Retrosynthesis from transforms to predictive sustainable chemistry and nanotechnology: a brief tutorial review

Alicja Mikolajczyk, Uladzislau Zhdan, Sylvain Antoniotti, Adam Smolinski, Karolina Jagiello, Piotr Skurski, Moussab Harb, Tomasz Puzyn, Jaroslaw Polanski

Article

Retrosynthesis from transforms to predictive sustainable chemistry and nanotechnology: a brief tutorial review

Alicja Mikolajczyk*[a,b], Uladzislau Zhdan[c], Sylvain Antoniotti[d], Adam Smolinski[e], Karolina Jagiello[a], Piotr Skurski[a,b], Moussab Harb[b], Tomasz Puzyn[a,b], Jaroslaw Polanski*[c]

Retrosynthesis is a tool initially developed to simplify planning the synthesis of organic molecules using a symbolic strategy involving disconnections to synthons. It can perform better when the initial strategy is supported by computer-assisted methods both in its strategy and tactic parts. With the progress of chemical knowledge management assisted by computer technologies, retrosynthesis got an opportunity to involve database mining, reaction prediction, machine learning (ML), and other data science tools, which allows for covering inorganic compounds and nanoparticles, for which strategy, e.g., the design of reaction conditions, is a critical issue. Retrosynthesis is also essential for green and sustainable chemistry. From one side, synthon representation makes it possible to select the green type processes and reactants among many possible, while recent computer technologies involving ML-based methods give a chance to more precise control of the green and sustainable metrics at the early stage of its design (before synthesis). A variety of such metrics were described in the literature. Many of them are intuitive heuristics, especially for sustainability evaluation. Green methods are among natural retrosynthesis goals since chemists searching for simplifications always preferred safer and cleaner methods than hazardous ones. Chemical intuition is more important than rigorous quantification in traditional approaches. With the growing availability of novel retrosynthetic tools controlled by green and sustainable metrics, we can hope to observe the significant development of predictive green and sustainability. As predictive greenness and sustainability engage broad chemical areas and contemporary software tends to be a black-box-like architecture, we designed this tutorial to provide an easily understandable background for the chemical and materials science audience involved in drug and material design and discovery.

Introduction

Improving chemical procedures to meet the needs of green and sustainable chemistry is a complex problem. Despite many efforts in this area in the last two decades, the results are below expectations. We still often lack information on the toxicity of chemicals to humans and the environment, their degradability, and recycling or reusing potential or life cycle analyses (LCA) which are essential to meet the supply limitations of chemicals or energy sources but has not attained yet a sufficient level of specialization to green and sustainable chemistry concepts.[1]

Chemical synthesis is a severe challenge in this field. The structural abundance of synthesized compounds complicates their toxicity evaluation or prediction. Chemical compounds can be obtained from different reactants by many different procedures in the presence of different solvents. In recent years, various catalytic and biocatalytic methods have been developed as critical innovations, reducing the toxicity and environmental nuisance of the procedures. The concept of green chemistry has been outlined clearly, focusing on toxicity, safety, energy preservation, preference of mild reaction conditions, suitable solvents, catalysts, etc. Green chemistry can concentrate on a single reaction, laboratory, country, or industry type. Sustainability is a much more complex conception because we should extend the analysis beyond a single system, preferentially to the global scale. Therefore, the precise evaluation of sustainable materials creates problems.

An essential dimension of improving chemical synthesis's green or sustainability dimension is to design and evaluate a library of potential synthetic routes targeting a certain chemical compound. Planning syntheses requires associating three domains: macroscopic, sub-microscopic and symbolic.[2,3] Retrosynthesis is a tool starting from a human chemist's by-hand symbolic transformations (transforms, disconnections) of the targeted molecule (target molecule) into smaller fragments, synthons (retrons). Corey defines a transformation as the exact reverse of a synthetic reaction (transforms), a retron – as the structural subunit for that transform must be present in the target molecule. He used a term synthon as the synonym of molecular fragment.[4] Disconnecting or reforming atomic bonds in molecules in the search of transforms has evolved into the sophisticated computer-aided synthesis design (CASD) system. It is not but recently that machine learning approaches significantly improved the CASD method. Notable, retrosynthesis and CASD, originating from the organic domain, tend to be extended to cover inorganic matter[5] and nano materials[6]. With the new CASD software, we could better assist green chemistry needs of a comprehensive evaluation of

the methods engaged in chemical substance preparation. The hazardous reagents, solvents, or auxiliaries could be avoided, while catalytic or biocatalytic processes should be preferred. Can we further upgrade the CASD to a method focusing directly on greenness and sustainability, which are critical issues for contemporary chemistry?

Various indexes were recently designed to measure the chemical efficiency of the process, minimizing waste and optimizing resource and energy use. Numerical values of the indexes enable us to sort reactions according to, for example, safety, environmental friendliness, or recyclability. We should optimize green or sustainable CASDs based on these indexes. Can recent learning machine approaches improve our understanding of green and sustainability issues? Despite the potential in this area, only a few examples appeared in the literature.[7,8] One of the obstacles can be a significant distance between the concerns of synthetic vs. computational chemistry. Retrosynthesis is still an autonomous outskirt of organic or inorganic chemistry rather than its central core. The software available is still not widespread enough among bench chemists, who still must care much more about the availability of the methods, lab potential, and staff experience to perform the reaction efficiently than about sustainability. Green aspects are often defeated by this complexity, especially when the synthetic scale is low. A focus on green and sustainability issues increases with the growing process scale. At the same time, sophisticated software engaging machine learning tends to form black box-like architectures of which we have little understanding. We present this review as a tutorial illustrating the retrosynthetic concept and its application in finding greener and more sustainable chemistry solutions.

Retrosynthesis: a method for diversifying and evaluating synthetic solutions

The core of retrosynthesis is the recognition of smaller fragments within the target molecule for which assembly is possible by known or predictable-plausible reactions. Originally, Corey designated these fragments as *synthons*.[9] Figures 1-4 and Table 1 explain the idea and basic lexicon of organic retrosynthesis. An informative introduction into the extended synthon concept and nomenclature is available in the reference.[10]

Table 1 Basic lexicon of retrosynthesis

	Definition
Target molecule (TM)	A molecule under design.
Synthon (S)	Any group of atoms indicated within the target molecule that converts the reagent (R) representing this synthon *in vitro* to a target molecule in the known or expected reaction(s). A heterolytic disconnection of the bonding within the target molecule forms either an electrophilic or nucleophilic center, an acceptor (a) or a donor (d) synthon in the lexicon of retrosynthesis. Homolytic disconnections controlling radical chemistry are not shown in Figure 3. A formal nomenclature of synthons involves functional group (heteroatom) identification which defines the location of the a or d center vs. this heteroatom. The a^{alkyl} or d^{alkyl} designate the synthons derived from the fragments that do not have functional groups. Synthons are virtual units needing conversion to reagents for in vitro operations (S $\rightarrow$R).
Functional group (FG)	An essential element controlling chemical reactivity. "*Proximity of reactive functional groups is of major importance in synthesis*"; Corey identifies potential FG operations as FG introduction, FG modification: removal, interconversion.[4,9,10] The reactivity type at certain atoms, or the synthon type (d or a) in the retrosynthesis lexicon, essentially depends on FG neighborhood. A direct relation of the d and a synthons to FGs allows for mapping the reactivity of FGs. A compact and consistent introduction to FG chemistry can be ref.[10]
Transform(ation)	The exact reverse of a synthetic reaction (*transforms*) to identify a *retron* – a structural subunit needed for a certain transform to be present in the target molecule. Often retrons and synthons are treated as synonyms.
Disconnection	Disconnection is a virtual breaking of one or more bonds in TM to form synthons. A wavy line indicates disconnection. The arrows showing transparently the (d) or (a) synthon polarity should guide heterolytic disconnections in a way analogous to the arrows marking electrophilic or nucleophilic reagents' behavior in chemical reactions.
Synthon chemistry	Connecting synthons d and a forms a chemical bond.
Reagent	A real chemical substance that represents corresponding synthon reactivity in vitro. A single synthon can be represented by a number of reagents showing its reactivity type. Sometimes, oxidation state of a given synthon can be modulated to help identifying the relevant reagent and associated chemistry.
	Operations on synthons
Functional group interconversion (FGI)	Individual FG stamps neighboring carbon atoms in synthon with the (d) or (a) polarity type. It also fine-tunes reactivity in the reagents corresponding to these synthons. FGIs are conversions of FG to FG1 planned in such a way that we know simple reactions transforming FG1 to FG for the reagents representing these synthons in vitro.
Functional group addition (FGA)	FG can be mounted on the unfunctionalized carbon atom of a synthon. If the TM should not contain the added FG, then, a chemistry should be known to remove this FG in the reaction(s) of reagents representing synthons in vitro.

Synthon to reagent conversion (S →R)	A virtual operation in which a synthon is transformed to a molecule. This usually needs adding to the synthon some fragment(s) with the opposed polarity which can be a simple stable ending fragment (Figure 3).
Synthon activation or blocking	S →R conversion could result in the reagents of the different reactivity levels involving these completely unreactive (blocked). Activation and blocking could take place selectively at different positions. Not favored in green or sustainable chemistry since it adds synthetic operations.
Umpolung	The (d) synthon is converted to the (a) synthon and vice versa (a) to (d). This operation results in the alteration of the reagent types and chemistry needed for in vitro reactions. In the reagent domain, umpolung converts electrophiles to nucleophiles and vice versa nucleophiles to electrophiles.

Synthon identification should provide a scheme of structural transformations designing synthesis *strategy* and not strict laboratory manipulations. We arranged retrosynthesis scheme in a vertical format to jointly map the disconnections, synthons, reactivity, and reagents (Figure 1).

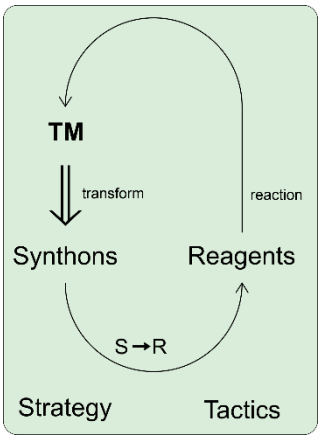

Figure 1. The idea of retrosynthesis. Target molecule (TM) transformed by bond disconnecting or reforming into fragments (synthons) in virtual operations (on paper or in silico). Converting (S -> R) into actual reagents which can be reacted in vitro to TM. Retrosynthesis initially focused on TM's symbolic disconnection strategies to synthons (left side). With the development of machine learning, the prediction of synthesis tactics, i.e., chemical context and operating conditions for the synthetic path from reagent to TM could be more predictive. This part is essential for extending retrosynthesis beyond the organic domain into inorganic compounds and nanoparticles and designing green and sustainable synthetic paths.

This scheme divides the synthon strategy (left) from reagent tactics (right) after synthon to reagent conversion. In other words, on the left side, we map transforms (retro-reactions), while on the right side, we note the current knowledge, which is represented, first of all, by the facts cataloged in databases and literature but also designed by analogy to them. The left side projects disconnections to predict optimal or promising reactions shown on the right side. Having a large reaction library (right factual side), we can also try to predict the nonfactual reaction outcomes for nonfactual reaction conditions, which is a fundamental goal of modern CASD systems.

Figure 2 outlines synthon nomenclature in which heteroatom X controls reactivity type and locants to indicate carbon atom positions. Alcohols are a common compounds type used in synthetic procedures. Figure 3 shows an example how we can illustrate a disconnection of a target molecule (TM) having an alcohol functional group (FG) to synthons S1 (d^{alkyl}) and S2 (a1).

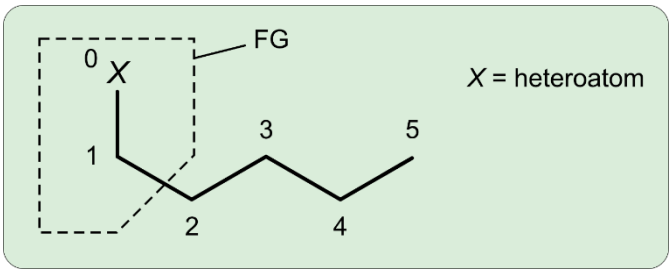

Figure 2. Functional group (FG) with heteroatom (X) labels the carbon chain atoms with locants.

Synthon to reagent conversion S1→R1 and S2→R2 gives; 1-hexyl magnesium bromide and pentanal, which will react in Grignard reaction. Heterolytic bond disconnection, an essential operation, defines the acceptor or donor synthons corresponding to the electrophilic or nucleophilic reactivity types. Radical disconnections were also broadly explored.[11] In vertical disconnections' flow we could realize how much information we need to support molecular symbols (left) to make them real in vitro chemistry (right). Retrosynthetic representation allows us to understand complex chemical reactivity problems. The retrosynthetic tree is a graph showing several (or all) possible routes to TM. Figure 4 compares a fuzzy chemical view with the strictly algorithmic representation of such a tree.

Retrosynthesis beyond the organic domain

Originally, retrosynthesis and CASD systems developed for organic synthesis extended into inorganic materials and nanostructures.[4,5] A short history decides that these domains are much less explored; however, expectations here are high. The exceptional particularity of organic molecules is that carbons and hydrogens are their two main components. Carbon catenation, i.e., the ability to form long chains, is another unique feature. Due to catenation, organic retrosynthetic trees have many deep branches. Usually, we need several reactions to create a long chain or a complex ring. These reactions can penetrate different areas of chemical space. However, on molecular level, substances of the defined structure synthesized as designed by different reactions are the same.

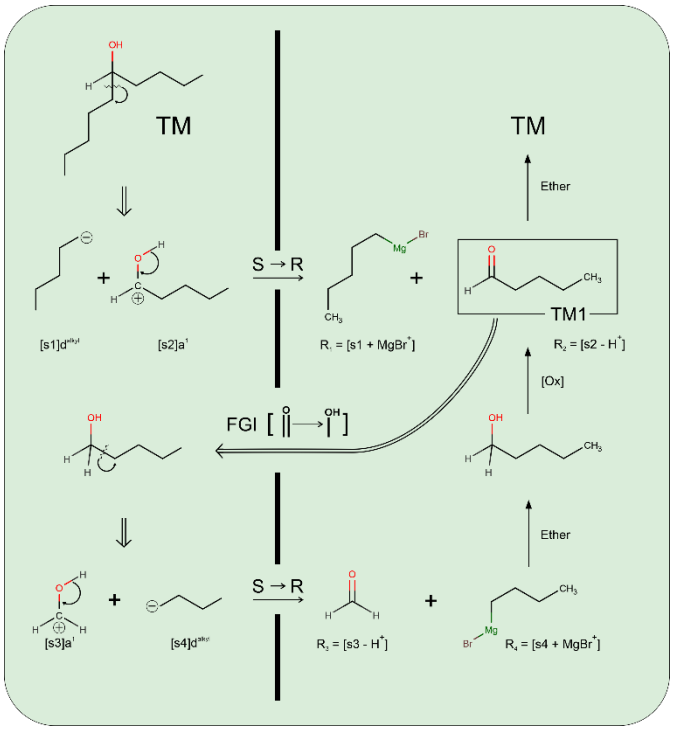

Figure 3. A vertical scheme illustrating disconnections and synthons (left), the synthon to reagent (S→R) conversion and reacting reagents (right) for an exemplary molecule. TM can be disconnected to S1 (d^{alkyl}) and S2 (a^1). S1→R1 and and S2→R2 conversion gives 1-bromohexane and pentanal. The S→R conversion reveals the reagents needed for in vitro reaction are pentyl bromide and pentanal TM1. TM1, after FGI conversion to alcohol, disconnects (left side) to synthons S3 [a1] and S4 [d^{alkyl}]. The S→R provides the respective reagents R3 and R4.

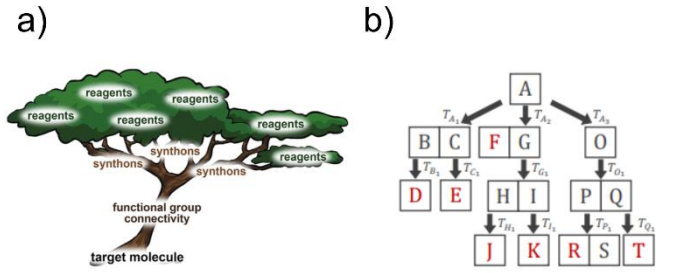

Figure 4. Synthetic tree addressing chemical (a) or informatics (b) representations.[56] Copyright © 2022 American Chemical Society and Division of Chemical Education, Inc. (a) and 2022 XXXX

Thus, that can be concluded that the resulting organic structure does not critically depend on the reaction path or reaction conditions used. Therefore, classical organic retrosynthesis, first of all, does not focus on reaction conditions but a chain or ring disconnections strategy. After finding the disconnection strategy, we decide tactics by choosing individual reactions, reactants and fine-tuning reaction conditions. In contrast, the structures in the inorganic or nanodomains can be more complex, having no restrictions on the types of atoms engaged. For example, in nanomaterials, the critical structural features deciding material functionality are agreed in the nanoscale, i.e., at a much larger size than the single-molecule level. Reaction conditions usually substantially influence such structures. Hence, to obtain the desired structural or functional properties, we focus much more on predicting reaction conditions. Practically, retrosynthesis of inorganic- or nano materials are computational approaches where we predict the influence of different reaction parameters for the structures yielded. Reproducibility and data quality[12] in these areas is much more problematic than in the organic domain, e.g., a small amount of dopants can critically influence a structure or function. The other factor hindering exploratory inorganic synthesis are anthropogenic biases in chemical reaction data.[13]

Since beyond organic domain we predict mainly in tactics, high throughput experimentation (HTE)[12] is one approach to the problem. An example can be the HTS screening of bimetallic catalysts targeting rational control of the geometry and composition of an active site for efficient chemical transformations.[14] The abundance of potential structures decides that novel machine learning methods are indispensable in inorganic or nano retrosynthesis. However, previously extensively explored methods, for example, QSAR in its predictive and robust[15,16] versions can contribute, e.g., by suggesting reasonable meaningful descriptors or indirectly predicting reaction conditions yielding targeted chemical compounds or chemical systems.

The representative example of nano-CASD is an autonomous retrosynthesis of gold nanoparticles via shape matching. The nanostructures were designed using the Bayesian optimization targeted at a specific nano-assembly structure, shape, and size. The reagents needed to be selected a priori while computations were formulated as the shape-matching minimizing the shape discrepancy. The highly computational analysis demonstrated that we are still at the beginning of this direction [6]. Could we expect disconnections in nanostructures similar to classical retrosynthesis needing to develop nanostructure representations?

Ab initio electronic structure methods in retrosynthesis

The main goal of the transform is a structural simplification of a target molecule to synthon representation. In turn, conversing synthons to synthon equivalents (reactans), we replace them with real in vitro reagents. How close do these equivalents resemble synthons, and how reliable could synthon-to-reagent transformation be? We recently pointed out that some synthons may be electronically, kinetically, and thermodynamically stable systems. In other words, we can nearly use them directly in the syntheses.[17] The representative examples of such stable synthons are, for example, borata-alkene anions. The borata-alkene anions (H2C=BR2)– are carbanionic synthons formally representing stabilized α-monoboryl carbanions.[18,19] We described the structures of the isolated negatively charged (H2C=BR2)– (R=H, CH3, C6H5, C6F5, Mes) systems (Mes = 2,4,6-trimethylphenyl) and characterized their electronic and thermodynamic stability[17] basing on theoretical ab initio models (the QCISD/aug-cc-pVTZ and MP2/aug-cc-pVDZ). The structurally smallest (H2C=BH2)– synthon adopts a planar C2v-symmetry equilibrium structure with the double-bond connecting the carbon and boron

atoms. Positively charged synthons are even more promising candidates for such systems because, in their case, the electronic stability would not be an issue (whereas the stability of the anions is always potentially jeopardized by the possibility of the excess electron auto detachment). Of course, even such stable synthons are ionic species; therefore, for the S → R conversion, we need a counterion to neutralize their negative or positive charge. The anionic borata-alkene synthon supported with sodium cation is an illustrative example.

In the more general context, currently, quantum chemistry probes reaction mechanisms efficiently; therefore, its application to retrosynthesis seems natural. Once standard retrosynthesis provides us with the retrosynthetic tree, we can determine by quantum methods the kinetic barriers involved in each reaction pathway and, thus, the corresponding reaction rates to choose the most efficient (i.e., Gibbs free energy-wise) synthetic route. Quantum chemistry methods, e.g., post-Hartree-Fock methods such as configuration interaction (CI)[20–22], Møller-Plesset perturbation theory (MP2, MP4)[23–25], coupled cluster (CC)[26–28], composite methods (G2, G4)[29,30] or the methods based on the density functional theory (DFT, e.g., B3LYP, CAM-B3LYP or wB97XD functionals)[31–34], together with the basis sets of double- or triple-zeta quality (e.g., 6-31++G(d,p), aug-cc-pVDZ, 6-311++G(d,p), aug-cc-pVTZ)[35–38] not only enables the evaluation of the overall reaction rate but also provides an insight into each elementary step of a chemical process at the molecular level. In order to gain such insight, the geometric structures of all stationary points (corresponding to the isolated substrates, initial reactant complex, intermediate products, transition states, and final products) involved in each reaction step must be obtained. In addition, the structure of each transition state (whose determination is often the most difficult part of the whole investigation) must be verified (usually by following the intrinsic reaction coordinate) to assure its relevance (i.e., to confirm that it connects the substrates and products of a given elementary step). Once such a comprehensive theoretical study is completed, the exact mechanism of a studied chemical reaction is revealed, applying to studying any chemical reaction, including those planned via retrosynthetic analysis.

Data and their processing by machine learning in retrosynthesis

Chemical descriptors or properties represent chemical compounds.[39,40] While descriptors can be calculated from molecular representations, properties need experiments to be measured. Alternatively, having a large library of the properties measured, we can attempt property predictions for novel structures. Among thousands of molecular descriptors available, SMILES are probably the most popular in recent computations for coding molecular structures.[41] First, the SMILES system is relatively easy to understand and interpret by computers and humans. Second, the SMILES system, in particular canonical SMILES, can be an unambiguous molecular representation, i.e., each SMILES represents a unique compound, and a unique SMILES code can represent each compound.

Coding molecular connectivity SMILES can efficiently generate novel molecules beyond factual space. However, novel descriptors still appear, e.g., for materials description.[41,42] For example, we can use fragment-related representations[43] or simplex codes[44] to code chemical data in the inorganic domain, while DeepSMILES[45] or SELFIES[46] are novel descriptors developed especially for novel structure generation avoiding vacant code to structure mapping. An example of the recent application of atomic descriptors coding atomic environments to retrosynthesis can be found in the reference.[47] A variety of other barriers in computer-assisted material design, e.g., data sharing in catalyst design, is discussed in the reviews.[48,49]

The increasing number of engaged molecular structures is typical for the current molecular design. Machine learning and deep learning involving neural networks are new methods needed to support computational approaches in these areas. Although drug design pioneered a number of *in silico* algorithms, recently, it is CASD that has significantly improved its efficiency. The lexicon of the current data science, especially for neural network approaches, is buzzily interfering in many areas. Usually, we identify three main learning architectures: unsupervised, supervised, or reinforcement learning. Figure 5 briefly illustrates these methods. In supervised learning, we support the molecular input data with the labels that the network should learn during the optimization. The unsupervised architecture is optimized by evaluating the similarity of the input signals; therefore, we need not show the labels. This is especially important when we do not know the label value, e.g., for a series of chemical compounds, we know their structures but do not know their reactivity or biological activity data. While supervised and unsupervised learning treats the data statically, reinforcement learning is a dynamic method investigating the interaction with the environment. The idea originated from the game theory, which searches for the best move. Therefore, in the language of this method, if the data (agent) moved randomly, interacting with the environment, a critique estimates an error. A low error value rewards a novel state, sitting better position as a result.[50]

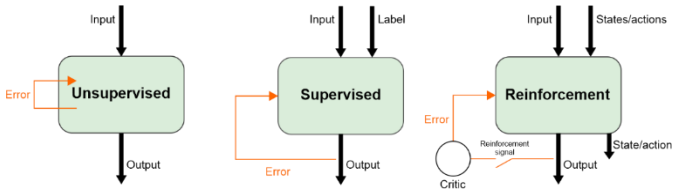

Figure 5. The unsupervised, supervised and reinforcement learning architectures. Details in text.

Each network is optimized on the known chemical data library to be used for predictions. Since we want to explore unknown areas of chemical space, generating novel chemical structures is a fundamental problem in molecular design. We need to create novel molecules and let the network simulate the output for them. The solution to this problem can be the so-

called generative approaches, e.g., generative adversarial network (GAN), in which the network maps the molecular representation of the known outputs by the discriminating block in the network. The random generator provides novel structures of the unknown output to be co-mapped to the discriminator, which gets experience in predicting. In Figure 6 we illustrated the idea of the GAN network, according to the reference.[41]

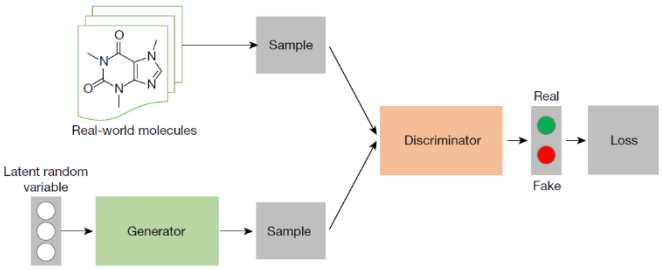

Figure 6. A complex learning architecture crossing the random generator of new structures in which chemical connectivity is coded by SMILES or related descriptor types. Reproduced after. Copyright © 2022 Springer Nature

Feature engineering and feature learning are two options for modeling and predicting (chemical) data.[51] The term feature follows the lexicon of informatics more than that of chemistry. Its meaning is between a (physical or chemical) property or descriptor and a variable. The feature can be, for example, a component-type representation resulting from principal component analysis, which does not have any particular chemical interpretation. When contrasting engineering vs. learning, we focus on the autonomic capabilities of a computer algorithm. In feature engineering, we need human intervention to design variables that algorithms analyze. In turn, computers should be fully autonomous in the feature learning mode, which means that the algorithm selects the features from among the raw data. In the chemical context, feature engineering asks humans how to construct a molecular representation. Which data should represent chemical compounds in a model? We can say that feature learning is an algorithm capable of autonomous feature engineering by a computer, enabling the molecular representation suitable for the individual computation to be determined.

Deep learning is the next term often used in chemical data science. Deep learning is a fully feature learning architecture using neural networks, usually the supervised back-propagation method. We get more and more experience in back-propagation routines; however, unsupervised architectures could be even more efficient.[52]

Beyond synthesis design, drug discovery is another area of molecular design widely explored by machine learning.[39,40] The extensive up-to-date review of the machine learning for molecular and material sciences the reader can find in the reference.[41] Early neural network applications can be illustrative examples for a better understanding of these methods[53,54], in particular comparing supervised vs. unsupervised architectures in deep chemistry is discussed in the reference.[52]

Computer-aided synthesis design

Retrosynthesis is a tool for splitting a target molecule into a synthon representation organized into a tree-like form. The synthon representation provides easy access to various reagents with the same reactivity type but having a spectrum of other chemical or physical properties. Accordingly, we can select the reagents with the lowest environmental impacts. Second, even more important, the number of possible synthetic routes (a strategy) under design by retrosynthesis is snowballing. The increase of potential transforms can be illustrated by the number of transformations per step, that can be as high as 80 to several thousands[55], extensively exploring chemical space and allowing the chance to avoid the adverse environmental impacts of specific paths and find green and sustainable options. The retrosynthetic tree (Figure 4) supports decision-making. A simple interpretation can be found in the reference.[56] One significant challenge is the availability of the algorithms allowing for extensive chemical space exploration needed for in silico CASD.[55,57,58]

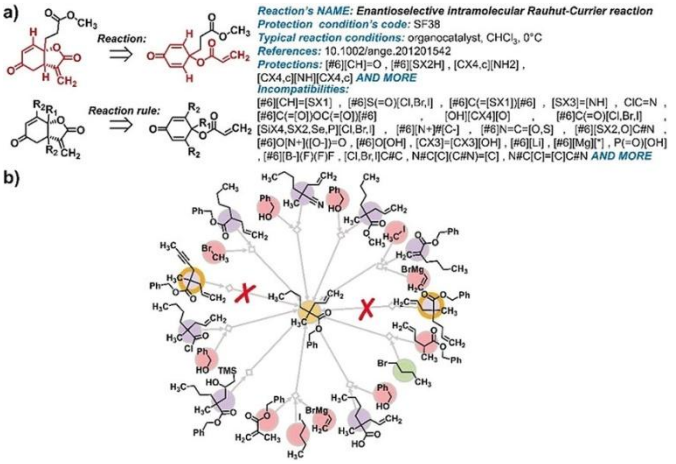

Figure 7. A reaction template (reaction rules) noted for an exemplary reaction. The template is programmed manually on factual data from databases. Data are carefully curated, and expert human knowledge is critical for the high efficiency of the feature-engineered CASD systems. Colors codes different molecules, yellow - target molecule, red – buyable, green – recorded in literature, violet – not recorded in the literature.[59] Copyright © 2022 Wiley Online Library

Corey developed the first software (LHASA, Harvard, Cambridge, MA, USA) to get computer assistance in CASD. Until recently, computers defeated the competition in this field, especially in finding the critical disconnections within complex natural products. However, machine learning is more and more competent here. Recent human-machine cooperation appeared especially successful in this competition.[59–61] In the recent study by the Grzybowski group, the Turing test cannot tell the difference between the human and machine-performed disconnections of the complex organic molecules.[62] Grzybowski also developed a highly efficient software in this field, Chematica (currently Synthia), which provides a full mode of synthesis design to arbitrary targets.[63] Notable, this is a feature-engineered architecture based on high-quality reaction data manually interpreted and programmed by chemists. The Grzybowski group started the

many-year efforts for efficient retrosynthesis by arranging organic chemistry into the network architecture. The nods represent reagents, while their connections correspond to organic reactions. They analyzed all available structures from the Beilstein (currently Reaxys) database as the library of chemical compounds as the knowledge database. This approach allowed them to define the rules of organic chemistry and its developments throughout history. More recently, they identified the similarity between natural language and chemical structures, formulating the reaction rules or so-called templates, and programming their use in the CASD software.[59] Figure 7 shows an example of such a template. Interestingly, fully autonomous deep-learned CASD systems appeared much less successful, despite many efforts.[55,64–66] However, a fully data-driven neural architecture, biosynthesis navigator (BioNavi-NP), was designed recently for computer-aided bio-retrosynthesis.[67]

The reader can find an extensive up-to-date review of CASD for chemical synthesis in reference[12] broadly reviewing the problems of data curation, descriptor, and algorithm selection, retrosynthesis extension to automated synthesis, and high-throughput automated synthesis systems or autonomous systems for chemical synthesis, risk and safety assessment, as well as reaction selectivity, yield, conditions and miniaturization in the CASD context. Modular automatic robotic system application in chemical synthesis is another trend that can significantly contribute to the safety and performance of modern chemistry, also improving reproducibly and data quality.[12]

A formal computer-oriented definition of the retrosynthetic components, e.g., reaction representations and operations, is available in the thorough review [Artificial Intelligence for Retrosynthesis Prediction, https://doi.org/10.1016/j.eng.2022.04.021]. In the same publication, the authors briefly described how retrosynthetic elements and operations are mounted into the individual machine learning (ML) methods available, indicating in particular, (i) the sequence-to-sequence models, (ii) graph neural networks, (iii) search algorithms and (iv) deep reinforcement learning.

Reaction prediction is a related problem of substantial importance for the CASD systems, especially in its tactics part. The reference[12] correlates retrosynthesis to the reaction prediction area, and an extensive review of machine learning for chemical reactions is available in the reference.[68] Popular interpretations and common understanding of the current state of the CASD method can be found in commentaries.[69,70]

CASD Software is still not broadly available, but several options exist. Synthia, developed from Chematica, is commercial retrosynthetic software from Sigma Aldrich, currently a branch of MERCK. The retrosynthetic option is available as a part of the Reaxys database.[71] Another package, the ASKCOS - software tools for organic synthesis, is a freeware option for retrosynthesis available from MIT, which enables one-step retrosynthesis, enabling drawing and searching within buyable compounds catalog. This package can be used directly at the internet site in the interactive mode.[8]

The first CASD for the inorganic domain, involving knowledge and prediction base, was a system integrating databases for the properties of inorganic substances and materials with data analysis having learning ability. The idea was published already in 2011.[72,73] The CASD for inorganic solids synthesis has been extensively explored recently.[74–80] More recently, rational solid-state synthesis routes for inorganic materials were designed using catalytic nucleation on crystalline reactants analysis with the reaction and interfacial energies to the nucleation barriers approximated from high-throughput thermochemical data and structural and interfacial features of crystals.[5] Bimetallic catalyst synthesis is another example using feature-engineered fingerprints with DFT calculated sp-states and localized d-states of adsorption sit, guiding machine learning HTS screening.[14] Recent examples in catalyst CASD can be found in the references.[76,81–84] The reader can compare the reference[85] for the extensive review of machine learning in predictive catalyst CASDs. The reference[86] discusses perspectives for machine learning in heterogeneous catalysis. A broad review of computational and machine learning methods assisting nanoparticle synthesis is reviewed in reference.[87] The extension of retrosynthesis to the nano-domain has been developed recently to design the particles of desired shapes or sizes.[6] QSAR is a related method that, in its predictive versions, could construct the CASD systems. QSAR can involve machine learning for modeling various materials' properties.[88] Table 2 gives illustrative examples of the descriptors for predictive CASDs.

Table 2 Descriptors for predictive CASDs.

Entry	Descriptors[a]	Domain	Material	Method	References
1	Binding energy	inorganic	Bimetallic catalysts	ML	14, 89, 90
2	Filling of a d-band				
3	Center of a d-band				
4	Width of a d-band				
5	Skewness of a d-band				
6	Kurtosis of a d-band				
7	Work function				
8	Local Pauling Electronegativity				
9	Ionization Potential				
10	Electron Affinity				
11	Pauling electronegativity				
12	Atomic Radius				

13	Orbital Radius				
14	mean_d, σ_d				
15	mean_iV, σ_iV				
16	mean_T, σ_T	inorganic	Zeolites	Random Forest	[91]
17	mean_ d2, σ_ d2				[92]
18	mean_iV2, σ_iV2				
19	mean_T2, σ_T2				
20	Ionization Energy				
21	Group				
22	Electronegativity				
23	Enthalpy fusion				
24	Surface Energy				
25	Melting point	inorganic	Calalysts	ML	[93]
26	Atomic Radius				
27	Density				
28	Period				
29	Atomic Number				
30	Atomic Weight				
31	Vacancy formation Energy				
32	Oxide formation Enthalpy				
33	Oxidation Energy				
34	Metal Atom	inorganic	heterogeneous catalysts	ML	[94]
35	Support				
36	HOMO				
37	LUMO				
38	Number of Valence Electrons				
39	Reaction Temperature				
40	Contact Time	inorganic	heterogeneous catalysts	ML	[95]
41	Pressure				
42	Slater-type orbitals				
43	Nuclear charge	inorganic	bimetallic catalysts	CASD	[96]
44	Shielding constant			(OPLS model)	
45	Effective quantum number				
46	Madelung potentials				
47	M-O bond length				
48	e_g electrons	inorganic	catalytic metal oxides	QSAR	[97]
49	d electrons				
50	Charge transfer energy				
51	Molecular size				
52	Bond lengths				
53	Interaction energies				
54	Proton Affinity				
55	Tolman cone angle				
56	Solid angle				
57	Buried volume	organic	organocatalysts	Quantitative analysis of ligand effects (QALE)	[98]
58	Ligand repulsive energy				
59	Molecular electrostatic potential				
61	Electronic parameters computed with semi-empirical methods				
62	Net donor parameters				
63	Infrared frequencies				
64	Free ligand HOMO				
65	Free energy between competing transition states leading todifferent enantiomers	organic	homogeneous catalysis	ML	[99]
66	Hammett parameters				
67	Temperature	organic	organometallic catalysts	QSAR	[100]
68	Reaction time				
69	Cartesian coordinates				
70	SMILES	organic	organometallic catalysts	ML	[101]
71	Reaction free energy				

a/ The parameters that we listed in Tables 2, 3, and 5 operate in the forward mode evaluating individual molecules or reaction s mode rather than molecular disconnections in the forward mode. For more details, the reader should compare the section (Retrosynthesis beyond TM); descriptors or (calculated/predicted)

properties. Mean_d, σ_d, mean_iV, σ_iV, mean_T, σ_T are six topological descriptors based on the first Delaunay shell. Group is the Periodic Table group of the element. e_g electrons is the nominal number of transition-metal electrons based on its formal oxidation state and spin state. SMILES is a line notation for entering and representing molecules and reactions.

Evaluating environmental nuisance of syntheses

While retrosynthesis alone provides insight into available synthesis options, it usually gives no systematic information on the greenness or sustainability of the processes or toxicity of reagents and solvents. With the increasing role of green chemistry, we needed an algorithmic method to evaluate the environmental benignness of chemistry. The **E factor** was designed as the first efficiency metric illustrating a fit to green chemistry criteria in the form of numerical value.

$$E = waste\ (kg) / product\ (kg)$$

Accordingly, the E-factor controls resource efficiency and waste minimization. Sheldon, who developed the concept of the E-factor discussed its history and perspectives in one of his recent publications.[102] We can observe that a massive production needs better E-factor optimization. For example, in the pharmaceutical industry, having a production of 10 —10^3 tons per year E factor takes a value between 25 and 100 kg waste/kg product, while in oil refining, yielding 10^6—10^8 tons per year, the E factor remains below 0. Fine chemicals and bulk chemicals of the production between the above-mentioned values also have E factors between the E factors for pharmaceuticals and oil refining.[103]

In particular, in Table 3 we specified several exemplary E factors calculated for individual syntheses, as reported in the literature.

Table 3 Examples of E factors values reported in literature for individual reaction

Process/Reaction	E-factor (kg waste/kg product)	References
Atropine synthesis	24	[104]
Diazepam synthesis	9	
Sonogashira cross-coupling	8 — 20	[105]
Ullman-type cross-coupling	9.70	[106]
Copper-catalyzed azide-alkyne cycloaddition	4.30	[107]
Suzuki reaction in azeotropic EtOH	3.5 — 3.9	[107]
Catalytic Cracking of Heavy Crude Oil	0.1	[108]

A variety of efficiency metrics for green chemistry are available currently. Illustrative examples are shown in Table 4. To realize the complexity of chemical problems, we can analyze such a simple metric as yield (Table 4, entry 1). In precise calculations, we involve the recovered starting material and the purity of the product. Typically, 95% spectroscopic purity is a level determining the isolated yield.[103] For a broader discussion of the metrics available, the reader should compare ref.[109]

Table 4 Quantitative indexes for GC-CASD

Index	Function	First publication
Yield	The ratio of the quantity of moles of a product formed to the limiting reactant consumed	
Conversion	The ratio of the quantity of moles of the limiting reactant consumed to its starting value	
Selectivity	The ratio of the quantity of moles of a desired product formed to the limiting reactant consumed	
Commercial availability	Reactant availability	
Atom Economy (AE)	The ratio of the atoms forming the product to all atoms in reagents	1991[110]
Atom Utilization (AU)	The ratio of the mass of product (kg) to the mass of all products (kg)	1992[111]
Environmental Impact Factor (E-factor)	The ratio of the total mass (kg) of all wastes to the mass of product (kg)	1992[112]
Mass Intensity (MI)	The total amount of mass required to produce a unit of product	2001[113]
Process Mass Intensity (PMI)	The ratio of the total materials used in a process (kg), with the exception of water, to the final product (kg)	2011[114]
Effective Mass Yield (EMY)	The ratio of the mass of products (kg) to the mass of non-benign reagents (kg)	1999[115]
Carbon Efficiency (CE)	The ratio of the carbon atoms in the product to all carbon atoms in substrates	2001[113]
Reaction Mass Efficiency (RME)	The ratio of the mass of isolated product (kg) to the total mass of reactants used in the reaction (kg)	2001[113]
Optimum Efficiency (OE)	The ratio of the Observed Reaction Mass Efficiency to theoretical Atom Economy	2015[116]
simple E-factor (sEF)	The ratio of the total mass of all raw materials and reagents (kg) excluded the final product to the mass of the product (kg)	2015[109]
complete E-Factor (cEF)	The ratio of the total mass of all raw materials and reagents (kg) excluded the final product, solvents, and water to the mass of the product (kg)	2015[109]
Solvent Intensity (SE)	The ratio of the total mass of solvents used (kg) to the mass of the product (kg)	2001[113]
Wastewater Intensity (WWI)	The ratio of the total mass of the process water (kg) to the mass of the product (kg)	2001[113]

Green Aspiration Level (GAL)	Transformation GAL times complexity	2015[109]
transformation GAL (tGAL)	The ratio of the sEF (or cEF) to average complexity	2015[109]
Relative Process Greenness (RPG)	The ratio of the GAL to actual sEF	2015[109]
Renewable Intensity (RI)	The ratio of the total mass of all renewable derivable materials used to the mass of the final product	2015[116]
Renewable Percentage (RP)	The ratio of the Renewable Intensity to a Reaction Mass Intensity	2015[116]
Innovative GAL (iGAL)	34.4% of the molecular weight of the salt-free form of the desired drug/product	2018[117]
Scale Risk Index (SRI)	Total scores for the human health, environmental and physical hazards times total mass of all substances and time for performing the synthesis	2017[118]

Green chemistry CASD (GC-CASD)

A natural goal of retrosynthesis is to simplify the synthetic availability of the targeted compound, making synthesis cheaper, less hazardous, and easier. Therefore, retrosynthesis naturally favors green solutions. For example, the Synthia application allowed to improve synthesis providing higher yields in shorter paths giving higher purity and decreasing the number of chromatography steps needed for product purification.[119]

To prioritize green chemistry in CASD, we should not only design a suitable synthetic path but also need to predict the chemical context of the reaction and operating conditions. Catalysts, reagents, solvents (chemical context), temperature, and pressure (operating conditions) are examples of the parameters critically important. The problem of the predictive design of reaction conditions is still not well explored; however, this issue has gained special interest recently.[120–124] Machine learning was used to predict suitable parameters for organic reactions to improve the accuracy and specificity of reaction outcome predictions.[125,126] In the context of green chemistry, efficient predictions would guide the search for green solutions, e.g., green solvents. Gao et al. designed a neural network model to predict up to one catalyst, two solvents, two reagents, and the temperature for a given organic reaction.[127] They trained the model on 10 million reactions extracted from the Reaxys to test it on 1 million reactions outside the training set. The prediction accuracy for combining chemical context with *the catalyst* and *at least one solvent* and *reagent was ca.* 69.6%.

CASD programs still prioritize synthetic success and chemical diversity of the pathways. Consideration of green chemistry rules is only scarce; for example, early examples in the references.[128,129] One reason for this is the complication of CASD itself. For example, the best first searches (BFS) or depth-first searches (DFS) synthesis tree search algorithms used in CASD needed manually formulated heuristics, which did not lead to sufficient CASD efficiency.[57,65] Therefore, making CASD more efficient needs more efficient algorithms with reasonable evaluation metrics. Trimming retrosynthetic tree complexity is an important objective being here a bottleneck. Segler applied the Monte Carlo Tree Search (MCTS) algorithm with symbolic architecture to reduce the complexity of the synthesis tree processing.[55] For the discussion of the technical problems of the MCTS method, the reader should see the reference.[130] Schreck et al. used similar method in their CASD system.[131,132] They recognized the CASD as a game performed by a chemist or a computer. The game rules to obtain the target molecule limits the use of chemicals for buyable compounds, i.e., these available at the market. The selection of the individual reaction depends upon the learning algorithm, i.e., the strategy an agent uses to pursue goals, or the so-called policy defined by the user. The user specifies the cost of performing a reaction, and this value is added to a running total, ultimately determining the overall synthesis cost. If the reactants identified are not buyable, we should plan their synthesis to count the total running cost. The authors used deep reinforcement learning to determine *policies* for the optimal reaction choices according to a user-defined cost metric in CASD.[131] A focus on buyable reactants optimizes economic cost but can also be a measure of green chemistry, for example, on a single laboratory scale.

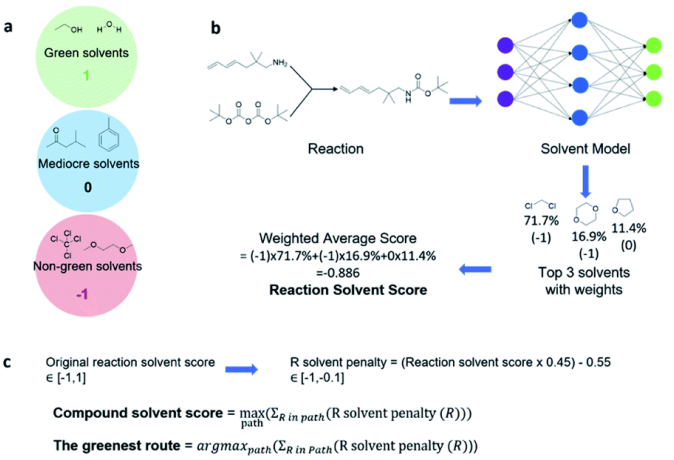

Figure 8. The MCTS architecture for CASD finding short paths with green solvents.[58]

We will refer to the CASD considering green chemistry as GC-CASD. Wang et al. adopted MCTS to design the GC-CASD architecture.[58] Unlike in classical CASD, a fundamental problem of GC-CASD is the rapid assessment of the environmental nuisance of reactants, solvents, catalysts, and so on. At the same time, since retrosynthesis often explores new synthetic pathways, it is necessary to predict the properties of compounds in specific chemical environments. We can refer to such operations as predictive green chemistry.

Solvents are usually simple chemicals, but optimization of their use is a complex multidimensional problem. An example is their design so that they have the potential to increase reagents' reactivity. Struebing et al. use quantum methods to select appropriate solvents.[133] Predictive catalyst design would be even more complicated. For example, could we extend a simple drug design heuristics of privilege structures[134] to privilege metal combinations[135], designing catalysts for environmental nanocatalysis?[136] Marcou et al. developed the system for the predictive design of catalysts for Micheal reaction.[137] Biocatalytic CASD is another problem explored in recent years.[67]

Technically, Wang's GC-CASD system uses the MCTS variant with reinforcement learning.[58] This CASD operates on reaction rules, or the so-called templates[57,138,139], and is limited to buyable chemicals, similar to other CASD algorithms.[55,131,140] The availability at the market estimates the *easiness of synthesis*. The catalog is a collection o ca. 100 000 low-price (under $100 per gram) compounds from Sigma Aldrich or eMolecules. The solvents were registered in the database to be represented by different scores depending upon their biosafety and flammability. Score values were defined after the solvent selection guides, as discussed by Byrne et al.[141] Green solvents such as water and ethanol obtained a score of 1, *mediocre* (poor), such as methyl isobutyl ketone or toluene, a score of 0, and *non-green* ones got -1. Figure 8 illustrates a neural network method used by authors to predict suitable solvents with a probability distribution given for a target reaction. The MCTS variant suggested shorter pathways with greener solvents than those reported earlier. Authors claim that the method could predict milder reaction temperatures and more economical catalysts. A similar procedure was designed by Gao et al.[126] In Table 4, we listed indicators that can be used to evaluate green chemistry aspects. Only some of them were used in the published CASD systems.

From green to sustainable CASD (S-CASD)

We need a global transformation of the chemical industry to sustainable routines. However, we still do not have generally efficient methods for predictive sustainability. Chemical intuition, manual data processing, and experience are still dominating practices. Weber et al. indicated the bottlenecks for automated predictive discovery in this area as (i) data, (ii) evaluation metrics, and (iii) decision-making.[142]

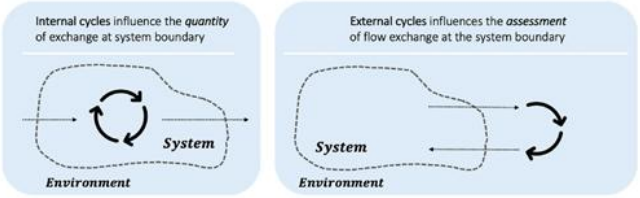

Figure 9. Sustainability needs the extension of the analysis beyond a single system. The circularity between the systems should influence the final analysis results. For example, the output flows (e.g., wastes) from outside could be used as the input to the system, potentially decreasing overall waste production.[142] Copyright © 2022 Royal Society of Chemistry

Not all GC-CASDs are sustainable. An example could be a search guided by buyable reactants. Although buying reagents, vs. their synthesizing, improves green chemistry locally for a specific lab, this is not necessarily true globally. The shift from *green* to *sustainable* systems means we should not care for a simple system. However, we should involve in an analysis a complex environment beyond the system boundary[143,144], as illustrated in Figure 9. Therefore, *sustainability* is much more complicated to predict than *greenness*. Weber et al. suggested using a reaction network (network of organic chemistry - NOC) for route selection and identifying strategic molecules for sustainable supply chains.[145–147] Notable, the NOC is a close analog to the network designed by Fialkowski et al. to discover the architecture of organic chemistry.[57,62,148–151] We should remember that this approach provided efficient knowledge for CASD design, resulting in the Chematica and Synthia.[7]

The new possibilities offered by the increasing computer power allowed for more efficient data mining and knowledge extraction from databases. For example, Voll and Marquardt designed Flux Analysis (RNFA) to identify optimal pathways for bio renewables on the literature extracted data.[152] An extensive review of this type of sustainable data exploration is given in the reference.[142]

Predictive sustainability is currently explored only slightly. Weber et al. indicated the criteria needed to evaluate the sustainability of the reaction network routes: (i) twelve principles of green chemistry, (ii) productivity scheme, (iii) extension of productivity scheme towards green engineering, (iv) improvement scheme.[142] Sustainability prediction should stimulate demand/supply outside the NOC boundaries. It should be automated. Jacob et al. predicting synthesis routes for converting a bio-waste feedstock, limonene, to a bulk intermediate, benzoic acid, developed a methodology for chemical route development and evaluation based on data mining.[145] They based these multi-criteria environmental sustainability evaluation on multiple indicators, including exergy, E-factor, solvent score, reaction reliability, and route redox efficiency. Thermodynamics-based metrics for ecological systems are discussed thoroughly in the book.[153] *Exergy* is an illustrative sustainability metric facilitating the second thermodynamics law in an ecological context.[153] Although theoretically, such metrics allow us to define better sustainability, we use much simpler practical measures for practical S-CASD evaluation. Exemplary heuristic rules for large-scale screening, can include carbon counts, catalysts, fragments, publication year, number of reaction records, reaction type, reagents, similarity, solvents, and yield. A function of these heuristics is to select the best possible solutions with the highest probability, e.g., remove old reactions, remove undesired solvents, remove undesired catalysts, remove reactions with yields lower than the threshold, etc.[142]

Table 5 Indexes for S-CASD

Metrics	Type	Effect	Reference
Total net primary energy usage rate (GJ/y)	Predictive	negative	154
Total net primary energy sourced from renewables (%)	Predictive	positive	154
Total net primary energy usage per kg product (kJ/kg)	Predictive	negative	154
Total net primary energy usage per unit value added (kJ/$)	Predictive	negative	154
Total weight of raw materials used per 1 kg of a product (kg/kg)	Predictive	negative	154
Total raw materials used per unit value added	Predictive	negative	154
Fraction of raw materials recycled	Predictive	positive	154
Hazardous raw material per kg product	Predictive	negative	154
Net water consumed per unit mass of product (kg/kg)	Predictive	negative	154
Net water consumed per unit value added	Predictive	negative	154
Waste reduction	Predictive	positive	154
Safety index	Predictive	positive	154
LCA – ozone layer depletion	Predictive	negative	154
LCA – photochemical oxidation	Predictive	negative	154
LCA – acidification	Predictive	negative	154
LCA – eutrophication	Predictive	negative	154
Carbon footprint – raw materials	Predictive	negative	154
Number of Records	Factual	positive	142
Publication Year	Factual	positive	142
Reaction Type	Factual	positive	142
Similarity	Predictive	positive	142
Yield	Predictive	positive	142
Number of hazardous materials input	Predictive	negative	155
Mass of hazardous materials input	Predictive	negative	155
Chemical exposure index	Predictive	negative	156
Health hazard, irritation factor	Predictive	negative	157,158
Health hazard, chronic toxicity factor	Predictive	negative	157,158
Safety hazard, fire/explosion	Predictive	negative	157,158
Safety hazard, reaction /decomposition	Predictive	negative	157,158
Safety hazard, acute toxicity	Predictive	negative	157,158
Fault tree assessment	Predictive	negative	159
Toxic release intensity	Predictive	negative	160
Environmental quotient	Predictive	negative	161
Environmental hazard, persistency of organic substances	Predictive	negative	157,158
Environmental hazard, air hazard	Predictive	negative	157,158
Environmental hazard, water hazard	Predictive	negative	157,158
Environmental hazard, solid waste	Predictive	negative	157,158
Environmental hazard, bioaccumulation	Predictive	negative	157,158
Global warming potential	Predictive	negative	162
Global warming intensity	Predictive	negative	155
Emergy to yield ratio	Predictive	positive	163
Emergy sustainability index	Predictive	positive	164
Renewability index	Predictive	positive	164
Total solid waste mass	Predictive	negative	155
Recycling mass fraction	Predictive	positive	155
Disposal mass fraction	Predictive	negative	155
Hazardous solid waste mass fraction	Predictive	negative	155
Total volume of liquid waste	Predictive	negative	155

A practical example of the extension of the green concept to sustainability is an application of the green aspiration level index iGAL 2.0 to reduce global API (active pharmaceutical ingredient) manufacturing wastes. Roschangar et al. developed iGAL for sustainability evaluation. The iGAL 2.0 extends the API (active pharmaceutical ingredient, drug substance) process waste, process mass intensity (PMI), and the complete E factor (cEF) metrics. They also showed how to adapt the yield (YD) and the convergence (CV) as sustainability measures.[165] Syntheses of similar yields arranged convergently provide better yields than those assembled linearly. Let us say that we have the reactions of the same yield value. Because the reaction yield is always a number lower than 100% (<1), the power function giving the total yield of linearly arranged syntheses decreases rapidly with an increasing power value. For the given number of starting materials (SM), the lower sum of subprocess steps (SSS) means a higher CV. Roschangar et al. modified simple CV metrics of SSS used previously by Hendrickson by using its relative value, SSavg, amounting to SSS/SM.

Yield is a measure of step productivity and is based on the molar limiting. For processes with two or more longest linear step sequences (LS) starting materials with the same step distance to the API, the starting material with the largest contribution to the API structure, or largest atom economical molecular weight (MW_{AE}), is prioritized. YD reflects the cumulative product of yields LS across steps (k).

$$YD = \prod_{k=1}^{LS} yield_k$$

Process Convergence (CV) indicates how directly the starting materials are assembled into the API and therefore reflects the efficiency of API process design. The appeal of using CV in combination with YD is that they are orthogonal and pertain to two complementary dimensions of process efficiency: design efficiency (CV) and productivity (YD).

$$CV_{iGAL} = \frac{1}{S_{avg}} \cdot 100\%$$

Accordingly, YD and CV could be crucial sustainability indicators for S-CASD algorithms if used for larger systems. Figure 10 shows the application of the CV and YD-optimized CASD. Authors claim that the iGAL 2.0 based CASD could significantly reduce global API (active pharmaceutical ingredient) manufacturing wastes. Actually, CV and YD evaluate green chemistry; however, the global dimension of the waste analysis upgrades the scale to sustainability.

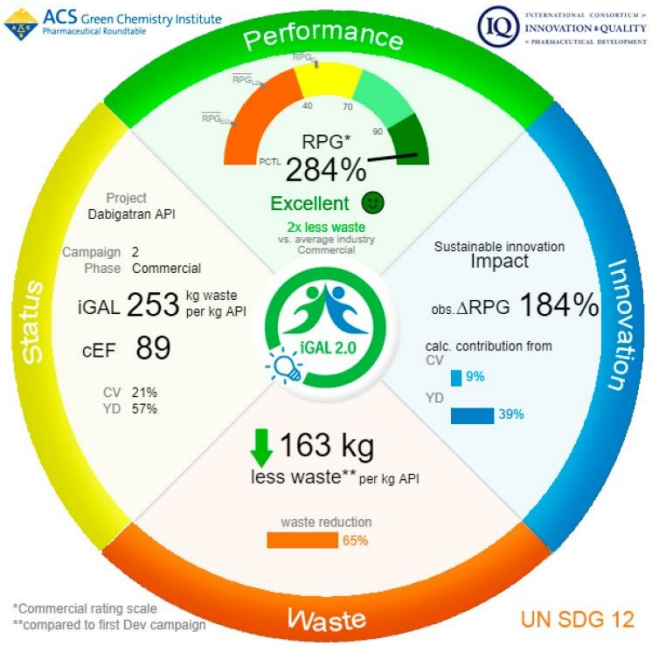

Figure 10. iGAL 2.0 scorecard output for 3rd generation Dabigatran API process. Copyright © 2022 American Chemical Society

The number of indicators explored for sustainability assessment is high. The United Nations framework of the Sustainable Development Goals (SDGs)[166] enumerates 231 unique indicators within 17 dimensions.[142] In Table 5, we specify indicators available in this area. A comparison between sustainability and greenness clearly illustrates how complex could be to define sustainability precisely enough. Many indexes are heuristics based on (chemical) intuition. The question is, can a large combination of such indexes efficiently control sustainability? The life cycle assessment (LCA) is a broader concept involved in sustainable development.[167]

To better understand the S-CASD potential, we illustrated in Figure 11a the first total synthesis of tropinone performed by Willstatter. The linear process performed in 1901 needed as many as 15 steps, yielding 0.5% of the product. Despite this small yield, Willstatter's synthesis was a masterpiece of chemistry designed by the Nobel Prize winner. A first total synthesis of tropinone. Could we design a better procedure that minimizes the number of steps, significantly increasing the process's greenness and sustainability? In Figure 11b, we illustrated the retrosynthesis of tropinone by our vertical hand-made mode, showing what is familiar to all current organic chemists. Tropinone can be obtained practically in a single step with almost 100% yield. In 1917, Robinson, another Noble Prize winner, reported this synthetic way. He obtained originally 17% yield, which was improved later to 90%

An original approach to sustainability is a search for bio-based building blocks of specific chemicals (solvents or reagents) from bio-based products. A represenattive example of bio-based retrosynthesis can be the publication of Moity et al. describing the

bio-based building block strategy in which authors were able to design a few-step synthesis of a large number of commodity chemicals. In particular, the authors used the CASD software, GRASS (Generator of Agro-based Sustainable Solvents), to design solvents from biomass. GRASS was programmed in 1980' to predict the products of flavor and food degradation products [R. Barone, C. Chanon, G. Vernin and C. Parkanyi, in Food Flavor and Chemistry; Explorations into the 21st Century, ed. A. M. Spanier, F. Shahidi, T. H. Parliment, C. Mussinan and E. Tratras Contis, RSC, Cambridge, UK, 2005, pp. 175–212.]. The study engaged as a starting chemical itaconic acid, a multifunctional bio-based building block. Figure XX briefly illustrates the results reported in reference [Moity et al., In silico design of bio-based commodity chemicals: application to itaconic acid based solvents, Green Chem., 2014, 16, 146 , DOI: 10.1039/c3gc41442f].

Figure 12 compares the GRASS (dotted arrow) vs. experimental (black arrow) pathways to compounds that are reported in the literatures as as solvents derived from itaconic acid; 2-MBDO, 2-MGBL, 3-MGBL, 3-MTHF and NACP are the codes of bolded compounds, the numbers above dotted arrows are references to the table entries in the original publication. We see a forward mode retrosynthetic visualization. © [In silico design of bio-based commodity chemicals: application to itaconic acid based solvents, Green Chem., 2014, 16, 146 , DOI: 10.1039/c3gc41442f].

Recently, the Grzybowski group published an entirely novel approach to sustainable chemistry, where authors designed novel synthetic routes to the essential drugs using the waste chemicals available by in silico retrosynthesis [Wolos et al., Computer-designed repurposing of chemical wastes into drugs, Nature https://doi.org/10.1038/s41586-022-04503-9]. Dapsone is a sulfone class antibiotic known from 1930'. Their current application is leprosy, for which FDA approved it in 2016. In Figure 12 we compared their approach which we can describe as the *waste building block* appropach to dapsone synthesis to the currently synthetic procedure.

Interestingly the comparison of retrosyntheses in Figures 11 and 12 also reveals the differences between the hand-made vs. in silico black box retrosyntheses. While the first approach is an essential issue to provide us with the backward disconnections comprehendible to a chemist working in the product-to-reactant strategy, the second one does not even show the backward process hidden from our view. Usually, in this mode we present a forward reaction scheme, even not pretending that the software user can intrude on the computer algorithm. Instead, the obtained scheme is even more accessible for human interpretation, while the usability of the results for substantial sustainability increase is impressive. The *waste-building block* strategy retrosynthesis is a significant breakthrough in sustainability. A variety of retrosyntheses with this strategy the reader can find in the reference [Nature (a) https://doi.org/10.1038/s41586-022-04503-9]

Retrosynthesis beyond target molecule

Virtually, retrosynthetic analysis involves two problems. The first is the design of the reaction sequence giving the target molecule, which is the primary goal in the organic chemistry domain. Optimization of disconnections leads to a comprehensive library of reactions that can be evaluated to select proper parameters for these reactions, in the search for brilliant green and sustainable reactants and conditions. The second problem, optimization of individual syntheses and processes is independent of the target molecule. An example of an decisively target independent green and sustainability level increase could be computer-aided sustainability-oriented process optimization. An excellent introduction to this problem the reader can find in the references

Similarly, the parameters that we listed in Tables 2, 3 and 5 ranges from these that can be directly engaged in the prediction or design procedures (in the target dependent and independent modes) to those that are vague heuristics. Let us indicate the publication year (Table 5), which is indicated in the literature as a sustainability index. It is well-defined but soft concerning a correlation to experimental data reliability.

Evaluating the toxicity of molecules and chemicals is an important metric for green chemistry or sustainable CASD. Therefore predictive toxicology is an essential issue for G-, S- or nano-CASD. Since we do not plan molecule synthesis here, predictive toxicology differs from classical retrosynthesis, being an example of the operation beyond TM. Great progress has been made in this direction, including the application of a variety of machine learning methods: support vector machines (SVMs), random forest (RF) and decision trees (DTs), neural networks, regression models, naïve Bayes, k-nearest neighbors or ensemble learning. The illustrative review in predictive toxicology the reader can find in the reference [Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models, Chem. Res. Toxicol. 2021, 34, 217−239].

Safe- and Sustainable-by-Design (SSbD) concept for nanomaterial S-CASD architectures

In the concept of Safe- and Sustainable-by-Design (SSbD), the design process of chemicals and (nano)materials should consider both safety assessment (i.e., identification of the potential risk that newly designed (nano)materials bring to the humans and the environment) and the sustainability, i.e., environmental, social or economic impact at the early stage of the design process of innovative (nano)materials.[168] The materials based on nanostructures offer unique physicochemical properties compared to the same micro or macro scale chemicals. However, in the literature the systematic knowledge about the influence of nanostructure modification or nanostructure combination on the functionality and safety of newly designed advanced (nano)materials is still limited.[169] Due to the high cost and time of experiments, investigating all possible structure combinations to design the most optimal (safe- and sustainable) products is impracticable. The most promising approaches that may support the design of Safe- and

Sustainable nanomaterials in a shorter time are based on in silico methods such as physics-based materials modeling (MM)[170] and data-based methods with Machine Learning (ML). Nano Quantitative Structure-Activity Relationship (nano-QSAR) is a related novel tool developed[171], but generally, this area still needs exploration. In Figure 11 we showed how nano-QSAR could be used for designing novel parameters for descriptive sustainable nano-retrosynthesis, while Table 6 specifies the descriptors and properties potentially useful for nano-CASD.

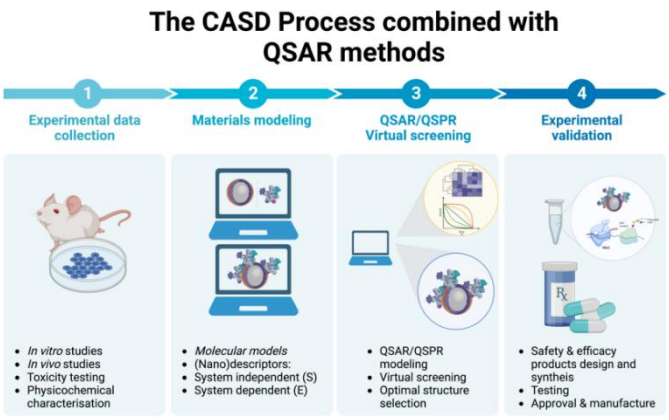

Figure 11. The schematic representation of CASD methods combined with MM/ML-based modeling.

One problem of nano-QSAR modeling are nanostructure representation and characterization.[171] Wyrzykowska et al. reviewed the most promising directions for developing the appropriated nanostructure representation described by a set of nanodescriptors that enhance the reliability of computational methods for Safe- and Sustainable-by-Design Strategy.[169] This study answers how to represent and describe nanomaterials in predictive nanoinformatics based on combined molecular modeling (MM) and machine learning (ML) techniques.[169] The concept[169] includes a system perspective that is crucial for nanomaterials characterization, i.e., (i) system-independent nanodescriptors that describe nanostructure (e.g., chemical composition, crystal structure, size, shape, surface structure), so-called S-descriptors; (ii) system dependent descriptors that describe particular elements of a nanoparticle and its environment in real-time (i.e., core, coating, ligands), so-called E-descriptors, and (iii) multicomponent structure represented by numerical combinations of descriptors for individual components of the nanostructure and descriptors related to interactions between these components.[172–174] According to JRC[168], one of the most promising data-driven methods to process the relationship between toxicological, environmental or physicochemical data and the nanomaterial structure is related to predictive QSAR, Nano-QSAR modeling and read-across methods, respectively. In the chemical sector, more than 200 000 chemicals have a limited number of information crucial for the design process of safe and sustainable nanoproducts.[168] Chemical-specific data gaps related to the specific process (including lack or inconsistency in data resources or synthesis process) may be overcome by nano-QSAR methods and read-across. These methods may predict missing datapoints (so-called endpoints, such as toxicity, ecotoxicity, or physicochemical properties) helping also to model quantitative relationship between synthesis conditions, nanostructure properties and its safety or sustainability profiles. The nano-QSAR methods combined with molecular modeling (MM) and ML screening may help to answer which structural features of nanomaterials are responsible for the observed toxic effects or physicochemical properties of industry interest.[175–178] Implementation of virtual screening in SSbD strategy may support the screening of a huge library of virtually created nanostructures and their combination, then manage ("design out") hazardous features (safety) as well as functionality (sustainability) at the earliest possible manufacturing step.[179] As a result of virtual screening, only the most optimal structures (i.e., components described by specific features and safety) may be finally selected for synthesis and experimental validation.[176,178] By the application of nano-QSAR methods and read-across, it is possible to reduce cost, time, and the number of necessary experiments and, at the same time, increase the efficiency of the design process. Thus, the combination of various in silico methods in the S-CASD architecture offers novel opportunities for knowledge-based optimization and development of new nanomaterials by improving their functionality and minimizing the potential unexpected risk to the human body and the environment.

Table 6 Descriptors for nano S-CASD

Descriptors	Reference
Electronegativity	
Sum of metal electronegativity for individual metal oxide divided by the number of oxygen atoms present in particular metal oxide	
Number of metal atoms	180
Number of oxygen atoms	
Charge of the metal cation corresponding to a given oxide	
Molecular Weight	
Enthalpy of formation of a gaseous cation having the same oxidation state as that in the metal oxide structure	181
Lattice energy	
Energy of highest occupied molecular orbital	182
Energy of lowest unoccupied molecular orbital	

Dipole moment
Mean polarizability
Largest Mulliken negative charge
Ratio of surface molecules to molecules in volume
Aggregation parameter
Covalent index
Cation polarizing power
mass density
Wigner–Seitz radius of oxide's molecule
Zeta potential in water
Number of electron shells in the metal of the oxides

183

184

Conclusions

When retrosynthesis started computers were newborns hardly adapted to chemical applications. The chemical audience received Corey's retrosynthesis idea with skepticism. Early computers were not ready to process chemical ideas and data. Therefore, teaching chemistry to computers was one of the first tasks of chemoinformatics. The view of the human chemists differed significantly from the computer needs. We required many efforts and ideas to transit chemical art to efficient chemical informatics. Today, in silico information processing and human brain chemistry arranges into complementary structures. The need to adapt chemistry to a form understandable to computers has brought much freshness to chemistry, allowing chemists to understand and clear out some ambiguities. In turn, by analyzing the work of the computer, chemists learned a different view of chemistry. It was the computer that became the teacher. In the case of retrosynthesis, we observe how the computational dimension makes the original symbolic product-to-reagents symbolic visualization layout lose meaning. The CASD strategy is a complex problem engaging both backward and forward strategies, and today, the results of retrosynthesis are often presented directly as the reagents-to-product process. Playing retro (backward) or forward strategy visualization is easily interchangeable with modern software. The classical retro-analysis stayed somewhere below the software level recognized by the user. However, the user accepts this black box-like architecture because we highly appreciate outstanding results of modcern CASDs. When analyzing chemists' sentiments, we can observe a bumpy road from initial reserve to enthusiasm. With suitable software, chemists can design synthesis efficiently.

Retrosynthesis is a tool allowing for finding reaction pathways to chemical compounds. Chemistry decides that in retrosynthesis, the molecules can be disconnected in many ways, indicating a variety of potential synthetic routes to products called retrosynthetic trees. From one side, the abundance of retrosynthetic trees is a severe complication. Therefore, in standard hand-made retrosynthesis, chemists tend to limit the number of possible solutions keeping it under the control of human brain capabilities. Even in computer-assisted synthesis design (CASD), retrosynthetic trees need trimming for efficient analysis. However, this abundance can be an opportunity for finding the path to reducing toxicity to humans and the environment, improving the degradability of chemicals, and their recycling or reusing potential. Drug design and retrosynthesis were the methods significantly improved recently to computer-assisted technologies. Machine learning and deep learning support this area.

Selecting retrosynthetic trees with a high fit with green chemistry requires unique methods. Similarly, we need particular strategies to support the new sustainable chemistry paradigm, which focuses not on a simple chemical reaction but on extending the analysis borders beyond individual processes. Data mining allows for exploring databases and literature for green and sustainable methods. However, in order to design novel chemistry, we need predictions. Predictive greenness and sustainability need efficient metrics for numerical evaluation. E-factor was the first index designed to estimate the reaction environmental fit. Various modifications were developed and used in the descriptive and predictive role. Retrosynthesis beyond the organic chemical subspace is a recent idea involving inorganic and nano domains, which targets nanostructures of the desired shapes and sizes. Predictive sustainability in (nano)materials needs novel methods that may speed up the design process. Combined MM with ML methods (including predictive QSAR or Nano-QSAR modeling) may significantly increase the *in silico* potential for designing safe and sustainable (nano) materials (i.e., more structure with different properties may be considered, and a more comprehensive description of nanostructure would be possible at the same time). The described *in silico* methods for CASD may reduce time, cost, and number of experiments, which is crucial for future optimization of the design process of safe and sustainable (nano)materials.

Although routine GC- S-or nano CASDs are still a matter of the future, some early birds have appeared in this area. The story of retrosynthesis resembles this of chess-playing software in the last decade. In 90’ human players used to win with AI algorithms. The Deep Blue chess-playing algorithm first brought a draw; nowadays, no human can compete with a machine in chess. In retrosynthesis, the machine and humans can produce similar results now that the Turing test cannot

distinguish. Similarly to classical organic retrosynthesis, the GC-, S-or nano CASDs will also profit from humans playing with computers who bring entirely novel ideas. The waste-oriented building blocks approach [Wolos et al., Computer-designed repurposing of chemical wastes into drugs, Nature https://doi.org/10.1038/s41586-022-04503-9] is an example of creatively overcoming sustainability problems with a fresh computer-oriented concept. In perspective, we can better mount the green, sustainability or nano indexes to the GC- S-or nano CASDs to essentially finetune the results. Because the parameter-oriented GC- S-or nano CASDs will need extensive data processing in the backward and forward modes, we can expect the human-computer interplay will provide the black-box-like architectures for the big data-dependent mode in which humans cannot easily compete with computers. These architectures will engage all available measured reactions and chemical data available for chemical compounds and reactions available in the literature as these listed in Tables 2-5. In conclusion, we can expect significant development and profit from the CASD strategies here.

Author Contributions

Conceptualization: JP, AM, PS, TP; Formal Analysis: JP, AM, TP, WZ, PS, SA, AS; Funding acquisition: JP; Visualization: WZ, JP, AM; Writing – original draft: JP, AM, PS, WZ, SA, AS, KJ, MH, TP; Writing – review & editing: JP, AM, PS, WZ, SA, AS, KJ, MH, TP

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Authors thanks National Science Center OPUS 2018/29/B/ST8/02303. The research activities co-financed by the funds granted under the Research Excellence Initiative of the University of Silesia in Katowice, Poland.

Notes and references

1 D. J. C. Constable, *Curr. Opin. Green Sustain. Chem.*, 2017, **7**, 60–62.
2 A. H. Johnstone, *J. Comput. Assist. Learn.*, 1991, **7**, 75–83.
3 A. H. Johnstone, *J. Chem. Educ.*, 1997, **74**, 262.
4 E. J. Corey and X.-M. Cheng, *Logic of chemical synthesis*, Wiley, New York, New ed., 1995.
5 M. Aykol, J. H. Montoya and J. Hummelshøj, *J. Am. Chem. Soc.*, 2021, **143**, 9244–9259.
6 K. Vaddi, H. Thart Chiang and L. D. Pozzo, *Digit. Discov.*, 2022, **1**, 502–510.
7 M.-A. Plourde and S. Hallé, in *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 207–211.
8 ASKCOS Homepage, https://askcos.mit.edu/, (accessed October 2, 2022).
9 E. J. Corey, *Pure Appl. Chem.*, 1967, **14**, 19–38.
10 J. Fuhrhop and G. Penzlin, *Organic Synthesis; Concepts, Methods, Starting Materials*, VCH, Weinheim, 2nd edn., 1994.
11 J. M. Smith, S. J. Harwood and P. S. Baran, *Acc. Chem. Res.*, 2018, **51**, 1807–1817.
12 Y. Shen, J. E. Borowski, M. A. Hardy, R. Sarpong, A. G. Doyle and T. Cernak, *Nat. Rev. Methods Primer*, 2021, **1**, 23.
13 X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler, A. J. Norquist and J. Schrier, *Nature*, 2019, **573**, 251–255.
14 Z. Li, S. Wang, W. S. Chin, L. E. Achenie and H. Xin, *J. Mater. Chem. A*, 2017, **5**, 24131–24138.
15 J. Polanski, A. Bak, R. Gieleciak and T. Magdziarz, *J. Chem. Inf. Model.*, 2006, **46**, 2310–2318.
16 J. Gasteiger and T. Engel, Eds., *Chemoinformatics: a textbook*, Wiley-VCH, Weinheim, 2003.
17 A. Cyraniak, S. Freza and P. Skurski, *Chem. Phys.*, 2022, **559**, 111543.
18 H. C. Brown and G. Zweifel, *J. Am. Chem. Soc.*, 1961, **83**, 2544–2551.
19 M. W. Rathke and R. Kow, *J. Am. Chem. Soc.*, 1972, **94**, 6854–6856.
20 J. A. Pople, M. Head-Gordon and K. Raghavachari, *J. Chem. Phys.*, 1987, **87**, 5968–5975.
21 J. Gauss and D. Cremer, *Chem. Phys. Lett.*, 1988, **150**, 280–286.
22 E. A. Salter, G. W. Trucks and R. J. Bartlett, *J. Chem. Phys.*, 1989, **90**, 1752–1766.
23 Chr. Møller and M. S. Plesset, *Phys. Rev.*, 1934, **46**, 618–622.
24 M. Head-Gordon, J. A. Pople and M. J. Frisch, *Chem. Phys. Lett.*, 1988, **153**, 503–506.
25 M. J. Frisch, M. Head-Gordon and J. A. Pople, *Chem. Phys. Lett.*, 1990, **166**, 275–280.
26 R. LeFebvre and C. Moser, Eds., *Advances in Chemical Physics: LeFebvre/Advances*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 1969.
27 G. D. Purvis and R. J. Bartlett, *J. Chem. Phys.*, 1982, **76**, 1910–1918.
28 G. E. Scuseria, C. L. Janssen and H. F. Schaefer, *J. Chem. Phys.*, 1988, **89**, 7382–7387.
29 L. A. Curtiss, K. Raghavachari, G. W. Trucks and J. A. Pople, *J. Chem. Phys.*, 1991, **94**, 7221–7230.
30 L. A. Curtiss, P. C. Redfern and K. Raghavachari, *J. Chem. Phys.*, 2007, **126**, 084108.
31 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
32 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
33 T. Yanai, D. P. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51–57.
34 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
35 A. D. McLean and G. S. Chandler, *J. Chem. Phys.*, 1980, **72**, 5639–5648.
36 R. Krishnan, J. S. Binkley, R. Seeger and J. A. Pople, *J. Chem. Phys.*, 1980, **72**, 650–654.
37 T. H. Dunning, *J. Chem. Phys.*, 1989, **90**, 1007–1023.
38 R. A. Kendall, T. H. Dunning and R. J. Harrison, *J. Chem. Phys.*, 1992, **96**, 6796–6806.
39 J. Polanski and J. Gasteiger, in *Handbook of Computational Chemistry*, Springer, Cham, 2nd edn., 2017, pp. 1997–2039.
40 J. Polanski, in *Encyclopedia of bioinformatics and computational biology*, Elsevier, Amsterdam, Netherlands, 2019, pp. 601–618.
41 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547–555.
42 S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191–201.

43 O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nat. Commun.*, 2017, **8**, 15679.
44 O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha and S. Curtarolo, *Chem. Mater.*, 2015, **27**, 735–743.
45 N. O'Boyle and A. Dalke, DOI:10.26434/chemrxiv.7097960.v1.
46 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn. Sci. Technol.*, 2020, **1**, 045024.
47 U. V. Ucak, I. Ashyrmamatov, J. Ko and J. Lee, *Nat. Commun.*, 2022, **13**, 1186.
48 D. Lach, U. Zhdan, A. Smolinski and J. Polanski, *Int. J. Mol. Sci.*, 2021, **22**, 5176.
49 Z. Wang and P. Hu, *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, 2016, **374**, 20150078.
50 V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis, *Nature*, 2015, **518**, 529–533.
51 K. V. Chuang, L. M. Gunsalus and M. J. Keiser, *J. Med. Chem.*, 2020, **63**, 8705–8722.
52 J. Polanski, *Int. J. Mol. Sci.*, 2022, **23**, 2797.
53 J. Gasteiger and J. Zupan, *Angew. Chem. Int. Ed. Engl.*, 1993, **32**, 503–527.
54 J. Polanski, F. Zouhiri, L. Jeanson, D. Desmaële, J. d'Angelo, J.-F. Mouscadet, R. Gieleciak, J. Gasteiger and M. Le Bret, *J. Med. Chem.*, 2002, **45**, 4647–4654.
55 M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
56 B. VanVeller, *J. Chem. Educ.*, 2021, **98**, 2726–2729.
57 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.
58 X. Wang, Y. Qian, H. Gao, C. W. Coley, Y. Mo, R. Barzilay and K. F. Jensen, *Chem. Sci.*, 2020, **11**, 10959–10972.
59 T. Badowski, E. P. Gajewska, K. Molga and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2020, **59**, 725–730.
60 B. A. Grzybowski, S. Szymkuć, E. P. Gajewska, K. Molga, P. Dittwald, A. Wołos and T. Klucznik, *Chem*, 2018, **4**, 390–398.
61 B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Mlynarski, M. Mrksich and B. A. Grzybowski, *Nature*, 2020, **588**, 83–88.
62 B. A. Grzybowski, K. J. M. Bishop, B. Kowalczyk and C. E. Wilmer, *Nat. Chem.*, 2009, **1**, 31–36.
63 B. Grzybowski, S. Szymkuć, K. Molga, E. P. Gajewska and A. Wolos, *CHIMIA*, 2017, **71**, 512–512.
64 K. Lin, Y. Xu, J. Pei and L. Lai, *Chem. Sci.*, 2020, **11**, 3355–3364.
65 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
66 I. V. Tetko, P. Karpov, R. Van Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 5575.
67 S. Zheng, T. Zeng, C. Li, B. Chen, C. W. Coley, Y. Yang and R. Wu, *Nat. Commun.*, 2022, **13**, 3342.
68 M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218–10239.
69 A. Extance, *Chem. World*.
70 S. Lemonick, *Chem. Eng. NEWS*, 2019, **97**, 6.
71 Reaxys Predictive Retrosynthesis | Award winner, https://www.elsevier.com/solutions/reaxys/predictive-retrosynthesis, (accessed December 1, 2022).
72 N. N. Kiselyova, A. V. Stolyarenko, V. V. Ryazanov, O. V. Sen'ko, A. A. Dokukin and V. V. Podbel'skii, *Pattern Recognit. Image Anal.*, 2011, **21**, 88–94.
73 N. Kiselyova, V. Dudarev and A. Stolyarenko, in *Data Analytics and Management in Data Intensive Domains*, eds. A. Pozanenko, S. Stupnikov, B. Thalheim, E. Mendez and N. Kiselyova, Springer International Publishing, Cham, 2022, pp. 151–165.
74 E. Kim, K. Huang, O. Kononova, G. Ceder and E. Olivetti, *Matter*, 2019, **1**, 8–12.
75 O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan and G. Ceder, *Sci. Data*, 2019, **6**, 203.
76 A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig and A. Mar, *Chem. Mater.*, 2016, **28**, 7324–7331.
77 M. Aykol, V. I. Hegde, L. Hung, S. Suram, P. Herring, C. Wolverton and J. S. Hummelshøj, *Nat. Commun.*, 2019, **10**, 2018.
78 S. A. Malik, R. E. A. Goodall and A. A. Lee, *Chem. Mater.*, 2021, **33**, 616–624.
79 J. Jang, G. H. Gu, J. Noh, J. Kim and Y. Jung, *J. Am. Chem. Soc.*, 2020, **142**, 18836–18843.
80 E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder and E. Olivetti, *Chem. Mater.*, 2017, **29**, 9436–9444.
81 S. Saxena, T. Suvra Khan, F. Jalid, M. Ramteke and M. Ali Haider, *J. Mater. Chem. A*, 2020, **8**, 107–123.
82 S. Back, K. Tran and Z. W. Ulissi, *ACS Catal.*, 2019, **9**, 7651–7659.
83 M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi and E. H. Sargent, *Nature*, 2020, **581**, 178–183.
84 O. Mamun, K. T. Winther, J. R. Boes and T. Bligaard, *Npj Comput. Mater.*, 2020, **6**, 1–11.
85 T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa and K. Shimizu, *ACS Catal.*, 2020, **10**, 2260–2297.
86 B. R. Goldsmith, J. Esterhuizen, J. Liu, C. J. Bartel and C. Sutton, *AIChE J.*, 2018, **64**, 2311–2323.
87 H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik and E. Kumacheva, *Nat. Rev. Mater.*, 2021, **6**, 701–716.
88 T. Le, V. C. Epa, F. R. Burden and D. A. Winkler, *Chem. Rev.*, 2012, **112**, 2889–2919.
89 Z. Li, X. Ma and H. Xin, *Catal. Today*, 2017, **280**, 232–238.
90 X. Ma, Z. Li, L. E. K. Achenie and H. Xin, *J. Phys. Chem. Lett.*, 2015, **6**, 3528–3533.
91 S. Yang, M. Lach-hab, I. I. Vaisman and E. Blaisten-Barojas, in *Computational Science – ICCS 2009*, eds. G. Allen, J. Nabrzyski, E. Seidel, G. D. van Albada, J. Dongarra and P. M. A. Sloot, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, vol. 5545, pp. 160–168.
92 S. Yang, M. Lach-hab, I. I. Vaisman and E. Blaisten-Barojas, *J. Phys. Chem. C*, 2009, **113**, 21721–21725.
93 T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. Shimizu and I. Takigawa, *J. Phys. Chem. C*, 2018, **122**, 8315–8326.
94 N. J. O'Connor, A. S. M. Jonayat, M. J. Janik and T. P. Senftle, *Nat. Catal.*, 2018, **1**, 531–539.
95 K. Suzuki, T. Toyao, Z. Maeno, S. Takakusagi, K. Shimizu and I. Takigawa, *ChemCatChem*, 2019, **11**, 4537–4547.
96 E.-J. Ras, M. J. Louwerse and G. Rothenberg, *Catal. Sci. Technol.*, 2012, **2**, 2456.
97 W. T. Hong, R. E. Welsch and Y. Shao-Horn, *J. Phys. Chem. C*, 2016, **120**, 78–86.
98 N. Fey, *Dalton Trans*, 2010, **39**, 296–310.
99 M. S. Sigman, K. C. Harper, E. N. Bess and A. Milo, *Acc. Chem. Res.*, 2016, **49**, 1292–1301.
100 E. Burello, D. Farrusseng and G. Rothenberg, *Adv. Synth. Catal.*, 2004, **346**, 1844–1853.
101 B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld and C. Corminboeuf, *Chem. Sci.*, 2018, **9**, 7069–7077.
102 R. A. Sheldon, *Green Chem.*, 2017, **19**, 18–43.
103 M. C. Pirrung, *Handbook of Synthetic Organic Chemistry*, Academic Press, 2016.

104 A.-C. Bédard, A. R. Longstreet, J. Britton, Y. Wang, H. Moriguchi, R. W. Hicklin, W. H. Green and T. F. Jamison, *Bioorg. Med. Chem.*, 2017, **25**, 6233–6241.

105 V. Kozell, M. McLaughlin, G. Strappaveccia, S. Santoro, L. A. Bivona, C. Aprile, M. Gruttadauria and L. Vaccaro, *ACS Sustain. Chem. Eng.*, 2016, **4**, 7209–7216.

106 F. Ferlin, V. Trombettoni, L. Luciani, S. Fusi, O. Piermatti, S. Santoro and L. Vaccaro, *Green Chem.*, 2018, **20**, 1634–1639.

107 F. Valentini and L. Vaccaro, *Molecules*, 2020, **25**, 5264.

108 E. Villamarin-Barriga, J. Canacuán, P. Londoño-Larrea, H. Solís, A. De La Rosa, J. F. Saldarriaga and C. Montero, *Catalysts*, 2020, **10**, 736.

109 F. Roschangar, R. A. Sheldon and C. H. Senanayake, *Green Chem.*, 2015, **17**, 752–768.

110 B. Trost, *Science*, 1991, **254**, 1471–1477.

111 R. A. Sheldon, in *Industrial Environmental Chemistry*, eds. D. T. Sawyer and A. E. Martell, Springer US, Boston, MA, 1992, pp. 99–119.

112 R. A. Sheldon, *Chem Ind*, 1992, **23**, 903–906.

113 A. D. Curzons, D. N. Mortimer, D. J. C. Constable and V. L. Cunningham, *Green Chem.*, 2001, **3**, 1–6.

114 C. Jimenez-Gonzalez, C. S. Ponder, Q. B. Broxterman and J. B. Manley, *Org. Process Res. Dev.*, 2011, **15**, 912–917.

115 T. Hudlicky, D. A. Frey, L. Koroniak, C. D. Claeboe and L. E. Brammer Jr., *Green Chem.*, 1999, **1**, 57–59.

116 C. R. McElroy, A. Constantinou, L. C. Jones, L. Summerton and J. H. Clark, *Green Chem.*, 2015, **17**, 3111–3121.

117 F. Roschangar, Y. Zhou, D. J. C. Constable, J. Colberg, D. P. Dickson, P. J. Dunn, M. D. Eastgate, F. Gallou, J. D. Hayler, S. G. Koenig, M. E. Kopach, D. K. Leahy, I. Mergelsberg, U. Scholz, A. G. Smith, M. Henry, J. Mulder, J. Brandenburg, J. R. Dehli, D. R. Fandrick, K. R. Fandrick, F. Gnad-Badouin, G. Zerban, K. Groll, P. T. Anastas, R. A. Sheldon and C. H. Senanayake, *Green Chem.*, 2018, **20**, 2206–2211.

118 R. C. C. Duarte, M. G. T. C. Ribeiro and A. A. S. C. Machado, *J. Chem. Educ.*, 2017, **94**, 1255–1264.

119 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.

120 B. J. Reizman and K. F. Jensen, *Chem. Commun.*, 2015, **51**, 13290–13293.

121 V. Sans, L. Porwol, V. Dragone and L. Cronin, *Chem. Sci.*, 2015, **6**, 1258–1264.

122 N. Holmes, G. R. Akien, R. J. D. Savage, C. Stanetty, I. R. Baxendale, A. John Blacker, B. A. Taylor, R. L. Woodward, R. E. Meadows and R. A. Bourne, *React. Chem. Eng.*, 2016, **1**, 96–100.

123 B. J. Reizman and K. F. Jensen, *Acc. Chem. Res.*, 2016, **49**, 1786–1796.

124 L. M. Baumgartner, C. W. Coley, B. J. Reizman, K. W. Gao and K. F. Jensen, *React. Chem. Eng.*, 2018, **3**, 301–311.

125 C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 434–443.

126 H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2018, **4**, 1465–1476.

127 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.

128 D. J. C. Constable, P. J. Dunn, J. D. Hayler, G. R. Humphrey, J. Johnnie L. Leazer, R. J. Linderman, K. Lorenz, J. Manley, B. A. Pearlman, A. Wells, A. Zaks and T. Y. Zhang, *Green Chem.*, 2007, **9**, 411–420.

129 S. G. Koenig, D. K. Leahy and A. S. Wells, *Org. Process Res. Dev.*, 2018, **22**, 1344–1359.

130 C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis and S. Colton, *IEEE Trans. Comput. Intell. AI Games*, 2012, **4**, 1–43.

131 J. S. Schreck, C. W. Coley and K. J. M. Bishop, *ACS Cent. Sci.*, 2019, **5**, 970–981.

132 jsschreck, 2022.

133 H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Siougkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo and C. S. Adjiman, *Nat. Chem.*, 2013, **5**, 952–957.

134 P. Schneider and G. Schneider, *Angew. Chem. Int. Ed.*, 2017, **56**, 7971–7974.

135 J. Polanski, D. Lach, M. Kapkowski, P. Bartczak, T. Siudyga and A. Smolinski, *Catalysts*, 2020, **10**, 992.

136 T. Siudyga, M. Kapkowski, P. Bartczak, M. Zubko, J. Szade, K. Balin, S. Antoniotti and J. Polanski, *Green Chem.*, 2020, **22**, 5143–5150.

137 G. Marcou, J. Aires de Sousa, D. A. R. S. Latino, A. de Luca, D. Horvath, V. Rietsch and A. Varnek, *J. Chem. Inf. Model.*, 2015, **55**, 239–250.

138 C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison and K. F. Jensen, *Science*, 2019, **365**, eaax1566.

139 M. H. S. Segler and M. P. Waller, *Chem. – Eur. J.*, 2017, **23**, 5966–5971.

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

141 F. P. Byrne, S. Jin, G. Paggiola, T. H. M. Petchey, J. H. Clark, T. J. Farmer, A. J. Hunt, C. Robert McElroy and J. Sherwood, *Sustain. Chem. Process.*, 2016, **4**, 7.

142 J. M. Weber, Z. Guo, C. Zhang, A. M. Schweidtmann and A. A. Lapkin, *Chem. Soc. Rev.*, 2021, **50**, 12013–12036.

143 P. Marion, B. Bernela, A. Piccirilli, B. Estrine, N. Patouillard, J. Guilbot and F. Jérôme, *Green Chem.*, 2017, **19**, 4973–4989.

144 P. T. Anastas and R. H. Crabtree, Eds., *Handbook of green chemistry*, John Wiley and Sons, Place of publication not identified, 2010.

145 P.-M. Jacob, P. Yamin, C. Perez-Storey, M. Hopgood and A. A. Lapkin, *Green Chem.*, 2017, **19**, 140–152.

146 A. A. Lapkin, P. K. Heer, P.-M. Jacob, M. Hutchby, W. Cunningham, S. D. Bull and M. G. Davidson, *Faraday Discuss.*, 2017, **202**, 483–496.

147 J. Marie Weber, P. Lió and A. A. Lapkin, *React. Chem. Eng.*, 2019, **4**, 1969–1981.

148 M. Fialkowski, K. J. M. Bishop, V. A. Chubukov, C. J. Campbell and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2005, **44**, 7263–7269.

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

150 P.-M. Jacob and A. Lapkin, *React. Chem. Eng.*, 2018, **3**, 102–118.

151 C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. Wei, B. Baytekin and B. A. Grzybowski, *Angew. Chem. Int. Ed.*, 2012, **51**, 7922–7927.

152 A. Voll and W. Marquardt, *AIChE J.*, 2012, **58**, 1788–1801.

153 S. E. Jørgensen and Y. M. Svirezhev, *Towards a Thermodynamic Theory for Ecological Systems*, Elsevier, 2004.

154 A. Carvalho, H. A. Matos and R. Gani, *Comput. Chem. Eng.*, 2013, **50**, 8–27.

155 D. Krajnc and P. Glavic, *Clean Technol. Environ. Policy*, 2003, **5**, 279–288.

156 J. T. Marshall and A. Mundt, *Process Saf. Prog.*, 1995, **14**, 163–170.

157 G. Koller, U. Fischer and K. Hungerbühler, *Ind. Eng. Chem. Res.*, 2000, **39**, 960–972.
158 H. Sugiyama, U. Fischer, K. Hungerbühler and M. Hirao, *AIChE J.*, 2008, **54**, 1037–1053.
159 M. R. Othman, J.-U. Repke, G. Wozny and Y. Huang, *Ind. Eng. Chem. Res.*, 2010, **49**, 7870–7881.
160 D. Tanzil and B. R. Beloff, *Environ. Qual. Manag.*, 2006, **15**, 41–56.
161 E. Heinzle, D. Weirich, F. Brogli, V. H. Hoffmann, G. Koller, M. A. Verduyn and K. Hungerbühler, *Ind. Eng. Chem. Res.*, 1998, **37**, 3395–3407.
162 A. Azapagic and S. Perdan, *Process Saf. Environ. Prot.*, 2000, **78**, 243–261.
163 S. Ulgiati and M. T. Brown, *Ecol. Model.*, 1998, **108**, 23–36.
164 Y. Zhang, A. Baral and B. R. Bakshi, *Environ. Sci. Technol.*, 2010, **44**, 2624–2631.
165 F. Roschangar, J. Li, Y. Zhou, W. Aelterman, A. Borovika, J. Colberg, D. P. Dickson, F. Gallou, J. D. Hayler, S. G. Koenig, M. E. Kopach, B. Kosjek, D. K. Leahy, E. O'Brien, A. G. Smith, M. Henry, J. Cook and R. A. Sheldon, *ACS Sustain. Chem. Eng.*, 2022, **10**, 5148–5162.
166 dpicampaigns, Take Action for the Sustainable Development Goals, https://www.un.org/sustainabledevelopment/sustainable-development-goals/, (accessed October 2, 2022).
167 L. Jacquemin, P.-Y. Pontalier and C. Sablayrolles, *Int. J. Life Cycle Assess.*, 2012, **17**, 1028–1041.
168 European Commission. Directorate General for Research and Innovation. and Research Council of Norway., *Future technology for prosperity: Horizon scanning by Europe's technology leaders.*, Publications Office, LU, 2019.
169 E. Wyrzykowska, A. Mikolajczyk, I. Lynch, N. Jeliazkova, N. Kochev, H. Sarimveis, P. Doganis, P. Karatzas, A. Afantitis, G. Melagraki, A. Serra, D. Greco, J. Subbotina, V. Lobaskin, M. A. Bañares, E. Valsami-Jones, K. Jagiello and T. Puzyn, *Nat. Nanotechnol.*, 2022, **17**, 924–932.
170 Roadmaps | EMMC | The European Materials Modelling Council, https://emmc.eu/emmc-roadmaps/, (accessed December 1, 2022).
171 T. Puzyn, A. Gajewicz, D. Leszczynska and J. Leszczynski, in *Recent Advances in QSAR Studies: Methods and Applications*, eds. T. Puzyn, J. Leszczynski and M. T. Cronin, Springer Netherlands, Dordrecht, 2010, pp. 383–409.
172 B. Fadeel, L. Farcal, B. Hardy, S. Vázquez-Campos, D. Hristozov, A. Marcomini, I. Lynch, E. Valsami-Jones, H. Alenius and K. Savolainen, *Nat. Nanotechnol.*, 2018, **13**, 537–543.
173 D. A. Winkler, *Small*, 2020, **16**, 2001883.
174 A. Rybińska-Fryca, A. Mikolajczyk and T. Puzyn, *Nanoscale*, 2020, **12**, 20669–20676.
175 A. Mikolajczyk, A. Gajewicz, B. Rasulev, N. Schaeublin, E. Maurer-Gardner, S. Hussain, J. Leszczynski and T. Puzyn, *Chem. Mater.*, 2015, **27**, 2400–2407.
176 E. Wyrzykowska, A. Rybińska-Fryca, A. Sosnowska and T. Puzyn, *Green Chem.*, 2019, **21**, 1965–1973.
177 A. Mikolajczyk, A. Gajewicz, E. Mulkiewicz, B. Rasulev, M. Marchelek, M. Diak, S. Hirano, A. Zaleska-Medynska and T. Puzyn, *Environ. Sci. Nano*, 2018, **5**, 1150–1160.
178 A. Sosnowska, M. Barycki, M. Zaborowska, A. Rybinska and T. Puzyn, *Green Chem.*, 2014, **16**, 4749–4757.
179 X. Yan, A. Sedykh, W. Wang, X. Zhao, B. Yan and H. Zhu, *Nanoscale*, 2019, **11**, 8352–8362.
180 S. Kar, A. Gajewicz, T. Puzyn, K. Roy and J. Leszczynski, *Ecotoxicol. Environ. Saf.*, 2014, **107**, 162–169.
181 T. Puzyn, B. Rasulev, A. Gajewicz, X. Hu, T. P. Dasari, A. Michalkova, H.-M. Hwang, A. Toropov, D. Leszczynska and J. Leszczynski, *Nat. Nanotechnol.*, 2011, **6**, 175–178.
182 T. Puzyn, N. Suzuki, M. Haranczyk and J. Rak, *J. Chem. Inf. Model.*, 2008, **48**, 1174–1180.
183 N. Sizochenko, B. Rasulev, A. Gajewicz, V. Kuz'min, T. Puzyn and J. Leszczynski, *Nanoscale*, 2014, **6**, 13986–13993.
184 E. Wyrzykowska, A. Mikolajczyk, C. Sikorska and T. Puzyn, *Nanotechnology*, 2016, **27**, 445702.