UC Berkeley
UC Berkeley Previously Published Works

Title
Toward autonomous design and synthesis of novel inorganic materials

Permalink
https://escholarship.org/uc/item/09v773qn

Journal
Materials Horizons, 8(8)

ISSN
2051-6347

Authors
Szymanski, Nathan J
Zeng, Yan
Huo, Haoyan
et al.

Publication Date
2021-08-01

DOI
10.1039/d1mh00495f

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Toward autonomous design and synthesis of novel inorganic materials

Nathan J. Szymanski[1,2], Yan Zeng[2], Haoyan Huo[1,2], Christopher J. Bartel[2*],
Haegyeom Kim[2*], and Gerbrand Ceder[1,2*]

Target: Materials Horizons
Last edited: 2/28/21 (NJS v3)

Abstract

Accelerating the discovery and optimization of novel inorganic materials is paramount to the development of next-generation technologies facilitating an increased reliance on sustainable energy. Despite the rapidly growing number of promising new compounds proposed using computational techniques, experimental realization through synthesis presents a major roadblock in the development process. With the goal of improving upon traditional approaches that require labor-intensive procedures, autonomous experimentation provides an exciting new opportunity to revolutionize materials discovery. Recent efforts have demonstrated that self-driving laboratories can efficiently perform experiments and learn from their outcomes using artificial intelligence, thereby enabling an efficient exploration of design variables to optimize production and processing at an unprecedented rate. However, due to several key limitations discussed in this review, the application of these approaches to solid inorganic compounds remains limited. To assist in the design of future platforms, we outline the current state-of-the-art methods available to automate each experimental component and discuss their integration to form a "closed loop" of synthesis, characterization, interpretation, and decision-making. Furthermore, we provide some perspectives regarding the next steps that can be taken to solve the major challenges associated with automating inorganic materials synthesis.

[1]Department of Materials Science & Engineering, UC Berkeley, Berkeley, CA 94720, USA

[2]Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

*correspondence to cbartel@berkeley.edu, haegyumkim@lbl.gov, and gceder@berkeley.edu

1. Introduction

Historically, great innovations in technologies have been driven by the discovery of novel materials. Current materials development largely relies on three key steps: (i) identification of a new composition and structure of interest, (ii) targeted and scalable synthesis of that compound, and (iii) post-processing of the product to carefully optimize its properties[1]. To accelerate this procedure, it is necessary to not only improve the efficacy of each step, but also to integrate all three into a closed loop so that they can occur in rapid succession and benefit from optimal feedback between them. While the initial step has been assisted by large-scale *ab initio* simulations[2,3], the latter two generally remain difficult and time-consuming owing to the iterative trial-and-error experimental approach required for both synthesis and property optimization. With the goal of overcoming these challenges, a breakthrough may be found in autonomous experimentation enabled by self-driving laboratories, which aim to aid the human researcher with robotic platforms guided by artificial intelligence (AI).

The automation of experiments has long been a topic of interest, with early examples of widespread utilization demonstrated in the pharmaceutical industry[4]. There, high-throughput (HT) chemistry platforms have been developed to accelerate drug discovery using combinatorial sampling of possible molecules and synthesis conditions, which can be performed in an automated and highly parallelized manner to save considerable time and costs[5-7]. More recently, the advent of AI has created a symbiosis between hardware and software, with active learning techniques guiding the exploration of design spaces and leading to increased efficiencies relative to combinatorial techniques[8-10]. This has opened the door to more sophisticated applications ranging from systematic inspection of retrosynthetic routes in small molecule manufacturing[11] to performance optimization in organic photovoltaics[12]. Furthermore, by automating the role of the experimenter as opposed to individual instruments, modern systems are becoming increasingly versatile as the underlying technology improves[13].

In contrast to organic chemistry, the development of autonomous experimentation for inorganic materials remains in its early stages. Given the challenges associated with handling solid powders, the limited availability of methods that can reliably characterize bulk samples with an abundance of impurities, and the lack of a rigorous theoretical framework describing the factors influencing synthesizability, the majority of existing work has demonstrated only partial automation of the experimental process. Within the thin film community, for example, HT automation of synthesis and characterization is routinely carried out to probe the effects of composition and processing conditions on the properties of resulting samples[14-16]. Similar methods have also been used to study bulk powders but are generally more limited with respect to the scope of compounds that can be dealt with[17-19]. Existing workflows are restricted to materials with readily available synthesis recipes, which precludes the discovery of completely novel systems with new and interesting properties. More recently, AI has been incorporated into the automation pipeline to achieve closed-loop optimization of synthetic routes for nanoparticles formed in continuous flow reactors[20] and nanotubes grown via chemical vapor deposition (CVD)[21,22]. While these platforms can be used to maximize the yield and purity of a target phase, they rely on a reasonable initial guess for the choice of precursors and synthesis conditions so that a measurable amount of the product is consistently obtained and used to guide the optimization. In practice, however, there is typically insufficient information available regarding suitable reaction pathways toward novel compounds, and consequently, the majority of synthesis trials fail to produce any amount of the target phase. Therefore, although current capabilities are indeed promising, considerable progress is necessary before a universally applicable platform enabling autonomous, end-to-end synthesis of inorganic materials can be realized.

Herein, we review the progress made toward "closing the loop" of experimental design, execution, and learning via the development of self-driving laboratories with a focus on applications in inorganic materials science. Accordingly, we consider three major aspects that must be automated to reach this goal. First, experimental procedures should be carried out by **modular, robotic platforms** with the capability of synthesizing and characterizing the materials of interest. Second, the data obtained from **characterization** should be interpreted by the machine and converted into simple, physically meaningful quantities providing insight into the experimental outcome. Last but not least, this information should then be passed to an **intelligent decision-making algorithm** that actively learns from previously tabulated data and suggests new experimental parameters to be tested. Successful design and integration of all three aspects is essential to complete the closed-loop workflow illustrated in **Figure 1**.

With the goal of implementing these components altogether to reach complete autonomy in the synthesis of inorganic materials, we offer perspectives regarding challenges and future directions. To this end, we outline promising techniques to automate solid-state synthesis, characterize the resulting samples by using deep learning algorithms to interpret X-ray diffraction (XRD) spectra, and make informed decisions regarding subsequent syntheses. We emphasize that the success of this process will rely on the ability to learn from all experimental outcomes – i.e., not only to discern whether a given synthesis attempt succeeded or failed, but more importantly to hypothesize and verify *why* it succeeded or failed. To formulate this prediction, insight can be gained from current work being conducted in several related areas including theories on synthesizability[23], *in situ* characterization of reaction pathways[24], and an increasing availability of synthesis data[25]. If developments in these areas are successful in enabling a self-driving synthesis laboratory, it would have wide-reaching impacts across the materials science community, providing the opportunity to efficiently generate new compounds at an unprecedented rate while reducing the amount of time and labor spent by the researcher.

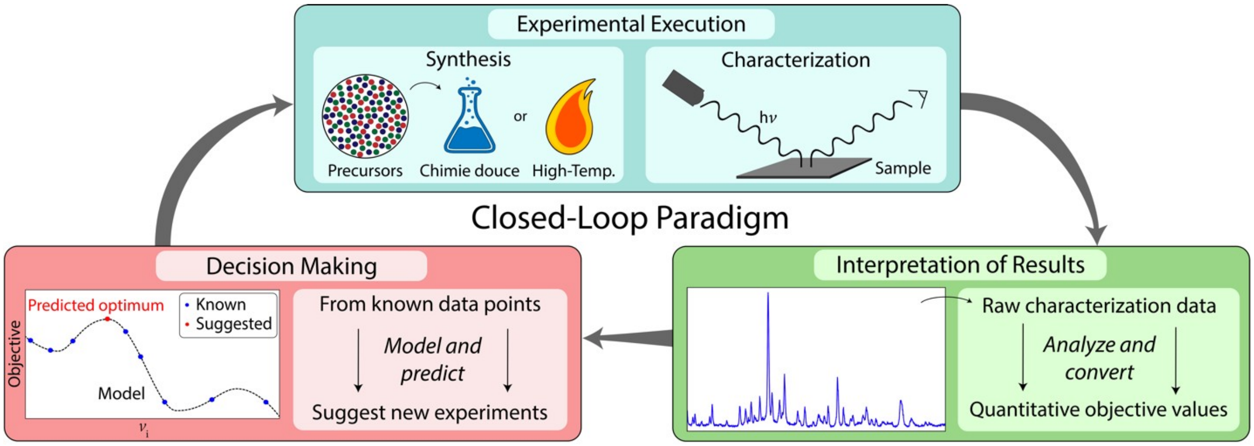


Figure 1: Schematic showing the general workflow of fully autonomous experimentation for the discovery and development of novel inorganic materials.

2. Synthesis & characterization

In the automated optimization of materials properties and processes, synthesizing samples is the first major step. We note that our initial discussion presented here is restricted to the hardware requirements necessary to carry out a synthesis procedure with a given set of parameters including the choice of precursors and conditions – algorithms designed to suggest these parameters will be reviewed in **Section 4**. After the samples have been prepared, appropriate characterization techniques should be employed to reveal the properties of interest and provide critical information regarding the experimental outcome. The execution of synthesis and characterization can be accomplished using robotic systems coupled with real-time and online monitoring to ensure high precision and the handling of any operational issues. The ease of automation, however, varies depending on the synthesis method and the form of the products. We therefore divide our discussion into three major categories: batch or continuous solution-based synthesis, thin film deposition, and solid-state synthesis of bulk powders. Moreover, while the focus of this review is placed on inorganic materials, we will often highlight the success of related platforms in organic chemistry, where automation is more common, to learn from their success and understand how similar methods can be extended to inorganic compounds.

2.1 Solution-based synthesis

The batch solution-based approach, whereby reagents are sequentially combined in appropriate solvents and subject to a series of carefully chosen experimental conditions to produce a desired target, is often the method chosen by organic chemists when synthesizing small molecules[26]. To efficiently explore the factors influencing the yields and properties of the resulting products, automated platforms enabling HT screening of samples have been widely adopted. As detailed in previous reviews[4,7,27], industrial drug discovery systems can routinely conduct thousands of experiments each day. For more complex molecules, however, subtle multi-step reaction sequences are required. To automate the step-by-step addition of reagents, a modular robotic system known as the “Chemputer” was developed by Steiner *et al.*[11] The backbone of their setup, shown in **Figure 2**, contains a series of syringe pumps and six-way selection valves used to transfer reagents between different components of the platform. Four modules are implemented to handle each aspect required for synthesis and characterization, including the main reactor, liquid-liquid separator, filtration apparatus, and rotary evaporator. At the end of the line, chromatography is employed to identify and quantify the resulting products. To verify the effectiveness of the platform, the authors employed their system to automatically synthesize three common drugs over the course of several days. For all procedures, the desired product was successfully obtained with a yield comparable to that generated by human chemists using a standard synthetic procedure.

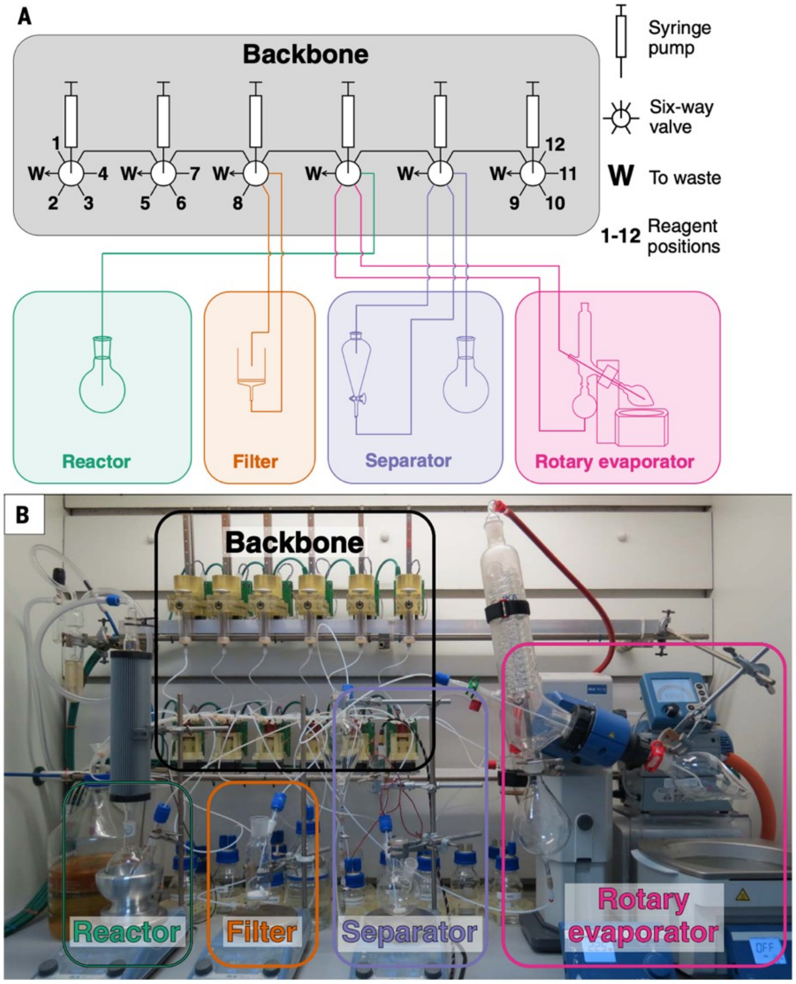


Figure 2: A (a) schematic and (b) photograph of the Chemputer, an automated platform enabling the synthesis of pharmaceutical compounds. The setup is comprised of four modules including the reactor, filter, separator, and rotary evaporation, all of which are connected through a series of syringe pumps and six-wave valves. Reproduced with permission[11]. Copyright 2019, AAAS.

To increase flexibility, Burger *et al.* have produced a mobile robot capable of replicating the actions performed by the traditional chemist – e.g., dispensing reagents, handling vials, and executing operations on lab hardware[13]. Because this platform focuses on automating the role of the researcher while allowing all other aspects of the lab to be interchangeable, it can in principle be applied to diverse sets of experiments by simply swapping out labware and re-programming the robot accordingly. This was demonstrated in the study of aqueous photocatalysts for hydrogen evolution, for which the robotic chemist successfully conducted 688 experiments over the course of eight days. The authors estimate that a comparable number of experiments would have taken a human researcher several months to complete, thus highlighting the benefits of automation.

Flow chemistry represents an alternative synthesis approach that is more widely implemented for large-scale manufacturing of organic compounds[28]. Continuous flow reactors pump reagents through a series of interconnected vessels, with reaction stoichiometries set by the reagent flow rates and conditions controlled using in-line modules. Rapid flow rates and excellent mixing ensure efficient production of target compounds. Moreover, because the systems can be pressurized, higher temperatures can be accessed to enable faster reaction rates. With respect to automation, flow reactors are readily adopted given their integrated design, which eliminates the need to manually transfer samples between different stations as is commonly required in batch chemistry[29,30]. For example, Bédard *et al.* built a reconfigurable system to autonomously optimize a variety of chemical transformations in a flow reactor[31]. An alternating series of tubing and reaction bays shown in **Figure 3** allow reagents to be added sequentially so that multi-step syntheses can be performed. To improve the versatility of the flow reactor, a “plug-and-play” approach is employed whereby six different modules can be interchanged to provide unique capabilities such as heating, cooling, and catalysis. Similarly, many different characterization techniques including high-performance liquid chromatography (HPLC), mass spectrometry (MS), and optical spectroscopy can be implemented for the analysis of reaction outcomes. Applying the platform to three synthetic procedures involving common pharmaceuticals, the authors investigated optimal reaction conditions across hundreds of experiments spanning a cumulative timespan of less than two days.

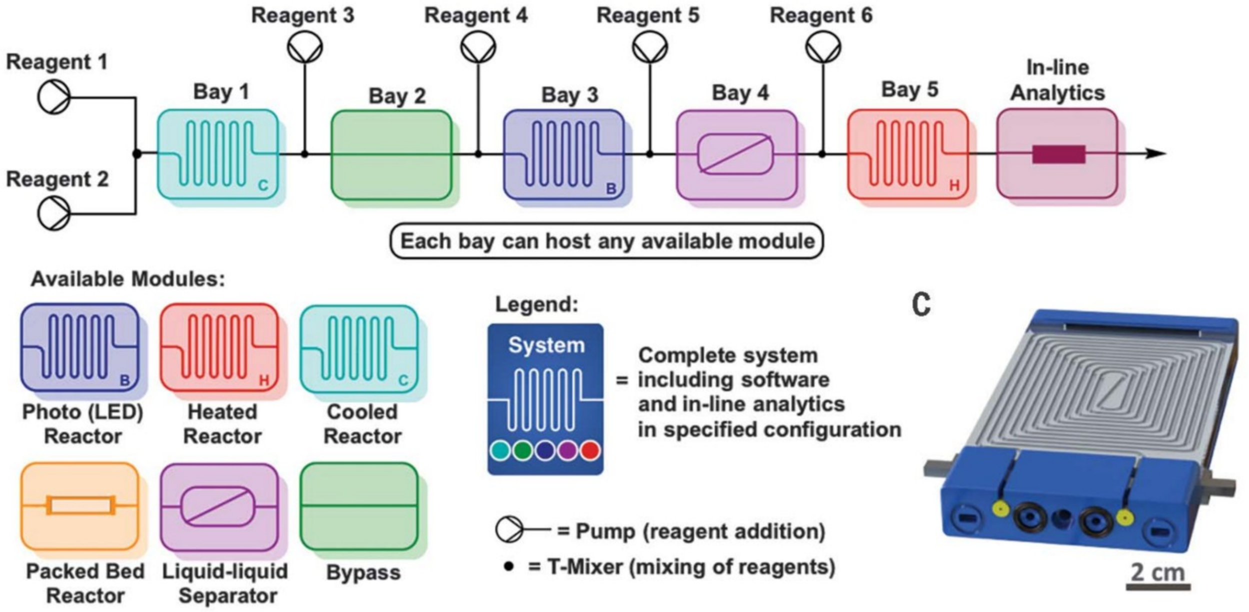


Figure 3: A schematic showing the alternating series of reaction bays implemented sequential throughout a continuous flow reactor designed to optimize organic reactions. A "plug-and-play" approach is used to swap out the modules listed to replicate varied reaction conditions such as heating, cooling, or photocatalysis. Reproduced with permission[31]. Copyright 2018, AAAS.

The promising developments for automated solution-based syntheses of organic molecules have led to several recent efforts that demonstrate the automation of inorganic materials syntheses using similar methods. To rapidly produce lithium and sodium metal oxides sampling varied compositions, for example, a batch solution-based approach was partially automated using the robotic system depicted in **Figure 4**[18,19]. This platform uses an electronic pipetting tool to transfer stock solutions of precursors into microplates, which are heated to mediate reactions between the starting materials. Depending on the choice of precursors and temperature, both co-precipitation and sol-gel routes can be tested at rates of hundreds of samples per day, with the characterization of each conducted by XRD. Comparable techniques have also been applied to automate the synthesis of metal halide perovskites from solution using inverse crystallization at high temperatures[32]. In existing workflows based on batch synthesis, however, manual intervention is required to transfer, dispose of, and replace sample containers between experimental iterations. To automate these processes and fully close the loop, future work may consider integrating a programmable robotic arm as demonstrated by the mobile robotic chemist[13].

In contrast to batch synthesis, flow chemistry is one of the few methods that has been proven successful in making inorganic synthesis fully autonomous. This was first shown by Krishnadasan *et al.* in the optimization of reactions producing CdSe nanoparticles[20]. Flow rates of CdO and Se precursors dissolved in organic solvents were controlled by electronic syringe pumps, combined using a Y-shaped reactor, and passed through a heated reaction vessel. Throughout this process, in-line spectrometry with a charge-coupled device (CCD) was employed to monitor product formation. In separate work, Li *et al.* used comparable techniques to automate the discovery of optically active perovskites. In their platform, precursor solutions were prepared by a rotation sampler and injected into a pipeline of temperature-controlled microfluidic reactors while *in situ* monitoring was conducted by optical spectroscopy[33]. For characterization of the synthesized materials, a robotic arm transferred samples from the flow reactor to a separate station where circular dichroism was measured using spectrometry. Each autonomous workflow was shown to be capable of performing hundreds of experiments at an accelerated rate relative to that obtained by a human researcher.

Despite these successes, the generality of solution-based synthesis for inorganic materials remains limited given the constraints that are imposed on the choice of precursors and reaction conditions. Both batch and flow syntheses require that the starting materials are soluble in an appropriate solvent, which precludes the use of compounds with low solubilities in available liquids. This limitation has little effect on the scope of suitable organic compounds, most of which have reasonable solubilities, but it is highly restrictive for inorganic materials because many cannot be dissolved in common solvents such as water or ethanol. Furthermore, to avoid evaporation of the liquid solvent, operating temperatures must be kept relatively low during synthesis. Even with the use of pressurization in flow reactors, $200\ ^{\circ}\mathrm{C}$ is an upper bound for most systems, suggesting that neither batch nor flow chemistry can be used for inorganic materials that are synthesizable only at high temperatures. Therefore, although solution-based methods are useful where applicable, they do not provide sufficient coverage of the entire chemical space to be used exclusively for automated inorganic synthesis.

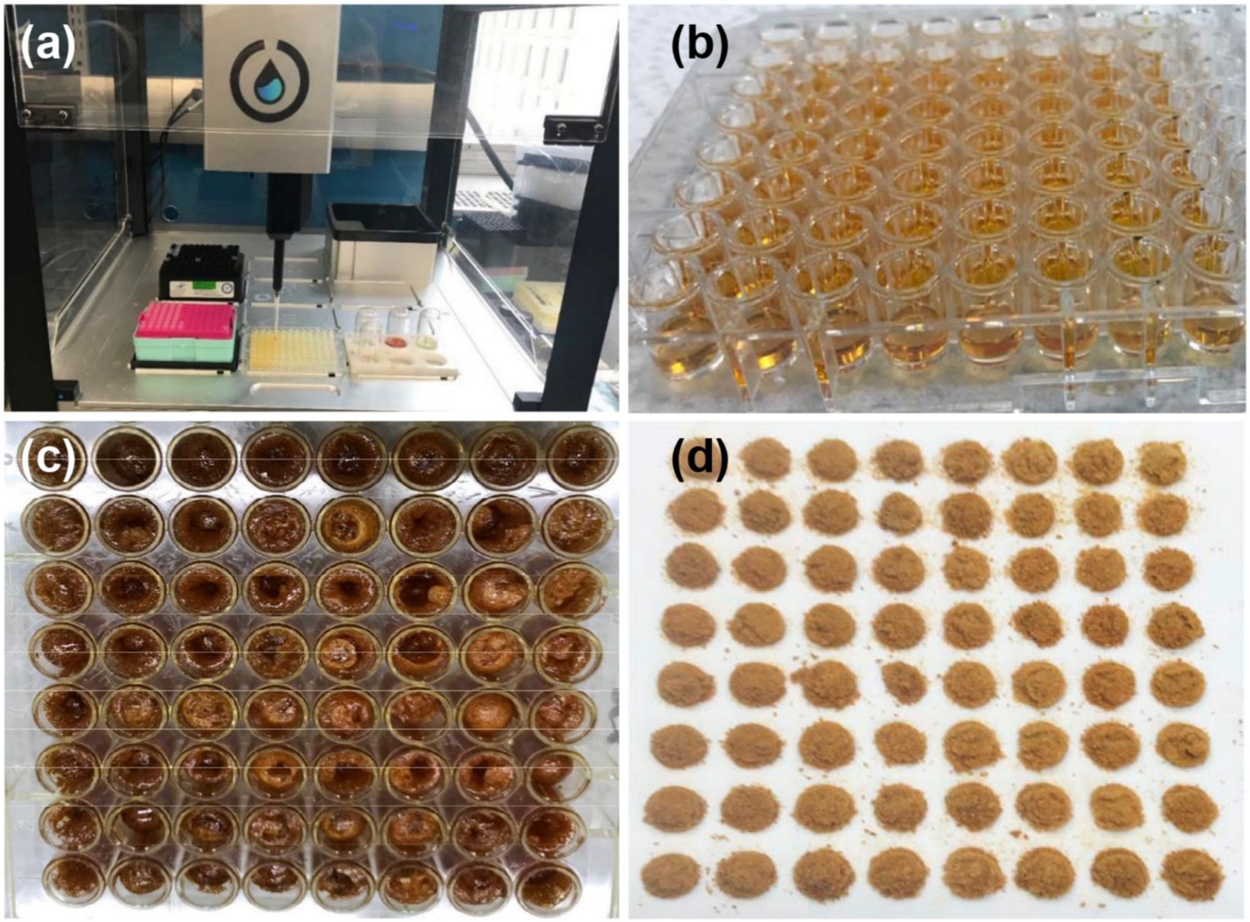


Figure 4: (a) A photograph of the robotic platform that partially automates the HT combinatorial synthesis of powder Na-Fe-Mn-O samples through a sol-gel approach. Panel (b) illustrates the precursors directly after mixing, (c) shows the samples after gel formation, and (d) the crushed product before firing at high temperature. Reproduced with permission[19]. Copyright 2020, American Chemical Society.

2.2 Thin film deposition

Partial automation of synthesis and characterization has become increasingly common throughout the thin film community, where combinatorial methods are employed to study a range of systems such as high-entropy alloys[34] and mixed metal chalcogenides[35,36]. These platforms rely on physical or chemical vapor depositions techniques to synthesize samples spanning a continuous range of compositions, either by sequentially depositing overlapping wedge-shaped layers from individual sources or by simultaneously depositing multiple elements (e.g., by co-sputtering) to achieve a compositional gradient with atomic-scale mixing[14]. The resulting thin film allows the effect of composition to be studied without requiring the synthesis of many individual samples. XRD, optical spectroscopy, or resistivity measurements are typically used to characterize a grid of points across the sample and build a combinatorial library of material properties for the system. By producing a large amount of data for a single targeted sample, automation can be constrained to one experimental cycle of synthesis and characterization, whereas the subsequent processes of analysis and planning of future experiments remain to the researcher's labor and intuition (i.e., the loop is not closed). For a detailed account of existing combinatorial techniques and their applications, we refer the reader to past review articles on the subject[14,37]. Here, we focus on several examples that demonstrate progress made toward closing the loop of synthesis, characterization, and decision-making for thin films.

In one of the first well-known examples of completely autonomous experimentation for inorganic materials science, Nikolaev *et al*. designed a platform to find the synthesis conditions that optimized the growth rate of carbon nanotubes[21]. This process was automated using wafers containing thousands of micron-sized silicon pillars that were coated in a thin layer of catalyst material. Each individual pillar served as a microreactor in a CVD process with ethylene as a source of carbon. By heating pillars one at a time using a highly focused laser and iteratively moving the wafer with a two-axis motion stage, the synthesis of individual samples was precisely controlled. Moreover, the same laser acted as an excitation source for Raman spectroscopy, allowing continuous *in situ* monitoring of growth rates. As illustrated in **Figure 5**, the system was shown to efficiently carry out experiments at a rate of 100 samples per day, a significant improvement over conventional methods[38]. We note, however, that the success of this platform relies on the use of wafers containing many carefully constructed pillars of precursor material, which must be formed in advance and are not necessarily suitable for applications outside of microelectronics.

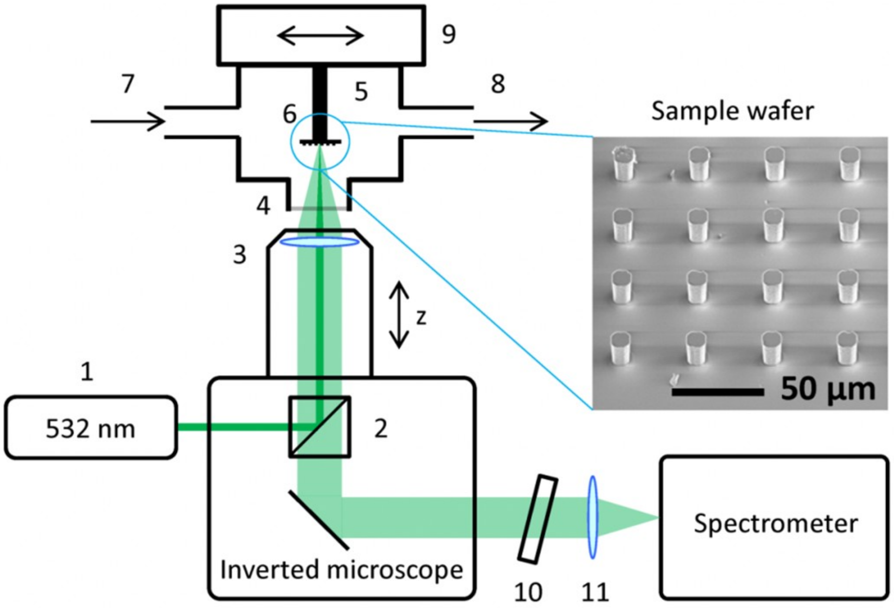


Figure 5: A schematic of the Autonomous Research System (ARES) built to study the formation of carbon nanotubes throughout varied synthesis conditions. A 532 nm laser was employed to heat individual samples of precursors on the wafer while simultaneously providing an excitation source for Raman spectroscopy. Reproduced with permission[21]. Copyright 2016, Nature Publishing Group.

To autonomously synthesize, process, and characterize organic thin films, MacLeod *et al.* introduced a self-driving laboratory named Ada[12]. As illustrated in **Figure 6**, Ada utilizes a robotic arm to transfer vials of fluid between stations on the platform, each of which provides a unique capability including sample storage, solution mixing, spin coating, annealing, and characterization. A combination of four-point probe resistivity measurements and ultraviolet-visible-near-infrared (UV-vis-NIR) spectroscopy were used to study the hole mobility of each sample. Focusing on Spiro-OMeTAD, an organic hole transport material used in photovoltaics, Ada was shown to successfully carry out 35 experiments in less than 30 hours with guidance provided by a Bayesian optimization algorithm (discussed in **Section 4.1.3**) to maximize hole mobility in the samples. Though their platform was constrained to organic films, the methods used can likely be extended to inorganic compounds where solution-based precursors are available for spin coating.

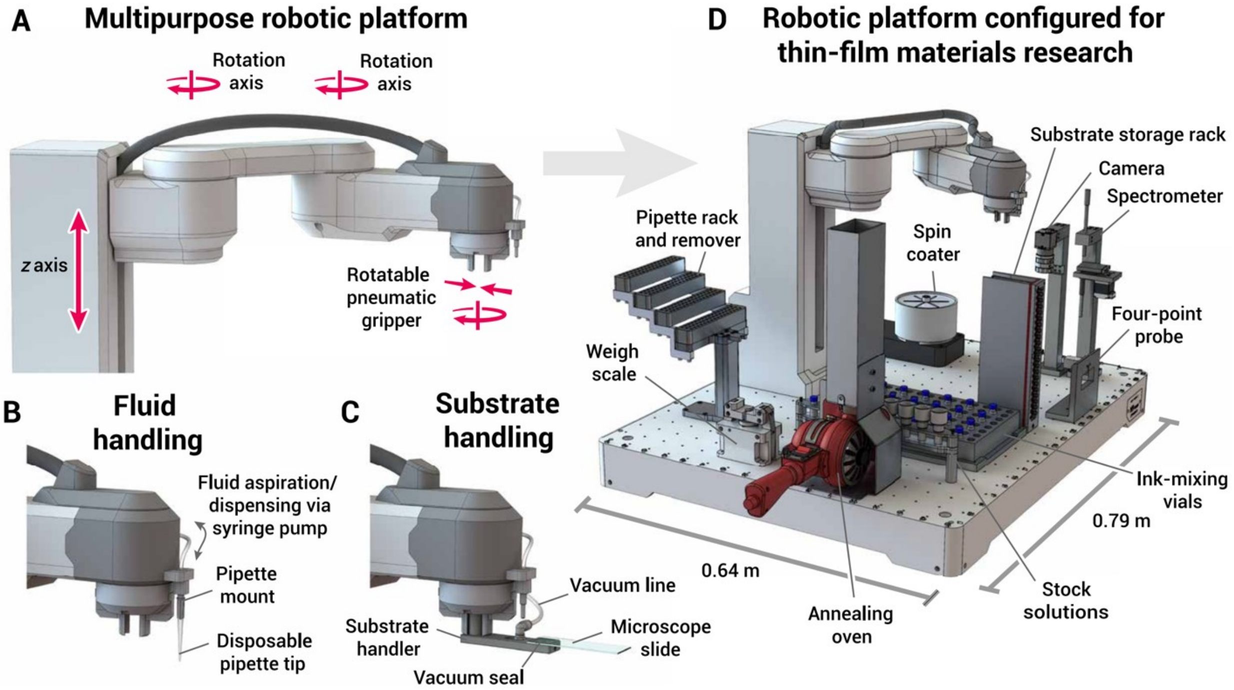


Figure 6: An illustration of the Ada platform used to optimize the optoelectronic properties of organic thin films. Handling and transferring of samples are performed by the robotic arm (a) using a combination of tools shown in (b) and (c), whereas storage, synthesis, and characterization of samples are conducted throughout the individual modules pictured in (d). Reproduced with permission[12]. Copyright 2020, AAAS.

R. Shimizu *et al.* extended autonomous synthesis to inorganic thin films using magnetron sputtering deposition[39]. With this method, TiO_2-based films with a varying concentration of Nb dopants were grown. The partial pressure of oxygen was chosen to be the single experimental variable, which set the reducing conditions during synthesis and thereby influenced the amount of dopant atoms implanted into the films. After each sample was grown, a robotic arm was utilized to transfer it to a separate station for characterization of its electric properties. Aiming to optimize the resistivity with respect to Nb concentration, the films were synthesized and characterized in a closed loop at a rate of twelve samples per day, whereas an equivalent procedure carried out manually is suggested to produce only two new samples per day. Although the observed rate of experiments is not as rapid as the workflows involving CNTs or Spiro-OMeTAD, magnetron sputtering deposition is more readily applicable to a wider range of materials and applications.

More recently, S. Ament *et al.* demonstrated that closed-loop experimentation can be used to synthesize metastable materials that would otherwise be difficult to access by trial-and-error[40]. In their approach, an amorphous layer with composition Bi_2O_3 was deposited onto a silicon substrate via reactive sputtering. To explore the formation of metastable polymorphs of Bi_2O_3 under different synthesis conditions, lateral-gradient laser spike annealing[41] (lg-LSA) was used to rapidly heat and crystallize samples at varied temperatures and dwell times. Within each experimental iteration, optical spectroscopy was applied to measure the reflectivity from a batch of samples. Large changes in reflectivity with respect to the synthesis conditions of the samples were assumed to signify phase transitions (i.e., formation of new Bi_2O_3 polymorphs). Hence, phase boundaries were determined by choosing subsequent experiments that were expected to maximize the gradient of the reflectance. A complete mapping of these boundaries was achieved from 617 samples that were autonomously synthesized, after which XRD was performed *a posteriori* to verify the corresponding phase identities. A key advantage of this approach was the use of lg-LSA, which allowed microscopic regions of the sample to be heated independently of one another. Therefore, many unique temperatures and dwell times could be efficiently tested from a single sample.

Because there are many available deposition techniques, the automation of thin film synthesis can be useful to make a wide range of inorganic materials. In addition to solution-based methods (e.g., spin coating[12]), which require precursors to be soluble in an appropriate solvent, deposition from a gaseous phase expands possible precursors to materials that can be vaporized through heating, sputtering, or irradiation (e.g., for CVD[21]). Techniques such as lg-LSA can also be used to produce reactions or phase transformations in a sample after it has been deposited, which further increases the number of accessible phases[40] However, the versatility of these methods is limited by the applications for which thin films are suitable, which include photovoltaics, protective coatings, and electronic circuitry. To generalize autonomous experimentation for the synthesis of compounds used for technologies such as batteries, catalysts, or functional materials, a solid-state approach must be considered to form bulk powders rather than thin films.

2.3 Solid-state synthesis

Solid-state synthesis carried out by mixing powder precursors and firing at high temperatures is a widely used and scalable approach to produce inorganic materials. Automating this process for HT or closed-loop experimentation, however, remains challenging due to the increased difficulty associated with handling solid powders as opposed to liquids or thin films. Working at high temperatures for long periods of time also poses potential problems caused by the melting of samples and the degradation of containers. Recent efforts have made steps to automate a few key aspects of solid-state synthesis for several classes of materials including PbTe-based thermoelectrics[42], yttrium-doped zirconia[43], and Zr-Ti-C-B ceramics[44]. These existing methods increase the rate at which solid-state syntheses are carried out by decomposing the entire procedure into modular components, each of which is either automated via robotic systems or designed to be conducted in a highly parallelized manner, thereby reducing the time spent by the human researcher per synthesized sample.

A frequently automated task is the dispensing of powder precursors performed by robotic dosing units capable of precisely weighing and dispensing a variety of materials. Mixing is typically carried out using a ball mill, which can be designed to accommodate many samples at once such that parallelization is possible[44]. If mechanochemical synthesis is desired, high-energy ball milling or highly reactive starting materials can be used to encourage the reaction[17]. If, instead, the goal of ball milling is to obtain a well-mixed sample while avoiding any reactions, then relatively inert precursors can be used with low-energy milling[42]. The parallelization of compacting and densification can be achieved by stacking samples on top of one another, separated by an inert material, and loading them altogether into a press. Firing of samples is readily parallelized, limited only by the size of the reaction vessel. However, unless separate furnaces are employed, all materials must be synthesized under the same conditions, which prohibits an efficient exploration of all synthesis parameters simultaneously. Ensuing characterization (e.g., by XRD) is usually conducted serially, though their operation periods are often short in comparison to the time required for synthesis and are therefore unlikely to represent the time-limiting step.

Using HT methods, solid-state syntheses can be performed at a rate of more than 200 reactions per day[17]. We stress, however, that all of the existing methods simply automate individual components of the synthesis process while still requiring a substantial amount of manual intervention between steps to transfer samples. As a result, human time and effort constitutes a large fraction of the total time allocated for the synthesis and characterization and solid powders, and closed-loop automation has not yet been established. This shortcoming is illustrated by the Sankey plot in **Figure 7**, which shows that of the total 328 minutes necessary to complete a full experimental iteration per sample, 105 minutes are consumed by human efforts. Much of this arises from preparative tasks, such as extraction and cleaning, that are difficult to automate. Future work is therefore needed to address these limitations and progress toward full autonomy, as will be discussed further in **Section 5**.

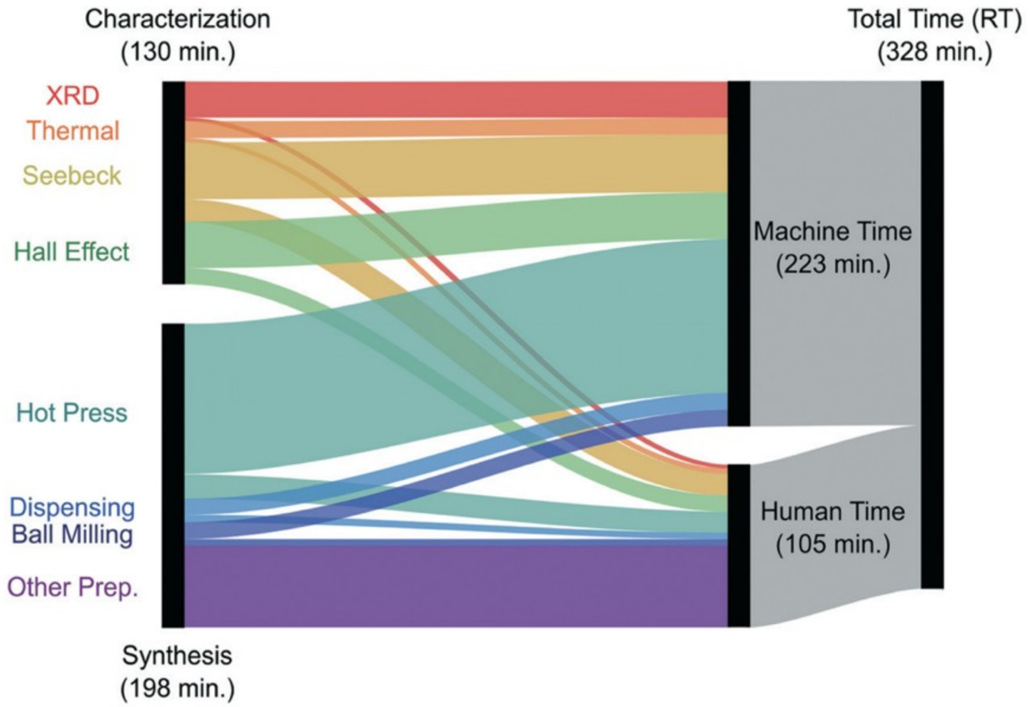


Figure 7: Sankey plot illustrating the time required to complete each component of synthesis and characterization (per sample) throughout the study of thermoelectric materials by Ortiz *et al.* Any effort that must be carried out manually by the human researcher is denoted “human time,” whereas all processes automated by the instrumentation is denoted “machine time.” As the workflow here is only partially automated to enable HT, but not fully autonomous, a substantial portion of the total time required is shown to be allocated to human efforts. Reproduced with permission[42]. Copyright 2019, Royal Society of Chemistry Publishing.

3. Interpretation

After a full cycle of synthesis and characterization is completed, the corresponding results should be exported to software responsible for interpreting the raw data and converting it into quantities that represent the properties of the system under consideration. In some cases, the process of interpretation is straightforward; for example, when measurements yield simple numerical quantities such as electrical resistivity or optical absorbance[12,45]. More generally, however, reliably interpreting characterization data is highly non-trivial, requiring detailed analysis by an expert. Such tasks may involve spectral data obtained from spectroscopic techniques[46], images captured via microscopy[47], or application-based measures of performance[48]. As part of the effort to realize self-driving laboratories in materials science, recent work has demonstrated the potential for machine learning models to analyze and interpret a variety of characterization data[49-51], thereby automating the interpretation component of closed-loop experimentation. The difficulty of this task depends on the number of variables associated with the data being analyzed – i.e., higher dimensionality and an increased amount of information tends to present greater challenges in the automation of interpretation. Accordingly, we break down our discussion into methods used for the analysis of univariate quantities, multivariate data in one dimension (e.g., spectra), and multivariate data in higher dimensions (e.g., images and tomograms).

3.1 Univariate

3.1.1 Surrogate properties

When the property of interest can be represented by a single scalar quantity obtained directly from an automated characterization procedure, interpretation is trivial and warrants no further discussion. It is often the case, however, that univariate properties are difficult to measure in an automated way. To deal with these situations, one may instead choose a “surrogate” property that is more easily measured and has some known relation to the property of interest. For example, MacLeod *et al.* designed Ada to optimize the hole mobility in organic thin films[12]. Considering that automating direct and reliable measurements of the hole mobility is difficult because it requires the construction of multilayer photovoltaic devices, the authors instead used a surrogate property known as the “psuedomobility.” This quantity was defined as the quotient of sheet conductance over absorbance of the thin film, which can be measured with four-point probe measurements and optical spectroscopy respectively. The actual hole mobility of each sample was shown to be directly related to its psuedomobility, and therefore the former can be optimized indirectly using automated characterization techniques.

Another example demonstrating the utility of surrogate properties is given by the determination of phase boundaries from combinatorial thin film libraries[37,52]. A direct approach to detect phase transitions would involve performing XRD measurements and interpreting the resulting spectra to identify the constituent phases. Although characterization by XRD can be carried out in an automated way, the subsequent process of phase identification (interpretation) is challenging given that it requires multivariate analysis (as will be discussed in **Section 3.2**). To avoid these difficulties, it is common to replace XRD by simpler characterization techniques yielding univariate quantities that vary when phase transitions occur. Common choices for these properties include electrical resistivity, optical reflectivity, and mechanical hardness[52], each of which are strongly dependent on the phase present. Accordingly, large changes in these properties are used to indicate a phase transition has occurred.

3.1.2 Reduction to univariate

In some cases, measurements generate data that are initially multivariate but that can be reduced into univariate quantities through dimensionality reduction. For example, optical spectra are commonly simplified by focusing on a single wavelength or by integrating across a range of wavelengths[53]. Similarly, stress-strain and hysteresis curves can be reduced to univariate quantities by deriving properties of interest such as the Young's modulus or saturation magnetization[54,55]. These methods are commonly implemented in combinatorial thin film studies but have also been extended to work with bulk materials in partially automated workflows. During the optimization of shape-memory alloys, for example, heat flow curves obtained from differential scanning calorimetry were simplified to a single value that represented thermal hysteresis of the samples[48]. For battery materials, HT characterization can be carried out by reducing voltage versus capacity curves into univariate quantities such as capacity and energy density[56].

To assist in phase identification, multivariate spectra (e.g., XRD or Raman) can be simplified by focusing only on a subset of peaks associated with a target phase. For example, Nikolaev *et al.* estimated the growth rates of CNTs by measuring the maximum intensities of two known Raman peaks shown in **Figure 8** as a function of time[21]. Similarly, Moosavi *et al.* monitored the phase purity and crystallinity of metal-organic frameworks by measuring the full-widths at half maximum (FWHM) of their XRD peaks[57]. By considering specific and well-defined features in the spectra, these approaches avoid the difficulties associated with automating phase identification from multivariate data. However, these techniques are applicable only if the desired product forms throughout most of the experimental trials, which may not be the case when attempting to synthesize novel compounds. Moreover, it disregards the formation of byproducts or impurities, which can provide useful insights into why a synthesis attempt failed.

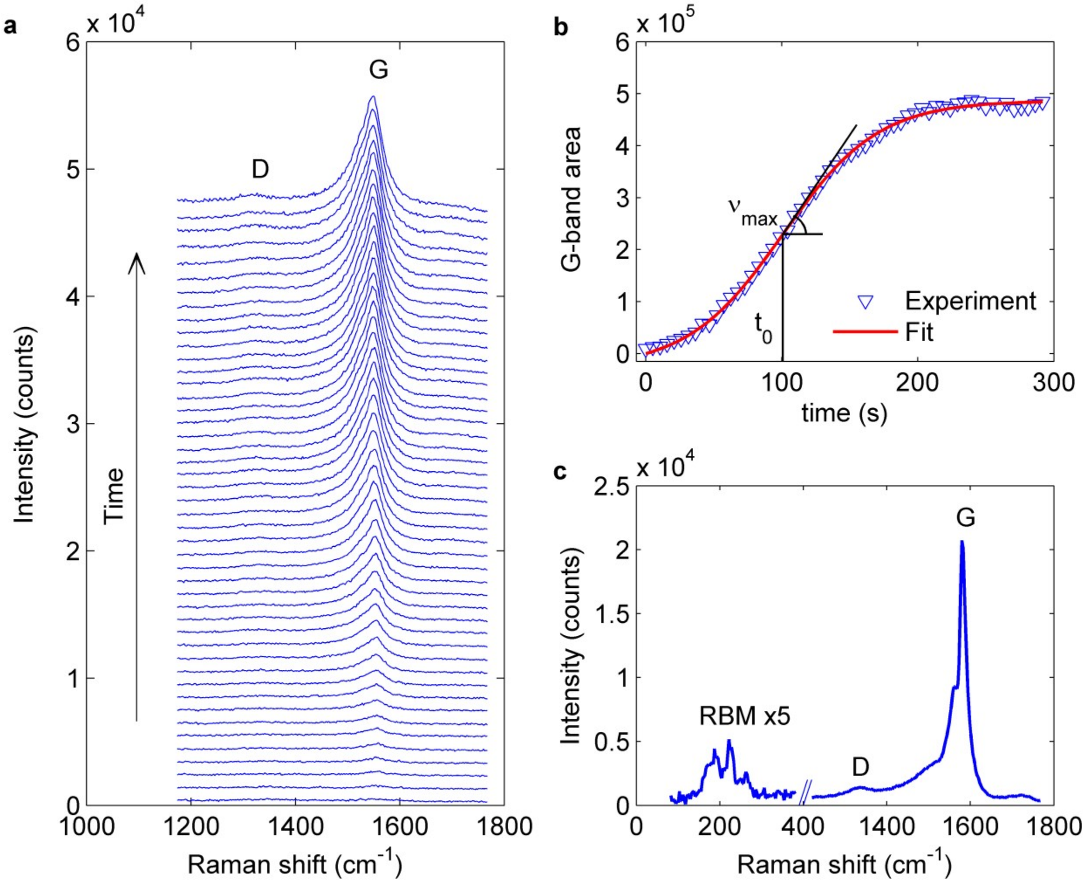


Figure 8: (a) To automate the characterization of carbon nanotubes, the intensity of the G and D bands, from the Raman spectra shown in (c), are continuously measured throughout each experiment. (b) By differentiating the Raman intensity with respect to time, the growth rate of each sample is obtained. Reproduced with permission[21]. Copyright 2016, Nature Publishing Group.

3.2 Multivariate in 1-D

A complete treatment of phase identification requires the analysis of multivariate spectra. For crystalline inorganic materials, this entails the application of XRD and comparing the sample's spectrum with reference data from sources such as the International Centre for Diffraction Data (ICDD)[58]. However, this comparison is complicated by variations that occur between measured and reference patterns due to defects, strain, off-stoichiometry, texture, and poor crystallinity. As a result, interpreting XRD spectra is generally an arduous process that must be carried out by an expert. Even with state-of-the-art tools, reliably automating phase identification for complex, multi-phase spectra remains a longstanding challenge. The available techniques currently used to complete this task are summarized in **Figure 9** and discussed below.

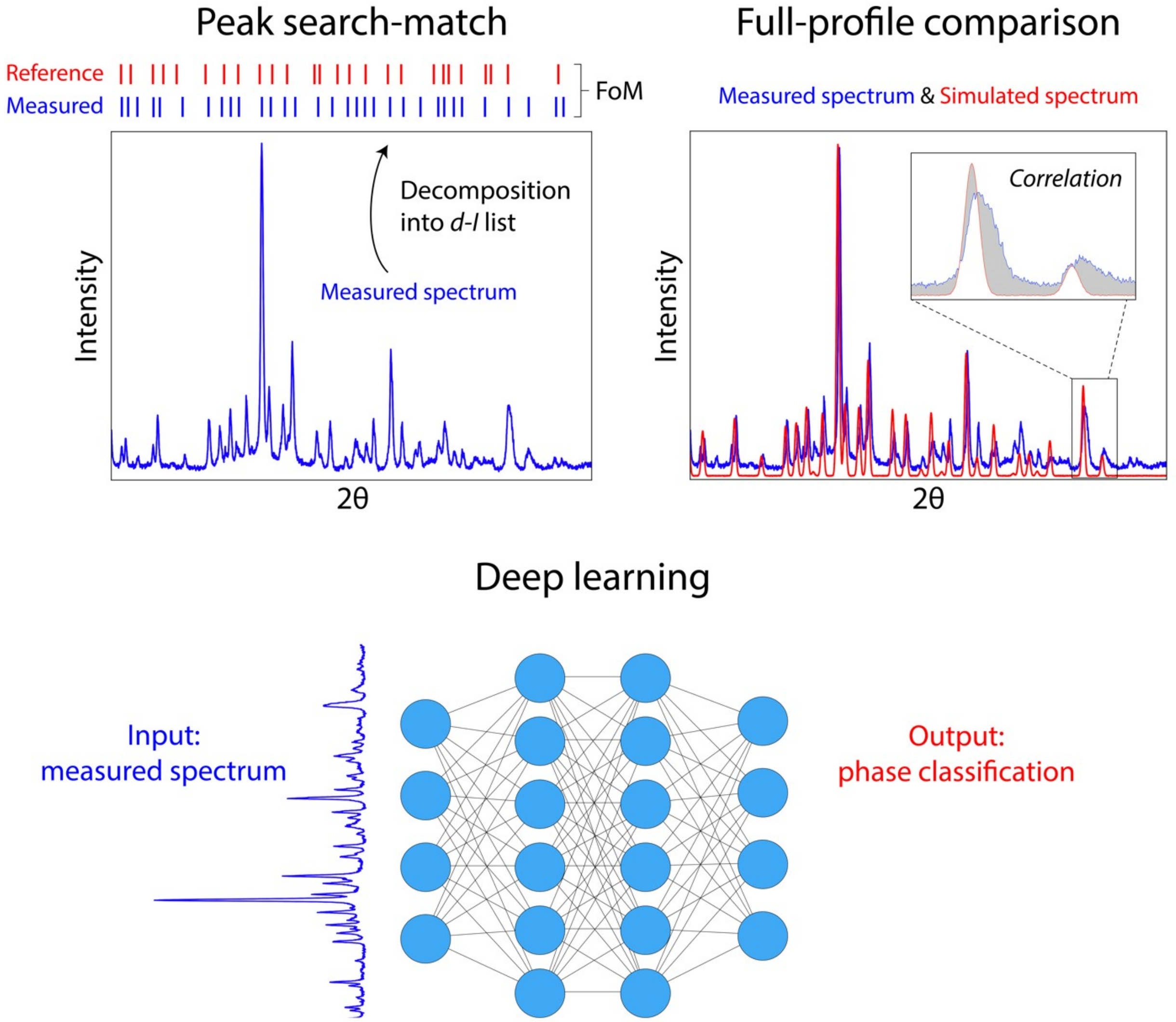


Figure 9: Possible techniques for automating the interpretation of XRD spectra. Peak search-match relies on the identification peaks and comparison with reference data using a Figure of Merit (FoM). Full-profile methods compare entire spectra measured experimentally with reference spectra, usually simulated, using a correlation metric. Deep learning employs neural networks trained on reference spectra to classify measured spectra into constituent phases.

Historically, the analysis of diffraction data has been conducted by decomposing spectra into discrete lists of peak positions (*d*) and intensities (*I*) which are compared with reference data[59]. Peak search-match algorithms rely on a Figure of Merit (FoM) to quantify the degree of similarity between pairs of *d*-*I* lists. A widely used metric is the de Wolff FoM, which is inversely related to the average discrepancy between observed and calculated *d*-spacings[60]. By calculating the FoM for all suspected reference phases, the compound with the highest value may be chosen for a given XRD pattern. However, the reliability of this method hinges on the ability to extract diffraction peaks from the measured spectrum, which becomes difficult in the presence of peak overlap, low peak intensity, or strong background signal[61,62]. These problems are exacerbated when a spectrum contains many peaks (e.g., in low-symmetry structures or multi-phase mixtures), and therefore the peak search-match approach generally produces limited accuracy. In a study conducted by Le Meins *et al.*[63], XRD spectra measured from ten distinct compounds were provided to the broader research community with the task of performing phase identification using peak search-match algorithms. Based on results collected from 25 participants, only 80% of phases were correctly identified on average, even with manual guidance by an expert, thus suggesting the need for improved methods if automation is to be attained.

An alternative to the discrete peak search-match approach is full-profile matching, where *entire spectra* are compared with those of reference phases using a measure of correlation such as cosine similarity, Pearson or Spearman coefficients, or dynamic time warping[64-66]. By removing the need to explicitly deconvolute individual peaks, analyzing the full profile provides a more robust treatment of complex and low-symmetry XRD patterns. Furthermore, this method can be combined with non-negative matrix factorization to identify the combination of compounds that best matches a measured spectrum, enabling classification of multi-phase mixtures[67]. Still, the reliability of full-profile matching remains limited when experimental artifacts cause large changes in peak positions, widths, and intensities. In a study by Iwasaki *et al.*, an accuracy of 70% was achieved with dynamic time warping for the classification of multi-phase mixtures comprised of alloys spanning the Fe-Co-Ni chemical space[68]. Misclassifications were largely attributed to variations in XRD patterns induced by off-stoichiometry of the samples. To improve upon existing methods based on full-profile matching, it is necessary to design an approach that can account for the possibility of experimental artifacts.

Deep learning has more recently been used to automate the interpretation of XRD spectra. In the initial study by Park *et al.*, a convolutional neural network (CNN) was trained to categorize the crystal symmetry of simulated patterns from 150,000 phases in the ICSD[69]. With 20% of spectra reserved for testing, accuracies of 81% and 95% were achieved for the classification of space groups and Bravais lattices respectively. Nevertheless, characterization of experimentally measured spectra is complicated by their differences from simulated patterns arising from various artifacts. In a later work by Vecsei *et al.*, a neural network was trained on simulated XRD patterns to classify symmetry as described previously[70]. The model was then applied to experimentally measured spectra extracted from the RRUFF database, producing a lower accuracy of 54% for space group classification. To resolve these shortcomings, simulated spectra in the training set can be augmented to include perturbations associated with experimental artifacts. For example, Oviedo *et al.* demonstrated that by stochastically varying peak positions and intensities in simulated spectra using for training, the resulting CNN correctly classified the space group for 80% of spectra measured from metal halide thin films[51].

Similar techniques based on deep learning and data augmentation have also been used to perform phase identification from experimentally obtained XRD patterns[71]. For example, Maffettone *et al.* trained an ensemble CNN using simulated spectra augmented with changes to peak widths, intensities, and background signals[72]. Their model was tested on patterns measured from samples in the Ni-Co-Al space, with 76% correctly identified. To handle multi-phase mixtures, Lee *et al.* trained a CNN using multi-phase spectra simulated from linear combinations of single-phase patterns for 170 compounds in the Sr-Li-Al-O space. Their model achieved a high accuracy of 98% when classifying experimentally measured spectra obtained from mixtures of high-purity powders including $SrAl_2O_4$, SrO, Li_2O, and Al_2O_3[73]. However, because the training procedure requires many linear combinations of phases with varied weight fractions to be sampled (1,785,405 in total), it restricts the inclusion of experimental artifacts owing to combinatorial explosion. Therefore, the model may fail when applied to characterize arbitrary samples obtained from a synthesis trial, which often contain substantial perturbations in their

XRD spectra caused by defects. With this limitation in mind, a more reliable approach is needed to characterize complex spectra produced by multi-phase mixtures, as will be discussed further in **Section 5**.

In addition to XRD spectra, deep learning has also been extended to automate the characterization of materials using Raman and Fourier transform infrared (FTIR) spectroscopy. In contrast to XRD, however, simulating Raman and FTIR spectra is more difficult because it involves the calculation of vibrational frequencies through *ab initio* methods. Accordingly, Liu *et al.* instead used experimentally measured spectra from the RRUFF database to train a CNN[49]. Data sparsity was overcome by augmenting the available spectra with stochastic changes in Raman shifts for each vibrational spectrum. The model correctly identified 88% of the chemical species tested, exceeding the accuracies of techniques that use similarity-based metrics[74]. These results show that a reliable classification of vibrational spectra can be automated in situations where suitable experimental data is available; however, this is generally not the case as clean and consistently measured vibrational spectra are difficult to find for most chemical spaces. Moreover, additional effort is necessary to determine whether deep learning can be used for spectra measured from multi-phase mixtures, where peak overlap is likely to be problematic.

3.3 Multivariate in higher dimensions

Imaging techniques have recently found use in partially automated workflows to accomplish tasks related to quality assurance or verification that the target materials were indeed synthesized. For example, images obtained from scanning electron microscopy (SEM) have been used to guide the optimization of fiber quality in polymeric samples by allowing the user to rank the images obtained after each experimental iteration[75]. SEM has also been used to check the quality of graphene[76] and carbon nanotubes[21] after automated growth procedures were carried out. Similarly, optical microscopy was employed to ensure a low defect density in organic thin films synthesized by Ada[12]. In all cases, however, manual analysis of the images was required. If this process could instead be automated, the loop between experimentation and interpretation could be closed.

Deep learning is well-suited to handle high-dimensional multivariate data including images and tomograms. Its application to inorganic materials was first demonstrated by Al-

Khedher in a study on CNTs, where a neural network was used to classify the degree of alignment and curvature of nanotubes imaged by SEM[77]. Training was performed on two datasets, each labeled with values describing their alignment and curvature: one based on idealized "rope" images representing nanotubes in various orientations and another containing experimentally obtained images depicting actual CNTs. By combining these images into a single dataset that was used to train the neural network, a classification accuracy comparable to that of a human researcher was achieved. A similar analysis of SEM images was designed by Modarres *et al.* to categorize the shape and morphology of nanomaterials using a CNN[47]. In their work, a training set was constructed by manually labeling thousands of experimentally measured SEM images based on ten categories including fibers, powders, and nanowires. The resulting model correctly classified 90% of test images that were obtained from experiment.

At higher spatial resolution, deep learning has also been used to automate the analysis of images obtained with scanning transmission electron microscopy (STEM). Several recent efforts have shown that CNNs are capable of mapping out atomic positions and identifying defective regions (e.g., vacancies or dislocations) in atomic-scale images[78-80]. This can be achieved by training the model on large sets of experimental data with the locations and sometimes chemical identities of atoms or defects manually labeled. Resulting accuracies can match or sometimes even exceed that of a human researcher, likely because the task being performed requires large amounts of data to be analyzed – i.e., a human mapping out hundreds of atoms in an image is likely to make occasional mistakes, whereas AI has the advantage of steady consistency.

In three dimensions, the analysis of mixtures imaged by atomic probe tomography (APT) has been automated with deep learning to identify interfacial regions between distinct phases[81]. Madireddy *et al.* employed a transfer learning technique where a CNN was trained on a large number of completely unrelated images with labeled edges such that it could then be applied to detect interfaces in APT images. Interestingly, even though the training set appeared very different from the test set, accurate phase segmentation was obtained by the model when applied to an alloy containing precipitates suspended in a matrix; though, occasional irregularities occurred when compositions smoothly varied and therefore the interfaces between phases were "blurred."

The largest obstacle faced throughout many of the existing deep learning methodologies is the limited availability of training data. The measurement and organization of novel data using methods such as electron microscopy and APT are often time-consuming and expensive. Furthermore, supervised training relies on manual labelling of features in images, which can be tedious and prone to errors – mistakes have been mitigated in past work by relying on more than one researcher to label a single set of images[79]. To avoid these difficulties, training data may be simulated rather than measured. However, as opposed to XRD spectra, simulating realistic images is challenging and sometimes not possible. Therefore, to generalize the application of deep learning for image analysis in materials science, a rapid and more reliable method of tabulating data is needed. We note that an automated platform may be suitable for this purpose, given that large amounts of data can be generated quickly without much human intervention. If sufficient training data cannot be obtained, then unsupervised learning may be employed to reveal similarities and differences throughout large sets of images, a particularly useful approach if data is obtained in HT.

4. Decision making

Upon conducting a set of experiments and interpreting the resulting data, the next step in the closed-loop automation process is to learn from this information and make decisions regarding the subsequent experiments to be performed. These decisions are usually made with the goal of optimizing some quantity[82]; for example, maximizing the yield of a product by modifying its synthetic procedure[83] or tuning the properties of a material with respect to its structure, composition, or processing conditions[12,48]. Alternatively, decisions can be made to formulate experimental tests that reveal information regarding a specific process[84]. In synthesis, for example, this may entail exploring various combinations of reactants and conditions followed by observation of their products to construct a network of possible reaction pathways in a system of interest. Regardless of the underlying motivations, the desired outcome of decision making remains the same: reaching a pre-defined optimum while minimizing time and costs. To this end, there exists a variety of active learning techniques facilitate optimization by iteratively learning from and querying data in the design space[9,85,86]. Before reviewing available active learning algorithms, it is first important to understand why active learning is necessary by considering the alternative methods listed in **Figure 10** and highlighting their shortcomings.

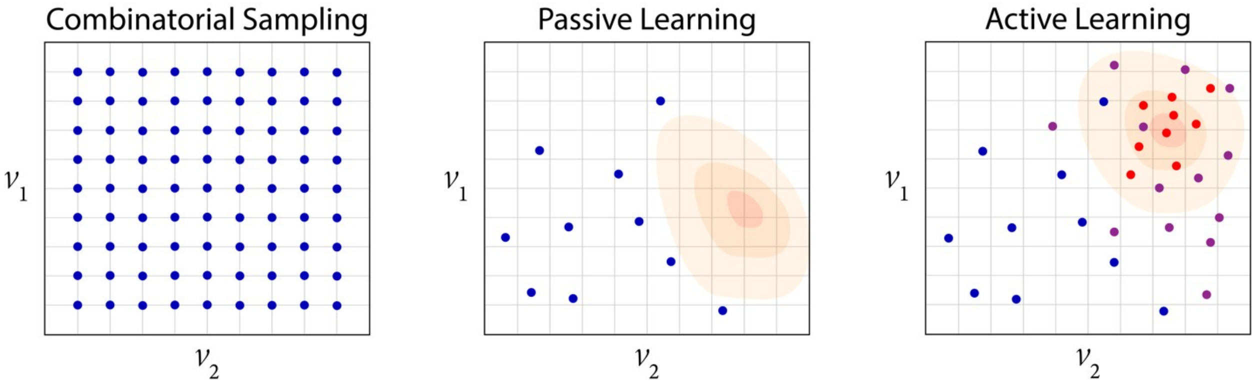


Figure 10: A schematic illustrating three possible optimization techniques. Combinatorial approaches sample many possible combinations of design variables (ν_i), sometimes chosen uniformly across the design space. Passive learning employs existing data points (blue dots) to form a model of the objective and make predictions regarding the location of its optimum (shaded region). Active learning builds upon this approach by suggesting new points at which to evaluate the objective (purple dots), from which the information is used to update the initial predictions and once again suggest new points to be queried (red dots), forming an iterative loop which is traversed until convergence to the true optimum is reached.

A simple and widely used optimization strategy is to perform a brute-force search of the design space, thus avoiding decision making altogether. Such is the concept underlying HT workflows, where a grid of data points is generated from combinatorial sampling of experimental parameters[87,88]. From this dataset, analysis may then be conducted *ex post facto* to identify relationships among variables and estimate any optima in the objective of interest. As the reliability of these conclusions depends on the density of points sampled across the design space, a large number of experiments are typically necessary to obtain satisfactory results. Consequently, successful applications of HT platforms have been limited to problems for which (a) the appropriate experiments are inexpensive, quick, and easily parallelized, or (b) the design space of interest is relatively narrow. A sufficiently dense sampling of compositions on thin films spanning ternary spaces, for example, can typically be achieved using several hundred samples[89,90]. In contrast, generating a grid of equal density for quaternary systems requires several thousand samples. Hence, as the number of necessary experiments scales exponentially with the dimension of the design space, combinatorial techniques quickly become intractable when additional variables are introduced. These problems are sometimes simplified by partitioning the design space and focusing on a much smaller subset of interest[91,92]; however, this solution is not generalizable because the most interesting region of the design space is generally unknown. Therefore, to efficiently explore the entire design space, active learning is required.

Contrary to HT experimentation, existing data can be used to learn trends and predict optima in the objective function without performing any new experiments. As the learner simply observes the environment without interacting with it to query new information, this technique is sometimes called *passive* learning to distinguish it from its *active* counterpart[93]. Enabled by the development of machine learning models, in addition to a growing amount of readily available data[2,3,94], passive learning has found widespread use throughout materials science. For a detailed overview of common machine learning algorithms and their application in materials science, we refer the reader to several recent reviews on the topic[95-97]. Here, we narrow our discussion to focus on two key limitations of passive learning as applied to optimization. First, the accuracy of the model is heavily reliant on both the volume and diversity of training data. In many situations, the design space of interest is sparsely populated. For example, applying machine learning to inorganic synthesis procedures remains difficult because there is a limited amount of accessible, labeled information regarding precursors, catalysts, and experimental conditions used throughout previous work[25,98,99]. Moreover, even in cases where more data is available, it tends to be highly biased toward specific regions of the design space. This bias commonly originates from a strong tendency for researchers to only publish positive results while leaving negative results unreported[100], or because many studies pursue minor modifications of an already successful material/procedure. This limits the diversity of the training set, and thus negatively affects the performance or applicability of the corresponding model. In addition to the limitations imposed by the sparsity of training data, passive learning models are inherently inept at predicting outliers, instead relying on the recognition of general trends in the data. While this capability is sufficient for many studies, it becomes problematic for optimization problems where the global extrema are of interest. To overcome these limitations, it is necessary to acquire new data so that the model can continuously learn and improve its accuracy, thus ensuring correct identification of optima.

As will be discussed below, active learning techniques are gaining traction throughout materials science, with multiple applications recently demonstrated in automated workflows[101-104]. We emphasize that the decision-making process used in active learning follows a natural approach, similar to that of a human expert – performing an appropriate set of tests, using the results to build knowledge of the system, and implementing that knowledge to intuitively select new tests – all in a highly serialized nature as to avoid unnecessary experiments. Considering this paradigm, we review two major categories of active learning as applied to decision making in optimization: *black-box* and *informed*.

4.1 Black-box optimization

In situations where little is known about the system at hand, the corresponding objective function can be treated as a "black box," i.e., an opaque function that must be queried at individual points through experimentation or simulation. Performing black-box optimization, a topic that has been studied extensively and applied throughout many areas of science and technology, requires the consideration of two key constraints[105]. First, as no analytical form of the objective function is available, optimization must be carried out without the use of exact derivatives. Second, the objective landscape may generally be non-convex, requiring global instead of local optimization. These properties exclude the application of explicit gradient-based and pure local search methods respectively. Moreover, extending black-box optimization specifically to experimentation presents an additional challenge: evaluation of the objective is usually expensive and time-consuming, suggesting the importance of reaching convergence in a minimal number of steps[8]. This further excludes algorithms that approximate derivatives via finite differences owing to their inefficiency with respect to the number of evaluations required. Instead, a variety of efficient and derivative-free techniques have been developed to perform global optimization on black-box functions where data acquisition is costly[106,107]. To shed light on which of these approaches are most suitable for accelerating inorganic materials discovery, we review those that have been implemented in experimental workflows and outline the major advantages and limitations of each.

4.1.1 Genetic algorithms

Many of the earliest methods used to replace brute-force optimization were based on genetic algorithms (GAs)[87]. In this approach, an initial batch (or generation) of experiments is conducted to evaluate the objective function(s). From the results, a new batch of experiments is suggested based on three processes: (i) *selection* dictates which samples are chosen to contribute to the next generation of experiments, (ii) *crossover* determines how the properties of selected samples are merged to suggest new experiments, and (iii) *mutation* applies random variations to the properties of suggested experiments. Each process is controlled by a set of hyperparameters (e.g., the rate at which mutation is applied) defined by the user. From the corresponding modifications, new generations of experiments are iteratively produced until convergence of the objective function(s) is reached. Because GAs impose a bias toward promising regions of the design space by selectively sampling experiments where the objective function is expected to be optimal, they generally provide increased efficiency relative to combinatorial techniques. GAs are also well-suited to handle a large number of variables, both qualitative and quantitative, and can perform well in multi-modal design spaces assuming that a sufficiently high mutation rate is used to escape local optima[108,109].

For experimentation in materials science and chemistry, GAs have found widespread application in the discovery of novel catalysts[110]. In such problems, where the aim is to maximize the yield of a desired product phase with respect to the choice of catalytic materials, traditional HT methods commonly become intractable because realistic systems have highly complex design spaces – industrial catalysts often contain as many as ten elements and can be further complicated by the choice of a support material[111,112]. Addressing this challenge, work in the early 2000s pioneered the implementation of GAs on batched heterogeneous catalysis experiments. Wolf *et al.* demonstrated a proof of concept by optimizing the composition of mixed metal oxide catalysts containing eight unique components for the oxidative dehydrogenation of propane[113]. The GA achieved significantly increased propene yield within only four generations, corresponding to a total of 224 experiments. This number represents an improvement over the previously used combinatorial methods, for which thousands of experiments were required in comparable systems[114]. However, the authors emphasize that the GA performs well only after its underlying hyperparameters were tuned through a series of tests conducted using a "pseudo-dataset" generated by heuristic relationships that approximate the effect of catalyst composition on the reaction yield. Alternatively, to improve upon the accuracy of heuristic relationships, machine learning models (e.g., neural networks) have been used to construct pseudo-datasets that assist in choosing the best-performing hyperparameters for a GA[111,115,116]. This method is applicable when a sufficient amount of experimental data is available to train the model on, and therefore may not be used in novel chemical spaces.

Outside of catalysis, GAs have also been applied to handle synthesis procedures where the concentrations of precursors and conditions are varied to optimize the yield and form of a target phase. Moosavi *et al.* designed a robotic platform guided by a GA to carry out experiments with the aim of maximizing crystallinity and phase purity in metal-organic frameworks[57]. Based on microwave-assisted synthesis, the search space consisted of nine parameters involving reactant ratios, solvent compositions, microwave power, environmental temperature, and reaction time. Exploration of these variables was conducted through three generations consisting of thirty experiments each. Between each generation, a random forest model was trained on data obtained from all previous experiments. The model was used to predict the results of suggested experiments and excluded any that were expected to yield unfavorable results (i.e., poor crystallinity or phase purity). The authors propose that this method improves the efficiency of the GA, which itself only considers results from the previous generation. Indeed, **Figure 11** shows that a large fraction of the population has converged to a narrow region of the design space by the third generation, with the corresponding samples having high crystallinity and phase purity. A similar approach was used by Nikolaev *et al.* to optimize the growth rate of carbon nanotubes with respect to the system's temperature, pressure, and partial pressures of ethylene, hydrogen, and water vapor[21]. Ten generations containing 84 samples each were produced under the guidance of a GA. A random forest model was trained between each generation and used to bias suggested experiments toward high expected growth rates. By the final batch of experiments, the samples were grown at a rate nearly 100× more rapid than in the initial experiments.

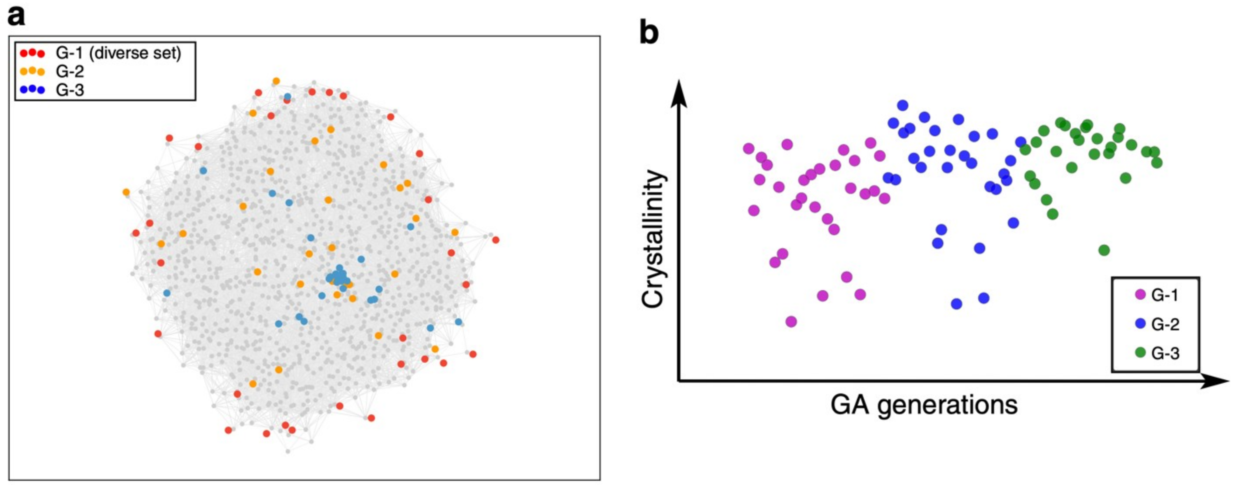


Figure 11: Results obtained during the optimization of synthesis conditions for metal-organic frameworks, guided by a genetic algorithm. (a) A dimension reduction of trials conducted in the 9-dimensional parameter space onto a 2-dimensional plane, showing the convergence of experiments to an optimal subspace. (b) Evolution of sample crystallinity throughout three generations. Reproduced with permission[57]. Copyright 2019, Nature Publishing Group.

Although GAs are capable of reaching convergence within relatively few generations, each population must contain a large number of individual experiments to achieve genetic diversity and ensure reliable performance. Depending on the complexity of the design space and the associated objective function, suitable population sizes may range from tens to hundreds of samples[117]. As a result, GAs excel when applied to problems for which parallelization of many experiments can be attained while keeping the associated time and costs reasonably low. In more general situations, however, experiments tend to be resource-intensive and large-scale parallelization can be difficult or impractical. Such cases therefore necessitate optimization algorithms that reach convergence in a minimal number of *total* experiments conducted either serially or in small batches.

4.1.2 Stable noisy optimization by branch and fit

The SNOBFIT (Stable Noisy Optimization by Branch and FIT) algorithm[118] combines aspects from local and global search strategies to efficiently optimize an objective function. From a given dataset, SNOBFIT employs a branching algorithm to partition the design space into unique sub-regions, each containing a single known datapoint. Within each region, a local model of the objective function is constructed via least-squares quadratic fitting of the contained datapoint and its nearest neighbors. These models represent the objective function *locally* but do not necessarily describe it *globally* – i.e., each quadratic fitting is performed independently using a subset of known datapoints. The resulting models are then used to predict and suggest sampling of new datapoints in regions where the objective function is expected to be optimal. At the same time, sampling is also suggested in sparsely populated regions to ensure the global optimum is not missed.

For automated experimentation, perhaps the most successful application of SNOBFIT is in the optimization of chemical reactions in continuous flow reactors[119,120]. This was first demonstrated by Krishnadasan *et al.*, who maximized the fluorescence of CdSe nanoparticles with respect to precursor flow rates in a microfluidic reactor[20]. SNOBFIT was used to minimize a "dissatisfaction coefficient" (DC) related to the difference between observed and targeted emission intensities produced by the samples. As shown in **Figure 12**, a minimum in the DC was found after performing 71 experiments, with the final samples showing a near four-fold increase in emission intensity relative to those initially measured. In a similar effort, Li *et al.* explored temperature-composition space to optimize circular dichroism (CD) in inorganic perovskite nanocrystals[33]. Experiments suggested by SNOBFIT quickly identified a local optimum in less than 50 trials, but then continued to converge toward an improved solution exhibiting a CD intensity twice that of the initial optimum. These results highlight the ability of SNOBFIT to escape local optima, which is enabled by sampling sparsely populated regions of the design space as opposed to relying only on greedy optimization.

Increasing the number of variables under consideration, Bédard *et al.* designed a reconfigurable flow reactor capable of probing multistep organic reaction sequences and maximizing the corresponding yield through optimization with SNOBFIT[31]. To demonstrate this capability, six different series of experiments were carried out for a variety of processes including amination, olefination, cross-coupling, and substitution. In all cases, high product yield was obtained in less than 50 experiments by varying temperature, flow rates, and catalyst compositions. However, the authors note that reasonably tight bounds were placed on the design variables using information obtained in previous work, thereby reducing the problem's complexity. A comparable performance of SNOBFIT was illustrated in the work of Cherkasov *et al.*, where the algorithm successfully identified an optimum in 61 trials to minimize a composite objective function incorporating product yield and substrate flow rate with respect to three synthesis parameters[121].

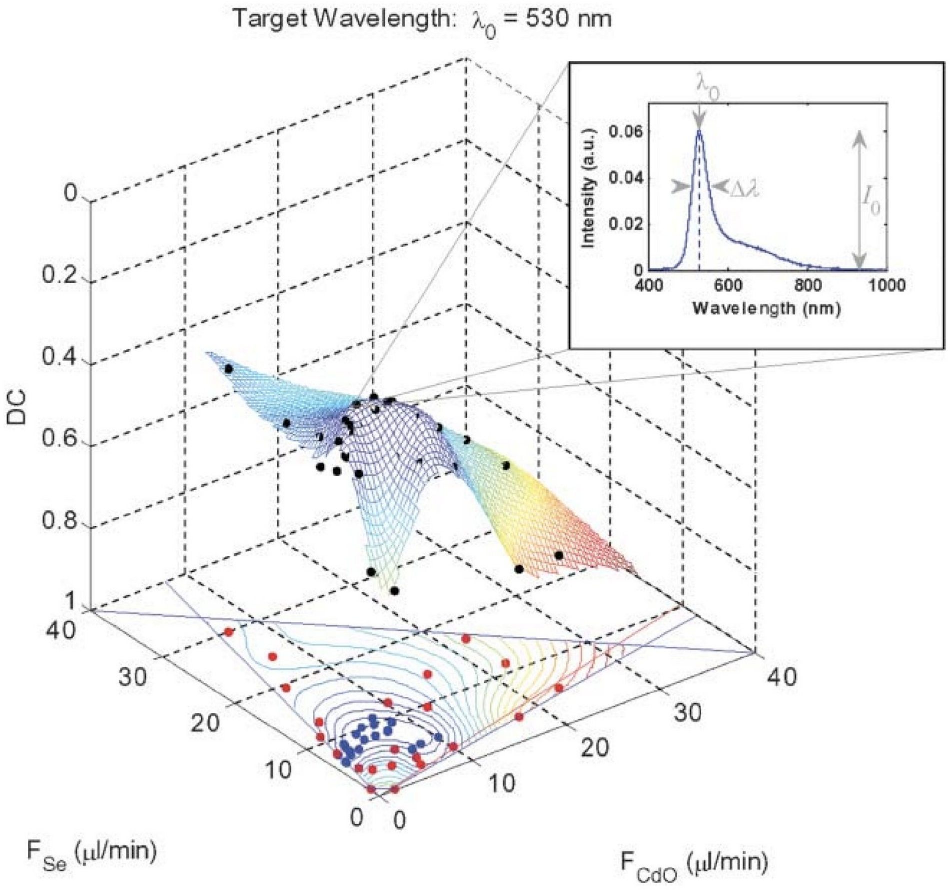


Figure 12: A visualization showing optimization of the dissatisfaction coefficient (DC) with respect to CdO and Se flow rates in the synthesis of CdSe nanoparticles. Black points represent experimental results, which are projected onto the bottom plane and colored such that blue/red dots correspond to samples with DC values greater/lesser than the median value of 0.26. The spectrum obtained from the best sample found (with a DC value of 0.19) is displayed by the

inset, showing a high emission intensity at 530 nm. Reproduced with permission[20]. Copyright 2007, Royal Society of Chemistry Publishing.

Although SNOBFIT often improves efficiency relative to combinatorial methods and GAs, it also displays a few key shortcomings that limit its applicability to certain optimization problems. First, its performance deteriorates when applied to problems with high-dimensional parameter spaces; studies have shown that SNOBFIT begins to underperform compared with many other global optimization techniques in cases where there are ten or more independent variables to consider[122]. Second, SNOBFIT is not capable of directly handling multi-objective optimization procedures[123]. An alternative route to deal with such problems is to instead optimize a composite objective function defined to capture changes in multiple quantities, however, this method gives no information regarding the trade-off between properties as would be revealed by the Pareto front. Last, SNOBFIT operates by establishing a set of individual models (i.e., quadratic functions) fit to approximate local regions of the objective function without providing a global model for the entire system. This approach therefore limits interpretability and makes it difficult to draw conclusions regarding general relationships between the variables and objectives.

4.1.3 Bayesian optimization

One of the most popular techniques for optimizing costly black-box objective functions is Bayesian optimization[124-126], which is designed to minimize the total number of experiments required to reach convergence. To accomplish this, optimization is carried out on a known and differentiable surrogate model rather than the objective function itself[105,106]. As shown in **Figure 13**, the surrogate model approximates the objective function using all available data points (e.g., from previously conducted experiments). This approximation is given by a probabilistic distribution of functions known as the *prior*, which is actively updated as new datapoints are sampled to form a *posterior* distribution that more closely resembles the true objective function. Calculating the prior and posterior are essentially regression problems that can be solved using a number of techniques, making Bayesian optimization versatile with respect to the types of data it can handle. Two models that are most commonly used for regression are Gaussian processes (GPs) and random forests (RFs), which typically work well with continuous and discrete search spaces respectively[8,124].

To analyze the surrogate model and choose the datapoints sampled during optimization, an acquisition function is used. The main task of the acquisition function is to query the objective function in a way that balances exploration and exploitation. Exploration aims to sample regions of the design space where uncertainty in the surrogate model is high, therefore aiming to improve the accuracy of the model and capture the global optimum. Exploitation uses predictions of the surrogate model to sample regions where the objective function is expected to be optimal. A purely explorative search requires an excessive number of evaluations to reach convergence whereas a purely exploitative search is prone to missing the global optimum, necessitating a trade-off between these two extremes. Some commonly used acquisition functions include *expected improvement* (EI), *entropy search* (ES), and *upper confidence bound* (UCB) as described in detail by past work[127]. It is important to keep in mind that no single approach is universally suited for all optimization problems (in accordance with the "no free lunch" theorem[128]). Rather, the acquisition function should be chosen to best suit the properties of the objective function at hand. For example, EI and UCB usually converge more rapidly than ES but are less explorative and therefore less suitable for objective functions with many local optima[127,129].

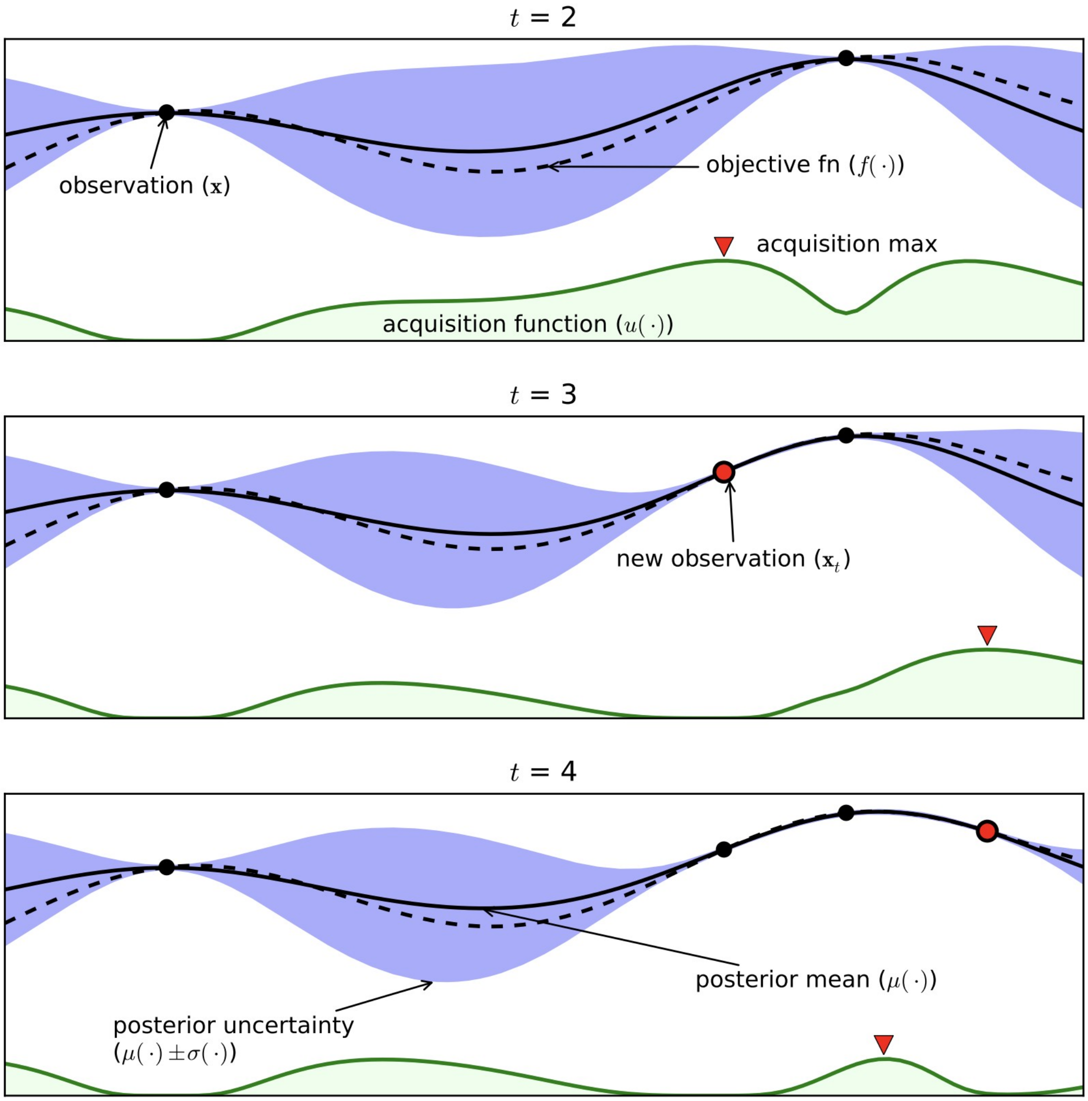


Figure 13: A schematic demonstrating the Bayesian optimization procedure. An approximation to the actual objective (solid black line) is given by the surrogate model (dashed black line), which is constructed to fit all known observations. Using the expected values and uncertainties (blue) given by this model, an acquisition function (green) is built. The maximum of the acquisition function is then identified to suggest new points at which to evaluate the objective, leading to increased refinement of the surrogate model. Here, a somewhat exploitative approach is used to select new evaluations near the predicted optimum. Reproduced with permission[130].

One of the first applications of Bayesian optimization to automated experimentation for materials design is described by Xue *et al.*[48] The goal of their work was to minimize thermal hysteresis effects in NiTi-based shape memory alloys by varying their composition. Although the corresponding samples could be represented using three simple variables (x, y, and z in $Ni_{50-x-y-z}Ti_{50}Cu_xFe_yPd_z$), the authors instead chose to map each composition into a higher-dimensional feature space containing information regarding the valence electron concentration, atomic radii, and electronegativity to capture electronic contributions to thermal hysteresis. To investigate the resulting six parameters, several surrogate models and acquisition functions were tested on a dataset of 22 randomly chosen experiments. A support vector regression model was found to outperform GPs as a surrogate model, and the most suitable acquisition function was identified as the *knowledge gradient* – a slight variant of EI that performs well when dealing with noisy objective functions[131]. Accordingly, this combination was used to perform a series of 58 new experiments that converged after ten iterations. Thermal hysteresis was decreased by as much as 42% relative to the initial samples. Furthermore, the surrogate model was found to accurately describe the objective landscape surrounding the optimum as reflected by a close agreement between predicted and experimentally measured values in the final few iterations.

A number of reports have extended Bayesian optimization to a greater variety of material systems and properties. This includes a study by C. Li *et al.*, where the synthesis procedure used to generate short polymer fibers was optimized to control the corresponding fiber size and shape with respect to precursor flow rates and reactor dimensions[75]. Using GPs and EI as the surrogate model and acquisition function, a five-dimensional parameter space was explored in three different series of trials, each toward a unique target fiber shape specified by the researchers. With only 20 experiments conducted per series, the results were shown to be highly dependent on the target to be optimized for. The authors tested whether changing the initial sampling of experimental parameters would improve the rate of convergence; however, they found that the results remained largely unchanged. We note that the mixed performance of Bayesian optimization in this situation can possibly be attributed to the low number of experiments conducted to search the broad, five-dimensional space of interest.

More recent work employed Bayesian optimization to build upon previous efforts involving autonomous carbon nanotube synthesis[22]. As discussed earlier, this problem was originally approached using a GA, which identified optimal nanotube growth conditions over the course of 840 experiments[21]. In contrast, Chang *et al.* revealed that optimization with GPs produced a similar set of optimal conditions in less than 200 experiments, therefore confirming the improved efficiency achieved using Bayesian techniques as opposed to genetic algorithms. However, the work also demonstrated that the performance of Bayesian optimization varies with respect to its underlying methods and hyperparameters. For example, when the three different kernel functions listed in **Figure 14 (a)** were used to define covariance, a key similarity measure employed in fitting the GPs, the optimized growth rates differed by as much as 15%. Even larger discrepancies were found to coincide with the choice of acquisition function, highlighting the importance of balancing exploration and exploitation. As shown in **Figure 14 (b)**, both UCB and EI perform well, whereas a purely exploitative approach known as the *maximum probability improvement* (MPI) converges to a sub-optimal solution.

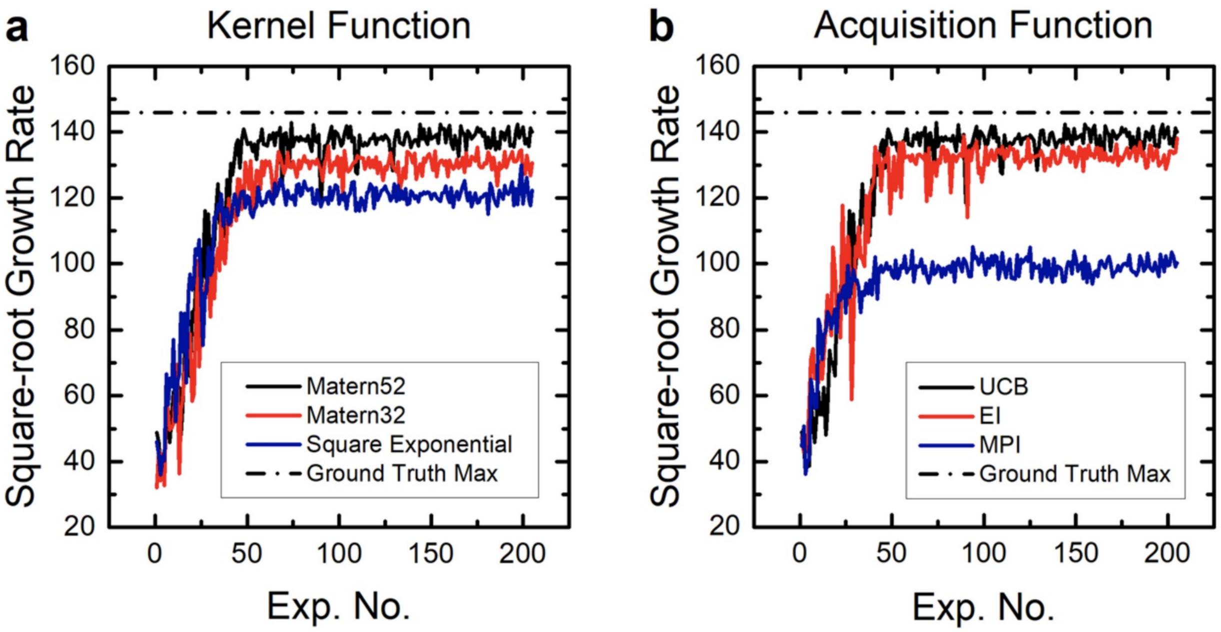


Figure 14: Results showing the evolution of growth rates measured during the synthesis of carbon nanotubes, with optimization carried out using a Bayesian approach. Gaussian processes are implemented as the surrogate model and three different kernel functions, listed in (a), are used to define covariance. In (b), the Matern52 kernel is employed with three different

acquisition functions: *upper confidence bound* (UCB), *expected improvement* (EI), and *maximum probability of improvement* (MPI). Reproduced with permission[22]. Copyright 2020, AAAS.

To improve upon existing methods and formalize a comprehensive tool for Bayesian optimization in the context of chemistry, the Probabilistic Harvard Optimizer Exploring Non-Intuitive Complex Surfaces (Phoenics) algorithm was developed[8]. The surrogate model and acquisition function in Phoenics are designed to evaluate costly and non-convex objective functions through iterative batches of experiments conducted in parallel. Because of its efficient rate of training, a Bayesian neural network (BNN) is chosen to approximate the objective function. While providing comparable performance to GPs, the time required to train a BNN scales linearly with the number of observations whereas GPs scale cubic[132]. Hence, using the BNN as a surrogate model ensures the training step is not rate-limiting in the optimization process, which becomes particularly important when the number of evaluations is large (e.g., in high-dimensional design spaces). A customized acquisition function is defined to balance exploration and exploitation through a control parameter set by the user. Batched experiments are chosen by randomly sampling the control parameter such that some experiments are highly explorative, while others favor exploitation. This method was tested on a set of multi-modal objective functions (e.g., the Ackley and Schwefel functions), with convergence achieved more rapidly than optimizations based on GPs and RFs in all cases.

The Phoenics algorithm has been implemented in the ChemOS software package[9] to carry out Bayesian optimization with fully autonomous experimental platforms in two studies. MacLeod *et al.* engineered a self-driving laboratory to maximize the hole mobility in organic thin films by tuning their annealing time and dopant concentration[12]. Given the low-dimensional search space involving only two parameters, convergence was obtained in less than 30 experiments, yielding a near ten-fold improvement in the mobility relative to initial samples. Applying the Phoenics algorithm to a more complex problem, Burger *et al.* designed a mobile robotic chemist that performed the necessary experiments to optimize the efficiency of aqueous photocatalysts for hydrogen evolution with respect to composition[13]. Based on a set of educated hypotheses, ten chemical species were identified as potentially promising components in the catalytic solutions. Bayesian optimization with Phoenics was used to explore the space of concentrations for each component. Despite the high dimensionality of this problem, the algorithm successfully reached convergence in about 600 evaluations conducted in batches of 16 experiments over the course of eight days. As illustrated in **Figure 15**, the resulting solutions displayed hydrogen evolution rates exceeding the baseline values by a five-fold margin.

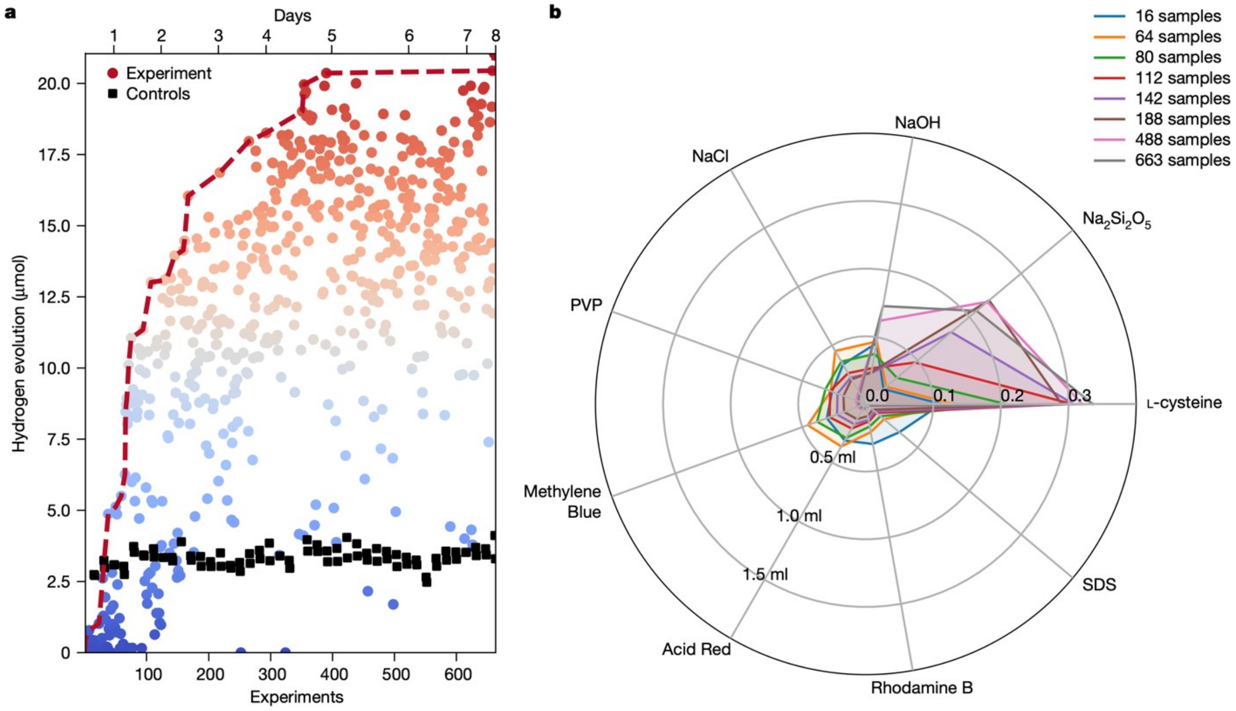


Figure 15: Results obtained during the automated optimization of hydrogen evolution mediated by a photocatalyst with varied composition. (a) Experimental measurements taken throughout a series of trial guided by Bayesian optimization. (b) A radar plot showing how the sampling of parameters changed over the course of optimization, with the shaded region boundaries representing the volume of each component dispensed after distinct numbers of experiments. Reproduced with permission[13]. Copyright 2020, Nature Publishing Group.

4.1.4 Summary

The examples reviewed in this paper support the efficacy of black-box optimization as applied in a variety of experimental workflows, with the Bayesian approach providing a particularly efficient route toward identifying global optima while minimizing the number of trials required. However, the performance of these techniques is limited by their inherently agnostic treatment of the objective, i.e., they make no assumptions regarding the underlying properties of the system at hand. As a result, black-box optimization methods tend to become intractable when (a) the design space is high-dimensional and therefore requires many evaluations to reach convergence, or (b) the majority of evaluations yield trivial results, which occurs if the objective function is constant-valued throughout large regions of the design space. Moreover, while black-box optimization methods are capable of illustrating trends in the data, further analysis must be carried out to relate these trends to physical phenomena.

4.2 Informed optimization

In contrast to black-box optimization techniques that learn and make decisions based solely on relationships between design variables and observed values of the objective function, informed optimization gains further insights by incorporating prior knowledge of the system into the optimization pipeline. In doing so, it may be possible to increase the efficiency with which the design space is explored, improve confidence in the identification of the global optimum, and provide results that are directly interpretable. As discussed by von Rueden *et al.*[133], devising an informed optimization algorithm requires the consideration of three questions: (1) what is the source of prior knowledge, (2) how is the corresponding information represented, and (3) where is the information integrated into the workflow? For experimentation in the physical sciences, possible solutions to these questions are summarized in **Table 1**.

Table 1: Taxonomy of the informed optimization approach. Listed are several common examples of knowledge sources, forms of representation, and routes for integration into the optimization pipeline. Adapted from von Rueden *et al.*[133]

***What:* Source**	***How:* Representation**	***When:* Integration**
Physics-based relations First principles calculations Past experimental results Expert intuition	Algebraic equations Differential equations Probabilistic relationships Logic-based rules Knowledge graphs	Parameter constraints Guided exploration Modified outcomes

Some common sources and representations of information include physics-based relations described by algebraic or differential equations, probabilistic relationships derived from previous experimental or theoretical findings, and logic-based rules or graphs developed using predictions from first principles calculations or expert intuition. To integrate this information into the optimization process, the selection of experiments can be biased toward subspaces where the objective is expected to be optimal[134], positive (negative) weights can be placed on particularly (un)desirable outcomes[135], and results can be validated through a comparison with known laws or empirical relationships[136]. In each case, all information can be incorporated prior to beginning the experimental procedure via *data fusion*. Alternatively, the algorithm may continuously update itself using a *hypothetico-deductive* model to iteratively form hypotheses and suggest experiments designed to confirm or disprove hypotheses[137]. This method is particularly promising because it provides a high degree of interpretability; however, it requires detailed planning regarding possible experimental outcomes and their implications for the system[138,139].To highlight the strengths and weaknesses of each informed optimization approach, several key examples of applications in materials science and chemistry are discussed below.

4.2.1 Data fusion

In the context of optimization, data fusion refers to the mapping of information from multiple sources onto an ensemble model where all knowledge is represented using a single composite function[140]. This method is commonly used to obtain an optimal balance between theory and experiment – the former is cheap to evaluate but prone to inaccuracy, whereas the latter is accurate but expensive to carry out. For example, we consider the relationship between the thermodynamic stability of materials, calculated via density functional theory (DFT), and their synthesizability observed in experiment under a given set of conditions. While these two properties tend to be correlated for many materials, they do not always agree with one another as kinetic barriers can prevent a system from reaching its equilibrium state (discussed in **Section 5.3**)[141]. Bridging the gap, Sun *et al.* employed data fusion to combine DFT-calculated Gibbs free energies of mixing with observed instabilities in halide perovskites to guide experiments toward maximizing the stability of samples with respect to their composition[134]. A layout of the process is shown in **Figure 16 (a)**. After initially performing DFT calculations to predict the free energy of mixing across a grid of compositions in the (Cs-MA-FA)PbI_3 space, a "physics-informed" Bayesian optimization technique was used to model stability and suggest experimental trials. Data fusion was accomplished by modifying the EI acquisition function to incorporate a probabilistic distribution constructed from a fit of the free energy data. This modification biased the suggested experiments toward compositions where the DFT-calculated free energy of mixing was highly negative and therefore stability was predicted to be strong. Following the data fusion approach, 112 samples were tested over the course of four experimental iterations. As illustrated in **Figure 16 (b)**, guidance by DFT enabled rapid convergence to a narrow subspace displaying high stability within the first batch of experiments, followed by minor variations in composition throughout the remainder of experiments to further optimize the objective. These results demonstrate that DFT calculations provide valuable prior insight that can be used to inform optimization and ensure a more efficient exploration of the design space.

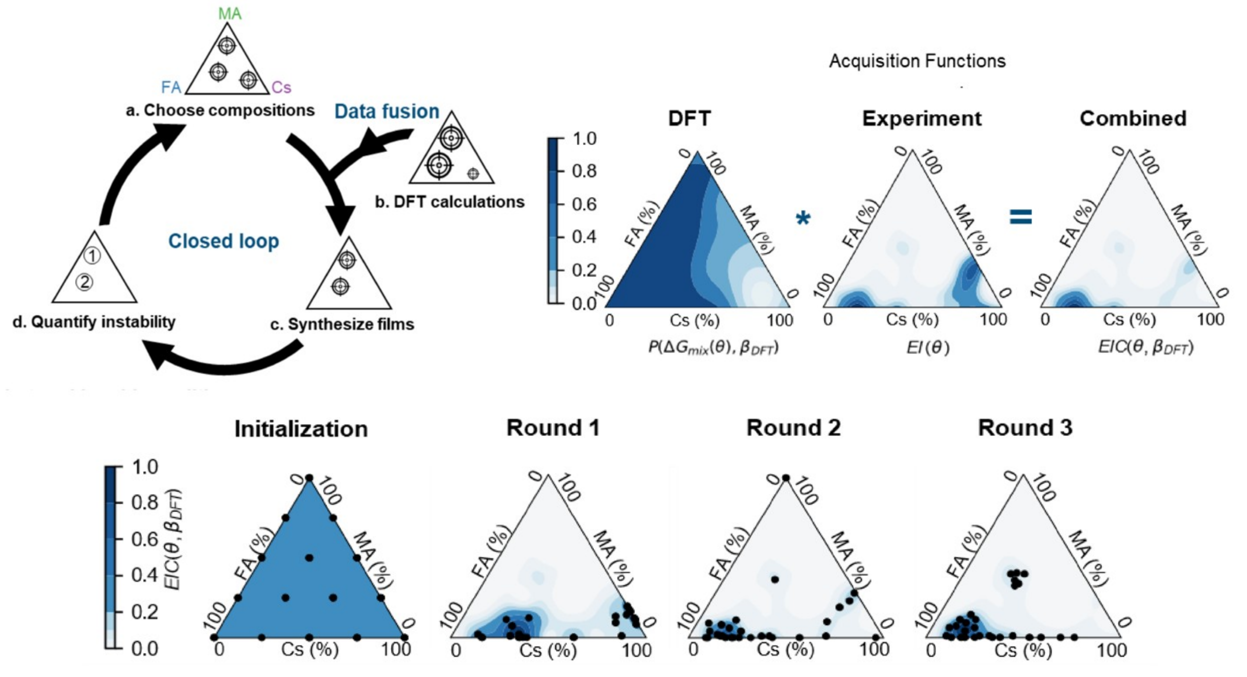


Figure 16: (a) Closed-loop process mediating the optimization of stability in halide perovskites, (Cs-MA-FA)PbI_3), for which free energy data obtained from DFT calculations is integrated into the traditional experimental procedure. (b) The *expected improvement* (EI) acquisition function is fit to experimental measurements of stability, then modified to account for predictions of ΔG_{mix} from DFT. (c) After an initial uniform sampling of compositions, experiments (black dots) are guided by the Bayesian optimization algorithm, showing convergence to narrow regions of the composition space with robust stability. Reproduced with permission[134].

Data fusion can also be used to improve the representation of design variables. Shields *et al.* recently developed an informed Bayesian optimization algorithm designed to optimize chemical reactions by maximizing the yield of a target phase with respect to the choice of precursors and synthesis conditions[142]. Given a set of promising reaction pathways, molecular precursors were represented by chemical descriptors (e.g., dipole moments, molar volumes, and electrophilicity) that were calculated in advance using DFT. The relationships between these descriptors and the objective function (the target yield) were modeled using Gaussian processes. For each experimental iteration, a batch of reactions were chosen according to Thompson sampling of the EI acquisition function. The authors first applied their algorithm to optimize the direct arylation of imidazoles, for which the design space included 12 ligands, four bases, three temperatures, and three concentrations – totaling 1,728 distinct reactions to be explored. As a benchmark, all of these reactions were carried out beforehand with HT experimentation, and the data was used to set up a game where 50 expert chemists were asked to optimize the reaction by iteratively selecting promising precursors and conditions until they believe they had found the true optimum (maximum target yield), or until a limit of 50 experiments was reached. Although human experts typically identified better-performing reaction pathways in early trials, the Bayesian optimization algorithm surpassed the average human performance within three batches of five experiments and obtained 99% target yield by the final batch. The algorithm was also applied to several research problems with increased design space complexity. For example, the Mitsunobu reaction[143] was optimized with respect to 180,000 possible reactions derived from 12 phosphines and five solvents with varied concentrations. In only four batches of 10 experiments, the optimizer identified three sets of reactions conditions that gave improved yield over the standard reaction, hence confirming the excellent performance of informed optimization. We note that the marked success of this approach relies on (i) a suitable choice of candidate reactions (precursors and conditions) based on expert knowledge, and (ii) DFT-calculated descriptors of the design variables, which captures electronic and steric relationships between molecular precursors and provides an improved representation over one-hot encoding (i.e., using dummy variables).

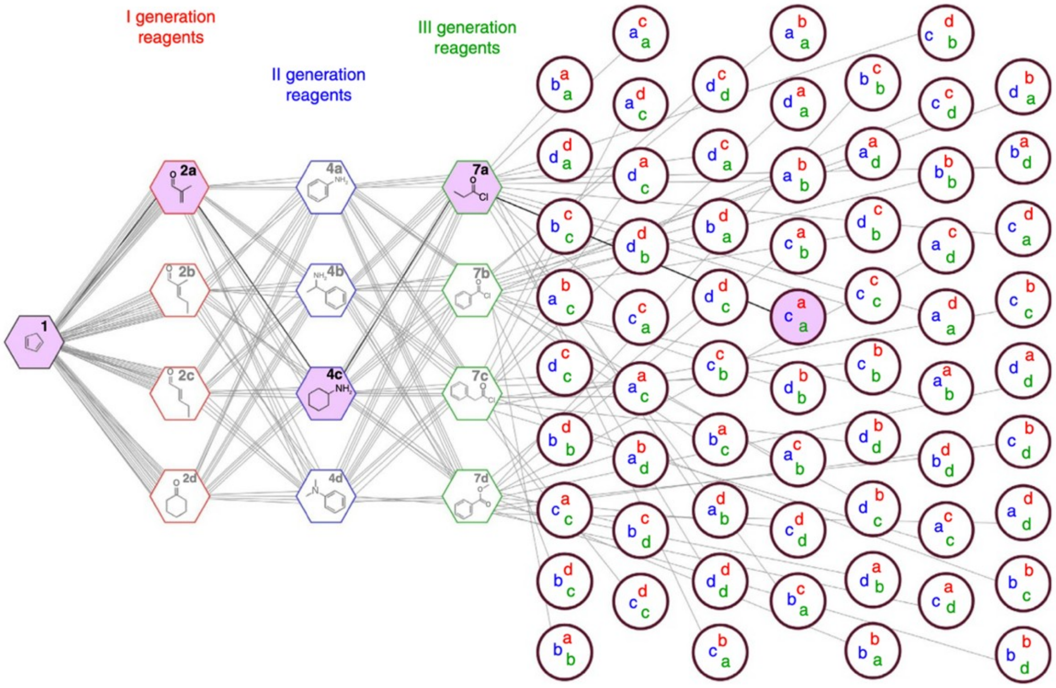
In the place of DFT calculations as an information source, one may alternatively employ physical relationships based on theoretical frameworks or empirical observations. For example, Ren *et al.* utilized a data fusion approach to embed physics domain knowledge into a Bayesian optimization procedure designed to maximize the efficiency of GaAs-based solar cells with respect to their growth temperature[136]. Given that only temperature was considered as a free variable, a simple black-box search method would likely suffice. However, the authors emphasized that such an approach does not reveal why that temperature maximizes solar cell efficiency. Instead, they developed a hierarchical Bayesian model to reveal relationships between the growth temperature and material descriptors including dopant concentrations, carrier lifetimes, and recombination velocities – each of which were related to the final performance of the solar cell. To accomplish this, temperature was mapped into a latent space represented by parameters that were used to calculate the material descriptors via Arrhenius equations, from which the efficiency of the corresponding solar cell was predicted using a neural network trained on simulated I-V curves. The model was used to guide a series of 25 experiments, with the final batch of samples showing an efficiency 6.5% higher than the baseline value obtained from a grid sampling of growth temperatures. For comparison, the authors performed similar tests without the influence of the Arrhenius equations, i.e., growth temperatures were mapped directly onto the descriptors in a black-box manner. This approach led to slower convergence and greater disagreement between predicted and experimentally observed properties, therefore supporting the advantages provided by data fusion.

4.2.2 Hypothetico-deductive models

For research problems where the application of data fusion is restricted (e.g., when relevant calculations cannot be performed prior to the experiment), hypothetico-deductive models may instead be used to actively learn from experimental results, interpret their implications for the system, and propose hypotheses to be tested by subsequent experiments. King *et al.* pioneered the automation of this approach in their design and application of a robotic scientist named Adam[137]. With the goal of identifying the genes encoding a group of enzymes, Adam was engineered to systematically probe metabolic networks describing the biochemical processes that may occur. These networks consist of nodes representing metabolites (i.e., intermediates and products) that are connected through edges representing reactions catalyzed by a known enzyme(s). If the genes encoding a specific enzyme are removed from the organism through mutation, any reaction catalyzed by that enzyme will be slowed[144]. Based on this concept, the authors constructed a decision tree with the following structure: high-level nodes pose fundamental questions (what genes encode an enzyme?), mid-level nodes form hypotheses related to these questions (a list of suspected genes), and low-level nodes suggest experiments to test the hypotheses (measurements of reaction rates in mutated organisms). By traversing this structure and iteratively proving or disproving a series of hypotheses, Adam analyzes low-level experimental data to propose solutions to high-level questions. The automated platform was demonstrated to identify the genes encoding 13 different enzymes in yeast, with all predictions confirmed by manual experimentation. However, because the success of Adam was enabled by the availability of detailed metabolic networks extracted from bioinformatic databases[145,146], extension to novel systems would require a re-design of the hypothetic-model to reflect the suspected genotype.

A counterpart to the metabolic network in synthetic organic chemistry is the reaction network, a graph used to illustrate possible connections between a set of reactants and products by considering likely intermediate reactions that may occur. Given the exceptionally large number of possible transformations deriving from or giving rise to any arbitrary molecule, conceiving detailed reaction networks is generally intractable for a human researcher. Instead, a variety of techniques have been developed for computer-aided synthesis planning (CASP)[138,147], whereby information taken from reaction databases along with chemistry-based rules and heuristics are employed. In the context of optimization, CASP methods are often used to maximize the yield of a desired product through retrosynthetic analysis, i.e., by starting from the target molecule and working "backwards" through the network to identify suitable precursors. To this end, a variety of state-of-the-art algorithms such as Monte Carlo tree search have been used to efficiently traverse promising branches in the reaction network, provide a ranking of the possible retrosynthetic routes, and suggest promising precursors[138]. However, the retrosynthetic process has not yet been fully integrated into a closed-loop workflow enabled by iterative experiment-theory feedback, likely owing to the large number of experiments that would be required and the difficulties associated with generalized identification of molecules resulting from synthesis.

Simplifying the problem described above, Dragone *et al.* created a robotic system designed to actively explore a reaction network by searching for the most reactive pathways[148]. Reactivity was quantified using a "reaction selection index" (RSI), defined as the mean squared error between the infrared spectrum of the product and that of the reactants. As a proof of concept, the authors worked with amide synthesis via the reaction network shown in **Figure 17**. One core molecule was subjected to a three-step reaction, where each step may involve four different reagents, resulting in 64 possible pathways ($4 \times 4 \times 4$). Possible intermediates and products were predicted using chemistry-based knowledge of likely transformations. To explore the network, four reactions were carried out between the core molecule and each reagent in the first generation. The reaction with the highest RSI was then selected and its product was combined with each reagent in the second generation to perform four new reactions. By repeating this process across the second and third generations, the most reactive pathway was identified (highlighted in **Figure 17**) with a final product yield of 27%. This task was accomplished in only 12 experiments, significantly reducing the workload relative to a brute-force approach where 64 experiments would be necessary to explore all possible pathways. Hence, the advantages of informed optimization are realized not only by constructing a reaction network to reach the desired product, but also by exploring the network more efficiently with reactivity as a guiding metric.



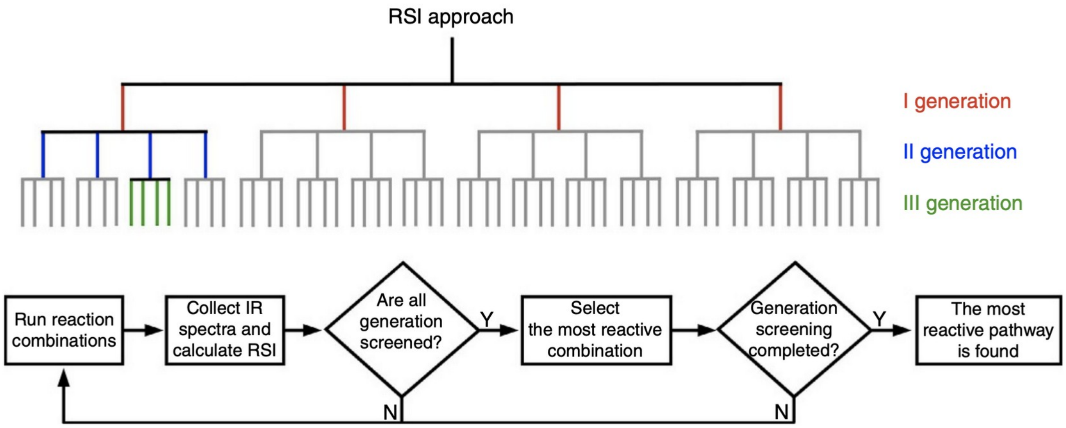


Figure 17: (a) An illustration of the reaction network considered in the optimization of reactivity for a single core molecule undergoing reactions with four possible reagents. Subsequent transformations are enumerated throughout three generations, with final products listed in the group of nodes labeled with a, b, c, or d. (b) A schematic illustrating the reaction selection index (RSI) approach; as opposed to probing all possible pathways, only those with the highest measured reactivity are investigated, therefore significantly reducing the experimental workload. Reproduced with permission[148]. Copyright 2017, Nature Publishing Group.

4.2.3 Summary

Both data fusion and hypothetico-deductive modeling have been demonstrated to accelerate optimization through integration of physics-based or data-driven knowledge. To choose between these techniques for future automated platforms in materials science, consideration may be given to the types of properties/processes in focus as well as the ability to extract relevant data from databases or calculations. Based on the examples reviewed here, we suggest that data fusion is well-suited to optimize materials *properties*, such as stability or optoelectronic performance, so long as they can be calculated (e.g., using DFT) or deduced from empirical relations with design variables (e.g., using Arrhenius equations). In contrast, although hypothetico-deductive modeling has so far been limited to applications in organic chemistry and biology, we propose that this approach may be used to optimize complex materials *processes*. The synthesis of novel inorganic compounds is a prime candidate for hypothetico-deductive modeling because the underlying reactions may be decomposed into sequences of reagents, intermediates, and products – each of which can be verified through experimentation. However, as will be discussed below, several key developments in theory, characterization, and data availability are required before informed optimization can be broadly extended to inorganic materials synthesis.

5. Perspectives

The design of novel inorganic materials for any application hinges upon the ability to synthesize desired compounds. The successful realization of candidate materials is a major hurdle before experimental testing for the application of interest and further engineering to optimize the material's performance. The need for an expert to carefully plan and execute synthesis trials, interpret their results, and design subsequent experiments makes synthesis a time- and labor-intensive process. This limits the amount of information-gathering experiments that can be performed, instead requiring that the product be obtained in a black-box fashion (i.e., by trial-and-error). An ideal autonomous synthesis system would be capable of producing materials across versatile platforms (e.g., solid-state or solution-based) without requiring any prior knowledge of successful synthesis procedures. With the integration of closed-loop interpretation and optimization, such a system has the potential to drastically improve the rate at which candidate materials can be prototyped for a variety of technologies. With this objective in mind, we outline the necessary advancements in hardware, interpretation techniques, and decision-making algorithms needed to realize automated and closed-loop synthesis of novel inorganic materials.

5.1 Synthesis

As discussed in **Section 2.1**, the full automation of solution-based syntheses can be achieved by using electronic and programmable syringe pumps to transfer samples between modules that perform unit operations such as mixing, heating, and filtration[11,13]. Although most existing applications deal with organic molecules, recent work has shown that similar techniques can be used to automate the synthesis of inorganic materials through sol-gel or precipitation methods[18-20], which are useful to produce nanoparticles (e.g., metal oxides) so long as there exist appropriate precursors with high solubilities in common solvents such as water. To expand the scope of compounds that can be made by solution-based techniques, hydrothermal synthesis may also be applied in future automated platforms. This approach permits a wider range of starting materials since high temperatures and pressures facilitate increased solubilities. Furthermore, it can be used to access compounds that are only metastable under ambient conditions[149]. A major challenge in the automation of hydrothermal synthesis is the robotic loading and unloading of the autoclave reaction vessel, which has recently been demonstrated in a commercial system (the Chemspeed FLEX autoplant[150]) that may soon assist in developing materials at the experimental stage.

For thin film synthesis, there are many reported workflows that can generate combinatorial libraries of samples with varied compositions in an automated and HT manner[14,88]. However, these platforms are distinguished from closed-loop experimentation by a lack of automation for data interpretation, decision making, and replacement of samples between experimental iterations. These shortcomings prevent an efficient probing of experimental variables beyond composition, including synthesis conditions such as temperature or pressure. Fortunately, the examples reviewed in **Section 2.2** demonstrate that automated thin film deposition can be integrated with robotics and optimization algorithms to develop closed-loop platforms. So far, autonomous thin film synthesis has been achieved with CVD[21,22], spin coating[12], magnetron sputtering[39], and reactive sputtering with lg-LSA[40]. These methods alone can be used to produce many types of materials, and increased adoption of other techniques such as molecular beam epitaxy and pulsed laser deposition support a promising future for the closed-loop automation of thin film synthesis.

In contrast to syntheses based on solutions or thin films, the automation of solid-state synthesis remains limited. As described in **Section 2.3**, several unit operations such as mixing, densification, and firing have been automated or parallelized; however, integrating these components without manual intervention between operations is challenging. To overcome these difficulties, we consider the advantages of solution-based and thin film syntheses with respect to ease of automation. On one hand, transferring of samples dissolved in a solution can be accomplished with electronic syringe pumps, whereas the handling of solid powders is more difficult. To solve this problem, a semi-solution-based approach may be taken whereby solid powders are handled as a slurry to allow transfer using syringe pumps. On the other hand, thin films avoid the difficulties associated with sample transfer by performing all experiments (including synthesis and characterization) on a single substrate that is more easily handled by robotic systems. Extending this concept to solid-state synthesis, automated platforms may rely on a multipurpose container that has robust mechanical, chemical, and heat resistances such that it can be used throughout the entire synthesis procedure without any degradation or contamination of the samples. This approach would therefore remove the need to extract and transfer the materials between unit operations.

5.2 Interpretation

After performing a synthesis trial, phase identification is needed to decide whether a planned reaction was successful, or to glean information regarding why it may have failed. Because of the wide availability of reference patterns for crystalline inorganic materials, XRD is often the tool of choice for this purpose[58,94]. Although automated loading of samples and analysis with XRD can be carried out with commercially available systems, interpreting the resulting spectra is a more difficult task. Of the methods discussed in **Section 3.2**, we suggest that those based on machine learning are most promising given three unique advantages. First, they provide a complete end-to-end treatment of raw spectra without requiring sensitive pre-processing steps such as peak extraction or baseline correction[71]. Second, machine learning models can account for possible non-idealities (e.g., from experimental artifacts) by performing data augmentation across spectra in the training set[51]. Last, ensemble models can be employed to generate probabilistic distributions associated with likely phases, therefore providing an estimation of uncertainty associated with the final classification[72].

Of the machine learning models previously developed, CNNs are particularly adept at handling XRD spectra because they use convolution to decompose complex patterns into feature maps representing distinct properties such as peak positions or intensities, which can then be related to corresponding phases through a neural network. Indeed, Oviedo *et al.* have reported improved performance of CNNs relative to several alternative machine learning algorithms and full-profile techniques when applied to experimentally measured single-phase patterns obtained from a combinatorial library of thin film samples[51]. We note that a key component of their algorithm was the incorporation of peak shifts in the training spectra, which reflect the epitaxial strain that is common in thin film samples – this augmentation is therefore denoted as "physics-informed”. A similar approach was employed in the work Maffettone *et al.*, where changes in peak intensities were related to possible texture in the samples[72]. Future work may expand upon these concepts to further improve the model’s accuracy and generality by incorporating data augmentation designed to account for all artifacts that commonly arise during sample preparation and synthesis. If proven successful by rigorous testing on experimentally measured patterns, CNNs will prove vital to facilitate phase identification in closed-loop experimental platforms enabling autonomous inorganic synthesis.

5.3 Decision making

Once the compounds produced by a reaction are known, a decision must be made regarding the next batch of experiments. Automated decision making for targeted syntheses requires either many experiments or advanced knowledge of the underlying objective function. For solid-state synthesis, the use of high temperatures and long heating times typically precludes a large number of experiments performed in parallel. Furthermore, unlike reactions between organic molecules, which can often be decomposed into a series of unit operations involving functional group additions/removals[151], solid-state reaction mechanisms are less well understood. Due to the bulk nature of materials, reaction sequences based on nucleation and growth are not easily predicted, and therefore the objective function that governs inorganic synthesis is difficult to describe. When a measurable amount of the desired product is consistently formed throughout a set of experiments, the objective function can be simplified by performing black-box optimization to maximize the target yield[31,113,121]. However, when dealing with more complex syntheses involving novel compounds, an informed optimization approach is better suited to overcome the low success rates of many trial reactions. As described in **Section 4.2.2**, the complexity of objection functions governing organic syntheses can often be reduced by designing reaction networks, which map input parameters (such as precursors) onto possible experiments outcomes (intermediates and products)[138,148]. With hypothetico-deductive modeling, promising pathways in the network can be hypothesized and verified by performing stepwise reactions and measuring the yields of expected products. While current applications of reaction networks remain mostly limited to small molecules, similar concepts may be extended to inorganic materials synthesis if an improved understanding of solid-state reactions is realized.

To build a reliable reaction network, it is necessary to predict which phases are likely to form from a set of specified precursors and synthesis conditions. This task is challenging for solid-state reactions because many factors can prevent a system from reaching thermodynamic equilibrium, and therefore inorganic materials synthesis is often treated as a black box that must be probed with trial-and-error experiments. Improving upon this limitation, recent work has sought to peak into the black box of solid-state synthesis using *in situ* characterization and computational modeling of thermodynamics. For example, Bianchini *et al.* monitored the synthesis of sodium metal oxides to highlight the importance of intermediate phase selection and its effects on the reaction products[23]. The authors argued for two types of reaction pathway controls. When the driving force is large, as in the initial reaction between precursors, the pathway is dictated by the compositionally unconstrained reaction energy – i.e., the first phase that forms during high-temperature synthesis is the one which locally maximizes the free energy reduction at the precursor interfaces, regardless of the overall composition of the mixture. This phase then evolves to the equilibrium ground state for the overall composition of the mixture through transformations that are either kinetically or thermodynamically controlled. When the remaining driving force is low, metastable intermediates were observed when facile transformation mechanisms from previous phases along the pathway allowed them to lower the free energy of the system faster than through the formation of the equilibrium phases.

Metastable intermediates that accompany reactions with low thermodynamic driving forces are often templated by structural similarities with preceding phases. For example, in the synthesis of $KBiS_2$ from K_2S and Bi_2S_3, McClain *et al.* showed that K_3BiS_3 initially formed as an intermediate because it shares a similar motif to the structures of the starting materials[152]. The authors propose that the distorted octahedral KS_6 and BiS_6 coordination environments in K_3BiS_3 serve as transition states between the distorted octahedral (tetrahedral) complexes of Bi_2S_3 (K_2S) and the ideal octahedral environments in $KBiS_2$. In addition to structure, the selectivity of intermediate phase formation can also be controlled by environmental conditions such as temperature and partial pressures of gaseous species. This was demonstrated by Todd *et al.* in the synthesis of $YMnO_3$ from a combinations of precursors Mn_2O_3, YCl_3, and Li_2CO_3[153]. They found that a high oxygen fugacity favors the formation of $LiMnO_2$ and YOCl as intermediates over Y_2O_3 and Mn_2O_3. Moreover, because the former pair of compounds are layered, their high diffusion rates enable rapid formation of the target ($YMnO_3$), further clarifying the role of kinetics in dictating reaction pathways between solid precursors. Alternatively, when compounds arise from molten phases rather than from solids, the first intermediate to form is usually the phase with the lowest barrier to nucleation. For example, Shoemaker *et al.* found that a single metastable phase, $K_5Sb_2S_8$, initially nucleated from a melt formed by Sn and K_2S_5[154]. This novel compound was transient in the reaction sequence, later decomposing to the thermodynamic ground state comprised by a mixture of $K_2Sn_2S_5$ and $K_4Sn_2S_8$.

The utility of the concepts described above are highlighted in the work of Miura *et al.*, where the influence of precursors on resulting reaction pathways taken throughout the synthesis of $YBa_2Cu_3O_{6+x}$ (YBCO) was studied[155]. In agreement with the work of Bianchini *et al.*[23], it was observed that initial reaction occurred at interfaces with the largest thermodynamic driving forces. The first intermediates were shown to drive the reaction along a pathway that led to rapid synthesis through a low-temperature liquid phase, which would have been inaccessible if other intermediates had formed. These findings suggest that precursors can be carefully selected to control reaction pathways and enable the formation of targeted products. However, accurately predicting the intermediates that form from a set of precursors is challenging as kinetic processes (e.g., diffusion, nucleation, and growth) in solids are difficult to calculate from first principles. As opposed to organic molecules, where reactions can be rationalized by considering molecular-level interactions (i.e., breaking and forming individual bonds), predicting kinetic contributions to solid-state reactions is challenging because they involve concerted displacements and interactions among many species over extended distances.

To simplify the analysis of solid-state reactions, McDermott *et al.* developed an algorithm that predicts all possible pathways between a set of precursors and final product(s), with each pathway decomposed into a sequence of pairwise reactions[156]. The thermodynamic driving forces associated with each pathway are used to calculate a cost function, from which suspected reaction sequences are ranked from most to least favorable. As an example, the authors demonstrated that their algorithm was able to predict the reaction pathway taken during the synthesis of $YMnO_3$. However, to extend this approach to more general syntheses of arbitrary compounds, it is necessary to incorporate factors beyond thermodynamic driving forces. As described in the previous three paragraphs, kinetic intermediates that form more rapidly than the equilibrium ground state often dominate when the driving forces for formation are comparable. Hence, although difficult to calculate directly, future algorithms may benefit from an estimation of transformation rates based on structural descriptions (i.e., structures can serve as templates for nucleated phases). An initial foray into this approach was recently published that derives a structural descriptor from classical nucleation theory to help indicate which reactions are plausible[157]. Further advances in the prediction of reaction pathways should address the potential for melting and preferential nucleation from molten phases that form during high-temperature syntheses.

It is clear that more work is needed to improve our understanding of the underlying objective function that governs solid-state synthesis, and an increased adoption of *in situ* characterization would be helpful for this purpose. The information gained from *in situ* measurements is vital to validate any predictions made by the aforementioned theories and could be used to modify or extend their rules accordingly[24]. Additionally, any knowledge regarding intermediates that form during a synthesis trial provide direct insight into why a synthesis attempt succeeded or failed[153,154,158,159]. Therefore, the identification of intermediate phases is particularly useful in the optimization of inorganic materials synthesis because it allows us to understand why the target phase was or was not formed. This knowledge can also be reincorporated into the experimental procedure to actively guide materials synthesis. For example, Rakita *et al.* used *in situ* synchrotron X-ray absorption spectroscopy to monitor the oxidation of copper samples placed in a controlled reaction environment[158]. By continuously observing changes in the average Cu oxidation state with respect to the partial pressure of the oxidant, more or less gas flow was fed into the reaction vessel to target an average oxidation state of Cu^{1+}. Although this state is difficult to obtain manually as it requires a precise balance of Cu^{0} and Cu^{2+} species, the autonomous workflow successfully formed samples with an average oxidation state near Cu^{1+} through active control of the reaction environment.

Unfortunately, the availability of *in situ* characterization techniques with high resolution and a fast scan rate remains limited, especially within the context of automated platforms. While methods such as thermogravimetric analysis or differential scanning calorimetry can sometimes be used to indicate when a reaction or phase change has occurred, they do not provide a means of identifying which intermediate phases formed. For this purpose, spectroscopic or diffraction-based techniques are necessary. However, obtaining reliable *in situ* data typically requires that high-intensity radiation or neutrons be used, which can be obtained only with high-energy sources (e.g., a synchrotron or nuclear reactor). Therefore, in the absence of improved in-lab diffractometers, we suggest that an automated synthesis platform can potentially replicate the information gained from *in situ* characterization by carrying out reactions across a range of temperatures, quenching after predetermined annealing times, and performing *ex situ* measurements on the resulting samples. This approach would provide discrete snapshots of the reaction pathway, and the throughput of automated platforms may enable sufficient measurements to mimic traditional *in situ* methods.

In the place of information obtained using new experiments, previously reported data can be tabulated and analyzed to reveal statistical trends, develop models, and validate predictions. In recent years, large-scale databases and application programming interfaces (APIs) that provide experimental and/or calculated materials data sets have become increasingly popular. For example, the ICSD contains hundreds of thousands of experimentally determined crystalline structures[94], and the Materials Project contains a variety of DFT-calculated properties across a comparable number of materials[2]. For information regarding experimental synthesis routes, however, there exists a limited amount of well-structured data with sufficiently high quality to be coupled with autonomous systems. To overcome this shortcoming, NLP efforts have been developed and applied to collect and curate synthesis data in relevant literature, such as synthesis conditions and resulting phases[25,98,99,160]. However, because of inherent biases in published results, the lack of negative examples (i.e., failed synthesis attempts) limits the ability to learn which factors contribute to an experiment's success[100,161]. Interestingly, the potential of coupling NLP with autonomous synthesis platforms has been demonstrated by the work of Mehr *et al.*, where a fully integrated system was built to carry out the synthesis of organic compounds with experimental parameters parsed directly from the literature[162]. With the increasing development of text-mined inorganic synthesis databases[160,163], we suggest that similar methods may be applied to power autonomous inorganic synthesis platforms.

5.4 Summary

With the growing development and adoption of self-driving laboratories in materials science, time can be freed up for the researcher to work on high-level tasks involving conceptual formulation and interpretation, while leaving low-level, manual efforts to be carried out by the robotic system[164,165]. These high-level tasks may include choosing the candidates to be synthesized based on prior knowledge such as structure-property relationships, designing the experiment such that all parameters and bounds are chosen to ensure maximal efficiency, and analyzing the corresponding results to ascertain broader scientific implications (e.g., a clarification of factors influencing synthesis) which inform further experimentation. Indeed, the sparsity of existing theories for solid-state reactivity emphasizes the ample work left for the human to develop these concepts and enable self-driving laboratories to meet their full potential in inorganic materials science. The lack of these established theories should not however precede the development of autonomous systems as expanding available datasets and automating time-consuming tasks will drive the development of more predictive theories for the directed synthesis of novel inorganic materials.

References

1 D. P. Tabor *et al.* Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature reviews materials* **3**, 5-20 (2018).

2 A. Jain *et al.* The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

3 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig & C. Wolverton. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **65**, 1501-1509 (2013).

4 M. Shevlin. Practical High-Throughput Experimentation for Chemists. *ACS Medicinal Chemistry Letters* **8**, 601-607 (2017).

5 C. A. Nicolaou, I. A. Watson, H. Hu & J. Wang. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *Journal of Chemical Information and Modeling* **56**, 1253-1266 (2016).

6 A. Weber, E. v. Roedern & H. U. Stilz. SynCar: An Approach to Automated Synthesis. *Journal of Combinatorial Chemistry* **7**, 178-184 (2005).

7 P. Szymański, M. Markowicz & E. Mikiciuk-Olasik. Adaptation of High-Throughput Screening in Drug Discovery - Toxicological Screening Tests. *International Journal of Molecular Sciences* **13**, 427-452 (2012).

8 F. Häse, L. M. Roch, C. Kreisbeck & A. Aspuru-Guzik. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Central Science* **4**, 1134-1145 (2018).

9 L. M. Roch *et al.* ChemOS: An orchestration software to democratize autonomous discovery. *PLoS ONE* **15** (2020).

10 N. S. Eyke, B. A. Koscher & K. F. Jensen. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends in Chemistry* **3**, 120-132 (2021).

11 S. Steiner *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363** (2019).

12 B. P. MacLeod *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* **6** (2020).

13 B. Burger *et al.* A mobile robotic chemist. *Nature* **583**, 237-241 (2020).

14 A. Ludwig. Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. *npj Computational Materials* **5** (2019).

15 F. Ren *et al.* Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances* **4** (2019).

16 S. Sun *et al.* Accelerated Development of Perovskite-Inspired Materials via High-Throughput Synthesis and Machine-Learning Diagnosis. *Joule* **3**, 1437-1451 (2019).

17 B. Li *et al.* Hydrogen Storage Materials Discovery via High Throughput Ball Milling and Gas Sorption. *ACS Combinatorial Science* **14**, 352-358, doi:10.1021/co2001789 (2012).
18 G. H. Carey & J. R. Dahn. Combinatorial Synthesis of Mixed Transition Metal Oxides for Lithium-Ion Batteries. *ACS Combinatorial Science* **13**, 186-189 (2011).
19 T. Adhikari *et al.* Development of High-Throughput Methods for Sodium-Ion Battery Cathodes. *ACS Combinatorial Science* **22**, 311-318, doi:10.1021/acscombsci.9b00181 (2020).
20 S. Krishnadasan, R. J. C. Brown, A. J. deMello & J. C. deMello. Intelligent routes to the controlled synthesis of nanoparticles. *Lab on a Chip* **7**, 1434-1441 (2007).
21 P. Nikolaev *et al.* Autonomy in materials research: a case study in carbon nanotube growth. *npj Computational Materials* **2** (2016).
22 J. Chang *et al.* Efficient Closed-loop Maximization of Carbon Nanotube Growth Rate using Bayesian Optimization. *Scientific Reports* **10** (2020).
23 M. Bianchini *et al.* The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature Materials* **19**, 1088-1095 (2020).
24 H. Kohlmann. Looking into the Black Box of Solid-State Synthesis. *European Journal of Inorganic Chemistry*, 4174-4180 (2019).
25 H. Huo *et al.* Semi-supervised machine-learning classification of materials synthesis procedures. *npj Computational Materials* **5** (2019).
26 J. R. Mohrig, D. Alberg, G. Hofmeister, P. F. Schatz & C. N. Hammond. *Laboratory Techniques in Organic Chemistry*. (W. H. Freeman and Company, 2014).
27 G. Schneider. Automating drug discovery. *Nature Reviews Drug Discovery* **17**, 97-113 (2018).
28 B. Gutmann, D. Cantillo & C. O. Kappe. Continuous-Flow Technology — A Tool for the Safe Manufacturing of Active Pharmaceutical Ingredients. *Angewandte Chemie* **54**, 6688-6728 (2015).
29 A. Adamo *et al.* On-demand continuous-flow production of pharmaceuticals in a compact, reconfigurable system. *Science* **352**, 61-67 (2016).
30 C. W. Coley *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566, doi:10.1126/science.aax1566 (2019).
31 A.-C. Bédard *et al.* Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220-1225 (2018).
32 Z. Li *et al.* Robot-Accelerated Perovskite Investigation and Discovery. *Chemistry of Materials* **32**, 5650-5663, doi:10.1021/acs.chemmater.0c01153 (2020).
33 J. Li *et al.* Autonomous discovery of optically active chiral inorganic perovskite nanocrystals through an intelligent cloud lab. **11** (2020).
34 Y. J. Li, A. Savan, A. Kostka, H. S. Stein & A. Ludwig. Accelerated atomic-scale exploration of phase evolution in compositionally complex

materials. *Materials Horizons* **5**, 86-92, doi:10.1039/C7MH00486A (2018).

35 Q. Yan *et al.* Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proceedings of the National Academy of Sciences* **114**, 3040-3043, doi:10.1073/pnas.1619940114 (2017).

36 S. Guerin, B. Hayden, D. W. Hewak & C. Vian. Synthesis and Screening of Phase Change Chalcogenide Thin Film Materials for Data Storage. *ACS Combinatorial Science* **19**, 478-491 (2017).

37 T. Gebhardt, D. Music, T. Takahashi & J. M. Schneider. Combinatorial thin film materials science: from alloy discovery and optimization to alloy design. *Thin Solid Films* **520**, 5491-5499 (2012).

38 P. Nikolaev, D. Hooper, N. Perea-López, M. Terrones & B. Maruyama. Discovery of Wall-Selective Carbon Nanotube Growth Conditions via Automated Experimentation. *ACS Nano* **8**, 10214-10222 (2014).

39 R. Shimizu, S. Kobayashi, Y. Watanabe, Y. Ando & T. Hitosugi. Autonomous materials synthesis by machine learning and robotics. *APL Materials* **8**, 111110 (2020).

40 S. Ament *et al.* Autonomous synthesis of metastable materials. arXiv:2101.07385v1. (2021).

41 R. T. Bell *et al.* Lateral Temperature-Gradient Method for High-Throughput Characterization of Material Processing by Millisecond Laser Annealing. *ACS Combinatorial Science* **18**, 548-558 (2016).

42 B. R. Ortiz, J. M. Adamczyk, K. Gordiz, T. Braden & E. S. Toberer. Towards the high-throughput synthesis of bulk materials: thermoelectric PbTe–PbSe–SnTe–SnSe alloys. *Molecular Systems Design & Engineering* **4**, 407-420 (2019).

43 T. A. Stegk, R. Janssen & G. A. Schneider. High-Throughput Synthesis and Characterization of Bulk Ceramics from Dry Powders. *Journal of Combinatorial Chemistry* **10**, 274-279, doi:10.1021/cc700145q (2008).

44 S. Shuang *et al.* High-throughput automatic batching equipment for solid state ceramic powders. *Review of Scientific Instruments* **90**, 083904 (2019).

45 S. Langner *et al.* Beyond Ternary OPV: High-Throughput Experimentation and Self-Driving Laboratories Optimize Multicomponent Systems. *Advanced Materials* **32**, 1907801 (2020).

46 R. Gautum, S. Venga, F. Ariese & S. Umpathy. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation* **2** (2015).

47 M. H. Modarres *et al.* Neural Network for Nanoscience Scanning Electron Microscope Image Recognition. *Scientific Reports* **7**, 13282 (2017).

48 D. Xue *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nature communications* **7**, 11241 (2016).

49 J. Liu *et al.* Deep convolutional neural networks for Raman spectrum recognition: a unified solution. *Analyst* **142**, 4067-4074 (2017).

50 H. S. Stein, S. Jiao & A. Ludwig. Expediting Combinatorial Data Set Analysis by Combining Human and Algorithmic Analysis. *ACS Combinatorial Science* **19**, 1-8 (2017).

51 F. Oviedo *et al.* Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials* **5** (2019).

52 S. Thienhaus, D. Naujoks, J. Pfetzing-Micklich, D. König & A. Ludwig. Rapid Identification of Areas of Interest in Thin Film Materials Libraries by Combining Electrical, Optical, X-ray Diffraction, and Mechanical High-Throughput Measurements: A Case Study for the System Ni–Al. *ACS Combinatorial Science* **16**, 686-694 (2014).

53 L. Yu, R. S. Kokenyesi, D. A. Keszler & A. Zunger. Inverse Design of High Absorption Thin-Film Photovoltaic Materials. *Advanced Energy Materials* **3**, 43-48 (2013).

54 R. Zarnetta, S. Kneip, C. Somsen & A. Ludwig. High-throughput characterization of mechanical properties of Ti–Ni–Cu shape memory thin films at elevated temperature. *Materials Science and Engineering: A* **528**, 6552-6557 (2011).

55 S. Salomon *et al.* Combinatorial synthesis andhigh-throughput characterization ofthe thin film materials systemCo–Mn–Ge: Composition, structure,and magnetic properties. *Physica status solidi A* **212**, 1969-1974 (2015).

56 Y. Lyu, Y. Liu, T. Cheng & B. Guo. High-throughput characterization methods for lithium batteries. *J Materiomics* **3**, 221-229 (2017).

57 S. M. Moosavi *et al.* Capturing chemical intuition in synthesis of metal-organic frameworks. *Nature communications* **10** (2019).

58 S. Gates-Rector & T. N. Blanton. The Powder Diffraction File: A Quality Materials Characterization Database. *Powder Diffraction* **34**, 352-360 (2019).

59 A. Altomare *et al.* Advances in powder diffraction pattern indexing: N-TREOR09. *Journal of Applied Crystallography* **42**, 768-775 (2009).

60 P. M. d. Wolff. A Simplified Criterion for the Reliability of a Powder Pattern Indexing *Journal of Applied Crystallography* **1**, 108-113 (1968).

61 A. Esmaeili, T. Kamiyama & R. Oishi-Tomiyasu. New functions and graphical user interface attached to powder indexing software CONOGRAPH. *Journal of Applied Crystallography* **50**, 651-659 (2017).

62 R. Oishi-Tomiyasu. Reversed de Wolff figure of merit and its application to powder indexing solutions. *Journal of Applied Crystallography* **46**, 1277-1282 (2013).

63 J.-M. L. Meins, L. M. D. Cranswick & A. L. Bail. Results and conclusions of the internet based ‘‘Search/match round robin 2002’’. *Powder Diffraction* **18**, 106-113 (2003).

64 C. J. Gilmore, G. Barr & J. Paisley. High-throughput powder diffraction. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles. *Journal of Applied Crystallography* **37**, 231-242 (2004).

65 E. Hernández-Rivera, S. P. Coleman & M. A. Tschopp. Using Similarity Metrics to Quantify Differences in High-Throughput Data Sets: Application to X-ray Diffraction Patterns. *ACS Combinatorial Science* **19**, 25-36 (2017).

66 L. A. Baumes, M. Moliner & A. Corma. Design of a Full-Profile-Matching Solution for High-Throughput Analysis of Multiphase Samples Through Powder X-ray Diffraction. *Chemistry - A European Journal* **15**, 4258-4260 (2009).

67 V. Stanev *et al.* Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Computational Materials* **43** (2018).

68 Y. Iwasaki, A. G. Kusne & I. Takeuchi. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Computational Materials* **4** (2017).

69 W. B. Park *et al.* Classification of crystal structure using a convolutional neural network. *IUCrJ* **4**, 486-494 (2017).

70 P. M. Vecsei, K. Choo, J. Chang & T. Neupert. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Physical Review B* **99**, 245120 (2019).

71 H. Wang *et al.* Rapid Identification of X-ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **60**, 2004-2011 (2020).

72 P. M. Maffettone *et al.* Crystallography companion agent for high-throughput materials discovery. arXiv:2008.00283v1. (2020).

73 J.-W. Lee, W. B. Park, J. H. Lee, S. P. Singh & K.-S. Sohn. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nature communications* **11** (2020).

74 S. S. Khan & M. G. Madden. New similarity metrics for Raman spectroscopy. *Chemometrics and Intelligent Laboratory Systems* **114**, 99-108 (2012).

75 C. Li *et al.* Rapid Bayesian optimisation for synthesis of short polymer fiber materials. *Scientific Reports* **7**, 5683 (2017).

76 H. Wahab *et al.* Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ Raman analysis. *Carbon* **167**, 609-619 (2020).

77 M. A. Al-Kheder, C. Pezeshki, J. L. McHale & F. J. Knorr. Quality classification via Raman identification and SEM analysis of carbon nanotube bundles using artificial neural networks. *Nanotechnology* **18**, 335703 (2007).

78 M. Ziatdinov *et al.* Deep Learning of Atomically Resolved Scanning Transmission Electron Microscopy Images: Chemical Identification and Tracking Local Transformations. *ACS Nano* **11**, 12742-12752 (2017).

79 W. Li, K. G. Field & D. Morgan. Automated defect analysis in electron microscopic images. *npj Computational Materials* **4** (2018).

80 A. Maksov *et al.* Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS2. *npj Computational Materials* **5** (2018).

81 S. Madireddy *et al.* Phase Segmentation in Atom-Probe Tomography Using Deep Learning-Based Edge Detection. *Scientific Reports* **9**, 20140 (2019).

82 J. H. Montoya *et al.* Autonomous intelligent agents for accelerated materials discovery. *Chemical Science* **11**, 8517-8532 (2020).

83 Z. Zhou, X. Li & R. N. Zare. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Central Science* **3**, 1337-1344 (2017).

84 L. Porwol *et al.* An Autonomous Chemical Robot Discovers the Rules of Inorganic Coordination Chemistry without Prior Knowledge. *Angewandte Chemie* **59**, 11256-11261 (2020).

85 D. A. Cohn, Z. Ghahramani & M. I. Jordan. Active learning with statistical methods. *Journal of Artificial Intelligence* **4**, 129-145 (1996).

86 E. J. Braham, R. D. Davidson, M. Al-Hashimi, R. Arróyave & S. Banerjee. Navigating the design space of inorganic materials synthesis using statistical methods and machine learning. *Dalton Transactions* **49**, 11480 (2020).

87 W. F. Maier, K. Stöwe & S. Sieg. Combinatorial and High-Throughput Materials Science. *Angewandte Chemie* **46**, 6016-6067 (2007).

88 H. Koinuma & I. Takeuchi. Combinatorial solid-state chemistry of inorganic materials. *Nature Materials* **3**, 429-438 (2004).

89 M.-X. Li *et al.* High-temperature bulk metallic glasses developed by combinatorial methods. *Nature* **569**, 99-103 (2019).

90 S. A. Kube *et al.* Combinatorial study of thermal stability in ternary nanocrystalline alloys. *Acta Materiala* **188**, 40-48 (2020).

91 R. Zarnetta *et al.* Identification of Quaternary Shape Memory Alloys withNear-Zero Thermal Hysteresis and Unprecedented Functional Stability. *Advanced Functional Materials* **20**, 1917-1923 (2010).

92 A. Marshal *et al.* Combinatorial evaluation of phase formation and magnetic properties of FeMnCoCrAl high entropy alloy thin flm library. *Scientific Reports* **9**, 7864 (2019).

93 S. Thrun. in *Handbook of Brain Science and Neural Networks* 381-384 (1995).

94 A. Belsky, M. Hellenbrandt, V. L. Karen & P. Luksch. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica* **B58**, 364-369 (2002).

95 J. Schmidt, M. R. G. Marques, S. Botti & M. A. L. Marques. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **5** (2019).

96 D. Morgan & R. Jacobs. Opportunities and Challenges for Machine Learning in Materials Science. *Annual Review of Materials Research* **50**, 71-103 (2020).

97 L. Himanen, A. Geurts, A. S. Foster & P. Rinke. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* **6**, 1900808 (2019).

98 E. Kim, K. Huang, S. Jegelka & E. Olivetti. Virtual screening of inorganic materials synthesis parameters with deep learning. *npj Computational Materials* **3** (2017).

99 E. Kim *et al.* Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials* **29**, 9436-9444 (2017).

100 P. Raccuglia *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73-76 (2016).

101 M. M. Flores-Leonar *et al.* Materials Acceleration Platforms: On the way to autonomous experimentation. *Current Opinion in Green and Sustainable Chemistry* **25**, 100370 (2020).

102 T. Lookman, P. V. Balachandran, D. Xue & R. Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials* **5** (2019).

103 A. G. Kusne *et al.* On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature communications* **11** (2020).

104 J. Li *et al.* AI Applications through the Whole Life Cycle of Material Discovery. *Matter* **3**, 393-432 (2020).

105 C. Audet & W. Hare. *Derivative-free and Blackbox Optimization*. (Springer, Cham, 2017).

106 A. Costa & G. Nannicini. RBFOpt: an open-source library for black-box optimization with costly function evaluations. *Mathematical Programming Computation* **10**, 597-629 (2018).

107 D. R. Jones, M. Schonlau & W. J. Welch. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization* **13**, 455-492 (1998).

108 K. S. Tang, K. F. Man, S. Kwong & Q. He. Genetic algorithms and their applications. *IEEE Signal Processing Magazine* **13**, 22-37 (1996).

109 D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. (Addison Wesley, 1989).

110 M. Holena. in *High-Throughput Screening in Chemical Catalysis: Technologies, Strategies and Applications* (eds A. Hagemeyer, P. Strasser, & A. F. Volpe Jr.) 153-174 (Wiley-VCH, 2006).

111 U. Rodemerck, M. Baerns, M. Holena & D. Wolf. Application of a genetic algorithm and a neural network for the discovery and optimization of new solid catalytic materials. *Applied Surface Science* **223**, 168-174 (2004).

112 Z. Ma & F. Zaera. in *Encyclopedia of Inorganic Chemistry* (John Wiley, 2006).

113 D. Wolf, O. V. Buyevskaya & M. Baerns. An evolutionary approach in the combinatorial selection and optimization of catalytic materials. *Applied Catalysis: General* **200**, 63-77 (2000).

114 K. W. Kuntz, M. L. Snapper & A. H. Hoveyda. Combinatorial catalyst discovery. *Current Opinion in Chemical Biology* **3**, 313-319 (1999).

115 A. Corma *et al.* Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (softcomputing techniques). *Journal of Catalysis* **229**, 513-524 (2005).

116 K. McCullough, T. Williams, K. Mingle, P. Jamshidi & J. Lauterbach. High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery. *Physical Chemistry Chemical Physics* **22**, 11174-11196 (2020).

117 J. T. Alander. On optimal population size of genetic algorithms. *Proceedings of CompEuro 1992 Computer Systems and Software Engineering* **1**, 65-70 (1992).

118 W. Huyer & A. Neumaier. SNOBFIT — Stable Noisy Optimization by Branch and Fit. *ACM Transactions on Mathematical Software* **35**, 1-25 (2008).

119 C. Mateos, M. J. Nieves-Remacha & J. A. Rincón. Automated platforms for self-optimization in flow. *Reaction Chemistry & Engineering* **4**, 1536 (2019).

120 A. D. Clayton *et al.* Algorithms for the self-optimisation of chemical reactions. *Reaction Chemistry & Engineering* **4**, 1545 (2019).

121 N. Cherkasov, Y. Bai, A. J. Expósito & E. V. Rebrov. OpenFlowChem – a platform for quick, robust and flexible automation and self-optimisation of flow chemistry. *Reaction Chemistry & Engineering* **3**, 769 (2018).

122 L. M. Rios & N. V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization* **56**, 1247-1293 (2013).

123 A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis & T. Lookman. Multi-objective Optimization for Materials Discovery via Adaptive Design. *Scientific Reports* **8** (2018).

124 B. Shahriari, K. Swersky, Z. Wang, R. P. Adams & N. d. Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE* **104** (2015).

125 P. I. Frazier & J. Wang. in *Information Science for Materials Discovery and Design* (eds T. Lookman, F. Alexander, & K. Rajan) 45-76 (Springer, 2015).

126 K. Higgins, S. M. Valleti, M. Ziatdinov, S. V. Kalinin & M. Ahmadi. Chemical Robotics Enabled Exploration of Stability in Multicomponent Lead Halide Perovskites via Machine Learning. *ACS Energy Letters* **5**, 3426-3436 (2020).

127 J. M. Hernández-Lobato, M. W. Hoffman & Z. Ghahramani. Predictive Entropy Search for Efficient Global Optimization of Black-Box Functions. *NIPS*, 918-926 (2014).

128 D. H. Wolpert & W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 67-82 (1997).

129 Z. Wang & S. Jegelka. Max-value entropy search for efficient Bayesian optimization. arXiv:1703.01968v3. (2018).
130 E. Brochu, V. M. Cora & N. d. Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. arxiv:1012.2599v1. (2010).
131 J. Wu & P. Frazier. The Parallel Knowledge Gradient Method for Batch Bayesian Optimization. arXiv:1606.04414v4. (2018).
132 J. Snoek *et al.* Scalable Bayesian Optimization using Deep Neural Networks. *International Conference on Machine Learning*, 2171-2180 (2015).
133 L. von Rueden *et al.* Informed Machine Learning--A Taxonomy and Survey of Integrating Knowledge into Learning Systems. *arXiv preprint arXiv:1903.12394* (2019).
134 S. Sun *et al.* A data fusion approach to optimize compositional stability of halide perovskites. *Matter* (2021).
135 S. Ariafar *et al.* Weighting is Worth the Wait: Bayesian Optimization with Importance Sampling. arXiv:2002.09927v1. (2020).
136 Z. Ren *et al.* Embedding physics domain knowledge into a Bayesian network enables layer-by-layer process innovation for photovoltaics. *npj Computational Materials* **6**, 1-9 (2020).
137 R. D. King *et al.* The automation of science. *Science* **324**, 85-89 (2009).
138 M. H. Segler, M. Preuss & M. P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604-610 (2018).
139 R. D. King *et al.* Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247-252 (2004).
140 L. A. Klein. *Sensor and data fusion: a tool for information assessment and decision making*. Vol. 138 (SPIE press, 2004).
141 W. Sun *et al.* The thermodynamic scale of inorganic crystalline metastability. *Science advances* **2**, e1600225 (2016).
142 B. J. Shields *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89-96 (2021).
143 O. Mitsunobu & M. Yamada. Preparation of esters of carboxylic and phosphoric acid via quaternary phosphonium salts. *Bulleltin of the Chemical Society of Japan* **40**, 2380-2382 (1967).
144 K. E. Whelan & R. D. King. Using a logical model to predict the growth of yeast. *BMC Bioinformatics* **9** (2008).
145 M. Kanehisa & S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27-30 (2000).
146 I. M. Keseler *et al.* EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic acids research* **33**, D334-D337 (2005).
147 C. W. Coley, W. H. Green & K. F. Jensen. Machine learning in computer-aided synthesis planning. *Accounts of chemical research* **51**, 1281-1289 (2018).

148 V. Dragone, V. Sans, A. B. Henson, J. M. Granda & L. Cronin. An autonomous organic reaction search engine for chemical reactivity. *Nature communications* **8**, 15733 (2017).

149 S. Feng & R. Xu. New Materials in Hydrothermal Synthesis. *Accounts of chemical research* **34**, 239-247 (2001).

150 Flex Autoplant from Chemspeed. https://www.chemspeed.com/flex-autoplant/.

151 A. F. d. Almeida, R. Moreira & T. Rodrigues. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry* **3**, 589-604 (2019).

152 R. McClain *et al.* Mechanistic insight of KBiQ2 (Q ¼ S, Se) using panoramic synthesis towards synthesis-by-design. *Chemical Science* **12**, 1378 (2021).

153 P. K. Todd, A. M. M. Smith & J. R. Neilson. Yttrium Manganese Oxide Phase Stability and Selectivity Using Lithium Carbonate Assisted Metathesis Reactions. *Inorganic Chemistry* **58**, 15166-15174 (2019).

154 D. P. Shoemaker *et al.* In situ studies of a platform for metastable inorganic crystal growth and materials discovery. *Proceedings of the National Academy of Science of the United States of America* **111**, 10922-10927 (2014).

155 A. Miura *et al.* Sequential pairwise reactions dictate phase evolution in the solid-state synthesis of multicomponent ceramics. arXiv:2009.10896. (2020).

156 M. J. McDermott, S. S. Dwarakanth & K. A. Persson. A graph-based network for predicting chemical reaction pathways in solid-state materials synthesis. *Preprint.* (2020).

157 M. Aykol, J. H. Montoya & J. S. Hummelshøj. Rational Solid-State Synthesis Routes for Inorganic Materials. *ChemRxiv* (2021).

158 Y. Rakita *et al.* Active Reaction Control of Cu Redox State Based on Real-Time Feedback from In Situ Synchrotron Measurements. *Journal of the American Chemical Society* **142**, 18758-18762 (2020).

159 M. R. Cosby *et al.* Salt Effects on Li-Ion Exchange Kinetics of Na2Mg2P3O9N: Systematic In Situ Synchrotron Diffraction Studies. *The Journal of Physical Chemistry C* **124**, 6522-6527 (2020).

160 O. Kononova *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Scientific data* **6** (2019).

161 O. Kononova *et al.* Opportunities and challenges of text mining in aterials research. *iScience* **24**, 102155 (2021).

162 S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan & L. Cronin. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101-108 (2020).

163 E. Kim *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Scientific data* **4** (2017).

164 C. W. Coley, N. S. Eyke & K. F. Jensen. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angewandte Chemie* **59**, 2-25 (2020).

165 P. Langley. The computational support of scientific discovery. *International Journal of Human-Computer Studies* **53**, 393-410 (2000).