How to explore chemical space using algorithms and automation

Piotr S. Gromski, Alon B. Henson, Jarosław M. Granda and Leroy Cronin

Abstract | Although extending the reactivity of a given class of molecules is relatively straightforward, the discovery of genuinely new reactivity and the molecules that result is a wholly more challenging problem. If new reactions can be considered unpredictable using current chemical knowledge, then we suggest that they are not merely new but also novel. Such a classification, however, requires an expert judge to have access to all current chemical knowledge or risks a lack of information being interpreted as unpredictability. Here, we describe how searching chemical space using automation and algorithms improves the probability of discovery. The former enables routine chemical tasks to be performed more quickly and consistently, while the latter uses algorithms to facilitate the searching of chemical knowledge databases. Experimental systems can also be developed to discover novel molecules, reactions and mechanisms by augmenting the intuition of the human expert. In order to find new chemical laws, we must seek to question current assumptions and biases. Accomplishing that involves using two areas of algorithmic approaches: algorithms to perform searches, and more general machine learning and statistical modelling algorithms to predict the chemistry under investigation. We propose that such a chemical intelligence approach is already being used and that, in the not-too-distant future, the automated chemical reactor systems controlled by these algorithms and monitored by a sensor array will be capable of navigating and searching chemical space more quickly, efficiently and, importantly, without bias. This approach promises to yield not only new molecules but also unpredictable and thus novel reactivity.

Algorithms in chemistry

The expansion of chemical knowledge by searching for new molecules and chemical reactions is an inherently practical endeavour, but it is increasingly becoming possible to conduct these searches with the help of computer algorithms[1,2]. Simply put, an algorithm is a process or set of rules to be followed in calculations or other problem-solving operations, and these rules can be formulated into instructions to be used by a computer. For instance, the molecular structures of a set of molecules could be placed into a database using a digital representation (for example, text-based SMILES or coordinates) and then investigated in terms of their similarity, calculated properties or potential reactivity (for example, suggesting new reactions based upon the functional groups present). In the synthesis of new molecules, it is common practice to use single-parameter optimization (changing the temperature, concentration, solvent polarity or other reaction parameters only one at a time) in order to find the best outcome, and this approach also spans the quest for new reactivity[3,4]. Despite this, synthetic chemists are wary of computer-based searches and prediction because searching through the vast number of possible molecules requires a very high level of expertise[5,6]. In this Perspective, we describe how machine learning, coupled with real-time chemistry[7,8], is set to change the way chemists discover molecules, reactions and reactivity as well as remove researcher bias. The first step in applying machine learning techniques in chemistry is to state the problem using a sequence of logical operations that can be built into an algorithm. Machine learning is a statistical approach that allows computers to learn patterns using provided training data and then make predictions on the basis of data not seen before[9]. Defining the problem using a flow chart of the chemical investigation (by identifying the inputs, process operations and outputs) can help with the development of a definition that expresses the problem algorithmically. This reframing opens many doors for using well-defined algorithms and other statistical approaches, with the goal of increasing the chances of discovery of new chemical knowledge.

Chemical space includes the entire universe of known and unknown molecules, as well as the transformations required to produce these molecules, and their relationship to each other, that is, a network of chemical reactions connecting the molecules in the space[10]. One of the primary difficulties of exploration is that chemical space is both large and sparse. Many parts of it contain only a few molecules, while other parts have highly dense clusters of molecules[11–14]. Furthermore, chemistry costs both time and physical resources, resulting in a limit on the number of experiments that can be conducted and the speed at which they can be safely performed. To alleviate these constraints, these characteristics of chemical space should be taken into account to design search approaches that are tailored for the application in mind. For instance, a self-imposed limitation of working on only a part of chemical space that contains a clustered, dense region is an intuitive approach. To use algorithms while taking advantage of the characteristics of chemical space, we must first consider how each possible point in chemical space has properties that can be measured (FIG. 1a). These properties could be directly linked to a molecule, for example, its molecular weight or other molecular characteristic, or could be an indirect probe, such as the colour change associated with a reaction. The parameters that define a chemical system — which could be a single-chemical reaction, a system of reactions or a supramolecular system — are the reaction inputs (reagents, catalysts and solvents) and the process conditions (addition rates, concentrations, temperature and time). By performing experiments and probing the possible points in chemical space, raw data are collected and subsequently transformed into specific desirable outputs. The relationship between the chemical outcome and the states in the search space, here called the mapping or utility function, can be very complex, but it simplifies the raw data into easily comparable values, for example, molecular weight, a change[15] in IR peak intensity[16] or a wavelength shift for the UV/vis peak wavelength[17]. To explore the search space effectively, a search algorithm must use the data collected from this space to then choose a new set of input parameters for the next experiment. This process is not dissimilar to how a human experimenter goes about collecting data and exploring a chemical system.

As an example of a simple use of chemical spaces and a mapping function, let us consider the mechanism-aided discovery of photocatalytic reactivity reported by Glorius et al.[18] (see FIG. 1b). Considering the mechanism of the reaction — excitation of a photocatalyst followed by quenching of the excited state by the reaction substrate — enabled the use of the quenching percentage of the photocatalyst excited state as the mapping function. Thus, rather than seek new reactivity by detecting a particular

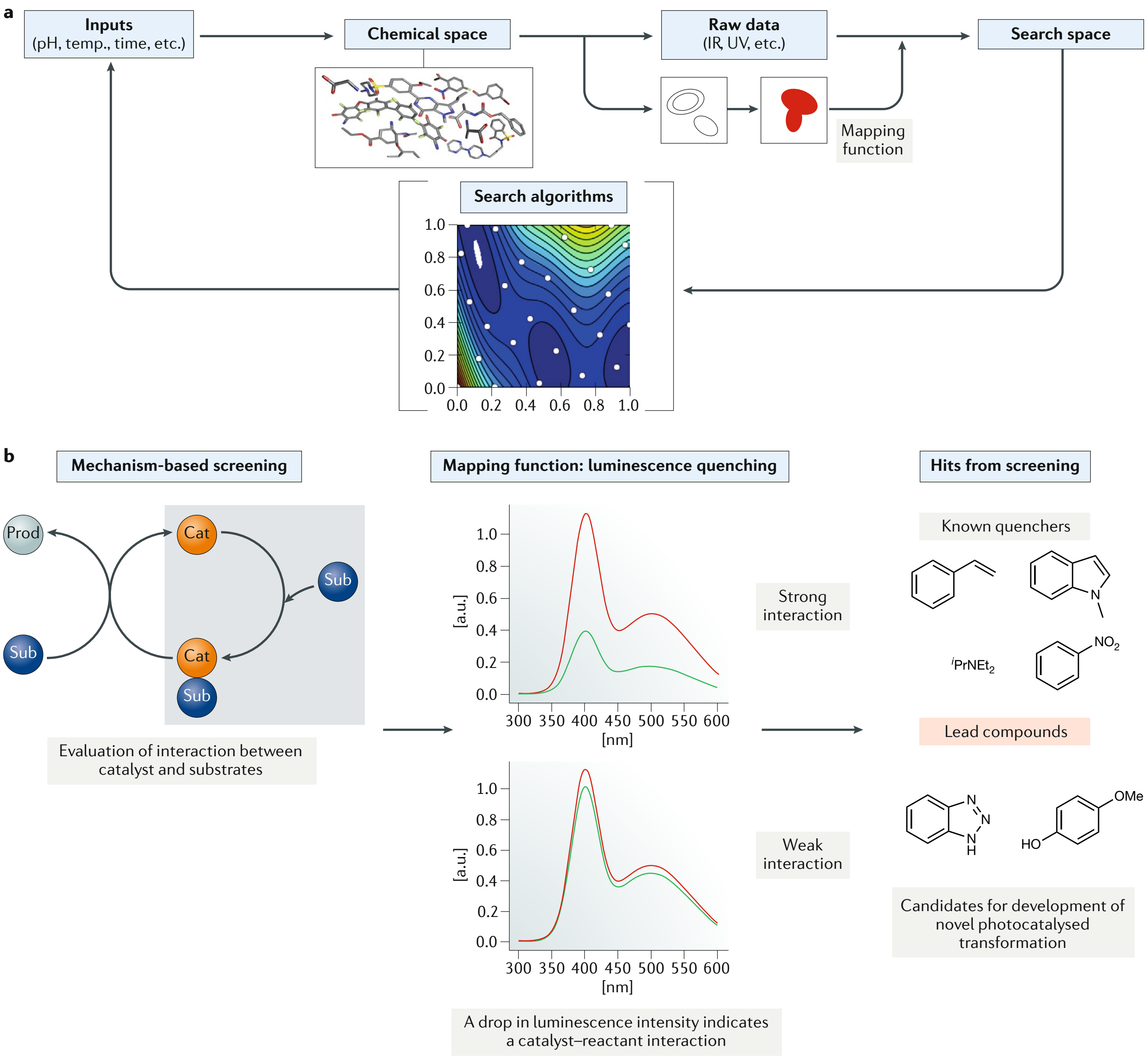


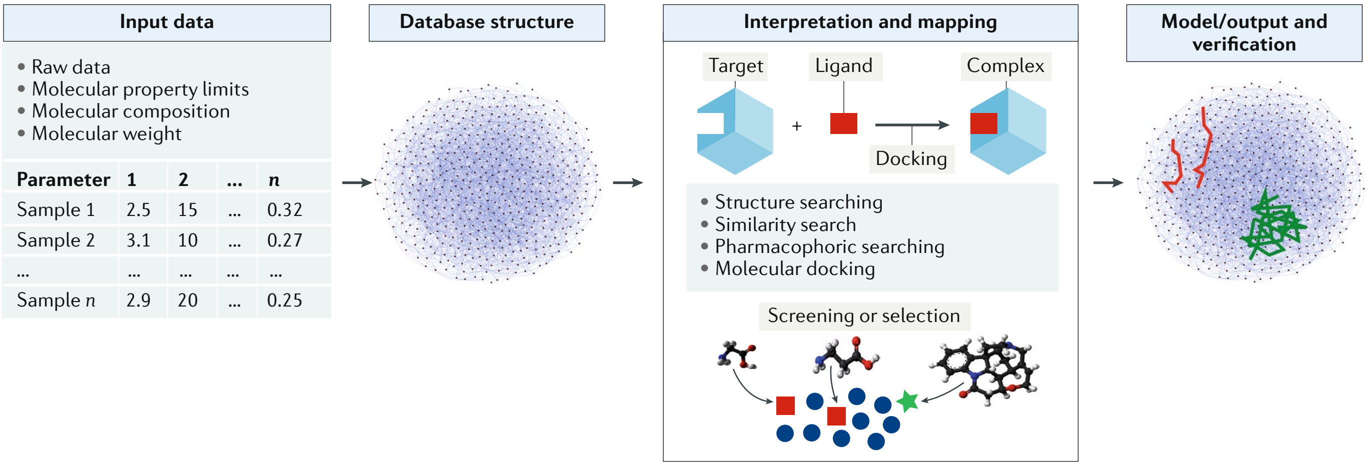
Fig. 1 | **Searching chemical space. a** | The schematic shows how the input parameters constrain chemical space to a limited section, which in turn is framed as a search space using a mapping function. Feedback from measurements can then be used by the algorithm to assess the outcome of the reaction for use in the search algorithm. **b** | An example using fluorescence-quenching efficiency as a mapping function to identify strong interactions between the catalyst and substrate in photocatalysed reactions. The left panel shows the underlying chemical mechanism, with the part relevant for the mapping function presented on a grey background showing the catalyst (Cat) and interaction with a substrate (Sub or Sub2), and an example of a photocatalyst used. From this information, a mapping function is constructed that assesses the level of interaction from the fluorescence quenching. In the right panel, there are several hits from the experimental search, including previously known and new quenchers. IR, infrared; Prod, product; Temp, temperature. Part **b** is adapted with permission from REF.[18], Wiley-VCH.

reaction product, the method detects the interaction strength between an organic substrate and a photocatalyst by evaluating the luminescence quenching. This approach helped in the discovery of two promising substrate classes after screening only 100 compounds, because the excited state quenching step in the mechanism determines the reaction efficiency. The approach focuses on a single-constituent mechanistic step by defining the mapping function as the quenching percentage — an indication of how effectively energy is transferred from the excited photocatalyst to organic molecules. Here, the selection of reactants at random defines the chemical space. Collecting the luminescence emission spectra from these reactions produces a search space in which the degree of quenching can be used to identify substrates of interest. Applying this approach to a wide variety of substrates in a secondary step accelerated the discovery of new substrates for the development of photocatalysed reactions from a large array (~100), as well as yielding mechanistic insights. Several classes of organic molecules were identified as effective quenchers, among them 1H-benzotriazole and 4-methoxyphenol. This example illustrates how selection of the right assay and acceptance criteria can enable algorithmically driven systems to explore chemical space.

Building databases

Although it may be desirable to explore only a small part of chemical space, it remains useful to have access to as large a body of chemical knowledge as possible from which to choose. Chemical databases today contain vast amounts of accessible data, though not all are freely accessible. Their benefits include a large number of data points containing many forms of chemical information[2], including reactivity, analytical data and sometimes metadata (information about the data). One key question is: how can knowledge of what reactions and molecules exist help us to discover novel molecules or reactions that are not predicted by the database entries? First, they can help us pinpoint the 'known-unknowns', worthy of research in their own right, but which will help when looking for 'unknown-unknowns'. Simply put, a good database can tell us what has not been explored as well as what has been. However, most databases currently do not contain negative results, nor do they offer an evaluation of the uncertainty of the source data, although multiple examples of the same chemical reactivity would allow a user to assess its validity[19]. The scale and diversity of the available data indicate that using databases for chemical discovery is not straightforward and depends on the use of suitable methodologies such as virtual screening, machine learning and statistics[20].

The two main methods used to build and explore chemical databases include virtual screening and molecular design (see FIG. 2). Virtual screening encompasses the use of computational algorithms and models to identify the properties and activity of molecules[21]. Virtual screening can also be combined with experimental work, such as high-throughput screening, to perform more cost-effective research[22,23]. By contrast, molecular design approaches the problem from the opposite perspective — using stored chemical data to select molecules that are likely to fit specific design criteria. Using these approaches, learning algorithms that can improve or predict new properties become practical and useful. While both virtual screening and molecular design are similar in that they search within databases of chemical information, they use different input parameters and mapping functions, as they have different goals. A substantial difference is that virtual screening does not allow for the prediction of the unknown. The stores of chemical data available today coupled with the variety of available algorithms can help in exploration and making discoveries that are inside the collected data[24]. Other approaches that use the chemical data in databases are explored in the field of chemometrics. Chemometrics uses a broad range of mathematical and statistical methods, on databases and live data, to improve the understanding of chemical systems and to correlate parameters or physical properties with instrumental data[25,26]. As an example, the use of multivariate linear regression models utilizing computationally and empirically derived physical organic molecular descriptors has been described by Santiago et al.[27], along with a discussion about methodological approaches from the perspective of reaction optimization and mechanistic interrogation.



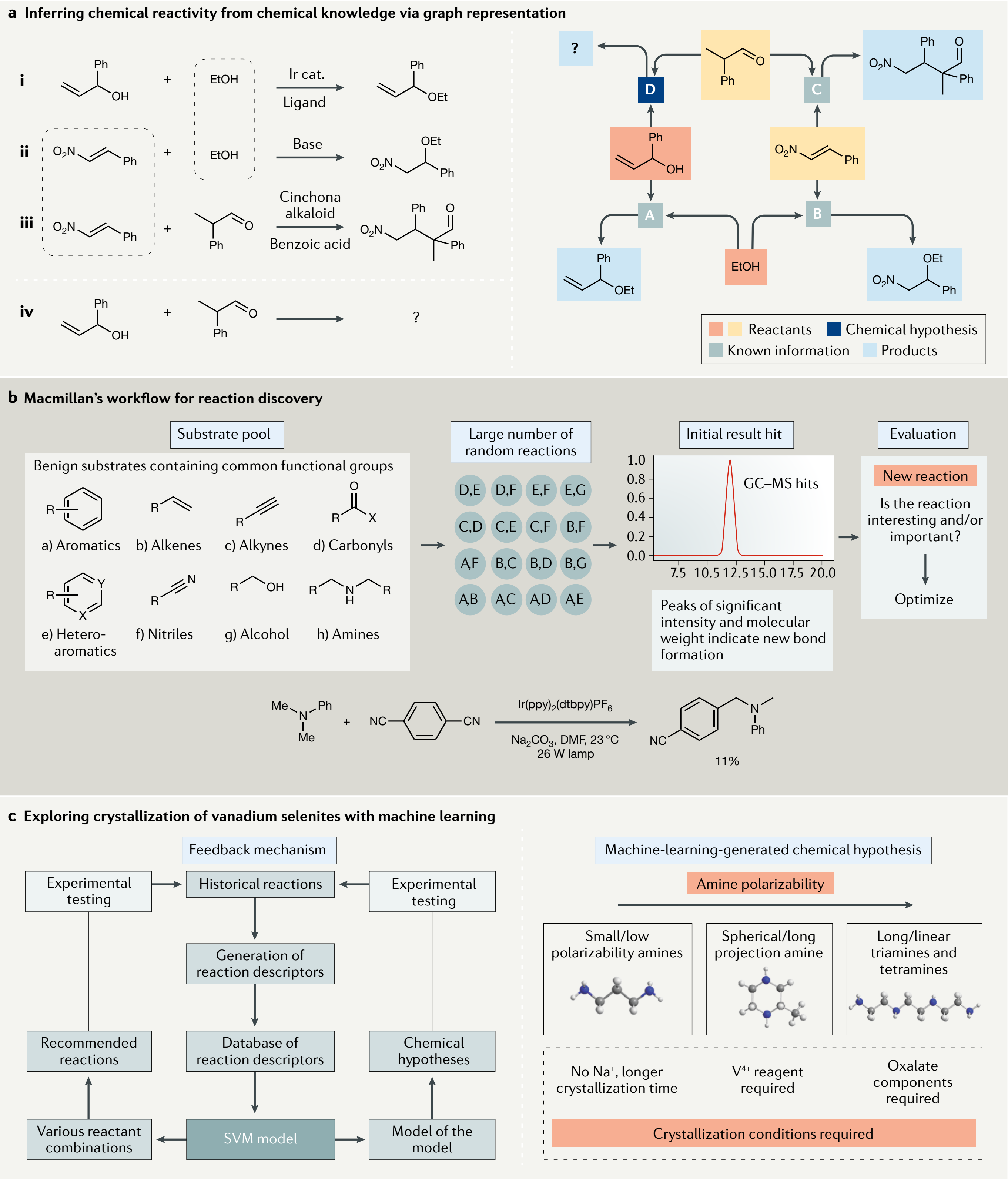
Parameter	1	2	...	*n*
Sample 1	2.5	15	...	0.32
Sample 2	3.1	10	...	0.27
...	...	...	...	...
Sample *n*	2.9	20	...	0.25

Fig. 2 | **Creation of databases and the extraction of data.** From left to right: input data are the collection of features, descriptors and raw data. Database structure is the structure holding the data based on the input classification and the intended use. Interpretation and mapping includes the various methods that can be used to identify molecular candidates, depending on the desired goal, as shown at the top with an example of ligand docking. The bottom shows computational molecular design utilizing the information gleaned from the databases in order to design molecules that are not outside the known collection. Model/output and verification — this is the final stage, in which the usefulness of the process is assessed through statistical verification methods; the red lines are candidate pathways, and the green network is a connected local portion of the chemical network.

A new approach to materials design for organic light-emitting diodes was demonstrated by Aspuru-Guzik and co-workers using high-throughput virtual screening. By combining theoretical computations, chemoinformatics with machine learning and organic synthesis, it was possible to successfully narrow down the space of 1.6 million possible molecules to thousands of promising novel organic light-emitting diode molecules and then to successfully synthesize new organic light-emitting diodes[28].

Searching chemical space

Chemical space can be searched for specific molecules or specific goals such as optimized yield or a biological function. This search can be performed through two different avenues: theoretical and experimental. Theoretical searches are an important use


a Inferring chemical reactivity from chemical knowledge via graph representation
i
Ph
OH
+
EtOH
Ir cat.
Ligand
Ph
OEt
ii
O₂N
Ph
+
EtOH
Base
OEt
O₂N
Ph
iii
O₂N
Ph
+
O
Ph
Cinchona alkaloid
Benzoic acid
Ph
O
O₂N
Ph
iv
Ph
OH
+
O
Ph
?
?
D
C
A
B
EtOH
Reactants
Chemical hypothesis
Known information
Products
b Macmillan's workflow for reaction discovery
Substrate pool
Benign substrates containing common functional groups
R
a) Aromatics
b) Alkenes
c) Alkynes
d) Carbonyls
X
Y
e) Hetero-aromatics
N
f) Nitriles
OH
g) Alcohol
N
H
h) Amines
Large number of random reactions
D,E D,F E,F E,G
C,D C,E C,F B,F
A,F B,C B,D B,G
A,B A,C A,D A,E
Initial result hit
1.0
0.8
0.6
0.4
0.2
0.0
7.5 10.5 12.5 15.0 17.5 20.0
GC–MS hits
Peaks of significant intensity and molecular weight indicate new bond formation
Evaluation
New reaction
Is the reaction interesting and/or important?
Optimize
Me
N
Ph
Me
+
NC
CN
Ir(ppy)₂(dtbpy)PF₆
Na₂CO₃, DMF, 23 °C
26 W lamp
NC
N
Ph
11%
c Exploring crystallization of vanadium selenites with machine learning
Feedback mechanism
Experimental testing
Historical reactions
Experimental testing
Generation of reaction descriptors
Recommended reactions
Database of reaction descriptors
Chemical hypotheses
Various reactant combinations
SVM model
Model of the model
Machine-learning-generated chemical hypothesis
Amine polarizability
Small/low polarizability amines
Spherical/long projection amine
Long/linear triamines and tetramines
No Na⁺, longer crystallization time
V⁴⁺ reagent required
Oxalate components required
Crystallization conditions required

case for databases of accumulated chemical knowledge. One way of expressing the chemical knowledge is by defining a graph with molecules as nodes and the reactions as edges[29]. This work, shown in FIG. 3a, created a new knowledge representation of chemical reactions. Once in this new form, Segler and co-workers were able to take techniques from the field of network analysis and apply them to the chemical problem of reaction prediction; the techniques allowed them to predict new plausible links in the graph. However, they had to adapt the techniques to account for the unique attributes that allow the edges to represent the reactant, reagent, catalyst, solvent and product of the molecules in chemical reactions that are missing in other types of graph. In addition to the prediction of reactions, they were able to predict the reaction conditions (that is, reactivity) by looking at the paths in the network that connect similar molecules. This means that the chemical information in databases can be used to perform theoretical searches for new knowledge among the existing chemical knowledge. These searches are made possible by reframing chemical information into new formats that extend the scope of tools that can be used. New developments[30] follow this path of extending the ways in which chemical information can be represented, thereby allowing for more methods from fields outside of chemistry to explore the information within chemical databases.

Retrosynthetic analysis can be considered as a reverse search of chemical space, starting from the target molecules and recursively transforming them into simpler precursors. Recently, retrosynthetic searches have been automated using deep neural networks and symbolic artificial intelligence by Waller and colleagues[31]. The neural networks were trained using databases that contain essentially all-known organic chemistry in order to efficiently guide and preselect the most promising retrosynthetic routes.

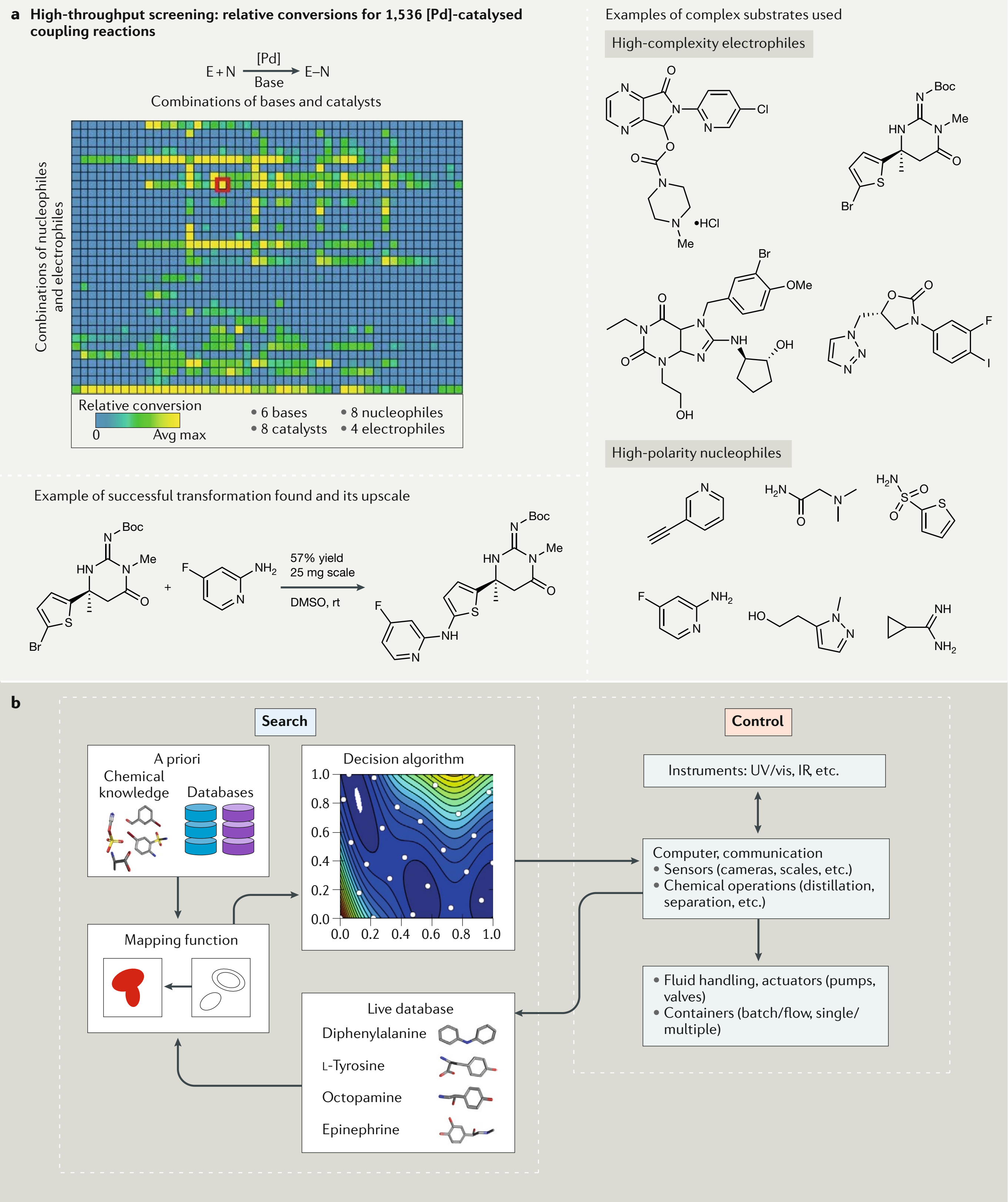
A common algorithm used for search problems is random search. It is both simple and limited because it does not make use of any feedback from experiments. A random search chooses the experiments to perform by picking randomly, with all possible choices equally probable. However, a random search cannot guarantee that the most interesting and informative parts of the space are being explored. The use of a random search for reaction discovery has been shown by MacMillan et al.[32]. However, it should be noted that their work would not be possible without a number of previous studies. For instance, Weber et al.[33] described how new multicomponent reactions can be rationally designed, discovered and optimized using automated approaches. Beeler et al.[34] developed these ideas further, building a workflow for reaction discovery constituting combinatorial screening, liquid chromatography/mass spectrometry/electrophoretic light scattering (LC/MS/ELS) screening, structure elucidation and reaction optimization. Multidimensional approaches to the high-throughput discovery of catalytic reactions have also been demonstrated in the work of Robbins and Hartwig[35]. They were able to develop a practical approach to chemistry involving a random search, avoiding known reagent combinations and using rapid analysis by GC–MS (see FIG. 3b). A careful selection of candidate pairings of the reagents, by deliberately excluding known reagent combinations, generates a large number of possible reactions to perform. Of these options, random search was used to carry out the number of reactions that could be performed practically. This so-called accelerated serendipity provides a good starting point for pinpointing new reactivity by further testing these potential combinations. This subsequent testing of the outcomes of the random search led to the discovery of a new C–H arylation reaction. Although random search is a trivial strategy, it can still lead to useful discovery if there is an intelligent selection of the search space. More examples of this approach can be found in the literature[36,37].

Intelligent search algorithm approaches that generate new experiments to be performed using feedback from reactions already tested are potentially much more efficient for use in chemistry. This is because there is inevitably a limited number of reactions that can be performed (a reaction budget); this budget is better spent on reactions that are more likely to match the search goal. Algorithms such as simulated annealing[38,39], genetic algorithms[17,40] and particle swarm optimization[41] choose the next experiments using data obtained from previous experiments and can be referred to as instance-based algorithms. These algorithms try to perform successive experiments, aiming for improvement in the overall trend by selecting each subsequent experiment using a statistical approach. The more likely directions are the ones that the algorithms have determined to have the highest chance to yield the biggest improvement using prior data. Advances in artificial intelligence have also been used to optimize chemical reactions using deep reinforcement learning. By using recurrent neural networks, which are able to retain memories of previous experiments, the system was able to outperform current optimization algorithms[42]. A further example of the application of such algorithms is presented in the work of Nikolaev et al. for carbon nanotube growth[43].

The second approach to search algorithms consists of model-based algorithms. These algorithms also use feedback from previous experiments, with the goal of improving the outcome of the ongoing search. However, the decision is made on the basis of a model of the

◀ Fig. 3 | **Searching for new reactivity, methods or properties. a** | Inferring reactivity from a graph structure. The traditional representation of a set of chemical reactions is shown in the left panel. Graphical representation (shown in the right panel) of the same chemical knowledge: the chemical knowledge can be represented as a graph $G=(M, R, E)$, where M denotes molecular nodes, R denotes reaction nodes and E denotes directed reaction edges. Searching for new reactivity has been reformulated as a problem of finding missing links in the knowledge graph between reactant molecules. The reframing enables the use of a mathematical graph methodology. **b** | Using high-throughput screening to improve the chances of novel reaction discovery. The process begins with substrates comprising well-known functional groups (generic organic substituent (R) and heteroatoms in positions X and Y), which is followed by a large number of reactions being performed by a robotic system. In the next step, the potential coupling for each reaction is estimated on the basis of gas chromatography–mass spectrometry (GC–MS) measurements. In the final step, initial results are assessed for their importance. The reaction shows a hit from the screening — the photocatalysed formation of an α-aminocyanobenzene coupling product. **c** | Machine-learning-assisted materials discovery. The feedback loop for learning from historical reactions (shown in the left panel), in which the prediction success is based on all possible combinations of the reaction and the conditions involved. The process starts with data entry acquired from notebooks, which allows for the generation of reaction and reactant descriptions (database of reaction descriptions), which is then used for training and testing using support vector machines (SVMs). From this point, a model of the model construction is used for the generation of interpretable decision trees, which allows generating human interpretable hypotheses about crystal formation or various other reactant combinations that are set for further recommendation and experimental testing, which effectively allow closing of the loop. An example of a chemical hypothesis (shown in the right panel) generated from the inversion of the machine learning model: polarizability and shape of the amine versus additivities and crystallization conditions. Cat, catalyst; DMF, dimethylformamide. Part **a** is adapted with permission from REF.[29], Wiley-VCH. Part **b** is adapted with permission from REF.[32], AAAS. Part **c** is adapted from REF.[47], Springer Nature Limited.

space. This model is constructed from the available data by an additional algorithm. Both during and at the end of a search, we possess a model that describes the space with improving fidelity; the system therefore develops intuition about the chemical space as the number of data points increases. Of the many different kinds of model-based algorithm, some common examples belong to the field of design of experiments, such as the two-level factorial, in which the identification of important factors is done at the beginning of the process; Plackett–Burman, which is similar to two-level factorial design and allows the


a
High-throughput screening: relative conversions for 1,536 [Pd]-catalysed coupling reactions
E + N → E–N
[Pd]
Base
Combinations of bases and catalysts
Combinations of nucleophiles and electrophiles
Relative conversion
0
Avg max
• 6 bases
• 8 catalysts
• 8 nucleophiles
• 4 electrophiles
Examples of complex substrates used
High-complexity electrophiles
•HCl
Boc
Me
Br
OMe
OH
F
I
High-polarity nucleophiles
Example of successful transformation found and its upscale
57% yield
25 mg scale
DMSO, rt
b
Search
A priori
Chemical knowledge
Databases
Mapping function
Decision algorithm
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Live database
Diphenylalanine
L-Tyrosine
Octopamine
Epinephrine
Control
Instruments: UV/vis, IR, etc.
Computer, communication
• Sensors (cameras, scales, etc.)
• Chemical operations (distillation, separation, etc.)
• Fluid handling, actuators (pumps, valves)
• Containers (batch/flow, single/multiple)

elimination of unimportant variables or factors through a type of screening design; full factorial, in which all the compositions of the levels of the factors are processed; and Box–Behnken, which is a surface methodology approach that has three or more levels and therefore can be used for the design of experiments with three or more factors. Finally, Doehlert designs, which, unlike Box–Behnken, are not rotatable, can give different qualities of estimates for different factors, which gives them very high efficiency that allows them to have different numbers of levels for different numbers of factors[44,45]. Nevertheless, these cannot be applied in a so-called black-box approach; it is essential to understand the application at hand and the data structure. This is due to the 'no free lunch theorem'[46], which states that in searching and optimization, the application of an algorithm to different data sets may result in diverse outputs and thus, in general, no single algorithm is optimal for solving all problems. This class of search algorithms in turn uses different model-building algorithms such as support vector machines (SVMs), a machine learning technique that can be used for both classification and regression, and which is particularly useful in studying small data sets. This is especially important in drug design for the prediction of various chemical properties, the optimization of chromatographic separation, fault detection, the modelling of industrial processing and much more[47]; self-organizing maps, an unsupervised learning method with links to artificial neural networks, which could be useful in identifying targets for both known drugs and computer-generated molecular scaffolds[48]; and kriging, a stochastic method for spatial predictions very often used for approximation problems[49]. Models of the search space are used in this case to enable efficient searching of the space. This allows the search algorithm to make predictions about the space. Furthermore, the models can also be used to characterize and understand the search space, with the overarching aim of uncovering novel parts of the chemical space.

A recent example of this approach has been the combination of prior knowledge with machine learning algorithms to predict the crystallization conditions of templated vanadium selenites[47], as shown in FIG. 3c. An SVM model, trained on data extracted from historical reactions after adding physicochemical descriptors, was able to predict crystallization outcomes and outperformed human experimenters. Inversion of the machine learning model, so that desired outcomes led to expected reaction conditions, provided recommendations for candidate crystallizations. FIGURE 3c shows one of the chemical hypotheses generated by the model that was tested experimentally and resulted in the identification of several new inorganic compounds. This model-based approach has also been used by Doyle and co-workers to navigate chemical space using high-throughput screening and machine learning[50]. By generating descriptors of Buchwald–Hartwig amination components, the authors were able to demonstrate an increase in efficiency of yield prediction accuracy using an ensemble learning method for classification and a regression model called random forest that substantially outperformed other linear regression models presented in the study. This result could be related to the fact that random forest is a nonlinear approach that randomly samples the data by constructing decision trees, which are then used to generate overall prediction. This was further exemplified by demonstrating the applicability of this protocol to deoxyfluorination with sulfonyl chlorides[51].

Model-based algorithms make use of feedback to generate a better model of chemical space and hence the potential points within it. The model contains the information gathered about the space, which might also be called its understanding. The system is given a chance to develop intuition about the chemical space as the number of data points increases and as the model is updated and refined. The bottom line is that using advanced algorithms to build models of the search space not only is more efficient but also increases the odds of generating new chemical knowledge because the model includes the collected knowledge. Understanding the chemical space can help find the parts that are most likely to contain the previously unknown. To gather sufficient experimental data to create good models, it is useful to use automation.

Automating the search

The use of automated approaches in chemistry could transform even the average laboratory by reducing the manual labour and time required for reaction preparation and work-up[52–55]. Robots also have the potential to improve the quality of experiments by decreasing variability[4]. In addition, the use of automation allows the exact operations undertaken for each experiment to be logged and linked with the results. Later, that information can be used for validation or repetition, as well as better knowledge archiving, transfer and reproducibility. Furthermore, the introduction of automation, especially in organic synthesis, might enable a reactivity-driven rather than target-driven search of chemical space[16].

The increasing availability of off-the-shelf robots capable of performing chemical operations allows for higher-throughput exploration: the increase in the practical number of experiments that can be performed can help when using a common search algorithm, that is, screening. When performing a screening search, the search space contains all the possible experiments within the constraints of the cost, time and resources that are available. Using a high-throughput robot greatly increases this total number of experiments. When the screening begins, all the experiments that are to be performed are already scheduled, and they are all executed, either in sequence or in parallel. An example of this approach was described by Santanilla et al.[23] (see FIG. 4a).

Standard off-the-shelf robots are inflexible in the chemistry that they can perform. Santanilla and co-workers were able to overcome this limitation by devising chemistry specifically suited to the capabilities of their robot[23]. For example, by adapting the chemistry to use dimethyl sulfoxide as a solvent — a low-volatility, plastic-compatible solvent that facilitates working at this scale — they were able to perform more than 1,500 experiments using as little as 0.02 mg of material. This

◀ Fig. 4 | **Optimizing reaction conditions. a** | Nanoscale high-throughput screening. High-throughput screening platform for Pd-catalysed cross-coupling reactions at ambient conditions and solubilizing, low-volatility, plastic-compatible solvents. To the left, the heat map shows data from 1,536 nanomole-scale reactions, with one example of a reaction discovered at room temperature (rt) (highlighted by a red box). To the right are examples of the high-complexity electrophiles and high-polarity nucleophiles used. **b** | Closed-loop robots work across two separate worlds. The control uses different operations needed to perform the desired reactions, such as fluid handling, measuring instruments and sensors, and communication and device-level control. The search algorithm operates entirely in silico with data from both a priori knowledge, such as chemical reactivity information and stored databases, and the live database of current experimental results. These two data sources are transformed into the desired conceptual representation using a mapping function that describes the search space. Inside this search space, a decision algorithm chooses the sequence of experiments to be performed by the control side. DMSO, dimethyl sulfoxide. Adapted with permission from REF.[23], AAAS.

screening step, followed by additional lower-throughput screenings scaled up to the milligram range, enabled the identification of dozens of successful reactions, including room temperature, metal-catalysed cross-coupling reactions, using several organic superbases in combination with biaryl palladium precatalysts. The benefits of using a high-throughput system in this case came with the trade-off of the limited chemistry. The advantages offered by robots increase greatly with the development of new off-the-shelf and bespoke systems that provide increased flexibility in the chemistry they can perform. By combining experimental and chemical design, robotics can be used for automatic, precise and highly reproducible investigations of reaction optimization and discovery. A contrasting example of the power of combining automation and innovative chemistry can be seen in the broadly applicable approach to small-molecule synthesis developed by Burke and co-workers[56]; for a general review of the subject, see Lehmann et al.[57]. The platform contains three parts: deprotection, coupling and purification modules. In the first stage, the MIDA boronate (*N*-methyliminodiacetic boronic acid ester) is deprotected to give free boronic acid, which is then coupled to alkyl and/or aryl halides in the coupling module in a Pd-catalysed reaction. The products of the reaction are then purified in a purification module employing the binary affinity of MIDA boronates towards silica gel in different solvents. Small molecules were synthesized in an iterative fashion analogously to peptide synthesis using MIDA boronates as the building blocks. The general applicability of this process was demonstrated by an automated synthesis of complex molecules, including natural products.

Although this system does not perform high-throughput experimentation, it makes great use of automation to push the envelope and the scope of automated chemistry. In another powerful approach, flow synthesis was combined with formulation to deliver a compact platform for the preparation of pharmaceutically active ingredients[58]. For a thorough review of the automation of small-molecule synthesis, see the recent review by Trobe and Burke[59].

Finally, automation is vital in closed-loop systems[60], as it allows the use of feedback-control algorithms as described by Gutierrez et al. (see FIG. 4b). A useful example of such an application is the autonomous system for the controlled synthesis of fluorescent nanoparticles described by Krishnadasan et al.[61].

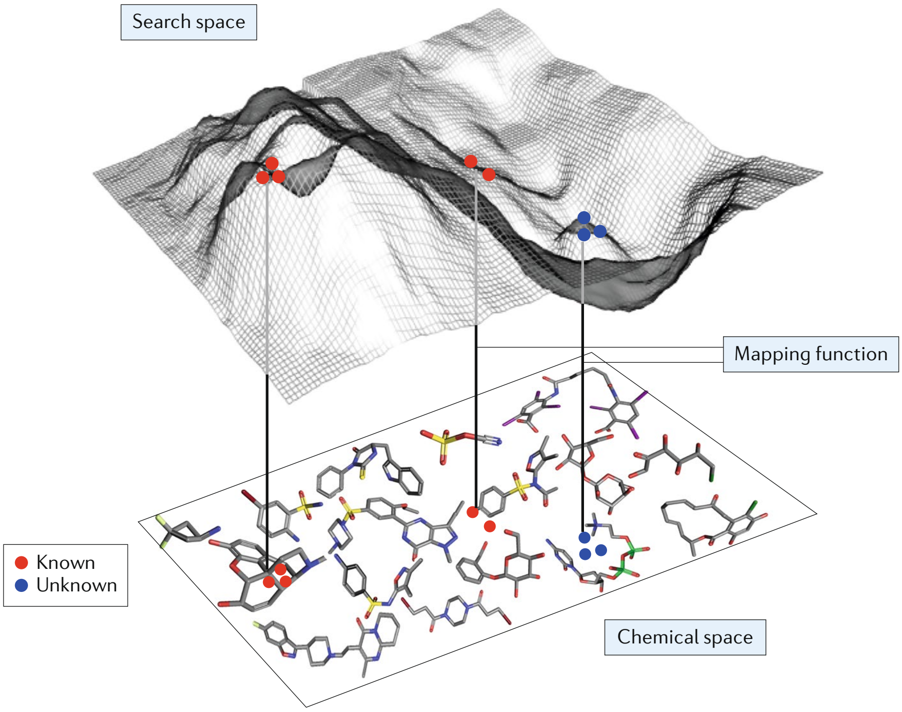


Fig. 5 | **A projected 3D search space.** The use of a mapping function can transform a 2D chemical space into a 3D search space, where the third dimension corresponds to the search goal. Red dots indicate known outcomes, whereas blue dots indicate new discoveries.

On the control side, the system uses a microfluidic reactor to carry out the synthesis of nanoparticles with an in-line spectrometer to monitor the emission spectra of the produced nanoparticles. In the search section, the data are collected and processed using a mapping function called the dissatisfaction coefficient, which runs linearly from zero to one, in which zero means complete satisfaction and one means complete dissatisfaction, and the search algorithm that was used was stable noisy optimization by branch and fit (SNOBFIT), which is designed to select continuous parameter settings for simulations or to optimize some user-specified criterion.

Recently, Aspuru-Guzik and co-workers[62] described a portable and modular software framework for operating closed-loop systems with integrated robotics, sensors and artificial intelligence. This single-software framework was applied to a number of different tasks, including learning the colour space of dyes and autonomously calibrating high-performance liquid chromatography analysis.

Uncovering novelty

Most chemical discoveries belong to three categories: new molecules, new reactions and new reactivity. Finding new reactivity enables a search for new reactions, which in turn aids the discovery of new molecules. The amount of chemical knowledge contributed by each of these types of discoveries means that this is also the order of their potential impact. Such findings must, by definition, belong outside the known or predictable; they are outliers and, as such, can oppose conventions, assumptions and biases. We can use the idea of an outlier to define a novel discovery in practical terms as any information about the chemical space that exhibits sufficiently different outcomes from prediction. This definition places the burden of the proof of novelty on the source of prediction, experimental or theoretical knowledge. It would be favourable to define a chemical system well enough to be able to make predictions based upon it — it then becomes possible to seek outliers. Whether a value is sufficiently different to be called an outlier depends on the characteristics of the system under study. This kind of anomaly detection is highly developed in many fields outside chemistry and may enable chemists to better define novelty criteria in the future[63–65].

The description of scientific work as novel can sometimes prove to be controversial. We therefore introduce here a practical, clear and useful approach to novelty. In order to ascertain if a discovery is novel, we have developed a simple but algorithmically programmable three-step approach (see FIG. 5). In general, it seeks to determine if an outcome is repeatable, new and predicted by the current knowledge. An implementation for an autonomous robot would, in the first step, evaluate the reproducibility of the data. This step filters for outcomes that simply result from a high degree of experimental variance. An abnormal outcome can result from, for example, contamination or an unlikely stochastic reaction. By enforcing repeatability on the outcome, such results would be excluded as candidates for novelty. Second, the chemical databases are searched to check if the outcome is new. Of course, the values in a database may not match exactly the experimental result, but as long as they are within a given margin, they should be used to conclude whether an outcome is new. Third, if the observation is new, then a determination about the predictability of the outcome from the current body of chemical knowledge is made. Put simply, if a result can be reached by whatever means from current available information, then it is predictable. It is possible to imagine that the rules of organic synthesis, including all the known transformations, could be encoded into a model[2] and then that novelty could be discovered when new reactions, molecules or reactivities are found when these rules are broken. If not all chemical rules are known, then novelty would involve adding new rules. The question of predictability is potentially very difficult to objectively determine, and hence, many experts often argue if an outcome is novel when considering all chemical knowledge. However, for the purposes of the system doing the measurements, the system can easily determine all three steps from its internal perspective. It is entirely plausible that chemical platforms could find outcomes that are novel given the information supplied to it, but when approached with the benefit of broader knowledge, the outcomes are no longer novel. This general approach serves two purposes: it frames a discussion of how to decide whether an outcome is to be considered novel, but more importantly, it also allows, in a practical way, a robot to conduct a deliberate search for novelty in an automated fashion. This approach can help chemists on two fronts: the former gives them a framework to evaluate chemical outcomes, while the latter enables the use of robots that can assist in exploring chemical space. Seeking outliers and then evaluating their novelty may seem straightforward, but we must still address the question of the best way to find outliers in chemical systems.

Outliers stand out because they contradict the knowledge gathered about the search space. Thus, a system needs a picture of the search space while the search is being conducted. In other words, outliers are easier to define when using a model. By using a model-based search algorithm, the system (and the chemist) will have a better understanding of the space, as the model can be used to explain and predict the space (see FIG. 5). By making predictions, we can subsequently find parts of the chemical space that do not match the results that are anticipated. This means that a search algorithm with a valid model (for example, of a reactivity pattern) allows the experimenter to make a clearer distinction between outliers that represent novel discoveries and ones that do not (FIG. 6). The more data the model is based on, the more precise the result. Thus, as the search algorithm performs more experiments and the model is dynamically updated as the results are obtained, the model is likely to have an improved ability to find outliers. A good model will help the search algorithm to prepare good experiments that add more information about the system. Using this flexible and adaptive approach, we can thus reach a far more comprehensive understanding of the chemical system under study and the chemical space that it embodies. This means that new information — not deducible using conventional techniques with small numbers of experiments — will become available, and by applying big data approaches and deep learning techniques, we will challenge our view of the rules of chemistry.

Conclusions

Major hurdles towards efficient exploration and discovery in chemistry include the difficulty in designing new ways to automatically search chemical space and the use of the appropriate search methodologies. The potential for improvement encompasses both basic and advanced chemical research, from basic reaction parameter optimization to novel reaction discovery. When combined with automated robotic systems, suitable algorithms can provide autonomous operation, allowing for larger chemical systems along with increased precision and reproducibility. However, the choice of the search algorithm and overarching framework must consider the characteristics of chemical space. The characteristics of this space, most notably its sparsity and practical limitations, offer two alternatives. The first and most common is to use heuristics to focus the space on areas that are easier to work with. The drawback of this approach is that it might suffer from biases that eliminate possible discoveries. The second is to use a search framework with a model-based search algorithm looking for outliers. By using a dynamic model search space, the chemical space can be explored with the presence of outliers directing the search towards the discovery of chemical novelty. To do this, we must not forget that chemists set out not only to discover new reactions but also to apply known design rules to make new molecules.

Ultimately, if chemical artificial intelligence (CAI) can be used to break the rules, find new reactivity and then update the rules, enabling the design of new molecules, then chemists will be able to replace serendipity with certainty. Once this becomes possible, the rate of discovery in chemistry will increase dramatically. This is because the use of CAIs enables chemists to challenge the current rules through the removal of bias. We predict that this will greatly accelerate the pace of discovery of new molecules, reactions and reactivity patterns. To do this, it is important to realize that a given reaction or reactivity pattern does not constitute a fixed law; although these reactions and patterns are extremely useful practical guides and use a consistent chemical language, they should be viewed as another layer of abstraction. The challenge for scientists is to step beyond their bias and instead use their expert knowledge together with CAI and chemical robots as a tool to venture into the unknown and to boldly go and explore chemical space.

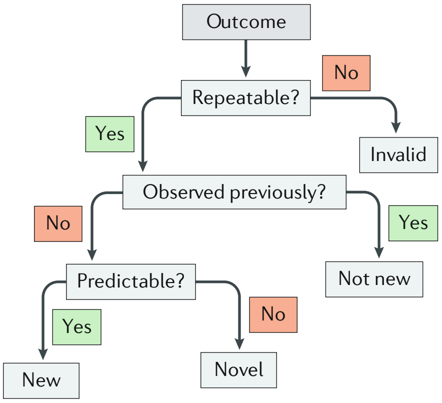


Fig. 6 | **A flow chart to assist in a strict definition of validity, newness and novelty.** A result is valid only if repeatable, new if not previously observed and novel only when the result would not be (easily) predicted on the basis of prior results.

*Piotr S. Gromski[1], Alon B. Henson[1], Jarosław M. Granda[1] and Leroy Cronin[1]**

WestCHEM, School of Chemistry, University of Glasgow, Glasgow, UK.

**e-mail: Lee.Cronin@glasgow.ac.uk*

https://doi.org/10.1038/s41570-018-0066-y

Published online: 15 January 2019

1. Miller, M. A. Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.* **1**, 220–227 (2002).
2. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
3. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
4. Richmond, C. J. et al. A flow-system array for the discovery and scale up of inorganic clusters. *Nat. Chem.* **4**, 1037–1043 (2012).
5. Carell, T. et al. New promise in combinatorial chemistry: synthesis, characterization, and screening of small-molecule libraries in solution. *Chem. Biol.* **2**, 171–183 (1995).
6. Ortholand, J.-Y. & Ganesan, A. Natural products and combinatorial chemistry: back to the future. *Curr. Opin. Chem. Biol.* **8**, 271–280 (2004).
7. Ingham, R. J. et al. A systems approach towards an intelligent and self-controlling platform for integrated continuous reaction sequences. *Angew. Chem. Int. Ed.* **54**, 144–148 (2015).
8. Sans, V., Porwol, L., Dragone, V. & Cronin, L. A self optimizing synthetic organic reactor system using real-time in-line NMR spectroscopy. *Chem. Sci.* **6**, 1258–1264 (2015).
9. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **4**, 468–481 (2014).
10. Oprea, T. I. & Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **3**, 157–166 (2001).
11. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
12. Goodnow, R. A. Jr, Dumelin, C. E. & Keefe, A. D. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nat. Rev. Drug Discov.* **16**, 131–147 (2017).
13. Reymond, J.-L., Ruddigkeit, L., Blum, L. & van Deursen, R. The enumeration of chemical space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2**, 717–733 (2012).
14. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **1**, 30–38 (2010).
15. Troshin, K. & Hartwig, J. F. Snap deconvolution: an informatics approach to high-throughput discovery of catalytic reactions. *Science* **357**, 175–181 (2017).
16. Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 15733 (2017).
17. Kreutz, J. E. et al. Evolution of catalysts directed by genetic algorithms in a plug-based microfluidic device tested with oxidation of methane by oxygen. *J. Am. Chem. Soc.* **132**, 3128–3132 (2010).
18. Hopkinson, M. N., Gómez-Suárez, A., Teders, M., Sahoo, B. & Glorius, F. Accelerated discovery in photocatalysis using a mechanism-based screening method. *Angew. Chem. Int. Ed.* **55**, 4361–4366 (2016).
19. Grzybowski, B. A., Bishop, K. J. M., Kowalczyk, B. & Wilmer, C. E. The 'wired' universe of organic chemistry. *Nat. Chem.* **1**, 31–36 (2009).
20. Soh, S. et al. Estimating chemical reactivity and cross-influence from collective chemical knowledge. *Chem. Sci.* **3**, 1497–1502 (2012).
21. Scior, T. et al. Recognizing pitfalls in virtual screening: a critical review. *J. Chem. Inf. Model.* **52**, 867–881 (2012).
22. Collins, K. D., Gensch, T. & Glorius, F. Contemporary screening approaches to reaction discovery and development. *Nat. Chem.* **6**, 859–871 (2014).
23. Santanilla, A. B. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
24. Ruddigkeit, L., Awale, M. & Reymond, J.-L. Expanding the fragrance chemical space for virtual screening. *J. Cheminform.* **6**, 27 (2014).
25. Brereton, R. G. The evolution of chemometrics. *Anal. Methods* **5**, 3785–3789 (2013).
26. Hopke, P. K. The evolution of chemometrics. *Anal. Chim. Acta* **500**, 365–377 (2003).
27. 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).
28. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
29. Segler, M. H. S. & Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chem. Eur. J.* **23**, 6118–6128 (2017).
30. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
31. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
32. McNally, A., Prier, C. K. & MacMillan, D. W. C. Discovery of an amino acid C-H arylation reaction using the strategy of accelerated serendipity. *Science* **334**, 1114–1117 (2011).
33. Weber, L., Illgen, K. & Almstetter, M. Discovery of new multi component reactions with combinatorial methods. *Synlett* **3**, 366–374 (1999).
34. Beeler, A. A., Su, S., Singleton, C. A. & Porco, J. A. Discovery of chemical reactions through multidimensional screening. *J. Am. Chem. Soc.* **129**, 1413–1419 (2007).
35. Robbins, D. W. & Hartwig, J. F. A. Simple, multidimensional approach to high-throughput discovery of catalytic reactions. *Science* **333**, 1423–1427 (2011).
36. Walker, B. E., Bannock, J. H., Nightingale, A. M. & deMello, J. C. Tuning reaction products by constrained optimisation. *React. Chem. Eng.* **2**, 785–798 (2017).
37. Chen, S., Reyes, K.-R. G., Gupta, M. K., McAlpine, M. C. & Powell, W. B. Optimal learning in experimental design using the knowledge gradient policy with application to characterizing nanoemulsion stability. *SIAM/ASA J. Uncertain. Quantif.* **3**, 320–345 (2015).
38. Kalivas, J. H., Roberts, N. & Sutter, J. M. Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry. *Anal. Chem.* **61**, 2024–2030 (1989).
39. Sutter, J. M., Dixon, S. L. & Jurs, P. C. Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing. *J. Chem. Inf. Comput. Sci.* **35**, 77–84 (1995).
40. Corma, A. et al. Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques). *J. Catal.* **229**, 513–524 (2005).
41. Chen, X., Du, W., Qi, R., Qian, F. & Tianfield, H. Hybrid gradient particle swarm optimization for dynamic optimization problems of chemical processes. *Asia Pac. J. Chem. Eng.* **8**, 708–720 (2013).
42. Zhou, Z., Li, X. & Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* **3**, 1337–1344 (2017).
43. Nikolaev, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *Comput. Mater.* **2**, 16031 (2016).
44. Hibbert, D. B. Experimental design in chromatography: a tutorial review. *J. Chromatogr. B* **910**, 2–13 (2012).
45. Murray, P. M., Tyler, S. N. G. & Moseley, J. D. Beyond the numbers: charting chemical reaction space. *Org. Process Res. Dev.* **17**, 40–46 (2013).
46. Wolpert, D. H. & Macready, W. G. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**, 67–82 (1997).
47. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
48. Reker, D., Rodrigues, T., Schneider, P. & Schneider, G. Identifying the macromolecular targets of de novo-designed chemical entities through self-organizing map consensus. *Proc. Natl Acad. Sci. USA* **111**, 4067–4072 (2014).
49. Sieg, S., Stutz, B., Schmidt, T., Hamprecht, F. & Maier, W. F. A. QCAR-approach to materials modeling. *J. Mol. Model.* **12**, 611–619 (2006).
50. 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).
51. Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **140**, 5004–5008 (2018).
52. Ley, S. V., Fitzpatrick, D. E., Myers, R. M., Battilocchio, C. & Ingham, R. J. Machine-assisted organic synthesis. *Angew. Chem. Int. Ed.* **54**, 10122–10137 (2015).
53. Pastre, J. C., Browne, D. L. & Ley, S. V. Flow chemistry syntheses of natural products. *Chem. Soc. Rev.* **42**, 8849–8869 (2013).
54. Straathof, N. J. W., Su, Y., Hessel, V. & Noël, T. Accelerated gas-liquid visible light photoredox catalysis with continuous-flow photochemical microreactors. *Nat. Protoc.* **11**, 10–21 (2016).
55. Ghislieri, D., Gilmore, K. & Seeberger, P. H. Chemical assembly systems: layered control for divergent, continuous, multistep syntheses of active pharmaceutical ingredients. *Angew. Chem. Int. Ed.* **54**, 678–682 (2015).
56. Li, J. et al. Synthesis of many different types of organic small molecules using one automated process. *Science* **347**, 1221–1226 (2015).
57. Lehmann, J. W., Blair, D. J. & Burke, M. D. Towards the generalized iterative synthesis of small molecules. *Nat. Rev. Chem.* **2**, 0115 (2018).
58. Adamo, A. et al. On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system. *Science* **352**, 61–67 (2016).
59. Trobe, M. & Burke, M. D. The molecular industrial revolution: automated synthesis of small molecules. *Angew. Chem. Int. Ed.* **57**, 2–25 (2018).
60. Gutierrez, J. M. P. et al. Evolution of oil droplets in a chemorobotic platform. *Nat. Commun.* **5**, 5571 (2014).
61. Krishnadasan, S., Brown, R. J. C., DeMello, A. J. & DeMello, J. C. Intelligent routes to the controlled synthesis of nanoparticles. *Lab. Chip* **7**, 1434–1441 (2007).
62. Roch, L. M. et al. ChemOS: an orchestration autonomous experimentation. *Sci. Robot.* **3**, eaat5559 (2018).
63. Goldstein, M. & Uchida, S. A. Comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE* **11**, e0152173 (2016).
64. Chandola, V., Banerjee, A. & Kumar, V. Anomaly detection: a survey. *ACM Comput. Surv.* **41**, 15 (2009).
65. Oprea, T. I. Chemical space navigation in lead discovery. *Curr. Opin. Chem. Biol.* **6**, 384–389 (2002).

Acknowledgements

The authors gratefully acknowledge financial support from the UK Engineering and Physical Sciences Research Council (EPSRC) (grant nos EP/H024107/1, EP/I033459/1, EP/J00135X/1, EP/J015156/1, EP/K021966/1, EP/K023004/1, EP/K038885/1, EP/L015668/1 and EP/L023652/1) and the European Research Council (ERC) (project 670467 SMART-POM).

Author contributions

P.S.G. and A.B.H. contributed equally to the article. L.C. conceived the framework and developed the novelty algorithm presented here; L.C., A.B.H., P.S.G. and J.M.G. performed the literature review; and all the authors wrote the article. The authors thank N. A. B. Johnson for the artistic depiction used in the graphical abstract.

Competing interests

The authors declare no competing interests

Publisher's note

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

Reviewer information

Nature Reviews Chemistry thanks M. Waller and the other anonymous reviewer(s) for their contribution to the peer review of this work.