ELSEVIER

Teaser To be able to predict chemical reactions is of the utmost importance for the pharmaceutical industry. Recent trends and developments are reviewed for reaction mining, computer-assisted synthesis planning, and QM methods, with an emphasis on collaborative opportunities.

Computational prediction of chemical reactions: current status and outlook

Ola Engkvist[1], Per-Ola Norrby[2], Nidhal Selmi[1], Yu-hong Lam[3], Zhengwei Peng[3], Edward C. Sherer[3], Willi Amberg[4], Thomas Erhard[4] and Lynette A. Smyth[4]

[1] Discovery Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, SE-43183 Mölndal, Sweden
[2] Pharmaceutical Sciences, Innovative Medicines and Early Development Biotech Unit, AstraZeneca R&D Gothenburg, SE-43183 Mölndal, Sweden
[3] Modeling and Informatics, MRL, Merck & Co., Rahway, NJ 07065, USA
[4] AbbVie Deutschland GmbH & Co. KG, Neuroscience Discovery, Medicinal Chemistry, Knollstrasse, 67061 Ludwigshafen, Germany

Over the past few decades, various computational methods have become increasingly important for discovering and developing novel drugs. Computational prediction of chemical reactions is a key part of an efficient drug discovery process. In this review, we discuss important parts of this field, with a focus on utilizing reaction data to build predictive models, the existing programs for synthesis prediction, and usage of quantum mechanics and molecular mechanics (QM/MM) to explore chemical reactions. We also outline potential future developments with an emphasis on pre-competitive collaboration opportunities.

Introduction

Small organic molecules are the bread and butter of drug discovery. To synthesize these small organic molecules, reaction predictions are practiced routinely by medicinal chemists, who make diverse sets of molecules on a small scale to efficiently probe the structure–activity relationship (SAR) through the design–make–test–analyze cycle, and by process chemists, who intend to discover the most efficient, cost-effective, and environmentally green routes to synthesize late-stage drug candidates in larger quantities. As such, the effectiveness of reaction prediction is a key factor contributing to the efficiency and success of drug discovery and development. Therefore, it is no surprise that there are many *in silico* tools available to assist chemists in reaction prediction and that this area has remained active in terms of research and development, especially in recent years. We have come together in a precompetitive fashion to further discussion of how the larger community can drive additional development in this space through data sharing and collaboration.

Corresponding author: Engkvist, O. (Ola.Engkvist@astrazeneca.com)

Ola Engkvist was awarded his PhD in computational chemistry by the University of Lund in 1997, and continued with postdoctoral research at the University of Cambridge and the Czech Academy of Sciences. Between 2000 and 2004, he worked for two biotech companies before joining AstraZeneca in Gothenburg, Sweden, in late 2004. He is currently a team leader in cheminformatics within the Discovery Sciences sector. His research interests include cheminformatics, molecular modeling, machine learning, phenotypic screening, and open innovation.

Yu-hong Lam was awarded his M. Chem. and PhD by the University of Oxford under the guidance of Veronique Gouverneur working in organofluorine chemistry. After he graduated, he carried out postdoctoral research at UCLA with Ken Houk in computational catalyst design and reaction discovery. He is currently a senior scientist at Merck Research Laboratories, and his research involves the collaborative application of modeling and informatics to streamline organic synthesis.

Willi Amberg studied at ETH Zfirich and was awarded his PhD in organic chemistry in 1989, followed by postdoctoral research at MIT and Scripps Research Institute. In 1992, he joined BASF Pharma, working in oncology and cardiovascular research, before taking up his current position as a group leader in neuroscience at Abbott GmbH and later at AbbVie GmbH. Ludwigshafen, Germany. His research activities are mainly focused on therapies for schizophrenia and Alzheimer's disease, and his interests include medicinal chemistry and drug design.

https://doi.org/10.1016/j.drudis.2018.02.014

There are several types of question to be addressed by reaction predictions: (i) forward reaction prediction: given a set of reaction building blocks, what could be the potential products? Which one might be the major product? What might be the most favorable reaction condition(s) for the putative major product? What is the potential yield of the putative major product? (ii) Retrosynthetic analysis: given a desired molecule, what are the possible synthetic route(s) to make this molecule based on available reaction building blocks on hand? How can we rank and filter these possible synthetic routines according to user-defined criteria? (iii) Reaction mechanism elucidation: given an overall reaction, what could the fundamental mechanistic reaction steps be? What are the major factors determining product yield or stereo- and regioselectivity?

Here, we discuss tools and methods to address these three types of question, with a focus on: (i) the latest machine learning (ML) approaches for both forward reaction prediction and retrosynthetic analysis; (ii) the utility of retrosynthetic analysis tools in the eyes of medicinal and process chemists; and (iii) the state of the art and outstanding problems in the application of quantum chemical calculations to elucidate reaction mechanisms, origins of selectivity, and spectroscopic properties.

Reaction knowledge mining

Background

Here, we focus on recent development in cheminformatics to better use historical reaction data for predicting synthetic pathways for novel molecules. With more sophisticated methods to extract data from in-house and literature sources, reaction knowledge mining is entering the 'big-data' era. This, in combination with ML methods, is creating a step change in the application of reaction knowledge mining.

Data sources, standardization, extraction, and reaction classification

As depicted in Fig. 1a, reaction data already published in scientific journals and patent literature are generally extracted, curated, aggregated, and hosted by data vendors and made available for users to access through vendors' proprietary tools (e.g., SciFinder from Chemical Abstracts Services and Reaxys from Elsevier). The vendor-provided reaction databases are not discussed here, because they have been recently reviewed elsewhere [1]. In general, end users do not have direct access to the full set of vendor reaction data for reaction knowledge mining. Only recently have several academic groups published reaction mining and predictive modeling works based on the reaction data content in *Reaxys*. Their work is discussed in the section titled 'Predictive reaction modeling'.

For proprietary reaction content generated by biotech, pharma, and chemical companies, it is common to have corporate electronic laboratory notebooks (ELN) for data and intellectual property (IP) capture (Fig. 1). Given that ELN applications are mainly designed for data and IP capture, they are not ideal environments for knowledge mining in general.

To perform in-depth reaction knowledge mining to address specific scientific questions using cheminformatics tools, the reaction data have to be hosted in an IT environment that is easy to access and of high performance (Fig. 1). AstraZeneca reported the successful extraction of its MedChem ELN pages and loaded them into an internal reaction DataMart to support web searches and other external applications [2]. In 2013, Roche reported that they had collaborated with both Elsevier and NextMove to extract reaction data content from more than ten internal reaction databases and its corporate chemistry ELN, and combined them with public reaction content from Elsevier to form an integrated reaction DataMart hosted behind the Roche firewall. Roche scientists can use a customized version of *Reaxys* to search and browse all of these reaction data sources in an integrated and streamlined manner. In addition, the integrated DataMart provides a larger and richer set of reactions to enable more powerful and effective knowledge mining [3].

The *HazELNut* suite of tools from NextMove Software is commonly used to extract reaction content from vendor-provided ELN systems, perform format conversion and data curation to fix common data entry issues seen in ELNs, and add additional annotation, such as reaction classification (Fig. 1b, [4,5]). These operations directly benefit downstream operations, such as knowledge mining and predictive model building. In addition to commercial software tools, there are open-source software tools available for basic reaction analysis [6].

Once the corporate ELN content is extracted and stored in a minable format, knowledge mining can be applied to address questions such as: (i) how many syntheses have been attempted using named reactions (e.g., Suzuki aryl C–N coupling reaction and Buchwald–Hartwig aryl C–N coupling reactions)? (ii) What are the distributions and trends observed in terms of success rate in these reactions? And, (iii) how frequently has a reaction building block been used for named reactions, and what were the associated reaction success rates [2,5]? This type of information can be readily used by chemists to make more-informed decisions during compound design and building-block selection. More in-depth analysis of knowledge mining has led to the publication of a set of most robust and commonly used reactions by the medicinal chemistry community [7] and extracted reaction rules to support either retrosynthetic analysis or reaction-based virtual library enumeration [8].

Scientific journals and patent literature are biased against negative data [9,10] and the same is expected to be true for the published reaction content. By contrast, corporate ELNs do contain negative data (failed reactions). Without such a bias against negative data, ELN reaction content is expected to be more suited for knowledge mining and predictive model building. However, even with millions of reaction records inside a typical corporate ELN system, the vast chemical reaction space (defined by reaction type, reactants, products, and more variables in reaction conditions) is still only sparsely explored [11]. Recent advances in miniaturization (down to the nanomolar scale) and workflow streamlining have demonstrated the potential to explore reaction space in a more systematic and well-controlled way with higher throughput [12]. The age of big-data might have finally arrived for organic synthesis [13]. It is also encouraging that a nonproprietary format has been developed (RInChI) for handling chemical reactions [14].

Predictive reaction modeling: machine learning

Given the increased availability of reaction data, reflected both in the number of different reactions and various successful conditions for a specific reaction, it is not surprising that there have been

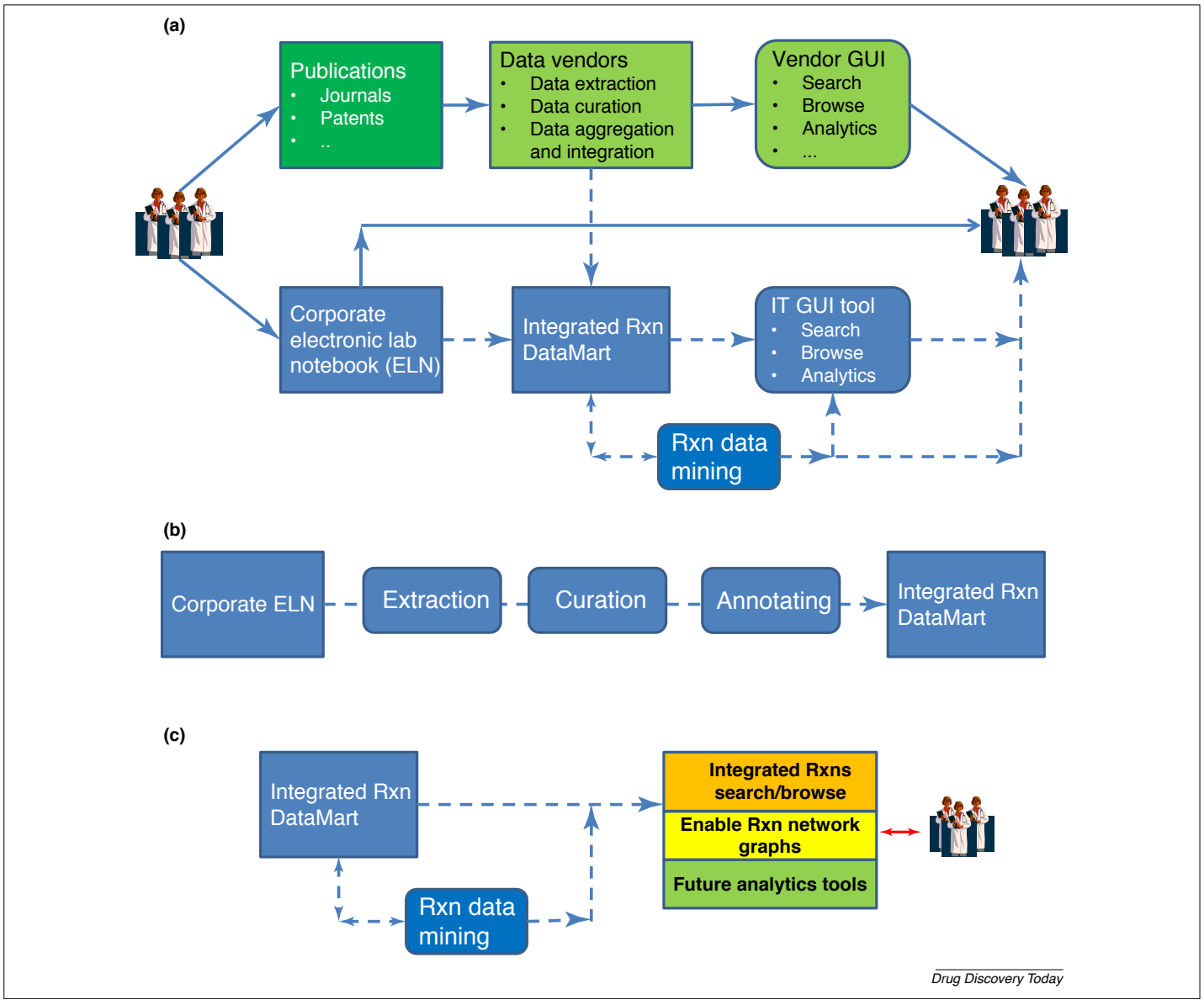

FIGURE 1

Generation, processing, and consumption of chemical reaction data content. (a) Experimental reaction data are either published in scientific journals of patents or captured by proprietary corporate electronic laboratory notebooks (ELNs). Rxn data go through various data extraction, curation, transformation, aggregation, and integration procedures before they can be searched and browsed by the broader scientific community. Finally, Rxn data mining can be used to address in-depth questions from scientists. Solid arrows represent the common practice, and dotted arrows represent more-advanced data flow with limited adoption. (b) Corporate ELN data captured or extracted, cleaned, and annotated before being loaded into a Rxn DataMart to facilitate high-performance search/browse/data mining and (c) exploitation.

several publications during recent years applying ML models to predict both which building blocks will react and under which conditions. Here, we discuss ML approaches, whereas other methods, such as rule-based and QM methods, are discussed in other sections. Besides setting up the input reaction data set and the ML algorithm, an important choice is how to define the chemical descriptors used to model the reaction. Both 2D reaction fingerprints as well as 3D QM descriptors, and levels in between, have been used. A difference between reaction prediction models and normal quantitative (Q)SAR modeling is that, in many cases, reactions are predicted in two steps, combining a feasibility prediction with a ranking of the feasible reactions. Although advanced ML methods have been applied in retrosynthetic analysis, good results have been obtained with similarity searches based on reaction networks extracted from the Beilstein database [15].

Baldi *et al.* used a two-step process combining molecular orbital (MO) theory and ML [16–18]. However, it is not clear how MO approaches will work for more complex reactions, such as metal-catalyzed reactions. Similar methods have been used by other groups [19,20].

Recently, deep learning methods have become popular outside of the chemistry community. One impressive example is the victory over a world champion in Go [21]. It is natural that deep learning is applied in drug discovery and in predicting reactions. Based on 15 000 reactions from granted US patents, Coley *et al.*

applied a neural network model to prioritize reaction candidates to predict products. In a fivefold cross-validation, the trained model assigned the major product rank 1 in 71.8% of cases, rank ≤3 in 86.7% of cases, and rank ≤5 in 90.8% of cases [22].

Another recent study created reaction fingerprints by concatenating the individual fingerprints of the reactants and reagents. The reaction fingerprints were used to predict the most likely reaction for a set of reactions derived from textbooks [23].

Segler *et al*. performed reaction modeling in a different way. They used two sets of reactions (hand-coded and automatically extracted) to classify reactions in *Reaxys*. With the hand-coded reactions, they classified 3 million reactions and with the automatically extracted reactions they classified almost 5 million reactions. Classification models were trained with neural networks and either the reactants or products were described with ECFP4 fingerprints depending on whether the models were used for reaction modeling or retrosynthetic analysis. The results with the neural networks were superior to those of rule-based methods, reaching 95% and 97% top-ten accuracies, respectively for both retrosynthesis and reaction prediction on a validation set of approximately 1 million reactions [24]. Although several publications have focused on predicting which the most likely product is given a set of reactants, less emphasis has been placed on predicting reaction conditions for well-known and established reactions. An exception is the study of Michael-type reactions by Varnek and co-workers [25]. Varnek *et al*. also assessed the reactivity of protecting groups [26]. Segler *et al*. showed that it is possible to invent new reactions when access to a large set of reactions from literature is available [27].

Reaction graphs for retrosynthetic analysis

Retrosynthetic analysis has been heavily dominated by rule-based methods and only recently was an alternative ML method presented. Segler *et al*. developed a system inspired by the program AlphaGo that beat a Go master [28]. As in Go (and different from chess), there are no good heuristics available in retrosynthetic analysis to estimate whether a specific reaction is useful. Thus, all options in the reaction graph from the final molecule back to the set of available building blocks need to be traversed. Starting from the molecule, reaction paths were investigated using a Monte Carlo Tree Search (MCTS) until a set of existing building blocks was identified. MCTS outperformed Best-First Search with rule-based heuristics, which have been used previously in retrosynthetic analysis [28]. There are also other examples available elsewhere [29].

Concluding remarks and outlook

The availability of big data sets, and ML to handle these data sets, has enabled remarkable progress during recent years and has the potential to transform society. Therefore, it is not surprising that the same trend can be seen within drug discovery and fields such as reaction knowledge mining. Recently, the number of relevant articles has increased significantly and the use of both large data sets and ML methods has been emphasized. Reaction knowledge mining is a true big-data field: the more reactions that are used to train the ML methods, the better the ML model will perform in terms of reaction prediction, reaction condition prediction, and retrosynthetic analysis.

To enable downstream reaction analytics and ML for predictive reaction modeling, it is important to capture all relevant data accurately. We envision the following: (i) future versions of chemistry ELNs will provide additional on-the-fly data checking/validation during the data entry stage to reduce and/or eliminate the need for downstream data clean-up; (ii) better role assignment for chemical samples used in reactions (e.g., reactants, reagents, acid/base, solvent, etc.) during the ELN data entry stage; (iii) systematic capture of reaction conditions and reaction steps in a machine-minable format; and (iv) the vendors of published reaction contents will be encouraged to allow their data content to be licensed and integrated with the proprietary reaction content from their customers. One need that has already been identified by the research community to facilitate reaction data exchange and integration is an open-source and standard reaction data exchange format. A reaction data exchange format, called Unified Data Model (UDM), is being actively worked on by Pistoia Alliance, an industrial precompetitive consortium [30].

Despite the progress made over the past decade, there are still several bottlenecks that should be addressed to progress the field further. There is still a need for a proper high-quality set of successful and failed reactions in the public domain that can be used for benchmarking, and for different ways of describing reactions and ML methods. If it were possible to share reaction data in a secure way without compromising IP, this would be beneficial to further improve predictions. We propose initiating the formation of precompetitive collaborations or consortiums to pursue building out data sets and new tools.

Computer-assisted retrosynthetic analysis tools

Computer-assisted synthesis planning

The art of performing retrosynthesis can be best explained as an iterative process of moving backwards from a particular target molecule in a synthesis tree of possible reactions. The ultimate goal is to identify viable chemical routes to available starting materials or published compounds. Computers have had a pivotal role in facilitating synthesis planning. There is a fundamental need for access to database sources that store all published reactions together with the available chemical matter and to make these searchable by researchers.

Here, we look at three commercially available retrosynthetic planning tools. Software tools, such as *SciFinder* and *Reaxys*, with their modules *SciPlanner* [31] and *Synthesis Planner* [32], are not considered as *de novo* retrosynthetic computer-assisted synthesis planning (CASP) tools in the scope of this review, because both use literature examples and do not attempt to identify reaction centers, and so on. Teaching a computer the expert knowledge of experienced organic chemists and their decision-making has been one of the major challenges in this field.

It is now almost 50 years ago since Corey and Wipke published their pioneering work in 1969 [33], which led to the development of Logic and Heuristics Applied to Synthetic Analysis (LHASA), the first 'true' retrosynthetic planning tool. It was an expert system and still needed manual input, but its general concepts are still found in programs today [34]. These implemented modules allowed the use of reaction transforms, mechanistic transforms, overlapping structure goals, and topological, stereochemical, and functional group-based strategies. Recent reviews summarizing

the history and past developments in this field are recommended for further reading [1,35–37].

Expectations of a retrosynthetic software tool

The ideal retrosynthesis tool

A small group of organic chemists from various companies and different areas (e.g., process and research chemists) was asked to define what their expectations would be of the ideal retrosynthetic tool. Based on the results from a survey (13 chemists, two companies, middle of 2017), the consolidated most-important factors required by a retrosynthetic program are that: (i) the program is user friendly; (ii) literature examples are given for suggested routes; (iii) the user can define single or multiple bond breaks and the order in which they should be broken; (iv) routes lead to commercially available building blocks; (v) unstable intermediates or problematic functional groups should be recognized and, optimally, a protecting group or synthetic strategy suggested to overcome these challenges; and (vi) a suitable scoring system is available to prioritize the results.

Other desirable options for such a program mentioned were that: (i) there is the possibility to capture in-house data and use them in addition to data already in the system (reactions and available building blocks); (ii) the user can define the starting materials for the synthesis; (iii) published full syntheses of target compounds should be shown as top scores; (iv) the program can recognize and suggest ideas for stereochemical transformations and isotope chemistry; (v) a diversity of routes is displayed; and (vi) the program allows exclusion of technologies that are not available on site (e.g., photochemistry).

The filtering and/or scoring options account for the largest difference between the requirements of chemists using the program. A process chemist, for example, might require a reliable cost and yield-driven synthesis for multiple (kilo) grams of material, whereas a research chemist might prefer a synthesis that is short and has high flexibility, with as many diverse substitutions as possible. Both groups of chemists might want to use completely different possibilities for filtering and/or scoring, which might include: (i) likelihood of success (number of examples present, literature yields for similar transformations); (ii) ease of chemistry (based on reaction type, multiple products); (iii) number of steps; (iv) green chemistry aspects; (v) availability of starting materials; (vi) cost of starting materials/entire route with solvents, and so on; and (vii) risk assessment of routes (toxicity, regulated substances).

Thus, an 'ideal retrosynthesis tool' will be difficult to develop because every group of users has its specific requirements and preferences. The ideal tool would be flexible, but address many of the main points important to most users. Here, we focus on the most important factors as defined above for all three retrosynthetic programs discussed below, and the order of description follows their commercial launch.

ICSYNTH

The computer-aided design tool IC*SYNTH* is a retrosynthesis software developed by InfoChem [38] and commercialized since 2005. The current version, 3.1, was launched in October 2017. The retrosynthesis synthetic analysis performed by IC*SYNTH* is founded on rule-based methods [39].

Data sources and how it works

The software is built around libraries of retrosynthetic transformations that have been generated from diverse literature and in-house sources. An atom-mapping and reaction center identification algorithm applied to reaction abstract data allows the creation of chemical transformation rules (transform) describing the combination of bond breaking and making in a given retroreaction. Atom-centered stereochemical information is captured and encoded in this data-processing step. All transforms generated are grouped by source of origins to form libraries that can be selected by the user while setting up a new search.

The largest transform libraries are derived from the 4.4 million literature reactions extracted from scientific literature and patents, covering the years 1974–2012 in the Speicherung und Recherche Strukturchemischer Information database (SPRESI) [40]. The Fundamental Organic Reactions (FOR) collection contains a set of 210 000 well-known reliable reactions, extracted from books and journals. Finally, the Name reaction transform library contains transforms generated from the most common name-reactions (5800 example reactions) selected from a variety of textbooks. These transform libraries are integral part of the IC*SYNTH* license. Additionally, transform libraries can be generated from any proprietary and/or commercial reaction database, provided valid licenses for the data are given.

InfoChem also provides software to customers for the in-house generation of transform libraries based on their own confidential reactions (e.g., from ELN data). These libraries of data can be interfaced with IC*SYNTH* searches alongside the provided libraries or as stand-alone searches.

Setting up a search

To prevent or favor certain disconnections, atoms and bonds can be specifically tagged as remaining constant or as reacting areas of the molecule. Once a structure is entered, it can be stored on the server for repeat study or used as template.

With version 3.1, IC*SYNTH* offers a simplified query form for beginners, but skilled users can use the expert form (Fig. 2) to define different parameters when setting up a new query. The first step involves choosing which transform libraries to use. Retrosynthesis results in IC*SYNTH* are presented via a synthesis tree, which requires, in a second step, the definition of dimensions of that tree by selecting a number of retro-steps, between one and ten, and a number of suggestions per step, between one and 5000. The third step involves setting up the precision of the search to low, medium, or high. A high-precision search will look for closely matching templates in the transform libraries, whereas a low-precision search will retrieve more fuzzy results but can generate more innovative ideas and unexpected and/or speculative suggestions. Finally, the last step entails selecting a disconnection strategy, which defines search and postprocessing evaluation protocols that determine the ranking of output suggestions.

Interpreting results

Suggestions returned with unacceptable structures, such as overly strained ring systems and unstable substructures, are identified and filtered out of the tree based on structural matches with predefined unwanted substructures. The synthesis tree summarizes all remaining plausible disconnections.

The display is interactive, any precursor can be blocked, and the search for suggested precursors can be continued on the same tree

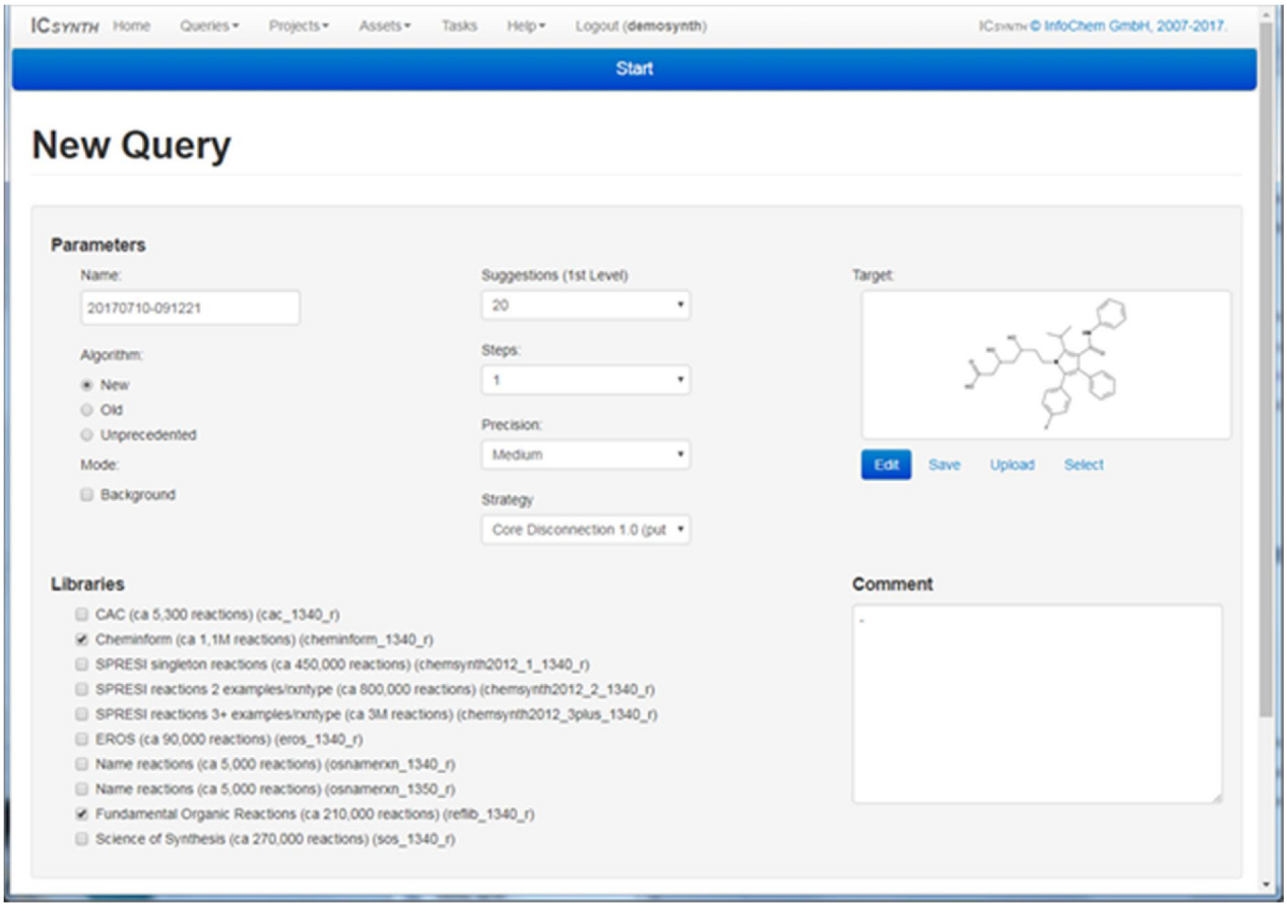

FIGURE 2

Setting up the search in IC*SYNTH*.

or selected to start a new one. Clicking on any of the molecules in the tree produces a separate reaction window presenting the suggested hypothetical forward reaction to the target or a suggested precursor, along with details on published precedent reactions and active links to the original literature.

The recently launched version 3.1 of IC*SYNTH* offers new visualization possibilities. Users can now also work via a reaction graph, where precursors, reactions, and final products appear as linked nodes (Figs 2 and 3). The graph is completely interactive and can self-reorganize to always provide the best possible view. A side panel provides detailed information about reactions and precursors when clicking on a node.

The new reaction graph visualization can be shared with team members, who can annotate and comment directly in the web application.

Applications

In 2015, InfoChem and AstraZeneca published the first comparison, conducted under controlled conditions, of organic chemists with and without a CASP tool, tasked with the retrosynthesis analyses of a series of target molecules [41]. A three-step workflow was devised for the routine usage of IC*SYNTH*: idea generation via the software; idea evaluation by chemists; and finally, detailed quantitative route evaluation covering aspects such as cost and greenness metrics. Through five case studies, it was demonstrated that IC*SYNTH* could provide new ideas that had not been previously considered by synthetic chemistry experts. Non-intuitive disconnections and nonobvious chemical reaction sequences were highlighted and led to the development of new routes.

ICFRP

Forward Reaction Prediction 5 (IC*FRP*) is a newer tool that operates in the opposite direction to IC*SYNTH*. It predicts reactions of a user's target molecule in the forward direction. The basic underlying reaction databases and technology are the same as IC*SYNTH*, with modifications to reflect intrinsic differences between retrosynthesis and reactivity. The status of IC*FRP* is currently experimental and is being developed in collaboration with the pharmaceutical industry, where applications to medicinal chemistry molecular design are evolving [42].

Chematica

The computer-assisted software package of *Chematica* was developed and commercialized in 2013 by Grzybowski Scientific Inventions [43], which recently became a part of the Merck KGaA business. *Chematica* allows both efficient exploration and scoring of reactions across the published chemical space using network analysis. It further can carry out *de novo* retrosynthetic design in a manual or fully automated mode based on expert-curated reaction rules (*Syntaurus*) [44].

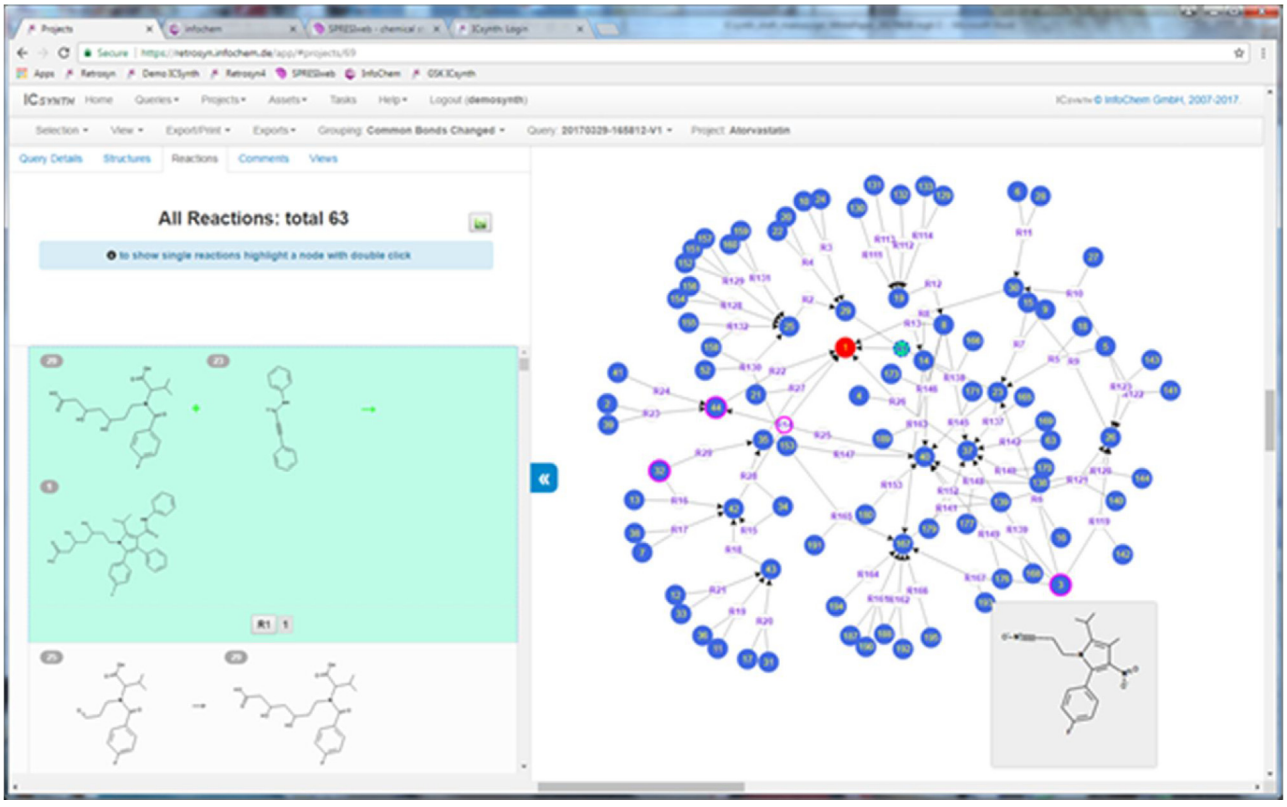

FIGURE 3

Visualization of search results with dynamic reaction graph from IC*SYNTH*.

Data sources and how it works

The proprietary knowledge base of the software, known as Network of Organic Chemistry (NOC), currently contains approximately 10 million compounds and a similar number of connecting reactions. Commercially available substances, along with their properties, are extracted from vendor catalogs and matched to NOC, which is, by default, connected to the Sigma-Aldrich collection, although other references can be easily added.

The core of *Syntaurus* is a collection of reaction rules that are applied during the synthetic planning process. With the knowledge of well-trained chemists involved during transform generation and data curation, the development of the knowledge base of *Syntaurus* culminated in more than 440 00+ so-called 'expert'-coded transforms reported as SMILES/SMARTS strings (SMILES: simplified molecular input line entry system; SMARTS: smiles arbitrary target specification). Each reaction ID contains not only its underlying synthetic fingerprint, but also information on functional groups that require protection or are not tolerated under typical reaction conditions and/or cause cross-reactivity conflicts. For each reaction rule, literature references describing the corresponding type of chemistry are included. Importantly, to overcome problems related to tracking of stereochemical changes and to ultimately ascribe proper reaction regiochemistry, two software modules, namely stereofix and regiofix, have been developed and implemented. Various sets of less reliable (i.e., machine-extracted) rules for heterocycle (30 000) and arene chemistry (100 00+), as well as a so-called 'specialized collection' (1 200 00+) can also be found in the software. Still, the vendor recommends that the users apply the auxiliary databases only after they have worked with the default expert database first. Moreover, the removal of structural inconsistencies among its predictions and a rapid evaluation of electron density of (hetero)arenes by simultaneous Hueckel calculations complete the *Syntaurus* toolbox.

Setting up a search

Chematica uses the *Marvin Sketch* editing tool, which enables drawing or uploading of structures from several commonly used file formats. Furthermore, structures can be generated from SMILES or automatically added by Beilstein Registry Number (BRN), Chemical Abstracts Service REGISTRY Number (CAS RN), or chemical name if known and indexed.

Having defined the target molecule, *Chematica* allows for two entirely independent search options. The first retrieves exclusively published reactions for experimentally synthesized molecules and the second performs a *de novo* retrosynthesis either manually step-by-step or fully automated. As a first step, the retrosynthetic rules to be used are chosen. The second step defines more precisely how *Syntaurus* evaluates synthetic choices and scores different pathways within the generated synthesis tree. Therefore, *Chematica* uses two different scoring functions in each chemical step: a chemical scoring function (CSF, 'synthetic position') and a reaction scoring function (RSF, 'synthetic move'). A lower total score indicates a more favorable pathway according to the user's predefined criteria (i.e., the scores are designed to keep track of the penalties). Both scoring functions are fully customizable to the

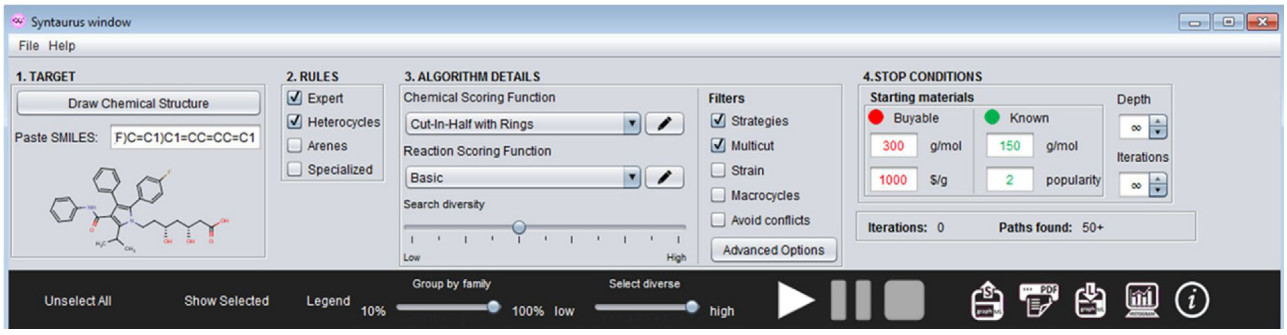

FIGURE 4

Setting up the search in *Syntaurus*.

desired search strategy, thus requiring more in-depth knowledge of the syntax used. A slider bar allows further adjustment of the degree of diversity of retrieved synthetic solutions. The final step involves the definition of reasonable stop criteria. The retrosynthetic pathway is terminated once commercially available fragments below a certain size (molecular weight) or cost (price per gram) are identified. NOC-known molecules below a certain size and/or with a certain minimal popularity score (how many times made) are further implemented as stop conditions. By default, *Syntaurus* is not restricted regarding the number of retrosynthetic steps or iterations, but both values can be limited (Fig. 4).

After the search is started, it can be paused at any time to review the first-generation results as the search resumes in the background.

Interpreting results

Suggestions returned by *Syntaurus* are ranked according to their overall score, which is the sum of CSF+RSF for all performed steps. Pathways with the lowest (best) score are shown on top of the list and thought to be most synergistic with the user's search strategy. For each list entry, a synthesis tree is generated (i.e., network graph representation), which can be manually inspected (Fig. 5). Clustering of search results 'by family' is useful to emphasize structurally diverse results.

A legend and the color coding help to navigate through the synthesis tree. The 'i'-button releases detailed information on the cost of commercially available starting materials and the popularity score of known chemicals. Pathway export generates PDF reports with all chemical information available, literature references, the search strategy, and basic transform information.

Applications

The developers of *Chematica* report examples for both of its synthesis modules [44]. A section devoted to synthesis optimization with constraints (SOCS) demonstrates search applications for 'optimal synthesis routes' within the experimentally verified NOC data. Depending on all the criteria imposed by the research scientists, the software can combine reaction data from several sources and propose an optimal route that, for instance could be most cost effective, avoids use of regulated or toxic starting materials, or is selective for intermediates and the product. Furthermore, the ability of *Chematica* to design novel synthetic routes was demonstrated both in the manual step-by-step mode and in the fully automated mode.

ChemPlanner

ChemPlanner from Wiley has been on the market since 2015 and is the successor of the previously known *ARChem*. It is a rule-based software that uses the ChemInform Reaction Database (CIRX) as its data source. In June 2017, Wiley licensed this technology and related content to CAS as part of a collaboration that will integrate *ChemPlanner* into *SciFinder*^{n} [45] and augment it with the larger CAS reaction content collection in addition to the information in CIRX [46].

Data sources and how it works

ChemPlanner generates its rule knowledge base through the use of CIRX, which contains over 2 million reactions and covers reaction data from 1990 to the present. Company reactions can be added to the system as well as in-house compound collections. Commercially available compounds are taken from several sources.

Rules generation is fully automated and begins with the loading of mapped reactions into the database. Molecular properties are then perceived (e.g., aromaticity, functional groups) and reaction cores are identified and extracted (bonds that change or are made or broken during the reaction). An extended core is then determined automatically as a third step. This includes the surrounding atoms and functional groups that are important for the reaction (e.g., where a carbonyl group is required for a reaction to proceed). Reactions that are perceived to share the same underlying chemistry are clustered together in a fourth step. Electronic properties are taken into account and leaving groups and functional groups on the reaction core are interchangeable as long as their electronic properties are similar. Generic rules are generated for each cluster and these are refined by selecting representative functional groups to replace the generic ones [39]. Further analysis allows the addition of more information (e.g., regioselectivity) for certain classes of reactions, and manual curation is carried out to compensate for any data and algorithmic limitations. By this procedure, the underlying database gives rise to more than 100 000 rules, which can then be used to propose syntheses for unknown molecules [47].

Setting up a search

ChemPlanner has multiple options for drawing and uploading molecules for retrosynthesis. After the structure has been entered, a known reaction search or the creation of a synthesis plan can be initiated.

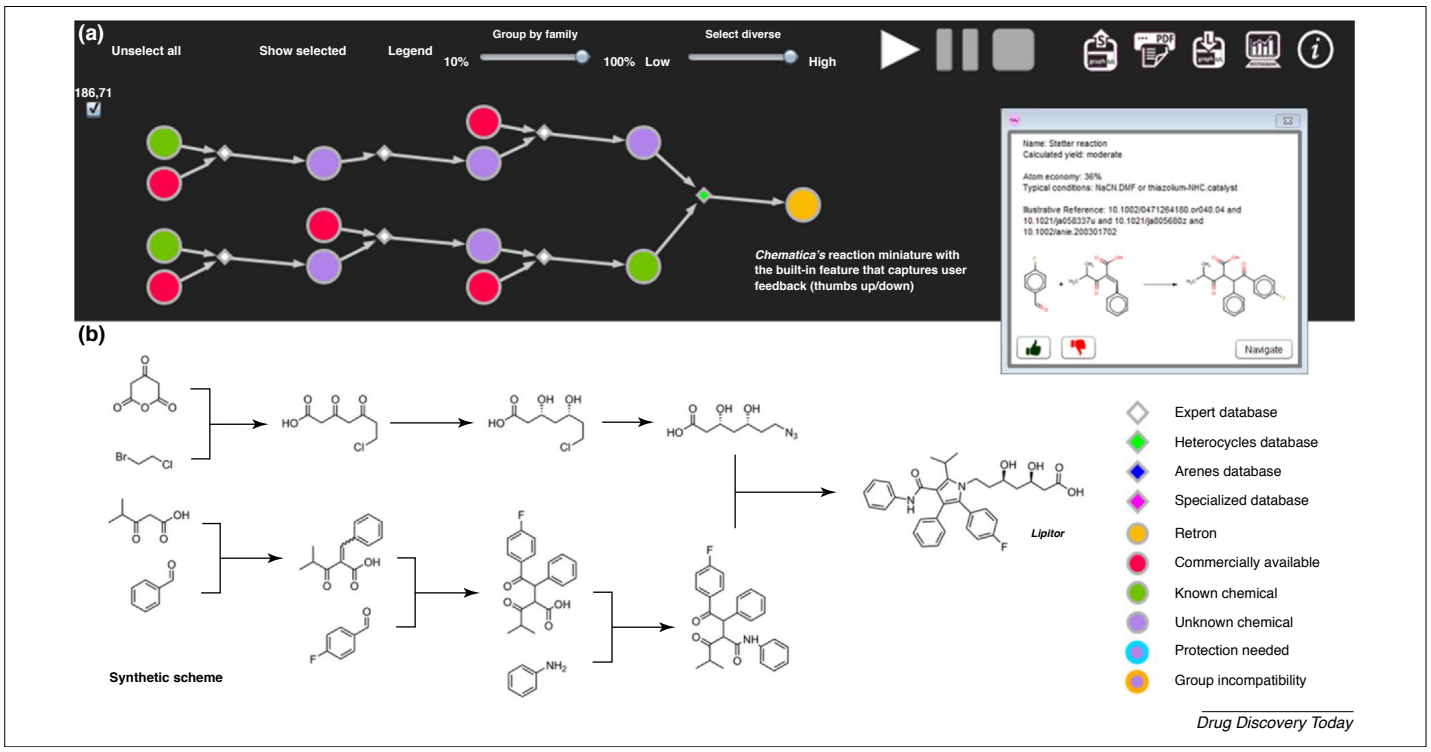

FIGURE 5

Primary output of search results with synthesis tree (a) and the corresponding possible retrosynthetic steps (b) manually compiled based on each individual reaction miniature.

The 'create a synthesis plan' is the retrosynthetic part of the program using the rules described above. Various parameters can be set for the search (Fig. 6) and it is possible to define whether rules should be used only where there are many examples, or whether all transformations should be used, including reactions that are rare and perhaps not as reliable. During the search, functional groups are screened for incompatibility by comparing with a list of compatible groups.

ChemPlanner is able to go back four synthetic steps, hence keeping search time low, and aims for commercially available starting materials. Where a molecule is more complex and this is not possible in four steps, a further run can easily be initiated for these precursors.

Interpreting results

Literature examples for exact or similar transformations are given for every step suggested. Yields are predicted based on the average yield for similar examples.

Each of the various routes suggested is ranked based on multiple criteria, including yield, cost, and number of literature reactions [47]. At first, only the top-scored route is shown, but the user can easily see relevant literature information and other possibilities for the synthesis, and switch to any of the possible synthesis suggestions given (Fig. 7). There is the option to block a particular precursor or transformation after the first prediction round and the program can rerank the routes.

When looking at a particular set of literature examples associated with a predicted synthesis step, the chemist can filter based on various criteria, such as reagent, solvent, or functional group, or sort by similarity to the predicted transformation.

Applications

Previous work has demonstrated the usefulness of *ChemPlanner* for predicting alternative routes for compounds as well as finding those routes that would be chosen by a chemist [48]. It was shown by one case study that the best retrosynthetic route of *ChemPlanner* was similar to that of the chemist, except for an alternative reaction for one step. This alerted the chemist to the availability of a more advanced intermediate, which could shorten the synthesis.

In addition, *ChemPlanner* almost always identified the chemist's favored routes, although the chemist's route was often not the most highly scored route suggested by *ChemPlanner*. Therefore, it is valuable to look at the other routes suggested by *ChemPlanner*, because they may be more suitable for the problem at hand.

Concluding remarks and outlook

Currently, there are approximately 130 million molecules in the CAS REGISTRY database. However, the number of small molecules possible is estimated to be somewhere around 10^{60} [49], which means only a small percentage has been reported in the literature so far. Taking this into account, one can also expect that the possible number of reactions will much higher than that currently captured in reaction databases. Therefore, a computer-assisted synthesis design tool will help chemists to identify alternative approaches to their target molecules, independent from the experience and educational background of individual chemists.

However, the success of commercially available software tools will also be influenced by their user friendliness and fulfillment of expectations. Depending on the area in which chemists work, such

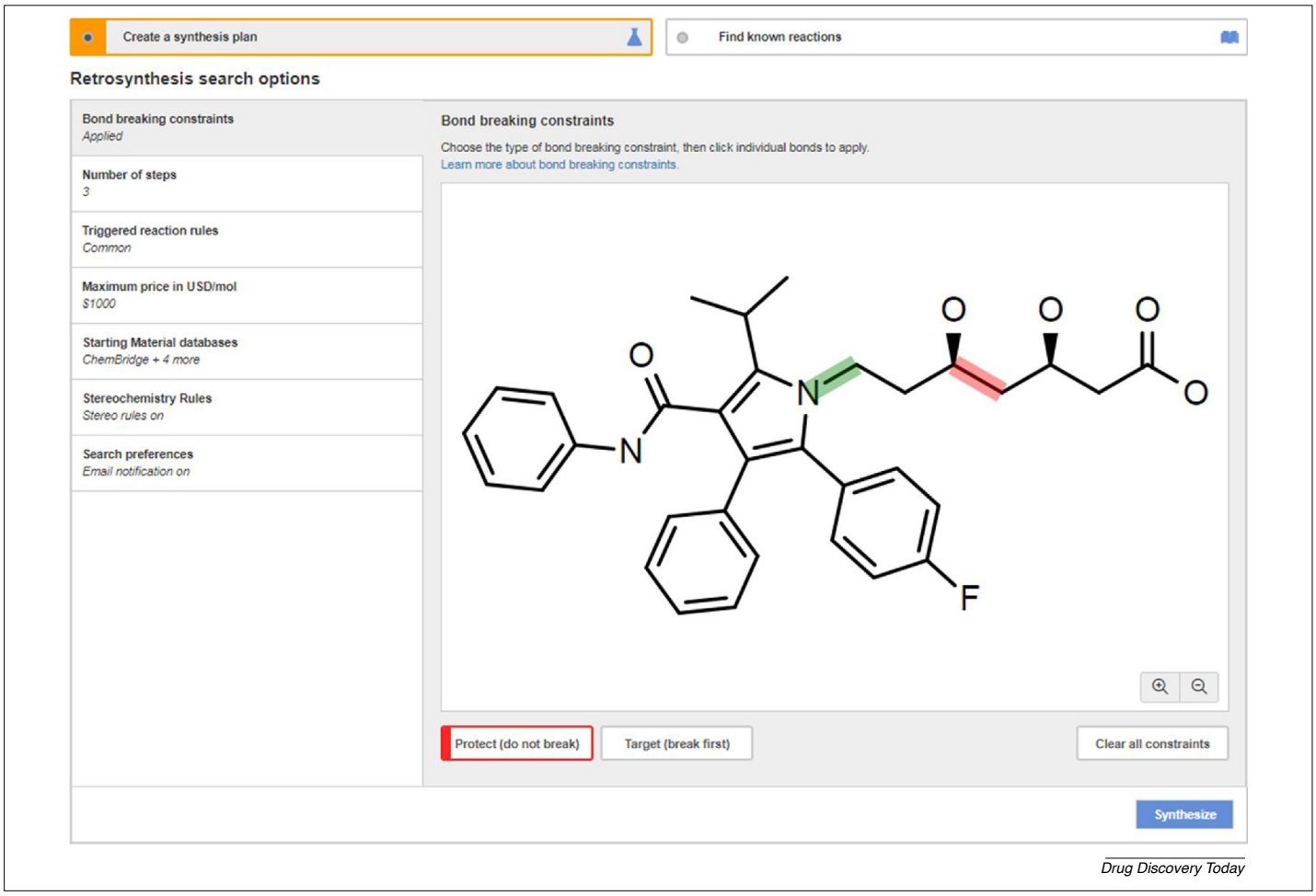

FIGURE 6

Setting up a search in *ChemPlanner*.

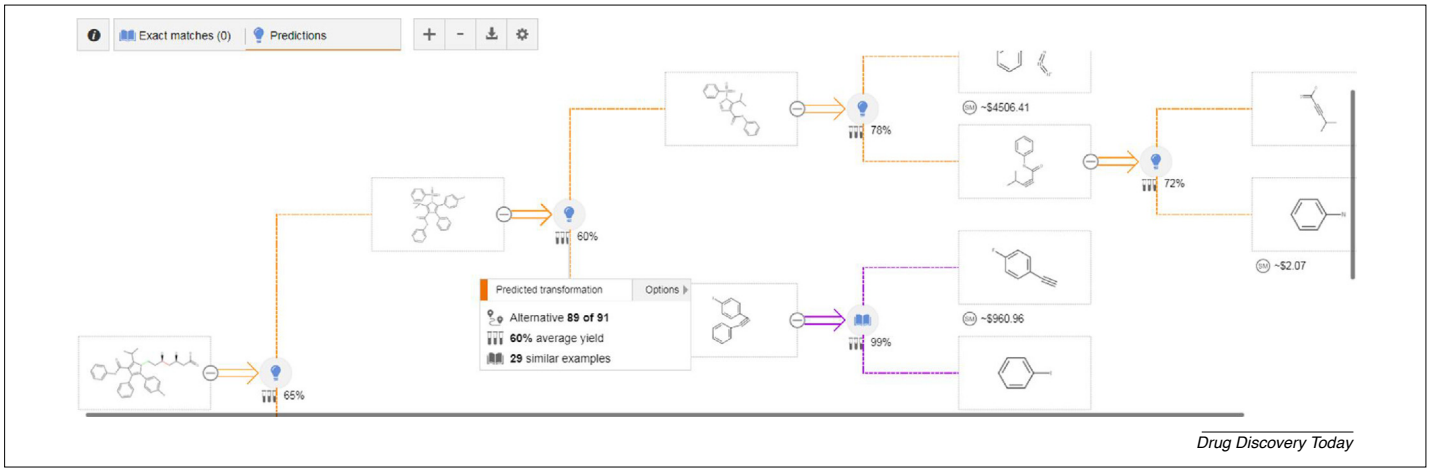

FIGURE 7

Primary output of search results with synthesis tree from *ChemPlanner*.

requirements might be different and a major challenge will be to combine these in a single tool. The three systems presented are already different in how they build up their rules, their search functions, and the illustration of results. Therefore, it is recommended to take the opportunity for trial periods offered by the different vendors, which should involve representatives from different areas who will be using the tool (e.g., research, process, and radiochemists).

An overview of computational methods for predicting chemical reactions have been presented. Current state-of-the-art have been

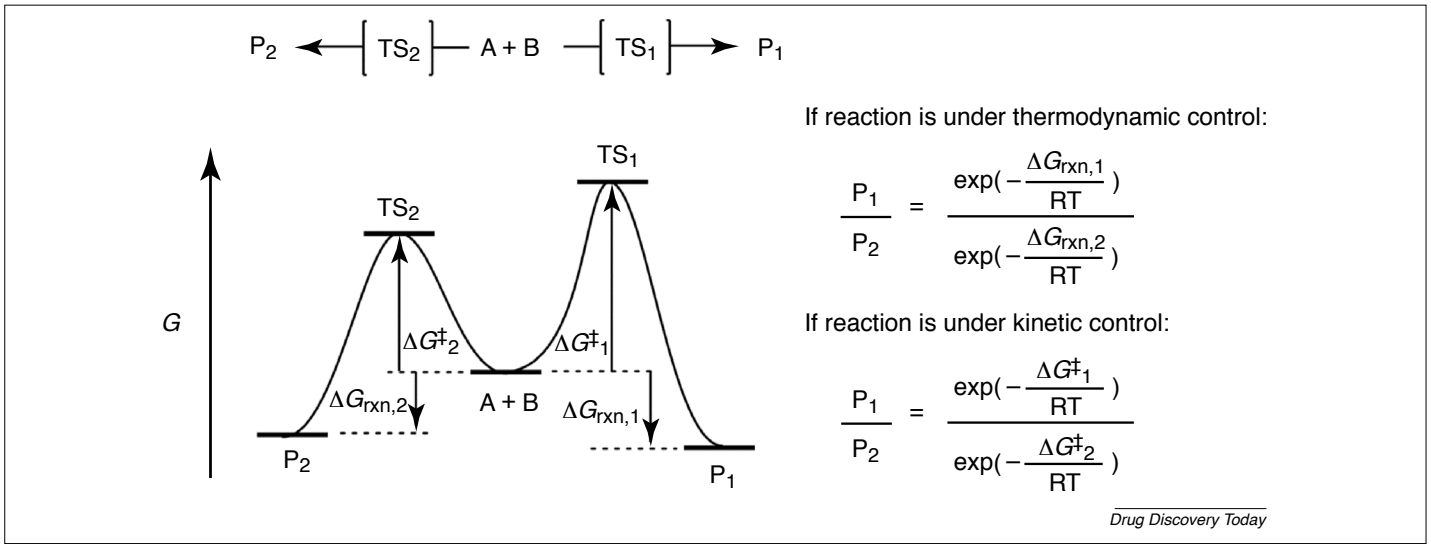

FIGURE 8

Relationship between product selectivities observed in experiments and energies accessible by quantum chemical computations.

described and improvement opportunities for the future highlighted. The authors strongly believe that computational prediction of chemical reactions will, with the increased amount of reaction data and computational resources, be an important area for the foreseeable future.

Quantitative modeling of reaction mechanisms

Introduction

Direct modeling of molecular structures and energetics is complementary to informatics-based methods and has attracted increasing attention in the chemical industry in recent years [50]. The generic reaction profile in Fig. 8, based on transition-state theory, illustrates the connection between reaction selectivity observed in experiment and the underlying reaction energetics provided by computations, which directly provide atomistic details and energetic information about reactants (A and B), products (P_1 and P_2) as well as the intervening transition structures (TS_1 and TS_2). The feasibility of a reaction can be predicted in terms of thermodynamic parameters and activation barriers, such as Gibbs energies of reaction (ΔG_{rxn}) and Gibbs energies of activation ($\Delta G^{\ddagger}$). Comparing the calculated ΔG_{rxn} or $\Delta G^{\ddagger}$ values for competing pathways allows one to predict reaction selectivity. Knowledge of the 3D structures of intermediates and TS then enables rational design of reagents and catalysts. In addition, molecular structures and spectra can be predicted with useful accuracies to facilitate structural assignments in synthesis. Therefore, the success of quantum chemistry in reaction prediction hinges on two crucial aspects: (i) accurate energy evaluation of a molecular system at a given geometry; and (ii) efficient geometry prediction of reactants, products, and TS.

Among the various quantum chemical computational methods, the most popular approaches are based on density functional theory (DFT) methods, which provide a balance of accuracy and speed that fits well with traditional pharmaceutical substrates. Although substantial success has been reported [51–60], significant hurdles remain for its wider adoption among both computational and experimental chemists. Here, we present a few recent applications to reaction mechanism elucidation, selectivity prediction, and the structural elucidation of organic molecules. Beyond QM/DFT reaction modeling, the application of QSAR methodologies in the catalyst design space is gaining momentum [61–72].

Recent advances in quantum chemical reaction modeling

Energy calculations

The B3LYP density functional [73–76] has dominated the study of organic reactions because of its balanced performance in modeling both minima and TS. However, early functionals were unable to treat London dispersion, a problem that is magnified as larger systems of more varying size are studied [77–79]. Modern functionals are dispersion corrected [74,80–86]. Recent studies commonly use ωB97X-D [87], M06-2X [88], or B3LYP-D3 [89] in geometry optimizations, in conjunction with large (at least triple-zeta) basis sets for accurate energies [57]. A connectivity-based correction scheme has been proposed to improve thermochemistry [90,91]. The entropic contribution to the free energy can be hard to evaluate, but mitigating schemes have been proposed [92,93]. For metal complexes, functional selection can be aided by several benchmark studies [94–98].

Geometry prediction

For the prediction of ground states of organic molecules and TS, modern density functionals generally give reliable predictions when used with a modest basis set [99]. Open-source [100,101] and commercial [100–102] software solutions are available for the conversion of SMILES strings or even structural formulas into reasonable geometries by molecular mechanics, which can be conformationally diversified by built-in conformational search routines if desired. The conformers can then be submitted to an appropriate DFT method to obtain refined geometries and energies. The global minimum, or a Boltzmann population, computed this way is typically reliable, so long as the conformational search is exhaustive and the quantum calculation method is appropriate. Indeed, Merck & Co. published a workflow for assigning absolute configurations of complex organic molecules based on rules-based

and force-field conformational searches combined with DFT calculations [103].

TS modeling requires substantially more user intervention than do minima calculations. One starts from a plausible hypothesis about the reaction mechanism and proceeds to search for the TS for each elementary step [57]. To obtain the TS with the correct bonds formed or broken, chemically reasonable input geometries are required. Moreover, it is necessary to visualize the results to verify that the output TS corresponds to the correct bonds being formed or broken. Some commercial software (e.g., Spartan [104]) features precomputed transition state libraries that aid in building good guess geometries of a variety of elementary reaction steps that can be elaborated into the reactants of interest for TS optimizations, whereas other software [105–109] provides algorithms that interpolate user-supplied reactant and product structures to a TS. However, in general, preparing appropriate input for TS calculations requires chemical expertise and, for mechanisms that are poorly understood, trial and error. Some progress has been made towards full automation of TS calculations [e.g., Zimmerman's growing-string method (GSM) [110], the global reaction route mapping (GRRM) strategy, developed by Maeda *et al*. [111], and Wheeler's automated reaction optimizer for new catalysts (AARON) [112]]. Another approach is to construct reaction-specific force fields based on model DFT calculations (Q2MM), and use the force fields for selectivity predictions [54]. The accuracy is generally better than expected from direct applications of DFT (mean unsigned error over several hundred examples is 2–3 kJ/mol without fitting to experiment at any point), thanks to the accurate dispersion in the force fields used, and complete conformational searches for the TS.

Spectroscopy

Analytical chemistry has a key role in mechanism elucidation and compound development in the pharmaceutical industry. QM software packages can calculate spectra such as nuclear magnetic resonance (NMR), infrared (IR), electronic circular dichroism (ECD/CD), ultraviolet (UV), Raman, and vibrational circular dichroism (VCD) [113–127]. One main use of predicted spectra is in structure determination, where NMR and IR can be used to match the experimental spectra of unknown molecules to known computed chemical structures. Computed VCD [103,115,117,118,122,123,126] and TD-DFT ECD spectra [75,128,129] also enable the assignment of absolute configuration. Additionally, QM-calculated proton affinities and solvation energies have been used in conjunction with molecular structural descriptors to formulate QSAR models to predict mass spectrometric response factors of drug-like molecules [130].

Outstanding challenges

Energy calculations

Although DFT will remain the method of choice in modeling organic reactions in the foreseeable future, numerous challenges need to be overcome before it can serve as a generally reliable predictive tool. Even though DFT is successful in predicting the relative rates for similar reaction pathways, such as competing pathways for the formation of different stereoisomers [51–60], current density functionals are incapable of predicting rate constants of synthetic reactions accurately enough to be useful. At room temperature, *RT* is about equal to 0.6 kcal/mol, whereas current density functionals commonly have error bars larger than this. As a result, errors in the computed rate constants can be substantial. A universal density functional with chemical accuracy has been named one of the 'holy grails' in computational chemistry [131]. The development of efficient density functionals that are applicable to molecular systems of increasing size and more complicated electronic structure is a vigorous area of research [132]. Impacting the accuracy of quantitative predictions are errors other than those from the density functional, which include adequate conformational sampling, three-body dispersion effects, anharmonicity, spurious imaginary frequencies, errors in the solvation free energy of ions, and explicit solvent (and ion) effects that are not well represented by continuum models [133–135].

Given that no known density functionals are universally accurate, a bewildering plethora of density functionals has proliferated. Although benchmarking data are critical to inform the choice of density functional, not all areas relevant to reaction modeling have been extensively benchmarked, including activation barriers [136], especially stereocontrolling transition states [137] and organotransition-metal systems [138].

These outstanding problems and gaps in understanding highlight the need for continued innovations that should be driven not only by theoretical chemists as developers of new methods, but also by experimentalists who are the end users of such methods. Industrial support of research in theoretical chemistry to develop robust, effective methodology to model industrially relevant organic molecules and reactions will be beneficial. The large variety of industrial reactions for which accurately measured thermodynamics and kinetics parameters are available can serve as valuable data sources against which computational methods can be benchmarked.

Geometry calculations

Automated transition structure software solutions are promising tools for studying reaction mechanisms, but some bottlenecks that are unique to TS modeling need to be overcome before these tools find wider popularity in industry. Besides the usual concerns of time efficiency and accuracy, issues such as how well the software tolerates an unreactive conformation (e.g., a global minimum conformer) as input and how exhaustive the conformational diversity is captured warrant testing by the wider communities of computational and, especially, synthetic chemists. The user interface of most of the available automated TS location software is command line based rather than graphical, posing a somewhat steep learning curve on the uninitiated. An 'automatic transition structure search' workflow was recently made available within Jaguar [105] that takes the structural formulas of reactants and products, searches for a transition structure, and reports the reaction thermochemistry and activation barrier. This is certainly a welcome direction, although its scope in addressing conformational isomerism and stereoisomeric pathways automatically remains to be seen.

Knowledge management

Besides the technical aspects in the generation of computational data, knowledge management is becoming an increasingly pressing issue. Even though most journals require computed structures (XYZ coordinates) and raw energies to be supplied as supporting information, these data are usually found in the form of PDF files, which has limited direct reusability from the computational chemist's point of view. Information frameworks analogous to

electronic lab notebooks or online web databases that capture these data and render them readily searchable and reusable will be popular to both computational and experimental chemists.

Precompetitive collaborations to advance predictive tools

To be able to increase the productivity of drug discovery, there is a need for better tools. This is relevant for medicinal chemistry from lead generation to process chemistry. Here, we have discussed the three most important approaches to synthesis prediction, with an emphasis on how improvements can be made to drive increased productivity. Starting with reaction mining, the field has changed significantly in recent years with ML methods becoming more popular and competing with, or even replacing, rule-based systems. There is an emerging trend to use larger data sets to increase the predictivity. Considering the increasing amount of data available and progress in ML, especially deep learning, we foresee a rapid future development in this area. We have reviewed three commercially available synthesis prediction tools and discussed a survey of medicinal chemist's wishes for a desktop tool for synthesis prediction. Finally, we have reviewed the state-of the-art applications of QM and MM to dig deeper into reaction mechanisms. Aligned with the general theme of this contribution covering computational prediction of chemical reactions, Maki *et al.* stated that 'while optimization of reaction conditions is inherently empirical, (computational) studies point the way to a more systematic analysis and provide a more predictive approach' [139].

Clearly, computational tools can be useful for computational, medicinal, and process chemists. However, in all three areas, we see the need for further improvements to fully embed computational tools in day-to-day work. Starting with reaction mining, the field is behind the related field of bioactivity prediction, where a public resource, such as ChEMBL provides a free and manually curated resource of bioactivity data to benchmark new algorithms on. Although a data set of reactions derived by text mining from patents in the public domain exists [140], there is not a manually curated large-scale set, although a small set has been published [141]. We believe that the lack of such a set severely impedes new algorithm and ML development. It would be of significant value to have a set of failed reactions in the public domain, because failed reactions are rarely reported in the literature.

There are precompetitive aspects of reaction mining and modeling that are currently underutilized. Certain information about chemical reactions can be shared to build better predictive models without compromising the IP positions of individual consortium members. However, this will only be possible if public standards are agreed and adhered to. Public standards would improve public–private data integration and would need to include reaction classification and consistent ways of describing reactions. Verras and co-workers from multiple pharma companies, a hospital, and a European Union research institute have published a malaria-related QSAR model as a consensus of several QSAR models built by participating scientists based on the proprietary experimental data in their own research organizations [142]. Roche and AstraZeneca formed the joint venture MedChemica to share data for building better predictive tools based on matched molecular pairs without disclosing full molecular structures or their experimental assay data [143]. As a research community, it is worth the effort to investigate ways to share certain aspects of reaction data at levels sufficient to build better predictive models without impacting IP positions. For example, statistics for success rates for named reactions (e.g., Suzuki coupling) and for publicly available reaction building blocks used in those named reactions can be shared with no IP consequences. To facilitate data exchange and sharing, a public data exchange format and a set of terms and standards will need to be developed, agreed upon, and adhered to by all partners engaging in such data sharing (e.g., method/tool used for reaction classification, ways of describing chemical reactions through reaction fingerprints).

Although cheminformaticians might be comfortable with accessing and manipulating reaction data and building predictive models from the command-line user interface, medicinal chemists and process chemists will be dependent on graphic user interfaces (GUI), which should be user friendly. The needs and preferences of different chemists will be so variable that a single interface is probably not achievable. Therefore, it is important to have an environment in which different software solutions can coexist and new features can be tested and evaluated by the user community. It would be beneficial to have certain modularity between the underlying reaction data and the user interface, so that the end-users can license an optimal solution for their needs. It would be desirable to have application programming interface (API) access to the different tools, so that they can more easily be integrated in IT environments within pharmaceutical companies.

It remains beyond the horizon to model all chemical reactions *ab initio*; therefore, modeling needs to be done with methods such as DFT and MM. To make progress, funding for developing new and more accurate methods that can be validated on pharmaceutically relevant molecules is critical. Better knowledge management and capture of published reaction data from QM calculations in such a way that the calculations are reproducible is also desirable. An interesting way forward might be to merge QM methods and ML [144].

References

1 Warr, W.A. (2014) A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Mol. Inf.* 33, 469–476

2 Tomkinson, N. (2014) *Reaction: Baby Steps in Data Exploitation.* AstraZeneca

3 Agnetti, F. *et al.* (2013) *Intuitive and Integrated Browsing of Reactions, Structures, and Citations: The Roche Experience.* Roche

4 Schneider, N. *et al.* (2015) Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* 55, 39–53

5 Schneider, N. *et al.* (2016) Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* 59, 4385–4402

6 Rahman, S.A. *et al.* (2016) Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* 32, 2065–2066

7 Hartenfeller, M. *et al.* (2011) A collection of robust organic synthesis reactions for in silico molecule design. *J. Chem. Inf. Model.* 51, 3093–3098

8 Christ, C.D. *et al.* (2012) Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* 52, 1745–1756

9 Gelernter, H. *et al.* (1990) Building and refining a knowledge base for synthetic organic-chemistry via the methodology of inductive and deductive machine learning. *J. Chem. Inf. Comput. Sci.* 30, 492–504
10 Matosin, N. *et al.* (2014) Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Dis. Model. Mech.* 7, 171–173
11 Cooper, T.W.J. *et al.* (2010) Factors determining the selection of organic reactions by medicinal chemists and the use of these reactions in arrays (small focused libraries). *Angew. Chem. Int. Ed. Engl.* 49, 8082–8091
12 Santanilla, A.B. *et al.* (2015) Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* 347, 49–53
13 Tetko, I.V. *et al.* (2016) Does 'Big Data' exist in medicinal chemistry, and if so, how can it be harnessed? *Future Med. Chem.* 8, 1801–1806
14 Grethe, G. *et al.* (2013) International chemical identifier for reactions (RInChI). *J. Cheminf.* 5, 45
15 Grzybowski, B.A. *et al.* (2009) The 'wired' universe of organic chemistry. *Nat. Chem.* 1, 31–36
16 Kayala, M.A. *et al.* (2011) Learning to predict chemical reactions. *J. Chem. Inf. Model.* 51, 2209–2222
17 Kayala, M.A. and Baldi, P. (2012) ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Model.* 52, 2526–2540
18 Sadowski, P. *et al.* (2016) Synergies between quantum mechanics and machine learning in reaction prediction. *J. Chem. Inf. Model.* 56, 2125–2128
19 Carrera, G.V. *et al.* (2009) Machine learning of chemical reactivity from databases of organic reactions. *J. Comput.-Aided Mol. Des.* 23, 419–429
20 Zhang, Q.-Y. and Aires-De-Sousa, J. (2005) Structure-based classification of chemical reactions without assignment of reaction centers. *J. Chem. Inf. Model.* 45, 1775–1783
21 Silver, D. *et al.* (2016) Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489
22 Coley, C.W. *et al.* (2017) Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443
23 Wei, J.N. *et al.* (2016) Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* 2, 725–732
24 Segler, M.H.S. and Waller, M.P. (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* 23, 5966–5971
25 Marcou, G. *et al.* (2015) Expert system for predicting reaction conditions: the Michael reaction case. *J. Chem. Inf. Model.* 55, 239–250
26 Lin, A.I. *et al.* (2016) Automatized assessment of protective group reactivity: a step toward big reaction data analysis. *J. Chem. Inf. Model.* 56, 2140–2148
27 Segler, M.H.S. and Waller, M.P. (2017) Modelling chemical reasoning to predict and invent reactions. *Chemistry* 23, 6118–6128
28 Segler, M. *et al.* Learning to plan chemical synthesis. https://arxiv.org/pdf/1708.04202.pdf. Accessed 9 March 2018.
29 Liu, B. *et al.* (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* 3, 1103–1113
30 www.pistoiaalliance.org/projects/udm/. [Accessed 26 February 2018]
31 www.cas.org/etrain/scifinder/sciplanner.html. [Accessed 26 February 2018]
32 https://service.elsevier.com/app/answers/detail/a_id/14597/supporthub/reaxys/. [Accessed 26 February 2018]
33 Corey, E.J. and Wipke, W.T. (1969) Computer-assisted design of complex organic syntheses. *Science* 166, 178–192
34 Corey, E. *et al.* (1985) Computer-assisted analysis in organic synthesis. *Science* 228, 408–418
35 Cook, A. *et al.* (2012) Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2, 79–107
36 Hanessian, S. (2005) Man, machine and visual imagery in strategic synthesis planning: computer-perceived precursors for drug candidates. *Curr. Opin. Drug Discov. Dev.* 8, 798–819
37 Todd, M.H. (2005) Computer-aided organic synthesis. *Chem. Soc. Rev.* 34, 247–266
38 www.infochem.de/. [Accessed 26 February 2018]
39 Ravitz, O. (2013) Data-driven computer aided synthesis design. *Drug Discov. Today: Technol.* 10, e443–e449
40 www.spresi.com/. [Accessed 26 February 2018]
41 Bøgevig, A. *et al.* (2015) Route design in the 21st century: the ICSYNTH software tool as an Idea generator for synthesis prediction. *Org. Process Res. Dev.* 19, 357–368
42 www.haxel.com/icic/2014/Programme/monday-13-oct-2014#knowledge-based-de-novo-molecular-design-using-icsynth-frp. [Accessed 26 February 2018]
43 http://chematica.net/. [Accessed 26 February 2018]
44 Szymkuć, S. *et al.* (2016) Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* 55, 5904–5937
45 www.cas.org/products/scifinder-n. [Accessed 26 February 2018]
46 CIRX. http://www.cheminform.com/reaction-library. (Accessed 7 March 2018).
47 http://news.wiley.com/ChemPlanner_Webinar. [Accessed 26 February 2018]
48 www.chemanager-online.com/en/whitepaper/wiley-chemplanner-predicts-experimentally-verified-synthesis-routes-medicinal-chemistry. [Accessed 26 February 2018]
49 Bohacek, R.S. *et al.* (1996) The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* 16, 3–50
50 Deglmann, P. *et al.* (2015) Application of quantum calculations in the chemical industry—an overview. *Int. J. Quantum Chem.* 115, 107–136
51 Ashley, E.R. *et al.* (2017) Ruthenium-catalysed dynamic kinetic resolution asymmetric transfer hydrogenation of β-chromanones by an elimination-induced racemization mechanism. *ACS Catal.* 7, 1446–1451
52 Cheong, P.H.-Y. *et al.* (2011) Quantum mechanical investigations of organo-catalysis: mechanisms, reactivities, and selectivities. *Chem. Rev.* 111, 5042–5137
53 Dirocco, D.A. *et al.* (2017) A multifunctional catalyst that stereoselectively assembles prodrugs. *Science* 356, 426–430
54 Hansen, E. *et al.* (2016) Prediction of stereochemistry using Q2MM. *Acc. Chem. Res.* 49, 996–1005
55 Ji, Y.N. *et al.* (2017) A rational pre-catalyst design for bis-phosphine mono-oxide palladium catalysed reactions. *Chem. Sci.* 8, 2841–2851
56 Mccabe Dunn, J.M. *et al.* (2017) The protecting-group free selective 3′-functionalization of nucleosides. *Chem. Sci.* 8, 2804–2810
57 Lam, Y.-H. *et al.* (2016) Theory and modeling of asymmetric catalytic reactions. *Acc. Chem. Res.* 49, 750–762
58 Sperger, T. *et al.* (2016) Computation and experiment: a powerful combination to understand and predict reactivities. *Acc. Chem. Res.* 49, 1311–1319
59 Tantillo, D.J. (2016) Speeding up sigmatropic shifts—to halve or to hold. *Acc. Chem. Res.* 49, 741–749
60 Wheeler, S.E. *et al.* (2016) Noncovalent interactions in organocatalysis and the prospect of computational catalyst design. *Acc. Chem. Res.* 49, 1061–1069
61 Denmark, S.E. *et al.* (2011) A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Application of quantitative structure activity/selectivity relationships. *J. Org. Chem.* 76, 4337–4357
62 Denmark, S.E. *et al.* (2012) Effects of charge separation, effective concentration, and aggregate formation on the phase transfer catalysed alkylation of phenol. *J. Am. Chem. Soc.* 134, 13415–13429
63 Harper, K.C. *et al.* (2012) Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.* 4, 366–374
64 Jensen, K.H. and Sigman, M.S. (2007) Systematically probing the effect of catalyst acidity in a hydrogen-bond-catalysed enantioselective reaction. *Angew. Chem. Int. Ed.* 46, 4748–4750
65 Jensen, K.H. and Sigman, M.S. (2010) Evaluation of catalyst acidity and substrate electronic effects in a hydrogen bond-catalysed enantioselective reaction. *J. Org. Chem.* 75, 7194–7201
66 Jensen, K.H. *et al.* (2010) Advancing the mechanistic understanding of an enantioselective palladium-catalysed alkene difunctionalization reaction. *J. Am. Chem. Soc.* 132, 17471–17482
67 Milo, A. *et al.* (2014) Interrogating selectivity in catalysis using molecular vibrations. *Nature* 507, 210–214
68 Milo, A. *et al.* (2015) Organic chemistry. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* 347, 737–743
69 Sigman, M.S. *et al.* (2016) The development of multidimensional analysis tools for asymmetric catalysis and beyond. *Acc. Chem. Res.* 49, 1292–1301
70 Sigman, M.S. and Jensen, D.R. (2006) Ligand-modulated palladium-catalysed aerobic alcohol oxidations. *Acc. Chem. Res.* 39, 221–229
71 Sigman, M.S. and Werner, E.W. (2012) Imparting catalyst control upon classical palladium-catalysed alkenyl C-H bond functionalization reactions. *Acc. Chem. Res.* 45, 874–884
72 Denmark, S.E. *et al.* (2011) A systematic investigation of quaternary ammonium ions as asymmetric phase-transfer catalysts. Synthesis of catalyst libraries and evaluation of catalyst activity. *J. Org. Chem.* 76, 4260–4336
73 Becke, A.D. (1993) Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* 98, 5648–5652
74 Lee, C.T. *et al.* (1988) Development of the Colle-Salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B* 37, 785–789
75 Stephens, P.J. *et al.* (1994) Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* 98, 11623–11627
76 Vosko, S.H. *et al.* (1980) Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* 58, 1200–1211
77 Check, C.E. and Gilbert, T.M. (2005) Progressive systematic underestimation of reaction energies by the B3LYP model as the number of C-C bonds increases: why

organic chemists should use multiple DFT models for calculations involving polycarbon hydrocarbons. *J. Org. Chem.* 70, 9828–9834
78 Hansen, A. *et al.* (2014) The thermochemistry of london dispersion-driven transition metal reactions: getting the 'right answer for the right reason'. *ChemistryOpen* 3, 177–189
79 Kruse, H. *et al.* (2012) Why the standard B3LYP/6-31G* model chemistry should not be used in DFT calculations of molecular thermochemistry: understanding and correcting the problem. *J. Org. Chem.* 77, 10824–10834
80 Biedermann, F. and Schneider, H.-J.R. (2016) Experimental binding energies in supramolecular complexes. *Chem. Rev.* 116, 5216–5300
81 Grimme, S. *et al.* (2016) Dispersion-corrected mean-field electronic structure methods. *Chem. Rev.* 116, 5105–5154
82 Li, A. *et al.* (2014) Quantum mechanical calculation of noncovalent interactions: a large-scale evaluation of PMx, DFT, and SAPT approaches. *J. Chem. Theory Comput.* 10, 1563–1575
83 Mardirossian, N. and Head-Gordon, M. (2016) How accurate are the Minnesota density functionals for noncovalent interactions, isomerization energies, thermochemistry, and barrier heights involving molecules composed of main-group elements? *J. Chem. Theory Comput.* 12, 4303–4325
84 Ramakrishnan, R. *et al.* (2014) Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1, 140022
85 Řezáč, J. and Hobza, P. (2016) Benchmark calculations of interaction energies in noncovalent complexes and their applications. *Chem. Rev.* 116, 5038–5071
86 Zheng, J. *et al.* (2007) Representative benchmark suites for barrier heights of diverse reaction types and assessment of electronic structure methods for thermochemical kinetics. *J. Chem. Theory Comput.* 3, 569–582
87 Chai, J.D. and Head-Gordon, M. (2008) Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* 10, 6615–6620
88 Zhao, Y. and Truhlar, D.G. (2008) The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* 120, 215–241
89 Grimme, S. *et al.* (2011) Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* 32, 1456–1465
90 Ramabhadran, R.O. and Raghavachari, K. (2011) Theoretical thermochemistry for organic molecules: development of the generalized connectivity-based hierarchy. *J. Chem. Theory Comput.* 7, 2094–2103
91 Ramabhadran, R.O. and Raghavachari, K. (2014) The successful merger of theoretical thermochemistry with fragment-based methods in quantum chemistry. *Acc. Chem. Res.* 47, 3596–3604
92 Grimme, S. (2012) Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem. Eur. J.* 18, 9955–9964
93 Ribeiro, R.F. *et al.* (2011) Use of solution-phase vibrational frequencies in continuum models for the free energy of solvation. *J. Phys. Chem. B* 115, 14556–14562
94 Harvey, J.N. (2006) On the accuracy of density functional theory in transition metal chemistry. *Annu. Rep. Prog. Chem. Sect. C: Phys. Chem.* 102, 203–226
95 Weymuth, T. *et al.* (2014) New benchmark set of transition-metal coordination reactions for the assessment of density functionals. *J. Chem. Theory Comput.* 10, 3092–3103
96 Hopmann, K.H. (2016) How accurate is DFT for iridium-mediated chemistry? *Organometallics* 35, 3795–3807
97 Sperger, T. *et al.* (2015) Computational studies of synthetically relevant homogeneous organometallic catalysis involving Ni, Pd, Ir, and Rh: an overview of commonly employed DFT methods and mechanistic insights. *Chem. Rev.* 115, 9532–9586
98 Sun, Y. and Chen, H. (2014) Performance of density functionals for activation energies of re-catalysed organic reactions. *J. Chem. Theory Comput.* 10, 579–588
99 Bock, D.A. *et al.* (2010) Crystal structures of proline-derived enamines. *Proc. Natl. Acad. Sci. U. S. A.* 107, 20636–20641
100 O'boyle, N.M. *et al.* (2011) Open Babel: an open chemical toolbox. *J. Cheminf.* 3, 33
101 Vainio, M.J. and Johnson, M.S. (2007) Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* 47, 2462–2474
102 Perkin Elmer (2017) *ChemOffice 16*. Perkin Elmer
103 Sherer, E.C. *et al.* (2014) Systematic approach to conformational sampling for assigning absolute configuration using vibrational circular dichroism. *J. Med. Chem.* 57, 477–494
104 Wavefunction (2016) *Spartan '16*. Wavefunction
105 Bochevarov, A.D. *et al.* (2013) Jaguar: a high-performance quantum chemistry software program with strengths in life and materials sciences. *Int. J. Quantum Chem.* 113, 2110–2142
106 Frisch, M.J. *et al.* (2016) *Gaussian 16*. Gaussian
107 Valiev, M. *et al.* (2010) NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* 181, 1477–1489
108 Shao, Y.H. *et al.* (2015) Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* 113, 184–215
109 Anon (2016) *TURBOMOLE V7.1*. University of Karlsruhe and Forschungszentrum Karlsruhe GmbH
110 Zimmerman, P. (2013) Reliable transition state searches integrated with the growing string method. *J. Chem. Theory Comput.* 9, 3043–3050
111 Maeda, S. *et al.* (2013) Systematic exploration of the mechanism of chemical reactions: the global reaction route mapping (GRRM) strategy using the ADDF and AFIR methods. *Phys. Chem. Chem. Phys.* 15, 3683–3701
112 Guan, Y. *et al.* (2017) *AARON: An Automated Reaction Optimizer for New Catalysts, v. 0.95*. Texas A&M University
113 Bally, T. and Rablen, P.R. (2011) Quantum-chemical simulation of H-1 NMR spectra. 2. Comparison of DFT-based procedures for computing proton-proton coupling constants in organic molecules. *J. Org. Chem.* 76, 4818–4830
114 Buevich, A.V. and Elyashberg, M.E. (2016) Synergistic combination of CASE algorithms and DFT chemical shift predictions: a powerful approach for structure elucidation, verification, and revision. *J. Nat. Prod.* 79, 3105–3116
115 Chavali, B. *et al.* (2007) Mid IR CD spectroscopy for medicinal chemistry: a pharmaceutical perspective. *Am. Pharm. Rev.* 10, 94–98
116 Cheeseman, J.R. and Frisch, M.J. (2011) Basis set dependence of vibrational Raman and Raman optical activity intensities. *J. Chem. Theory Comput.* 7, 3323–3334
117 Freedman, T.B. *et al.* (2003) Absolute configuration determination of chiral molecules in the solution state using vibrational circular dichroism. *Chirality* 15, 743–758
118 He, Y.A. *et al.* (2011) Determination of absolute configuration of chiral molecules using vibrational optical activity: a review. *Appl. Spectrosc.* 65, 699–723
119 Hwang, T.L. *et al.* (2016) Application of 1,1-ADEQUATE, HMBC, and density functional theory to determine regioselectivity in the halogenation of pyridine N-oxides. *Org. Lett.* 18, 1956–1959
120 Kutateladze, A.G. and Reddy, D.S. (2017) High-throughput in silico structure validation and revision of halogenated natural products is enabled by parametric corrections to DFT-computed C-13 NMR chemical shifts and spin-spin coupling constants. *J. Org. Chem.* 82, 3368–3381
121 Mevers, E. *et al.* (2016) Homodimericin A: a complex hexacyclic fungal metabolite. *J. Am. Chem. Soc.* 138, 12324–12327
122 Minick, D.J. *et al.* (2007) Strategies for successfully applying vibrational circular dichroism in a pharmaceutical research environment. *Am. Pharm. Rev.* 10, 118–123
123 Nafie, L.A. *et al.* (1976) Vibrational circular-dichroism. *J. Am. Chem. Soc.* 98, 2715–2723
124 Navarro-Vazquez, A. (2017) State of the art and perspectives in the application of quantum chemical prediction of H-1 and C-13 chemical shifts and scalar couplings for structural elucidation of organic compounds. *Magn. Reson. Chem.* 55, 29–32
125 Smith, S.G. and Goodman, J.M. (2010) Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: the DP4 probability. *J. Am. Chem. Soc.* 132, 12946–12959
126 Stephens, P.J. *et al.* (2012) *VCD Spectroscopy for Organic Chemists*. CRC Press
127 Willoughby, P.H. *et al.* (2014) A guide to small-molecule structure assignment through computation of (H-1 and C-13) NMR chemical shifts. *Nat Protoc.* 9, 643–660
128 Sherer, E.C. *et al.* (2015) Absolute configuration of remisporines A & B. *Org. Biomol. Chem.* 13, 4169–4173
129 Stephens, P.J. *et al.* (2004) Determination of absolute configuration using concerted ab initio DFT calculations of electronic circular dichroism and optical rotation: bicyclo[3.3.1]nonane diones. *J. Org. Chem.* 69, 1948–1958
130 Cramer, C.J. *et al.* (2017) Prediction of mass spectral response factors from predicted chemometric data for druglike molecules. *J. Am. Soc. Mass. Spectrom.* 28, 278–285
131 Houk, K.N. and Liu, F. (2017) Holy grails for computational organic chemistry and biochemistry. *Acc. Chem. Res.* 50, 539–543
132 Peverati, R. and Truhlar, D.G. (2014) Quest for a universal density functional: the accuracy of density functionals across a broad spectrum of databases in chemistry and physics. *Philos. Trans. A Math. Phys. Eng. Sci.* 372, 20120476
133 Jensen, J.H. (2015) Predicting accurate absolute binding energies in aqueous solution: thermodynamic considerations for electronic structure methods. *Phys. Chem. Chem. Phys.* 17, 12441–12451
134 Liu, Z. *et al.* (2017) Mechanism and reactivity in the Morita-Baylis-Hillman reaction: the challenge of accurate computations. *Phys. Chem. Chem. Phys.* 19, 30647–30657

135 Plata, R.E. and Singleton, D.A. (2015) A case study of the mechanism of alcohol-mediated Morita Baylis–Hillman reactions. The importance of experimental observations. *J. Am. Chem. Soc.* 137, 3811–3826
136 Xu, X. *et al.* (2011) How well can modern density functionals predict internuclear distances at transition states? *J. Chem. Theory Comput.* 7, 1667–1676
137 Simón, L. and Goodman, J.M. (2011) How reliable are DFT transition structures? Comparison of GGA, hybrid-meta-GGA and meta-GGA functionals. *Org. Biomol. Chem.* 9, 689–700
138 Steinmetz, M. and Grimme, S. (2013) Benchmark study of the performance of density functional theory for bond activations with (Ni,Pd)-based transition-metal catalysts. *ChemistryOpen* 2, 115–124
139 Maki, B.E. *et al.* (2009) Impact of solvent polarity on N-heterocyclic carbene-catalysed beta-protonations of homoenolate equivalents. *Org. Lett.* 11, 3942–3945
140 Lowe, D.M. (2012) *Extraction of Chemical Structures and Reactions from the Literature.* University of Cambridge
141 Kraut, H. *et al.* (2013) Algorithm for reaction classification. *J. Chem. Inf. Model.* 53, 2884–2895
142 Verras, A. *et al.* (2017) Shared consensus machine learning models for predicting blood stage malaria inhibition. *J. Chem. Inf. Model.* 57, 445–453
143 https://sciencebusiness.technewslit.com/?p514386. [Accessed 26 February 2018]
144 Schütt, K.T. *et al.* (2017) Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* 8, 13890