A Journal of the Gesellschaft Deutscher Chemiker

Angewandte International Edition

GDCh

Chemie

www.angewandte.org

Accepted Article

Title: Autonomous discovery in the chemical sciences part I: Progress

Authors: Klavs F. Jensen, Connor W Coley, and Natalie S Eyke

To be cited as: *Angew. Chem. Int. Ed.* 10.1002/anie.201909987
Angew. Chem. 10.1002/ange.201909987

Link to VoR: http://dx.doi.org/10.1002/anie.201909987
http://dx.doi.org/10.1002/ange.201909987

WILEY-VCH

10.1002/anie.201909987

Autonomous discovery in the chemical sciences part I: Progress

Connor W. Coley*† Natalie S. Eyke* Klavs F. Jensen*‡

Keywords: automation, chemoinformatics, machine learning, drug discovery, materials science

Author bios:

Connor W. Coley completed his B.S. in chemical engineering at the California Institute of Technology and his M.S.C.E.P. and Ph.D. in chemical engineering at the Massachusetts Institute of Technology (MIT). His research focuses on how data science and laboratory automation can be used to streamline discovery in the chemical sciences.

Natalie S. Eyke completed her B.S. in chemical engineering at the University of Michigan in 2014. After graduating, she joined the Chemical Engineering Research & Development department at Merck & Co., Inc., where she worked on process development for small molecule pharmaceuticals. In 2017, she began a Ph.D. in chemical engineering at the Massachusetts Institute of Technology (MIT), where she works for Professors Klavs F. Jensen and William H. Green. Her research focuses on combining active machine learning and high-throughput experimentation to facilitate reaction screening.

Klavs F. Jensen is the Warren K. Lewis Professor in Chemical Engineering and Materials Science and Engineering at the Massachusetts Institute of Technology. He is a co-director of MIT's Pharma AI consortium that aims to bring machine learning technology into pharmaceutical discovery and development. He received his MSc in Chemical Engineering from the Technical University of Denmark (DTU) and his Ph.D. in chemical engineering from the University of Wisconsin-Madison. His research interests include on-demand multistep synthesis, methods for automated synthesis, and machine learning techniques for chemical synthesis and interpreting large chemical data sets. Catalysis, chemical kinetics and transport phenomena are also topics of interest along with development of methods for predicting performance of reactive chemical systems.

*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139
†ccoley@mit.edu
‡kfjensen@mit.edu

Contents

1 Abstract

This two-part review examines how automation has contributed to different aspects of discovery in the chemical sciences. In this first part, we describe a classification for discoveries of physical matter (molecules, materials, devices), processes, and models and how they are unified as search problems. We then introduce a set of questions and considerations relevant to assessing the extent of autonomy. Finally, we describe many case studies of discoveries accelerated by or resulting from computer assistance and automation from the domains of synthetic chemistry, drug discovery, inorganic chemistry, and materials science. These illustrate how rapid advancements in hardware automation and machine learning continue to transform the nature of experimentation and modelling.

Part two reflects on these case studies and identifies a set of open challenges for the field.

2 Introduction

The prospect of a robotic scientist has long been an object of curiosity, optimism, skepticism, and job-loss fear, depending on who is asked. As computing was becoming mainstream, excitement grew around the potential for logic and reasoning—the underpinnings of the scientific process—to be codified into computer programs; as hardware automation became more robust and cost effective, excitement grew around the potential for a universal synthesis platform to enhance the work of human chemists in the lab; and as data availability and statistical analysis/inference techniques improved, excitement grew around the potential for statistical models (machine learning included) to draw new insights from vast quantities of chemical information [1–7].

The confluence of these factors makes that prospect increasingly realistic. In organic chemistry, we have already seen proof-of-concept examples of the "robo-chemist" [8] able to intelligently select and conduct experiments [9–11]; there have even been strides made toward a universal synthesis platform [12–15], theoretically capable of executing most chemical processes but highly constrained in practice. While there have been fewer success stories in automating drug discovery holistically [16], the excitement around machine learning in this application space is especially apparent, with dozens of start-up companies promising to revolutionize the development of new medicines through artificial intelligence [17].

A more pessimistic view of automated discovery is that machines will never be able to make real "revolutions" in science because they necessarily operate within a specific set of instructions [18]. This attitude is exemplified by *Lady Lovelace's objection*: "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of

anticipating any analytical relations or truths" [1]. Some have expressed a milder sentiment, perhaps in light of advances in computing, cautioning that an increasing reliance on robotic tools might reduce the odds of a serendipitous discovery [6]. Muggleton is more declarative, stating that "science is an essentially human activity that requires clarity both in the statement of hypotheses and their clear and undeniable refutation through experimentation" [19]. However, there is little disagreement that automation and computation in science has improved productivity through efficiency, reduction of error, and the ability to address large-scale problems [20].

In the remainder of Part 1, we will discuss the different types of discovery typically reported in the chemical sciences and how they can be unified as searches in a high-dimensional design space. Along with this definition comes a recommended set of questions to ask when evaluating the extent to which a discovery can be attributed to automation or autonomy. We will then discuss a number of case studies arranged in terms of the type of discovery being pursued and the nature of the approach used to do so. Part 2 will reflect on these case studies and make explicit what we believe to be the primary obstacles to autonomous discovery.

3 Defining discovery

3.1 Classifications of discoveries

There is no single definition of what constitutes a scientific discovery. Valdés-Pérez defines discovery as "the generation of novel, interesting, plausible, and intelligible knowledge" [5]. Data-driven knowledge discovery, specifically, has been defined as the "nontrivial extraction of implicit, previously unknown, and potential useful information" [21]. Each of these criteria, however, is inherently subjective. "Novel" is simultaneously ambiguous and considered distinct from "new"; it is generally meant to indicate some level of nonobviousness or, by one definition, a lack of predictability [22]. However, if we artificially limit what we consider to be known and demonstrate a successful extrapolation to a conclusion that really *was* known, it would be reasonable to argue that this does not constitute a discovery. This connects to the question of what it might mean for a discovery to be "interesting" or "useful", for which we avoid providing a precise definition.

For the purposes of this review, we instead define three broad types of discoveries in the chemical sciences (Figure 1) and provide examples of each.

Physical matter. Often, the ultimate result of a discovery campaign is the identification of a molecule (not discounting macromolecules), material, or device that achieves a desired function. This category encompasses most drug discovery efforts, where the output may be new chemical matter that could later become

10.1002/anie.201909987

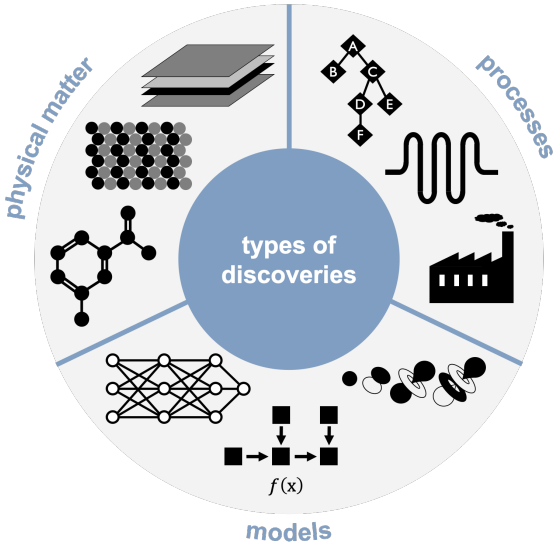


Figure 1: The three broad categories of discovery described in this review: physical matter, processes, and models.

part of a therapeutic, as well as materials discovery for a variety of applications.

Processes. Discoveries may also take the form of processes. These may be abstract, like the Haber-Bosch process, pasteurization, and directed evolution. They are more often concrete, like synthetic routes to organic molecules or a specific set of reaction conditions to achieve a chemical transformation.

Models. Our definition of a model includes empirical models (such as those obtained through regression of experimental data), structure-function relationships, symbolic regressions, natural laws, and even conceptual models that provide mechanistic understanding. It is common for models to be *part* of the discovery of the other two types as surrogates for experiments, as will be seen in many examples below.

The most famous examples of scientific discoveries in chemistry tend to be natural laws or theories that are able to rationalize observed phenomena that previous theories could not. Mendeleev's periodic table of the elements, Thomson's discovery of the electron, Rutherford's discovery of atomic nuclei, Kekulé's structure of benzene, etc. In their time, these represented radical departures from previous frameworks. Identifying such models through computational or algorithmic approaches would require substantially more open-ended hypothesis generation than what is currently possible.

3.2 Discovery as a search

We argue that the process of scientific discovery can always be thought of as a search problem, regardless of the nature of that discovery [7, 23, 24].

Molecular discovery is a search within "chemical space" [25–28]–an enormous combinatorial design space of theoretically-possible molecules. An oft-cited estimate of its size, considering only small molecules made up of CHONS atoms, is 10^{60} [29]; for any one application or with reasonable restrictions (e.g., on drug-likeness or synthetic accessibility), the size of the *relevant* chemical space will be significantly smaller [30, 31]. Biological compounds exist in an even larger space if one considers that there are, e.g., 20^{100} theoretically-possible 100-peptide proteins using only canonical amino acids, although again the number that are foldable and biologically relevant will be significantly smaller. Materials discovery is another combinatorial design space, where structural composition must be defined by both discrete variables (e.g., elemental identities) and continuous variables (e.g., stoichometric ratios) and processing conditions. The design space for a device is even larger, as it compounds the complexity of its constituent components with additional considerations of its geometry.

Discovering a chemical or physical process is the result of searching a design space defined by process variables and/or sequences of operations. For example, optimizing a chemical reaction for its yield might involve changing species' concentrations, the reaction temperature, and the residence time [32]. It may also include selecting the identity of a catalyst as a discrete variable [33], or changing the order of addition [34]. A new research workflow can be thought of as the identification of actions to be taken and their timing, such as the development of split-and-pool combinatorial chemistry for diversity-oriented synthesis [35] or a screening and selection strategy for directed evolution [36].

The majority of models that are "discovered", under our broad definition, are empirical relationships that come from data fitting. In these cases, the search space is well-defined once an input representation (e.g., a set of descriptors or parameters) and a model family (e.g., a linear model, a deep neural network) are selected. While this can present a massive search space when considering all possible values of all learned parameters (e.g., for deep learning regression techniques), the final model is often the result of a simplified, *local* search from a random initialization (e.g., using stochastic gradient descent). Symbolic regressions are searches in a combinatorial space of input variables and mathematical operations [37]. More abstract models, like mechanistic explanations of natural phenomena, exist in a high-dimensional hypothesis space that is difficult to formalize; automated discovery tools that are able to generate causal explanations do so using simplified terminology and well-defined ontologies [38].

In virtually every case of computer-assisted discovery, the actual search space is significantly larger than

what the program or platform is allowed to explore. We might decide to focus our attention on a specific set of compounds (e.g., a fixed scaffold), a specific class of materials (e.g., perovskites), a specific step in a catalyst synthesis process with a finite number of tunable process variables (e.g., the temperature and time of an annealing step), or a specific hypothesis structure (e.g., categorizing a ligand's effect on a protein as an agonist, antagonist, promoter, etc.). Constraining the search space is one way of integrating domain expertise/intuition into the discovery process. Moreover, it can greatly simplify the search process and mitigate the practical challenges of automated validation and feedback.

3.3 The role of validation and feedback

The way that we navigate the search space in a discovery effort is often iterative. Classically, the discovery of physical matter, such as in lead optimization for drug discovery, is divided into stages of design, make, test. An analogous cycle for searching hypothesis space could be described as hypothesize, validate, revise beliefs.

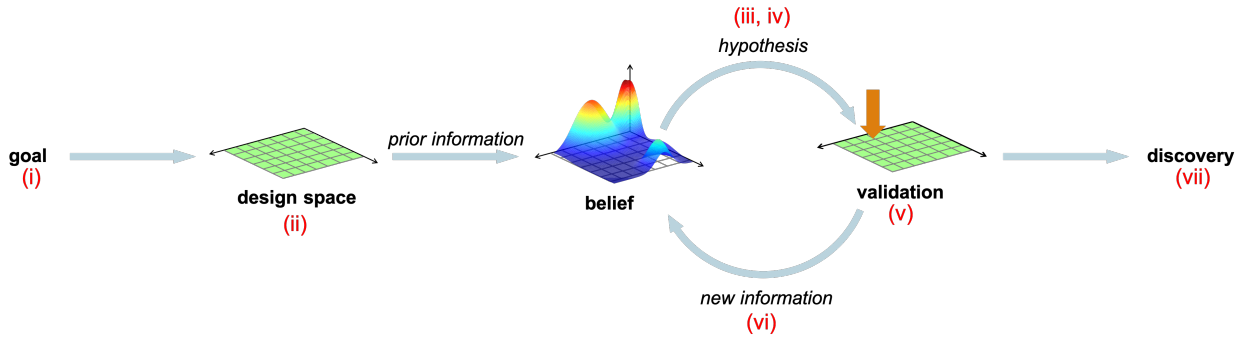


Figure 2: Simplified schematic of a hypothesis-driven (or model-driven) discovery process. When not proceeding iteratively, *new information* is not used to revise our belief (current knowledge). Lowercase roman numerals (red) correspond to the questions for assessing autonomy in discovery.

This third step–test or revise beliefs–helps to explain the role of validation and feedback in discovery: experiments, physical or computational, serve to support or refute hypotheses. When information is imperfect or insufficient to lead to a confident prediction, it is important to collect new information to improve our understanding of the problem. This might mean taking an empirical regression fit to a small number of data points, evaluating our uncertainty, and performing follow-up experiments to reduce our uncertainty in regions where we would like to have a more confident prediction (Figure 2). Purely virtual screening is not sufficient for drug discovery [39], where experimental validation continues to be essential [40]; Schneider and Clark describe experimental testing of drugs designed using *de novo* workflows as a "non-negotiable" criterion [41]. Similarly, Halls and Tasaki consider synthesis, characterization, and testing as critical components of

10.1002/anie.201909987

materials discovery [42]. The scope of hypotheses that lend themselves to automated validation has limited the scope of discovery tasks that are able to be automated.

Consider a scenario where we have a large data set of molecular structures and a property of interest, like their *in vitro* binding affinity for a particular protein target. We can perform a statistical regression to correlate the two and represent our understanding of the structure-function landscape. Based on that model, we may propose a new structure–a compound not yet tested–that is predicted to have high activity. Whether that constitutes discovery of the compound is ambiguous. Using scientific publication as the bar, it is reasonable to expect a high degree of confidence, regardless of whether that confidence arises from a statistical analysis of existing data or from confirmation through acquisition of new data. Even with a highly accurate model, performing a large virtual screen could lead to thousands of false positive results [31]. For a philosophical discussion about the nature of knowledge and need for confidence, correctness, and justification, see the Gettier problem in ref. 43.

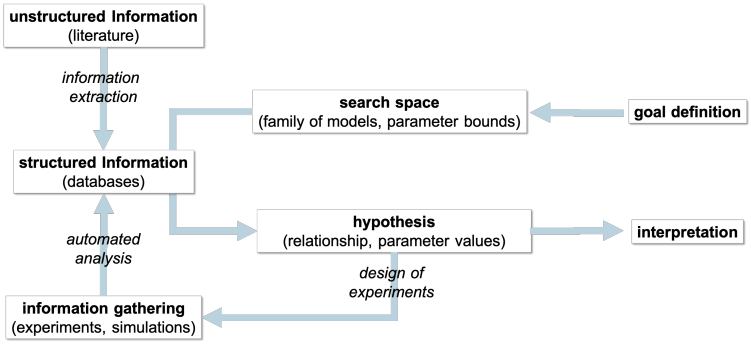


Figure 3: One way to visualize the discovery process. The goal definition will implicitly or explicitly define the search space within which we operate. Available structured information can be used to generate or refine a hypothesis within that search space. Often, when we are doing more than pure data analysis, there will be an iterative process of information gathering prior to the final output, or interpretation.

We note here that this hypothesis-first approach to discovery (Figure 3) is consistent with the philosophy of Popper [44]. This is in contrast to an observation- or experiment-first approach, which is more consistent with the philosophy of Bacon [45]; data mining studies tend to be Baconian [46]. In practice, when discovery proceeds iteratively, the distinction between the two is simply where one enters the cycle. Both are types of model-guided discovery, which is distinct from brute-force screening or approaches relying solely on serendipity (*vide infra*).

4 Elements of autonomous discovery

It is impossible to imagine conducting research without some degree of machine assistance, defining "machine" broadly. We rely on computers to organize, analyze, and visualize data; analytical instruments to queue samples, perform complex measurements, and convert them into structured data. However, it is important to consider precisely what is facilitated by automation or computer-assistance in terms of the broader discovery process. Many technologies (e.g., NMR sample changers) add a tremendous amount of convenience and reduce the manual burden of experimentation, but provide only a modest acceleration of discovery rather than a fundamental shift in the way we approach these problems. Considering the cognitive burden of experimental design and analysis connects to the distinction between autonomy and automation. A toy slot car that sets its own speed as it proceeds through a fixed track is qualitatively different from a self-driving car in the city, yet each successfully operates within its defined environment. Though there is no precise threshold between automation and autonomy, autonomy generally implies some degree of decision-making and adaptability in response to unexpected outcomes.

4.1 Assessing autonomy in discovery

Here, we propose a set of questions to ask when evaluating the extent to which a discovery process or workflow is autonomous: (i) How broadly is the goal defined? (ii) How constrained is the search/design space? (iii) How are experiments for validation/feedback selected? (iv) How superior to a brute force search is navigation of the design space? (v) How are experiments for validation/feedback performed? (vi) How are results organized and interpreted? (vii) Does the discovery outcome contribute to broader scientific knowledge? These questions are mapped onto the schematic for hypothesis-driven discovery in Figure 2.

(i) **How broadly is the goal defined?** While algorithms can be made to exhibit creativity (e.g., coming up with a unique strategy in Go or Chess [47, 48]), at some level, they do so for the sake of maximizing a human-defined objective. Is the goal defined at the highest level possible (e.g., find an effective therapeutic)? Or is it narrow (e.g., find a molecule that maximizes this black-box property for which we have an assay and preliminary data)? The higher the level at which the mission can be defined, the more compelling the discovery becomes. That requires platforms to understand what experiments can be performed and how they are useful for the task at hand.

(ii) **How constrained is the search/design space?** An unconstrained search space is one that *we* operate in as human researchers. There are many ways in which humans can artificially constrain the search space available to an autonomous platform. A maximally constrained search space in the discovery

of physical matter could be a (small) fixed list of candidates over which to screen. Limitations in the experimental and computational capabilities of an autonomous platform have the effect of constraining the search space as well; some have described the scientific process as a dual search in a hypothesis space and experimental space [24, 49]. How these constraints are defined influences the difficulty of the search process, the likelihood of success, and the significance of the discovery. The fewer the constraints placed on a platform, the greater the degree to which it can be said to be operating autonomously.

(iii) **How are experiments for validation/feedback selected?** Unconstrained experimental design is a complex process requiring evaluation of local decisions as well as a global strategy for the overall timeframe, coherency, and scientific merit of a proposed experiment [50]. When operating within a restricted experimental space, design can be simplified to local decisions of specific implementation details without these high-level decisions. Cummings and Bruni define a taxonomy for human-automation collaboration in terms of the three primary roles played by a human or computer: moderator (of the overall decision-making process), generator (of feasible solutions), and decision-maker (of which action to take) [51]. Their levels of automation include ones where humans must take all decisions/actions, where the computer narrows down the selection, where the computer executes one if the human approves, and where the computer executes automatically and informs the human if necessary. The second level is typical for the discovery of new physical matter, where computational design algorithms may propose compounds that are subjected to a manual assessment of synthesizability before being manually synthesized. The smaller the search space and the cheaper the experiments–including considerations of time and risk of failure–the less human intervention is required in selecting experiments.

(iv) **How superior to a brute force search is navigation of the design space?** This question seeks to identify the extent to which there is "intelligence" in the search strategy. Langley et al.'s notion of discovery as a heuristic search emphasizes this criterion [23]. Whether or not the strategy is more effective than a brute force search depends on the size of the space and how experiments are selected. For example, a high throughput screen of compounds from a fixed library is equivalent to a brute-force search. An active learning strategy designed to promote exploration might require only 20% of the experiments to find an optimal solution. When dealing with continuous (e.g., process variables) or virtually infinite (e.g., molecular structure) design spaces, it is not possible to quantify meaningfully the number of experiments in a brute-force search.

(v) **How are experiments for validation/feedback performed?** Being able to automatically gather new information to support/refute a hypothesis is an important aspect of an autonomous discovery workflow. At one extreme, experiments are performed entirely by humans (regardless of how they are

proposed); in the middle, experiments are performed semi-automatically but require significant human set-up between experiments; at the other extreme, experiments can be performed entirely without human intervention. This question is tightly coupled to that of who chooses the experiments and the size of the search space. The narrower the experimental design space, the more likely it is that validation/feedback can be automated. In computational studies, it is relatively straightforward to automate simulations if we are willing to discard failures without manual inspection (e.g., simulations that fail to converge).

(vi) **How are results organized and interpreted?** In an iterative workflow, the results of information gathering (experiments, simulations) are organized as structured information and used to update our prior knowledge and revise our beliefs before the next round of experimental design. Provided that the experiments/simulations can be designed to produce information that is already in a compatible format (e.g., quantifying a reaction yield to build a model of yield as a function of process variables), this is simply a practical step toward closing the loop. In a few specialized workflows, experimental results naturally drive the selection of subsequent experiments, as in directed evolution and phage-assisted continuous evolution [52].

(vii) **(optional) Does the outcome contribute to broader scientific knowledge?** Though not necessarily related to the concept of autonomy, this question speaks to impact and intelligibility. Does it require extensive interpretation after the fact to evaluate *how* or *what* it has learned, or is it self-explanatory? Intelligibility is one of the criteria for discovery put forward by Valdés-Pérez [5], among others. Describing physical phenomena requires far less domain knowledge than does explaining those phenomena [53]. Especially in empirical modeling, there is often a dichotomy between models built for accurate predictions and models built for explanatory predictions [54, 55]. Turing made note of this as early as 1950, saying that "an important feature of a learning machine is that its teacher will often be very largely ignorant of quite what is going on inside, although he may still be able to some extent to predict his pupil's behavior" [1]. The past few years have seen an interest in the transparency, interpretability, and explainability of machine learning models, not just the accuracy [56].

Several of these questions probe the extent to which discovery is "closed loop", which implicitly assumes an iterative process of multiple hypothesize-test-revise beliefs cycles. Iterative refinement is crucial when operating inside poorly-explored design spaces (e.g., using an uncommon scaffold) or with new objective functions (e.g., maximizing binding to a new protein target *in vitro*). Most of the case studies described in the following sections are better described as "open loop" and involve only certain aspects of the workflow in Figure 2. For example, a common paradigm of computer-aided discovery is to define an objective function,

perform a large-scale data mining study, propose new molecules or materials, and manually validate a small number of those predictions. Waltz and Buchanan describe many early computational discovery programs as merely running calculations, rather than trying to close the loop [20].

4.2 Enabling factors

A confluence of improved data availability, computing abilities, and experimental capabilities have brought us substantially closer to autonomous discovery (Figure 4). These improvements contribute to two categories of methodological progress: (1) techniques for navigating the search space more effectively, and (2) techniques for accelerating validation/feedback. Many machine learning techniques, for example, have been used to build empirical models within the search space to enable or accelerate the search; mapping the design space for a molecule, material, device, or process to relevant performance metrics is a prerequisite for any "rational design".

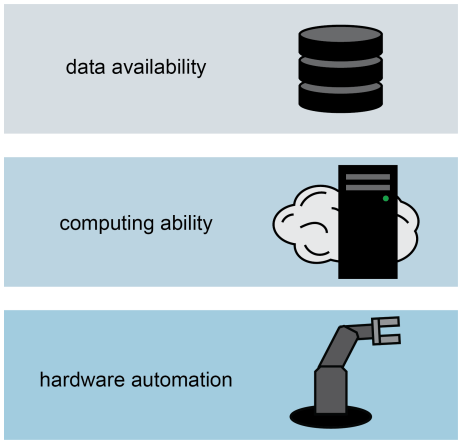


Figure 4: The factors that have enabled autonomous discovery fall into one of three main categories.

As Claus and Underwood point out, effective discovery requires assimilation of knowledge contained in large quantities of data of diverse types [57]. The quantity of chemical property and process data available in journals, the patent literature, and online databases makes it challenging to analyze by hand. Digitization of organic reaction information into computer-readable databases like Reaxys, SPRESI, CASREACT, and Lowe's USPTO dataset has not just facilitated searching that information, but has enabled new analyses thereof [58]. Millions of bioactivity measurements are found in databases like ChEMBL and PubChem [59]. There are also many repositories for experimental and computational properties of materials, which have facilitated the construction of empirical models to predict new material performance [60, 61]. Gil et al. [62] discusses the utility of AI techniques in searching and synthesizing large amounts of information as part of "discovery informatics" [57, 63, 64]. Even now, an enormous amount of untapped information remains

housed in laboratory notebooks and journal articles. For such information to be directly usable, someone must undertake the challenge of compiling the data into an accessible, user-friendly format and overcome any intellectual property restrictions. Image and natural language processing techniques can make this task less burdensome, thus there is increasing interest in applying such techniques to the chemical sciences [65–70].

Autonomous discovery systems rely on a variety of computational tools to generate hypotheses from data without human intervention. This includes both the software that makes the recommendations (e.g., proposes correlations, regresses models, selects experiments) as well as the underlying hardware that makes using the software tractable. Our discussion of the advances in this area focuses on software developments with an emphasis on machine learning algorithms, which have elicited cross-disciplinary excitement [71–74].

Typically, search domains that are of interest for discovery are characterized by high dimensionality (e.g., chemical space). In such domains, the patterns within the available data may be beyond the capacity of humans to infer *a priori*. Machine learning and pattern recognition algorithms can be used to discover these regularities automatically, e.g., by using the available data to parameterize a neural network model [75]. Varnek and Baskin and Mitchell provide overviews of machine learning techniques as applicable to common cheminformatics problems [76, 77] and brief tutorials can be found in a number of reviews [78–81]. It is becoming increasingly common to use machine learning to develop empirical quantitative structure-activity/property relationships (QSARs/QSPRs) for virtual screening and as part of broader discovery frameworks [82]. These models can be used to distinguish promising compounds from unpromising ones and prioritize molecules for synthesis and testing, thus facilitating the extrapolation of information about existing molecules to novel molecules that exist only *in silico* [31].

Algorithms that enable efficient navigation of design spaces represent an important set of computing advances. Even with a model representing our belief about a physical structure-property relationship, an algorithmic framework is needed to apply that belief to experimental design. These frameworks include active learning strategies [83] that aim to maximize the accuracy of predictive models while minimizing the required training data, as well as goal-directed strategies such as Bayesian optimization [84] and genetic algorithms [85]. These iterative techniques can reduce the experimental burden associated with discovery in domains or search spaces where exhaustive testing is not practical.

Algorithms that are capable of directly proposing candidate molecules or materials (physical matter) as a form of experiment selection are worth special emphasis. Recently, deep generative models [86] such as generative adversarial networks (GANs) [87] and variational autoencoders (VAEs) [88] have attracted a great deal of interest, as they facilitate the creation of diverse molecular libraries without relying on systematic enumeration [89–91]. Many of the case studies below leverage these and related frameworks for the discovery of physical matter.

10.1002/anie.201909987

Experimental advances toward autonomous discovery include automation (along with parallelization and miniaturization) of well-established laboratory workflows as well as entirely novel synthetic and analytical methodologies. Aspects of experimental validation (Figure 5) have existed in an automated format for decades (e.g., addition to and sampling from chemical reactors [92, 93]), and many of the requisite hardware units have been commercialized (e.g., liquid handling platforms and plate readers available through companies such as Beckman, Hamilton, BioTek, and Tecan). However, moving beyond piecemeal automation to the entire experimental burden of discovery workflows is challenging. Each process step, which may include synthesis, purification, assay preparation, and analysis, must be seamlessly integrated for the platform to operate without manual intervention; each interface presents new potential points of failure [94].

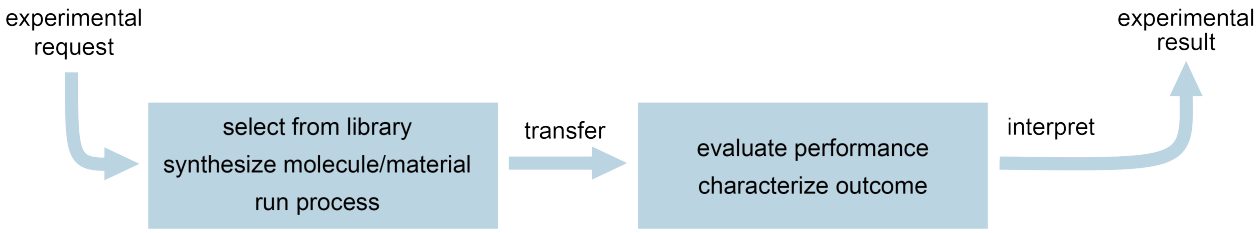


Figure 5: Generic workflow for experimental validation.

The complexity of software required for hardware automation ranges from sequencing commands from a fixed schedule [15], to real-time control and optimization [10], to higher-level scheduling and orchestration [95]; user interface driven software such as LabVIEW [96] can aid the creation of fit-for-purpose control systems with minimal programming experience. Although end-to-end automation of an experimental discovery workflow is uncommon, there are numerous benefits to be gained even from partial automation, chief among these being standardization and increased throughput [97].

In addition to automation, novel experimental methodologies have been developed that lend themselves particularly well to autonomous discovery workflows by facilitating the exploration of broad design spaces, a helpful feature that increases the likelihood of discovery [98]. These include synthesis-focused methodologies, such as DNA-encoded libraries [99] and diversity-oriented synthesis [35, 100], as well as analysis-focused methodologies, such as ambient mass spectrometry [101] and MISER for accelerating liquid chromatographic analysis [102].

The three categories of enabling factors described herein facilitate discovery in different ways: data is leveraged to create models that inform and predict, computational tools are used to create models from data and reason about which experiments to perform next, and physical (or computational) experiments validate hypotheses and facilitate refinement thereof. These factors can be strategically combined to give rise to different types of studies. For example, the experimental capabilities described here, in isolation, can be used

for high-throughput, brute-force screening; computational tools can be used for data generation (through, e.g., DFT simulations); virtual screening is achieved through the combination of data and algorithms; and integration of all three is needed for fully autonomous discovery.

5 Examples of (partially) autonomous discovery

In this section, we summarize a series of case studies that demonstrate how automation and machine autonomy influence discovery in various research domains. The extent to which techniques in automation and computation have enabled each case varies. Some only benefit from automated laboratory hardware, others learn underlying trends from large or complex data, and still others use computational techniques to efficiently explore high dimensional design spaces.

5.1 Foundational computational reasoning frameworks

There has been a long-standing fascination with the question of whether it is possible to codify and automate the process of discovery [103]. In the 1980s and 1990s, several programs were developed to mimic a codifiable approach to discovery and to reproduce specific quintessential discoveries *of models*, led by Langley and Zytkow [6]. These programs focused on questions of model induction and hypothesis generation (as a form of data analysis) rather than experimental selection and automated validation/feedback.

BACON is a rule-based framework introduced in 1978 to formalize the Baconian method of inductive reasoning to discover empirical laws, supplemented with data-driven heuristics [104]. BACON.4, a later iteration specifically designed for chemical problems, searched for arithmetic combinations of input variables to identify regularities in data (e.g., noting that pressure times volume is invariant for constant temperature in a closed gas system) [7]. This approach was able to recapitulate Ohm's law, Archimedes' law of displacement, Snell's law, conservation of momentum, gravitation, and Black's specific heat law [105]. The search for an empirical relationship was greatly simplified by excluding any irrelevant variables (i.e., all input variables were known to be important) and eliminating all measurement noise. Extensions of this approach included describing piecewise functions (FARENHEIT [106]) and coping with irrelevant observations and noise (ABACUS [107]). More recently, Schmidt and Lipson demonstrated that using a symbolic regression framework similar to BACON, it is possible to rediscover Hamiltonians, Lagrangians, and geometric conservation laws from empirical motion tracking data [37]. Much like its predecessors, their program uses a two-part process of generating and scoring hypothesized analytical laws.

The STAHL program developed by Zytkow and Simon in the mid-1980s sought to automate the construction of compositional models to, e.g., rediscover Lavoisier's theory of oxygen [108]. It operates on a

list of chemical reactions to produce a list of proposed chemical elements and the compounds they make up by making inferences like "A + B + C $\longrightarrow$ B + D" $\implies$ "D is composed of A and C". While the program was arguably successful in formalizing a specific form of scientific reasoning, the lack of any consideration for stoichometry, phase changes, and ability to consider uncertainty, competing hypotheses, and request information makes such a logic framework limited in utility. The KEKADA program [109] was designed with those abilities in order to replicate the discovery of the Krebs cycle. Using seven heuristic operators (hypothesis proposers, problem generators, problem choosers, expectation setters, hypothesis generators, hypothesis modifiers, and confidence modifiers) and simulated experiments of metabolic reactions, KEKADA was able to rediscover the Krebs cycle from the same empirical data that would have been obtainable at the time.

The knowledge bases for these early programs were comprised of expert-defined relationships, rules, and heuristics designed to reflect *prior knowledge* and bring the programs up to the level of domain experts. Programs based entirely on user-defined axioms have proved successful in automatic theorem generation in graph theory [110]. However, these rules bring at least two drawbacks in the context of inductive reasoning. The first is that it is more difficult for experts to recapitulate their knowledge through rules than by providing examples from which an algorithm can generalize [111]. The second is that too stringent priors may restrict the model from deviating far enough from existing theory to make a substantial discovery and merely "fill in the gaps" of what is known. Kulkarni and Simon argue that a lack of prior knowledge about allowed/disallowed reactions actually served to *benefit* Krebs, as a formally trained chemist might not have pursued a hypothesis that was–at the time–believed to be highly unlikely [109].

5.2 Discovery of mechanistic models

5.2.1 Discovery of detailed kinetic mechanisms

Computer assistance has proved useful in the exploration and simulation of reaction pathways [112–115]. The vast number of possible elementary reactions creates a combinatorial space of hypothesized pathways that is difficult to explore manually in an unbiased manner, making it a prime candidate for algorithmic approaches. Millions of possible elementary reactions can be generated even with species of just a few atoms [116]. One such approach, MECHEM, enumerates elementary reactions in catalytic reaction systems to identify series of mechanistic steps able to rationalize an observed global reaction [117–119]. Ismail et al. have demonstrated a similar approach to identifying multi-step reaction mechanisms for catalytic reactions using a ReaxFF potential energy surface [121] to guide the search toward kinetically-likely pathways [120].

The Reaction Mechanism Generator (RMG) fills a similar role in developing detailed kinetic mechanisms for combustion and pyrolysis processes [122]. Expert-defined reaction templates enumerate potential ele-

mentary reactions between a set of user-defined input molecules; rate constants for the forward and reverse reactions are estimated from a combination of first principles calculations (e.g., DFT) and group additivity rules regressed to experimental data.

The ability to estimate kinetic and thermodynamic parameters enables the identification of new elementary reactions and pathways and, e.g., exploration of untested fuel additives' effects on ignition delay [123]. Earlier studies by Broadbelt et al. used a similar approach to develop detailed kinetic models for pyrolysis reactions [124].

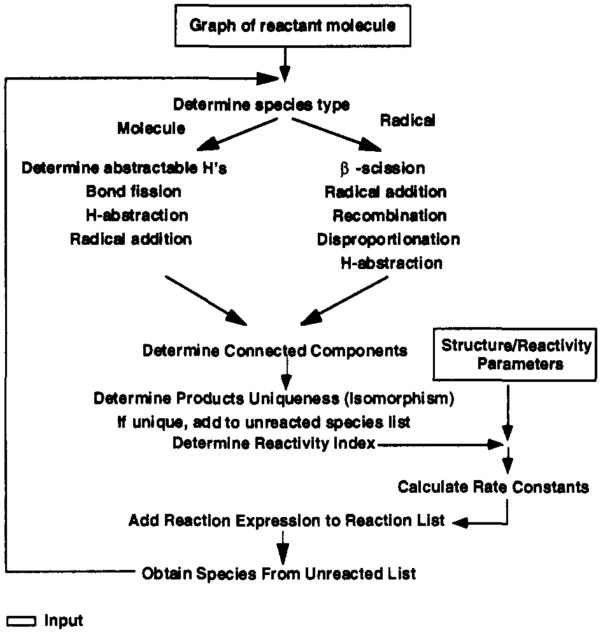


Figure 6: Discovery of detailed kinetic models through iterative selection of important elementary reaction steps. Figure reproduced from Broadbelt et al. [124].

Mechanistic enumerations/searches have been applied extensively to the discovery of transition states and reaction channels [125–127]. These methods represent a search in the $(3N-6)$-dimensional potential energy surface landscape implicitly defined by an N-atom pool of reacting species. Approaches like Berny optimization [128] are used to identify transition state (TS) geometries for the purposes of estimating energetic barrier heights. Double-ended search methods like the freezing string method (FSM [129]) or growing string method (GSM [130]) require knowledge of the product structure and run iterative electronic structure calculations to identify a plausible reaction pathway; these can be applied to the discovery of new elementary reactions by systematically enumerating potential product species [131–133] (Figure 7). Single-ended search methods operate on reactant species only and perturb the geometry along reactive coordinates, including, e.g., the artificial force induced reaction method (AFIR [134]).

An alternate approach to reaction discovery is by direct simulation of reactive mixtures using molecular

dynamics (MD) [135–137]. Wang et al. describe the use of an "*ab initio* nanoreactor" to find unexpected products from starting materials similar to the Urey-Miller origin of life experiment [135]. Importantly, this approach does not require heuristics to define reaction coordinates or enumeration rules to define possible products. Instead, molecules in an MD simulation are periodically pushed toward the center to impart kinetic energy and encourage collisions at a rate that enables the observation of rare events over tractable simulation timescales. In principle, these can be to applied to the prospective prediction of novel reaction types and, ultimately, the development of new synthetic methodologies.

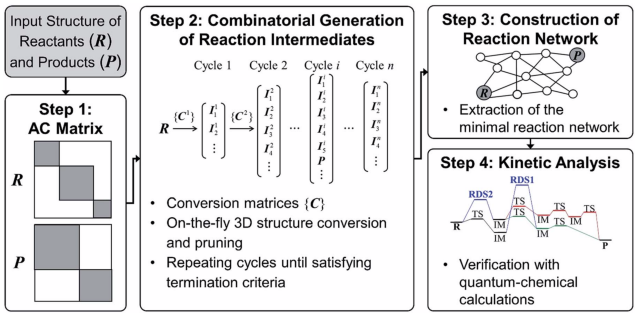


Figure 7: Workflow for identification of reaction networks between known reactants R and known products P through combinatorial enumeration of possible mechanistic steps pruned by calculated transition state energies. Figure reproduced from Kim et al. [133].

5.3 *Noniterative* discovery of chemical processes

5.3.1 Discovery of new synthetic pathways

Synthetic pathways are a prerequisite for physically producing a molecule of interest, whether for experimental validation of a predicted property or for production at scale. Retrospective analyses of known single-step chemical reactions can yield hypothesized synthetic pathways as combinations thereof. Gothard et al. describe an analysis of seven million reactions from the Beilstein database for the discovery of one-pot reactions; their search space comprised any consecutive sequence of known reactions where the product of one is a reactant of another [138]. Candidate sequences were evaluated using eight filters, including a 322×322 table of functional groups and their cross-reactivity and a 322×97 table of their compatibility under 97 categories of reaction conditions. Through application of these expert heuristics to millions of candidate sequences, the authors identified multi-step chemistries that could potentially be run without an intermediate purification, choosing a handful of such pathways for experimental validation. While their filters were all hand-encoded, data mining techniques can also be used to estimate functional group reactivity [139, 140]. Selecting pathways within a search space defined by combinations of known single-step reactions has taken on other forms as well, including the identification of cyclic pathways [141], the optimization of process cost [142], and the

optimization of estimated process mass intensity [143].

Generating yet-unseen chemical reactions for a synthesis plan–a necessity for the synthesis of novel molecules–is a harder search problem than when searching within a fixed reaction network [144]. Because the number of states in a naive retrosynthetic expansion will scale as b^d for branching factor b and depth d, guiding the search is an essential aspect of computer-aided synthesis planning (CASP) programs. The breadth of the search depends on the coverage of the rule sets: abstracted enzymatic reactions tend to number in the hundreds [145], expert transformation rules often number in dozens or hundreds [146, 147] but can extend into the tens of thousands in contemporary programs [148], and algorithmically-extracted templates generally number in the thousands to hundreds of thousands [149–152]. To the extent that reaction rules and synthetic strategies can be codified, synthesis planning is highly conducive to computational assistance [58, 153–157] (Figure 8). CASP approaches that generate retrosynthetic suggestions without the use of pre-extracted template libraries [158, 159] still result in a large search space of possible disconnections.

Even the earliest CASP programs emphasized the importance of navigating the search space of possible disconnections [153, 160]. The search in OCSS was guided by five subgoals for structural simplification: reduce internal connectivity, reduce molecular size, minimize functional groups, remove reactive or unstable functional groups, and simplify stereochemistry [160]. Starting material oriented retrosynthesis introduces additional constraints in the search, as the goal state is a specific starting material, rather than one of many from a database of available compounds [161]. It is only fairly recently that CASP tools have started to be used more widely for discovery of synthetic routes. Development is stymied by the complexities of validation and feedback, which can only occur by experimental implementation [162] or review by expert chemists [163].

There are two main approaches to navigating the search space during retrosynthetic expansion to determine which disconnections are most promising: value functions and action policies. Value functions estimate the synthetic complexity of reactant molecules as a proxy for how close they are to being purchasable [164–168]. Despite their limitations, these are widely used in virtual screening libraries as a rapid means of prioritizing compounds that appear more synthetically tractable. While even simple user-defined heuristics that attempt to break a molecule into the smallest possible fragments can be successful in planning full synthetic routes, learned value functions can offer some advantages in finding shorter pathways or being tailored to a user-defined cost function [169]. Action policies directly predict which transformation rule to apply based on literature precedents in a knowledge base; this can be accomplished through a simple nearest-neighbor strategy [170] or through a trained neural network model for classification [152]. The latter approach has been integrated into a Monte Carlo tree search framework to rapidly generate and explore the space of candidate pathways, resulting in recommendations that chemists considered equally plausible to literature pathways in a double-blind study [163]. Less common approaches to navigating the search space include

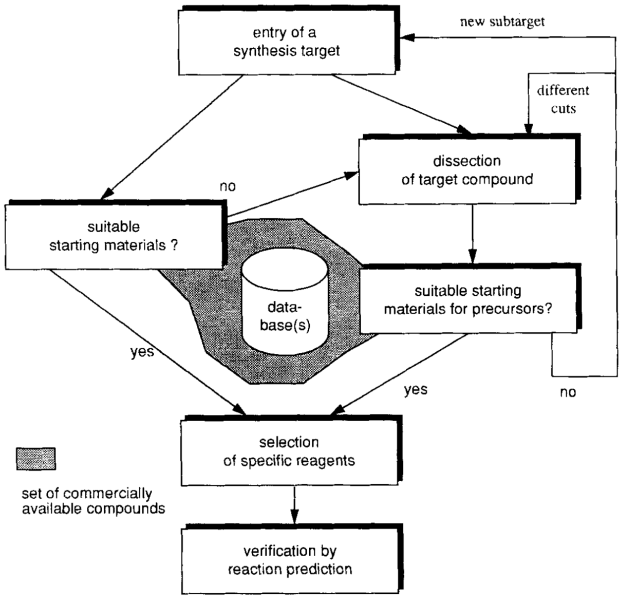


Figure 8: Workflow used by the WODCA program for computer-aided synthesis planning. Figure reproduced from Ihlenfeldt and Gasteiger [153].

proof-number search [171].

Reaction pathway discovery is relevant in synthetic biology and metabolic engineering contexts as well. For example, one study by Rangarajan et al. describes the application of Rule Input Network Generator (RING, [173]) to identify plausible production biosynthetic pathways through a heuristic-driven network generation and analysis [172]. Kim et al. review algorithms and heuristics used to explore metabolic networks and find optimal pathways [145]. A broader review of machine learning for biological networks can be found in ref. 174.

5.3.2 Discovering models of chemical reactivity

Identifying synthetic pathways is but one step toward fully automated synthesis. For any theoretical robochemist capable of synthesizing any molecule on demand [8, 14], these ideas must be able to be acted upon and executed in the laboratory. Even without automated synthesis, hypothesized synthetic pathways are of little use without experimental validation. This requires additional models of chemical reactivity that can, among other things, propose suitable reaction conditions, estimate the confidence in the reactions it proposes, and have some notion of why one set of substrates might achieve a higher yield than others. Models for these tasks can be trained directly on experimental data using a variety of statistical techniques.

Given a set of combinations of successful and unsuccessful reaction examples (i.e., high and low yielding), one can train a binary classifier model to predict whether a proposed set of reaction conditions will be success-

ful [175]. The same task can also be treated as a regression of reaction yields, rather than as a classification, as a function of substrate descriptors; a virtual screen of known conditions as a fixed search space can then propose substrate-dependent optimal conditions [176]. When *only* successful reaction examples are present, one can treat the selection of reaction conditions as a recommendation problem comprising a classification subproblem (for reagent, catalyst, solvent identity) and a regression subproblem (temperature) under the assumption that the "true" published conditions are adequate. This was Gao et al.'s approach using the Reaxys database to produce a model that proposes conditions at the level of species identity and temperature based on reactant and product structures [177]. In the process of learning the relationship between reactants/products and suitable reaction conditions, the model learns a continuous embedding for chemicals that reflects their function in organic synthesis, similar to how semantic meaning is captured by word2vec models [178]. Formulating condition selection as a data-driven classification problem has also been used in a more focused manner as an alternative to expert recommender systems [179], e.g., to choose phosphine ligands for Buchwald-Hartwig aminations [143] or catalysts for deprotections [140].

In some cases, computational prediction of solvation free energies can meaningfully assist in the selection of reaction solvents [180]. To a first approximation, solvation energy can be estimated by a linear model describing potential solute-solvent interactions [181, 182]. When those interaction parameters can be predicted via DFT, one can estimate the performance of a large virtual set of solvents, e.g., to optimize the rate constant for a particular reaction of interest [183].

Similar models for *a priori* evaluation of reaction conditions can be found in materials applications. In one instance, Raccuglia et al. used a combination of 3955 reaction successes and failures from laboratory notebooks to train an SVM model to predict outcomes for the crystallization of vanadium selenites [184] (Figure 9). Recasting the model as a decision tree led to correlations that reflected expert intuition, which arguably contributed to the synthesis of five previously-unseen compounds [185]. A similar study applied a much smaller dataset of 54 conditions to predict whether a process would produce atomically precise gold nanocrystals, using a siamese neural network architecture to relate proposed conditions to precedents [186]. For larger scale analyses, the literature serves as an unstructured data source of inorganic reactions and has been used to populate a structured database of synthesis conditions and outcomes via natural language processing of over 640,000 manuscripts [187]; virtual screening and synthesis planning pipelines have been built on top of such data to guide the experimental realization of computationally-proposed materials [68, 188–190].

Anticipating the outcomes of organic reactions is a very different modelling task. The space of possible results is high dimensional (chemical space) rather than low dimensional (e.g., the phase of the resulting material or a boolean measure of success/failure). The ability to accurately prediction reaction products would

10.1002/anie.201909987

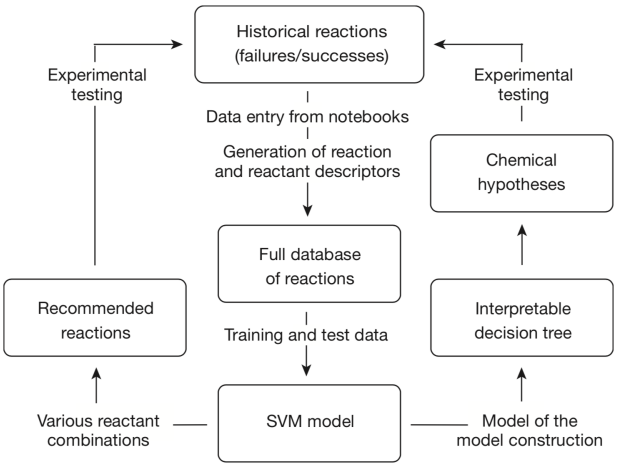


Figure 9: Workflow used by Raccuglia et al. for training an interpretable predictive model of the success/failure of vanadium selenite crystallization. Figure reproduced from Raccuglia et al. [184].

be powerful in combination with CASP to improve the likelihood that proposed reactions are experimentally realizable. The task of predicting reaction outcomes *in silico* has been approached through several heuristic and computational techniques over the years [191–196] but has seen renewed interest as a supervised learning problem as a result of increased data availability [157].

Segler and Waller treat reaction discovery as an edge prediction problem in a knowledge graph of known chemistry [197]. Specifically, they predict the products of bimolecular reactions through the application of algorithmically-extracted half reactions that similar substrates underwent. Novel combinations of half reactions that had not been observed previous could be accurately predicted, albeit with a modest rate of success. With a similar goal, Jacob and Lapkin build a stochastic block model (SBM) for the classification of reactions into true or false using reactions in Reaxys (true) and ones randomly generated from known chemicals (false) [198]. Other machine learning-based methods include ones that rank enumerated mechanistic [199–201] or pseudo-mechanistic [202] steps, score/rank reaction templates [152, 203], score/rank candidate products generated from reaction templates [204], propose reaction products as resulting from sets of graph edits [205, 206], and translate reactant SMILES strings to product SMILES strings using models built for natural language processing tasks [207–209]. These all formulate reaction prediction differently; for example, the model in ref. 206 learns to enumerate likely changes in bond order and learns to rank candidate products generated through combinatorial enumeration of those sub-reactions (Figure 10).

The *quantitative* prediction of reaction outcomes is closer to a standard regression task; when only one chemical species is varied–a single substrate or a single catalyst–the problem is exactly that of developing a QSAR/QSPR model. The historical approach in physical organic chemistry is again the development of linear free energy relationships [181], for which group contribution approaches are particularly attractive due

10.1002/anie.201909987

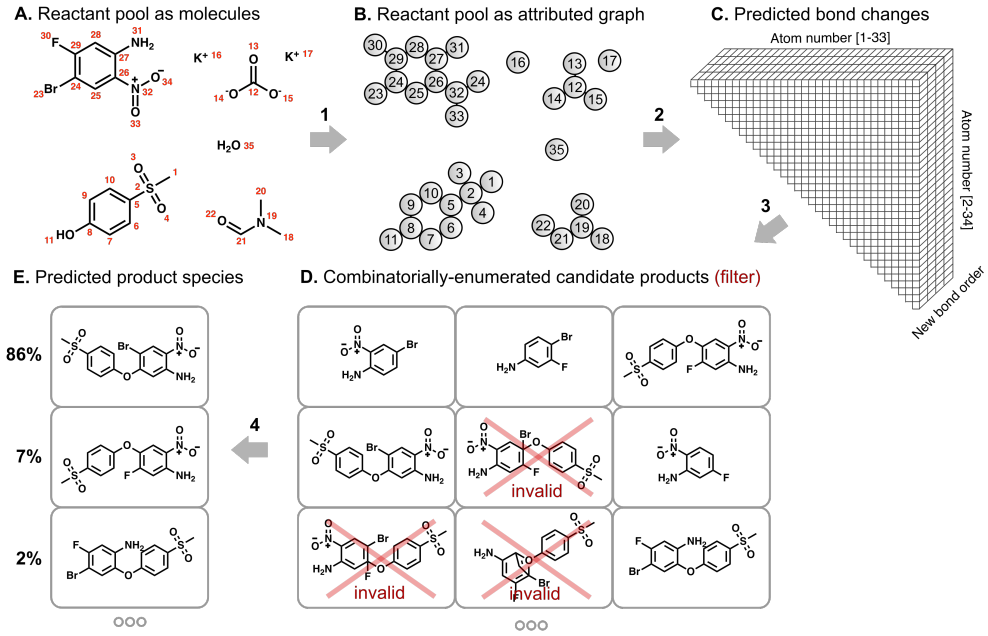


Figure 10: Workflow used by Coley et al. for predicting the products of organic reactions. Figure reproduced from Coley et al. [206].

to their interpretability [210]. Hammett parameters are a classic example of correlating molecular structures with reactivity [211]. Computational prediction of organic reaction rates has been demonstrated using simple regressions on expert descriptors [212] and using structure-derived descriptors [213–217] with much of the latter work coming from Varnek and coworkers.

Even with increasingly powerful machine learning techniques to describe patterns in experimental data, computational chemistry has a significant role in developing predictive models of chemical reactivity [218]. Using informative electronic (e.g., Fukui functions [219, 220]) and steric (e.g., Sterimol [221, 222]) descriptors can help model generalization and performance, especially in low data environments. Given suitable descriptors and holding other process parameters constant, complex properties have been described with linear or nearly-linear models, e.g., catalyst performance and enantioselectivity [223–227] (Figure 11). Descriptors tailored to a specific reaction class can be effective representations for predicting regioselectivity [228] and yield [229] among other performance metrics, although they may not be broadly applicable across reaction and substrate classes. In principle, these descriptors could be calculated with greater universality than expert-selected ones already known to be relevant [230]. Similarly, selectivity in complex synthetic steps can be explained by expert-defined DFT calculations [231] that could, in principle, be automated.

If these models are truly describing the underlying patterns of chemical reactivity, they could be applied prospectively to the discovery of new synthetic methods. This is yet to be demonstrated. Time-split validations arguably demonstrate this generalization ability, however, a separate algorithm (a hypothesis

10.1002/anie.201909987

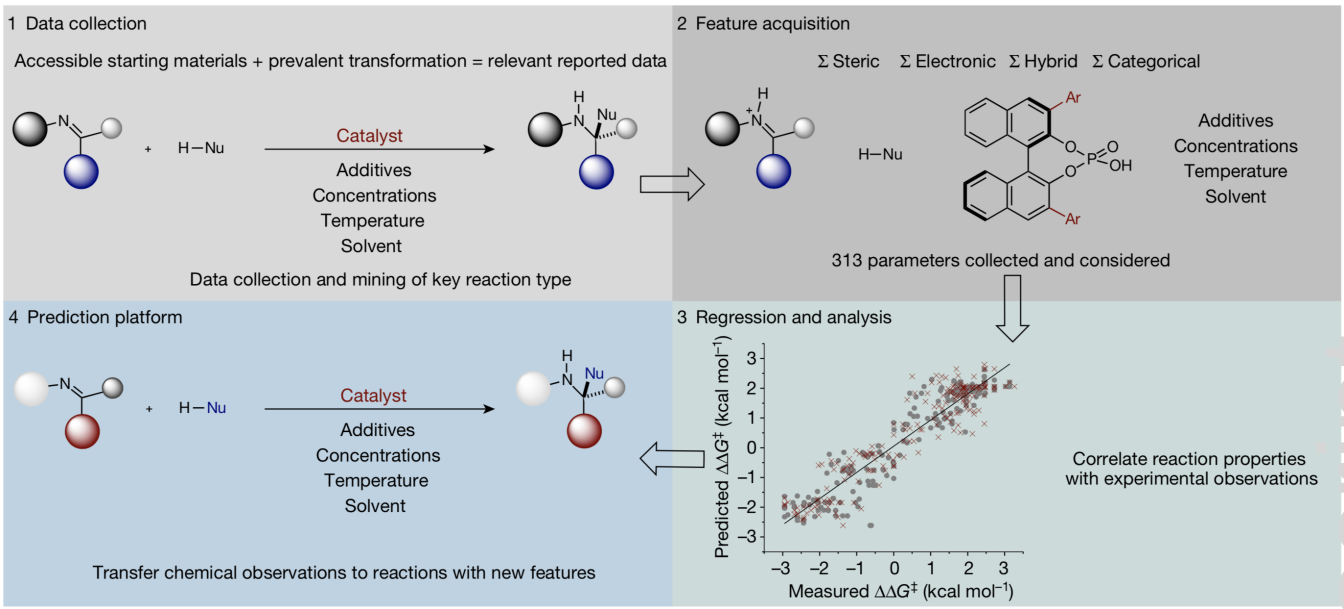


Figure 11: Discovery of catalysts for enantioselective catalysis using a surrogate model trained on experimental $\Delta\Delta G^{\ddagger}$ values to screen unseen reaction conditions. Figure reproduced from Reid and Sigman [227].

generator) would be required to "steer" these models toward the combination of reactants mostly likely to result in new chemistry.

5.3.3 Discovery of new chemical reactions from experimental screening

The discovery of new chemical reactions can widen synthetically-accessible chemical space and allow us to realize molecules that were previously difficult to access [232, 233]. The rise of combinatorial chemistry in the 1990s opened up new means of discovering new chemical reactions and functional physical matter through experimental screening [99, 234]. Low-volume liquid handling and rapid analysis by HPLC or ESI-MS or even fluorescent readouts [235] have enabled material-efficient reaction screening toward this end [236]. Microplate reaction screening has advanced to the point where it requires only nanomole quantities of material and achieves throughputs of thousands of reactions per hour [237, 238]; related technologies using continuous flow [239] and electrospray ionization [240] can achieve similar throughputs and material consumption.

These technologies have accelerated the rate at which candidate reactions (different substrates, conditions) can be tested, but still navigate a search space in a brute force manner. High throughput experimentation *can* be hypothesis-driven and used to investigate a narrower search space [241, 242] or be informed by mechanistic knowledge [243] and functional diversity [244, 245] (Figure 12), though this is less common in practice. Beyond improving the speed of experimentation and sensitivity of analysis, progress toward

automated discovery of new chemical reactions has included developing new techniques for exploring the vast space of possible chemical reactions with fewer individual experiments: either by an active search (*vide infra*) or by pooling. Clever pooling strategies allow for the simultaneous evaluation of multiple hypotheses through techniques like mass-encoded libraries [246], DNA-templated synthesis [247], and substrate combinations designed to enable straightforward deconvolution [248, 249] for "accelerated serendipity" [250]. Multicomponent reactions represent a particularly large space and have historically been discovered either through serendipity or pooled/combinatorial screening [251–253]. Collins et al. review screening approaches to reaction discovery and development [254].

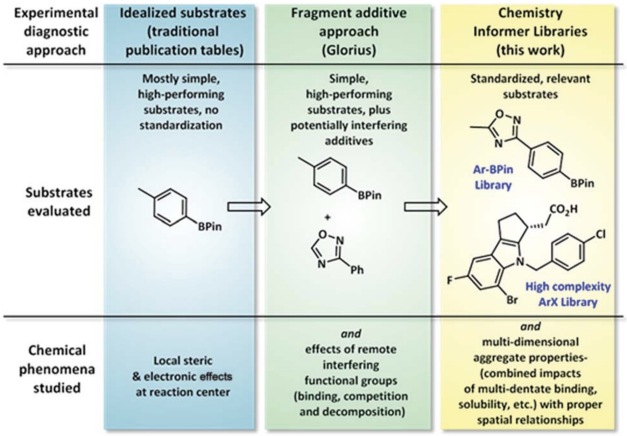


Figure 12: Chemistry informer approach to reaction screening to understand substrate compatibility, emphasizing the use of complex substrates to understand more complex chemical phenomena. Figure reproduced from Kutchukian et al. [244].

5.4 *Iterative* discovery of chemical processes

5.4.1 Discovery of optimal synthesis conditions

Automatic discovery of optimal synthesis conditions is a task where closed-loop experimentation is frequently applied. With a platform able to perform reactions under a wide range of operating conditions and automatically analyze and interpret the outcomes, one can use an optimization algorithm to guide a search within a pre-defined process parameter space. As with many other examples of automated discoveries, the search space is highly constrained by expert human operators. Standard numerical optimization routines are sufficient to explore the narrow search space of interest when an expert is able to define a narrow range of conditions that is likely to lead to promising results.

The earliest automated platforms for organic reactions used batch reactors and computer-controlled valves or pumps to automatically add reagents according to computer-selected experiments [92, 93]. Automated

control of continuous process variables (e.g., residence time, temperature, reactant ratios) is simplified when using flow platforms that eliminate the need to physically replace or clean batch vessels. Due to the ease of sampling a crude product stream with an inline valve, they are frequently used to screen arrays of different process conditions to map out an experimental space and the corresponding parameter-performance relationship [255–258]. Automated optimization of organic reactions in flow (Figure 13) has been extensively reviewed and is an excellent entry point for those interested in automated chemistry [9, 11, 259–265].

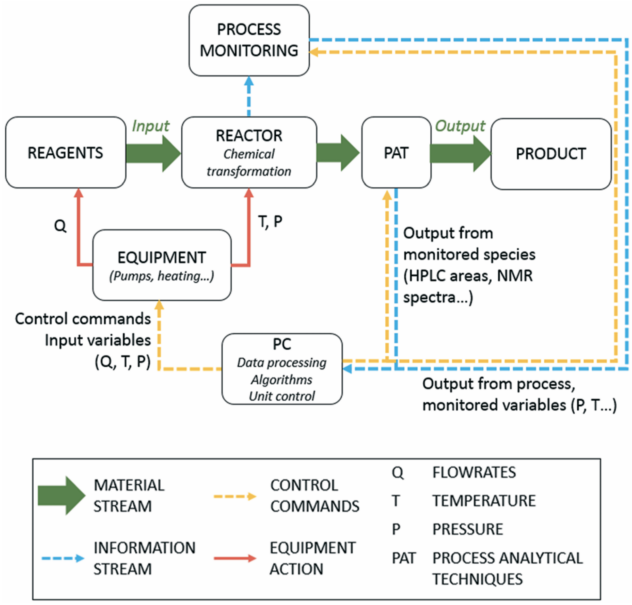


Figure 13: General platform schematic for the iterative optimization of synthetic processes in flow with respect to continuous variables (flowrates, temperature, pressure). Figure reproduced from Mateos et al. [264].

Optimization routines that have been widely employed include conjugate gradient [266], simplex [267], genetic algorithms (GAs) [268], Stable Noisy Optimization by Branch and Fit (SNOBFIT) [269], adaptive response surface methods [270], Bayesian optimization approaches [271], and reinforcement learning [272]. Optimizations over continuous variables generally use black box methods like genetic algorithms, Simplex, and SNOBFIT [32, 273–279], gradient-based methods like steepest descent and conjugate gradient [280], or explicit model-based methods like an adaptive response surface [33, 281, 282]. In a recent study, Bédard et al. describe a reconfigurable flow platform that uses the SNOBFIT algorithm to optimize several common organic transformations [283]. Discrete process variables can be varied through the use of selector valves or liquid handlers [239, 284] to also optimize, e.g., catalyst/ligand identities [33, 282, 285] and reaction solvent [286]. Despite their hype, "modern" machine learning approaches to reaction optimization [287, 288] have not demonstrated any clear advantages over previously-used statistical methods. One underexplored opportunity is to embed prior chemical knowledge into the model through pretraining; Zhou et al. do this

not with chemical knowledge, but with knowledge about the geometry/roughness of the expected regression surface to improve the hill-climbing efficiency of a reinforcement learning optimization routine [287].

When the performance of a chemical process is measured by multiple objectives, it is important to understand their associated tradeoffs [289]. Rather than combining them into a single scalar metric to optimize over [10, 290, 291], one can optimize for knowledge of the Pareto front–settings of process variables where one performance metric cannot be increased without decreasing another [292, 293]. Multi-step reactions are particularly challenging to optimize, because the effects of changing one parameter can propagate through downstream process steps. They are typically broken up into individual synthetic steps to improve the tractability of the problem [294, 295] or optimized approximately through screening, rather than true closed-loop feedback [296].

Similar closed-loop optimizations have been demonstrated for materials-focused applications. Different properties of interest necessitate different analytical endpoints, but the overall workflow is the same. Optimization goals have included the emission intensity of quantum dots [297], the conversion and particle size resulting from a copolymerization [290], the identification of crystallization conditions for polyoxometalates [298, 299], the production of Bose-Einstein condensates [300], and the realization of a metal-organic framework (MOF) with high surface area [301]. The MOF synthesis optimization by Moosavi et al. is particularly noteworthy in that prior data on syntheses of other MOFs were used to estimate the relative importance of synthetic parameters to enable a maximally diverse initial design of experiments, jump-starting the phase of iterative empirical optimization [301].

The challenge for these discoveries is often practical, not methodological. Experimental platforms must be able to analyze the relevant performance metrics and to control process variables across a search space that is broad enough to make computational assistance worthwhile. The ARES (autonomous research system) is an example of how complex instrumentation can enable optimizations of processes that are traditionally difficult to automate [302, 303]. ARES can perform up to 100 carbon nanotube growth experiments via chemical vapor deposition (CVD) per day under different temperatures, pressures, and gas compositions with real-time monitoring of growth rates using Raman spectroscopy. After fitting a model with 84 expert-defined experiments as prior knowledge, a genetic algorithm was used to achieve a user-defined target growth rate through automated control of process conditions.

Iterative discovery of quantitative models of process performance (e.g., experimentation to estimate kinetic parameters) differs from optimization only in how experiments are selected. Instead of selecting experiments with the ultimate goal of maximizing yield or achieving optimal product properties, experiments can be selected to minimize uncertainty in regressed parameters or discriminate between multiple hypothesized models [304]. The acquisition functions needed for these goals–to quantify how useful a proposed experiment

would be–can be directly imported from work in statistics on parameter estimation and model discrimination [305]. There are still challenges for multi-step reactions, as deconvoluting the effects of kinetic parameters from individual steps may not be straightforward even when the rate laws are known [306].

5.4.2 Discovery of new chemical reactions through an active search

There are far fewer examples of trying to discover of new chemical reactions through *active* searches than through noniterative screening strategies. Amara et al. describe one example of discovering new reaction pathways in a catalytic reactor system by reformulating the problem as a reaction optimization [307]. Using a modified Simplex algorithm, they were able to optimize the yield of then-uncharacterized side products; mechanistic pathways were proposed by experts based on evaluation of the conditions leading to different product distributions.

Granda et al. instead treat reaction discovery as a natural consequence of building a quantitative model of chemical reactivity [308] (Figure 14). Specifically, they describe a platform for evaluating the reactivity of two- and three-component reactions among a set of 18 hand-picked building block molecules (969 possible experiments) using two empirical models: one makes a boolean prediction of whether a reaction has taken place based on NMR, MS, and ATIR data before/after mixing; the second makes a boolean prediction of whether a given combination of substrates is likely to be reactive, using a one-hot representation for chemical species. The physical apparatus is reminiscent of MEDLEY–an automated reaction system employing computer-controlled pumps connected to a round-bottom flask [309]. However, the goal of reaction discovery and training this binary classifier are misaligned: the algorithmic exploration of the search space of possible reactions does not direct experiments toward those likely to lead to novel reactions; the reactions claimed to have been discovered were identified only through (error-prone [310]) manual analysis of product mixtures. Moreover, all 969 reaction combinations could have been performed in a miniaturized well-plate format while using less time and materials (cf. item iv of section 4.1) and the one-hot encoding of substrates precludes prediction of reactivity for unseen substrates. An earlier version of this platform was used by the same group to explore the 64 possible pathways defined by a three-step synthesis with one of four reagents added at each step [311]. The "most reactive" pathway was found through a step-wise greedy search by identifying the reagent whose addition led to the largest change in the ATIR spectrum of the product mixture. While this too did not explicitly bias experiments towards new reactions, the concept of selecting experiments in a non-brute force manner for reaction discovery is worth further investment.

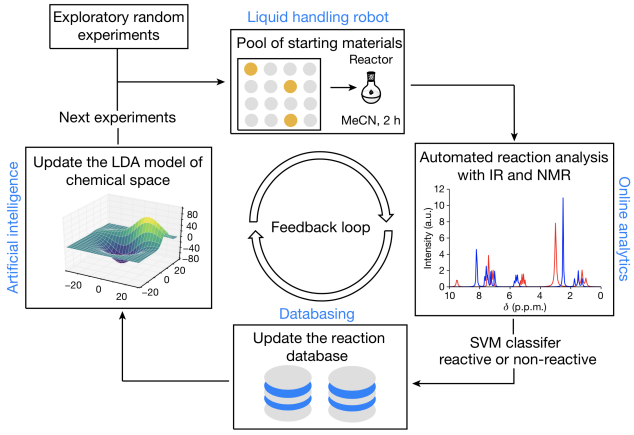


Figure 14: Workflow for iteratively training a binary classifier of whether a reaction mixture is reactive, using experimental validation and feedback. Figure reproduced from Granda et al. [308].

5.5 *Noniterative* discovery of structure-property models

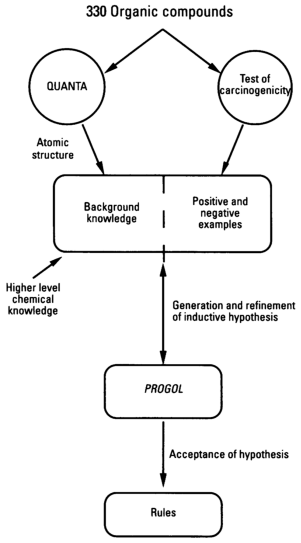
Models capable of relating the structural and compositional features of a molecule or material to its properties are of substantial utility in discovery. These relationships are often learned directly from data, whether via standard multivariate regression or machine learning algorithms. To the extent that they are interpretable, they can yield insight into how the fundamental features of a chemical entity or system influence its properties or performance, thus informing design. Quantitative structure-activity/property relationships (QSARs/QSPRs) can act as our belief about a performance landscape (cf. "belief" in Figure 2) for the sake of a specific discovery task like the discovery of new physical matter. While there is only a weak distinction between developing a QSAR/QSPR model for its own sake or for the purposes of exploring a design space, this section will focus on studies where the primary discovery is of the model itself. General considerations and trends in QSAR/QSPR are discussed in refs. 312, 77, and 313.

5.5.1 Discovery of important molecular features

Given a QSAR/QSPR model, one can investigate how the model perceives different structural attributes to reveal which are most informative of the prediction task. Substructure filters are commonly employed to process screening hits [314–316] and flag reactive or toxic functional groups [317–322]. This is a problem of interpretability that has received significant attention in the machine learning community [323]. When the form of the desired interpretation is restricted to molecular substructures, standard approaches of feature selection can be applied to representations based on the presence/absence of certain substructures [324, 325]. Polishchuk provides a recent review of interpretability for QSAR/QSPR models, including this category [326].

An early attempt to correlate predicted function directly with structural attributes was PROGOL [327],

an inductive logic programming algorithm. In its original demonstration, PROGOL identified a set of five criteria for determining whether a compound is likely to be mutagenic based on the presence of hypothesized toxicophores defined by connectivity and partial charge values; subsequent studies pursued similar explanations for carcinogenicity [328] and ACE inhibition activity [329], among others (Figure 15).



A compound is carcinogenic if it has a(n)			
(1)	positive Ames test	[99-40]	or
(2)	ester oxygen and ≤ 1 methyl group	[8-2]	or
(3)	ester oxygen and an aromatic hydrogen with a partial charge ≤ 0.041	[9-2]	or
(4)	ether oxygen with a partial charge ≥ –0.182	[5-1]	or
(5)	ether oxygen with a partial charge of –0.358	[2-0]	or
(6)	chlorine, an hydroxyl group and a benzyl ring.	[9-1]	or
(7)	bromine with a partial charge ≤ –0.086	[5-0]	or
(8)	aldehyde oxygen	[3-0]	or
(9)	aromatic amine group and an unsaturated carbon with a partial charge ≤ –0.181	[5-0]	or
(10)	unsaturated carbon with a partial charge of ≥ 0.4	[4-0]	or
(11)	unsaturated carbon and a 6-membered carbon ring	[6-1]	or
(12)	carbon atom in a 6-membered aromatic ring with a partial charge of 0.005	[3-0]	or
(13)	carbon atom in a 6-membered aromatic ring with a partial charge of 0.211	[4-1]	or
(14)	carbon atom in a 6-membered aromatic ring with a partial charge of –0.135	[3-1]	or
(15)	aliphatic carbon with a partial charge ≥ 0.507	[7-1]	or
(16)	aliphatic carbon with a partial charge of –0.085.	[2-0]	or
(17)	four halide atoms attached to tetrahedral carbons	[4-1]	or
(18)	carbon in a 5-membered aromatic ring with the same charge as a carbon in a 6-membered aromatic ring	[3-0]	

Figure 15: Workflow for the application of PROGOL [327] to the induction of structural alerts for carcinogenicity and the resulting rules. [x-y] indicates that the rule correctly applied x times and incorrectly applied y times in the dataset of 330 organic compounds. Figure reproduced from King and Srinivasan [328].

One approach to interpretability is to rely on few-parameter regressions with interpretable descriptors [330, 331], which can provide explanations as meaningful as the descriptors themselves. Decision trees provide a natural mode of assessing descriptor importance, though ensembling methods (e.g., Random Forest models) can obfuscate analysis [332, 333]. More general techniques exist to extract symbolic rules from trained machine learning models that are relatively agnostic to the type of model used [334–336]; ref. 184 provides an example of a decision tree extracted from an SVM model trained to predict the success of an inorganic synthesis procedure. There are numerous other examples of QSAR/QSPR studies that estimate descriptor importance in an attempt to rationalize predictions [143, 326, 337–340]. Other approaches instead aim to identify the training examples most relevant to a given test example [341, 342].

Visualizing explanations can be more intuitive than looking at quantitative feature importance metrics. One popular approach is to approximate a model by a fragment-contribution approach by looking at how the predicted property changes when part of the input molecule is masked [343–345]. If the value decreases when masking a certain substructure, that substructure is assumed to positively contribute to the property. This per-atom or per-substructure importance metric is usually an oversimplification of what is being learned, though sometimes it is exactly what is being learned [346]. The accuracy of machine learning models is usually

at least partially attributable to the nonlinearities between the input featurization and output property.

5.5.2 Discovery of models for spectral analysis

A natural application of data science techniques is to the analysis of spectral data for computer-aided structural elucidation (CASE). The underlying function that maps a molecule or material to the results of an assay is no different than a standard structure-property relationship, except the property might be high dimensional. CASE will become increasingly important for structure confirmation and quantitation as autonomous systems start to explore new areas of chemical and reactivity space.

The DENDRAL program is an early example of a program designed for structural elucidation of organic structures from mass spectrometry (MS) data [347, 348]. It crossreferences the mass loss between peaks with a list of known fragments to identify the likely substituents of the original molecule, enumerates possible molecular structures, predicts the MS spectra of those candidate structures, and makes its final proposal based on consistency with the observed spectrum. DENDRAL proved useful in its ability to perform many rapid calculations (spectral simulations and matching), but still required expert heuristics to explore the vast space of possible structures, including a "badlist" to prune unrealistic ones. Hufsky and Böcker provide a recent review of computational analysis of MS fragmentation patterns [349], which continues to be the subject of supervised learning approaches [350] and has seen renewed interest in the context of metabolomics [351–354].

Unsurprisingly, other types of analytical data are also commonly evaluated using computational or machine learning models, including UV circular dichroism to elucidate protein secondary structure [355]. The reverse problem of spectral prediction is also popular and includes techniques to predict NMR shifts [356, 357], IR spectra [358], and protein fluoresence [359]. Materials-focused studies have looked at predicting the optical properties of metal oxides [360] and analyzing microstructure from SEM data [361], among others. Tables 1 and 2 of ref. 362 summarize many early examples of applying neural networks to MS, NMR, IR, NIR, UV, and fluoresence spectra leading up to the mid-1990s. A more recent overview of learning structure-spectrum relationships and CASE can be found in ref. 363.

5.5.3 Discovery of potential energy surfaces and functionals

There is tremendous interest in using machine learning techniques to build surrogate models for computationally-expensive *ab initio* calculations. The accurate prediction of electronic properties is directly useful for the discovery of organic electronic materials and is a central focus of the Harvard Clean Energy Project [364]. Models can replace either the entire energy calculation [365–368] or specific parameterized components (e.g., functionals or correlation energies) [369–372]. A prominent example from Roitberg, Isayev, and coworkers

is ANAKIN-ME or ANI (accurate neural network engine for molecular energies); ANI is a neural network surrogate model of an energy potential trained on roughly 60,000 DFT calculations [373] and its second generation, ANI-1ccx, is further refined on CCSD(T)/CBS calculations [374] (Figure 16). Active learning strategies can be used to strategically acquire costly training data when training such models [374–376].

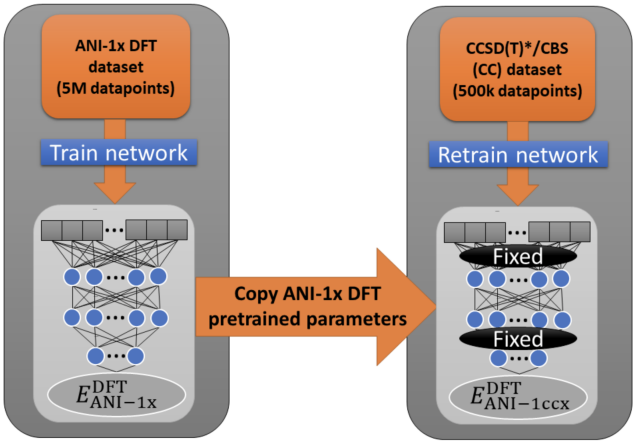


Figure 16: Workflow for refining a surrogate model of electronic structure calculations, originally trained on ωB97x/6-31G* data, on higher-quality CCSD(T)*/CBS data. Figure reproduced from S Smith et al. [374].

The desire to create computationally-inexpensive surrogate models has underpinned the development of classical force field models for molecular dynamics (MD) simulations [377–379]. Perhaps unsurprisingly, machine learning models can serve as drop-in replacements for heuristic force fields if so trained [380–382] and can assist in coarse-graining for larger scale simulations [383, 384] or as a post-processing step to analyze simulation results [385, 386]. Structure-aided drug design relies on similar parametric functions for predicting protein-ligand binding. There are many molecular docking programs that propose and score different poses describing ligand interactions with protein targets [387–389]. Scoring functions–meant to provide quantitative measures that correlate with binding affinity–are ideal applications of machine learning techniques. Nonlinear statistical models can help bridge the divide between our pseudo-first principles models of the underlying chemical interactions and the actual behavior we observe experimentally [389–396].

5.5.4 Discovery of models for phase behavior

QSAR/QSPR models that describe a molecule or material's phase behavior can aid computational design by predicting whether a proposed compound is physically realizable in its desired form. For hypothesized metallic alloys, for example, one can predict crystal structures [397–400] and phase behavior [401–406]. For organic molecules, one can similarly predict whether compounds are likely to crystallize easily [407] and their preferred processing-dependent polymorph [408]. Machine learning models can also reduce the number of evaluations required, e.g., for finding minimum energy configurations [399].

10.1002/anie.201909987

5.6 *Noniterative* discovery of new physical matter

The noniterative discovery of new physical matter is a common application of computational learning techniques or automated experimental platforms. This category encompasses experimentation strategies in which search spaces are predefined and explored exhaustively and virtual screening with or without the use of a surrogate model to approximate a structure-function landscape (right half of Figure 17).

A quintessential paradigm in this category is the use of a large dataset from experiments or simulations to train a QSAR/QSPR model, often using some form of machine learning for nonlinear regression, which is then used to screen a large number of candidate compounds or materials. A handful of candidates may be selected for synthesis and validation of the prediction, but the results of that validation are not used to revise the model. This approach essentially constructs a fixed "map" with which to explore the search space and identify promising candidates. It leaves little room for serendipity, as compounds that are not predicted to be useful–even if accounting for uncertainty–are generally not tested, unless the algorithm is explicitly biased toward random exploration.

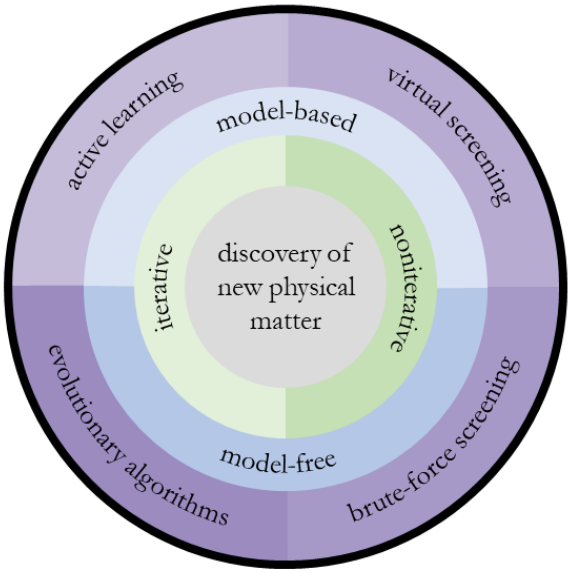


Figure 17: Taxonomy of strategies for the discovery of new physical matter.

5.6.1 Discovery through brute-force experimentation

In several studies, the part of the discovery process that is automated is not the hypothesis (model building or selection of compounds to test) but the experiment (initial data generation or validation). Experimental automation addresses the practical challenge of validation but not the methodological challenge of how to guide the scientific process or constrain the search space. In general, brute-force experimentation is a productive discovery strategy only when the experimentation is high-throughput in nature. High-throughput experimentation platforms are capable of searching broad design spaces, which makes serendipitous discoveries more likely and places less emphasis on the experiment-selection faculties of the researchers. Note, however, that *manual* constraint of the design space remains a critical aspect of the process.

Among the more interesting developments in HTE for drug discovery are entirely novel methodologies uniquely suited to rapid data generation. Despite the advances in achieving greater throughput with traditional HTE efforts [409, 410], the space that can be feasibly screened using single-compound-per-well synthesis approaches is often too small to provide many promising bioactive leads. DNA-encoded libraries (DELs) [99] enable synthesis of compounds for screening at rates of hundreds of compounds per well [411, 412] using a split-and-pool synthesis strategy [413]. Many modern DEL case studies report theoretical library complexities of hundreds of millions [414–416] or even billions [417–419] of compounds, exceeding the size of the search space by several orders of magnitude over traditional HTE approaches [420]. In light of these impressive synthesis rates, it should be noted that analysis and (if necessary) purification can be rate-limiting. DNA-encoded chemistry is reviewed in ref. 420.

Another strategy that has proved useful in this area is diversity-oriented synthesis (DOS) which aims to generate structurally (and thereby functionally) diverse collections of small molecules [100]. These strategies may involve reacting a starting material with diverse arrays of reagents in series, or coupling an array of starting materials to one another across strategic functional groups [421]; multicomponent reactions are particularly useful for this application [422]. DOS strategies and their application to the discovery of lead drug compounds and biological probes have been reviewed extensively [100, 421–424].

Other novel approaches to making HTE for drug discovery more efficient focus not on synthesizing larger and/or more diverse libraries, but on screening them more efficiently [425–429]. Development of information-rich, efficient assays is a complex challenge. If the disease target has already been identified and it is possible to isolate, stabilize, and accurately dispense the target, then *in vitro* biochemical assays, which may involve the assessment of target-ligand binding affinity or augmentation of enzymatic activity, are useful [430–432]. These target-based assaysassays can be easily miniaturized and serve as the workhorse of many screening campaigns. They tend to be efficient, but assess activity against a single target in isolation and ignore the

complexities of human physiology and polypharmacology. Cell-based assays can do a better job capturing activity (because, for example, relevant cofactors are present) while also providing a measure of toxicity and other off-target effects. Despite the added complexity of automatically maintaining and dispensing cell populations, cell-based assays have been adequately automated and miniaturized for compatibility with highe-density plates [433]. A variety of easily-automated well measurement tools have been developed for compatibility with cell-based assays, including fluorescent detection; the automation and miniaturization of this type of assay for compatibility with HTS has been well-reviewed by An and Tolliday [434]. Recently, there has been an increased interest in phenotypic screening [435] and computational tools for the high-throughput analysis of its results [436, 437].

Many high-throughput synthesis and analysis tools have been developed to facilitate experimentation and discovery in materials science. Combinatorial synthesis methods that yield a single sample containing continuous composition gradients [438] are one notable example. In the seminal demonstration of this technique, Xiang et al. describe a parallelized synthesis method for superconducting copper oxide thin films that varies composition, stoichiometry, and deposition sequence to identify promising compositions [439]. As Senkov et al. point out, this type of experimentation can present unique obstacles to miniaturization in subdomains such as metal alloy design, where experiments of a certain scale are required to observe emergent macro- and mesoscale properties [440]. Since this early demonstration, combinatorial and other HTE methods have been developed in a wide array of materials subfields: to screen solid-state catalyst libraries [441, 442] as well as to discover cobalt-based MOFs [443], photosensitizers for catalyzing photoinduced water reduction [444], mixed metal oxide catalysts [445], adhesive coatings for automotive applications [446], polymers for gene delivery [447], ternary alloys that have high glass forming ability [448], and others.

Development of high-throughput analytics is critical to avoid bottlenecks; one method highly amenable to parallelization is infrared imaging [441, 449, 450], which has been used to screen arrays of heterogeneous catalysts [451]. Potyrailo and Takeuchi emphasize the diverse properties materials exhibit and the need for a correspondingly diverse set of characterization tools [452]. As an alternative to developing high-throughput characterization techniques, Sun et al. recently reported the rapid exploration of a 96-member library of perovskite-inspired compositions (Figure 18) in which they accelerated screening in part by replacing rate-limiting analytics with a cheaper analysis and a machine learning model [453]. Their study included the first reporting of four lead-free compositions in thin-film form. Reviews pertinent to high-throughput experimental screening for materials discovery are provided by Meier and Schubert [454], Zhao [455], Rajan [456], Hook et al. [457], Potyrailo et al. [458], and Green et al. [438].

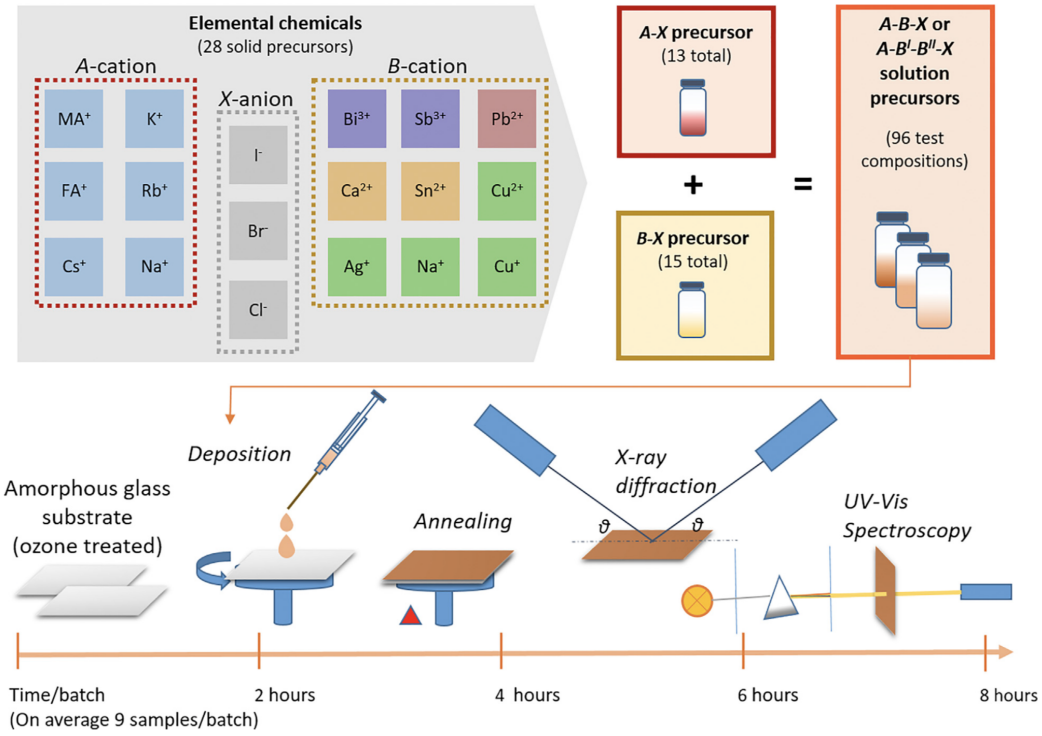


Figure 18: Workflow for the (relatively) rapid screening of a combinatorial space of perovskite-like compositions. Figure reproduced from Sun et al. [453].

5.6.2 Discovery through computational screening

There are a number of noniterative approaches to materials discovery [459] that rely on computational workflows like virtual screening or high-throughput simulation techniques [460–462]; Ong et al. developed a Python library (pymatgen) specifically designed to facilitate these workflows [463]. Curtarolo et al., Pyzer-Knapp et al., and Himanen et al. provide recent reviews on high throughput virtual screening for materials discovery, specifically describing the process of generating large databases and identifying meaningful trends within them [61, 461, 464].

The predominant use of machine learning in computational materials discovery has been to fit surrogate models to existing (often, experimental) data and screen a large design space [465–471]. To the extent that performance can be correlated to structure, these models can reveal opportunities for the design of new catalysts/ligands for organic synthesis [223, 225–227, 472] (Figure 11), metallic catalysts [473, 474], Heusler compounds [475], metal organic frameworks (MOFs) [476], hybrid organic-inorganic perovskites [477], superhard materials [478], thermal materials [479], organic electronic materials [480–484], polymers for electronic applications [485, 486], porous crystalline materials for gas storage [487, 488], and reductive additives for battery electrolyte formulations [42]. Computational models have also been used to determine

when calculations are likely to fail [489] and to identify associations between materials and specific property keywords through text mining [490].

A few trends are apparent in the experimental validation of these frameworks. First, the confidence of computational predictions is intimately coupled to the quality and applicability of the model. Second, experimental validation is often preceded by extensive manual filtering of the computationally-prioritized compounds to take into account factors such as synthesizability, laboratory capabilities, and (human-)perceived suitability for the discovery objective [482, 491, 492]. As with organic molecules, there can be a misalignment between the compounds one would like to test (that are computationally predicted to achieve a desired function) and what can be realized (synthesized) experimentally. In one example, as a conservative filter for synthesizability, Sumita et al. require that proposed molecules have at least one known synthetic route reported in SciFinder [493]. Third, there are often discrepancies between the predictions made by a surrogate model and the values determined experimentally, and in some cases the discrepancies are large enough to have a substantial bearing on the desired performance [491]. These latter two trends imply that the pertinent features of promising materials are often not fully captured by the algorithms developed to date. Thus, experimental validation is acutely relevant in this area.

Computational screening, with or without experimental validation, is also a common strategy for identifying promising therapeutic candidates. Many reviews on the use of virtual screening in drug discovery exist, including Schneider [39], Sliwoski et al. [494], Macalino et al. [495], Lavecchia [496], Wingert and Camacho [497], Zhang et al. [82], and Panteleev et al. [498]; these emphasize the use of machine learning methods to generate the surrogate QSAR/QSPR models that guide the VS process. Walters recently reviewed strategies for library enumeration in the drug discovery space [31]. These range from applying known reaction transformations to available molecules in order to define make-on-demand libraries [146, 499–504] to generative strategies, which are discussed later in the context of *de novo* design of singular lead compounds. Make-on-demand libraries have the advantage that candidates selected for follow-up experimental validation should be readily synthesizable, with some exceptions (10-20% of compounds, anecdotally) due to imperfect enumeration rules.

For drug discovery applications, virtual screening is often divided into two categories: structure-based [505–507] and ligand-based [508, 509]. Structure-based VS relies on scoring functions that relate information about a molecule and the target protein to binding affinity between the two. Docking analysis is a common paradigm in structure-based VS. Many software packages for this purpose exist, including AutoDock [510], FlexX [511], GOLD [512], and Glide [513], and have been extensively reviewed and compared [387, 514]. Ligand-based strategies, in contrast, make no direct consideration of the structure of the target protein. Instead, they rely on QSAR models and/or direct similarity assessments [515] that compare library compounds

to a reference compound that exhibits desired properties. Many algorithms exist for making the similarity comparisons [516, 517].

Studies that validate virtual screening strategies by synthesizing and testing the compounds identified by their workflow include [504, 518, 519]; we note that such validation is routine and expected in industrial drug discovery campaigns. In one example, Hoffer et al. combine virtual screening with partially-automated synthesis and testing in a workflow for hit-to-lead optimization that they call diversity-oriented target-focused synthesis, or DOTS [520, 521] (Figure 19). The DOTS framework begins with a hit fragment around which a virtual library is enumerated through the *in silico* application of common synthetic reactions that combine the hit with commercially-available building blocks. Next, the authors apply their docking program, S4MPLE [522], which uses an evolutionary algorithm for conformational sampling, to select the compounds with favorable target interactions. Finally, the high-priority set is synthesized using a Chemspeed robotic synthesis platform to carry out expert-defined syntheses and subjected to *in vitro* evaluation. In one application, all seventeen of the high-priority compounds had higher pIC50 values than the initial hit.

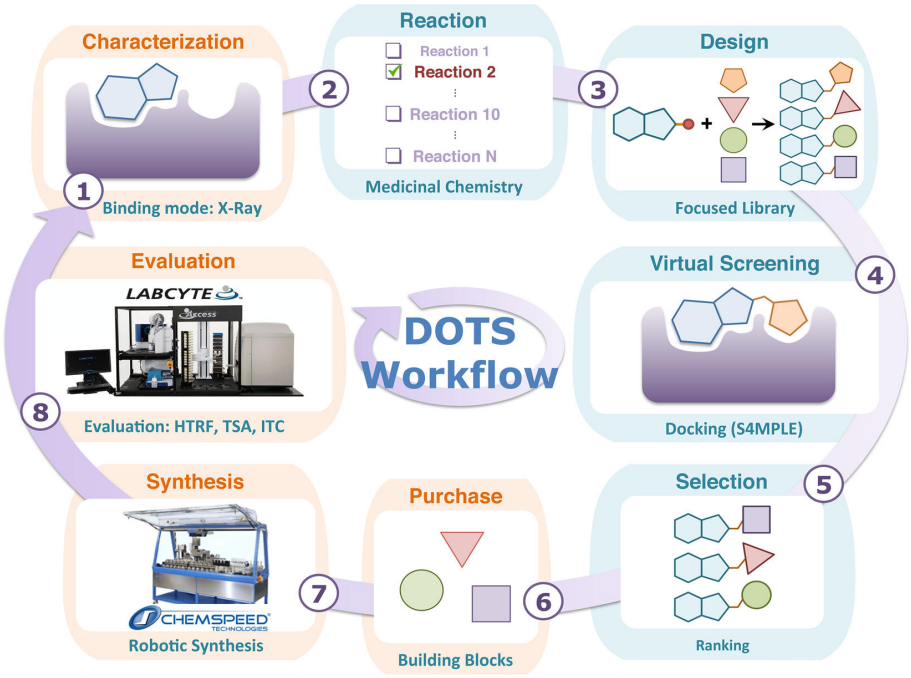


Figure 19: Workflow for diversity-oriented target-focused synthesis (DOTS). Experimental steps are in orange; computational steps are in blue. Figure reproduced from Hoffer et al. [520].

5.6.3 Discovery through molecular generation

All of the case studies above examine search spaces defined by discrete sets of candidates. These candidate sets consist either of the compounds already existing in a database or library of interest, or they are somehow

10.1002/anie.201909987

systematically enumerated. While some of these libraries are quite large, for example, the 170 million make-on-demand compounds from ref. 504 or the 11 billion in the REAL database [523], their discrete nature constrains the search space. Computational techniques such as deep generative models in which molecules are generated, manipulated, and/or optimized in a continuous latent space (Figure 20) have emerged and represent a means of overcoming the finiteness of discrete candidate sets (and, more specifically, as an alternative to earlier design approaches based, e.g., on genetic algorithms [524]). These models are predicated on the assumption that the generated compounds, by virtue of being drawn from the same prior distribution as the training molecules, will inherit the training molecules' important properties such as stability and synthesizability while being biased toward a specific property of interest (e.g., bioactivity) [89, 91]. Experimental validation is uniquely relevant for these techniques since they are not based on first-principles calculations, interpretable QSARs, or well-vetted heuristics, but rather neural models that create an obfuscated approximation of the distribution within chemical space and an underlying structure-function landscape.

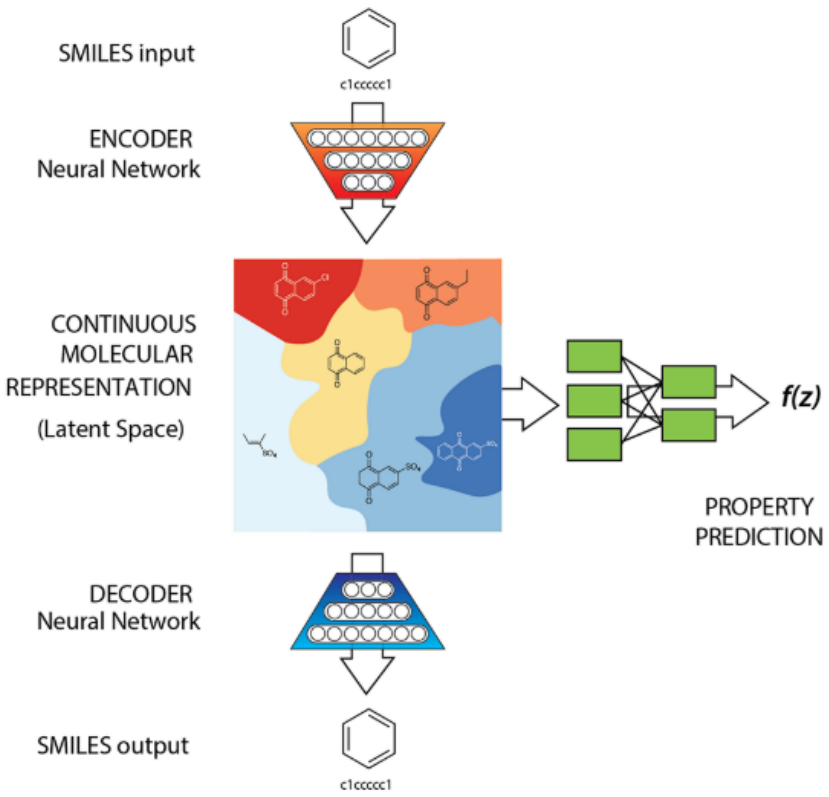


Figure 20: Diagram of an autoencoder for molecular discovery. Figure reproduced from Gómez-Bombarelli et al. [525].

In an early example of the adaptation of deep generative networks to the pharmaceutical space, Kadurin et al. describe the development of an adversarial autoencoder (AAE), which wraps an autoencoder in the generative adversarial network (GAN) training framework [87] to identify antitumor agents based on existing MCF-7 cell line assay data [526] (see ref. 527 for the original implementation and ref. 528 for the improved technique, DruGAN). In another early example, Gómez-Bombarelli et al. applied a VAE operating on SMILES strings (following the decoding approach of Bowman et al. [529]) for the latent-space optimization of molecules with respect to druglikeness and synthetic accessibility heuristics, demonstrating superior performance to random search and a genetic algorithm when initialized on low-performing molecules [525].

Generative models with RNN encoding-decodings have emerged as one of the major paradigms in *de novo* drug design [530, 531]. For example, Yuan et al. use a character-level RNN [533] to generate virtual libraries of SMILES strings [532]. By training their model on 25,000 known VEGFR-2 inhibitors, they were able to generate a library enriched with high-affinity ligands relative to target-agnostic screening libraries, as judged through a computational docking program. Five of the highest-affinity ligands were selected for synthesis and testing and two were found to be more potent than vatalanib, a known inhibitor. Bjerrum and Threlfall adopted a similar approach using the ZINC12 database to train their model [534]. Their emphasis on evaluating the synthetic accessibility of the molecules that their model designed reflects the extent to which generative models have failed in this area historically. Combining this strategy with reinforcement learning to generate molecules that are similar to a seed compound [535], molecules that have high predicted bioactivity against a particular target [531, 535, 536], molecules that otherwise have desirable druglike properties (such as chemical beauty and Lipinski) [537], and molecules that represent an internally diverse set [530] have proven fruitful, as have applications to peptide design [538, 539]. Transfer learning in the form of model pretraining has also been useful to successfully overcome the disadvantages inherent in low-data domains, for example to design modulators of therapeutically-relevant nuclear receptors [540, 541].

As Jin et al. point out, a failing of the SMILES string representation in the molecular generation context is that a single molecule can usually be mapped to several distinct, valid SMILES strings, which complicates the creation of a latent space that varies smoothly from one molecule to another, similar one [542]. They contribute an alternative approach, the *junction tree variational autoencoder*, that generates molecular graphs rather than SMILES strings, demonstrating the ability to generate both a library of valid molecules as well as optimize in the latent space.

Several of the generative model case studies cited herein include experimental validation [530, 532, 538, 539, 541, 543, 544], although the validation was not automated and the sets of generated molecules often required extensive filtering before selection for synthesis. Schneider and Clark review fragment-based *de novo* drug discovery efforts that specifically include experimental validation [41]. They also highlight the

fact that *de novo* efforts are often plagued by synthesizability issues and advocate for the incorporation of CASP software into the workflow to help address this. In lieu of experimentation, some studies validate the capabilities of generative models by comparing distributions of properties of the generated molecules to those of the training set [545, 546]. See [89–91] for detailed reviews of the methods for and applications of generative models in chemistry and molecular design.

5.7 *Iterative* discovery of new physical matter

One rarely has a perfect understanding of a structure-function landscape, particularly in discovery applications where data can be limited. In this section, we focus on case studies in which computer-assistance is applied to *at least* the experimental selection aspect of an iterative discovery workflow (left half of Figure 17). In general, iterative discovery of new physical matter centers around a structure-function model that is used to reason about which experiments to perform. The results of the experiments are then used to devise the subsequent round and update the structure-function model such that more accurate predictions can be made. Iterative strategies like active learning and Bayesian optimization can be used to augment the set of available data efficiently, focusing on informative and/or promising experiments within the search space [83, 84]. Model-free iterative strategies include evolutionary algorithms, which operate by mutating candidates directly based on validation data from an experiment or simulation. Strategies where validation and feedback inherently drives experimental selection (e.g., directed evolution and continuous evolution techniques [52, 411]) are out of the scope of this review, but do represent an important class of autonomous experimental platforms.

5.7.1 Discovery for pharmaceutical applications

Given the complexity of the assays involved in characterizing new drug compounds, active learning is especially beneficial for reducing the experimental burden associated with drug discovery and accelerating the search [547, 548]. Pool-based active learning (Figure 21) refers to the selection of candidates from a discrete, pre-enumerated set of options [83]. Active learning has been deployed in this setting both for information–to build accurate models of the activity of potential drug compounds against specified targets [549, 550]–as well as for performance–to identify active compounds in as few experiments as possible [549, 551]. These aims are not mutually exclusive; in order to identify high-performing candidates, it is often the case that these algorithms must be designed to select experiments where activity is expected to be highest and experiments that support overall model accuracy [552–555].

A particularly noteworthy example of the pool-based active learning strategy is the platform Eve, which

10.1002/anie.201909987

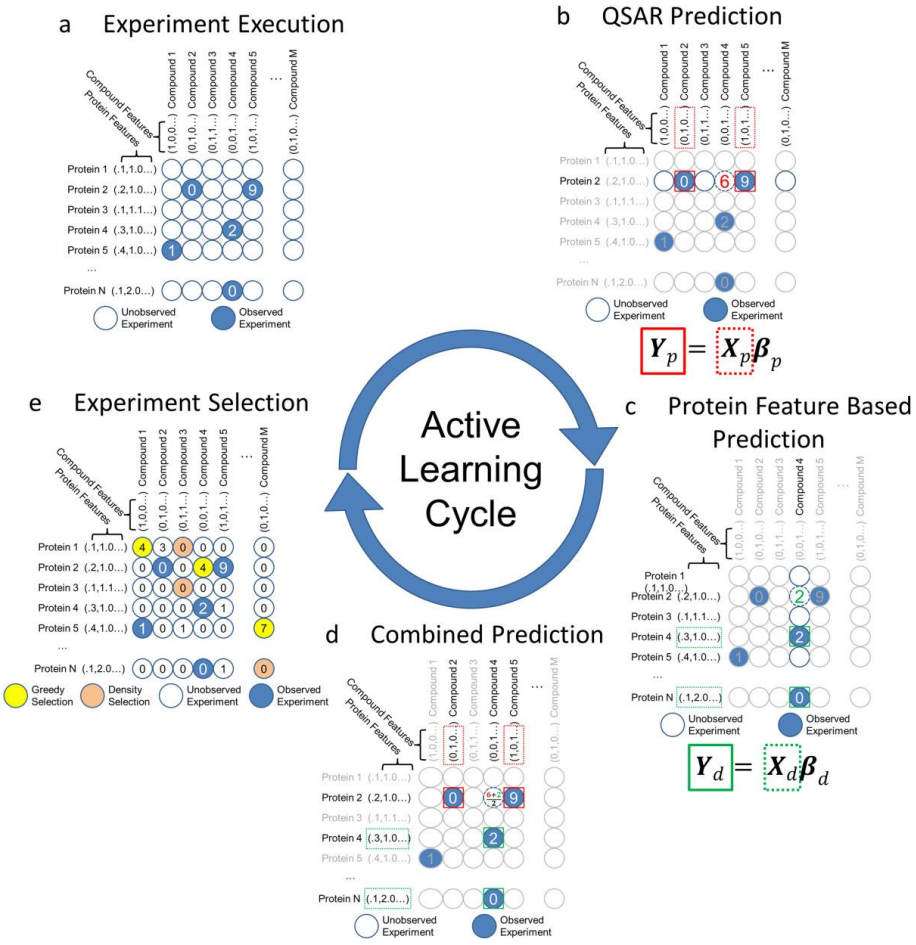


Figure 21: Workflow for pool-based active learning to identify compounds that bind strongly to proteins within an N compound $\times$ M protein space of interactions. Figure reproduced from Kangas et al. [552].

was designed to conduct closed-loop hit identification [556, 557]. Experimentally, Eve has the capacity to rapidly screen predefined compound libraries against a variety of biological assays at a rate of >10,000 compounds per day. The platform uses collected data to create a surrogate model of the structure-activity landscape and then selects subsequent rounds of compounds with high predicted activity, selectivity, and/or prediction variance, rather than exhaustively exploring its search space. The authors created an econometric model of the drug discovery process that accounts for (a) the utility of a hit, (b) the utility of the reduction in experimental space that needs to be screened, and (c) the cost of missed hits, among other factors, and found that it is typically more economical to use active learning than brute-force screening.

Increasing the size of the search space increases the likelihood there is a high-performing global optimum (although it may be difficult to identify). In experimental settings, the ability to synthesize compounds on-demand expands the search space beyond the set of in-stock compounds. Desai et al. describe a microfluidic platform able to produce 27×10 compounds on-demand via a one-step Sonogashira coupling [554]. The plat-

form integrates synthesis with online purification, dilution, and activity assay against Abl1 and Abl2 kinases. A random forest model was created to approximate the structure-activity landscape and guide experiment selection. Experiments were chosen via one of two approaches–one greedy approach to maximize expected activity and one to explore undersampled chemical space–and results were used to update the surrogate model [558]. Despite the expansion in the search space achieved by incorporating synthesis capabilities, the design space explored in this example remains extremely narrow; a brute-force search would have been tractable and potentially faster if parallelized. Still, this study is an excellent proof-of-concept for closed-loop synthesis, purification, and testing. More flexible synthesis-purification and synthesis-purification-testing platforms have been developed but only applied to open-loop discovery with manual compound design and synthesis planning [12, 425, 559, 560] (Figure 22). Table 2 in ref. 561 reviews some additional examples of integrated synthesis and testing.

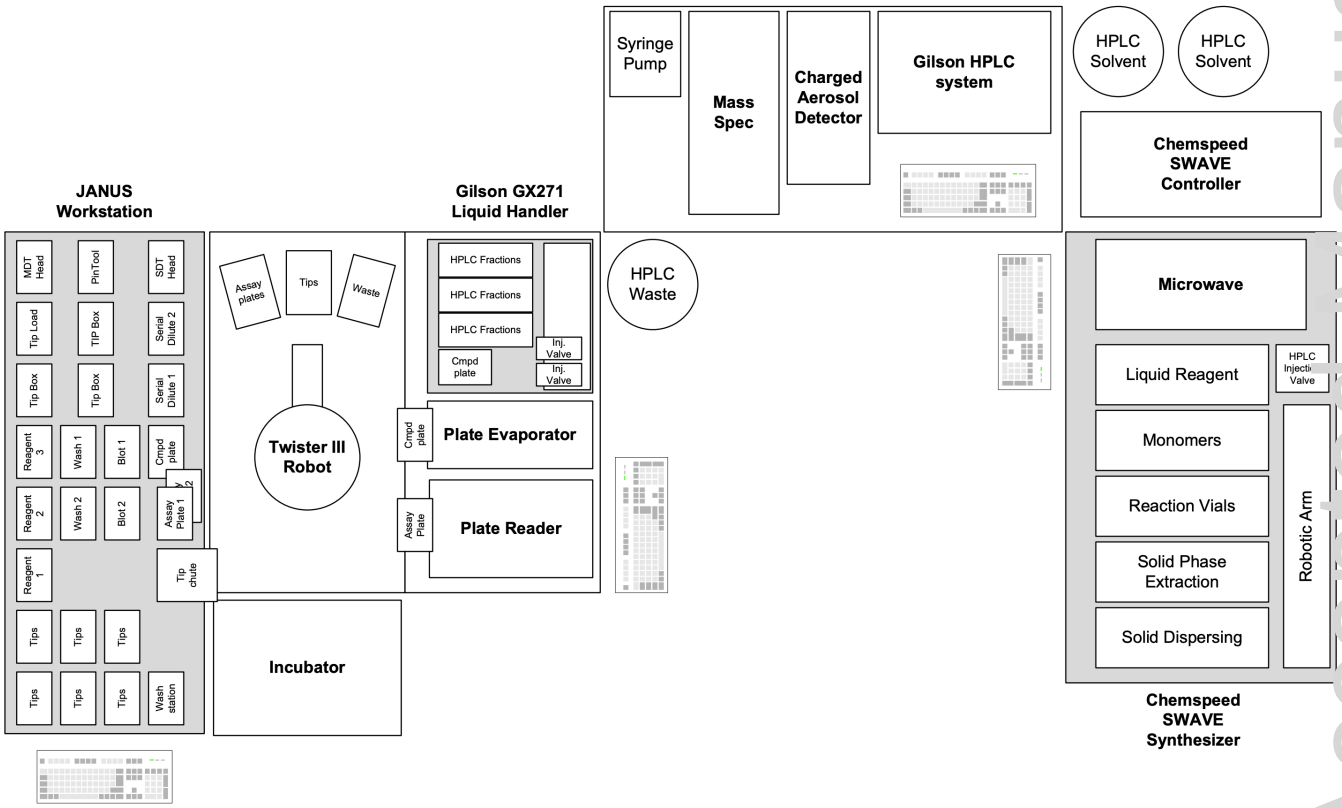


Figure 22: Integrated platform for open-loop synthesis, purification, and testing from AbbVie. Figure reproduced from Baranczak et al. [560].

Evolutionary strategies are another means of expanding the search space beyond a set of pre-enumerated candidates. Besnard et al. employ one such technique in which candidate compounds evolve *in silico* as part of the iterative process [562]. Specifically, on each iteration, new candidates are evolved from the highest performers from the previous generation through the application of transformations from the medicinal

chemistry literature. Here, the discovery workflow relies on the development of accurate QSAR models trained on ChEMBL data to ensure alignment between the *in silico* performance and experimental activity. Firth et al. use a related technique that evolves molecules using a fragment replacement routine (RATS, rapid alignment of topological scaffolds), which enables a less constrained exploration of chemical space than does reaction enumeration; this approach was demonstrated on a surrogate multi-objective function (including a model of CDK2 activity) starting from a known active scaffold, with manual experimental validation of a small library of recommendations [563].

Genetic algorithms (GAs) are related to the approaches used by Besnard et al. and Firth et al. Candidate compounds are proposed as mutations of a parent compound whose performance is known; allowed mutations serve as a constraint on the search space and the optimization trajectory [564] (Figure 23). In contrast to active learning, GAs tend to use *static* fitness functions for compound scoring (although a few rely on experimental outcomes instead [565]). In a very early example of iterative molecular optimization using a genetic algorithm, Weber et al. describe the identification of inhibitors of the serine protease thrombin; 16 generations led to the identification of sub-micromolar inhibitors [251]. The key to this approach is that the $10 \times 40 \times 10 \times 40$ design space was defined by discrete substrate choices in a 4-component Ugi-type reaction to ensure straightforward (albeit manual) synthesis and testing. Other iterative strategies for drug discovery include *in silico* application of synthetic transformations to generate molecules that are scored based on their similarity to a target molecule and are, in principle, synthetically-accessible [567].

GAs have been successfully used to conduct both single- and multi-objective optimizations that account for factors including target protein binding affinity [569, 570], cost and bioavailability [571], and similarity to a chosen compound [568, 571–573]. Many strategies are fragment-based, operating in the manner described above, although some also allow atom-level mutations [574]. Strategies that operate directly on molecular graphs have also been proposed [575], and in the case of peptide design, one can operate directly on sequences [576]. Despite the increased interest in deep learning techniques, genetic algorithms remain a powerful strategy for exploring chemical space [577]. A number of reviews describe drug discovery applications of GAs and evolutionary algorithms more generally [524, 578–581].

5.7.2 Discovery for materials applications

Iterative experimental design strategies are increasingly being adopted to guide discovery in the materials space. For example, Xue et al. use the Bayesian optimization framework EGO [583] to select experiments from a discretized composition space that ultimately led them to NiTi-based shape memory alloys delivering low thermal hysteresis [582]. Similar approaches have been used for the same [584] and other applications to discover $BaTiO_3$-based piezoelectric materials with large electrostrains [585] and to optimize melting

temperature [586]. Compared to other domains, materials discovery tends to be relatively conducive to fully computational approaches since the properties we can calculate are more directly relevant to the functions we wish to optimize. As a result, several studies have used Bayesian optimization to select compounds for evaluation with calculations or simulations, rather than experiments, e.g., to optimize thermal conductivities [587] and elastic moduli [588]. Additionally, some instances of generative models–described in the noniterative discovery section above–incorporate Bayesian optimization to optimize compounds for desirable performance [525].

Related active learning strategies have been deployed in materials development as well. For example, Tran and Ulissi use DFT to validate candidates proposed through active learning from a fixed pool of intermetallic compounds from the Materials Project, with the goal of identifying high-performance electrocatalysts for CO_2 reduction and H_2 evolution [589]. Gubaev et al. use active learning to efficiently create a DFT surrogate for predicting the convex hulls of metallic alloy systems, ultimately discovering previously unknown stable alloys [590]. Even iterative greedy searches have proven effective in prioritizing simulations of a fixed library of candidate materials for hydrogen storage [591].

Several recent studies combine GAs, surrogate models for electronic structure calculations, and active learning for the discovery of spin crossover complexes and transition metal catalysts [592, 593] (Figure 24). These studies represent fully autonomous discovery within the space of organometallic complexes the genetic algorithm is able to explore. Among the many additional applications of GAs to materials discovery have been polymer design [564], the identification of stable four-component alloys [594], promising polymers for OPVs [595], MOFs for carbon capture [596], and polymer dielectrics with user-defined dielectric constants and bandgaps [597]. An excellent review of applications of active learning and Bayesian optimization to materials development is provided in ref. 598.

The utility of computational validation for materials discovery is somewhat offset by the complexity of experimental validation. The synthesis and analysis of materials and devices can be difficult to automate. A recent study by MacLeod et al. (Figure 25) serves as an excellent example of automating more than simple solution-phase chemistry [599]. Precursor solutions are spin cast into thin films, which are thermally annealed and analyzed for their optical and electronic properties. The platform, Ada, uses ChemOS [95] for hardware orchestration and Pheonics [291] for Bayesian optimization to explore a two dimensional continuous search space of composition and annealing temperature; its objective is to optimize a pseudomobility metric that correlates with hole mobility. While what Ada measures is still a proxy for the true performance of a multilayer device, the ability to miniaturize and automate fabrication processes like thin film casting expands the scope of problems able to be tackled by autonomous platforms.

10.1002/anie.201909987

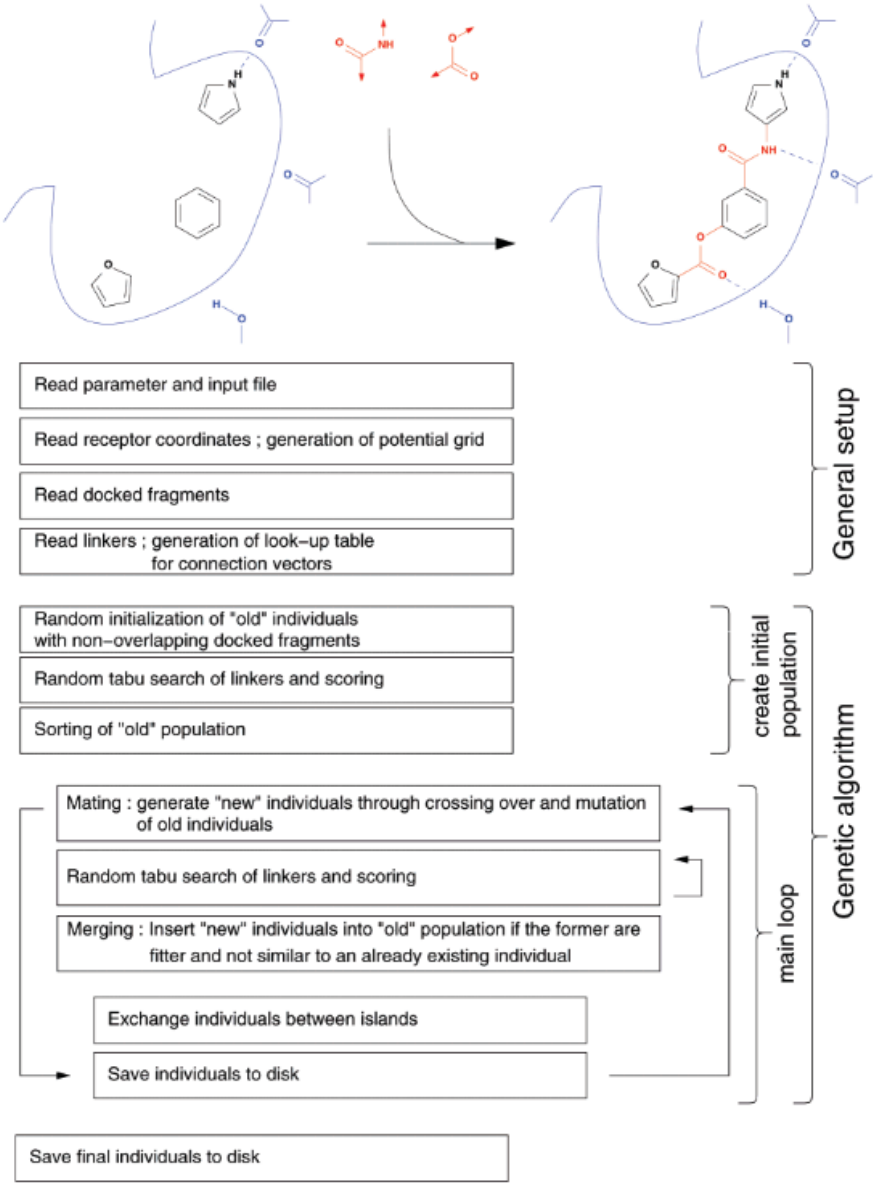


Figure 23: (Top) Schematic illustration of one approach to genetic algorithm-based drug design: predocked fragments (black) are linked to fragments (red) from a user-supplied list, with the target protein and its polar groups indicated in blue. (Bottom) The flow chart of the algorithm. Figure reproduced from Dey and Caflisch [568].

10.1002/anie.201909987

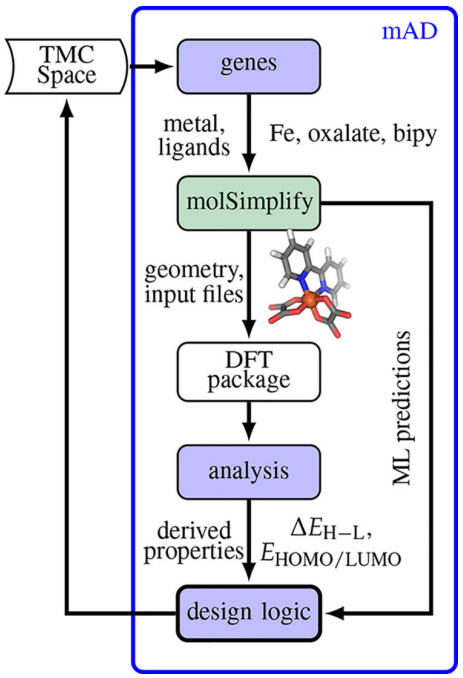


Figure 24: Workflow for the iterative design of transition metal complexes (TMCs) using a genetic algorithm and automated DFT calculations. Figure reproduced from Nandy et al. [593].

10.1002/anie.201909987

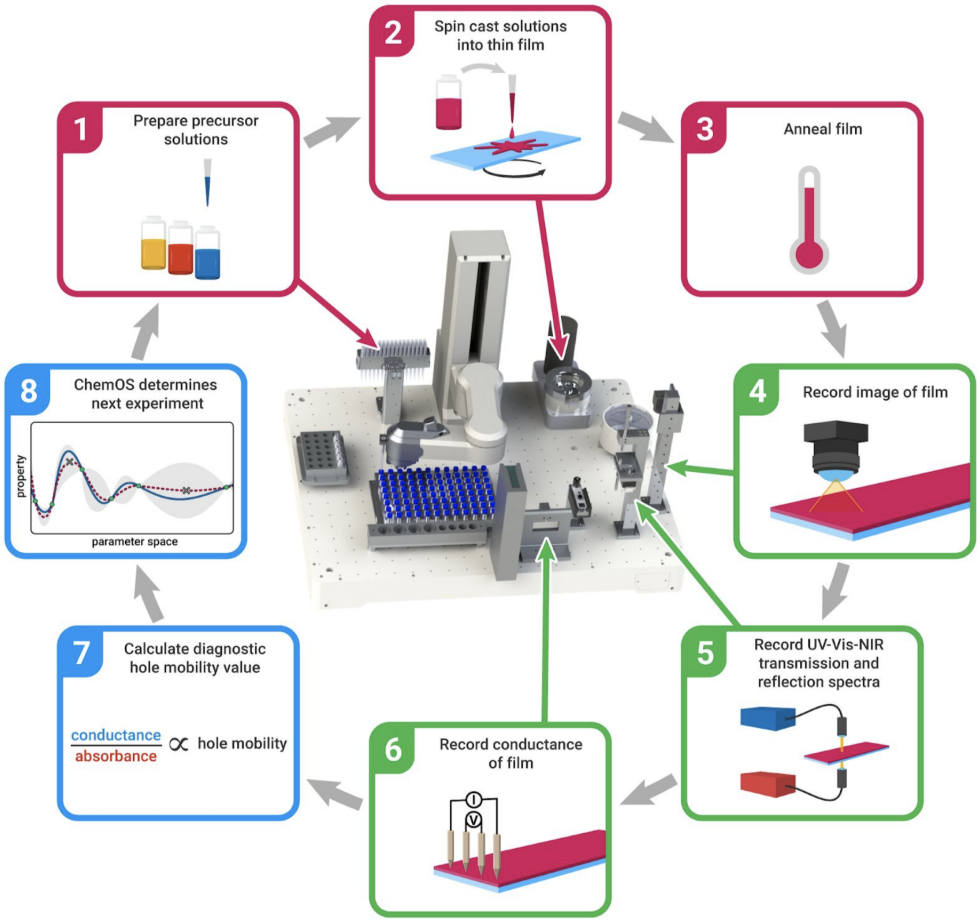


Figure 25: The autonomous platform Ada for optimizing optoelectronic properties of spin cast thin films. Figure reproduced from MacLeod et al. [599].

5.8 Brief summary of discovery in other domains

There are many more attempts to automate aspects of the discovery process and incorporate machine learning into scientific workflows than can be mentioned here. Some pertinent to biology and human health are summarized below. A more comprehensive, collaborative review can be found in ref. 600.

To identify correlations from text mining: The enormous size of the scientific literature makes it challenging to analyze holistically. Information retrieval tools are needed to bring together relevant pieces of information, either for automated analysis or manual inspection. ARROWSMITH was an early system for the latter use [601] that identified MEDLINE abstracts with overlapping terms potentially indicating an implicit causal relationship. This was used to discover a number of testable hypotheses including a link between magnesium and migraines [602]. Literature mining is an essential tool for organizing biological data and enabling computational studies [38, 603–607].

To identify trends in genomics data: The vast quantity of structured genetic information brought about by the Human Genome Project and advances in DNA sequencing is well-suited for data mining. As one example, probabilistic graph models can be built from genetic, protein-interaction, and metabolic pathway information to propose hypotheses for gene functions [608, 609]. Two practical introductions to machine learning for genomics can be found in ref. 610 and ref. 611.

To engineer new proteins: Protein engineering through directed evolution requires navigating a high dimensional and discontinuous structure-function landscape. Random mutagenesis navigates this space blindly, one step at a time, while site-directed mutagenesis requires knowledge of which amino acid positions to perturb. Supervised machine learning models and other statistical techniques can assist in the selection of mutants, as reviewed by Yang et al. [612]. Computational techniques are used for protein engineering in many other ways [613] including protein structure prediction [614, 615].

To identify gene/enzyme relationships: A high-profile example of an automated platform for molecular genetics is King et al.'s Adam [616–618]. In the original demonstration, Adam was made aware of the aromatic amino acid synthesis pathway in yeast, hypothesized which of 15 open reading frames (ORFs) encoded which enzyme, and selected growth experiments to perform (choosing one knockout mutant out of 15 options and one to two metabolites out of 9 options). While this almost represented a truly closed-loop system, there were still manual steps involved in transferring well plates between the liquid handler, incubator, and plate reader. Additionally, these experimental and hypothesis spaces are extraordinarily narrow: the model's accuracy using an active search strategy was 80.1% compared to 74.0% for a naive method that chose the cheapest experiment yet to be performed. In ref. 618, King acknowledges the criticism that "the new scientific knowledge was implicit in the formulation of the problem, and is therefore not novel".

6 Conclusion

In the first part of this review, we have defined three broad categories of discovery–physical matter, processes, and models–and suggested guidelines for evaluating the extent to which a scientific workflow can be described as autonomous: (i) How broadly is the goal defined? (ii) How constrained is the search/design space? (iii) How are experiments for validation/feedback selected? (iv) How superior to a brute force search is navigation of the design space? (v) How are experiments for validation/feedback performed? (vi) How are results organized and interpreted? (vii) Does the discovery outcome contribute to broader scientific knowledge?

As illustrated by the case studies we have included, there has been substantial progress in developing methods that build toward autonomous discovery. Yet there are few examples of true closed-loop discovery for all but the narrowest design spaces–often a consequence of the complexity of automating experimental validation.

We continue to face both practical and methodological challenges in our quest for autonomous discovery. The second part of this review will reflect on a selection of case studies in terms of the questions we pose and then describe remaining challenges where further development is required.

7 Acknowledgements

We thank Thomas Struble for providing comments on the manuscript and our other colleagues and collaborators for useful conversations around this topic. This work was supported by the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium and the DARPA Make-It program under contract ARO W911NF-16-2-0023.

References

[1] A. M. Turing, *Mind* **1950**, *LIX*, 433–460.

[2] G. F. Bradshaw, P. W. Langley, H. A. Simon, *Science* **1983**, *222*, 971–975.

[3] A. M. Turing, *Computers & Thought*, (Eds.: E. A. Feigenbaum, J. Feldman), MIT Press, Cambridge, MA, USA, **1995**, pp. 11–35.

[4] P. Langley, *Int. J. Hum. Comput. Stud.* **2000**, *53*, 393–410.

[5] R. E. Valdés-Pérez, *Artif. Intell.* **1999**, *107*, 335–346.

[6] A. Sparkes et al., *Automated Experimentation* **2010**, *2*, 1.

[7] P. D. Sozou, P. C. Lane, M. Addis, F. Gobet in *Springer Handbook of Model-Based Science*, Springer Handbooks, Springer International Publishing, **2017**, pp. 719–734.

[8] M. Peplow, *Nature* **2014**, *512*, 20–22.

[9] C. Houben, A. A. Lapkin, *Curr. Opin. Chem. Eng.* **2015**, *9*, 1–7.

[10] D. E. Fitzpatrick, C. Battilocchio, S. V. Ley, *Org. Process Res. Dev.* **2016**, *20*, 386–394.

[11] B. J. Reizman, K. F. Jensen, *Acc. Chem. Res.* **2016**, *49*, 1786–1796.

[12] A. G. Godfrey, T. Masquelin, H. Hemmerle, *Drug Discov. Today* **2013**, *18*, 795–802.

[13] J. Li, S. G. Ballmer, E. P. Gillis, S. Fujii, M. J. Schmidt, A. M. E. Palazzolo, J. W. Lehmann, G. F. Morehouse, M. D. Burke, *Science* **2015**, *347*, 1221–1226.

[14] Lowe, Derek, Automated Chemistry: A Vision, en-US, **2018**.

[15] S. Steiner et al., *Science* **2018**, *363*, eaav2211.

[16] G. Schneider, *Nat. Rev. Drug Discov.* **2017**, *17*, 97–113.

[17] S. Smith, 141 Startups Using Artificial Intelligence in Drug Discovery, `https://blog.benchsci.com/startups-using-artificial-intelligence-in-drug-discovery` (visited on 07/30/2019).

[18] P. W. Anderson, E. Abrahams, *Science* **2009**, *324*, 1515–1516.

[19] S. H. Muggleton, *Nature* **2006**, *440*, 409–410.

[20] D. Waltz, B. G. Buchanan, *Science* **2009**, *324*, 43–44.

[21] A. Sharafi, *Knowledge Discovery in Databases*, Springer Fachmedien Wiesbaden, Cambridge, MA, USA, **2013**.

[22] P. S. Gromski, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Rev. Chem.* **2019**, *3*, 119–128.

[23] P. Langley, H. A. Simon, G. L. Bradshaw, J. M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Process*, MIT Press, Cambridge, MA, USA, **1987**.

[24] D. Klahr, *Cogn. Sci.* **1988**, *12*, 1–48.

[25] T. I. Oprea, J. Gottfries, *J. Comb. Chem.* **2001**, *3*, 157–166.

[26] C. Lipinski, A. Hopkins, *Nature* **2004**, *432*, 855–861.

[27] C. M. Dobson, *Nature* **2004**, *432*, 824–828.

[28] J.-L. Reymond, *Acc. Chem. Res.* **2015**, *48*, 722–730.

[29] R. S. Bohacek, C. McMartin, W. C. Guida, *Med. Res. Rev.* **1996**, *16*, 3–50.

[30] K. L. M. Drew, H. Baiman, P. Khwaounjoo, B. Yu, J. Reynisson, *J. Pharm. Pharmacol.* **2012**, *64*, 490–495.

[31] W. P. Walters, *J. Med. Chem.* **2018**, *62*, 1116–1124.

[32] J. P. McMullen, K. F. Jensen, *Org. Process Res. Dev.* **2010**, *14*, 1169–1176.

[33] B. J. Reizman, Y.-M. Wang, S. L. Buchwald, K. F. Jensen, *React. Chem. Eng.* **2016**, *1*, 658–666.

[34] S. E. Denmark, B. L. Christenson, D. M. Coe, S. P. O'Connor, *Tetrahedron Lett.* **1995**, *36*, 2215–2218.

[35] S. L. Schreiber, *Science* **2000**, *287*, 1964–1969.

[36] F. H. Arnold, *Acc. Chem. Res.* **1998**, *31*, 125–131.

[37] M. Schmidt, H. Lipson, *Science* **2009**, *324*, 81–85.

[38] B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, P. K. Sorger, *Mol. Syst. Biol.* **2017**, *13*, 954.

[39] G. Schneider, *Nat. Rev. Drug Discov.* **2010**, *9*, 273–276.

[40] S. K. Saikin, C. Kreisbeck, D. Sheberla, J. S. Becker, A.-G. A., *Expert Opin. Drug Discovery* **2019**, *14*, 1–4.

[41] G. Schneider, D. E. Clark, *Angew. Chem. Int. Ed.* **2019**, *58*, 10792–10803.

[42] M. D. Halls, K. Tasaki, *J. Power Sources* **2010**, *195*, 1472–1478.

[43] E. L. Gettier, *Analysis* **1963**, *23*, 121.

[44] K. Popper, *Conjectures and Refutations, The Growth of Scientific Knowledge*, Routledge, **1962**.

[45] F. Bacon, *Novum organum*, Google-Books-ID: tH4_AAAAYAAJ, Clarendon Press, **1878**, 644 pp.

[46] P. Giza, *J. Exp. Theor. Artif. Intell.* **2017**, *29*, 1053–1069.

[47] D. Silver et al., *Nature* **2017**, *550*, 354–359.

[48] D. Silver et al., *Science* **2018**, *362*, 1140–1144.

[49] D. Klahr, A. Fay, K. Dunbar, *Cognitive Psychology* **1993**, *25*, 111–146.

[50] L. M. Baker, K. Dunbar in Proceedings of the 18th Annual Conference of the Cognitive Science Society, Erlbaum, Mahwah, NJ, **1996**, pp. 154–159.

[51] M. L. Cummings, S. Bruni in *Springer Handbook of Automation*, Springer Berlin Heidelberg, **2009**, pp. 437–447.

[52] M. S. Packer, D. R. Liu, *Nat. Rev. Genet.* **2015**, *16*, 379–394.

[53] P. Langley, J. M. Zytkow, *Artif. Intell.* **1989**, *40*, 283–312.

[54] L. Breiman, *Statist. Sci.* **2001**, *16*, 199–231.

[55] G. Shmueli, *Statist. Sci.* **2010**, *25*, 289–310.

[56] R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, *arXiv preprint arXiv:1905.08883* **2019**.

[57] B. L. Claus, D. J. Underwood, *Drug Discov. Today* **2002**, *7*, 957–966.

[58] W. A. Warr, *Mol. Inf.* **2014**, *33*, 469–476.

[59] A. Gaulton et al., *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

[60] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, *MRS Bull.* **2016**, *41*, 399–409.

[61] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *arXiv:1907.05644 [cond-mat physics:physics]* **2019**.

[62] Y. Gil, M. Greaves, J. Hendler, H. Hirsh, *Science* **2014**, *346*, 171–172.

[63] V. G. Honavar, *Review of Policy Research* **2014**, *31*, 326–330.

[64] Y. Gil, H. Hirsh in 2012 AAAI Fall Symposium Series, **2012**.

[65] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, *J. Cheminform.* **2015**, *7*, S1.

[66] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **2016**, *56*, 1894–1904.

[67] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, A. Valencia, *Chem. Rev.* **2017**, *117*, 7673–7761.

[68] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436–9444.

[69] Z. Zhai, D. Q. Nguyen, S. A. Akhondi, C. Thorne, C. Druckenbrodt, T. Cohn, M. Gregory, K. Verspoor, *arXiv:1907.02679 [cs]* **2019**.

[70] S. Zheng, S. Dharssi, M. Wu, J. Li, Z. Lu in *Methods in Molecular Biology*, (Eds.: R. S. Larson, T. I. Oprea), Methods in Molecular Biology, Springer New York, New York, NY, **2019**, pp. 231–252.

[71] M. Musib et al., *Science* **2017**, *357*, 28–30.

[72] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, T. Blaschke, *Drug Discov. Today* **2018**, *23*, 1241–1250.

[73] G. B. Goh, N. O. Hodas, A. Vishnu, *J. Comput. Chem.* **2017**, *38*, 1291–1307.

[74] J. Vamathevan et al., *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477.

[75] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer Science & Business Media LLC, **2006**.

[76] A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413–1437.

[77] J. B. O. Mitchell, *WIREs Comput Mol Sci* **2014**, *4*, 468–481.

[78] J. Luts, F. Ojeda, R. Van de Plas, B. De Moor, S. Van Huffel, J. A. Suykens, *Anal. Chim. Acta* **2010**, *665*, 129–145.

[79] M. Rupp, *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.

[80] T. Mueller, A. G. Kusne, R. Ramprasad, *Rev. Comput. Chem.* **2016**, *29*, 186–273.

[81] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.

[82] L. Zhang, J. Tan, D. Han, H. Zhu, *Drug Discov. Today* **2017**, *22*, 1680–1685.

[83] B. Settles, *Synthesis Lectures on Artificial Intelligence and Machine Learning* **2012**, *6*, 1–114.

[84] P. I. Frazier, *arXiv preprint arXiv:1807.02811* **2018**.

[85] J. H. Holland, *Adaptation in Natural and Artificial Systems, An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, The MIT Press, **1992**.

[86] R. Salakhutdinov, *Annu. Rev. Stat. Appl.* **2015**, *2*, 361–385.

[87] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio in Advances in neural information processing systems, **2014**, pp. 2672–2680.

[88] D. P. Kingma, M. Welling.

[89] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360–365.

[90] D. Schwalbe-Koda, R. Gómez-Bombarelli, *arXiv:1907.01632 [physics stat]* **2019**.

[91] D. C. Elton, Z. Boukouvalas, M. D. Fuge, P. W. Chung, *arXiv:1903.04388 [physics stat]* **2019**.

[92] S. N. Deming, H. L. Pardue, *Anal. Chem.* **1971**, *43*, 192–200.

[93] H. Winicov, J. Schainbaum, J. Buckley, G. Longino, J. Hill, C. Berkoff, *Anal. Chim. Acta* **1978**, *103*, 469–476.

[94] J. Y. Pan, *ACS Med. Chem. Lett.* **2019**, *10*, 703–707.

[95] L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. E. Yunker, J. E. Hein, A. Aspuru-Guzik, *Sci. Robot.* **2018**, *3*, eaat5559.

[96] C. Elliott, V. Vijayakumar, W. Zink, R. Hansen, *JALA: Journal of the Association for Laboratory Automation* **2007**, *12*, 17–24.

[97] T. Chapman, *Nature* **2003**, *421*, 661–663.

[98] T. Kodadek, *Nat. Chem. Biol.* **2010**, *6*, 162–165.

[99] S. Brenner, R. A. Lerner, *PNAS* **1992**, *89*, 5381–5383.

[100] D. S. Tan, *Nat. Chem. Biol.* **2005**, *1*, 74–84.

[101] R. G. Cooks, *Science* **2006**, *311*, 1566–1570.

[102] C. J. Welch, X. Gong, W. Schafer, E. C. Pratt, T. Brkovic, Z. Pirzada, J. F. Cuff, B. Kosjek, *Tetrahedron: Asymmetry* **2010**, *21*, 1674–1681.

[103] H. A. Simon, P. W. Langley, G. L. Bradshaw, *Synthese* **1981**, *47*, 1–27.

[104] P. Langley, *Proc. 2nd National Conference of the Canadian Society for Computational Studies of Intelligence 1978* **1978**, 173–180.

[105] P. Langley, G. L. Bradshaw, H. A. Simon in *Machine Learning*, (Eds.: R. S. Michalski, J. G. Carbonell, T. M. Mitchell), Symbolic Computation, Springer Berlin Heidelberg, Berlin, Heidelberg, **1983**, pp. 307–329.

[106] J. M. Zytkow in *Proceedings of the Fourth International Workshop on MACHINE LEARNING*, (Ed.: P. Langley), Elsevier, **1987**, pp. 281–287.

[107] B. C. Falkenhainer, R. S. Michalski, *Mach. Learn.* **1986**, *1*, 367–401.

[108] J. M. Zytkow, H. A. Simon, *Mach. Learn.* **1986**, *1*, 107–137.

[109] D. Kulkarni, H. A. Simon, *Cogn. Sci.* **1988**, *12*, 139–175.

[110] S. Fajtlowicz in *Graph Theory and Applications, Proceedings of the First Japan Conference on Graph Theory and Applications*, (Eds.: J. Akiyama, Y. Egawa, H. Enomoto), Graph Theory and Applications, Elsevier, **1988**, pp. 113–118.

[111] I. H. Witten, B. A. MacDonald, *Int. J. Man Mach. Stud.* **1988**, *29*, 171–196.

[112] W. H. Green in *Chemical Engineering Kinetics*, (Ed.: G. B. Marin), Chemical Engineering Kinetics, Elsevier, **2007**, pp. 1–313.

[113] G. N. Simm, M. Reiher, *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.

[114] G. N. Simm, A. C. Vaucher, M. Reiher, *J. Phys. Chem. A* **2019**, *123*, 385–399.

[115] J. P. Unsleber, M. Reiher, *arXiv:1906.10223 [physics]* **2019**.

[116] J. T. Margraf, K. Reuter, *ACS Omega* **2019**, *4*, 3370–3379.

[117] R. E. Valdés-Pérez, *Catal. Lett.* **1994**, *28*, 79–87.

[118] R. E. Valdés-Pérez, *Artif. Intell.* **1994**, *65*, 247–280.

[119] R. E. Valdés-Pérez, *Artif. Intell.* **1995**, *74*, 191–201.

[120] I. Ismail, H. B. V. A. Stuttaford-Fowler, C. Ochan Ashok, C. Robertson, S. Habershon, *J. Phys. Chem. A* **2019**, *123*, 3407–3417.

[121] T. P. Senftle et al., *Npj Comput. Mater.* **2016**, *2*, 15011.

[122] C. W. Gao, J. W. Allen, W. H. Green, R. H. West, *Comput. Phys. Commun.* **2016**, *203*, 212–225.

[123] P. Zhang, N. W. Yee, S. V. Filip, C. E. Hetrick, B. Yang, W. H. Green, *Phys. Chem. Chem. Phys.* **2018**, *20*, 10637–10649.

[124] L. J. Broadbelt, S. M. Stark, M. T. Klein, *Ind. Eng. Chem. Res.* **1994**, *33*, 790–799.

[125] P. M. Zimmerman, *J. Comput. Chem.* **2013**, *34*, 1385–1392.

[126] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*, 14621.

[127] S. Maeda, Y. Harabuchi, *J. Chem. Theory Comput.* **2019**, *15*, 2111–2115.

[128] H. B. Schlegel, *J. Comput. Chem.* **1982**, *3*, 214–218.

[129] A. Behn, P. M. Zimmerman, A. T. Bell, M. Head-Gordon, *J. Chem. Phys.* **2011**, *135*, 224108.

[130] A. Goodrow, A. T. Bell, M. Head-Gordon, *J. Chem. Phys.* **2009**, *130*, 244108.

[131] Y. V. Suleimanov, W. H. Green, *J. Chem. Theory Comput.* **2015**, *11*, 4248–4259.

[132] C. A. Grambow, A. Jamal, Y.-P. Li, W. H. Green, J. Zádor, Y. V. Suleimanov, *J. Am. Chem. Soc.* **2018**, *140*, 1035–1048.

[133] Y. Kim, J. W. Kim, Z. Kim, W. Y. Kim, *Chem. Sci.* **2018**, *9*, 825–835.

[134] S. Maeda, T. Taketsugu, K. Morokuma, *J. Comput. Chem.* **2014**, *35*, 166–173.

[135] L.-P. Wang, A. Titov, R. McGibbon, F. Liu, V. S. Pande, T. J. Martínez, *Nature Chem.* **2014**, *6*, 1044–1048.

[136] L.-P. Wang, R. T. McGibbon, V. S. Pande, T. J. Martinez, *J. Chem. Theory Comput.* **2016**, *12*, 638–649.

[137] T. Lei, W. Guo, Q. Liu, H. Jiao, D.-B. Cao, B. Teng, Y.-W. Li, X. Liu, X.-D. Wen, *J. Chem. Theory Comput.* **2019**, *15*, 3654–3665.

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

[139] S. Soh, Y. Wei, B. Kowalczyk, C. M. Gothard, B. Baytekin, N. Gothard, B. A. Grzybowski, *Chem. Sci.* **2012**, *3*, 1497.

[140] A. I. Lin, T. I. Madzhidov, O. Klimchuk, R. I. Nugmanov, I. S. Antipin, A. Varnek, *J. Chem. Inf. Model.* **2016**, *56*, 2140–2148.

[141] M. D. Bajczyk, P. Dittwald, A. Wołos, S. Szymkuć, B. A. Grzybowski, *Angew. Chem. Int. Ed. Engl.* **2018**, *57*, 2367–2371.

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

[143] J. Li, M. D. Eastgate, *React. Chem. Eng.* **2019**, DOI 10.1039/c9re00019d.

[144] W. D. Smith, **1997**, 7.

[145] S. M. Kim, M. I. Peña, M. Moll, G. N. Bennett, L. E. Kavraki, *Journal of Cheminformatics* **2017**, *9*, 51.

[146] M. Hartenfeller, H. Zettl, M. Walter, M. Rupp, F. Reisen, E. Proschak, S. Weggen, H. Stark, G. Schneider, *PLoS Comput. Biol.* **2012**, *8*, e1002380.

[147] S. Avramova, N. Kochev, P. Angelov, *Data* **2018**, *3*, 14.

[148] S. Szymkuć, E. P. Gajewska, T. Klucznik, K. Molga, P. Dittwald, M. Startek, M. Bajczyk, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2016**, *55*, 5904–5937.

[149] J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Khew, A. P. Johnson, S. Major, R. A. Wade, H. Y. Ando, *J. Chem. Inf. Model.* **2009**, *49*, 593–602.

[150] C. D. Christ, M. Zentgraf, J. M. Kriegl, *J. Chem. Inf. Model.* **2012**, *52*, 1745–1756.

[151] A. Bøgevig, H.-J. Federsel, F. Huerta, M. G. Hutchings, H. Kraut, T. Langer, P. Löw, C. Oppawsky, T. Rein, H. Saller, *Org. Process Res. Dev.* **2015**, *19*, 357–368.

[152] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966–5971.

[153] W.-D. Ihlenfeldt, J. Gasteiger, *Angew. Chem. Int. Ed. Engl.* **1996**, *34*, 2613–2633.

[154] M. H. Todd, *Chem. Soc. Rev.* **2005**, *34*, 247.

[155] A. Cook, A. P. Johnson, J. Law, M. Mirzazadeh, O. Ravitz, A. Simon, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 79–107.

[156] O. Ravitz, *Drug Discov. Today Technol.* **2013**, *10*, e443–e449.

[157] C. W. Coley, W. H. Green, K. F. Jensen, *Acc. Chem. Res.* **2018**, *51*, 1281–1289.

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

[159] S. Zheng, J. Rao, Z. Zhang, J. Xu, Y. Yang, *arXiv:1907.01356 [physics]* **2019**.

[160] E. J. Corey, W. T. Wipke, *Science* **1969**, *166*, 178–192.

[161] A. P. Johnson, C. Marshall, P. N. Judson, *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 310–316.

[162] T. Klucznik et al., *Chem* **2018**, *4*, 522–532.

[163] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604–610.

[164] S. H. Bertz, *J. Am. Chem. Soc.* **1981**, *103*, 3599–3601.

[165] P. Ertl, A. Schuffenhauer, *J. Cheminform.* **2009**, *1*, 8.

[166] R. P. Sheridan, N. Zorn, E. C. Sherer, L.-C. Campeau, C. (Chang, J. Cumming, M. L. Maddess, P. G. Nantermet, C. J. Sinz, P. D. O'Shea, *J. Chem. Inf. Model.* **2014**, *54*, 1604–1616.

[167] J. R. Proudfoot, *J. Org. Chem.* **2017**, *82*, 6968–6971.

[168] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2018**, *58*, 252–261.

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

[170] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 1237–1245.

[171] A. Heifets, PhD thesis, **2014**.

[172] S. Rangarajan, A. Bhan, P. Daoutidis, *Comput. Chem. Eng.* **2012**, *46*, 141–152.

[173] S. Rangarajan, A. Bhan, P. Daoutidis, *Ind. Eng. Chem. Res.* **2010**, *49*, 10459–10470.

[174] D. M. Camacho, K. M. Collins, R. K. Powers, J. C. Costello, J. J. Collins, *Cell* **2018**, *173*, 1581–1592.

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

[176] M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *J. Am. Chem. Soc.* **2018**, *140*, 5004–5008.

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

[178] T. Mikolov, K. Chen, G. Corrado, J. Dean, *arXiv:1301.3781 [cs]* **2013**.

[179] R. Banãres-Alcántara, E. Ko, A. Westerberg, M. Rychener, *Comput. Chem. Eng.* **1988**, *12*, 923–938.

[180] C. Reichardt, T. Welton, *Solvents and Solvent Effects in Organic Chemistry*, Wiley-VCH Verlag GmbH & Co. KGaA, **2011**, 425 pp.

[181] P. R. Wells, *Chemical Reviews* **1963**, *63*, 171–219.

[182] R. W. Taft, J.-L. M. Abboud, M. J. Kamlet, M. H. Abraham, *J. Solution Chem.* **1985**, *14*, 153–186.

[183] H. Struebing, Z. Ganase, P. G. Karamertzanis, E. Siougkrou, P. Haycock, P. M. Piccione, A. Armstrong, A. Galindo, C. S. Adjiman, *Nature Chem.* **2013**, *5*, 952–957.

[184] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73–76.

[185] R. J. Xu, J. H. Olshansky, P. D. F. Adler, Y. Huang, M. D. Smith, M. Zeller, J. Schrier, A. J. Norquist, *Mol. Syst. Des. Eng.* **2018**, *3*, 473–484.

[186] J. Li, T. Chen, K. Lim, L. Chen, S. A. Khan, J. Xie, X. Wang, *Advanced Intelligent Systems* **2019**, *1*, arXiv: 1811.02771, 1900029.

[187] E. Kim, K. Huang, A. Tomala, S. Matthews, E. Strubell, A. Saunders, A. McCallum, E. Olivetti, *Sci. Data* **2017**, *4*, 170127.

[188] E. Kim, K. Huang, S. Jegelka, E. Olivetti, *Npj Comput. Mater.* **2017**, *3*, 53.

[189] E. Kim et al., *arXiv:1901.00032 [cond-mat stat]* **2018**.

[190] Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS Cent. Sci.* **2019**, DOI 10.1021/acscentsci.9b00193.

[191] J. Gasteiger, C. Jochum in *Organic Compunds*, *Vol. 74*, Springer-Verlag, Berlin/Heidelberg, **1978**, pp. 93–126.

[192] I. Ugi, J. Bauer, J. Brandt, J. Friedrich, J. Gasteiger, C. Jochum, W. Schubert, *Angew. Chem. Int. Ed.* **1979**, *18*, 111–123.

[193] T. D. Salatin, W. L. Jorgensen, *J. Org. Chem.* **1980**, *45*, 2043–2051.

[194] G. Sello, *J. Chem. Inf. Model.* **1992**, *32*, 713–717.

[195] H. Satoh, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 34–44.

[196] I. M. Socorro, K. Taylor, J. M. Goodman, *Org. Lett.* **2005**, *7*, 3541–3544.

[197] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, *23*, 6118–6128.

[198] P.-M. Jacob, A. Lapkin, **2018**, DOI 10.26434/chemrxiv.6954908.v1.

[199] M. A. Kayala, C.-A. Azencott, J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2011**, *51*, 2209–2222.

[200] M. A. Kayala, P. Baldi, *J. Chem. Inf. Model.* **2012**, *52*, 2526–2540.

[201] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, N. Huynh, D. Van Vranken, P. Baldi, *Mol. Syst. Des. Eng.* **2017**, *3*, 442–452.

[202] J. Bradshaw, M. J. Kusner, B. Paige, M. H. S. Segler, J. M. Hernández-Lobato, *arXiv:1805.10970 [physics stat]* **2018**.

[203] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Central Science* **2016**, *2*, 725–732.

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

[205] W. Jin, C. Coley, R. Barzilay, T. Jaakkola, *NeurIPS* **2017**, 2604–2613.

[206] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370–377.

[207] J. Nam, J. Kim, *arXiv:1612.09529 [cs]* **2016**.

[208] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, T. Laino, *arXiv:1711.04810 null* **2017**, *9*, 6091–6098.

[209] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. Bekas, A. A. Lee, **2018**, DOI 10.26434/chemrxiv.7297379.v1.

[210] J. A. Platts, D. Butina, M. H. Abraham, A. Hersey, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835–845.

[211] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96–103.

[212] J. Gasteiger, U. Hondelmann, P. Röse, W. Witzenbichler, *J. Chem. Soc. Perkin Trans. 2* **1995**, *0*, 193–204.

[213] J. Gálvez, M. Gálvez-Llompart, R. García-Domenech, *Green Chem.* **2010**, *12*, 1056.

[214] T. I. Madzhidov, P. G. Polishchuk, R. I. Nugmanov, A. V. Bodrov, A. I. Lin, I. I. Baskin, A. A. Varnek, I. S. Antipin, *Russ. J. Org. Chem.* **2014**, *50*, 459–463.

[215] P. Polishchuk, T. Madzhidov, T. Gimadiev, A. Bodrov, R. Nugmanov, A. Varnek, *J. Comput.-Aided Mol. Des.* **2017**, *31*, 829–839.

[216] M. Glavatskikh, T. Madzhidov, D. Horvath, R. Nugmanov, T. Gimadiev, D. Malakhova, G. Marcou, A. Varnek, *Mol. Inf.* **2019**, *38*, 1800077.

[217] T. I. Madzhidov, T. R. Gimadiev, D. A. Malakhova, R. I. Nugmanov, I. I. Baskin, I. S. Antipin, A. A. Varnek, *J. Struct. Chem.* **2017**, *58*, 650–656.

[218] Q. N. N. Nguyen, D. J. Tantillo, *Chem. Asian J.* **2013**, *9*, 674–680.

[219] K. Fukui, H. Fujimoto, *Frontier Orbitals and Reaction Paths, Selected Papers of Kenichi Fukui*, Google-Books-ID: azpkDQAAQBAJ, WORLD SCIENTIFIC, **1997**, 563 pp.

[220] P. W. Ayers, R. G. Parr, *J. Am. Chem. Soc.* **2000**, *122*, 2010–2018.

[221] A. Verloop in *Pesticide Chemistry: Human Welfare and Environment*, (Eds.: P. Doyle, T. Fujita), Elsevier, **1983**, pp. 339–344.

[222] S. H. Unger, C. Hansch in *Progress in Physical Organic Chemistry*, John Wiley & Sons, Inc., **2007**, pp. 91–118.

[223] J. D. Oslob, B. Åkermark, P. Helquist, P.-O. Norrby, *Organometallics* **1997**, *16*, 3015–3021.

[224] A. Milo, A. J. Neel, F. D. Toste, M. S. Sigman, *Science* **2015**, *347*, 737–743.

[225] M. S. Sigman, K. C. Harper, E. N. Bess, A. Milo, *Acc. Chem. Res.* **2016**, *49*, 1292–1301.

[226] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631.

[227] J. P. Reid, M. S. Sigman, *Nature* **2019**, *571*, 343–348.

[228] S. Banerjee, A. Sreenithya, R. B. Sunoj, *Phys. Chem. Chem. Phys.* **2018**, *20*, 18311–18318.

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

[230] G. A. Landrum, J. E. Penzotti, S. Putta, *Meas. Sci. Technol.* **2004**, *16*, 270–277.

[231] M. Elkin, T. R. Newhouse, *Chem. Soc. Rev.* **2018**, *47*, 7830–7844.

[232] D. C. Blakemore, L. Castro, I. Churcher, D. C. Rees, A. W. Thomas, D. M. Wilson, A. Wood, *Nature Chem.* **2018**, *10*, 383–394.

[233] J. Boström, D. G. Brown, R. J. Young, G. M. Keserü, *Nat. Rev. Drug Discov.* **2018**, *17*, 709–727.

[234] J.-M. Lehn, *Chem. Eur. J.* **1999**, *5*, 2455–2463.

[235] K. H. Shaughnessy, P. Kim, J. F. Hartwig, *J. Am. Chem. Soc.* **1999**, *121*, 2123–2132.

[236] M. T. Reetz, M. H. Becker, H.-W. Klein, D. Stöckigt, *Angew. Chem. Int. Ed.* **1999**, *38*, 1758–1761.

[237] A. Buitrago Santanilla et al., *Science* **2014**, *347*, 49–53.

[238] S. Lin et al., *Science* **2018**, *361*, eaar6236.

[239] D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson, N. W. Sach, *Science* **2018**, *359*, 429–434.

[240] M. Wleklinski, B. P. Loren, C. R. Ferreira, Z. Jaman, L. Avramova, T. J. P. Sobreira, D. H. Thompson, R. G. Cooks, *Chem. Sci.* **2018**, *9*, 1647–1653.

[241] E. S. Isbrandt, R. J. Sullivan, S. G. Newman, *Angew. Chem. Int. Ed.* **2019**, *58*, 7180–7191.

[242] M. Shevlin, M. R. Friedfeld, H. Sheng, N. A. Pierson, J. M. Hoyt, L.-C. Campeau, P. J. Chirik, *J. Am. Chem. Soc.* **2016**, *138*, 3562–3569.

[243] M. Teders, A. Gómez-Suárez, L. Pitzer, M. N. Hopkinson, F. Glorius, *Angew. Chem. Int. Ed.* **2017**, *56*, 902–906.

[244] P. S. Kutchukian et al., *Chem. Sci.* **2016**, *7*, 2604–2613.

[245] A. B. Beeler, S. Su, C. A. Singleton, J. A. Porco, *J. Am. Chem. Soc.* **2007**, *129*, 1413–1419.

[246] H. M. Geysen, C. D. Wagner, W. M. Bodnar, C. J. Markworth, G. J. Parke, F. J. Schoenen, D. S. Wagner, D. S. Kinder, *Chem. Biol.* **1996**, *3*, 679–688.

[247] M. W. Kanan, M. M. Rozenman, K. Sakurai, T. M. Snyder, D. R. Liu, *Nature* **2004**, *431*, 545–549.

[248] D. W. Robbins, J. F. Hartwig, *Science* **2011**, *333*, 1423–1427.

[249] K. Troshin, J. F. Hartwig, *Science* **2017**, *357*, 175–181.

[250] A. McNally, C. K. Prier, D. W. C. MacMillan, *Science* **2011**, *334*, 1114–1117.

[251] L. Weber, K. Illgen, M. Almstetter, *Synlett* **1999**, *1999*, 366–374.

[252] A. Dömling, *Curr. Opin. Chem. Biol.* **2000**, *4*, 318–323.

[253] B. Ganem, *Acc. Chem. Res.* **2009**, *42*, 463–472.

[254] K. D. Collins, T. Gensch, F. Glorius, *Nature Chem.* **2014**, *6*, 859–871.

[255] A. Sugimoto, T. Fukuyama, M. T. Rahman, I. Ryu, *Tetrahedron Lett.* **2009**, *50*, 6364–6367.

[256] K. Koch, B. J. A. van Weerdenburg, J. M. M. Verkade, P. J. Nieuwland, F. P. J. T. Rutjes, J. C. M. van Hest, *Org. Process Res. Dev.* **2009**, *13*, 1003–1006.

[257] P. J. Nieuwland, R. Segers, K. Koch, J. C. M. van Hest, F. P. J. T. Rutjes, *Org. Process Res. Dev.* **2011**, *15*, 783–787.

[258] D. L. Browne, S. Wright, B. J. Deadman, S. Dunnage, I. R. Baxendale, R. M. Turner, S. V. Ley, *Rapid Commun. Mass Spectrom.* **2012**, *26*, 1999–2010.

[259] R. L. Hartman, K. F. Jensen, *Lab Chip* **2009**, *9*, 2495.

[260] D. C. Fabry, E. Sugiono, M. Rueping, *Isr. J. Chem.* **2014**, *54*, 341–350.

[261] S. V. Ley, D. E. Fitzpatrick, R. J. Ingham, R. M. Myers, *Angew. Chem. Int. Ed.* **2015**, *54*, 3449–3464.

[262] D. K. B. Mohamed, X. Yu, J. Li, J. Wu, *Tetrahedron Lett.* **2016**, *57*, 3965–3977.

[263] V. Sans, L. Cronin, *Chem. Soc. Rev.* **2016**, *45*, 2032–2043.

[264] C. Mateos, M. J. Nieves-Remacha, J. A. Rincón, *React. Chem. Eng.* **2019**, DOI 10.1039/c9re00116f.

[265] C. S. Horbaczewskyj, C. E. Willans, A. A. Lapkin, R. A. Bourne in *Handbook of Green Chemistry*, American Cancer Society, **2019**, pp. 329–374.

[266] R. Fletcher, *Comput. J.* **1964**, *7*, 149–154.

[267] J. A. Nelder, R. Mead, *Comput. J.* **1965**, *7*, 308–313.

[268] J. H. Holland, *J. ACM* **1962**, *9*, 297–314.

[269] W. Huyer, A. Neumaier, *ACM Trans. Math. Softw.* **2008**, *35*, 1–25.

[270] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar, L. A. Escaleira, *Talanta* **2008**, *76*, 965–977.

[271] M. Pelikan, D. E. Goldberg, E. Cantú-Paz in Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation - Volume 1, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, **1999**, pp. 525–532.

[272] R. S. Sutton, A. G. Barto, *Reinforcement Learning*, Springer US, **2018**, 549 pp.

[273] J. P. McMullen, K. F. Jensen, *Annual Rev. Anal. Chem.* **2010**, *3*, 19–42.

[274] A. J. Parrott, R. A. Bourne, G. R. Akien, D. J. Irvine, M. Poliakoff, *Angew. Chem. Int. Ed.* **2011**, *50*, 3788–3792.

[275] R. A. Skilton, A. J. Parrott, M. W. George, M. Poliakoff, R. A. Bourne, *Appl. Spectrosc.* **2013**, *67*, 1127–1131.

[276] V. Sans, L. Porwol, V. Dragone, L. Cronin, *Chem. Sci.* **2015**, *6*, 1258–1264.

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

[278] K. Poscharny, D. Fabry, S. Heddrich, E. Sugiono, M. Liauw, M. Rueping, *Tetrahedron*, Engineering Chemistry for the Future of Organic Synthesis **2018**, *74*, 3171–3175.

[279] M. Rubens, J. H. Vrijsen, J. Laun, T. Junkers, *Angew. Chem. Int. Ed.* **2019**, *58*, 3183–3187.

[280] J. S. Moore, K. F. Jensen, *Org. Process Res. Dev.* **2012**, *16*, 1409–1415.

[281] A. Echtermeyer, Y. Amar, J. Zakrzewski, A. Lapkin, *Beilstein J. Org. Chem.* **2017**, *13*, 150–163.

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

[283] A.-C. Bédard, A. Adamo, K. C. Aroh, M. G. Russell, A. A. Bedermann, J. Torosian, B. Yue, K. F. Jensen, T. F. Jamison, *Science* **2018**, *361*, 1220–1225.

[284] Y.-J. Hwang, C. W. Coley, M. Abolhasani, A. L. Marzinzik, G. Koch, C. Spanka, H. Lehmann, K. F. Jensen, *Chem. Commun.* **2017**, *53*, 6649–6652.

[285] J. E. Kreutz, A. Shukhaev, W. Du, S. Druskin, O. Daugulis, R. F. Ismagilov, *J. Am. Chem. Soc.* **2010**, *132*, 3128–3132.

[286] B. J. Reizman, Thesis, Massachusetts Institute of Technology, **2015**.

[287] Z. Zhou, X. Li, R. N. Zare, *ACS Cent. Sci.* **2017**, *3*, 1337–1344.

[288] D. Reker, G. J. L. Bernardes, T. Rodrigues, *chemRxiv* **2018**, DOI 10.26434/chemrxiv.7291205.v1.

[289] B. E. Walker, J. H. Bannock, A. M. Nightingale, J. C. deMello, *React. Chem. Eng.* **2017**, *2*, 785–798.

[290] C. Houben, N. Peremezhney, A. Zubov, J. Kosek, A. A. Lapkin, *Org. Process Res. Dev.* **2015**, *19*, 1049–1053.

[291] F. Häse, L. M. Roch, C. Kreisbeck, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 1134–1145.

[292] S. Garg, S. K. Gupta, *Macromol. Theory Simul.* **1999**, *8*, 46–53.

[293] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, A. A. Lapkin, *Chem. Eng. J.* **2018**, *352*, 277–282.

[294] D. Cortés-Borda et al., *J. Org. Chem.* **2018**, *83*, 14286–14299.

[295] E. Wimmer, D. Cortés-Borda, S. Brochard, E. Barré, C. Truchet, F.-X. Felpin, *React. Chem. Eng.* **2019**, DOI 10.1039/c9re00096h.

[296] P. Sagmeister, J. D. Williams, C. A. Hone, C. O. Kappe, *React. Chem. Eng.* **2019**, DOI 10.1039/c9re00087a.

[297] S. Krishnadasan, R. J. C. Brown, A. J. deMello, J. C. deMello, *Lab Chip* **2007**, *7*, 1434.

[298] V. Duros, J. Grizou, W. Xuan, Z. Hosni, D.-L. Long, H. N. Miras, L. Cronin, *Angew. Chem. Int. Ed.* **2017**, *56*, 10815–10820.

[299] V. Duros, J. Grizou, A. Sharma, S. H. M. Mehr, A. Bubliauskas, P. Frei, H. N. Miras, L. Cronin, *J. Chem. Inf. Model.* **2019**, *59*, 2664–2671.

[300] P. B. Wigley et al., *Sci. Rep.* **2016**, *6*, 25890.

[301] S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou, B. Smit, *Nat. Commun.* **2019**, *10*, 539.

[302] P. Nikolaev, D. Hooper, N. Perea-López, M. Terrones, B. Maruyama, *ACS Nano* **2014**, *8*, 10214–10222.

[303] P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, *Npj Comput. Mater.* **2016**, *2*, 16031.

[304] J. P. McMullen, K. F. Jensen, *Org. Process Res. Dev.* **2011**, *15*, 398–407.

[305] C. M. Anderson-Cook, C. M. Borror, D. C. Montgomery, *J. Stat. Plan. Inference* **2009**, *139*, 629–641.

[306] B. J. Reizman, K. F. Jensen, *Org. Process Res. Dev.* **2012**, *16*, 1770–1782.

[307] Z. Amara, E. S. Streng, R. A. Skilton, J. Jin, M. W. George, M. Poliakoff, *Eur. J. Org. Chem.* **2015**, *2015*, 6141–6145.

[308] J. M. Granda, L. Donina, V. Dragone, D.-L. Long, L. Cronin, *Nature* **2018**, *559*, 377–381.

[309] A. Orita, Y. Yasui, J. Otera, *Org. Process Res. Dev.* **2000**, *4*, 333–336.

[310] J. K. Sader, J. E. Wulff, *Nature* **2019**, *570*, E54–E59.

[311] V. Dragone, V. Sans, A. B. Henson, J. M. Granda, L. Cronin, *Nat. Commun.* **2017**, *8*, 15733.

[312] A. Dudek, T. Arodz, J. Galvez, *Comb. Chem. High Throughput Screen.* **2006**, *9*, 213–228.

[313] A. Cherkasov et al., *J. Med. Chem.* **2014**, *57*, 4977–5010.

[314] B. C. Pearce, M. J. Sofia, A. C. Good, D. M. Drexler, D. A. Stock, *J. Chem. Inf. Model.* **2006**, *46*, 1060–1068.

[315] J. B. Baell, G. A. Holloway, *J. Med. Chem.* **2010**, *53*, 2719–2740.

[316] J. B. Baell, J. W. M. Nissink, *ACS Chem. Biol.* **2018**, *13*, 36–44.

[317] D. Sanderson, C. Earnshaw, *Hum Exp Toxicol* **1991**, *10*, 261–273.

[318] J. Kazius, R. McGuire, R. Bursi, *J. Med. Chem.* **2005**, *48*, 312–320.

[319] I. Sushko, E. Salmina, V. A. Potemkin, G. Poda, I. V. Tetko, *J. Chem. Inf. Model.* **2012**, *52*, 2310–2316.

[320] N. Basant, S. Gupta, K. P. Singh, *J. Chem. Inf. Model.* **2015**, *55*, 1337–1348.

[321] J.-P. Métivier, A. Lepailleur, A. Buzmakov, G. Poezevara, B. Crémilleux, S. O. Kuznetsov, J. L. Goff, A. Napoli, R. Bureau, B. Cuissart, *J. Chem. Inf. Model.* **2015**, *55*, 925–940.

[322] X. Li, X. Yan, Q. Gu, H. Zhou, D. Wu, J. Xu, *J. Chem. Inf. Model.* **2019**, *59*, 1044–1049.

[323] F. Doshi-Velez, B. Kim, **2017**.

[324] R. Tibshirani, *J. Royal Stat. Soc.* **1996**, *58*, 267–288.

[325] I. Guyon, A. Elisseeff, *Journal of Machine Learning Research* **2003**, *3*, 1157–1182.

[326] P. Polishchuk, *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.

[327] R. D. King, S. H. Muggleton, A. Srinivasan, M. J. Sternberg, *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 438–442.

[328] R. D. King, A. Srinivasan, *Environ. Health Perspect.* **1996**, *104*, 1031–1040.

[329] P. Finn, S. Muggleton, D. Page, A. Srinivasan, *Machine Learning* **1998**, *30*, 241–270.

[330] C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178–180.

[331] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.

[332] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.

[333] V. E. Kuz'min, P. G. Polishchuk, A. G. Artemenko, S. A. Andronati, *Mol. Inf.* **2011**, *30*, 593–603.

[334] G. G. Towell, J. W. Shavlik, *Mach. Learn.* **1993**, *13*, 71–101.

[335] N. Barakat, A. P. Bradley, *Neurocomputing*, Artificial Brains **2010**, *74*, 178–190.

[336] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, T. Unterthiner, *arXiv:1903.02788 [cs q-bio stat]* **2019**.

[337] E. Byvatov, G. Schneider, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993–999.

[338] M. Eklund, U. Norinder, S. Boyer, L. Carlsson, *J. Chem. Inf. Model.* **2014**, *54*, 837–843.

[339] C. W. Coley, R. Barzilay, W. H. Green, T. S. Jaakkola, K. F. Jensen, *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.

[340] A. A. Lee, Q. Yang, A. Bassyouni, C. R. Butler, X. Hou, S. Jenkinson, D. A. Price, *PNAS* **2019**, *116*, 3373–3378.

[341] K. Hansen, D. Baehrens, T. Schroeter, M. Rupp, K.-R. Müller, *Mol. Inf.* **2011**, *30*, 817–826.

[342] P. W. Koh, P. Liang in Proceedings of the 34th International Conference on Machine Learning - Volume 70, JMLR.org, **2017**, pp. 1885–1894.

[343] S. Riniker, G. A. Landrum, *J. Cheminform.* **2013**, *5*, 43.

[344] P. G. Polishchuk, V. E. Kuz'min, A. G. Artemenko, E. N. Muratov, *Mol. Inf.* **2013**, *32*, 843–853.

[345] P. Polishchuk, O. Tinkov, T. Khristova, L. Ognichenko, A. Kosinskaya, A. Varnek, V. Kuz'min, *J. Chem. Inf. Model.* **2016**, *56*, 1455–1469.

[346] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.

[347] B. G. Buchanan, G. L. Sutherland, E. A. Feigenbaum, Heuristic DENDRAL: A Program for Generating Explanatory Hypotheses in Organic Chemistry, English, Monographs, **1969**.

[348] E. A. Feigenbaum, B. G. Buchanan, J. Lederberg, On Generality and Problem Solving: A Case Study Using the DENDRAL Program, English, Monographs, **1971**.

[349] F. Hufsky, S. Böcker, *Mass Spec. Rev.* **2017**, *36*, 624–633.

[350] J. N. Wei, D. Belanger, R. P. Adams, D. Sculley, *arXiv:1811.08545 [physics stat]* **2018**, *5*, 700–708.

[351] F. Allen, R. Greiner, D. Wishart, *Metabolomics* **2015**, *11*, 98–110.

[352] I. Blaženović, T. Kind, J. Ji, O. Fiehn, *Metabolites* **2018**, *8*, 31.

[353] M. A. Samaraweera, L. M. Hall, D. W. Hill, D. F. Grant, *Anal. Chem.* **2018**, *90*, 12752–12760.

[354] M. Ludwig, K. Dührkop, S. Böcker, *Bioinformatics* **2018**, *34*, i333–i340.

[355] G. Böhm, R. Muhr, R. Jaenicke, *Protein Eng. Des. Sel.* **1992**, *5*, 191–195.

[356] J. Aires-de-Sousa, M. C. Hemmer, J. Gasteiger, *Anal. Chem.* **2002**, *74*, 80–90.

[357] Y. Binev, M. M. B. Marques, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2007**, *47*, 2089–2097.

[358] M. Gastegger, J. Behler, P. Marquetand, *Chem. Sci.* **2017**, *8*, 6924–6935.

[359] C. Nantasenamat, C. Isarankura-Na-Ayudhya, N. Tansila, T. Naenna, V. Prachayasittikul, *J. Comput. Chem.* **2007**, *28*, 1275–1289.

[360] H. S. Stein, D. Guevarra, P. F. Newhouse, E. Soedarmadji, J. M. Gregoire, *Chem. Sci.* **2019**, *10*, 47–55.

[361] J. Ling, M. Hutchinson, E. Antono, B. DeCost, E. A. Holm, B. Meredig, *Materials Discovery* **2017**, *10*, 19–28.

[362] B. G. Sumpter, D. W. Noid, *Annu. Rev. Mater. Sci.* **1996**, *26*, 223–277.

[363] J. Aires de Sousa in *Applied Chemoinformatics*, (Eds.: T. Engel, J. Gasteiger), Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2018**, pp. 133–163.

[364] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.

[365] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[366] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. Anatole von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.

[367] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. von Lilienfeld, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

[368] G. Schmitz, I. H. Godtliebsen, O. Christiansen, *J. Chem. Phys.* **2019**, *150*, 244113.

[369] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **2011**, *108*, DOI 10.1103/physrevlett.108.253002.

[370] J. Behler, *Angew. Chem. Int. Ed.* **2017**, *56*, 12828–12840.

[371] M. Welborn, L. Cheng, T. F. Miller, *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.

[372] L. Cheng, M. Welborn, A. S. Christensen, T. F. Miller, *J. Chem. Phys.* **2019**, *150*, 131103.

[373] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192–3203.

[374] J. S Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. Roitberg, DOI 10.26434/chemrxiv.6744440.

[375] E. V. Podryabinkin, A. V. Shapeev, *Comput. Mater. Sci.* **2017**, *140*, 171–180.

[376] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, *J. Chem. Phys.* **2018**, *148*, 241733.

[377] S. L. Mayo, B. D. Olafson, W. A. Goddard, *J. Phys. Chem.* **1990**, *94*, 8897–8909.

[378] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, W. M. Skiff, *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.

[379] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, D. A. Case, *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[380] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[381] L. Zhang, J. Han, H. Wang, R. Car, W. E, *Phys. Rev. Lett.* **2018**, *120*, 143001.

[382] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, G. Csányi, *J. Phys. Chem. Lett.* **2018**, *9*, 2879–2885.

[383] W. Wang, R. Gómez-Bombarelli, *arXiv:1812.02706 [physics stat]* **2018**.

[384] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, C. Clementi, *ACS Cent. Sci.* **2019**, DOI 10.1021/acscentsci.8b00913.

[385] W. F. Reinhart, A. W. Long, M. P. Howard, A. L. Ferguson, A. Z. Panagiotopoulos, *Soft Matter* **2017**, *13*, 4733–4745.

[386] A. Mardt, L. Pasquali, H. Wu, F. Noé, *Nat. Commun.* **2018**, *9*, 5.

[387] G. L. Warren et al., *J. Med. Chem.* **2006**, *49*, 5912–5931.

[388] N. S. Pagadala, K. Syed, J. Tuszynski, *Biophys. Rev.* **2017**, *9*, 91–102.

[389] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, R. Wang, *J. Chem. Inf. Model.* **2019**, *59*, 895–913.

[390] P. J. Ballester, J. B. O. Mitchell, *Bioinformatics* **2010**, *26*, 1169–1175.

[391] Q. U. Ain, A. Aleksandrova, F. D. Roessler, P. J. Ballester, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2015**, *5*, 405–424.

[392] J. C. Pereira, E. R. Caffarena, C. N. dos Santos, *J. Chem. Inf. Model.* **2016**, *56*, 2495–2506.

[393] E. J. Bjerrum, *Comput. Biol. Chem.* **2016**, *62*, 133–144.

[394] M. Wójcikowski, P. J. Ballester, P. Siedlecki, *Sci. Rep.* **2017**, *7*, 46710.

[395] J. Jiménez, M. Škalič, G. Martínez-Rosell, G. De Fabritiis, *J. Chem. Inf. Model.* **2018**, *58*, 287–296.

[396] M. M. Stepniewska-Dziubinska, P. Zielenkiewicz, P. Siedlecki, *Bioinformatics* **2018**, *34*, 3666–3674.

[397] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, G. Ceder, *Physical review letters* **2003**, *91*, 135503.

[398] C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, *Nature Mater.* **2006**, *5*, 641–646.

[399] Z. W. Ulissi, A. R. Singh, C. Tsai, J. K. Nørskov, *J. Phys. Chem. Lett.* **2016**, *7*, 3931–3935.

[400] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nat. Commun.* **2018**, *9*, DOI 10.1038/s41467-018-05169-6.

[401] O. Levy, G. L. W. Hart, S. Curtarolo, *J. Am. Chem. Soc.* **2010**, *132*, 4830–4833.

[402] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, G. Ceder, *Chem. Mater.* **2010**, *22*, 3762–3767.

[403] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *89*, 094104.

[404] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier, A. Zunger, *Nature Chem.* **2015**, *7*, 308–316.

[405] N. Artrith, A. Urban, G. Ceder, *J. Chem. Phys.* **2018**, *148*, 241711.

[406] P. Nguyen, T. Tran, S. Gupta, S. Rana, S. Venkatesh, *arXiv:1811.06060 [cond-mat stat]* **2018**, arXiv: 1811.06060.

[407] J. G. P. Wicker, R. I. Cooper, *CrystEngComm* **2015**, *17*, 1927–1934.

[408] A. M. Hiszpanski, C. J. Dsilva, I. G. Kevrekidis, Y.-L. Loo, *Chem. Mater.* **2018**, *30*, 3330–3337.

[409] L. M. Mayr, D. Bojanic, *Curr. Opin. Pharmacol.*, Anti-infectives/New technologies **2009**, *9*, 580–588.

[410] W. P. Janzen, *Chem. Biol.* **2014**, *21*, 1162–1170.

[411] J. Scheuermann, C. E. Dumelin, S. Melkko, D. Neri, *J. Biotechnol.* **2006**, *126*, 568–581.

[412] L. Mannocci, M. Leimbacher, M. Wichert, J. Scheuermann, D. Neri, *Chemical Communications* **2011**, *47*, 12747.

[413] L. A. Thompson, J. A. Ellman, *Chem. Rev.* **1996**, *96*, 555–600.

[414] M. A. Clark et al., *Nat. Chem. Biol.* **2009**, *5*, 647–654.

[415] A. Litovchick et al., *Sci. Rep.* **2015**, *5*, 1–8.

[416] Y. Ding et al., *ACS Med. Chem. Lett.* **2015**, *6*, 888–893.

[417] P. A. Harris et al., *J. Med. Chem.* **2016**, *59*, 2163–2178.

[418] C. S. Kollmann et al., *Bioorg. Med. Chem.* **2014**, *22*, 2353–2365.

[419] H. Deng et al., *J. Med. Chem.* **2012**, *55*, 7061–7079.

[420] R. A. Goodnow, C. E. Dumelin, A. D. Keefe, *Nat. Rev. Drug Discov.* **2016**, *16*, 131–147.

[421] W. R. Galloway, A. Isidro-Llobet, D. R. Spring, *Nat. Commun.* **2010**, *1*, 1–13.

[422] J. E. Biggs-Houck, A. Younai, J. T. Shaw, *Curr. Opin. Chem. Biol.* **2010**, *14*, 371–382.

[423] R. J. Spandl, M. Díaz-Gavilán, K. M. G. O'Connell, G. L. Thomas, D. R. Spring, *The Chemical Record* **2008**, *8*, 129–142.

[424] M. Garcia-Castro, S. Zimmermann, M. G. Sankar, K. Kumar, *Angew. Chem. Int. Ed.* **2016**, *55*, 7586–7605.

[425] N. J. Gesmundo, B. Sauvagnat, P. J. Curran, M. P. Richards, C. L. Andrews, P. J. Dandliker, T. Cernak, *Nature* **2018**, *557*, 228–232.

[426] P. O. Krutzik, G. P. Nolan, *Nat. Methods* **2006**, *3*, 361–368.

[427] J. Swann et al., *ACS Infect. Dis.* **2016**, *2*, 281–293.

[428] J. Inglese, D. S. Auld, A. Jadhav, R. L. Johnson, A. Simeonov, A. Yasgar, W. Zheng, C. P. Austin, *React. Chem. Eng.s* **2006**, *103*, 11473–11478.

[429] M. T. Guo, A. Rotem, J. A. Heyman, D. A. Weitz, *Lab Chip* **2012**, *12*, 2146.

[430] W. P. Walters, M. Namchuk, *Nat. Rev. Drug Discov.* **2003**, *2*, 259–266.

[431] M. Schenone, V. Dančík, B. K. Wagner, P. A. Clemons, *Nat. Chem. Biol.* **2013**, *9*, 232–240.

[432] J. A. Frearson, I. T. Collie, *Drug Discov. Today* **2009**, *14*, 1150–1158.

[433] P. E. Brandish, *J. Biomol. Screening* **2006**, *11*, 481–487.

[434] W. F. An, N. Tolliday, *Mol Biotechnol* **2010**, *45*, 180–186.

[435] J. G. Moffat, J. Rudolph, D. Bailey, *Nature Reviews Drug Discovery* **2014**, *13*, 588–602.

[436] A. E. Carpenter et al., *Genome Biology* **2006**, *7*, R100.

[437] T. R. Jones et al., *PNAS* **2009**, *106*, 1826–1831.

[438] M. L. Green, I. Takeuchi, J. R. Hattrick-Simpers, *J. Appl. Phys.* **2013**, *113*, 231101.

[439] X. .-.-D. Xiang, X. Sun, G. Briceno, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, P. G. Schultz, *Science* **1995**, *268*, 1738–1740.

[440] O. Senkov, J. Miller, D. Miracle, C. Woodward, *Nat. Commun.* **2015**, *6*, 1–10.

[441] S. M. Senkan, *Nature* **1998**, *394*, 350–353.

[442] S. Senkan, K. Krantz, S. Ozturk, V. Zengin, I. Onal, *Angew. Chem. Int. Ed.* **1999**, *38*, 2794–2799.

[443] P. Wollmann et al., *Chem. Commun.* **2011**, *47*, 5151.

[444] J. I. Goldsmith, W. R. Hudson, M. S. Lowry, T. H. Anderson, S. Bernhard, *J. Am. Chem. Soc.* **2005**, *127*, 7502–7510.

[445] S. Bergh, S. Guan, A. Hagemeyer, C. Lugmair, H. Turner, A. F. Volpe, W. Weinberg, G. Mott, *Appl. Catal. A* **2003**, *254*, 67–76.

[446] R. A. Potyrailo, B. J. Chisholm, W. G. Morris, J. N. Cawse, W. P. Flanagan, L. Hassib, C. A. Molaison, K. Ezbiansky, G. Medford, H. Reitz, *J. Comb. Chem.* **2003**, *5*, 472–478.

[447] A. Akinc, D. M. Lynn, D. G. Anderson, R. Langer, *J. Am. Chem. Soc.* **2003**, *125*, 5316–5323.

[448] P. Tsai, K. M. Flores, *Acta Mater.* **2016**, *120*, 426–434.

[449] A. Holzwarth, H. W. Schmidt, W. F. Maier, *Angew. Chem. Int. Ed.* **1998**, *37*, 2644–2647.

[450] C. M. Snively, G. Oskarsdottir, J. Lauterbach, *Angew. Chem. Int. Ed.* **2001**, *40*, 3028–3030.

[451] J. Caruthers, J. Lauterbach, K. Thomson, V. Venkatasubramanian, C. Snively, A. Bhan, S. Katare, G. Oskarsdottir, *Journal of Catalysis* **2003**, *216*, 98–109.

[452] R. A. Potyrailo, I. Takeuchi, *Meas. Sci. Technol.* **2005**, *16*.

[453] S. Sun et al., *Joule* **2019**, *3*, 1437–1451.

[454] M. A. R. Meier, U. S. Schubert, *J. Mater. Chem.* **2004**, *14*, 3289.

[455] J. Zhao, *Prog. Mater Sci.* **2006**, *51*, 557–631.

[456] K. Rajan, *Annu. Rev. Mater. Res.* **2008**, *38*, 299–322.

[457] A. L. Hook, D. G. Anderson, R. Langer, P. Williams, M. C. Davies, M. R. Alexander, *Biomaterials* **2010**, *31*, 187–198.

[458] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, H. Lam, *ACS Comb. Sci.* **2011**, *13*, 579–633.

[459] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, *Nature Chem.* **2009**, *1*, 37–46.

[460] W. Setyawan, S. Curtarolo, *Comput. Mater. Sci.* **2010**, *49*, 299–312.

[461] E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, A. Aspuru-Guzik, *Annu. Rev. Mater. Res.* **2015**, *45*, 195–216.

[462] A. Jain et al., *APL Materials* **2013**, *1*, 011002.

[463] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314–319.

[464] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nature Mater.* **2013**, *12*, 191–201.

[465] H. K. D. H. Bhadeshia, R. C. Dimitriu, S. Forsik, J. H. Pak, J. H. Ryu, *Mater. Sci. Technol.* **2009**, *25*, 504–510.

[466] C. D. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Panté, R. E. W. Hancock, A. Cherkasov, *J. Med. Chem.* **2009**, *52*, 2006–2015.

[467] T. Le, V. C. Epa, F. R. Burden, D. A. Winkler, *Chem. Rev.* **2012**, *112*, 2889–2919.

[468] K. Rajan, *Annu. Rev. Mater. Res.* **2015**, *45*, 153–169.

[469] L. Zhang, Z. Chen, J. Su, J. Li, *Renewable and Sustainable Energy Reviews* **2019**, *107*, 554–567.

[470] G. H. Gu, J. Noh, I. Kim, Y. Jung, *J. Mater. Chem. A* **2019**, *7*, 17096–17117.

[471] J. G. Freeze, H. R. Kelly, V. S. Batista, *Chem. Rev.* **2019**, *119*, 6595–6612.

[472] B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, C. Corminboeuf, *Chem. Sci.* **2018**, *9*, 7069–7077.

[473] Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5*, 24131–24138.

[474] R. Jinnouchi, R. Asahi, *J Phys Chem Lett* **2017**, *8*, 4279–4283.

[475] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, A. Mar, *Chem. Mater.* **2016**, *28*, 7324–7331.

[476] A. S. Rosen, J. M. Notestein, R. Q. Snurr, *J. Comput. Chem.* **2019**, *40*, 1305–1318.

[477] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.

[478] A. Mansouri Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, J. Brgoch, *J. Am. Chem. Soc.* **2018**, *140*, 9844–9853.

[479] H. Zhang, K. Hippalgaonkar, T. Buonassisi, O. M. Løvvik, E. Sagvolden, D. Ding, *ES Energy Environ.* **2019**, DOI 10.30919/esee8c209.

[480] J. Hachmann et al., *Energy Environ. Sci.* **2014**, *7*, 698–704.

[481] E. O. Pyzer-Knapp, K. Li, A. Aspuru-Guzik, *Adv. Funct. Mater.* **2015**, *25*, 6495–6502.

[482] R. Gómez-Bombarelli et al., *Nature Mater.* **2016**, *15*, 1120–1127.

[483] S. Sun et al., *arXiv:1812.01025 [cond-mat physics:physics]* **2018**.

[484] S. Lu, Q. Zhou, L. Ma, Y. Guo, J. Wang, *Small Methods* **2019**, *0*, 1900360.

[485] M. Zeng, J. N. Kumar, Z. Zeng, R. Savitha, V. R. Chandrasekhar, K. Hippalgaonkar, *arXiv:1811.06231 [cond-mat]* **2018**, arXiv: 1811.06231.

[486] L. Wilbraham, R. S. Sprick, K. E. Jelfs, M. A. Zwijnenburg, *Chem. Sci.* **2019**, *10*, 4973–4984.

[487] Y. J. Colón, D. Fairen-Jimenez, C. E. Wilmer, R. Q. Snurr, *J. Phys. Chem. C* **2014**, *118*, 5383–5389.

[488] A. Pulido et al., *Nature* **2017**, *543*, 657–664.

[489] C. Duan, J. P. Janet, F. Liu, A. Nandy, H. J. Kulik, *J. Chem. Theory Comput.* **2019**, *15*, 2331–2345.

[490] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95–98.

[491] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.* **2018**, *9*, 2639–2646.

[492] C. E. Wilmer, M. Leaf, C. Y. Lee, O. K. Farha, B. G. Hauser, J. T. Hupp, R. Q. Snurr, *Nature Chem.* **2011**, *4*, 83–89.

[493] M. Sumita, X. Yang, S. Ishihara, R. Tamura, K. Tsuda, *ACS Cent. Sci.* **2018**, *4*, 1126–1133.

[494] G. Sliwoski, S. Kothiwale, J. Meiler, E. W. Lowe, *Pharmacol. Rev.* **2013**, *66*, 334–395.

[495] S. J. Y. Macalino, V. Gosu, S. Hong, S. Choi, *Arch. Pharm. Res.* **2015**, *38*, 1686–1701.

[496] A. Lavecchia, *Drug Discov. Today* **2015**, *20*, 318–331.

[497] B. M. Wingert, C. J. Camacho, *Curr. Opin. Chem. Biol.* **2018**, *44*, 87–92.

[498] J. Panteleev, H. Gao, L. Jia, *Bioorg. Med. Chem. Lett.* **2018**, *28*, 2807–2815.

[499] F. Chevillard, P. Kolb, *J. Chem. Inf. Model.* **2015**, *55*, 1824–1835.

[500] L. Humbeck, S. Weigang, T. Schäfer, P. Mutzel, O. Koch, *ChemMedChem* **2018**, *13*, 532–539.

[501] H. M. Vinkers, M. R. de Jonge, F. F. D. Daeyaert, J. Heeres, L. M. H. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. Van Aken, P. A. J. Janssen, *J. Med. Chem.* **2003**, *46*, 2765–2773.

[502] C. A. Nicolaou, I. A. Watson, H. Hu, J. Wang, *J. Chem. Inf. Model.* **2016**, *56*, 1253–1266.

[503] N. van Hilten, F. Chevillard, P. Kolb, *J. Chem. Inf. Model.* **2019**, *59*, 644–651.

[504] J. Lyu et al., *Nature* **2019**, *566*, 224–229.

[505] P. D. Lyne, *Drug Discov. Today* **2002**, *7*, 1047–1055.

[506] S. Ghosh, A. Nie, J. An, Z. Huang, *Curr. Opin. Chem. Biol.* **2006**, *10*, 194–202.

[507] T. Cheng, Q. Li, Z. Zhou, Y. Wang, S. H. Bryant, *AAPS J* **2012**, *14*, 133–141.

[508] P. Willett, *Drug Discov. Today* **2006**, *11*, 1046–1053.

[509] P. Ripphausen, B. Nisius, J. Bajorath, *Drug Discov. Today* **2011**, *16*, 372–376.

[510] D. S. Goodsell, A. J. Olson, *Proteins* **1990**, *8*, 195–202.

[511] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, *J. Mol. Biol.* **1996**, *261*, 470–489.

[512] G. Jones, P. Willett, R. C. Glen, A. R. Leach, R. Taylor, *J. Mol. Biol.* **1997**, *267*, 727–748.

[513] R. A. Friesner et al., *J. Med. Chem.* **2004**, *47*, 1739–1749.

[514] N. S. Pagadala, K. Syed, J. Tuszynski, *Biophys. Rev.* **2017**, *9*, 91–102.

[515] P. J. Ballester, I. Westwood, N. Laurieri, E. Sim, W. G. Richards, *J. R. Soc. Interface* **2009**, *7*, 335–342.

[516] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, G. Pujadas, *Methods* **2015**, *71*, 58–63.

[517] Y.-C. Lo, S. E. Rensi, W. Torng, R. B. Altman, *Drug Discov. Today* **2018**, *23*, 1538–1546.

[518] L. Friedrich, T. Rodrigues, C. S. Neuhaus, P. Schneider, G. Schneider, *Angew. Chem. Int. Ed.* **2016**, *55*, 6789–6792.

[519] J. Fang, X. Pang, R. Yan, W. Lian, C. Li, Q. Wang, A.-L. Liu, G.-H. Du, *RSC Adv.* **2016**, *6*, 9857–9871.

[520] L. Hoffer et al., *J. Med. Chem.* **2018**, *61*, 5719–5732.

[521] L. Hoffer, M. Saez-Ayala, D. Horvath, A. Varnek, X. Morelli, P. Roche, *J. Chem. Inf. Model.* **2019**, *59*, 1472–1485.

[522] L. Hoffer, J.-P. Renaud, D. Horvath, *J. Chem. Inf. Model.* **2013**, *53*, 836–851.

[523] REAL Compounds - Enamine, `https://enamine.net/library-synthesis/real-compounds` (visited on 07/25/2019).

[524] G. Schneider, U. Fechner, *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.

[525] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268–276.

[526] A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, A. Zhavoronkov, *Oncotarget* **2016**, *8*, 10883–10890.

[527] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, B. Frey, **2016**.

[528] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.

[529] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, S. Bengio, *arXiv preprint arXiv:1511.06349* **2015**.

[530] E. Putin, A. Asadulaev, Q. Vanhaelen, Y. Ivanenkov, A. V. Aladinskaya, A. Aliper, A. Zhavoronkov, *Mol. Pharmaceutics* **2018**, *15*, 4386–4397.

[531] M. H. S. Segler, T. Kogej, C. Tyrchan, M. P. Waller, *ACS Cent. Sci.* **2017**, *4*, 120–131.

[532] W. Yuan et al., *J. Chem. Inf. Model.* **2017**, *57*, 875–882.

[533] Andrej, *Multi-layer Recurrent Neural Networks (LSTM, GRU, RNN) for character-level language models in Torch: karpathy/char-rnn*, **2015**.

[534] E. J. Bjerrum, R. Threlfall, *arXiv:1705.04612 [cs q-bio]* **2017**.

[535] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, *J. Cheminform.* **2017**, *9*, 1–14.

[536] M. Popova, O. Isayev, A. Tropsha, *Sci. Adv.* **2018**, *4*, eaap7885.

[537] B. Sanchez-Lengeling, C. Outeiral, G. L. Guimaraes, A. Aspuru-Guzik, *ChemRxiv* **2017**, 1–18.

[538] A. T. Müller, J. A. Hiss, G. Schneider, *J. Chem. Inf. Model.* **2018**, *58*, 472–479.

[539] F. Grisoni, C. S. Neuhaus, G. Gabernet, A. T. Müller, J. A. Hiss, G. Schneider, *ChemMedChem* **2018**, *13*, 1300–1302.

[540] A. Gupta, A. T. Müller, B. J. H. Huisman, J. A. Fuchs, P. Schneider, G. Schneider, *Mol. Inf.* **2017**, *37*, 1700111.

[541] D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Mol. Inf.* **2018**, *37*, 1700153.

[542] W. Jin, R. Barzilay, T. Jaakkola, *arXiv:1802.04364* **2018**.

[543] E. Putin, A. Asadulaev, Y. Ivanenkov, V. Aladinskiy, B. Sanchez-Lengeling, A. Aspuru-Guzik, A. Zhavoronkov, *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.

[544] D. Polykovskiy, A. Zhebrak, D. Vetrov, Y. Ivanenkov, V. Aladinskiy, P. Mamoshina, M. Bozdaganyan, A. Aliper, A. Zhavoronkov, A. Kadurin, *Mol. Pharmaceutics* **2018**, *15*, 4398–4405.

[545] P. Ertl, R. Lewis, E. Martin, V. Polyakov, **2017**, *07449*.

[546] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, G. Klambauer, *J. Chem. Inf. Model.* **2018**, *58*, 1736–1741.

[547] R. F. Murphy, *Nat. Chem. Biol.* **2011**, *7*, 327–330.

[548] D. Reker, G. Schneider, *Drug Discov. Today* **2015**, *20*, 458–465.

[549] M. K. Warmuth, J. Liao, G. Rätsch, M. Mathieson, S. Putta, C. Lemmen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.

[550] A. W. Naik, J. D. Kangas, C. J. Langmead, R. F. Murphy, *PLoS One* **2013**, *8*, e83996.

[551] Y. Fujiwara, Y. Yamashita, T. Osoda, M. Asogawa, C. Fukushima, M. Asao, H. Shimadzu, K. Nakao, R. Shimizu, *J. Chem. Inf. Model.* **2008**, *48*, 930–940.

[552] J. D. Kangas, A. W. Naik, R. F. Murphy, *BMC Bioinf.* **2014**, *15*, 143.

[553] D. Reker, P. Schneider, G. Schneider, *Chem. Sci.* **2016**, *7*, 3919–3927.

[554] B. Desai et al., *J. Med. Chem.* **2013**, *56*, 3033–3047.

[555] B. Li, S. Rangarajan, *arXiv:1906.10273 [physics]* **2019**.

[556] E. Bilsland et al., *Open Biol.* **2013**, *3*, 120158.

[557] K. Williams et al., *J. R. Soc. Interface* **2015**, *12*, 20141289.

[558] W. Czechtizky et al., *ACS Med. Chem. Lett.* **2013**, *4*, 768–772.

[559] J. E. Hochlowski, P. A. Searle, N. P. Tu, J. Y. Pan, S. G. Spanton, S. W. Djuric, *J. Flow Chem.* **2011**, *1*, 56–61.

[560] A. Baranczak, N. P. Tu, J. Marjanovic, P. A. Searle, A. Vasudevan, S. W. Djuric, *ACS Med. Chem. Lett.* **2017**, *8*, 461–465.

[561] S. Chow, S. Liver, A. Nelson, *Nat. Rev. Chem.* **2018**, *2*, 174–183.

[562] J. Besnard et al., *Nature* **2012**, *492*, 215–220.

[563] N. C. Firth, B. Atrash, N. Brown, J. Blagg, *J. Chem. Inf. Model.* **2015**, *55*, 1169–1180.

[564] V. Venkatasubramanian, K. Chan, J. Caruthers, *Comput. Chem. Eng.*, An International Journal of Computer Applications in Chemical Engineering **1994**, *18*, 833–844.

[565] S. D. Pickett, D. V. S. Green, D. L. Hunt, D. A. Pardoe, I. Hughes, *ACS Med. Chem. Lett.* **2010**, *2*, 28–33.

[566] L. Weber, S. Wallbaum, C. Broger, K. Gubernator, *Angew. Chem. Int. Ed. in English* **1995**, *34*, 2280–2282.

[567] A. Button, D. Merk, J. A. Hiss, G. Schneider, *Nat. Mach. Intell.* **2019**, *1*, 307–315.

[568] F. Dey, A. Caflisch, *J. Chem. Inf. Model.* **2008**, *48*, 679–690.

[569] R. Wang, Y. Gao, L. Lai, *J. Mol. Model.* **2000**, *6*, 498–516.

[570] S. C.-H. Pegg, J. J. Haresco, I. D. Kuntz, *J. Comput.-Aided Mol. Des.* **2001**, *15*, 911–933.

[571] V. J. Gillet, P. Willett, P. J. Fleming, D. V. Green, *J. Mol. Graphics Modell.* **2002**, *20*, 491–498.

[572] R. P. Sheridan, S. K. Kearsley, *J. Chem. Inf. Model.* **1995**, *35*, 310–320.

[573] G. Schneider, M.-L. Lee, M. Stahl, P. Schneider, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 487–494.

[574] D. Douguet, E. Thoreau, G. Grassy, *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449–466.

[575] N. Brown, B. McKay, F. Gilardoni, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087.

[576] S. Kamphausen, N. Höltge, F. Wirsching, C. Morys-Wortmann, D. Riester, R. Goetz, M. Thürk, A. Schwienhorst, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 551–567.

[577] J. H. Jensen, *Chem. Sci.* **2018**, *10*, 3567–3572.

[578] C. A. Nicolaou, N. Brown, *Drug Discov. Today Technol.* **2013**, *10*, e427–e435.

[579] D. E. Clark, D. R. Westhead, *J. Comput-Aided Mol. Des.* **1996**, *10*, 337–358.

[580] L. Terfloth, *Drug Discov. Today* **2001**, *6*, 102–108.

[581] C. A. Nicolaou, N. Brown, C. S. Pattichis, *Curr. Opin. Drug Discov. Devel.* **2007**, *10*, 316–324.

[582] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **2016**, *7*, 1–9.

[583] D. R. Jones, M. Schonlau, W. J. Welch, *J. Global Optim.* **1998**, *13*, 455–492.

[584] A. Solomou, G. Zhao, S. Boluki, J. K. Joy, X. Qian, I. Karaman, R. Arróyave, D. C. Lagoudas, *Mater. Des.* **2018**, *160*, 810–827.

[585] R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, *Adv. Mater.* **2018**, *30*, 1702884.

[586] A. Seko, T. Maekawa, K. Tsuda, I. Tanaka, *Phys. Rev. B* **2014**, *89*, DOI 10.1103/physrevb.89.054303.

[587] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, *Phys. Rev. Lett.* **2015**, *115*, DOI 10.1103/physrevlett.115.205901.

[588] P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, *Sci. Rep.* **2016**, *6*, DOI 10.1038/srep19660.

[589] K. Tran, Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696–703.

[590] K. Gubaev, E. V. Podryabinkin, G. L. Hart, A. V. Shapeev, *Comput. Mater. Sci.* **2019**, *156*, 148–156.

[591] A. W. Thornton et al., *Chem. Mater.* **2017**, *29*, 2844–2854.

[592] J. P. Janet, L. Chan, H. J. Kulik, *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

[593] A. Nandy, C. Duan, J. P. Janet, S. Gugler, H. J. Kulik, *Ind. Eng. Chem. Res.* **2018**, *57*, 13973–13986.

[594] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, J. K. Nørskov, *Phys. Rev. Lett.* **2002**, *88*, 255506.

[595] N. M. O'Boyle, C. M. Campbell, G. R. Hutchison, *J. Phys. Chem. C* **2011**, *115*, 16200–16210.

[596] Y. G. Chung et al., *Sci. Adv.* **2016**, *2*, e1600909.

[597] A. Mannodi-Kanakkithodi, G. Pilania, T. D. Huan, T. Lookman, R. Ramprasad, *Sci. Rep.* **2016**, *6*, DOI 10.1038/srep20952.

[598] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, *Npj Comput. Mater.* **2019**, *5*, 21.

[599] B. P. MacLeod et al., *arXiv:1906.05398 [cond-mat physics:physics]* **2019**.

[600] T. Ching et al., *J. R. Soc. Interface* **2018**, *15*, 20170387.

[601] D. R. Swanson, N. R. Smalheiser, *Artif. Intell.*, Scientific Discovery **1997**, *91*, 183–203.

[602] D. R. Swanson, *Perspect. Biol. Med.* **1988**, *31*, 526–557.

[603] H.-M. Müller, E. E. Kenny, P. W. Sternberg, *PLOS Biol.* **2004**, *2*, e309.

[604] M. Krallinger, A. Valencia, *Genome Biol.* **2005**, *6*, 224.

[605] L. J. Jensen, J. Saric, P. Bork, *Nat. Rev. Genet.* **2006**, *7*, 119–129.

[606] M. Krallinger, A. Valencia, L. Hirschman, *Genome Biol.* **2008**, *9*, S8.

[607] E. W. Sayers et al., *Nucleic Acids Res.* **2019**, *47*, D23–D28.

[608] S. M. Leach, H. Tipney, W. Feng, W. A. Baumgartner, P. Kasliwal, R. P. Schuyler, T. Williams, R. A. Spritz, L. Hunter, *PLoS. Comput. Biol.* **2009**, *5*, e1000215.

[609] L. Bell, R. Chowdhary, J. S. Liu, X. Niu, J. Zhang, *PLoS One* **2011**, *6*, e21474.

[610] M. W. Libbrecht, W. S. Noble, *Nat. Rev. Genet.* **2015**, *16*, 321–332.

[611] G. Eraslan, Ž. Avsec, J. Gagneur, F. J. Theis, *Nat. Rev. Genet.* **2019**, *20*, 389–403.

[612] K. K. Yang, Z. Wu, F. H. Arnold, *Nature Methods* **2019**, *16*, 687.

[613] R. Verma, U. Schwaneberg, D. Roccatano, *Computational and Structural Biotechnology Journal* **2012**, *2*, e201209008.

[614] R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. F. G. Green, C. Qin, A. Zidek, A. Nelson, A. Bridgland, H. Penedones, *Annu. Rev. Biochem.* **2018**, *77*, 363–382.

[615] M. AlQuraishi, *Cell Systems* **2019**, *8*, 292–301.e3.

[616] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, S. G. Oliver, *Nature* **2004**, *427*, 247–252.

[617] R. D. King et al., *Science* **2009**, *324*, 85–89.

[618] R. D. King, V. Schuler Costa, C. Mellingwood, L. N. Soldatova, *IEEE Technol. Soc. Mag.* **2018**, *37*, 40–46.