Check for updates

Automation and computer-assisted planning for chemical synthesis

Yuning Shen[1,4], Julia E. Borowski[2,4], Melissa A. Hardy[3,4], Richmond Sarpong[3]✉, Abigail G. Doyle[2]✉ and Tim Cernak[1]✉

Abstract | The molecules of today — the medicines that cure diseases, the agrochemicals that protect our crops, the materials that make life convenient — are becoming increasingly sophisticated thanks to advancements in chemical synthesis. As tools for synthesis improve, molecular architects can be bold and creative in the way they design and produce molecules. Several emerging tools at the interface of chemical synthesis and data science have come to the forefront in recent years, including algorithms for retrosynthesis and reaction prediction, and robotics for autonomous or high-throughput synthesis. This Primer covers recent additions to the toolbox of the data-savvy organic chemist. There is a new movement in retrosynthetic logic, predictive models of reactivity and chemistry automata, with considerable recent engagement from contributors in diverse fields. The promise of chemical synthesis in the information age is to improve the quality of the molecules of tomorrow through data-harnessing and automation. This Primer is written for organic chemists and data scientists looking to understand the software, hardware, data sets and tactics that are commonly used as well as the capabilities and limitations of the field. The Primer is split into three main components covering retrosynthetic logic, reaction prediction and automated synthesis. The former of these topics is about distilling the strategy of multistep synthesis to a logic that can be taught to a computer. The section on reaction prediction details modern tools and models for developing reaction conditions, catalysts and even new transformations based on information-rich data sets and statistical tools such as machine learning. Finally, we cover recent advances in the use of liquid handling robotics and autonomous systems that can physically perform experiments in the chemistry laboratory.

[1]*Department of Medicinal Chemistry, University of Michigan, Ann Arbor, MI, USA.*

[2]*Department of Chemistry, Princeton University, Princeton, NJ, USA.*

[3]*Department of Chemistry, University of California, Berkeley, CA, USA.*

[4]*These authors contributed equally: Yuning Shen, Julia E. Borowski, Melissa A. Hardy.*

✉*e-mail: rsarpong@berkeley.edu; agdoyle@princeton.edu; tcernak@med.umich.edu*

https://doi.org/10.1038/s43586-021-00022-5

In 1948, Claude Shannon reported that information can be encoded and transferred as ones and zeros[1], which launched the field of information theory and set the stage for the merger of data science and chemical synthesis. Within two decades, information theory had become ingrained in organic chemistry, as evidenced by the translation of retrosynthetic logic to a computer code[2]. By this time, linear free energy relationships such as the Hammett[3] and Brønsted[4] equations were well established, and the commercialization of computers enabled predictive reaction calculations to be performed with increasing sophistication. In 1966, the automation of chemical reactions with a computer-driven robot enabled a high-throughput synthesis of peptides[5,6]. It would seem that the information age of chemical synthesis enjoyed its pinnacle decades ago. Yet, in 2021, the retrosynthesis of complex molecules, the high-fidelity prediction of reaction outcomes and the automation of chemical reactions very much remain developing areas of research. Computational power has accelerated, automation hardware and software have advanced and algorithms for chemical synthesis have been refined, paving the way for an exciting future where molecules can be automatically designed, synthesized and tested. Building upon seven decades of discovery since the seminal report on information theory, there is still ample basic science to develop to realize the information age of chemical synthesis.

The data available to the synthetic chemist to tackle research problems have increased significantly in the past decade. In order to apply data science in chemical synthesis, computers must understand the information encoded by molecular structures. Representation of molecules was non-trivial in the early days of the field, and to begin to address this long-standing challenge, programs such as ChemDraw emerged to help communicate chemical structures to the computer[7]. Today, molecules are commonly input and processed as string-based representations including the simplified molecular input line entry system (SMILES)

Linear free energy relationships
Linear relationships between the free energy of activation or free energy change of a reaction induced by a substituent of a molecule and a parameter that describes the electronic or steric properties of that substituent. Linear free energy relationships are a subset of structure–function (or structure–activity) relationships.

Simplified molecular input line entry system
(SMILES). A string notation to represent chemical structures that can be generated from a two-dimensional or three-dimensional graph notation. Notably, the same molecule can sometimes be represented by multiple different SMILES codes depending on the drawing that was input. These notations are human understandable and variable in length.

string[8] (FIG. 1). The SMILES string is a concise and accessible format that enables the rapid transmission of chemical structures into and out of the computer. Nonetheless, the SMILES notation has limitations — for example, the output can be dependent upon how the input structure is drawn, which can be amended by canonicalization, transition metal complexes are not well described by SMILES and, although relative stereochemistry is encoded, absolute stereochemistry can be lost. Other common string-based inputs include the International Chemical Identifier (InChI) and the SMILES arbitrary target specification (SMARTS), which includes connectivity and stereochemical information while allowing consideration of generic R-groups, and provides a way to efficiently store chemical structure information and interact with software.

This Primer aims to introduce non-experts to the current state of the field of chemical information theory, capturing both experimental and theoretical aspects, as well as automation software and hardware currently in use. The field entered a renaissance about 5 years ago, and continues to evolve rapidly. This Primer groups the merger of chemical synthesis with information theory into the three themes of retrosynthetic logic, reaction prediction and automated synthesis (FIG. 1). These three themes capture a majority of the recent activity of the field, and each theme is discussed in terms of techniques for experimentation, the types of results obtained and modern applications, followed by some overall considerations and the outlook for the field. For additional details, the reader is referred to reviews on retrosynthetic software[9–15], reaction prediction[16–19] and automated synthesis[20–25]. This Primer discusses these three themes independently in the Experimentation, Results and Applications sections, and collectively as we discuss data, limitations and the outlook for the field.

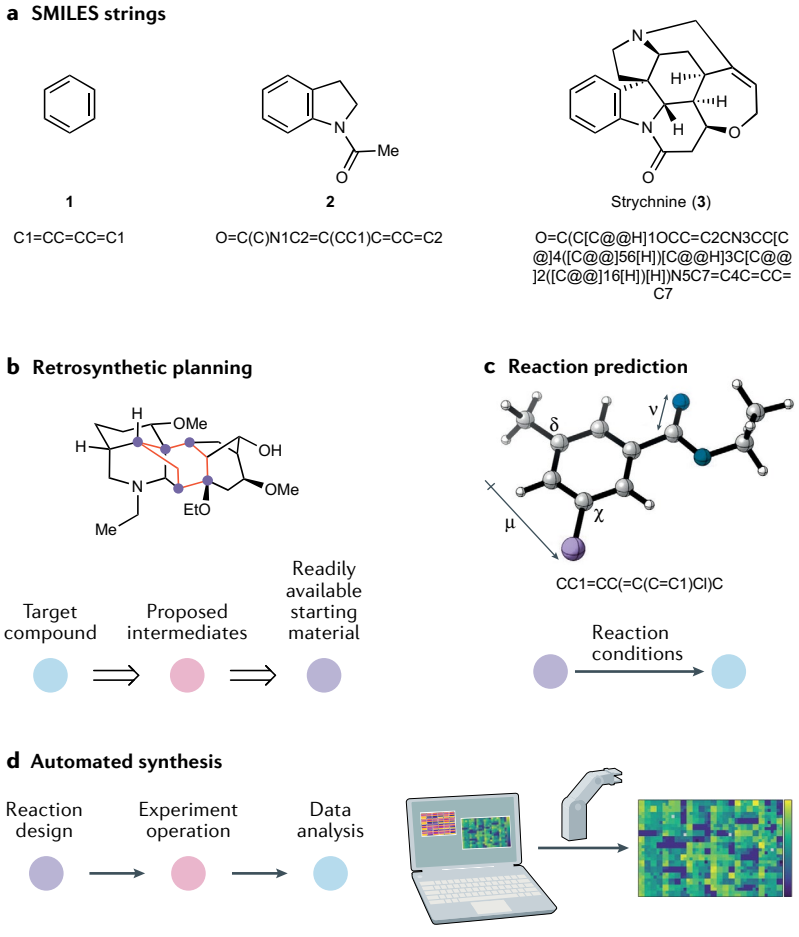


Fig. 1 | **General tools underlying chemical synthesis with information theory.** Molecular representation such as simplified molecular input line entry system (SMILES) strings (panel **a**), retrosynthetic planning (panel **b**), reaction prediction (panel **c**) and automated synthesis (panel **d**) as discussed in this Primer.

Experimentation

Chemical synthesis in the information age involves an understanding of organic chemistry and data science experimentation techniques. In this section, we cover the software, hardware and data formats needed to perform research in each thematic area of computer-assisted synthesis — retrosynthetic logic, reaction prediction and automated synthesis.

Retrosynthetic logic

We begin with a focus on retrosynthetic logic, and consider both the manner by which computers disconnect a molecule and the strategic value of a given transform. Ultimately, our goal is to introduce a practising synthetic chemist to the considerations, capabilities and limitations of modern retrosynthetic computer programs. This area of research has a long history[2,26–28], and today enjoys a resurgence that has been enabled in part by the availability of digitized reaction data from sources such as Reaxys and SciFinder.

Traditional logic. Retrosynthesis was defined by Corey in the 1960s to describe the iterative process of reducing a complex target molecule to a simple precursor by breaking bonds[2], to arrive at a compound that is readily available. This recursive analysis revolutionized complex molecule synthesis and introduced rules that were codified in an express attempt to develop computer-assisted synthesis[29]. After all, understanding the logic of retrosynthesis is necessary to program a computer to automate the process.

The game of chess is often employed as a metaphor for organic synthesis[9,30]. Similar to chess, several attempts have been made to apply computer-assisted logic to organic synthesis. However, unlike games such as chess, one reason chemical synthesis cannot be easily solved is the non-trivial evaluation of whether a given retrosynthetic transformation will make the route more efficient overall, both strategically and experimentally. Therefore, rather than concrete rules for retrosynthesis, the heuristics or guidelines that were first codified in Logic and Heuristics Applied to Synthetic Analysis (LHASA) are favoured[31]. Also unlike chess, the value of each move, and the overall objective of the route, may be open to interpretation through a chemical lens. In any event, chess was a logical proving ground for the advancement of artificial intelligence with the Deep Blue chess program[32]. More recently, the more sophisticated game of Go has succumbed to computational planning, with the AlphaGo program able to beat the

International Chemical Identifier
(InChI). A fixed-length, 27-character line notation that is designed to allow easy searches of chemical compounds. These are derived from the full length that encodes layers of information about a given molecular structure including connectivity, charge, stereochemistry and atomic isotopes. These notations are not human understandable.

SMILES arbitrary target specification
(SMARTS). An extension of the simplified molecular input line entry system (SMILES) notations that allows for the specification of generic atoms and bonds to allow for substructures for searching databases.

Reaction rules
Descriptions of chemical transforms that can be applied in a retrosynthetic module. These encode the substructures of the products and starting materials for a given synthetic step, and also include additional layers to express the scope and limitations of when the transform can be applied.

Reaction templates
Descriptions of chemical transforms that include the substructures of the reactants and products and highlight structural changes. These contain somewhat less context than a reaction rule and often require additional strategy to select which of the numerous templates to apply in a retrosynthetic module to minimize computational cost.

Sequence-to-sequence
A family of machine learning algorithms developed for natural language processing (language translation, image captioning and so on) that relies on recurrent neural networks to transform one sequence into another sequence.

Transformer
An algorithm developed for natural language processing (language translation, image captioning and so on). This algorithm does not rely on recurrent neural networks and can process data in any order, thus allowing for reduced training times .

world's best Go players[33]. The logical and creative nature of organic synthesis has made it attractive as a higher computational bar to clear.

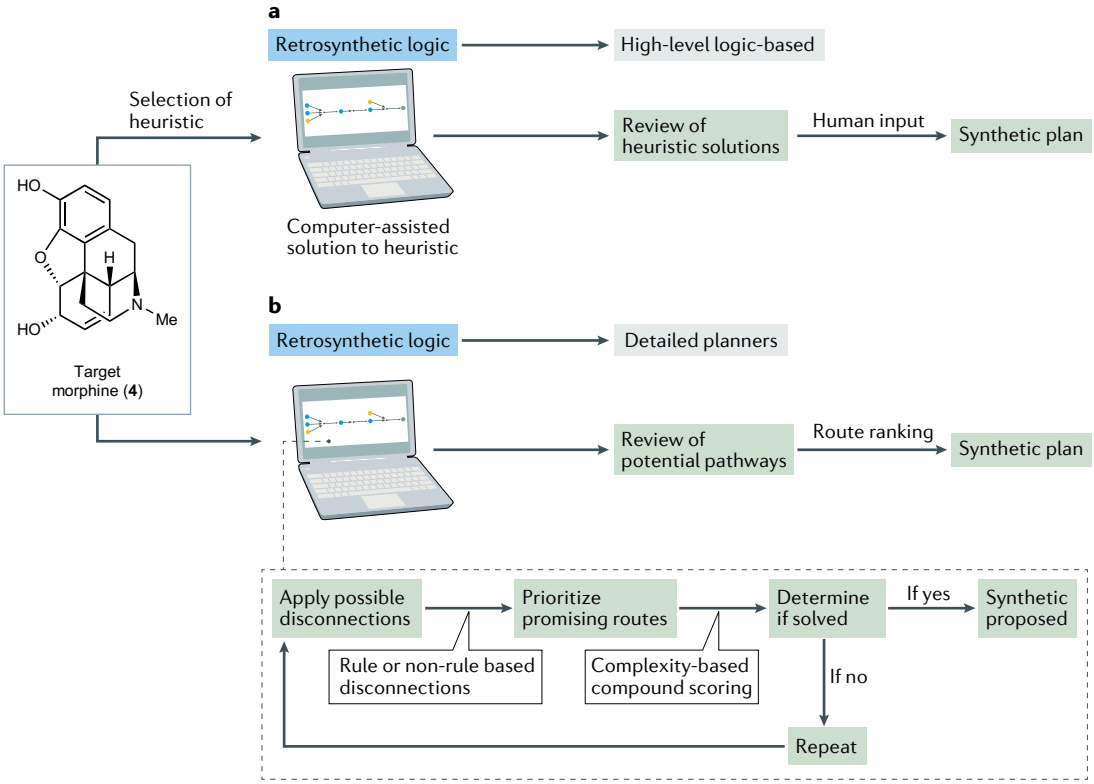
High-level logic-based programs. High-level logic-based retrosynthetic programs (FIG. 2a) are designed to apply a particular heuristic to a given compound and are intended to be used with significant input from an experienced user. First, this requires the identification of a specific heuristic followed by the application of an algorithm that can modify the molecular representation and present routes or key disconnections. For example, it can be advantageous for a synthetic route towards a chiral, enantio-enriched molecule to start from readily available enantio-enriched starting materials that obviate the need to install stereocentres using asymmetrical reactions. To this end, many programs have been developed to map readily available enantio-enriched starting materials to a particular target[31,34–37]. For any high-level logic-based program, the algorithms used are defined by the heuristic to be applied and can be highly specialized. For example, in the case of starting material-based programs, the software is based on similarity recognition whereas application of bond-network analysis requires software to analyse the molecule as a graph. These programs present the solution to the heuristic but typically require significant input from the chemist to arrive at an overall synthetic route.

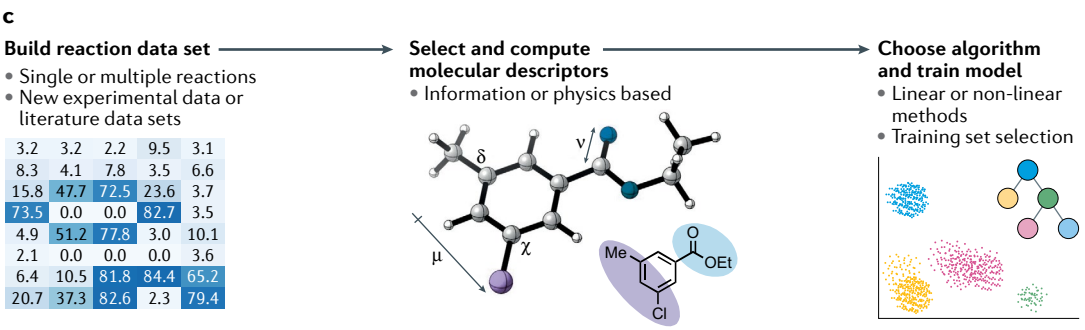
Detailed retrosynthetic planners. An orthogonal strategy is the development of synthetic planners (FIG. 2b). Instead of identifying strategic elements, these programs propose discrete intermediate compounds that can be used as stepping stones to reach the target molecule. This process requires a module that disconnects the target compound into potential precursors and reports a proposed structure to the user[14]. Internally, these modules typically apply either rule-based or rule-free methods to propose possible transforms.

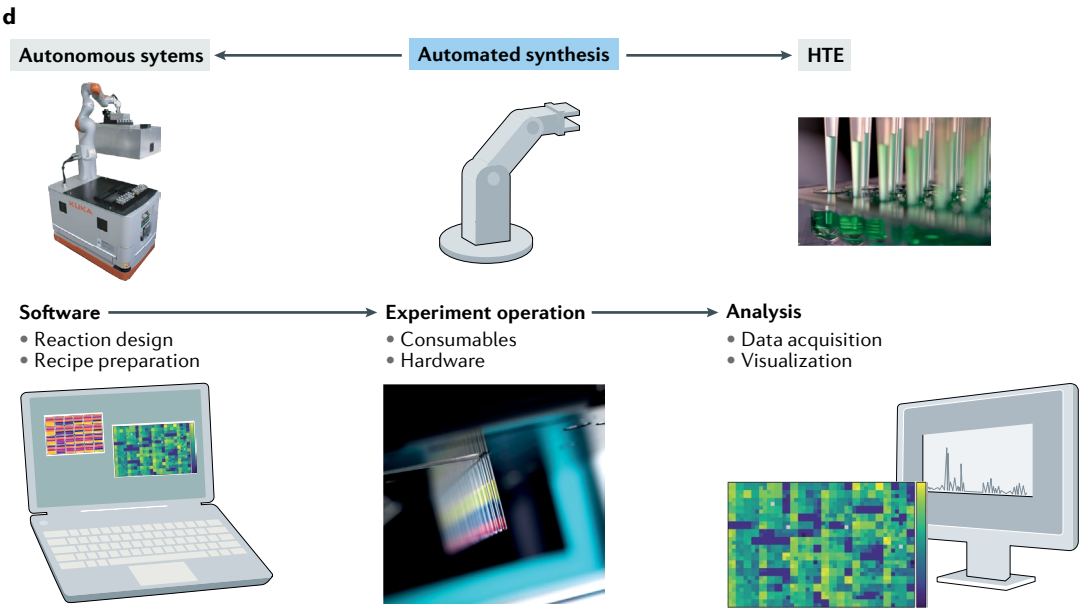
Rule-based methods are conceptually akin to the process of an organic chemist selecting a known reaction type to apply to a particular synthetic target. It follows that one way to build the necessary reaction rules is to have expert organic chemists encode a transform defined by the substructures of the products and reactants as well as necessary molecular context, such as functional group compatibility, stereoselectivity and so on. This approach has been well implemented in state-of-the-art detailed synthesis planners[38], but building a library of expert-coded rules is laborious and inherently dependent on the expertise of the coders. As a result, a growing area of research is the automatic generation of the reaction rules from accessible reaction databases[12]. In this process, a library of reaction templates is typically extracted from each reaction in the database. In a rule-based approach, these templates are then clustered and processed with additional molecular context to automatically generate explicit reaction rules[39,40]. Other approaches apply templates directly to the target where filters, such as similarity-based neural networks, are often used to apply only a chemically relevant subset of the template library to reduce the required computational power[41–44]. Although these methods are common in state-of-the-art detailed synthesis planners, there is significant computational cost in extracting rule or template libraries; additionally, such libraries inherently demote or exclude rare reactions with sparse literature examples.

By contrast, rule-free methods bypass the need to build a library of reaction rules by directly mapping products to potential starting materials. Early examples in this field recognized that by representing molecules as text (for example, SMILES strings), retrosynthesis could be treated as a natural language processing problem in which the target molecule (one language) is translated to reactants (another language)[45,46]. Implementation of sequence-to-sequence neural networks was able to effectively generate single-step retrosyntheses[47]. One inherent challenge to this approach is that not all generated SMILES strings lead to a valid chemical structure[48]. Subsequent research built upon this work by applying the newly developed transformer neural networks that have been able to more accurately produce valid SMILES strings[49–51]. Rule-free approaches that represent the molecules as information-rich subgraphs rather than as a string of text have also been developed[52,53]. Compared with rule-based approaches, rule-free methods are more generalizable and have a lower associated computational cost; however, they lack comparative interpretability. Today, rule-based or template-based methods are more common in detailed retrosynthetic planners but multistep rule-free planners are emerging[49] and remain an important area of development.

As one considers multistep syntheses, applying these approaches to generate simplifying transforms will lead to an exponential increase in the number of precursor compounds that must be analysed[30]. A program must rank the most promising disconnections to navigate the synthetic tree to prioritize among the many possible precursor compounds; this navigation strategy will significantly impact the output synthetic route. One example is the use of a Monte Carlo tree search with reinforcement learning to balance exploration against exploitation in navigating the synthesis search space[54,55]. Although promising disconnections typically reduce the complexity of the target molecule, developing an objective measure for ranking retrosynthetic transformations can be difficult as the most concise syntheses are concerned with target-relevant synthetic complexity, which cannot be locally determined. There is also no guarantee that a synthesis that uniformly increases target-relevant complexity can be achieved for a given target with currently available methods or that the application of a particular transform will be synthetically feasible. For this reason, many measures of structural complexity[56] have been applied[30]; however, recent efforts have focused towards implementing measures of synthetic complexity[57,58]. Other programs have eliminated this issue by incorporating human interaction at each step to determine the feasibility of a proposed strategy, albeit in a less automated way[37,59]. Considerations for generating and prioritizing disconnections are vital when building, evaluating or applying a retrosynthetic program.


a
Retrosynthetic logic
High-level logic-based
Selection of heuristic
Review of heuristic solutions
Human input
Synthetic plan
Computer-assisted solution to heuristic
HO
O
H
N
Me
HO
Target morphine (4)
b
Retrosynthetic logic
Detailed planners
Review of potential pathways
Route ranking
Synthetic plan
Apply possible disconnections
Prioritize promising routes
Determine if solved
If yes
Synthetic proposed
Rule or non-rule based disconnections
Complexity-based compound scoring
If no
Repeat


c
Build reaction data set
• Single or multiple reactions
• New experimental data or literature data sets
3.2 3.2 2.2 9.5 3.1
8.3 4.1 7.8 3.5 6.6
15.8 47.7 72.5 23.6 3.7
73.5 0.0 0.0 82.7 3.5
4.9 51.2 77.8 3.0 10.1
2.1 0.0 0.0 0.0 3.6
6.4 10.5 81.8 84.4 65.2
20.7 37.3 82.6 2.3 79.4
Select and compute molecular descriptors
• Information or physics based
δ
ν
μ
χ
Me
O
OEt
Cl
Choose algorithm and train model
• Linear or non-linear methods
• Training set selection


d
Autonomous sytems
Automated synthesis
HTE
Software
• Reaction design
• Recipe preparation
Experiment operation
• Consumables
• Hardware
Analysis
• Data acquisition
• Visualization

◀ Fig. 2 | **Retrosynthetic planners, reaction prediction and automated synthesis workflows. a** | Workflow for high-level logic-based retrosynthetic planners showcasing significant human input in the planning stages. **b** | General workflow employed iteratively for detailed retrosynthetic planners. **c** | Building a statistical model for a single-step reaction. **d** | Automated syntheses and main components in an automation platform. Automated synthesis generally falls into two categories: autonomous systems or high-throughput experimentation (HTE). Autonomous system example (top-left image) in panel **d** reprinted from REF.[154], Springer Nature Limited. Multichannel pipette dispensing liquid (top-right image) in panel **d**, credit to red_moon_rise/Getty Images Plus. Image of automated liquid handler (bottom middle) in panel **d**, image courtesy of Carrey Rooks (SPT Labtech).

Reaction prediction

To construct a molecule, one must know not only the retrosynthetic sequence of steps but also the reaction conditions to execute each step. Identifying effective reaction conditions experimentally can consume significant time and resources because reaction space is highly dimensional. Therefore, chemists have sought tools for predicting the optimal conditions and reagents to ensure that a reaction yields the desired product. In this subsection, we discuss modern reaction prediction (FIG. 2c), with a focus on its dependence on the availability of high-quality data and meaningful descriptors that parameterize the reaction components and translate chemical knowledge. These factors influence the algorithm choice required for successful modelling. Our goal is to introduce the capabilities and limitations of current reaction prediction tactics. The methods covered model experimental reaction data for the prediction of yields, selectivities and reaction conditions, although additional work in the field has been done for the prediction of reaction products[60,61] in the development of computational workflows for virtual reaction screening protocols[62–65].

Data sets for reaction prediction. Data sets can be built by data mining digital records of pre-existing experiments or from new experimental data. Obtaining data from sources such as Reaxys or SciFinder, patents, published chemical literature or proprietary databases enables the creation of large data sets, on the order of hundreds to millions of data points, without the need for experimental resources. However, the quality of these data can be inconsistent between sources or limited by omission of critical reaction metadata such as temperature or solvent identity[14]. The data available in the literature are also often biased away from perceived 'negative' reactions, leaving out data points with low yields or selectivities, which can be critical for accurate modelling[66]. To aid in the extraction of data from publications, platforms that can translate text from experimental protocols into tabulated data have emerged[67–69]. Synthetic chemistry experimental protocols are generally written in loosely structured prose, and progress in text extraction has centred on harmonization of experimental instructions. For instance, a human can recognize that the phrase 'the reaction was quenched by the addition of 100 ml of water' is equivalent to 'water (100 ml) was added as a quenchant', but a computer must be trained to recognize the equivalency of these statements.

Newly generated data sets allow for internal consistency and enable exploration of new chemical space not yet represented in the literature. Experimental researchers can usually access data sets in the scale of tens of data points using standard screening protocols, but generation of larger data sets (hundreds or thousands of data points) requires the use of high-throughput experimental set-ups as discussed below (see Automation). When feasible, it is advantageous to design a data set that covers a broad area of chemical space. That is, to improve model performance, the data used to build the model should include data points that are representative of the range of possible parameter values to avoid overfitting to the training set or introducing biases[68]. Curation of a diverse data set can be facilitated by using a dimensionality reduction method, such as principal component analysis, in tandem with clustering algorithms, such as *K*-means, to group together data points that exist in a similar area of chemical space — that is, ones that have similar properties based on the descriptors used[70–72]. A subset of potential reactions can be chosen from these clusters for use in modelling. For commonly used substrates or catalysts, the development of representative subsets of molecules that provide good coverage of chemical space, referred to as universal training sets, using design of experiments principles is an active area of interest[73,74]. For smaller data sets or novel methodologies, however, such a purposeful design of the data set may not be practical.

Descriptor selection. It is essential to consider how many dimensions, or reaction variables, are being modelled and how they can be most effectively described when building a model for reaction prediction. Common variables for reaction prediction include the substrate, solvent, temperature, additive, base and ligand. Modelling can be performed on a single reaction variable, for example, a study of ligand effects[75,76], or can be performed on several variables simultaneously to achieve more comprehensive predictions of 'over the arrow' conditions[60,64,73,77,78]. These variables can be made machine-readable by one-hot encoding[79], where each category (for example, each ligand) is transformed into a binary variable (present or not present in a reaction)[80]. Molecular descriptors are employed to provide chemical and structural context when building a model.

Information-based descriptors provide a means of converting the two-dimensional structures of molecules into a machine-readable format. Many of these descriptors are rooted in the representation of molecules as molecular graphs, in which atoms are treated as nodes with bonds defined as the edges that connect them. A 'walk' through the molecular graph collects the identity and connectivity of each node and edge. This information is stored as matrices such that it can be used in machine learning algorithms[81]. Other information-based descriptor sets have been developed that include chemical information about the atoms or functional groups present in a molecule, such as electronic or topological properties, and provide snapshots of local and global chemical environments. One commonly used type of descriptor for quantitative structure–activity relationship modelling is molecular fingerprints that provide substructures of a molecule and describe the neighbourhood

Monte Carlo tree search
An algorithm for navigating search trees in which search steps are selected randomly, without branching, until a solution has been found or a maximum depth is reached. Algorithms of this type have emerged as strategic in applications of sequential decision problems without clear heuristics.

Quantitative structure–activity relationship
A statistical modelling method used to relate molecular structure to biological and physico-chemical properties and predict these properties in new molecules.

Density functional theory
(DFT). A computational method for modelling the electronic structure of atoms and molecules using quantum mechanics. In synthetic chemistry, density functional theory is used to compute and study molecular structures and their corresponding energies that cannot be obtained through experimental methods.

Molecular mechanics
A computational method for modelling molecular structure using classical mechanics. Bonds are treated as springs from which a potential energy can be determined. Molecular mechanics is a less computationally expensive method relative to density functional theory.

HOMO–LUMO energies
(Highest-occupied molecular orbital–lowest-unoccupied molecular orbital energies). These values correspond to the energetics of the molecular orbitals that are most involved in bond-making and bond-breaking processes, commonly referred to as the frontier molecular orbitals.

Sterimol parameters
Three steric parameters — B_1, B_5 and L — for molecular substituents determined from three-dimensional structures. B_1 and B_5 represent the minimum and maximum widths, respectively, of the molecule perpendicular to the primary bond axis. L is the total length of the substituent measured along the primary bond axis.

Buried volume
A steric parameter for ligands in transition metal complexes. The volume of a ligand, bonded to a metal at a fixed distance, enclosed by a sphere of a defined radius r. Provided as a percentage, representing the percentage of the sphere that is filled by a single bound ligand.

Conformers
(Also known as conformational isomers). Structures of a molecule that differ by the rotation of groups about one or more single bonds in the molecule. Conformers can interconvert without making or breaking bonds and will have different relative energies based on the presence of attractive or repulsive interactions.

in which certain atoms or functional groups reside[82,83]. In practice, software libraries such as RDKit or Mordred[84] generate tabulated values for a given descriptor set from a one-dimensional (string) or two-dimensional (structure) representation of a molecule. A developing strategy for representing molecules directly from two-dimensional structures is the use of neural networks to provide a molecular graph representation that can be directly fed into the model for reaction prediction[85–88]. These learned molecular representations remove the step of choosing and obtaining a set of descriptors but may have difficulty accounting for information about a molecule's three-dimensional structure.

Experimental or computational methods — including density functional theory (DFT) and molecular mechanics — are used to formulate physically meaningful descriptors, enabling access to a diverse array of electronic and steric descriptors and considering the molecule in three-dimensional space. These descriptors include HOMO–LUMO energies, torsion angles, Sterimol parameters[89,90], buried volume[91,92] and nuclear magnetic resonance (NMR) shifts. When using computational methods for acquiring these descriptors, a conformational search is often performed to find one or more conformers that may be active in a reaction and need to be accounted for in parameter acquisition[89]. Including individual descriptors for different conformers (for example, highest and lowest energy conformers) may be necessary to accurately reflect a flexible molecule's three-dimensional properties, as it can be challenging to determine which conformer represents the reactive species[74].

Descriptors can be thought of as information-based or physics-based. Physics-based parameters, in concert with experimentally derived parameters, have the benefit of interpretability, allowing a chemist to intuit additional physical information from the model[93]. However, acquisition can be a time and resource-intensive process, owing to the need for computationally expensive structure optimizations or experimental resources, limiting throughput. These physical parameters are not general to all molecular scaffolds and reaction components, thus rendering the creation of a comprehensive descriptor set challenging[94]. For example, describing a monodentate phosphine on the basis of buried volume does not provide a means of direct comparison with bidentate phosphine structures. In this regard, information-based descriptors provide a more general and easily accessible means of storing chemical information, as they can be accessed using computer software from string or two-dimensional structural representations. However, this generalizability comes at the cost of interpretability and falls short for organometallic complexes that are not accurately translated into the string-based representations used to generate these descriptors.

Algorithm selection. The machine learning algorithms applied in synthetic chemistry can be categorized generally as either linear or non-linear. When possible, multiple algorithms will be tested and tuned for a given data set before the best performer, as determined by model validation (described in the Reaction prediction section of Results), is tested on new data.

Linear and multivariate linear regression have been used as tools for probing linear free energy relationships and modelling reaction selectivities[16,95]. This method can be used on smaller data sets (magnitude of tens of data points), making it suitable for use with traditional experimental screening methods. Although stepwise methods for parameter selection can provide meaningful linear models, regression algorithms such as ridge regression, least absolute shrinkage and selection operator (LASSO) regression and elastic net methods may be necessary to improve model performance and to prevent overfitting.

Non-linear methods have been popular strategies for modelling using larger data sets. Many of these machine learning algorithms have been previously implemented on non-chemical data sets where large quantities of data are easily available, such as photographs or text. Non-linear algorithms that have been tested or applied in synthetic chemistry include ensemble random forest, *k*-nearest neighbours, support vector machines and neural networks[77,93,96]. Using a non-linear algorithm can be beneficial for multidimensional reaction predictions involving several reaction variables, such as the catalyst, solvent, additive and temperature, whereas linear methods have commonly been used for modelling data sets with a focus on a single reaction condition variable for one class of reaction.

Automation

A long-term vision of synthesis is to prepare samples of complex molecules automatically. Retrosynthetic logic and reaction prediction algorithms would provide the recipe, which would then be translated into reality by an automated hardware platform. Much in the way the automated synthesis of peptides[6,97,98] and oligonucleotides[99] is routine today, we envision a future where complex pharmaceuticals and natural products can be automatically produced. Although much effort has been invested in broadening the menu of automatable reactions and synthetic targets, automated synthesis of diverse products using an assortment of reaction types is in its infancy. Automating complex molecule synthesis will augment human creativity in the design of new functional molecules such as medicines[22,100], materials[101] and energy sources[102]. Automated synthesis generally falls into one of two broad objectives: to either increase reaction throughput or increase user autonomy (FIG. 2d). The former objective is commonly referred to as high-throughput experimentation (HTE), an area that has been heavily developed in recent years. HTE allows rapid, efficient, miniaturized and systematic generation of reaction data points, which can be used to inform reaction prediction. The latter objective aims to automate as much of the synthetic experimentation process as possible, such that little to no user input is required to synthesize molecules once the experiment is started. Such autonomous systems have the potential to realize the multistep recipes of a retrosynthetic algorithm, or invent new reactions. In this section, we review the hardware, software, consumables and reaction types for automated synthesis under the umbrellas of HTE or

High-throughput experimentation (HTE). A technique used for screening chemical experiments, typically in a miniaturized format. Common formats for HTE include 24-well, 96-well and 384-well arrays, whereas ultraHTE refers to arrays of 1,536 experiments or more.

autonomous systems. We cover both robotic and manual tactics for performing HTE and focus on label-free approaches, excluding reactions on resin beads[103,104], on DNA[105] or on other supporting media[106].

High-throughput automated synthesis systems. When setting objectives for an HTE campaign, a first analysis should consider the desired reaction throughput, reaction type and engineering requirements, such as heating or cooling, as well as budget. In recent years, HTE has become highly accessible in a manual format enabling multiplexing of dozens of reactions at a time in 24-well or 96-well reactors[107,108]. Meanwhile, nanoscale ultraHTE has made it possible to run thousands of reactions in a single campaign; but this process requires specialized equipment[100,109–111]. Homogeneous reactions performed at ambient temperature in low-volatility solvents are the easiest to automate. Exclusion of air requires an inert atmosphere enclosure for reaction set-up, which is typically accomplished in a glovebox. Heating reactions are generally straightforward in HTE, but operations such as cooling, stirring, photo-irradiation and gas-handling require additional engineering[20]. Solutions to nearly every common synthetic operation have been developed for HTE, each tackling a different engineering challenge. However, in many cases, budgets can be significantly reduced by engineering the chemistry to fit the automation platform, rather than engineering the platform to fit the chemistry.

With respect to the hardware and consumables, HTE systems are divided broadly into well-plate (miniaturized batch)[21] or microfluidic (miniaturized flow) formats[112], and typically operate on milligram to microgram reaction scales. In general, efforts are made to maintain reaction concentrations of ~0.1 M, such that a typical reaction volume is 1–100 µl (REF.[113]). Handling these small volumes can be done manually using single-channel or multichannel micropipettes, or using liquid handling robotics[109,114–118]. Among the commercial systems in use for HTE are Chemspeed SWING[115,119], Unchained Labs Junior[21], Tecan Freedom EVO[21], Labcyte Echo[111,114], Thermo Matrix[120] and SPT Labtech mosquito[93,109,120–122]. Low solvent vapour pressure facilitates liquid transfer, such that volatile solvents like dichloromethane and diethyl ether are rarely used, modestly volatile solvents such as tetrahydrofuran, acetonitrile or toluene are commonly used in microvial HTE and high boiling solvents such as dimethylsulfoxide, *N*-methyl-2-pyrrolidinone or ethylene glycol are preferable for handling minute liquid volumes (<3 µl). A typical workflow involves the preparation of concentrated reagent stock solutions or vigorously stirred suspensions in source vials, which are then multiplexed into the desired reaction array in a well plate or microfluidic reaction vessel. Prepared reagent stocks may be stored in an inert atmosphere, but long-term stock solution storage conditions should be informed by ageing studies. Although the automated dispensing of liquids is straightforward, solid chemicals may be dense, flocculent, waxy, crystalline or granular. This diversity of properties renders automated solid dispensing challenging[123]. One recent solution to the solid dispensing problem emerged by coating solid chemicals on the surface of tiny glass beads such that the chemical assumes the uniform bulk properties of the beads[124,125]. In flow systems, pumps are required and offer tunable pressure and flow rates. Homogeneous reactions are strongly preferred in flow to prevent system clogging[110,126,127]. HTE in 24-well, 96-well or 384-well plates and ultraHTE in 1,536-well plates can be executed in glass or plastic reaction vessels. Glass shell microvials with parylene-coated metal stir dowels sealed inside aluminium reaction blocks are routine for 24-well to 96-well reaction campaigns[100,113]. For 384-well to 1,536-well reaction campaigns, glass well plates are available but expensive[120], so consumable plastic polypropylene or cyclic octane copolymer plates are more common.

Diverse chemical reaction types have been studied, and reactivity trends become apparent when multiple reaction parameters are varied. Among reactions commonly used in HTE campaigns, metal-catalysed cross-coupling reactions have emerged as a popular choice, perhaps because many reaction variables must typically be explored in the development of such reactions[109,128–130]. Suzuki–Miyuara[110,131] and Buchwald–Hartwig[93,132] couplings have been studied considerably, owing to their prominent role in medicinal chemistry[133]. A benefit to miniaturized HTE is that precious catalysts, ligands or complex substrates are needed in only small amounts, so they can be easily conserved or used in more experiments than a traditional format. Miniaturization of photoredox reactions is now somewhat commonplace[120,134,135], and miniaturized electrochemistry in flow has been reported recently[136].

HTE in well plates and in flow requires careful selection of solvents to facilitate liquid transfers or avoid clogs in the reactor coils. Homogeneous reaction mixtures are desirable, but stirring capabilities can handle reaction suspensions and slurries. Solvent evaporation must be minimized when handling small liquid volumes. Variation in temperature is routine in flow systems, but atypical in a well-plate format. By contrast, the well-plate format is ideal for varying discrete reaction variables such as the catalyst, ligand, additive or substrate in parallel. Notably, automated tasks can also prevent the direct exposure of human operators to dangerous chemicals. Flow systems have been broadly used to incorporate hazardous reagents[137,138], and miniaturized HTE may also lower the risk of handling dangerous chemicals by virtue of the small reaction scale[139]. One area of considerable interest is the use of algorithms to optimize reaction variables. The SNOBFIT algorithm[131,140] has been popular for the optimization of continuous variables, such as temperature and concentration, whereas LabMate.ML merges random reaction selection with active learning and considers discrete variables such as the catalyst, solvent or additive[141].

Autonomous systems for chemical synthesis. At the highest level of sophistication, fully autonomous systems enable self-driven chemistry optimization and discovery[142]. In contrast to HTE, which has largely repurposed liquid handling equipment originally designed for biochemical high-throughput screening, many autonomous systems are built de novo. The custom design of robotic

systems for chemistry can lead to high efficiency and flexibility, but a significant investment in the build of hardware and software is typically required and few systems have been commercialized[126]. At some point, human input is required — for instance, to load reagents into the system — and a key question when designing an autonomous platform is the level to which tasks should be automated. For simple tasks, such as moving a reaction vessel from one instrument to the next or uncapping a reaction vial, automated engineering solutions are available but may significantly increase a project's complexity and budget.

With autonomous systems, the aim is to emulate traditional organic synthesis that takes place in the fume hood. In most cases, despite the objective to minimize human intervention, manual operation of some tasks, such as loading physical reagents into the system, is required[143]. Generalized systems include a reactor, a separator for reaction workup and an evaporator to remove volatiles while collecting the products[80,144]. An approach to miniaturize a chemical production factory that enables multistep synthesis[145] and a cartridge-based platform with preloaded reaction recipes have been recently reported[146]. Compound purification, for instance by column chromatography, can be included but purification remains challenging. A notable exception reminiscent of solid-supported synthesis is the use of *N*-methyliminodiacetic acid (MIDA) boronates, which selectively elute from silica gel when tetrahydrofuran is used as an eluent[147,148]. The apparatus in autonomous systems is typically based on customized flow equipment, or retrofitted synthesis glassware, connected using chemical-resistant polytetrafluoroethylene tubing. Pumps transfer reaction mixtures between each module. In-line analysis to collect high-performance liquid chromatography coupled to mass spectrometry (HPLC-MS)[100,149], NMR[144], infrared and UV spectroscopy, pH[80], biochemical assay[122] and other analytical data in real time have been implemented, creating closed-loop systems.

The automated formal synthesis of paclitaxel was reported more than a decade ago, highlighting the diversity of accessible reaction types on autonomous platforms[150]. Amide coupling, nucleophilic substitutions, cycloadditions, olefinations, reductive aminations and Grignard reactions have all been performed both as discrete steps and as part of multistep synthesis campaigns. The autonomous syntheses of diverse drug molecules, using popular reactions from the medicinal chemists' toolbox[133,151,152], have also been reported, in addition to the synthesis of natural products through tandem coupling and cycloaddition reactions[148]. To complement autonomous synthesis of target molecules, autonomous reaction discovery has emerged with robotic platforms discovering and optimizing new multicomponent coupling reactions or photocatalytic methods[153]. The engineering investment for these systems focuses on emulating traditional organic synthesis; therefore, most solvent and temperature regimes are supported. Phase homogeneity remains highly desirable to facilitate transfer of reaction mixtures as monophasic liquids.

Although fully autonomous systems typically perform one reaction at a time, the minimization of human intervention means that they can run continuously, often learning from the result of one reaction to inform the design of the next. In a recent example, a self-driving robot performed 688 sequential reactions over 8 days to discover a new photocatalytic method[154]. Related discoveries have been made by self-optimizing robots varying substrates and catalysts[80].

Software for automated synthesis. Software in an automated workflow falls into one of three categories: software to select target molecules and design a synthetic route; scheduling software to control the robotics during the experimental operation; and laboratory information management software to process experimental inputs and outputs including reaction metadata. Metadata may include information such as well location, reagent SMILES or molar mass inputs for a mass spectrometer, from one instrument to the next, and ultimately document and report the experimental outcome. For HTE, research has so far relied on commercial packages or custom macros in Excel. Numerous academic software packages specifically for HTE[110,155] and autonomous synthesis[156,157] (IBM RXN for Chemistry) are emerging. To pursue a fully fledged automation platform, a variation on all three software components is typically integrated. For example, a recent flow system determines a target molecule, self-optimizes the routes based on a retrosynthetic algorithm[129] and drives a fully autonomous system. This system merged discrete flow modules for reaction execution, reaction workup and volatile removal with a robotic arm to orchestrate the physical movement of the flow modules depending on the configuration recommended by the self-driving software. In another system, named Chemputer, reaction and scheduling inputs are encoded from published experimental protocols, and integration of in-line NMR and mass spectrometry analysis enables the autonomous discovery of multicomponent coupling reactions[144]. In reality, the human chemists' intuition remains quite necessary to evaluate the reagents, substrates, concentrations and other reaction parameters, but the autonomous system can remove the tedious experimental tasks.

Safety

Our discussion on retrosynthetic logic and reaction prediction focuses on the computational aspects. Hazards are clearly minimized in such research, although safety and risk assessment can be a key motivator in the selection of predicted routes from a retrosynthetic analysis. For instance, if one route suggests the use of a particularly explosive reagent, efforts to replace this reaction or at least the explosive reagent may be undertaken. With respect to automated synthesis, the miniaturized reactions used in HTE are often less hazardous to the operator given their small scale. By contrast, care must be taken to protect robotic equipment, which is generally not designed to handle corrosive reagents. The use of automation introduces a new hazard in the form of moving mechanical parts, but engineering controls such as inertion chambers physically separate the operator from the moving parts, and many modern robots are outfitted with sensors to stop moving if they encounter an obstruction, such as an

operator's hand. Autonomous systems further distance the operator from hazardous experimental operations. An exciting recent demonstration utilized a motorized robot on wheels to physically move around a laboratory and deliver samples from reaction stations to analytical stations with no human present. This sophisticated platform harnessed scheduling software reminiscent of that used in the automotive industry[154].

Results

The output of an experiment in retrosynthetic logic, reaction prediction or automated synthesis may be physical samples or simply data. In any event, the information density is often higher than a traditional experiment. Here, we cover what the output of information-rich synthesis experiments looks like, and how to interpret the results.

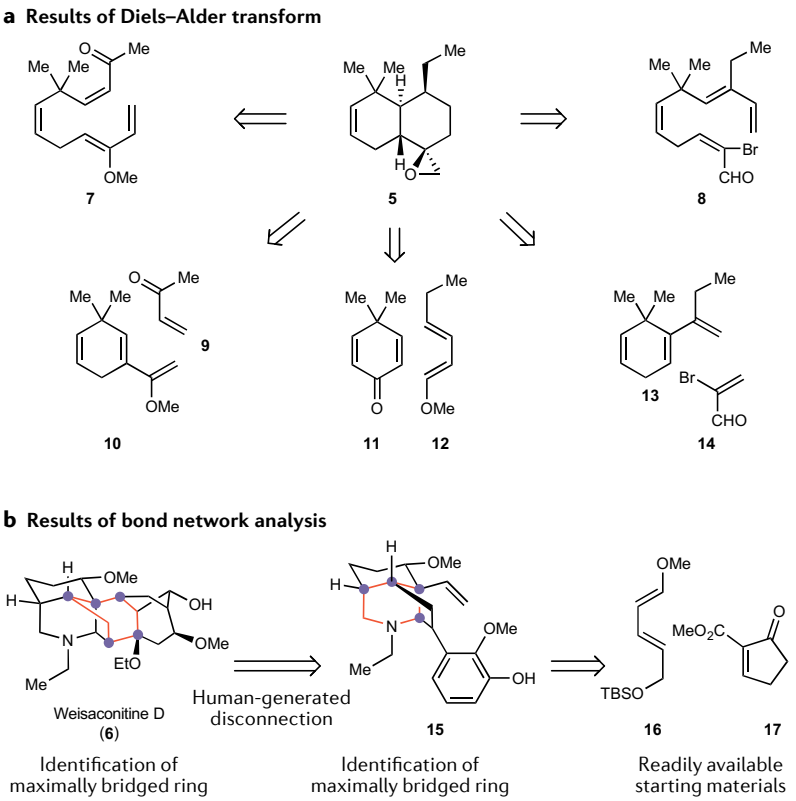


Fig. 3 | **Results from retrosynthetic planning programs. a** | Results of a Diels–Alder transform from application of Logic and Heuristics Applied to Synthetic Analysis (LHASA) to a fused bicyclic compound (**5**)[31]. **b** | Results of bond-network analysis by the program Maxbridge as applied to synthesis of the diterpenoid alkaloid weisaconitine D (**6**)[158]. **c** | Cartoon depiction of results from a detailed retrosynthetic planner that proposes discreet intermediates en route to a target compound. Each node represents an intermediate and each arrow represents a transformation. Depending on the program, this may include proposed conditions for each transformation.

Retrosynthetic logic

The output of retrosynthesis programs is meant to be a guide or recipe for a chemist to synthesize a particular compound. In this section, we detail the variations in outputs of high-level logic-based planners and detailed planners, as well as methods used to validate predicted routes and disconnections.

High-level logic-based programs. Each high-level logic-based retrosynthetic program is designed to apply a particular heuristic. Therefore, the nature of the output also depends on the strategy applied. The level of detail varies from program to program.

For example, some programs allow the chemist to apply a particular reaction, in which case the software will output any proposed instance of, for example, Diels–Alder cycloaddition to build the core of the molecule[31] (FIG. 3a). Ultimately, in this case, the user selects the desired transform-based strategy (such as a Robinson annulation, sigmatropic rearrangements and so on) and often must select the most promising proposal manually. Another well-recognized strategy for the synthesis of topologically complex molecules is the use of network analysis to identify the most impactful bond to disconnect in simplifying a topologically complex structure[158] (FIG. 3b). Programs have been developed to automate this analysis and propose disconnections that can lead to a concise synthetic sequence, and proposals of particular transformations would be left to the chemist user.

It is worth noting that a given heuristic, by definition, does not perform equally well on all molecules. For example, as an analysis that seeks to reduce structural complexity by breaking apart bridged structures, applications of bond-network analysis can only be invoked in the synthesis of bridged molecules. Therefore, evaluating and directly comparing different programs is a challenge and each program is, by intention, applicable or advantageous only in certain cases.

Detailed retrosynthetic planners. The output of retrosynthetic planners is relatively uniform, as each planner generally proposes a synthetic route through several intermediates[54,126] (FIG. 3c). Additional rigour can be added to a program through the use of reaction prediction software[41,53,60,159,160], either as part of the algorithm or through post-processing, to indicate how likely the corresponding forward synthesis is to proceed. As each molecule will generate multiple solutions, another necessary feature is the ability to rank and highlight the most promising of several possible routes that were able to return a 'successful' synthesis. Typically, metrics such as anticipated selectivity and expected yield based on literature precedents, as well as step count, are prioritized in ranking syntheses as these are likely to correlate with whether the synthesis will be viable and economical — in terms of cost, time and waste produced — in the laboratory.

To determine whether a program has successfully identified a viable route, synthetic validation of the proposal is preferable, but this requires significant investment of resources. Nonetheless, several computer-generated syntheses have been successfully

carried out in the laboratory[38,126]. Other methods of validation include comparisons of the proposed route with existing literature that was not in the program's data set[41,42], or double-blind surveys with trained chemists[54] who can probe the perceived effectiveness, as well as the elegance, of a route.

Each aspect of retrosynthesis programs, such as the development of the module to generate disconnections, the prioritization of different routes and the maximum allowed search depth, is designed to balance performance and computational cost to maximize success within the intended application. The computational cost associated with successfully identifying a synthetic route to a target varies widely depending on the complexity of the molecule and the algorithms used. State-of-the-art programs perform well on reasonably complex targets and can identify many well-precedented reactions such as cycloadditions as well as many modular routes that include common reactions such as Suzuki and amide couplings. The synthesis of topologically complex targets remains at the forefront of the field. Recently, computationally predicted retrosynthesis of the fused-ring alkaloid (*R*,*R*,*S*)-tacamonidine was experimentally vetted, marking one of the most complex molecules synthesized with the aid of a retrosynthesis program to date[161].

Reaction prediction

Models must first be vetted by statistical validation metrics before being applied towards the prediction of out-of-sample reactions and interpreted for their chemical significance. At its simplest, the output of these models is a reaction yield or selectivity for a specific regioisomer or stereoisomer of the product (given as the difference in free energy of activation ($\Delta\Delta G^{\ddagger}$) value between the transition states to reach the major and minor isomers) for a given set of conditions that have been parameterized. More complex reaction prediction algorithms can provide the user with the proposed product or conditions to successfully transform a set of reactants to products. These models can then be used to virtually screen proposed reactions, predict higher yielding or more selective conditions, or construct forward routes for retrosynthetic planners.

Model validation. To construct a model for reaction prediction, the data must first be split into a training set and a test set; the former is used in building the model, whereas the latter set is used to assess the predictive power of the model based on data for which there are known outputs. Best practice requires the use of cross-validation (for example, *k*-fold cross-validation); this process creates several training test splits with which to construct a model, such that the model learns from different subsets of the data set. Cross-validation is performed in an effort to avoid overfitting to a single training set and improve a model's predictive ability. When it is necessary to tune the settings for a given algorithm, a training–validation test split is performed, with the model built using the training set and evaluated using the validation set. The settings of the model, commonly referred to as hyperparameters, are adjusted and the model retrained until the outcome for the validation set is satisfactory; at this point, keeping the hyperparameters constant, the model is evaluated against the test set. Cross-validation can be performed on the training–validation data to improve the model in a process known as nested cross-validation[17].

Common metrics used to assess a model include the coefficient of determination R^2 value, which assesses the fit of the model when comparing the values predicted by the model against those empirically observed (for example, predicted versus experimental yields). Models that overfit the training data can have a high R^2 value for training data but a low R^2 value for test data, indicating that they are likely unable to successfully predict outputs beyond the data from which the model learned (see the Limitations and optimizations section for guidance regarding avoidance of overfitting). Further, despite being a common metric used to assess models for reaction prediction, the R^2 value has been shown to be limited in its ability to characterize model predictive power and should be used in tandem with additional statistical metrics[162]. The root mean squared error is one metric that can be applied, which determines the averaged value of the difference between observed and predicted outputs. A low root mean squared error indicates good fidelity between predicted and observed values. Importantly, measures of model fit on training data cannot be used as reliable indicators for how a model will perform against new data.

To test the robustness of a model more rigorously for out-of-sample prediction, especially in cases where these out-of-sample data points are dissimilar to the data points on which the model was trained, purposeful skewing of the training set has been performed. For example, in training a neural network model to predict enantioselectivities for a chiral phosphoric acid-catalysed transformation, a training set composed entirely of reactions with low selectivities was used to predict the higher selectivities of a test set[73].

Mechanistic insights. The descriptors that are classified as most important to modelling reactivity or selectivity can be further investigated in mechanistic experimentation, providing a means of quantitatively interrogating structure–function relationships. Interpretability of a model and its descriptors is essential for mechanistic analysis. In a best-case scenario, the model can point to the descriptors most influencing the prediction, which are in turn rationalized by the chemist on the basis of their understanding of chemical reactivity. Linear models, providing the chemist with the most important descriptors in the form of dependent variables with coefficients, are the most transparent and readily amenable to interpretation. By contrast, non-linear methods can be challenging to interpret owing to the complexity of these algorithms. For example, although a single decision tree can be rationalized by a chemist, the ensemble that constitutes a random forest model becomes challenging to read. Feature importance plots (available with SciKit-learn and other open-source software libraries) that indicate the relative contribution of a descriptor to the predictive ability of the model, for example, can provide insights that facilitate mechanistic interrogation[93].

k-fold cross-validation
A method for evaluating model performance on limited data. The data are split into *k* groups; one group is a test set, whereas the other is used as the training set. This is repeated *k* times to train and test the several groupings of the data.

R^2 value
(Also known as the coefficient of determination). A measure of how well a model fits the data when comparing the measured values against predicted values for the training set. An R^2 value of 0.8 means that the model can account for 80% of the observed variance in the data.

Automation

In terms of automation, outputs can be physical if the product sample is isolated or information-based if the reaction is analysed solely to produce performance data. In the latter case, products are typically not isolated, especially on a small scale, although miniaturized chromatographic purification, as well as the direct submission of reaction mixtures to a bioassay, are emerging areas of research. In terms of exporting information, a minimum data output would typically include an identifier describing each reaction component, such as a SMILES string, and the reaction metadata the experiment has been designed to capture, such as the temperature, concentration, pressure or stirring rate. The advantage of automated synthesis is that such metadata, which is critical for reaction prediction, is often captured and systematically reported. By contrast, for many manually performed reactions reported in the literature, such metadata cannot be easily accessed. As described above (Experimentation), a custom software package or a set of scripts to unite multiple commercial software packages is often needed. Electronic laboratory notebooks can complement custom software, but are typically not designed to capture all necessary details and dataflow for an automated experiment.

In some cases, an isolated yield of physical product is reported, but in most HTE studies the output is an analytical read-out[24] such as the peak area in a HPLC or mass spectrometry analysis. Distinct compounds will have different UV extinction coefficients or mass ionization abilities, such that reaction results are only qualitative unless calibration curves are performed to achieve quantitative results. NMR has been used to generate quantitative results in automated systems but the throughput is low. Evaporative light-scattering detection[163], charged aerosol detection[164] or matrix-assisted laser desorption/ionization[120,165] with deuterated internal standards that are chemically similar to the analytes have been used to provide semi-quantitative results in a high-throughput manner. Most typically, an HTE experiment is used to generate information and select reaction wells are targeted for confirmation on a traditional reaction scale, with traditional gravimetric analysis and NMR characterization of purified products. In addition to enabling the systematic study of defined reaction space, increased experimental throughput can accelerate the serendipitous discovery of new reactions. The initial conditions are often suboptimal for new reaction discovery when the products appear only as minor products and can be easily overlooked in a high-throughput data set. Deconvolution algorithms have been developed to highlight these minor peaks as potentially novel reaction products[166,167].

Applications

Retrosynthetic logic

Computer-assisted retrosynthesis has been an active area of research for more than half a century that has seen renewed interest because of the significant opportunities offered by the revolution in data science and new machine learning techniques. This renaissance has culminated in both excellent software to inspire chemists through logic-based programs and fully automated generation of retrosyntheses that have been experimentally validated in the laboratory.

High-level logic-based retrosynthesis. The LHASA program is a forerunner in the field that is intimately tied to the development of retrosynthesis as a whole. Although the strategies implemented in LHASA have been refined, they form the basis of high-level logic approaches. LHASA is no longer available, but serves as an important foundation in the field. For this reason, our first case study is the application of LHASA's chiral starting material program to the synthesis of a bicyclic compound (**18**) (FIG. 4a) where the mapping heuristic was able to identify a non-obvious starting material, citronellol (**19**)[168]. After the program identified the material, the chemist users interfaced with LHASA's detailed planner to arrive at a complete retrosynthesis. Independently, comparisons with published work[169] indicated this was a viable strategy that had been previously carried out successfully in the laboratory. This case exemplifies that applying retrosynthetic programs can help find non-intuitive solutions that may not be obvious to the human observer, and can afford elegant and efficient syntheses.

Detailed retrosynthetic planners. In the past decade, there have been numerous reports on the implementation of detailed retrosynthetic planners and various components of the software. In particular, we review two retrosynthesis programs and their approach and successes in identifying routes to complex molecules. The reader is also referred to other literature detailing synthetic planners that highlight developments in this area[39,54,170–173].

SYNTHIA, formerly known as Chematica[30], has been developed over the past decade and has recently been commercialized by MilliporeSigma. This software boasts a collection of more than 100,000 expert-coded rules that are recursively applied to synthetic targets. In each hand-coded rule, there also exist general reaction conditions and identification of potential functional group conflicts, which provides insight into whether protecting groups are necessary (SYNTHIA). Two important components in multistep synthetic planning are how the algorithm selects which branches of the synthetic tree to pursue, that is determining which disconnections are valuable, and the way in which the algorithm selects the 'best' of many multistep syntheses. In SYNTHIA, these functions are called the chemical scoring function and the reaction scoring function, respectively, and are editable by the user to allow prioritization based on factors such as the overall step count, starting material cost and use of protecting groups.

In a recent example, several drug-like molecules of interest to MilliporeSigma were analysed by SYNTHIA (FIG. 4b) and the routes were vetted experimentally[38] (FIG. 4c). In these cases, the average computation time was around 15–20 min per molecule and the first or second top-rated route was chosen to vet without allowing significant changes from the proposed reagents. The restriction was implemented that only 70 h of bench work would be allowed to develop the

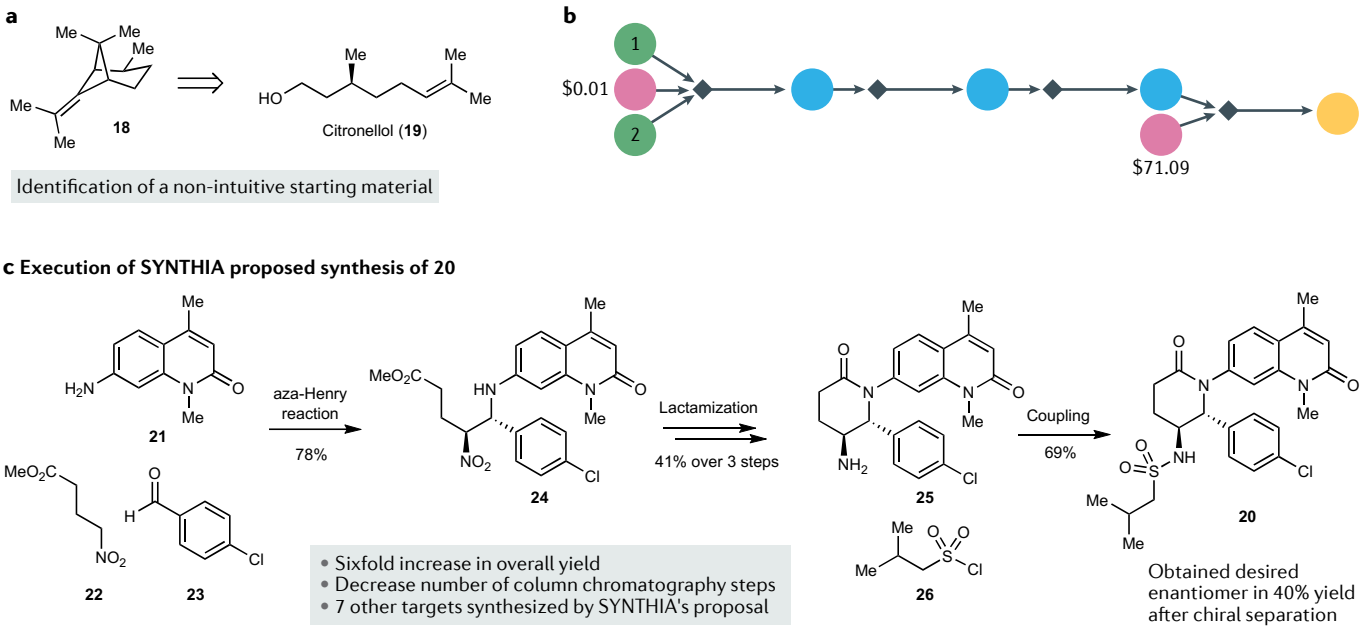


Fig. 4 | **Applications of retrosynthetic planning validated by laboratory efforts. a** | Identification of a non-obvious enantio-enriched, abundant starting material citronellol (**19**) by Logic and Heuristics Applied to Synthetic Analysis (LHASA)[168]. **b** | Chematica/SYNTHIA-generated results for lactam **20**. Each diamond represents a reaction step, showing the map of known starting materials (green) and commercial starting materials (pink) through intermediates (blue) to the product (yellow)[38]. **c** | Laboratory-validated forward synthesis of **20** with only minor variations from Chematica/SYNTHIA-generated results[38]. Panel **b** adapted with permission from REF.[38], Elsevier.

experimental execution of predicted routes. In particular, for a quinoline-fused lactam (**20**) selected for its biological activity, SYNTHIA's proposed route was successful experimentally, decreased the required column chromatography steps from five to three and led to an overall yield improvement of more than sixfold that previously reported. Seven additional targets were also synthesized using the output from SYNTHIA and each included some benefit over the previous route, including higher yields, shorter step counts, higher purity or more facile separation.

Another well-known package is the ASKCOS suite, an open-source software package meant to aid in many aspects of computer-assisted synthesis including reaction planning and one-step retrosynthesis[14,42,52,58,60,77,160]. Several of these tools individually fit into either a logic-based approach or an iterative automated set-up that has recently been applied to retrosynthetic planning for the synthesis of 15 drug molecules[126] (FIG. 5). Notably, by demonstrating the efficacy of retrosynthetic software with integrated reaction prediction tools and by validating the retrosynthetic proposals with fully automated syntheses, this work represents a union of the three topics covered in this Primer.

Building on earlier developments in detailed retrosynthetic planning[54], ASKCOS was trained to automatically extract reaction templates from the US Patent and trademark office as well as the Reaxys database, arriving at a library of 163,723 transforms. Within the algorithm, the templates extracted from similar molecules and, thus, predicted to work well on the target molecule are applied to identify possible disconnections[41]. At this stage, another module of the software tests forward reaction prediction software to ascertain whether reaction conditions that are expected to furnish the desired product from the proposed starting material exist (FIG. 5b–d, see Reaction prediction case study below). Another module that predicts the major product is used to determine whether there are likely to be side products or other considerations that affect the feasibility of a given transformation (FIG. 5e). Limits are set on the maximum search depth to explore the synthetic tree and it is explored through a Monte Carlo tree search to balance the exploration of 'promising' routes as well as the exploration of less frequently visited branches. Notably, the program successfully developed routes to targets it had never seen before, as well as useful active pharmaceutical ingredients.

The proposed synthesis of safinamide (**27**) (FIG. 5a), as well as 14 other medicinal compounds, was completed from inexpensive, commercially available reagents using a fully automated system (FIG. 5f, see Automation case study below).

Ultimately, these examples showcase the capabilities of retrosynthetic planning software to arrive at new syntheses of useful molecules in a quick and efficient manner.

Reaction prediction

Here, we present a sampling of case studies that highlight successful implementations of the various tools and approaches to reaction prediction, demonstrating the strengths of each approach and describing potential for future development. The reader is referred to other reaction prediction studies that highlight progress in this area[64,73,75,78,96,174,175].

Reaction selectivity. A two-step approach to modelling using multivariate linear regression was used to predict enantioselectivities of various chiral phosphoric acid-catalysed nucleophilic additions to imines[94] (FIG. 6a). A comprehensive multivariate linear regression model for a data set of 367 literature reactions, with enantioselectivity as the output ($\Delta\Delta G^{\ddagger}$), was built using a combination of two-dimensional quantitative structure–activity relationships and three-dimensional computational descriptors to parametrize the catalysts and substrates (see, for example, **40**–**43**, **45**, **46**). The sign of $\Delta\Delta G^{\ddagger}$ was interpreted to assign each reaction to one of two possible transition state geometries (*E*-imine or *Z*-imine). Subsequently, the data were partitioned by transition state geometry and used to construct two focused multivariate linear regression models (FIG. 6b). For out-of-sample predictions, the parameters were first passed through the general model before using the appropriate focused model based on the predicted transition state geometry.

This approach allowed for improved predictive ability (as measured by the average absolute error of $\Delta\Delta G^{\ddagger}$) and more nuanced mechanistic insights that account for the different catalyst–substrate interactions incurred in the two transition states. The coefficients of each linear model highlighted the steric and electronic properties of both the substrate and the catalyst that influence selectivity in both transition states. Both models had common steric terms for imine and catalyst structure, with the *Z*-imine transition state having an additional steric term associated with the nucleophile. The strength of this approach is its ability to predict outcomes for diverse reactions and novel catalyst structures within a common manifold (FIG. 6c), while also applying a linear algorithm that facilitates analysis.

Reaction yields. In 2017, a random forest algorithm was implemented to predict reaction yields in a palladium-catalysed Buchwald–Hartwig amination reaction of aryl halides (**49**) with 4-methylaniline (**48**) to give **50** (REF.[93]) (FIG. 7a). One objective of this work was to model and predict reactivity in the presence of challenging heterocyclic motifs; however, screening reactions across multiple substrates traditionally necessitates the individual synthesis of each product to enable quantitation, thus limiting the scope of unique reactions that can be run. As an alternative, the Glorius fragment additive screening approach[176] was applied, where fragment additives representing different functional groups are added to the reaction to interrogate the influence of such discrete functional groups on the reaction. In the present case, 23 unique isoxazole additives were screened against a matrix of 4 palladium catalysts, 15 aryl halides and 3 bases to measure the inhibitory effects of the additives. Using the ultraHTE (1,536-well plate) set-up at Merck & Co., Inc., this strategy resulted in a data set of 4,608 unique reactions for modelling.

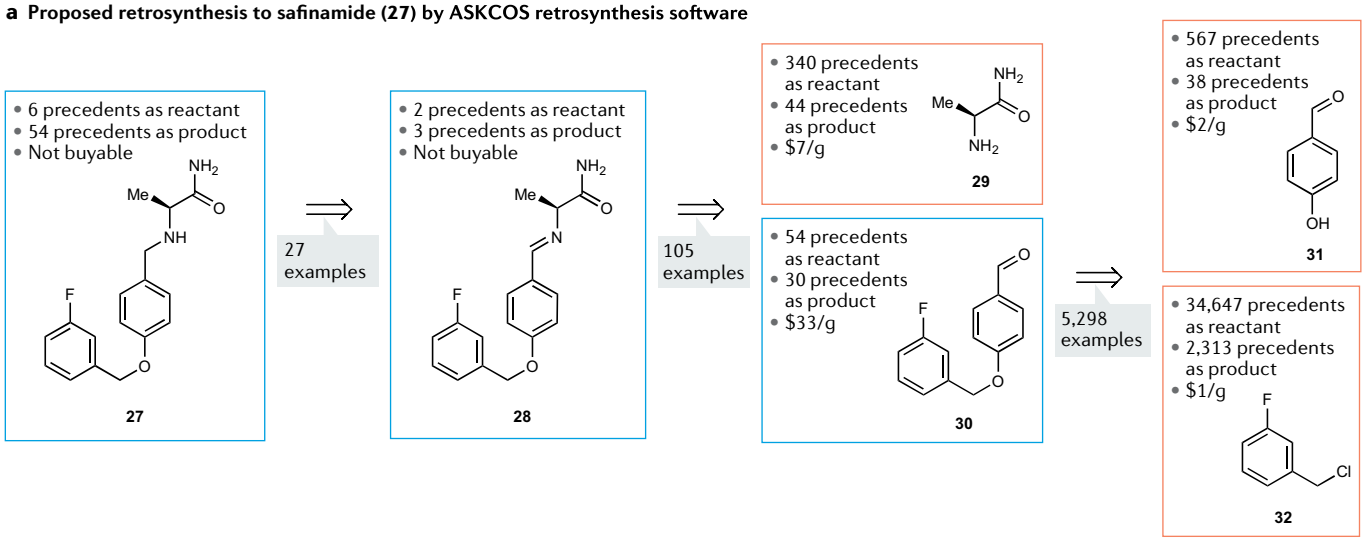
Molecular, atomic and vibrational properties were computed by DFT using automated feature-generation software. The random forest model proved to have the highest predictive performance (test set prediction $R^2 = 0.92$, root mean squared error = 7.8%) (FIG. 7b) over other linear regression and machine learning methods, and was used to predict the yields for a series of out-of-sample additives. Although the model itself is not readily interpretable in the same sense as a linear model, the construction of the model was probed by determining relative feature importance as demonstrated by the increase in error that occurs upon shuffling the values for a given descriptor and retraining the model (FIG. 7c). The descriptors that caused the greatest increase in error — in this case, a series of descriptors that reflect the electrophilicity of an isoxazole additive such as **51** or **52** — guided stoichiometric mechanistic studies into side reactivity (FIG. 7d). Specifically, for the more electrophilic isoxazole additive, N–O oxidative addition followed by Kemp elimination was shown to be a potential competitive pathway preventing oxidative addition of the aryl bromide.

Reaction conditions. In a larger-scale implementation of reaction prediction, a general platform for the prediction of organic reaction conditions (catalyst, solvent, reagents and temperature)[77] was developed, and has subsequently been implemented as part of the fully automated ASKCOS platform[126]. The data set consists of approximately 11 million reactions obtained from the Reaxys database. It is important to note that, as is common practice in generating data sets from the literature, they employed a series of 'data cleaning' protocols. These filters helped eliminate incomplete and uninterpretable entries, homogenize the syntax of input data and filter out 'rare' reaction components for which data are sparse.

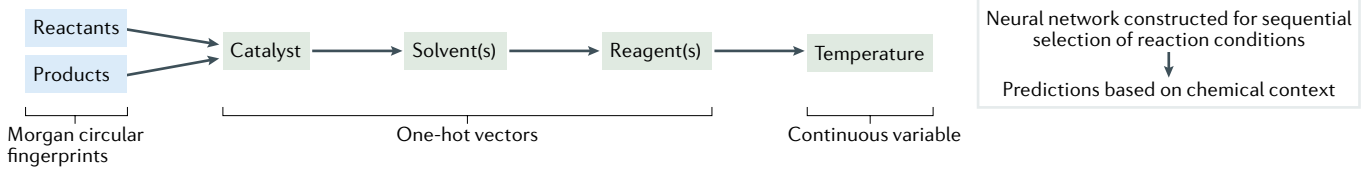
Starting from their respective SMILES string, reactants and products were described using Morgan circular fingerprints. Temperature was included as a continuous variable, whereas the catalyst, reagent(s) and solvent(s) were represented as one-hot vectors. These data were used to train a neural network model, which was constructed to allow for sequential prediction of each condition — catalyst, followed by solvents, reagents and temperature — to enable context-based recommendations (FIG. 5b,c). This structure was created to reflect the way in which chemists commonly approach constructing conditions for a reaction and facilitate the selection of chemically compatible conditions. Overall, in assessing the top-10 predictions made for 1 million test set reactions, they were able to accurately predict the catalyst and at least 1 solvent or reagent 69.6% of the time (FIG. 5d). The breadth of reactivity that this model can assess makes this platform suitable for use in the prediction of reaction conditions for multistep syntheses.

The descriptor set chosen for this model prioritizes throughput over interpretability, given the need to parameterize approximately 11 million reactions that involve structurally diverse reactants and products[77]. By contrast, the DFT-based descriptors used in the respective reports by Reid and Sigman[94] and Doyle and colleagues[93] are more interpretable for mechanistic interrogation but are computationally more expensive. As these reports focus on a specific reaction or are manifold of reactivity employing a limited scope of reactants, catalysts and reagents, it is feasible to perform the requisite molecular structure optimizations needed to get these descriptors for all variables of interest.

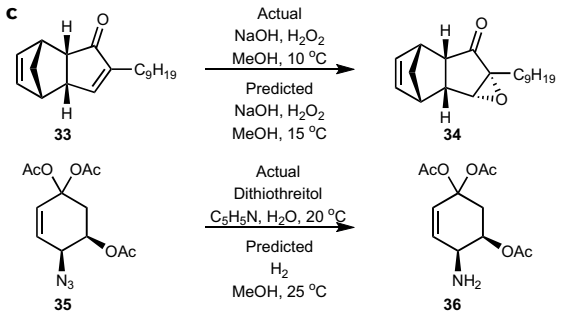
a **Proposed retrosynthesis to safinamide (27) by ASKCOS retrosynthesis software**



b **Schematic of neural network model implemented in ASKCOS for prediction of conditions**



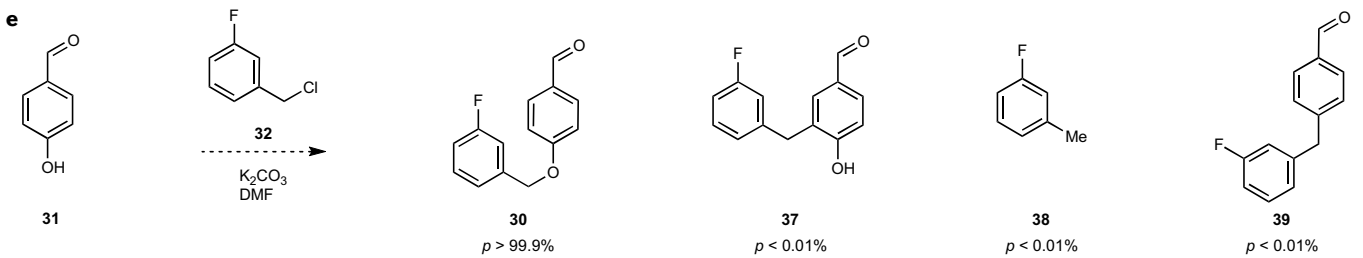
c



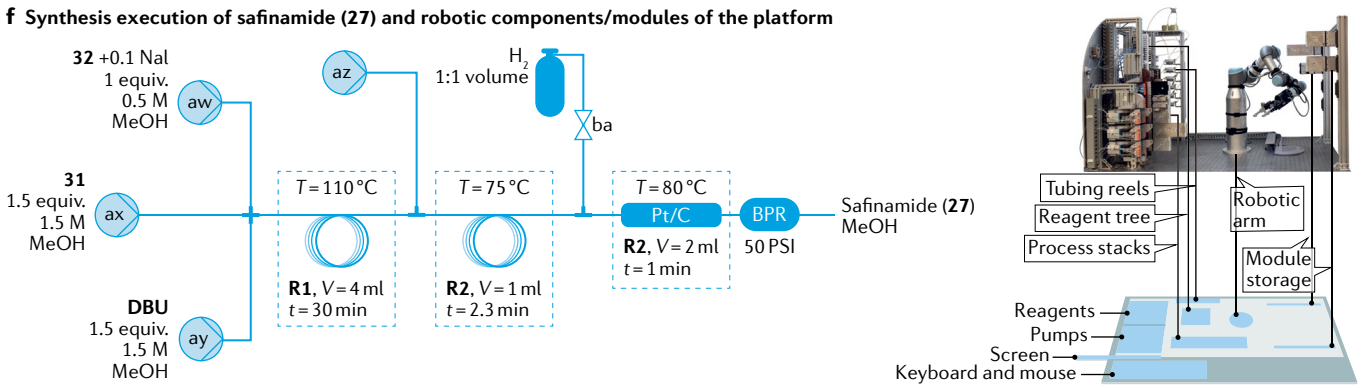
d **Accuracy of prediction tasks (%)**

Prediction task	Top 3 exact	Top 10 exact	Top 3 close	Top 10 close
Catalyst (c)	93.6	94.9	94.9	96.4
Solvent 1 (s1)	75.8	83.0	78.2	85.4
Solvent 2	90.1	91.7	90.2	91.9
Reagent 1 (r1)	73.2	83.1	74.8	84.9
Reagent 2	89.3	91.8	89.3	92.1
c, s1, r1	57.3	66.0	60.4	69.6
All	50.1	57.3	53.2	60.3

e



f **Synthesis execution of safinamide (27) and robotic components/modules of the platform**



◀ Fig. 5 | **Retrosynthetic planning, reaction prediction and automated synthesis platform directed by ASKCOS. a** | Proposed retrosynthesis of safinamide (**27**) by automated retrosynthesis software[126]. Compounds that are not commercially available (blue box) can be built from available starting materials (orange box). **b** | Schematic of the sequential selection of reaction conditions performed by the neural network model used by ASKCOS[126]. **c** | Example of correct and incorrect reaction condition prediction made by the neural network model[77]. **d** | Accuracy of neural network model for various reaction condition prediction tasks[77]. **e** | Example forward reaction prediction for the base-mediated coupling of phenol **30** and arene **37**. This validation was performed for each reaction in the proposed retrosynthesis. **f** | Automated synthesis of **27** in the workflow of the ASKCOS platform (left); specific fluid streams labelled alphabetically (blue circles). Mechanical components and unit operations of the ASKCOS flow synthetic platform (right)[126]. BPR, back-pressure regulator; DBU, 1,8-diazabicyclo[5.4.0]undec-7-ene; MeOH, methanol; R1, reactor 1; R2, reactor 2. Panel **a** adapted with permission from REF.[126], AAAS. Panel **d** reprinted with permission from REF.[77] (https://pubs.acs.org/doi/10.1021/acscentsci.8b00357), ACS; further permissions related to the material excerpted should be directed to the ACS. Panel **f** reprinted with permission from REF.[126], AAAS.

Automation

The field of automated synthesis has advanced rapidly in the past decade. We highlight a handful of examples, and refer the reader to other recent developments[114,126,144,177].

Reaction selectivity. A dehydrogenation step required to generate the key intermediate (**55**) for the synthesis of the hepatitis C treatment elbasvir (**53**) was studied using HTE in 96-well plates with the aim of finding an environmentally friendly replacement for $KMnO_4$ in a key oxidation[177] (FIG. 8). Here, 4 oxidants, 12 photocatalysts and 2 solvents were screened to find a promising photoredox system. Following detailed mechanistic work, results from scouting reactions run on a 2.5-µmol scale in glass shell vials were translated to a 100-g scale in a photochemical flow reactor.

Reaction miniaturization. Reaction miniaturization has emerged as a key technology enabled by modern automated synthesis. Recently, thousands of transition metal-catalysed coupling reactions were run in 1.2 µl droplets and the reaction products directly submitted to a bio-affinity assay against diverse kinase proteins[122]. Reaction screening on a similar scale was also performed in a custom flow apparatus assembled inside an inert atmosphere glovebox[110]. Acoustic droplet ejection has become a popular technology for low-volume liquid handling in biochemical experimentation, and has recently been used in automated chemical synthesis[111]. An Ugi four-component coupling reaction was used to synthesize indoline analogues[114] (FIG. 9). The stock solutions employed in the reaction were prepared and manually pipetted into a 384-source plate, then dosed by an Echo robot into the 384-well reaction plate and, finally, the reactions were analysed by supercritical fluid chromatography–mass spectrometry and thin-layer chromatography–UV–mass spectrometry. The results were further validated by synthesis of select products on a 10-g scale.

Organic synthesis in a modular robotic system. An automated modular synthesis platform, Chemputer, executed the syntheses of diphenhydramine hydrochloride (**62**), rufinamide (**68**) and sildenafil (**69**) with minimal human intervention[144]. The system was coded with a chemical markup language (XDL), which reduces common laboratory tasks such as mixing or filtration into machine-readable operations and includes metadata for future analysis. The platform has a backbone structure that enables facile switching of modules for routine synthesis tasks such as heating or phase separation (FIG. 10). The backbone has a six-port valve that connects pumps to the modules so reagents or reaction mixtures can be flowed to the appropriate module. The three drugs (**62**, **68** and **69**) were synthesized in 38–100 h for the overall sequence, with yields comparable with reported manual syntheses. An earlier generation of this system was used in a machine learning-driven autonomous search for new chemical reactivity[80].

Following fully automated retrosynthetic planning using the ASKCOS suite (see Retrosynthetic logic case study), the fully automated syntheses of 15 drugs were achieved in a sophisticated flow reactor[126] (FIG. 5f). A robotic arm assembled required modules such as reactors or separators into an appropriate flow path. The reaction types varied from amide couplings to reductions, and the platform yielded the products at a rate of 100 mg h^{-1}.

Reproducibility and data deposition

High-quality data sets are essential for reaction prediction and synthesis planning. Although proprietary databases, such as Reaxys and SciFinder, can yield large amounts of reaction data, many entries can be incomplete or missing key information such as stereochemistry or yields. Mining data from the chemical literature can be a time-intensive process, often requiring the manual extraction of data from supplemental text. Consistency of data can also be uncertain and unsuccessful reactions are often under-represented. The quality of data influences the quality of the reaction prediction models and reaction templates used in retrosynthesis planners. A reaction prediction model built on incorrect or incomplete data seeks correlations that may not exist and, thus, lacks predictive power. The development of robust models trained on large data sets, if available, can overcome a few aberrant data points. In short, ideal data are complete, cover a wide chemical space and are reproducible. HTE is one solution to obtain higher quality data for many applications because it intrinsically generates large data sets in a systematic way. HTE excels in elucidating trends for parallel reactions that interrogate sets of condition variables. Given the miniaturized nature of HTE, reproducibility on a larger scale needs to be assessed, which is often done through the scale up of a selection of wells to validate translation to more traditional conditions.

In the context of both literature-mined and HTE data, the processing and visualization of large amounts of metadata require software or custom scripts. As a result, it has become common for academic scientists to upload the scripts for data processing alongside algorithms and data sets on open-source websites such as GitHub to enable the transfer of these methods between researchers. Reproducibility of reaction data is vital, but even data-driven methods of reaction prediction and retrosynthetic planning are not fully translatable.

Acoustic droplet ejection
A technology that uses precise ultrasound waves to move or transfer nanolitre volumes of solutions. Acoustic droplet ejection transfers the droplets from the source plate into an inverted receiving plate above the source plate.

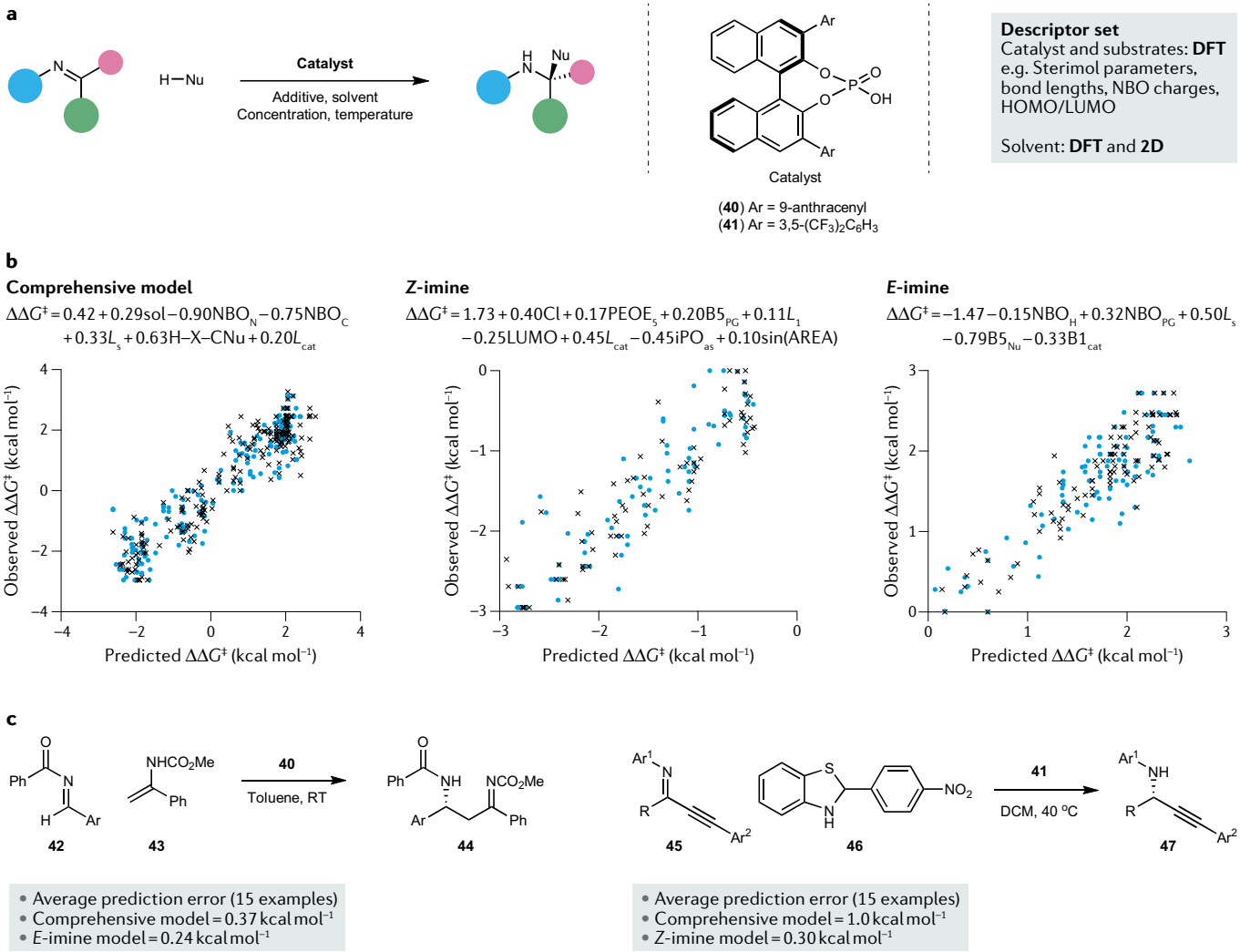


Fig. 6 | **Prediction of reaction enantioselectivities for chiral phosphoric acid-catalysed nucleophilic additions to imines[94]. a** | General reaction scheme for transformations included in the data set. Catalyst and substrates were parameterized using molecular descriptors obtained by density functional theory (DFT) calculations, whereas solvent was described using both two-dimensional (2D) and DFT molecular descriptors. **b** | Comprehensive, *Z*-imine and E-imine regression models. Each model contains a combination of descriptors for both catalyst and substrate properties; for a more detailed explanation of each descriptor, see REF.[94]. Data represent training (black) and validation (blue) sets. **c** | Out-of-sample prediction. Performance metrics demonstrate the improved predictive power of the focused correlations over the comprehensive correlations. $\Delta\Delta G^{\ddagger}$, difference in free energy of activation; NBO, natural bond orbital. Adapted from REF.[94], Springer Nature Limited.

In many workflows, decisions may be made by the user that factor in so-called chemical intuition in the construction of the model — for example, trying to get a linear predictive model to use a particular descriptor based on prior observations or selection of a particular retrosynthetic disconnection. Furthermore, both retrosynthetic planning programs and reaction prediction software often contain a randomized element, such as a randomized initialization in the set of reaction templates or a randomized train/test split that can lead to slightly different results and model validation metrics. Seed data can be provided to aid in the reproducibility of 'random' selections. For these reasons, recreating the exact model from a publication can be challenging, as can duplicating an exact synthetic proposal using step by step algorithms. However, the overall trends are likely to be reproducible for a robust model. In cases where a retrosynthesis program can identify a particularly efficient reaction sequence, it is likely to converge on that route, regardless of randomized elements, if given enough search time.

Limitations and optimizations

Successful implementation of these methods in synthetic chemistry requires an understanding of their inherent limitations. By keeping in mind the challenges incurred in their use, one can avoid common pitfalls and better understand the utility of these techniques.

Retrosynthesis

Although many programs have successfully designed routes to a target compound, there is wide variability in the complexity of the target molecules accessible.

For example, natural product synthesis often focuses on complex sp^3-rich molecules as targets, which often have additional stereoelectronic challenges associated with building their framework. By contrast, a program designed to identify pathways to drug-like molecules is often trained on data specific to sp^2-rich molecules. By virtue of a close literature precedent being intimately tied to a program's perceived likelihood of success for a particular reaction, detailed retrosynthesis programs are generally unable to propose novel disconnections that, if proposed, would be viewed sceptically by the user. Until recently, retrosynthetic planners were only accessible to a handful of groups. Today, accessibility is increasing as many platforms have been made open source (such as AiZynthFinder, ASKCOS, IBM RXN for Chemistry) whereas many others have been developed commercially (such as SciFinder^n, SYNTHIA, Reaxys, ICSYNTH, Chemical.AI and Iktos spaya.ai).

Reaction prediction

Data overfitting can occur when more descriptors are used than there are available data points and this is an inherent challenge of the modelling techniques used for reaction prediction. Overfitting in modelling of chemical reactivity exists in part because reaction data are available in low quantity relative to the data sets with which machine learning algorithms are commonly applied, such as online images or text, emphasizing the limitations presented by reaction data availability. Also, given the high dimensionality of a chemical reaction, the available chemical information is relatively sparse except for a handful of popular reactions. The addition of more descriptors will improve the model's fit to the training data; however, an overfit model will not be effective in exploring chemical space beyond data it has already seen. Using cross-validation and limiting the number of descriptors can help prevent overfitting; the use of a test set can help identify when overfitting has occurred. The use of a larger data set is a conceptually simplistic means of avoiding overfitting, but cannot be accomplished in cases where only a small amount of data are available. By contrast, if a model is not given enough information from which to learn, for example if there are insufficient descriptors to account for the chemical complexity of a system, this model can be underfit, likely limiting its predictive ability.

Automation

In general, the first concern users have when approaching miniaturized HTE is the ability to translate results of small-scale reactions to a traditional scale. For homogeneous reactions, this is generally not problematic, and examples of scaling from microgram up to gram scale are known[109], but in the case of a heterogeneous reaction, issues of particle size or stirring rate may impact fidelity across multiple scales. Solvents must be carefully chosen in automated synthesis, unless additional engineering controls have been implemented. Ideally, a reaction

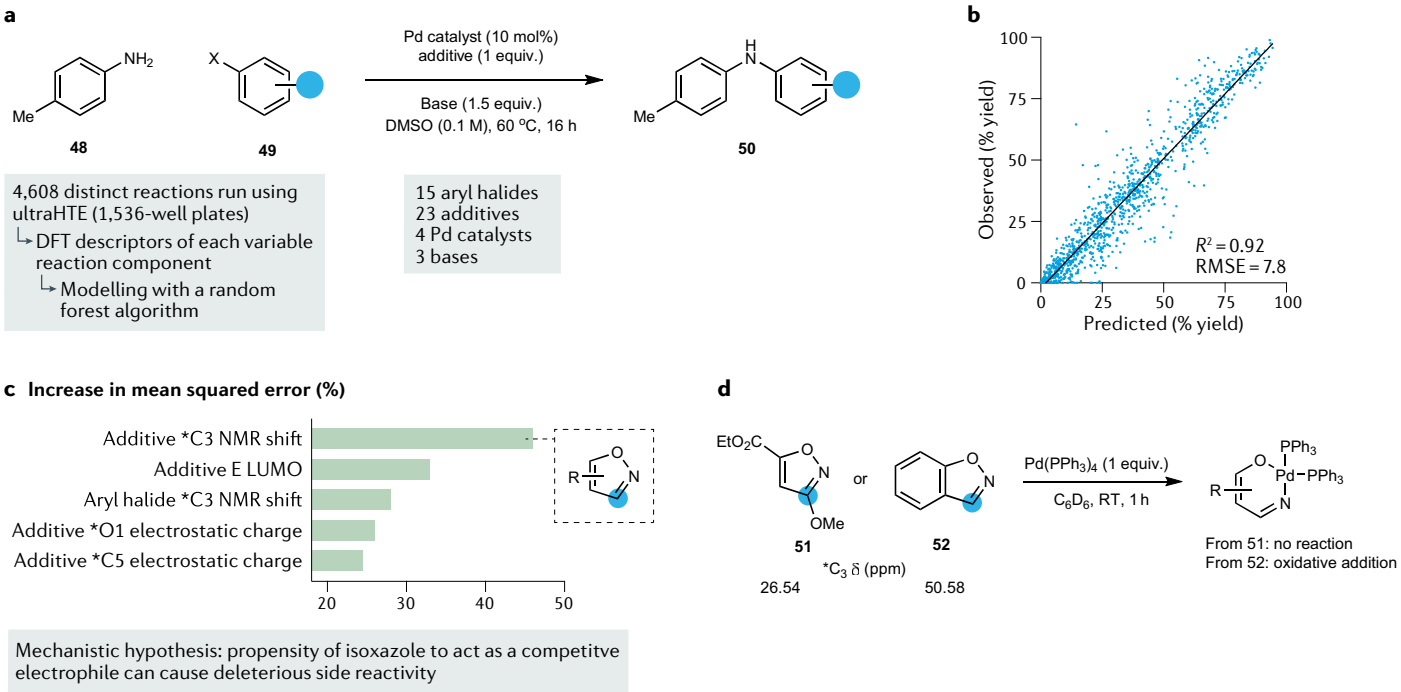
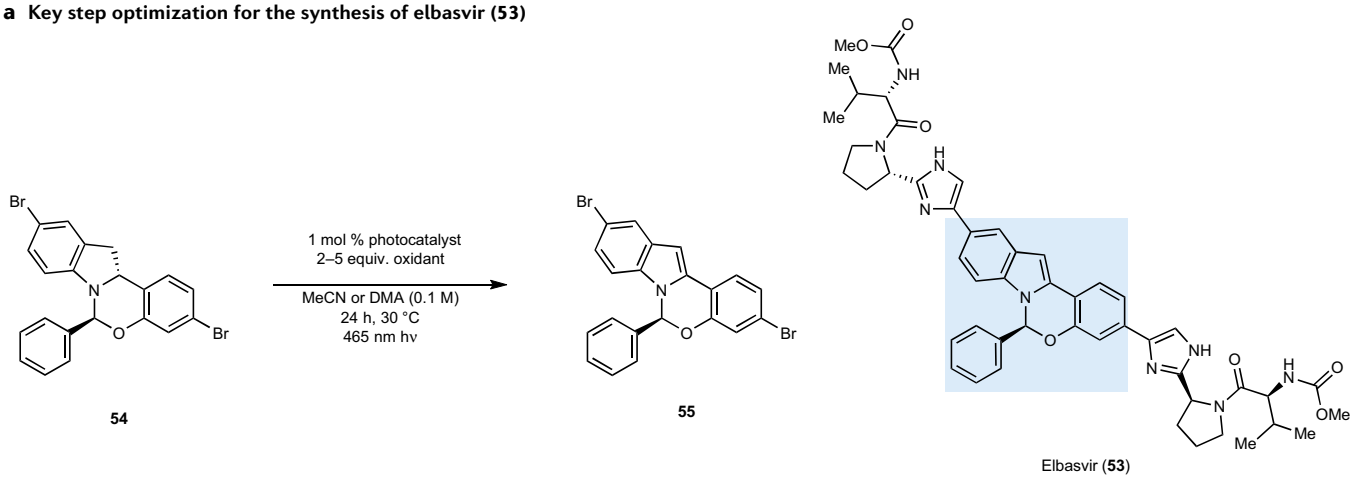
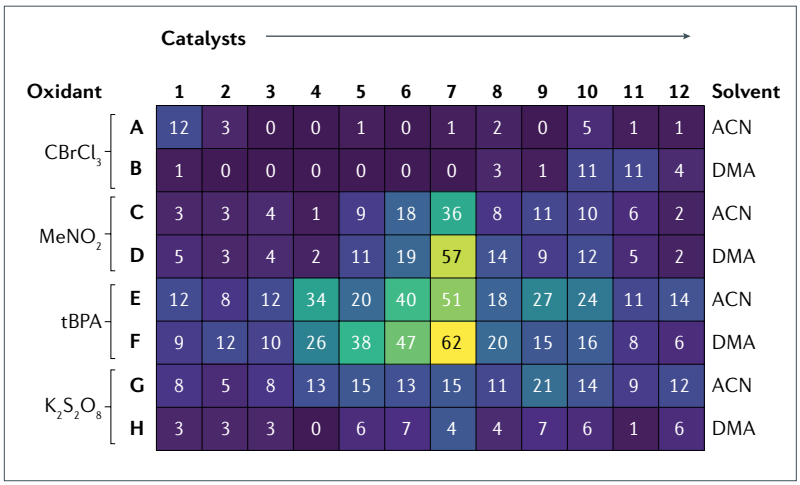


Fig. 7 | **Prediction of reaction yields for palladium-catalysed Buchwald–Hartwig aminations[93]. a** | General reaction scheme and variables employed to build a data set using ultra-high-throughput experimentation (ultraHTE). **b** | Workflow used to develop a random forest model and performance of test set data. **c** | Relative feature importance plot used to interrogate the model. Analysis of the relative importance of various descriptors revealed that the ^{13}C nuclear magnetic resonance (NMR) shift of the isoxazole additive was the most significant contributor to model performance. It was hypothesized that this descriptor served as a read-out of the degree to which the isoxazole can act as a competitive electrophile. **d** | Mechanistic study to assess the competitive N–O oxidative addition of the isoxazole additive to Pd(0). The observation of both oxidative adducts suggests that more electrophilic isoxazoles (**51**) promote this side reactivity that causes diminished yields of the desired Buchwald–Hartwig amination product. DFT, density functional theory; R^2, coefficient of determination; RMSE, root mean squared error. Adapted with permission from REF.[93], AAAS.

a Key step optimization for the synthesis of elbasvir (53)



b HTE using parallel photoredox screening platform and the design of conditions for the key dehydrogenation step



Catalysts

1 $[Ru(bpy)_3](PF_6)_2$
2 $[Ru(phen)_3]Cl_2$
3 $[Ru(dip)_3]Cl_2$
4 $Ir(ppy)_3$
5 $Ir(dF\text{-}ppy)_3$
6 $[Ir(ppy)_2(dtbpy)](PF_6)$
7 $[Ir(dF\text{-}CF_3\text{-}ppy)_2(dtbpy)](PF_6)$
8 Fluorescein
9 Eosin Y
10 Eosin B
11 Rose Bengal
12 None

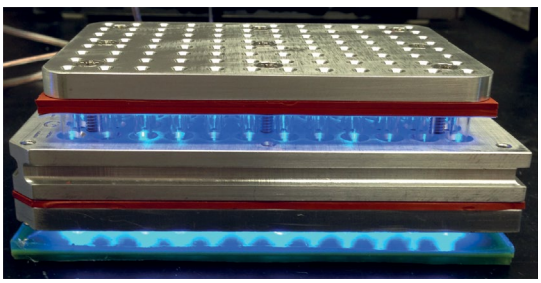


Fig. 8 | **Probing reaction selectivity using high-throughput experimentation**[177]**. a** | Optimization of the dehydrogenation step in the synthesis of elbasvir (**53**). **b** | Photoredox conditions designed and screened by high-throughput experimentation (HTE) in a 96-well plate. Numbers in each well in the colour map (middle) are percentage assay yields based on high-performance liquid chromatography analysis, relative to an internal standard; blue represents lower yields and yellow represents higher yields. Panel **b** adapted with permission from REF.[177] (https://doi.org/10.1039/C5SC03350K), RSC.

mixture will have low volatility and be homogeneous, compatible with plastic components of the robot and consumables (typically polypropylene or cyclic octane copolymer). Increasing the reaction volume can increase solubility and minimize evaporation, but most reactions have favourable kinetics at a substrate concentration of 0.1 M or above, so dilution is not a general fix[118]. Solvent screening is also a challenge because multiple stock solutions must be prepared for the same reagents. Another limitation is the accessibility of expensive robots and hardware, as many systems are built de novo and only a handful of commercial systems have been used by more than one research group. Solid handling robots are considerably less developed, and for exploratory chemistry where diverse solid properties are encountered, manual solid handling remains common[124,125]. Reconfiguration of procedures is needed when switching reaction types or consumables and human intervention is generally required, especially the first time a system is run.

Outlook

As the field of synthetic chemistry works to incorporate emerging methods for retrosynthesis, reaction prediction and automation into the synthetic chemist's toolkit, it is necessary to take stock of what has been accomplished. The work highlighted in this Primer represents a limited cross-section of achievements in the field, new developments are emerging. Automated and high-information content solutions should not be viewed or implemented as a replacement for the chemist, nor as a 'fix' to every challenge of synthesis. Rather, these methods enable chemists to streamline the tedious steps of synthesis and focus on creativity.

Towards this goal, there are several avenues for targeted development that will prove important across retrosynthetic planning, reaction prediction and automation. Continued work to develop generalizable molecular representations will improve the ability with which a computer can learn chemistry and assess

reactivity. An idealized descriptor or descriptor set can be applied across classes of molecules and provide complete chemical context without requiring extensive computational methods. Upgrading of software, hardware and equipment is required to pave the road to increased automation. Some vendors offer commercially available consumables such as glass vials, reaction blocks and customized plates for specific reaction conditions. Vendors now provide easy entry points for new users, such as commercialized consumable kits for screens in 24-well and 96-well plates. Increased commercial availability, with good selection and competitive pricing, is critical to advance any new area of science. Instruments from different vendors often cannot be easily integrated with each other, limiting the scope of accessible experiments. The instrument data and hardware interoperability gap will become increasingly recognized as more chemists request instrument integration from vendors, and either suitable instrument drivers are made available or instrument back-end software becomes more accessible for customization.

A challenge the community must tackle is the lack of access to uniform data. Literature repositories contain a wealth of chemical data, yet it remains challenging to organize this information into a usable format. Data mining of literature sources often requires the manual collection of data from the literature. Supporting information, for example, is typically housed as a portable document file (.pdf), from which data extraction can require considerable effort, although automated extraction is emerging[178]. The data harmonization challenge can be addressed with the development of open-access databases for reaction data and the creation of a standardized data submission format in journals. As much as the community has adopted a standard for reporting NMR characterization data, there is momentum to standardize the documentation of routine reaction metadata as high-dimensional data formats become readily accepted by journals. Similarly, automated synthesis has the capability to generate a large volume of new data, which will only be useful to the field if well formatted. Similar to X-ray crystal structures submitted to the Cambridge Crystallographic Data Centre (CCDC) or Protein Data Bank (PDB), or DNA sequences submitted to GenBank, uniform reporting of reaction data will hopefully become a community standard.

With respect to retrosynthesis planners, many of the available detailed planners will allow for synthesis applications focused on building moderately complex compounds. We envision that these planners can augment the way we search the literature, much in the same way that database searching using Reaxys or SciFinder proliferates data. In this way, we can accelerate the process of navigating many potential routes and enable practitioners to focus on bigger overall questions in their research.

Although the use of these retrosynthetic planning programs rely on known, well-precedented reactions, computer-assisted synthesis also fosters innovation. The use of heuristic-based programs often invites and

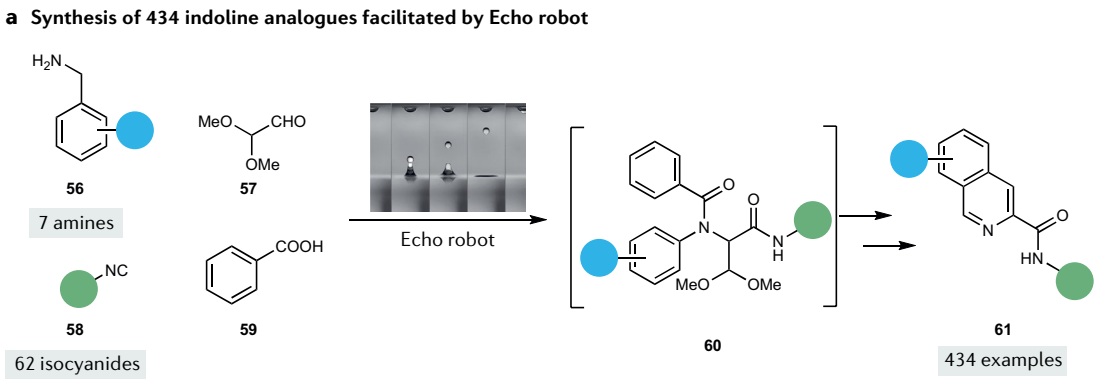


Fig. 9 | **Reaction miniaturization and validation**[114]**. a** | Discovery of a miniaturized Ugi coupling reaction for the synthesis of indoline analogues (**61**) facilitated by an Echo robot. **b** | Reaction progress from the discovery of the hit in 384-well plates, to confirmation of selected experiments at a scale of 80–1,000 mg. HTE, high-throughput experimentation. Echo robot image in panel **a** reprinted with permission from REF.[111], RSC.

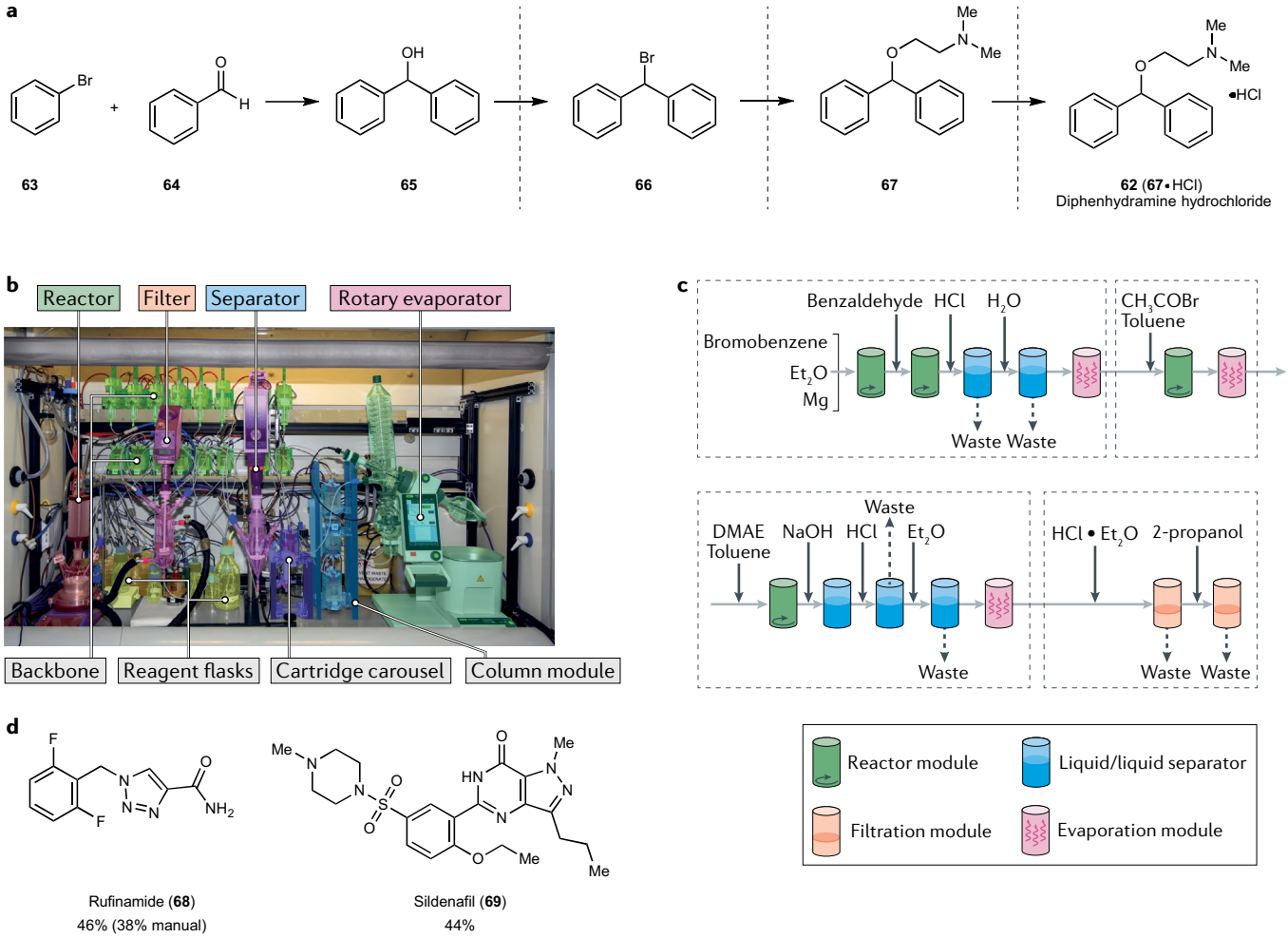


Fig. 10 | **Organic synthesis using a modular automated robotic system**[144]**.** Schematic route (panel **a**) for the synthesis of diphenhydramine hydrochloride (**62**) in an autonomous platform directed by Chemputer, with the apparatus for the platform (panel **b**) and a cartoon illustration of the platform (panel **c**). Two other medicinal active molecules synthesized on this platform were rufinamide (**68**) and sildenafil (**69**) (part **d**). Image of Chemputer in panel **b** courtesy of L. Cronin. Panel **c** reprinted with permission from REF.[144], AAAS.

necessitates creativity from chemists to arrive at a more ideal route. Additionally, step by step retrosynthetic programs rely on definitions of synthetic complexity and a reduction in complexity, where there is still considerable room for consensus and improvement.

In order for these methods to be adopted broadly, the technical developments made in the laboratory will need to be met with changes in the classroom where chemists are trained to use these tools. Traditional chemical education does not include statistics or computer science in a way that will provide students with the foundational knowledge used to develop and implement computer-assisted synthesis methods. The means of addressing this challenge are twofold. Chemistry curricula can introduce information science to help cultivate a generation of 'bilingual' chemists who are conversant in data science, whereas the tools themselves can be strategically designed to be accessible and user-friendly to enable their adoption in synthesis. Every chemist does not need to be a statistician or computer scientist but, rather, needs to be provided with the knowledge necessary to implement these tools at a high level or possess the ability to converse with data-savvy collaborators, all while maintaining the highest standards of creativity and rigour in synthetic organic chemistry.

We expect that another challenge in entering this field is the broad range of knowledge required. In retrosynthetic planning, reaction prediction and automation, there is wide variability in the methods used. In the coming years, we expect that the community will come to a consensus on which of the various methods to build upon to improve and demonstrate their value rather than building new tools from scratch. Therefore, fair comparisons of different approaches are needed to assess the strengths, shortcomings and best uses of these technologies.

Ultimately, information-rich techniques will become part of the practising synthetic chemist's toolkit towards the dream of realizing the synthesis of any given molecule from route prediction to preparation of target compounds in high yield with minimal human effort. In reality, we are far from realizing this vision, and it is unlikely to ever be realized for all of chemical space. However, a system that could produce terpene,

alkaloid and polyketide natural products with the ease that peptides and nucleotides are currently synthesized automatically would be a new paradigm for the field. Incrementally, we will continue to work towards this goal by developing new tools, standardizing those currently available, educating, and streamlining the inefficiencies of synthesis to allow chemists to focus on new and creative aspects of synthesis.

Published online: 18 March 2021

1. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
2. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **166**, 178 (1969).
3. Hammett, L. P. *Physical Organic Chemistry; Reaction Rates, Equilibria, and Mechanisms* 1st edn (McGraw-Hill, 1940).
4. Brønsted, J. N. & Pedersen, K. J. Die katalytische Zersetzung des Nitramids und ihre physikalisch-chemische Bedeutung [German]. *Zeitschrift für Phys. Chemie Stochiometrie und Verwandtschaftslehre* **108**, 185–235 (1924).
5. Merrifield, R. B., Stewart, J. M. & Jernberg, N. Instrument for automated synthesis of peptides. *Anal. Chem.* **38**, 1905–1914 (1966).
6. Merrifield, R. B. in *Hypotensive Peptides* 1–13 (Springer, 1966).
7. Evans, D. A. History of the Harvard ChemDraw project. *Angew. Chem. Int. Ed.* **53**, 11140–11145 (2014).
8. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
9. Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **34**, 247–266 (2005).
10. Ihlenfeldt, W.-D. & Gasteiger, J. Computer-assisted planning of organic syntheses: the second generation of programs. *Angew. Chem. Int. Ed. Engl.* **34**, 2613–2633 (1996).
11. Cook, A. et al. Computer-aided synthesis design: 40 years on. *Wiley Iinterdiscip. Rev. Comput. Mol. Sci.* **2**, 79–107 (2012).
12. Ravitz, O. Data-driven computer aided synthesis design. *Drug Discov. Today Technol.* **10**, e443–e449 (2013).
13. Engkvist, O. et al. Computational prediction of chemical reactions: current status and outlook. *Drug Discov. Today* **23**, 1203–1218 (2018).
14. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
15. Johansson, S. et al. AI-assisted synthesis prediction. *Drug Discov. Today Technol.* **32-33**, 65–72 (2019).
16. Zahrt, A. F., Athavale, S. V. & Denmark, S. E. Quantitative structure–selectivity relationships in enantioselective catalysis: past, present, and future. *Chem. Rev.* **120**, 1620–1689 (2020).
17. Strieth-Kalthoff, F., Sandfort, F., Segler, M. H. S. & Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **49**, 6154–6168 (2020).
18. Reid, J. P. & Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2**, 290–305 (2018).
19. de Almeida, A. F., Moreira, R. & Rodrigues, T. Synthetic organic chemistry driven by artificial intelligence. *Nat. Rev. Chem.* **3**, 589–604 (2019).
20. Shevlin, M. Practical high-throughput experimentation for chemists. *ACS Med. Chem. Lett.* **8**, 601–607 (2017).
21. Mennen, S. M. et al. The evolution of high-throughput experimentation in pharmaceutical development and perspectives on the future. *Org. Process. Res. Dev.* **23**, 1213–1242 (2019).
22. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97 (2018).
23. Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The evolution of chemical high-throughput experimentation to address challenging problems in pharmaceutical synthesis. *Acc. Chem. Res.* **50**, 2976–2985 (2017).
24. Welch, C. J. High throughput analysis enables high throughput experimentation in pharmaceutical process research. *React. Chem. Eng.* **4**, 1895–1911 (2019).
25. Allen, C. L., Leitch, D. C., Anson, M. S. & Zajac, M. A. The power and accessibility of high-throughput methods for catalysis research. *Nat. Catal.* **2**, 2–4 (2019).
26. Vlěduts, G. Ě. Concerning one system of classification and codification of organic reactions. *Inform. Stor. Retr.* **1**, 117–146 (1963).
27. Ugi, I. et al. Models, concepts, theories, and formal languages in chemistry and their use as a basis for computer assistance in chemistry. *J. Chem. Inf. Comput. Sci.* **34**, 3–16 (1994).
28. Ugi, I. et al. Computer-assisted solution of chemical problems — the historical development and the present state of the art of a new discipline of chemistry. *Angew. Chem. Int. Ed. Engl.* **32**, 201–227 (1993).
29. Corey, E. J. *The Logic of Chemical Synthesis* (Nobel Foundation, [Nobelstiftelsen], 1991).
30. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
31. Pensak, D. A. & Corey, E. J. in *Computer-Assisted Organic Synthesis* Vol. 61 Ch. 1 1–32 (American Chemical Society, 1977).
32. Campbell, M., Hoane, A. J. & Hsu, F.-H. Deep Blue. *Artif. Intell.* **134**, 57–83 (2002).
33. Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
34. Hanessian, S., Franco, J. & Larouche, B. The psychobiological basis of heuristic synthesis planning, man, machine, and the chiron aproach. *Pure Appl. Chem.* **62**, 1887–1910 (1990).
35. Wipke, W. T. & Rogers, D. Artificial intelligence in organic synthesis. SST: starting material selection strategies. An application of superstructure search. *J. Chem. Inf. Comput. Sci.* **24**, 71–81 (1984).
36. Mehta, G., Barone, R. & Chanon, M. Computer-aided organic synthesis — SESAM: a simple program to unravel "hidden" restructured starting materials skeleta in complex targets. *Eur. J. Org. Chem.* **1998**, 1409–1412 (1998).
37. Corey, E. J., Long, A. K. & Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **228**, 408 (1985).
38. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
39. Law, J. et al. Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **49**, 593–602 (2009).
40. Christ, C. D., Zentgraf, M. & Kriegl, J. M. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* **52**, 1745–1756 (2012).
41. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
42. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**, 1237–1245 (2017).
43. Segler, M. H. S. & Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chem. Eur. J.* **23**, 6118–6128 (2017).
44. Baylon, J. L., Cilfone, N. A., Gulcher, J. R. & Chittenden, T. W. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model.* **59**, 673–688 (2019).
45. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
46. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
47. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
48. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
49. Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**, 3355–3364 (2020).
50. Karpov, P., Godin, G. & Tetko, I. V. in *Artificial Neural Networks and Machine Learning — ICANN 2019: Workshop and Special Sessions* (eds Kůrková, V., Karpov, P. & Theis, F.) 817–830 (Springer International, 2019).
51. Schwaller, P. et al. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).
52. Somnath, V. R., Bunne, C., Coley, C. W., Krause, A. & Barzilay, R. Learning graph models for template-free retrosynthesis. Preprint at https://arxiv.org/abs/2006.07038 (2020).
53. Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P. & Jastrzębski, S. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. Preprint at https://arxiv.org/abs/2006.15426 (2020).
54. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604 https://www.nature.com/articles/nature25978#supplementary-information (2018).
55. Segler, M., Preuß, M. & Waller, M. P. Towards "AlphaChem": chemical synthesis planning with tree search and deep neural network policies. Preprint at https://arxiv.org/abs/1702.00020 (2017).
56. Bertz, S. H. The first general index of molecular complexity. *J. Am. Chem. Soc.* **103**, 3599–3601 (1981).
57. Huang, Q., Li, L.-L. & Yang, S.-Y. RASA: a rapid retrosynthesis-based scoring method for the assessment of synthetic accessibility of drug-like molecules. *J. Chem. Inf. Model.* **51**, 2768–2777 (2011).
58. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
59. Gasteiger, J. et al. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspect. Drug Discov. Des.* **20**, 245–264 (2000).
60. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**, 434–443 (2017).
61. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2**, 725–732 (2016).
62. Rosales, A. R. et al. Rapid virtual screening of enantioselective catalysts using CatVS. *Nat. Catal.* **2**, 41–45 (2019).
63. Burai Patrascu, M. et al. From desktop to benchtop with automated computational workflows for computer-aided design in asymmetric catalysis. *Nat. Catal.* **3**, 574–584 (2020).
64. Marcou, G. et al. Expert system for predicting reaction conditions: the Michael reaction case. *J. Chem. Inf. Model.* **55**, 239–250 (2015).
65. Walker, E. et al. Learning to predict reaction conditions: relationships between solvent, molecular structure, and catalyst. *J. Chem. Inf. Model.* **59**, 3645–3654 (2019).
66. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
67. Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **59**, 4385–4402 (2016).
68. Jia, X. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
69. Mehr, S. H. M., Craven, M. S., Leonov, A. I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101 (2020).
70. Martin, T. M. et al. Does rational selection of training and test sets improve the outcome of QSAR

modeling? *J. Chem. Inf. Model.* **52**, 2570–2578 (2012).
71. Murray, P. M. & Forfar, L. C. The application of advanced design of experiments for the efficient development of chemical processes. *Chem. Inform.* https://doi.org/10.21767/2470-6973.100023 (2017).
72. Luque Ruiz, I., Cerruela Garcī a, G. & G ómez-Nieto, M. Ā. in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* Ch. 7 (eds Varmuza, K., Dehmer, M. & Bonchev, D.) 201–228 (Wiley, 2012).
73. Zahrt, A. F. et al. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* https://doi.org/10.1126/science.aau5631 (2019).
74. Henle, J. J. et al. Development of a computer-guided workflow for catalyst optimization. descriptor validation, subset selection, and training set analysis. *J. Am. Chem. Soc.* **142**, 11578–11592 (2020).
75. Zhao, S. et al. Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* **362**, 670 (2018).
76. Woods, B. P., Orlandi, M., Huang, C. Y., Sigman, M. S. & Doyle, A. G. Nickel-catalyzed enantioselective reductive cross-coupling of styrenyl aziridines. *J. Am. Chem. Soc.* **139**, 5688–5691 (2017).
77. Gao, H. et al. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **4**, 1465–1476 (2018).
78. Lin, A. I. et al. Automatized assessment of protective group reactivity: a step toward big reaction data analysis. *J. Chem. Inf. Model.* **56**, 2140–2148 (2016).
79. Casari, A. & Zheng, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists* 1st edn (O'Reilly Media, 2018).
80. Granda, J. M., Donina, L., Dragone, V., Long, D. L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
81. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **12**, 56 (2020).
82. Cherkasov, A. et al. QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**, 4977–5010 (2014).
83. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
84. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminform.* **10**, 4 (2018).
85. Merkwirth, C. & Lengauer, T. Automatic generation of complementary descriptors with molecular graph networks. *J. Chem. Inf. Model.* **45**, 1159–1168 (2005).
86. Duvenaud, D. K. et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv. Neural Inf. Process. Syst.* **28**, 2224–2232 (2015).
87. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Mol. Des.* **30**, 595–608 (2016).
88. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
89. Brethomé, A. V., Fletcher, S. P. & Paton, R. S. Conformational effects on physical-organic descriptors: the case of Sterimol steric parameters. *ACS Catal.* **9**, 2313–2323 (2019).
90. Harper, K. C., Bess, E. N. & Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.* **4**, 366–374 (2012).
91. Clavier, H. & Nolan, S. P. Percent buried volume for phosphine and N-heterocyclic carbene ligands: steric properties in organometallic chemistry. *Chem. Commun.* **46**, 841–861 (2010).
92. Hillier, A. C. et al. A combined experimental and theoretical study examining the binding of N-heterocyclic carbenes (NHC) to the Cp*RuCl (Cp* = η^5-C_5Me_5) moiety: insight into stereoelectronic differences between unsaturated and saturated NHC ligands. *Organometallics* **22**, 4322–4326 (2003).
93. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
94. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
95. Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
96. Li, X., Zhang, S. Q., Xu, L. C. & Hong, X. Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew. Chem. Int. Ed.* **59**, 13253–13259 (2020).
97. Chan, W. & White, P. *Fmoc Solid Phase Peptide Synthesis: a Practical Approach* Vol. 222 (OUP Oxford, 1999).
98. Seeberger, P. H. Automated oligosaccharide synthesis. *Chem. Soc. Rev.* **37**, 19–28 (2008).
99. Kaplan, B. E. The automated synthesis of oligodeoxyribonucleotides. *Trends Biotechnol.* **3**, 253–256 (1985).
100. Cernak, T. et al. Microscale high-throughput experimentation as an enabling technology in drug discovery: application in the discovery of (piperidinyl)pyridinyl-1*H*-benzimidazole diacylglycerol acyltransferase 1 inhibitors. *J. Med. Chem.* **60**, 3594–3605 (2017).
101. Hook, A. L. et al. High throughput methods applied in biomaterial development and discovery. *Biomaterials* **31**, 187–198 (2010).
102. Yan, Y., Robinson, S. G., Sigman, M. S. & Sanford, M. S. Mechanism-based design of a high-potential catholyte enables a 3.2 V all-organic nonaqueous redox flow battery. *J. Am. Chem. Soc.* **141**, 15301–15306 (2019).
103. Francis, M. B. & Jacobsen, E. N. Discovery of novel catalysts for alkene epoxidation from metal-binding combinatorial libraries. *Angew. Chem. Int. Ed.* **38**, 937–941 (1999).
104. Taylor, S. J. & Morken, J. P. Thermographic selection of effective catalysts from an encoded polymer-bound library. *Science* **280**, 267–270 (1998).
105. Kölmel, D. K., Loach, R. P., Knauber, T. & Flanagan, M. E. Employing photoredox catalysis for DNA-encoded chemistry: decarboxylative alkylation of α-amino acids. *ChemMedChem* **13**, 2159–2165 (2018).
106. Geri, J. B. et al. Microenvironment mapping via Dexter energy transfer on immune cells. *Science* **367**, 1091–1097 (2020).
107. Bellomo, A. et al. Rapid catalyst identification for the synthesis of the pyrimidinone core of HIV integrase inhibitors. *Angew. Chem. Int. Ed.* **51**, 6912–6915 (2012).
108. Dreher, S. D., Dormer, P. G., Sandrock, D. L. & Molander, G. A. Efficient cross-coupling of secondary alkyltrifluoroborates with aryl chlorides — reaction discovery using parallel microscale experimentation. *J. Am. Chem. Soc.* **130**, 9257–9259 (2008).
109. Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49 (2015).
110. Perera, D. et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429 (2018).
111. Shaabani, S. et al. Automated and accelerated synthesis of indole derivatives on a nano-scale. *Green Chem.* **21**, 225–232 (2019).
112. Trobe, M. & Burke, M. D. The molecular industrial revolution: automated synthesis of small molecules. *Angew. Chem. Int. Ed.* **57**, 4192–4214 (2018).
113. Wong, H. & Cernak, T. Reaction miniaturization in eco-friendly solvents. *Curr. Opin. Green Sustain. Chem.* **11**, 91–98 (2018).
114. Wang, Y. et al. Acoustic droplet ejection enabled automated reaction scouting. *ACS Cent. Sci.* **5**, 451–457 (2019).
115. Boga, S. B. et al. Selective functionalization of complex heterocycles via an automated strong base screening platform. *React. Chem. Eng.* **2**, 446–450 (2017).
116. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
117. Lee, G. M., Clément, R. & Baker, R. T. High-throughput evaluation of in situ-generated cobalt (III) catalysts for acyl fluoride synthesis. *Catal. Sci. Technol.* **7**, 4996–5003 (2017).
118. Qiu, J., Albrecht, J. & Janey, J. Solubility behaviors and correlations of common organic solvents. *Org. Process. Res. Dev.* **24**, 2702–2708 (2020).
119. Christensen, M. et al. Data-science driven autonomous process optimization. Preprint at https://doi.org/10.26434/chemrxiv.13146404.v2 (2020).
120. Lin, S. et al. Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236 (2018).
121. Uehling, M. R., King, R. P., Krska, S. W., Cernak, T. & Buchwald, S. L. Pharmaceutical diversification via palladium oxidative addition complexes. *Science* **363**, 405 (2019).
122. Gesmundo, N. J. et al. Nanoscale synthesis and affinity ranking. *Nature* **557**, 228–232 (2018).
123. Bahr, M. N. et al. Collaborative evaluation of commercially available automated powder dispensing platforms for high-throughput experimentation in pharmaceutical applications. *Org. Process Res. Dev.* **22**, 1500–1508 (2018).
124. Martin, M. C. et al. Versatile methods to dispense submilligram quantities of solids using chemical-coated beads for high-throughput experimentation. *Org. Process Res. Dev.* **23**, 1900–1907 (2019).
125. Tu, N. P. et al. High-throughput reaction screening with nanomoles of solid reagents coated on glass beads. *Angew. Chem. Int. Ed.* **58**, 7987–7991 (2019).
126. Coley, C. W. et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* https://doi.org/10.1126/science.aax1566 (2019).
127. Noel, T. et al. Palladium-catalyzed amination reactions in flow: overcoming the challenges of clogging via acoustic irradiation. *Chem. Sci.* **2**, 287–290 (2011).
128. Boele, M. D. K. et al. Selective Pd-catalyzed oxidative coupling of anilides with olefins through C–H bond activation at room temperature. *J. Am. Chem. Soc.* **124**, 1586–1587 (2002).
129. McMullen, J. P., Stone, M. T., Buchwald, S. L. & Jensen, K. F. An integrated microreactor system for self-optimization of a heck reaction: from micro- to mesoscale flow systems. *Angew. Chem. Int. Ed.* **49**, 7076–7080 (2010).
130. Zhang, J., Bellomo, A., Creamer, A. D., Dreher, S. D. & Walsh, P. J. Palladium-catalyzed C(sp^3)–H arylation of diarylmethanes at room temperature: synthesis of triarylmethanes via deprotonative-cross-coupling processes. *J. Am. Chem. Soc.* **134**, 13765–13772 (2012).
131. Reizman, B. J., Wang, Y.-M., Buchwald, S. L. & Jensen, K. F. Suzuki–Miyaura cross-coupling optimization enabled by automated feedback. *React. Chem. Eng.* **1**, 658–666 (2016).
132. Kashani, S. K., Jessiman, J. E. & Newman, S. G. Exploring homogeneous conditions for mild Buchwald–Hartwig amination in batch and flow. *Org. Process Res. Dev.* **24**, 1948–1954 (2020).
133. Boström, J., Brown, D. G., Young, R. J. & Keserü, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709–727 (2018).
134. Twilton, J. et al. Selective hydrogen atom abstraction through induced bond polarization: direct α-arylation of alcohols through photoredox, HAT, and nickel catalysis. *Angew. Chem. Int. Ed. Engl.* **57**, 5369–5373 (2018).
135. Dirocco, D. A. et al. Late-stage functionalization of biologically active heterocycles through photoredox catalysis. *Angew. Chem. Int. Ed. Engl.* **53**, 4802–4806 (2014).
136. Mo, Y., Rughoobur, G., Nambiar, A. M. K., Zhang, K. & Jensen, K. F. A multifunctional microfluidic platform for high-throughput experimentation of electroorganic chemistry. *Angew. Chem. Int. Ed.* **59**, 20890–20894 (2020).
137. Deadman, B. J., Collins, S. G. & Maguire, A. R. Taming hazardous chemistry in flow: the continuous processing of diazo and diazonium compounds. *Chemistry* **21**, 2298–2308 (2015).
138. Movsisyan, M. et al. Taming hazardous chemistry by continuous flow technology. *Chem. Soc. Rev.* **45**, 4892–4928 (2016).
139. Selekman, J. A. et al. High-throughput automation in chemical process development. *Annu. Rev. Chem. Biomol. Eng.* **8**, 525–547 (2017).
140. Hwang, Y. J. et al. A segmented flow platform for on-demand medicinal chemistry and compound synthesis in oscillating droplets. *Chem. Commun.* **53**, 6649–6652 (2017).
141. Reker, D., Hoyt, E. A., Bernardes, G. J. L. & Rodrigues, T. Adaptive optimization of chemical reactions with minimal experimental information. *Cell Rep. Phys. Sci.* **1**, 100247 (2020).
142. Jiang, T. et al. An integrated console for capsule-based, fully automated organic synthesis. Preprint at https://doi.org/10.26434/chemrxiv.7882799.v1 (2019).
143. Wang, C. & Glorius, F. Controlled iterative cross-coupling: on the way to the automation of organic synthesis. *Angew. Chem. Int. Ed.* **48**, 5240–5244 (2009).

144. Steiner, S. et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).
145. Collins, N. et al. Fully automated chemical synthesis: toward the universal synthesizer. *Org. Process Res. Dev.* **24**, 2064–2077 (2020).
146. Wanner, B. M., Nichols, P. L. & Jiang, T. Cartridge-based automated synthesis — a new tool for the synthetic chemist. *Chimia* **74**, 808–813 (2020).
147. Gillis, E. P. & Burke, M. D. Multistep synthesis of complex boronic acids from simple MIDA boronates. *J. Am. Chem. Soc.* **130**, 14084–14085 (2008).
148. Li, J., Grillo, A. S. & Burke, M. D. From synthesis to function via iterative assembly of N-methyliminodiacetic acid boronate building blocks. *Acc. Chem. Res.* **48**, 2297–2307 (2015).
149. Sun, S. & Kennedy, R. T. Droplet electrospray ionization mass spectrometry for high throughput screening for enzyme inhibitors. *Anal. Chem.* **86**, 9309–9314 (2014).
150. Doi, T. et al. A formal total synthesis of taxol aided by an automated synthesizer. *Chem. Asian J.* **1**, 370–383 (2006).
151. Roughley, S. D. & Jordan, A. M. The medicinal chemist's toolbox: an analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **54**, 3451–3479 (2011).
152. Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
153. Hsieh, H.-W., Coley, C. W., Baumgartner, L. M., Jensen, K. F. & Robinson, R. I. Photoredox iridium–nickel dual-catalyzed decarboxylative arylation cross-coupling: from batch to continuous flow via self-optimizing segmented flow reactor. *Org. Process Res. Dev.* **22**, 542–550 (2018).
154. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
155. Mahjour, B., Shen, Y., Liu, W. & Cernak, T. A map of the amine–carboxylic acid coupling system. *Nature* **580**, 71–75 (2020).
156. Roch, L. M. et al. ChemOS: an orchestration software to democratize autonomous discovery. *PLoS ONE* **15**, e0229862 (2020).
157. Pendleton, I. M. et al. Experiment specification, capture and laboratory automation technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* **9**, 846–859 (2019).
158. Marth, C. J. et al. Network-analysis-guided synthesis of weisaconitine D and liljestrandinine. *Nature* **528**, 493 (2015).
159. Schwaller, P., Gaudin, T., Lányi, D., Bekas, C. & Laino, T. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
160. Coley, Connor W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
161. Mikulak-Klucznik, B. et al. Computational planning of the synthesis of complex natural products. *Nature* **588**, 83–88 (2020).
162. Alexander, D. L. J., Tropsha, A. & Winkler, D. A. Beware of R^2: simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *J. Chem. Inf. Model.* **55**, 1316–1322 (2015).
163. Chung, R. & Hein, J. E. Automated solubility and crystallization analysis of non-UV active compounds: integration of evaporative light scattering detection (ELSD) and robotic sampling. *React. Chem. Eng.* **4**, 1674–1681 (2019).
164. Baranczak, A. et al. Integrated platform for expedited synthesis–purification–testing of small molecule libraries. *ACS Med. Chem. Lett.* **8**, 461–465 (2017).
165. Hoogenboom, R., Wiesbrock, F., Leenen, M. A. M., Meier, M. A. R. & Schubert, U. S. Accelerating the living polymerization of 2-nonyl-2-oxazoline by implementing a microwave synthesizer into a high-throughput experimentation workflow. *J. Comb. Chem.* **7**, 10–13 (2005).
166. Troshin, K. & Hartwig, J. F. Snap deconvolution: an informatics approach to high-throughput discovery of catalytic reactions. *Science* **357**, 175 (2017).
167. McNally, A., Prier, C. K. & MacMillan, D. W. C. Discovery of an α-amino C–H arylation reaction using the strategy of accelerated serendipity. *Science* **334**, 1114 (2011).
168. Johnson, A. P., Marshall, C. & Judson, P. N. Some recent progress in the development of the LHASA computer system for organic synthesis design: starting-material-oriented retrosynthetic analysis. *Recl. Trav. Chim. Pays Bas* **111**, 310–316 (1992).
169. Snider, B. B. & Kulkarni, Y. S. Preparation of unsaturated. α.-chloro acids and intramolecular [2 + 2] cycloadditions of the chloroketenes derived from them. *J. Org. Chem.* **52**, 307–310 (1987).
170. Schwaller, P. et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
171. Genheden, S. et al. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **12**, 70 (2020).
172. Nicolaou, C. A., Watson, I. A., LeMasters, M., Masquelin, T. & Wang, J. Context aware data-driven retrosynthetic analysis. *J. Chem. Inf. Model.* **60**, 2728–2738 (2020).
173. Bøgevig, A. et al. Route design in the 21st century: the ICSYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* **19**, 357–368 (2015).
174. Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379–1390 (2020).
175. Mirő, J. et al. Enantioselective allenoate-Claisen rearrangement using chiral phosphate catalysts. *J. Am. Chem. Soc.* **142**, 6390–6399 (2020).
176. Collins, K. D. & Glorius, F. Intermolecular reaction screening as a tool for reaction evaluation. *Acc. Chem. Res.* **48**, 619–627 (2015).
177. Yayla, H. G. et al. Discovery and mechanistic study of a photocatalytic indoline dehydrogenation for the synthesis of elbasvir. *Chem. Sci.* **7**, 2066–2073 (2016).
178. Vaucher, A. C. et al. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).

Acknowledgements
A.G.D., R.S., M.A.H. and J.E.B. were supported by the National Science Foundation (NSF) under the Center for Computer Aided Synthesis (C-CAS) (CHE-1925607). M.A.H. is grateful for funding from the NSF graduate research fellowship program (DGE-1752814). Y.S. and T.C. were supported by the University of Michigan College of Pharmacy.

Author contributions
Introduction (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Experimentation (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Results (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Applications (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Reproducibility and data deposition (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Limitations and optimizations (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Outlook (Y.S., J.E.B., M.A.H., R.S., A.G.D. and T.C.); Overview of the Primer (T.C.).

Competing interests
T.C. has received mosquito robotics from SPT Labtech and Merck & Co., Inc. T.C. and R.S. receive research support from MilliporeSigma, the company that owns the retrosynthetic software SYNTHIA. All other authors declare no competing interests.

Peer review information
Nature Reviews Methods Primers thanks O. Ravitz, M. Segler, S. Trice and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RELATED LINKS
AiZynthFinder: https://github.com/MolecularAI/aizynthfinder
ASKCOS: https://github.com/connorcoley/ASKCOS
Chemical.AI: https://Chemical.AI
IBM RXN for Chemistry: https://rxn.res.ibm.com/
ICSYNTH: https://www.deepmatter.io/products/icsynth/
Iktos spaya.ai: https://beta.spaya.ai/
RDKit: https://www.rdkit.org/
Reaxys: https://www.elsevier.com/solutions/reaxys/features-and-capabilities/synthesis-planner
SciFinder^n: https://www.cas.org/products/scifinder
SciKit-learn: https://scikit-learn.org/stable/
SYNTHIA: https://www.sigmaaldrich.com/chemistry/chemical-synthesis/synthesis-software.html

© Springer Nature Limited 2021