1]-[c:2].[#7;a:3]:[cH;D2;+0:4]:[c:5].[C:6]-[NH2;D1;+0:7]
```

```
Based on retrosynthetic transform template(s):
Template 59c5118d05581eb9f5753d1f (5624 examples)
Template 59c5125c05581eb9f575fa62 (579 examples)
Template 59c513c905581eb9f576e0d1 (451 examples)
Template 59c513ca05581eb9f576e125 (130 examples)
```

Figure 10. Examples of incorrect predictions based on automatically extracted reaction rules. **a)** Electrophilic aromatic substitution of pyrroles occurs at positions marked with arrows (*left*). MCTS algorithm described in [12] incorrectly suggests the possibility of functionalization at a less reactive position (*red*). **b)** Incorrect prediction for Suzuki coupling caused by lack of rules accounting for incompatible functional groups. More reactive aryl iodide cannot be a spectator of

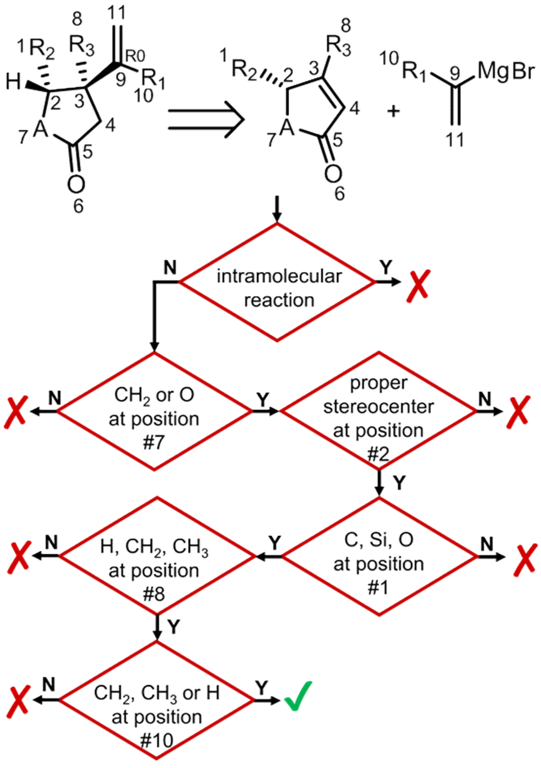
Suzuki coupling of an aryl bromide. **c)** Erroneous prediction for the preparation of *p*-iodophenyl boronic acid– this compound cannot be prepared as proposed from an aryl bromide due to higher reactivity of aryl iodide also present in the substrate. **d)** Improper handling of stereochemistry in Mitsunobu reaction. Proposed stereoretentive process is possible only if anchimeric assistance – missing in the extracted reaction transform – is accounted for. Extracted reaction core (*grey frame*) is common for primary, secondary (reacting with inversion of configuration) and tertiary (hardly reactive) alcohols. **e**) Incorrect predictions of stereochemical outcomes are not limited to Mitsunobu displacements. Another incorrectly stereoretentive process is proposed for the simple alkylation of a carboxylic acid a secondary mesylate. In this case, the algorithmically extracted reaction template (*grey frame*) is also too general and allows for substrates bearing primary, secondary and tertiary mesylates. **f)** Faulty predictions *en route* to a valerolactam derivative. The extracted reaction core is too narrow and does not account for the reaction being substrate-controlled (cf. **Figure 8d**). Lack of any directing groups in the substrate makes application of such a template to this target molecule incorrect. **g)** Incorrect prediction for a chiral-catalyst-controlled conjugate addition. The extracted reaction core is too narrow and, in this specific case, allows for mismatched (cf. **Figure 8c,d**) substituent. **h)** To achieve any appreciable yields, the Bartoli indole synthesis requires[60] *ortho*-substitution which is missing in this automatically extracted reaction template. **i)** Synthesis of tetrahydroisoquinolines via Pictet-Spengler cyclisation is feasible only when electron-rich (hetero)arylethylamines are used as substrates. Automatically extracted reaction core is too narrow and allows for annulation of electron–poor pyridine. **j)** Information related to the presence of isotopically labelled atom is lost during generation of the synthetic precursors. Examples **b-j** were taken from ASKCOS software [11,14].For additional examples and details, see **Figures S1-S25**.

Mechanism-based rule coding.

Per our discussion in the preceding section and also quantification in our previous works[7], there are on the order of 100,000 distinct reaction classes constituting the body of modern organic chemistry. The encouraging thing – and contrary to what some authors claimed[12] – is that while the number of specific reactions published in the literature grows exponentially and doubles approximately every 10-15 years [8,61], the number of new reaction classes/types with distinct mechanisms is not increasing nearly as rapidly (based on our experience and estimates, there are ca. 3,000-5,000 such examples per year). This makes the task of coding the rules manually manageable, at least in principle – in our case, it took over a decade and gave rise to, currently, over 75,000 reaction transforms incorporated into Chematica.

Coding each of these transforms begins with a thorough study and understanding of the reaction mechanism. Assume, for example, that we wish to code diastereoselective Michael addition of terminal vinyl magnesium bromides to cyclic enones (top-left portion of **Figure 11**; for clarity, Mg and Br atoms from the Grignard reagent are not shown on the left side of the reaction scheme). The first condition to be met is the intermolecular character of the reaction. This is motivated by the fact that only intermolecular cyclisation can ensure desired *cis* arrangement of R_2 and R_3 groups in the product. In the SMARTS notation in the reaction record shown in the figure, (field 'Reaction's SMARTS'), this requirement is indicated by the 'R0' sign specifying that atom #9 is not allowed to be a part of any ring. On the other hand, the presence of a ring between substituents R_2 and R_3 is allowed. Next, we need to specify the substituents present on the five-membered ring. Available literature on the topic indicates that CH_2 and oxygen are allowed at position #7 as cyclic ketones and esters could serve as reacting partners.

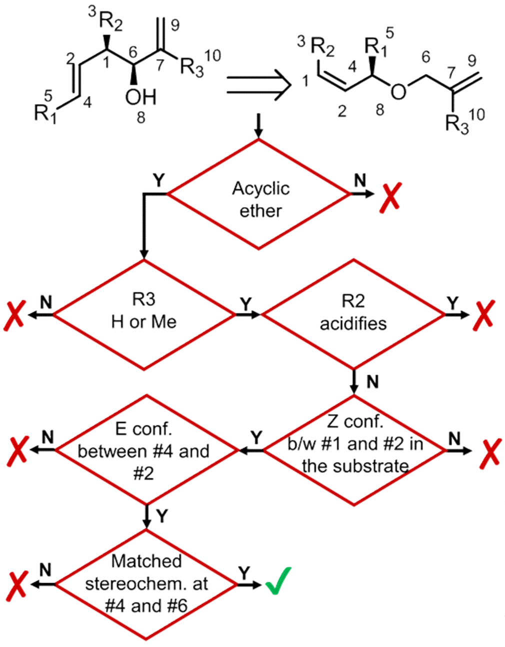
Also, proper chirality of atom #2 has to be specified, as the stereoselective outcome is dictated by orientation of the substituent present at this position. Because atom #7 is indicated as either CH_2 or oxygen, the substituents at position #1 are limited to those that are admissible for both ketones and lactones serving as Michael acceptors. The substituents at position #8 are limited to unsubstituted alkyl or a hydrogen because: (i) bulky groups on the β-carbon reduce reactivity of Michael acceptor, and (ii) it is necessary to avoid an additional chiral center that might influence stereoselectivity of the reaction, especially in the presence of a ring between substituents R_2 and R_3. To prevent bulky groups that may cause steric hindrance and disturb the addition, atom present at position #10 of the vinyl magnesium halide is limited to hydrogen or unsubstituted alkyls. Other substituents, such as vinyl group, need to be specified in separate SMARTS lines. The reaction record also includes additional fields specifying groups that need to be protected (e.g., aldehydes, primary amines and thiols, in total 14 groups), groups that are always incompatible in the reaction (e.g., acid chlorides, other Michael acceptors, etc., in total more than 100 groups of which 47 are listed in the figure), typical reaction conditions, representative literature sources, and 10 other fields (18 in total). We note that different variants of stereoselective Michael addition (e.g. with 6-membered enone serving as a Michael acceptor or other nucleophiles reacting as Michael donors – e.g., aryl Grignard reagents, amines, more substituted vinyl magnesium halides, thiols etc.) are also included in Chematica.



Reaction's NAME:
Diastereoselective addition of vinyl magnesium bromides to Michael acceptors
Reaction's SMARTS:
[CH2:11]=[CR0:9]([CX4H2,CX4H3,#1:10])[C@@:3]1([CX4H3,CX4H2,#1:8])[CH2:4][C:5](=[O:6])[CH2,O:7][C@@H:2]1[OH0,#6,SiX4H0:1].[Mg:88][Br:87]>>[O:6]=[C:5]1[*:7][C@H:2]([*:1])[C:3]([*:8])=[C:4]1.[Br:87][Mg:88][C:9]([*:10])=[C:11]
Protection condition's code:
SB23
Protections:
[#6][CH2][OH] , [#6][CH]([#6])[OH] , [#6][C]([#6])([#6])[OH] , [c][OH] , [OH][c][c][OH] , [#6][CH]=O , [#6][C](=[O])[CX3]=[CX3] , [#6]C([OH])=O , [#6][SX2H] , [nH] , [OH][CX4][CX4][OH] , [OH][CX4]C[CX4][OH] , [CX4,c][NH2] , [CX4,c][NH][CX4,c]
Incompatibilities:
[#6][CH]=[SX1] , [#6]S(=O)[Cl,Br,I] , [#6]C(=[SX1])[#6] , [SX4]=[NH] , [PX4]=[NH] , [SX3]=[NH] , [#6][SX3](=O)[OH] , ClC=N , [CX4]1[SX2][CX4]1 , [#6][S](=O)(=O)[OH] , [OH][CX4][O] , [#6]C(=O)[Cl,Br,I] , [#6]C(=[O])OC(=[O])[#6] , [#6][N+]#[C-] , [#6]N=C=[O,S] , [#6][SX2,O]C#N , [#6]O[N+]([O-])=O , [#6]O[OH] , [#6]OO[#6] , [CX3]=[CX2]=O , [#6]=[N+]=[N-] , c[N+]#[N] , [CX3]([#6,#1])([#6,#1])=[NX2H] , [CX3]=[CX3][OH] , [#6][Li] , [#6][Mg][*] , P(=O)[OH] , [#6][B-](F)(F)F , [O,N,P,S,n,Si,Se]C(=O)[C](C(=O)[O,N,P,S,n,Si,Se])=[C] , N#C[C](C#N)=[C] , N#C[C]=[C]C#N , [SiX4,SX2,Se,P][Cl,Br,I] , [CX4]([OH0,PH0,BH0,SH0,Se])[Cl,Br,I] , [Cl,Br,I]C#C , [O,S,Se,P,NX3]C#C , [CX4H2][Br,I] , [CX4H1][Br,I] , [CX4H0][Br,I] , [CX3]=[CX3]C#N , [CX3]=[CX3]C(=O)[O,N,S] , [CX2]#[CX2]C#N , [CX2]#[CX2]C(=O)[O,N,S] , [O-][N+](=O)[CX3]=[CX3] , [O-][N+](=O)[CX3]=[CX2]=[CX3] , FC(F)(F)[CX3]=[CX3!H0] , [CX3]=[CX3]S(=O)(=O)[O,#6] , [CX3]=[CX2]=[CX3]S(=O)(=O)[O,#6] ***AND MORE***
Typical reaction conditions: 1. CuI, low temperature, 2. NH_4Cl
References: 10.1016/j.tet.2011.09.022 and 10.1021/ol061510m

Figure 11. Mechanism based translation of diastereoselective Michael addition into a machine readable format. The upper left part of the figure presents the general scheme of the reaction in retrosynthetic direction and the "decision tree" guiding the coding of a corresponding reaction rule. The reaction record also contains information about groups that need protection, incompatibilities, reaction conditions etc.

Another example, in **Figure 12**, illustrates encoding of a more chemically advanced 2,3-Wittig rearrangement with chirality transfer. The first requirement is the acyclic character of the ether (variants for cyclic substrates are also known, but they need to be coded in separate lines delineating the scope of substituents on a given ring system). The condition is met by denoting carbon #6 as R0 (which means it cannot participate in any ring). Another condition describes structure of the migrating group. In our specific example, it is limited to an unsubstituted allyl or allyl with methyl substituent at carbon #7. We note that although many groups can migrate in the [2,3]-Wittig rearrangement, and it is possible to write one, general SMARTS (e.g., [CH2:6][#6,Sn:7], where #6, Sn- means any carbon or tin atoms), it is not advisable to code the rule in this way since α-(allyloxy)carboanions need to be generated at low temperatures (usually -78^{o}C) to suppress the competing [1,2]-Wittig shift[62]. Even for the limited SMARTS transcription encompassing only generally accepted migrating groups such as phenyl, propargyl, or allyl, writing one SMARTS line (e.g. [CH2:6][CX3,CX2,c:7]) is not a good solution because these groups differ in terms of incompatibility requirements. Additionally, the line is restricted to the unsubstituted or 2-methyl allyl (and not to *any* allyl) because (i) it ensures *syn* selectivity of the newly created stereocenters at atoms #1 and #6; and (ii) less substituted allyl is a better migrating group. Another condition describes non-acidifying character of the substituent R2 (comprised of atoms #5,11,12,13 in the SMARTS line). The chemical rationale is that increasing the acidity of protons adjacent to carbon #4 might result in its deprotonation (instead of deprotonation of atom #6) and migration of the second allyl group, thus resulting in a different product. Another condition is the *E* configuration of the resulting alcohol. [2,3]-Witting rearrangement is known to be *E*-selective and the *Z* isomer is observed as the minor one. The last requirement that needs to be considered is proper relative stereochemistry at atoms #1 and #6, as *syn* alcohol derives from *Z*-alkene while *anti* is obtained with significantly lower selectivity from the *E*-isomer[63]. As in the example from **Figure 11**, the reaction record also contains the list of groups that require protection, those that are outright incompatible with the reaction, as well as ten other fields. We note that filling in many of these fields requires careful analysis – for instance, there are some 430 incompatible groups we routinely consider during coding, and the particular selection for a given reaction must reflect reaction mechanism as well as reaction conditions.



3 R2
9
2
1
6
7
R3 10
5
R1
4
OH
8
3 R2
R1 5
6
9
1
4
7
O
2
8
R3 10
Acyclic ether
Y
N
R3 H or Me
N
Y
R2 acidifies
Y
N
E conf. between #4 and #2
N
Y
Z conf. b/w #1 and #2 in the substrate
N
Y
Matched stereochem. at #4 and #6
N
Y
✓

Reaction's NAME: ***[2,3]-Wittig Rearrangement***

Reaction's SMARTS: [CH2:9]=[C:7]([#1,CH3:10])[C@@HR0:6]([OH:8])[C@H:1]([CX4!H0:3])/[CHR0:2]=[CH:4]/[CX4:5]([CH2,CH3,#1:13])([#1,CH2,CH3:11])[#1,CH2,CH3:12]>>[CH:1](\[*:3])=[CH:2]\[C@H:4]([O:8][C:6][C:7]([*:10])=[C:9])[C:5]([*:13])([*:11])[*:12]

Protection condition's code: SB27

Protections: [#6][CH2][OH],[#6][CH]([#6])[OH], [#6][C]([#6])([#6])[OH], [c][OH][OH][c][c][OH], [#6][CH]=O, [#6]C([#6])=O, [#6]C([OH])=O, [#6][SX2H], [OH][CX4][CX4][OH], [OH][CX4]C[CX4][OH], [CX4,c][NH2], [CX4,c][NH][CX4,c], [nH]

Incompatibilities: [#6][CH]=[SX1] , [#6]S(=O)[Cl,Br,I] , [#6]C(=[SX1])[#6] , [SX4]=[NH] , [PX4]=[NH], [SX3]=[NH] , [#6][SX3](=O)[OH],[CX4]1[SX2][CX4]1, [#6][S](=O)(=O)[OH] , [#6][N+]#[C-], [#6]N=C=[O,S] [#6]C(=O)[Cl,Br,I] , [#6]C(=[O])OC(=[O])[#6] , ClC=N , [#6]O[N+]([O-])=O , [#6]O[OH] , [#6]OO[#6] , [CX3]=[CX2]=O, [#6]=[N+]=[N-] , c[N+]#[N] , [CX4][Cl,Br,I], c[I], [Cl,Br,I]C#C, [#6]S(=O)[Cl,Br,I], [CX4]1[O,N][CX4]1, [CX4]1[SX2][CX4]1, [#6][SX2][SX2][#6] , [#6]C(=[O])OC(=[O])[#6], [#6]C(=O)[NH2], [#6]C(=[O,S])[NH][#6], [#6]C(=[S])[NH2], [#6]C(=[S])[NH][#6], [#6][S](=O)(=O)[NH2], [CX3]=[NX2H], [CX4,c][NX3][NH2], [#6]=[N+]=[N-], c[N+]#[N].[#6][Li], [#6][Mg][*], [#6][Zn][*], [#6][SX2,O]C#N, [#6]O[N+]([O-])=O, [#6][NX2]=O, cO[CX4!H0]c, [NX3H][C]=[C], [NX3H2][C]=[C], ,[CX4!H0][C](=[O])[OH0], [CX4!H0]C(=O)N([#6])[#6], [#6]C(=O)n, [CX3,c,CX2][CX4][O][CX4!H0][CX3,c,CX2,P,S], [#6]C(=[O])[NH][#6], [#6][S](=O)(=O)[#6], [#6]C(=O)N([OH0])[#6], [*][O][C](=[O])[O][*], [#6,N,O,P,S][NH]S([#6])(=O)=O, [n][c;r6](=[O])[n,c], [n][c;r6]([Cl,F])[n,c], [#6][N+]([O-])=O, [CX4!H0][C](=[O])[OH0], [CX4][O][CX4!H0][CX2]#[CX2], [n][C](=[O])[O][#6], [c]1[n][n][n][n]1[CX4!H0], [#6,P][O,NX3,SX2,n,SeX2][CX4H2][c],[CX4][O][CX4H2][C]=[CH2], [CX4!H0][c]1[n][n][n][n]1 ***AND MORE …***

Typical reaction conditions: nBuLi or other base.low temp.

References: 10.1021/jo00184a040 and 10.1021/ol0612612

Figure 12. Mechanism-based translation of 2,3-Wittig rearrangement into a machine readable format. The upper-left part of the figure shows general scheme of the reaction in retrosynthetic direction and the "decision tree" guiding the coding of a specific reaction rule. The reaction record also contains information about groups that need protection, incompatibilities, reaction conditions, etc.

While the coding protocols such as those we outlined in this section offer a substantial improvement – in terms of chemical correctness – compared to indiscriminate machine-extraction of rules, they do not yet cover all aspects of rules' applicability, as it is often impossible to capture all the relevant effects just in the SMILES/SMARTS notation. These additional effects generally require augmentation by QM or MM calculations or by AI methods which we discuss in specific sections below.

The importance of structural context.

In our discussion above, one manifestation of "molecular context" has been the treatment of groups that present cross-reactivity problems or those that need to be protected. In addition, there are context dependencies that relate to the skeletal structure of the retron and/or the synthons. As an example, let's consider a Wittig or a metathesis reaction – these powerful reactions have very broad scopes but cannot proceed – due to strain – to install the double C=C bonds at the bridgehead atoms of small bicyclic systems (the so-called Bredt's rule). Specifying such outcomes at the level of reaction transforms is unfeasible whereas performing detailed calculations on-the-fly would take too much time (in particular, during retrosynthetic planning whereby very large numbers of intermediates are inspected). A more practical option is to prohibit in all molecules considered during planning those motifs that violate the Bredt's rules and, more generally, all those that are excessively strained. Some of such "prohibited motifs" are recognized as impossible by inspection, and some require more careful evaluation of strain via molecular mechanics calculations (see **Figure 13a** for a small selection of such motifs from our library of ca. 600).

A slightly more subtle variation of this problem is what to do with motifs which, in principle, can exists but only under rather extreme conditions – e.g., bicyclo[1.1.0]butanes or cyclopropenes. In such cases, we make the motif prohibited unless it is explicitly present in the target one wishes to synthesize.

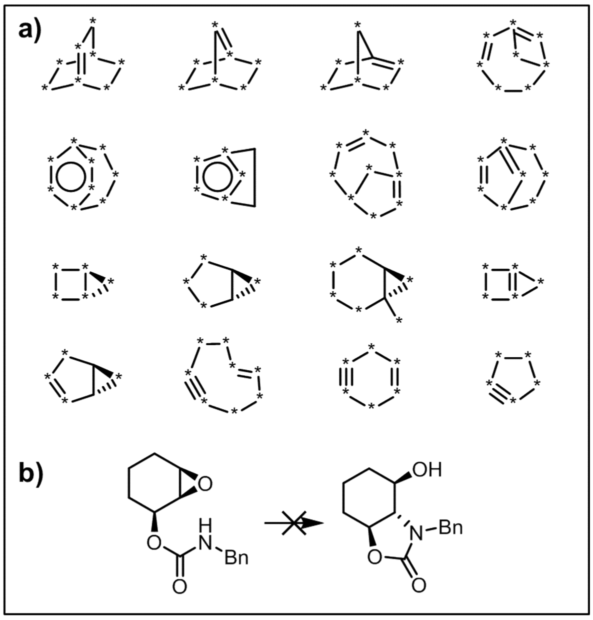


Figure 13. Strained molecules and reactions. **a)** A small sample of strained motifs forbidden in retrosynthetic planning. * denotes any atom. **b)** A bicyclic system cannot be prepared via base-induced ring opening of epoxide due to high strain along the reaction coordinate.

A related problem arises when synthon/retron molecules are not themselves strained but the reaction cannot occur because strain develops in the transition state, often during intramolecular cyclization reactions. Some of such cases can be captured by prohibiting pairs of motifs in the retron and in the synthon(s) such that the former cannot form from the latter. In the example shown in **Figure 13b**, a *trans*-fused bicyclic system cannot be formed via S_N2 reaction involving base-induced ring opening of epoxide with carbamate. Such clear-cut cases, however, are rare and, in general, one has to quantify strain along the reaction coordinate. This problem is, of course, not a new one but when combined with synthetic planning in which large numbers of molecules are evaluated, it has a distinct flavour – all such calculations have to be performed very rapidly (in Chematica, well below 100 msec to allow inspection of tens of thousands upon thousands of intermediates during a typical retrosynthetic search[7,9]). In some cases, classic physical organic chemistry knowledge is very helpful as it provides us with the information about the angles at which the cyclizing groups approach one another – in this way, the mutual orientations of the reacting species can be restricted, greatly simplifying the problem. For instance, if a cyclization reaction is between a nucleophile and a carbonyl group, the trajectory of approach is specified by the so-called Bürgi-Dunitz[64] and Flippin-Lodge[65,66] angles. Calculation of energy along such a well-defined coordinate by Molecular Mechanics is quite rapid and when the threshold strain is calibrated against examples of reactions that are known *not* to proceed, one can eliminate a sizeable proportion of impossible cyclizations (cf. examples in **Figure 14**). This being said, we emphasize that we do not yet have a general method available to deal with all cyclizations, since in many cases a unique "angle of approach" is hard to define and one then has to sample more mutual orientations and molecular conformations; these nuances will be described in our upcoming papers on the topic.

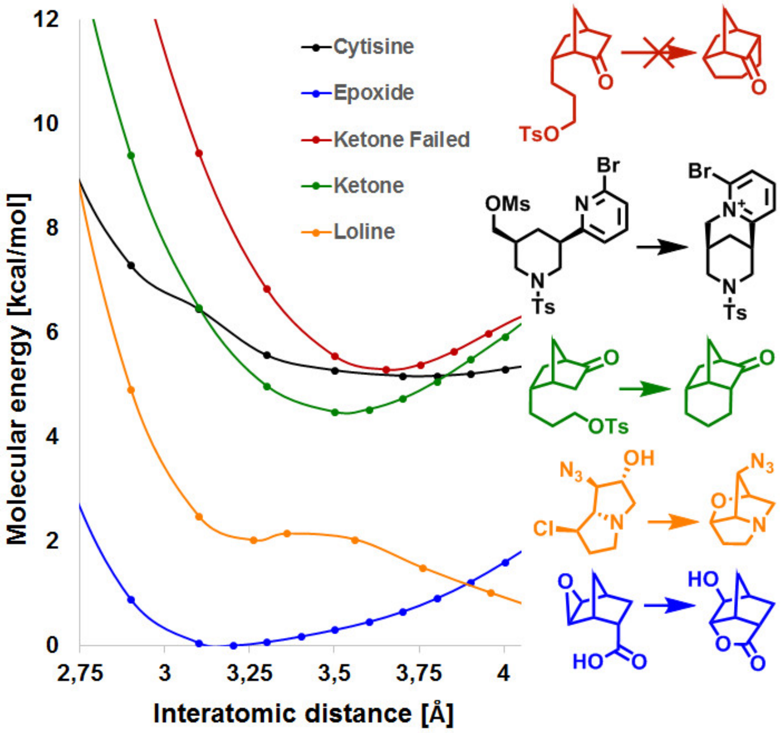


Figure 14. Energy profiles for several experimentally attempted S_N2 cyclizations with either positive or negative outcomes. The examples span alkylation of a ketone (red and green curves; from ref [67]), opening of an epoxide (blue, ref [68]), cyclisation forming loline's skeleton (orange; [69]), and cyclisation executed *en route* to cytysine (black, ref. [70]). The admissible energies must be below the red curve which corresponds to a cyclization that is known not to proceed in experiment. The energies were calculated using Merck Molecular Force Field 94 (MMFF) and were close to energies obtained from more precise HF/6-311+G** calculations. Figure reproduced with permission from [71].

Accounting for non-local electronic effects.

In some very popular reaction classes, the number of possible substituents is not only extremely large – too large to enumerate exhaustively – but their mutual placement is essential. The problem here is that on cannot just specify in the reaction transforms the lists of substituents at different positions since not all their combinations are permissible. For instance, for electrophilic aromatic substitutions on a benzene ring, we cannot just specify lists of, say, H, Cl, Br, NH_2, NO_2, etc. for every surrounding position because, as is well known from organic chemistry classes, different substituents have different ortho/para vs. meta-directing abilities and, depending on a specific arrangement of these groups present on the ring, they might collectively activate or deactivate our position of interest. The problem is further compounded by the fact that the degrees of such influence can be different depending on the aromatic or heteroaromatic system being substituted. In such cases, the only legitimate strategy is to define a reaction core very narrowly (i.e., an electrophile plus an aromatic carbon being substituted) but supplement such a reaction rule by a routine calculating the propensity of this specific carbon atom to undergo the substitution reaction.

In our early works[7], we used a classical Hückel method to calculate electron densities over aromatic systems and allowed substitutions only at positions for which the densities were above certain threshold values. This method, however, had only ~75% accuracy over diverse aromatic systems. The “diverse” is an important keyword here because methods parameterized only on certain types of aromatic systems (see [72,73]) are of limited use for generalized synthetic planning (and the more accurate proton/electrophile affinity approaches are prohibitively slow). In the end, we implemented a “hybrid approach” combining Hammett substituent constants, proton affinities averaged over all aromatic carbons within a specific ring type (pre-calculated at the DFT level of theory using B3LYP functional with 6-31+G* basis set), the Hückel method (with parameters for heteroatoms taken mostly from [74]), as well as some additional heuristics. This model, described in detail in the Supplementary Information to ref [9], evaluates aromatic substitution reactions with speeds commensurate with synthetic planning (10-100 msec per reaction) and offers accuracy above 90%.

Augmenting reaction rules by AI models.

The most general problem one might face when considering rules' applicability is when both electronic and steric effects are important. In situations when there are multiple reaction precedents/examples available for a given, well defined reaction type, one might fine-tune substituent scope by machine learning, ML, methods. This type of approach has been used recently by Doyle and co-workers[75] for predicting the substituent-dependent outcomes of Buchwald-Hartwig couplings, although the specific implementation of AI methods subsequently received critique[76] for the lack of statistical rigor in model testing.

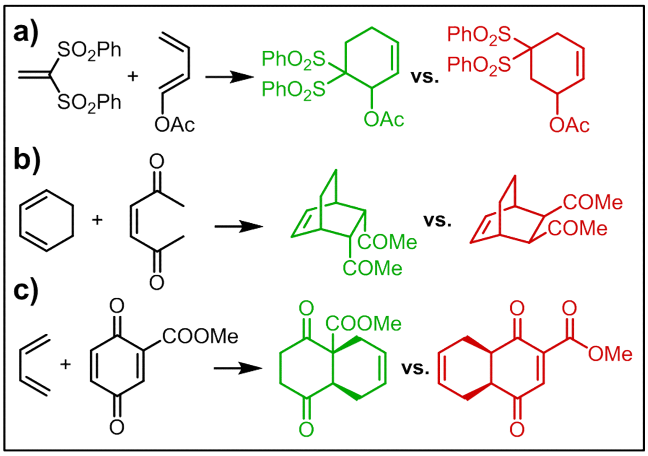


Figure 15. Selectivity of Diels-Alder reaction. **a)** Formation of regioisomers from unsymmetrical dienes. **b)** Formation of *endo-*/*exo-* diastereoisomers. **c)** Formation of different products from substrates with multiple reaction sites. Green – observed product, red – possible by-product.

To illustrate some key aspects of AI rule augmentation, we consider a synthetically powerful[77-78] Diels-Alder reaction for which ~ 20,000 reaction precedents are readily available in Reaxys. As we discussed in detail in our recent work on this topic[27], unsymmetrical dienes and dienophiles can react in different orientations to give different regioisomers (**Figure 15a**) or diastereoisomers (**Figure 15b**) or, when multiple diene/dienophile sites are present in the molecule, altogether different products (**Figure 15c**). These outcomes are dictated by the substituents present on the diene (maximum six substituents) and the dienophile (maximum four substituents) – in other words, the reaction core is small and well defined but much of the game is in the atoms/groups decorating the core. Clearly, enumerating all possible substituent combinations is impractical and, indeed, quite pointless, as the outcomes for the majority of them would have no experimental benchmark to compare with. QM-based calculation are not only slow but, as we showed in [27], offer an accuracy of only ~80% in terms of predicting correct outcomes. On the other hand, the number of literature precedents is sufficient to train ML models assigning certain "features" to the substituents and learning how the combinations of these features translate into reaction outcome (the leading regio-, diastereo-, or site-isomer). Without repeating the entire discussion from ref [27], few main points are worth re-emphasizing as their applicability is beyond the Diels-Alder example:

(i) It is essential to use features that capture electron donating/withdrawing propensities of substituents as well as their steric bulk. With, respectively, Hammett constants[79] and the TSEI indices[80] assigned to the substituents, the ML models can offer accuracy well above 90% and are also applicable to structurally diverse examples, including cases not seen during model training.

(ii) Although atom-connectivity-based descriptors (e.g., ECFP4[81], MACCS, or RDKit[44] fingerprints) or even physically meaningless descriptors (e.g. random numbers assigned to substituents) can offer high accuracies when trained on structurally related examples, such models fail when confronted with examples structurally different than those seen during training.

(iii) The AI methods perform significantly better when provided some "insight" about the reaction. When the reaction cores are not specified and the algorithms are learning from just the structures of the products and reagents, their performance is significantly worse.

Ensuring rules' quality and estimating scope.

As evidenced by our discussion so far, the number of possible factors that need to be considered during reaction coding is quite substantial. Of course, not all of the aspects we highlighted need to be considered simultaneously: for instance, calculations of electron densities usually are relevant only to systems of delocalized π electrons (but see **Figure 16**), whereas AI augmentation routines should be considered only for rules having large numbers (say, more than ~1,000) of reliable literature precedents. Still, the coding process is certainly very time-consuming an also requires tremendous care to eliminate possible human errors. In our team, we have implemented several levels of quality control – special scripts checking for rules' proper syntax, scripts testing applicability of rules on some test-molecules, a peer-review cross-checking system (chemists checking each other's transforms), and ultimate verification of the results by a super-user who inputs the reaction rules into Chematica's database. We also have in place an error reporting systems from Chematica's end users and standardized procedures how such errors are fixed. Over the years, we followed these procedures to encode more than 70,000 reaction transforms whose ultimate validation has been their correct performance in experimental execution of computer-planned routes[9].

In parallel, we have been keeping track of the retro-/synthetic scope of our collection. One illustrative metric has been how many transforms with specific substituents at each position our rules cover – the most recent number stands at ~ 4.5 million reactions (~ 70, 000 rules/transforms, with a median of distinct 64 specific reactions per transform) expanding to several tens of millions when '*' symbols denoting "any atom" are considered (with a very conservative assumption that each '*' admits just one substituent type). This is significantly more than 320,000 unique radius = 2 transforms extracted[29] from the USPTO collection, and more than ~13 million literature precedents in the Reaxys database, altogether attesting to the synthetic latitude of our collection. The bottom-line message of this section is that, contrary to some claims[12], expert-coding of rules can not only keep up with the "current literature" (as deposited in reaction databases) but can exceed its scope and make high-quality, mechanism-based extensions to reaction variants not yet carried out.

Some unsolved problems.

Of course, there are still rules that need to be added and we estimate that our ultimate collection will comprise some 100,000 transforms to address even the most complex synthetic problems. At least some of these rules will benefit from additional electronic density calculations, probably at levels of theory higher than those used to predict aromatic substitutions (cf. earlier in the text and ref [9]). Such calculations could, for instance, better delineate the applicability of Prins cyclisations suffering from competing Oxonia-Cope rearrangement for electron rich benzyl-alcohol substrates [82] (**Figure 16a**), determine electron density thresholds for benzaldehydes participating in aldol reaction[83-84] (**Figure 16b,c**), penalize substitutions occurring at electron rich benzylic positions commonly suffering from competitive S_N1 process leading to erosion of optical purity[85-86] (**Figure 16d,e**), or help assure the electron withdrawing character of styrenes alkylated with diesters[87] (**Figure 16f**) .

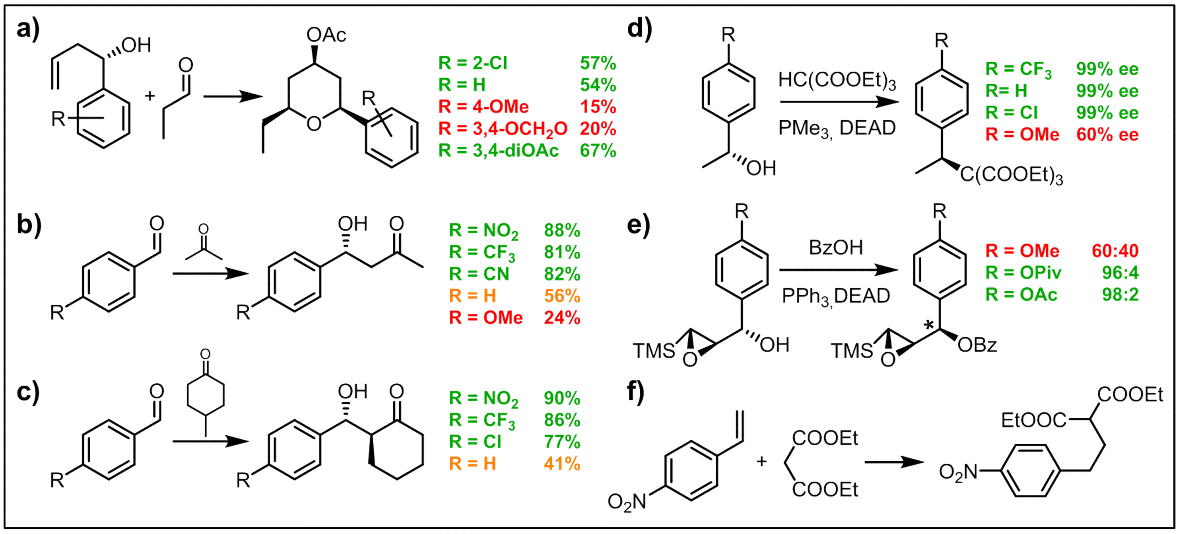


Figure 16. Reactions other than electrophilic aromatic substitutions that would benefit from electronic density calculation. **a**) Prins cyclisations of electron-rich benzylic alcohols suffers from competitive Oxonia-Cope rearrangement; **b,c**) Organocatalytic aldol reactions proceed well for electron deficient benzaldehydes; **d,e**) Mitsunobu reaction of electron-rich secondary alcohols suffers from competitive S_N1 process leading to erosion of *ee*; **f**) addition of nucleophiles to vinylarenes requires electron deficient arene to proceed. Substituents and yields listed in orange and red correspond to cases that one would probably want to eliminate from the reactions' scopes.

Another challenge is to couple the reaction rules with MM calculations of molecules' conformations to ensure that the reaction sites are available and not sterically hindered – such evaluation would be of most relevance to the synthesis of natural products for which there are many examples of conformational effects dictating reactivity[88]. The specific molecules formed as a result of rules' application should also be evaluated for the stability of certain stereochemical motifs. For instance, *cis*-substituted dihydrobenzofuran (**Figure 17a**) is known to be unstable during BBr_3 mediated demethylation of phenol[89], certain piperidines epimerize via retro-Michael/Michael addition during HWE olefination [90] or conversion of ester to methyl ketone [91] (**Figure 17b,c**), while trans fused Diels-Alder adduct shown in **Figure 17d** epimerizes easily to a more stable *cis*-lactone when treated with silica[92]. It is presently unclear to us whether such subtle effects can be reliably calculated/predicted or whether it is better to create and curate a growing, literature-based list of "unstable motifs" that would be matched against the outcomes of reaction rules.

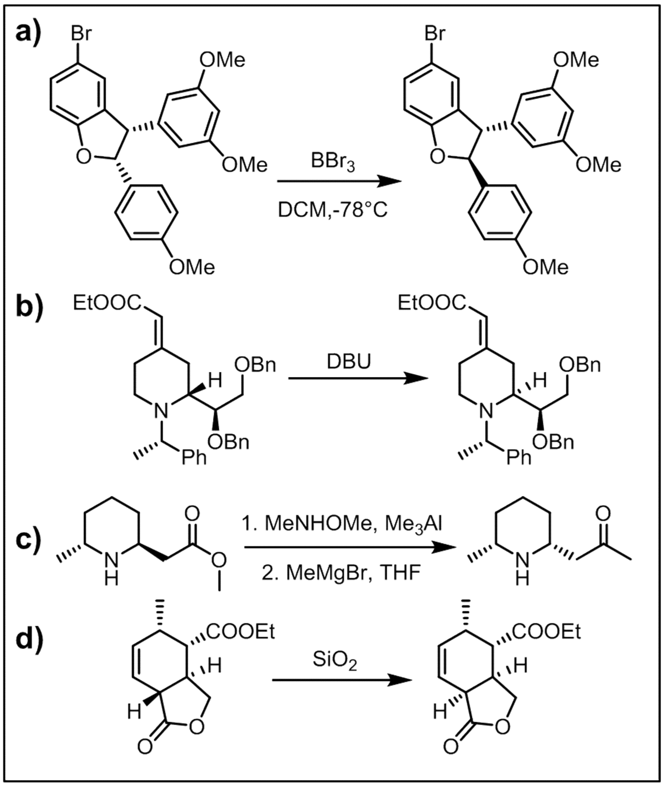


Figure 17. Unexpected epimerizations of thermodynamically unstable molecules. **a)** Epimerization of the *cis*-diaryl tetrahydrofuran via *p*-quinonemethide observed during attempted demethylation of phenols [89]. **b)** Epimerization of a piperidine derivative through retro 1,6/1,6 addition observed during attempted HWE olefination [90]. **c)** Epimerization of the *trans*-piperidine

via retro 1,4/1,4 addition observed during attempted conversion of an ester to a methyl ketone[91]. **d)** Epimerisation of strained *trans*- fused lactone[92] upon treatment with SiO_2.

Finally, an important problem to address is the estimation of the pK_a of CH acids and protic groups. For instance, during alkylation of a valerolactam[93] anion with phenethyl bromide, the low yield can be ascribed to the acidity of benzylic H (red in **Figure 18a**) and competing E2 elimination forming styrene by-product. In the same genre but for a more complex target, unexpected acidity of vinyl triflate thwarted the desired deprotonation of lactone during Vanderval's synthesis of Ineleganolide[94] (**Figure 18b**). Problems of this sort can be solved be estimating the pK_a values within the molecule and inspecting if there are protons more acidic than the one at our desired reaction centre. Although pK_a values for some specific series of structurally related compounds have been described in the literature[95], a solution applicable to arbitrary molecules in organic solvents is still missing. There are some models developed internally by large pharma companies or commercial packages from companies[96-99] such as Schrödinger, but it is unclear how accurate these tools are (e.g., Schrödinger's package appears quite accurate for some cases, but in some others – e.g., PhOMe calculated with Jaguar's DFT calculation with empirical corrections – gives predictions missing the experimental values by almost five pK_a units!). We have been working on developing our own pK_a predictor but it is too early to estimate its ultimate accuracy.

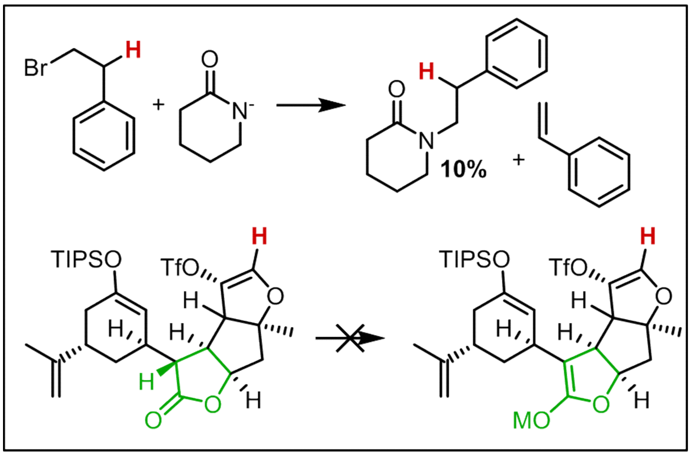


Figure 18. Presence of acidic H's causing problems in attempted **a)** alkylation of lactam with phenethyl bromide, and **b)** enolisation of lactone (green). Interfering, acidic protons are highlighted in red.

Conclusions.

In summary, translation of organic-chemical knowledge into machine readable rules is much more than just writing out SMILES/SMARTS strings describing reaction cores. The various types of considerations and calculations that accompany the coding process are, in fact, a modern embodiment of physical-organic chemistry and an interesting junction between this seasoned area of chemical research and contemporary computing methods. Above all, the protocols we strived to illustrate in this Perspective are a cornerstone on which all higher-level routines to find complete synthetic pathways rest – as we mentioned in the introduction and wish to reiterate here, machine's ability to design high-quality synthetic routes will be only as good as the quality of the underlying rules describing individual reactions.

Acknowledgements

The authors thank the U.S. DARPA for generous support under the "Make-It" Award, 69461-CH-DRP #W911NF1610384. B.A.G. also gratefully acknowledges personal support from the Institute for Basic Science Korea, Project Code IBS-R020-D1.

Conflicts of interest

While the Chematica retrosynthesis platform, mentioned in the text, was originally developed and owned by B.A.G.'s Grzybowski Scientific Inventions, LLC, neither he nor the co-authors no longer hold any stock in this company, which is now property of Merck KGaA, Darmstadt, Germany. The authors continue to collaborate with Merck within the DARPA "Make-It" award. All queries about access options to Chematica (now rebranded as Synthia^{TM}), including academic collaborations, should be directed to Dr. Sarah Trice at sarah.trice@sial.com.

References:

1. E. J. Corey and W. T. Wipke, *Science* 1969, **166**, 178–192.
2. E. Corey, A. Long and S. Rubenstein, *Science* 1985, **228**, 408–418.
3. E. J. Corey, R. D. Cramer and W. J. Howe, *J. Am. Chem. Soc.* 1972, **94**, 440–459.
4. T. H. Varkony, D. H. Smith and C. Djerassi, *Tetrahedron* 1978, **34**, 841–852.
5. I. Ugi, J. Bauer, R. Baumgartner, E. Fontain, D. Forstmeyer and S. Lohberger, *Pure Appl. Chem.* 1988, **60**, 1573–1586.
6. J. B. Hendrickson, D. L. Grier and A. G. Toczko, *J. Am. Chem. Soc.* 1985, **107**, 5228–5238.
7. S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk and B. A. Grzybowski, *Angew. Chem. Int. Ed.* 2016, **55**, 5904–5937.
8. M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem. Int. Ed.* 2005, **44**, 7263–7269.
9. 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 and B. A. Grzybowski, *Chem* 2018, **4**, 522–532.
10. J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.* 2009, **49**, 593–602.
11. ASKCOS, http://askcos.mit.edu, (accessed 04 June 2019).
12. M. H. S. Segler, M. Preuss and M. P. Waller, *Nature* 2018, **555**, 604–610.
13. M. Kowalik, C. M. Gothard, A. M. Drews, N. A. Gothard, A. Weckiewicz, P. E. Fuller, B. A. Grzybowski and K. J. M. Bishop, *Angew. Chem. Int. Ed.*, 2012, **51**, 7928–7932.

14. C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.* 2018, **51**, 1281–1289.

15. S. Lemonick, *C&EN Glob. Enterp.*, 2018, **96**, 16–20.

16. Reaxys, https://www.reaxys.com, (accessed 04 June 2019).

17. SciFinder, https://scifinder.cas.org, (accessed 04 June 2019).

18. SPRESI, http://www.infochem.de/products/databases/spresi.shtml, (accessed 04 June 2019).

19. D. M. Lowe, Chemical reactions from US pat.(1976-Sep2016), 2017, https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, (accessed 04 June 2019).

20. F. A. Carey and S. R. J., *Advanced Organic Chemistry*, Springer US, Boston, MA, 2007.

21. E. V. Anslyn and D. Dennis, *Modern Physical Organic Chemistry*, University Science Books, Herndon, VA, 2005.

22. S. Kenis, M. D'hooghe, G. Verniest, T. A. Dang Thi, C. Pham The, T. Van Nguyen and N. De Kimpe, *J. Org. Chem.* 2012, **77**, 5982–5992.

23. K. C. Nicolaou, G. Vassilikogiannakis, W. Mägerlein and R. Kranich, *Angew. Chem. Int. Ed.*, 2001, **40**, 2482–2486.

24. W. Jin, C. Coley, R. Barzilay and T. Jaakkola, *NIPS*, 2017, pp.2607–2616.

25. P. Schwaller, T. Gaudin, D. Lányi, C. Bekas and T. Laino, *Chem. Sci.* 2018, **9**, 6091–6098.

26. B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.* 2017, **3**, 1103–1113.

27. W. Beker, E. P. Gajewska, T. Badowski and B. A. Grzybowski, *Angew. Chemie Int. Ed.*, 2019, **58**, 4515–4519.

28. W. Jaworski, S. Szymkuć, B. Mikulak-Klucznik, K. Piecuch, T. Klucznik, M. Kaźmierowski, J. Rydzewski, A. Gambin and B. A. Grzybowski, *Nat. Commun.*, 2019, **10**, 1434.

29. I. A. Watson, J. Wang and C. A. Nicolaou, *J. Cheminform.* 2019, **11**, 1.

30. Y. Masaki, H. Arasaki and M. Shiro, *Chem. Lett.* 2000, **29**, 1180–1181.

31. K. Spielmann, R. M. de Figueiredo and J.-M. Campagne, *J. Org. Chem.* 2017, **82**, 4737–4743.

32. S. P. Chavan and H. S. Khatod, *Tetrahedron: Asymmetry* 2012, **23**, 1410–1415.

33. D. Mal, B. Senapati and P. Pahari, *Tetrahedron Lett.* 2006, **47**, 1071–1075.

34. W. R. Roush, M. R. Michaelides, D. F. Tai and W. K. M. Chong, *J. Am. Chem. Soc.* 1987, **109**, 7575–7577.

35. D. Mal and J. Roy, *Tetrahedron* 2015, **71**, 1247–1253.

36. D. Mal and J. Roy, *Org. Biomol. Chem.* 2015, **13**, 6344–6352.

37. J. R. Del Valle and M. Goodman, *J. Org. Chem.* 2003, **68**, 3923–3931.

38. H. Becker, M. A. Soler and K. Barry Sharpless, *Tetrahedron* 1995, **51**, 1345–1376.

39. A. Tanaka, H. Okamoto and M. Bersohn, *J. Chem. Inf. Model.*, 2010, **50**, 327–338.

40. R. R. Karimov, D. S. Tan and D. Y. Gin, *Tetrahedron* 2018, **74**, 3370–3383.

41. M. Gao, Y.-C. Wang, K.-R. Yang, W. He, X.-L. Yang and Z.-J. Yao, *Angew. Chem. Int. Ed.* 2018, **57**, 13313–13318.

42. T. Tsukiyama, A. Kinoshita, R. Ichinose and K. Sato, *Biosci. Biotechnol. Biochem.* 2002, **66**, 1407–1411.

43. B. M. Trost, M. R. Machacek and H. C. Tsui, *J. Am. Chem. Soc.* 2005, **127**, 7014–7024.

44. RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org/, (accessed 04 June 2019).

45. RDChiral. https://github.com/connorcoley/rdchiral, (accessed 04 June 2019).

46. J. R. Coombs, F. Haeffner, L. T. Kliman and J. P. Morken, *J. Am. Chem. Soc.* 2013, **135**, 11222–11231.

47. W. R. Roush, A. D. Palkowitz and M. J. Palmer, *J. Org. Chem.* 1987, **52**, 316–318.

48. H. Kim, S. Ho and J. L. Leighton, *J. Am. Chem. Soc.* 2011, **133**, 6517–6520.

49. R. W. Hoffmann, *Chem. Rev.* 1989, **89**, 1841–1860.

50. G. Schmid, T. Fukuyama, K. Akasaka and Y. Kishi, *J. Am. Chem. Soc.* 1979, **101**, 259–260.

51. J. Shi, H. Shigehisa, C. A. Guerrero, R. A. Shenvi, C.-C. Li and P. S. Baran, *Angew. Chem. Int. Ed.*, 2009, **48**, 4328–4331.

52. N.-H. Lin, L. E. Overman, M. H. Rabinowitz, L. A. Robinson, M. J. Sharp and J. Zablocki, *J. Am. Chem. Soc.* 1996, **118**, 9062–9072.

53. H. Chiba, S. Oishi, N. Fujii and H. Ohno, *Angew. Chem. Int. Ed.* 2012, **51**, 9169–9172.

54. G. Sabitha, P. AnkiReddy and S. Das, *Synthesis (Stuttg)* 2014, **47**, 330–342.

55. J.-F. Brazeau, A.-A. Guilbault, J. Kochuparampil, P. Mochirian and Y. Guindon, *Org. Lett.* 2010, **12**, 36–39.

56. L. C. Dias and A. G. Salles, *J. Org. Chem.* 2009, **74**, 5584–5589.

57. https://scifinder.cas.org/scifinder/view/link_v1/reaction.html?l=T5OKcO0Ri0Vxs5SgbYOjGp9biqO0mXFn0S569XvHjrkzEryFRL2Nr9NcFsIqYvcr, (accessed 04 June 2019).

58. W. Kleist, S. S. Pröckl, M. Drees, K. Köhler and L. Djakovitch, *J. Mol. Catal. A Chem.* 2009, **303**, 15–22.

59. B. Mishra, M. Neralkar and S. Hotha, *Angew. Chem. Int. Ed.* 2016, **55**, 7786–7791.

60. G. Bartoli, G. Palmieri, M. Bosco and R. Dalpozzo, *Tetrahedron Lett.*, 1989, **30**, 2129–2132.

61. K. J. M. Bishop, R. Klajn and B. A. Grzybowski, *Angew. Chem. Int. Ed.* 2006, **45**, 5348–5354.

62. T. Nakai and K. Mikami, in *Organic Reactions*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1994, pp. 105–209.

63. D. J. S. Tsai and M. M. Midland, *J. Org. Chem.* 1984, **49**, 1842–1843.

64. H. B. Bürgi, J. D. Dunitz, J. M. Lehn and G. Wipff, *Tetrahedron* 1974, **30**, 1563–1572.

65. C. H. Heathcock and L. A. Flippin, *J. Am. Chem. Soc.* 1983, **105**, 1667–1668.

66. E. P. Lodge and C. H. Heathcock, *J. Am. Chem. Soc.* 1987, **109**, 3353–3361.

67. J. G. Henkel and L. A. Spurlock, *J. Am. Chem. Soc.* 1973, **95**, 8339–8351.

68. H. Tan and J. H. Espenson, *J. Mol. Catal. A Chem.* 1999, **142**, 333–338.

69. M. Cakmak, P. Mayer and D. Trauner, *Nat. Chem.* 2011, **3**, 543–545.

70. V. Barát, D. Csókás and R. W. Bates, *J. Org. Chem.* 2018, **83**, 9088–9095.

71. K. Molga, P. Dittwald and B. A. Grzybowski, *Chem* 2019, **5**, 460–473.

72. J. C. Kromann, J. H. Jensen, M. Kruszyk, M. Jessing and M. Jørgensen, *Chem. Sci.* 2018, **9**, 660–665.

73. A. Tomberg, M. J. Johansson and P.-O. Norrby, *J. Org. Chem.*, 2019, **84**, 4695–4703.

74. F. A. Van-Catledge, *J. Org. Chem.* 1980, **45**, 4801–4802.

75. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science* 2018, **360**, 186–190.

76. K. V. Chuang and M. J. Keiser, *Science* 2018, **362**, eaat8603.

77. K. C. Nicolaou, S. A. Snyder, T. Montagnon and G. Vassilikogiannakis, *Angew. Chem. Int. Ed.* 2002, **41**, 1668–1698.

78. M. Juhl and D. Tanner, *Chem. Soc. Rev.* 2009, **38**, 2983–2992.

79. C. Hansch, A. Leo and R. W. Taft, *Chem. Rev.* 1991, **91**, 165–195.

80. C. Cao and L. Liu, *J. Chem. Inf. Comput. Sci.* 2004, **44**, 678–687.

81. D. Rogers and M. Hahn, *J. Chem. Inf. Model.* 2010, **50**, 742–754.

82. S. R. Crosby, J. R. Harding, C. D. King, G. D. Parker and C. L. Willis, *Org. Lett.* 2002, **4**, 577–580.

83. L. Li, S. Gou and F. Liu, *Tetrahedron: Asymmetry* 2014, **25**, 193–197.

84. S. Luo, H. Xu, J. Li, L. Zhang, X. Mi, X. Zheng and J.-P. Cheng, *Tetrahedron* 2007, **63**, 11307–11314.

85. M. C. Hillier, J.-N. Desrosiers, J.-F. Marcoux and E. J. J. Grabowski, *Org. Lett.* 2004, **6**, 573–576.

86. R. F. C. Brown, W. R. Jackson and T. D. McCarthy, *Tetrahedron Lett.* 1993, **34**, 1195–1196.

87. K. K. Gnanasekaran, J. Yoon and R. A. Bunce, *Tetrahedron Lett.* 2016, **57**, 3190–3193.

88. E. M. Carreira and L. Kvaerno, *Classics in Stereoselective Synthesis*. Wiley – VCH, Weinheim, 2009.

89. Y. Natori, M. Ito, M. Anada, H. Nambu and S. Hashimoto, *Tetrahedron Lett.* 2015, **56**, 4324–4327.

90. P. Etayo, R. Badorrey, M. D. Díaz-de-Villegas and J. A. Gálvez, *Chem. Commun.* 2006, 3420–3422.

91. K. Csatayová, I. Špánik, V. Ďurišová and P. Szolcsányi, *Tetrahedron Lett.* 2010, **51**, 6611–6614.

92. J. Wu, H. Yu, Y. Wang, X. Xing and W.-M. Dai, *Tetrahedron Lett.* 2007, **48**, 6543–6547.

93. T. Fujii, S. Yoshifuji and K. Yamada, *Chem. Pharm. Bull. (Tokyo)* 1978, **26**, 2071–2080.

94. E. J. Horn, J. S. Silverston and C. D. Vanderwal, *J. Org. Chem.* 2016, **81**, 1819–1838.

95. F. G. Bordwell, *Acc. Chem. Res.* 1988, **21**, 456–463.

96. R. Fraczkiewicz, M. Lobell, A. H. Göller, U. Krenz, R. Schoenneis, R. D. Clark and A. Hillisch, *J. Chem. Inf. Model.* 2015, **55**, 389–397.

97. C. Liao and M. C. Nicklaus, *J. Chem. Inf. Model.* 2009, **49**, 2801–2812.

98. J. C. Shelley, A. Cholleti, L. L. Frye, J. R. Greenwood, M. R. Timlin and M. Uchimaya, *J. Comput. Aided. Mol. Des.* 2007, **21**, 681–691.

99. A. D. Bochevarov, M. A. Watson, J. R. Greenwood and D. M. Philipp, *J. Chem. Theory Comput.* 2016, **12**, 6001–6019.