In Pursuit of the Exceptional: Research Directions for Machine Learning in Chemical and Materials Science

Joshua Schrier[1], Alexander J. Norquist,[2] Tonio Buonassisi,[3] Jakoah Brgoch[4]

[1] Department of Chemistry, Fordham University, The Bronx, NY 10458, USA

[2] Department of Chemistry, Haverford College, Haverford PA, 19041, USA

[3] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4] Department of Chemistry and Texas Center for Superconductivity, University of Houston, Houston, TX, 77204, USA

Abstract

Exceptional molecules and materials with one (or more) extraordinary properties are both technologically valuable and fundamentally interesting because they often involve new physical phenomena or new compositions that defy expectations. Historically, exceptionality has been achieved through serendipity, but recently, machine learning (ML) and automated experimentation have been widely proposed to accelerate target identification and synthesis planning. In this Perspective, we argue that the data-driven methods commonly used today are well-suited for optimization but not for realizing new exceptional materials or molecules. Finding such outliers should be possible using ML, but only by shifting away from using traditional ML approaches that tweak the composition, crystal structure, or reaction pathway. We highlight case studies of high-T_c superconductors and superhard materials to demonstrate the challenges of ML-guided discovery and discuss the limitations of automation for this task. We then provide six recommendations for the development of ML methods capable of exceptional materials discovery: (i) Avoid the tyranny of the middle and focus on extrema; (ii) When data is limited, qualitative predictions that provide direction are more valuable than interpolative accuracy; (iii) Sample what can be made and how to make it, and defer optimization; (iv) Create room (and look) for the unexpected while pursuing your goal; (v) Try to fill-in-the-blanks of input and output space; (vi) Do not confuse human understanding with model interpretability. We conclude with a description of how these recommendations can be integrated into automated discovery workflows that should enable the discovery of exceptional molecules and materials.

I. Introduction

Machine learning (ML) is contributing to many areas of chemistry and materials research, as diverse as solar cells,[1] photoresist,[2] high-entropy alloys,[3] drug design[4,5] and formulation[6] discovery, and biomedical polymers.[7] Many introductory texts[8–10] and review articles[11–15] provide tutorials and explications of applications of ML to chemistry and materials. Significant effort has been invested in identifying compositions, crystal structures, reaction conditions, etc., to yield highly desirable materials, molecules, or unique physical properties. These efforts are valuable both scientifically and industrially, and automating this process increases the efficiency and productivity of research and development efforts. However, these applications have been mainly demonstrated in the context of incremental improvement and optimization. *Incremental* does not mean *easy*—ML optimizations are often in high-dimensional spaces that would have otherwise required months or years of traditional experimentation to achieve the same results.[16–19] In a mathematical sense, the underlying property-response landscapes are well-behaved (i.e., the surfaces are approximately smooth, convex,[20] and elementary[21]), so local information gathered from stepwise changes can find optima efficiently. This makes these problems computationally "easy". However, as we argue below, transformative discoveries are seldom achieved by this type of incremental approach.

In this perspective, we suggest that there are fundamental limitations hindering the application of ML today for the discovery of *exceptional materials* that shift the research paradigm (in the Kuhnian sense[22]) while addressing pressing technological and economic needs for human health, energy, and the environment. We highlight some current state-of-the-art examples in ML, iterative-optimization, and high-throughput/autonomous experimentation approaches. We also focus on limitations of using these methods with regards to exceptional materials discovery by

considering historical challenges in high-T_c superconductor and superhard materials discovery, and the extent to which existing ML methods have contributed to these efforts. We then provide six recommended research directions for ML that can address this challenge based on our survey of the field. Finally, we conclude with a vision of what the materials research process would look like if these research directions were successful.

II. The Challenge of the Exceptional

II.A. What is Exceptional?

We define an *exceptional material or molecule* as one that achieves such extraordinary properties through unexpected means, like new physics or chemistry. These are *black swan events* — unpredicted surprises that have a significant effect on the field but are only rationalized only after the first observation.[23] For example, the discovery of high-T_c cuprate superconductors rejected the conventional wisdom of condensed matter physics and rapidly transformed the field (*vide infra*). Similarly, organic chemistry has had numerous scientific discoveries that have gone against deep-seated textbook notions, transforming our molecular control and greatly enhancing our synthetic toolbox.[24]

Nevertheless, materials with extraordinarily low or high values of a single property may still be insufficient to enable transformative technologies. Instead, what is truly exceptional is a constellation of properties, many of which are mutually exclusive.[15,25] Tradeoffs between desirable properties are common in many fields. These may be purely empirical trends observed within a materials class, such as Ashby plots (solid mechanics) or Robeson plots (in membrane separation materials). Alternatively, they may be first-order approximations to underlying rigorous theoretical relationships, such as Wiedemann-Franz (proportionality of electrical and thermal conductivity)

or Dulong-Petit (inverse relationship between heat capacity and molar mass). Increasingly, technology demands materials that can withstand coupled extremes, such as simultaneous mechanical, thermal, radiation, and corrosive attack required for next-generation fission and fusion reactors[15,26] or simultaneous high photoconversion efficiency and mechanical durability for photovoltaics.[27] An exceptional material may be surprising by displaying properties that have not yet been observed simultaneously. Alternatively, an exceptional material may be surprising by violating a contraindicated physical relationship (e.g., in the case of thermoelectric materials, finding materials with high electrical conductivity but low thermal conductivity, in violation of the Wiedemann-Franz law), for example, by taking advantage of nonlinear effects or a careful balance of contradictory influences. This is a distinct problem from multi-objective optimization,[28–30] discussed in Section III.C.

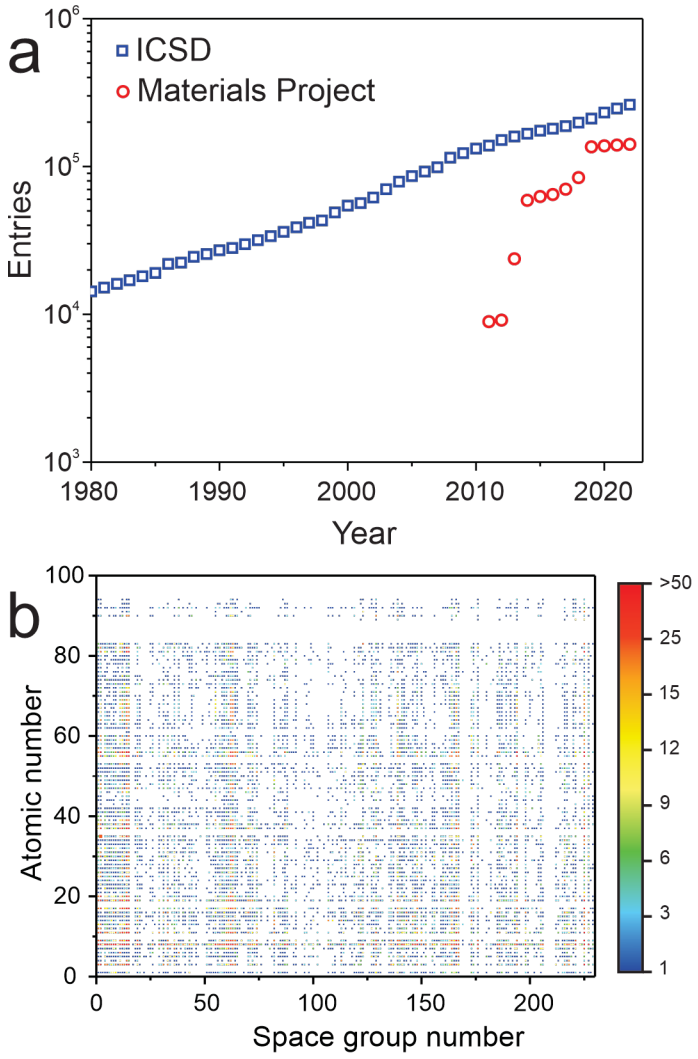

FIGURE 1: (a) Structure entries in the ICSD and Materials Project as a function of time, plotted on a logarithmic scale; (b) Number of compounds in the Materials Project (color) as a function of both chemistry (atomic number) and structure (space group number). Adapted from Ref. [31].

Genuine surprise would not be possible or necessary if we already had an adequate sample of all possible materials. However, an empirical analysis suggests that humans have barely scratched the surface of possible compositions. If one considers only stoichiometric quaternary solid-state inorganic compounds satisfying conservative valency and electronegatively constraints, there are approximately 10^{10} compositions,[32] many orders of magnitude below the 10^5 compounds in the entire Inorganic Crystal Structure Database (ICSD). Similar estimates exist for the chemical space of synthesizable organic molecules.[33] The reported number of new structures deposited in the ICSD shows exponential growth (Figure 1a),[31] and thermodynamic stability network calculations indicating an increasing trend in the discovery rate of new materials.[34] Nevertheless, the compositional variations are certainly mapped unevenly. Computational datasets, such as the Materials Project database, exhibit wide disparities in the prevalence of certain elements (Figure 1b).[31] A significant portion (~80%) of these materials are also found to fall above their respective (0 K) convex hulls, suggesting there is some level of metastability to their thermodynamic favorability, further expanding the complexity of materials discovery.[35] However, the ability to study metastability in computational datasets is limited by dataset biases in the distribution of formation energies for different structure types.[36] There remains plenty of room to discover new materials and molecules, and we are far from the regime of pure interpolation.

II. B. Why is Finding an Exceptional Compound Difficult?

Finding an exceptional compound is intrinsically a low probability event, as the compositions and combination of synthesis and processing conditions needed to produce them are rare and unique. But rarity alone is not the problem. Consider a golf course: The probability of a

randomly placed ball occupying the hole is small, yet golfers regularly guide the ball on the green to the hole with (ideally) few attempts by taking advantage of the landscape and "reading" its many properties. Similarly, it is easy to solve research problems if there exists a clear gradient towards the goal (by analogy, this may arise from the inherent topography which causes balls to roll down hills of the landscape or the golfer's knowledge of the map of the course). It is harder if there are many traps where these gradient-based heuristics fail. More formally, mathematicians have devised many ways of characterizing the ease and difficulty of finding optima on high-dimensional response surfaces. For example, cases where the inputs are continuous can be characterized in terms of *smoothness* (the number of continuous derivatives a function has over its domain) and *convexity* (continuous functions where values at the midpoint of every interval do not exceed the values of the function at its endpoint).[20] Cases where the inputs are discrete can be characterized in terms of *elementariness* (those which can be realized as an eigenvector of the Laplacian of the neighborhood diagraph).[21] Whether the inputs are discrete or continuous, the underlying idea is to characterize functions for which local information gathered from stepwise changes can find optima efficiently.[20,37] In general, many practical algorithms are also capable of efficiently finding solutions even when the response surface only approximately obey these criteria.

Materials discovery can be both easy and hard, depending on the problem. Empirically, it has been observed that many successful materials ML problems are approximately smooth and convex response surfaces, with a broad basin of attraction towards a few local optima,[38] like the schematic example plotted in Figure 2a. Thus, it is not surprising that ML-based approaches for representing the landscape can be successful and gradient descent is an efficient strategy. In contrast, exceptional materials are often comprised of much harder "needle in a haystack" problems,[39] where the response surface behaves as shown in Figure 2b. The response function is

no longer smooth, and any approximate information about a local environment is no longer a good guide to the behavior of new candidates. Whereas a mathematician would apply a preconditioner to transform the problem into a more suitable form, experimental scientists typically lack prior knowledge about the nature of the response function, and consequently are limited in what types of *a priori* transformations can be applied. Acquiring more data about the system (either by physics-based simulation or by high-throughput experimentation) or improving the nature of the search process is one of the few solutions.

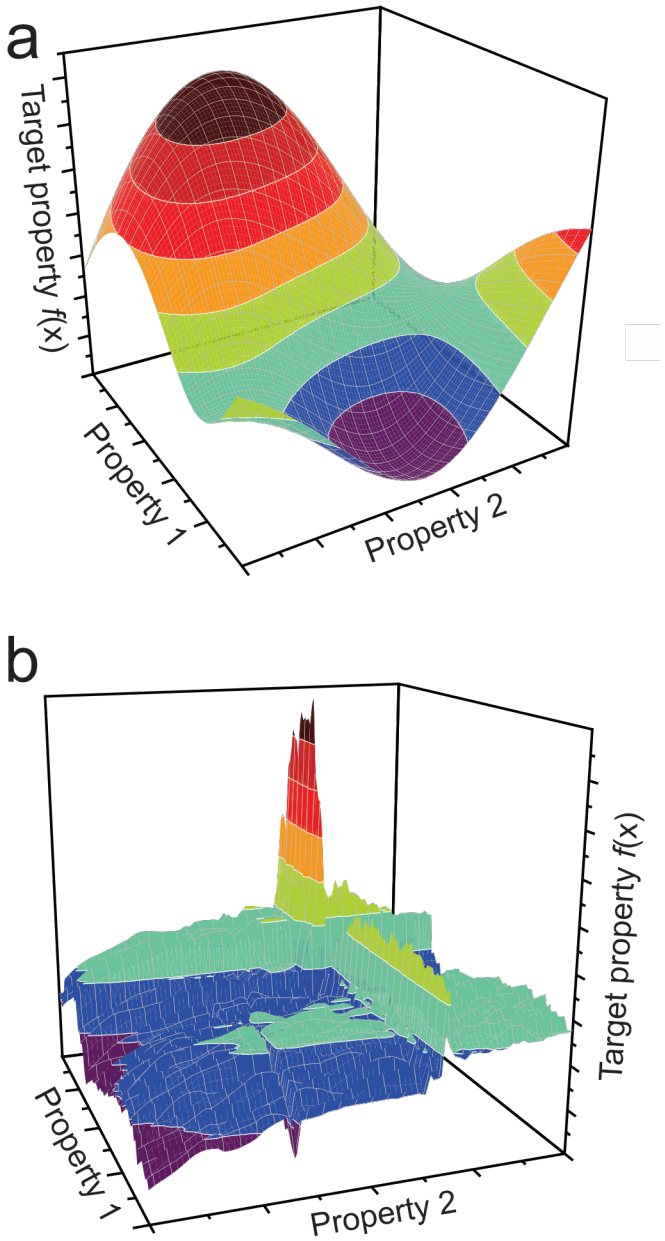

Figure 2: Iterative versus exceptional materials. (a) Previous work has focused on optimizations on smooth, convex response surfaces; (b) Exceptional material properties are often characterized by very sharp discontinuities in as composition and reaction conditions are changed. Adapted from Ref. [38].

III. The State of Current Machine Learning Approaches

III.A. Computational Screening and Early ML

Multi-agency funding efforts like the Materials Genome Initiative (MGI),[6] and similar efforts worldwide,[41] were premised on combining physics-based computation, data resources, and high-throughput experimentation to provide more data and accelerate the discovery of new materials. Many of the highest profile early efforts based on high-throughput density functional theory (DFT) calculations leading to databases such as the Materials Project, AFLOW, OQMD, etc. As the information contained in these repositories grew and the desire to expedite this process took over, researchers began to augment DFT with ML approaches. They started by first performing simple regression or classification predictions and used these results to screen known crystal structure databases looking for materials with superb properties.[42] The efforts have since expanded dramatically using experimental datasets, autonomous experimentation, multitask and transfer learning, among numerous other approaches. A tremendous number of predictions have been made using this procedure, although there have been far fewer experimental validations of these models.[42] Nevertheless, in nearly every example, the model's predictions are modest improvements of known systems rather than new state-of-the-art, transformative materials.

III.B. ML as an Experimental Optimization Tool

Many of the current demonstrations of ML for chemistry and materials are essentially *optimizations* of the composition, reaction conditions, and processing conditions to maximize or minimize a desired property. ML is used as a low-cost proxy for experimental input-output relationships. For example, the design of catalysts for chemical reactions has involved performing density functional theory (DFT) calculations to determine the optimal catalyst composition and

reaction conditions. ML models can be trained on this data to accelerate the screening process, thereby reducing the time it takes to map the response surface and achieve the desired result.[43] Incorporating model uncertainty with the predicted outcome enables guiding the experimental process in an algorithmic way to achieve the desired experimental goal. One classic illustrative example comes from the seminal 2016 work of Nikolaev et al. on ML-optimized carbon nanotube growth in an autonomous system.[44] A random forest model trained on a small initial dataset served as a proxy for the dependence of observed nanotube growth rate as a function laser heating and the partial pressures of four gases. Active learning methods were used to sample uncertain new points, and the algorithm was then employed to select the optimal set of input parameters to rapidly achieve a desired growth rate. Other illustrative ML-enhanced materials optimization examples include nanocrystal growth and optical properties in a microfluidic system,[45] mechanical properties of 3d-printed structures,[46] crystal growth conditions,[47–49] and halide alloy stability[19,50], and superconductivity.[51] See Refs. [42] & [52] for a more comprehensive review.

Limitations of data-driven strategies have been noted in the literature to some extent. Many of these papers have been premised on the argument that more data is require, or more computationally intensive approaches are needed to generate higher-quality training data.[53] Algorithmic performance can also depend on the initial dataset (the "cold start" problem), and available datasets often exhibit sampling biases.[54] This problem can be partially mitigated by adding additional constraints to maximize the explored input space[49] or by incorporating human expertise in the loop.[55] While previous research articles have benchmarked computational methods and metrics for this task,[56,57] and a recent perspective discussed types of machine-learning guided iterative experimentation towards this goal,[15] a more critical view of the field is that regardless of

the accuracy produced by these methods, they will not generate the materials necessary to enable paradigm shifts.

ML-based organic (retro)synthesis prediction and planning face similar issues.[2,58] There is a tremendous power (and computational complexity) associated with selecting a sequence of known reactions into a new arrangement. This presents an immense combinatorial challenge where ML-derived heuristics can make the problem tractable,[59,60] with recent reviews discussing these efforts.[61,62] Although it has been suggested that deep-learning-based template-free methods can propose genuinely inventive new reactions,[63] performance can be poor outside the training set (even for undergraduate textbook reactions) and often reflects the most common reactions in the training set rather than finding optimal reactions.[64] Some ML researchers characterize large language models (LLMs) as "stochastic parrots" because of their tendency to generate outputs that merely have the same statistical local structure as the training corpus (and thus perpetuate or amplify training set bias) without incorporating long-term structure or meaning.[65] Regardless, empirical evidence suggests that suitably trained LLMs can learn meaningful internal representations of the problem,[66] and the nature of the formal grammatical expressiveness of different model types is an area of active research.[67] Within the context of chemistry, there is evidence that transformer-based LLMs models learn relevant atom mapping rules, implying that the learned representations are physically meaningful.[63] Additionally, general purpose LLMs are surprisingly effective at predicting molecular and material properties with only small amounts of example data.[68] Yet there remains the problem of optimizing the reaction conditions and stoichiometries. Again, the tremendous technical challenges and practical benefits of this are immense—as demonstrated by exciting recent work on the optimization of heteroaryl Suzuki-

Miyaura reactions[69] and reviewed more comprehensively in Ref. [70]. But again, this is in the domain of incremental optimization.

III.C High-throughput Experimentation and Autonomous Operation to Discover Exceptional Compounds

The importance of high throughput experimentation (HTE) for data generation that will enable materials discovery has a long history[71,72] The importance of data management, informatics, and ML for HTE has also been discussed extensively.[73,74] Increasingly, this has taken the form of closed-loop autonomous research systems or "self-driving laboratories[52,75–77] like the ARES system[44] mentioned above. An autonomous system performs the entire workflow of design, synthesis, characterization, and optimization under algorithmic control. Similar types of design/build/test/learn cycles exist in many scientific domains besides materials, such as protein design, synthetic biology, and drug discovery.[78] Each workflow step can be accelerated by incorporating ML tools—including inverse design, accelerated synthesis, characterization, and optimization. Indeed, characterization—rather than synthesis and processing—is often a bottleneck in this process, and ML can accelerate the characterization process. There are many technical challenges and opportunities related to miniaturization, continuous versus batch operation, etc. that will be discussed in Section VI. Nevertheless, recent demonstrations of autonomous materials research include metallic thin films,[79] wear-resistant metallic glasses,[80] organic laser materials,[81] acid generators for photoresists relevant to semiconductor device fabrication,[2] multi-modal materials characterization,[82] and reversible addition–fragmentation chain transfer (RAFT) polymerization.[83] For comprehensive recent reviews on autonomous systems for materials, see Refs. [52,84–86], and for a recent perspective on the current state of the art

of these systems for exceptional materials see Ref. [15] while organic synthesis is discussed in Refs. [61] & [87].)

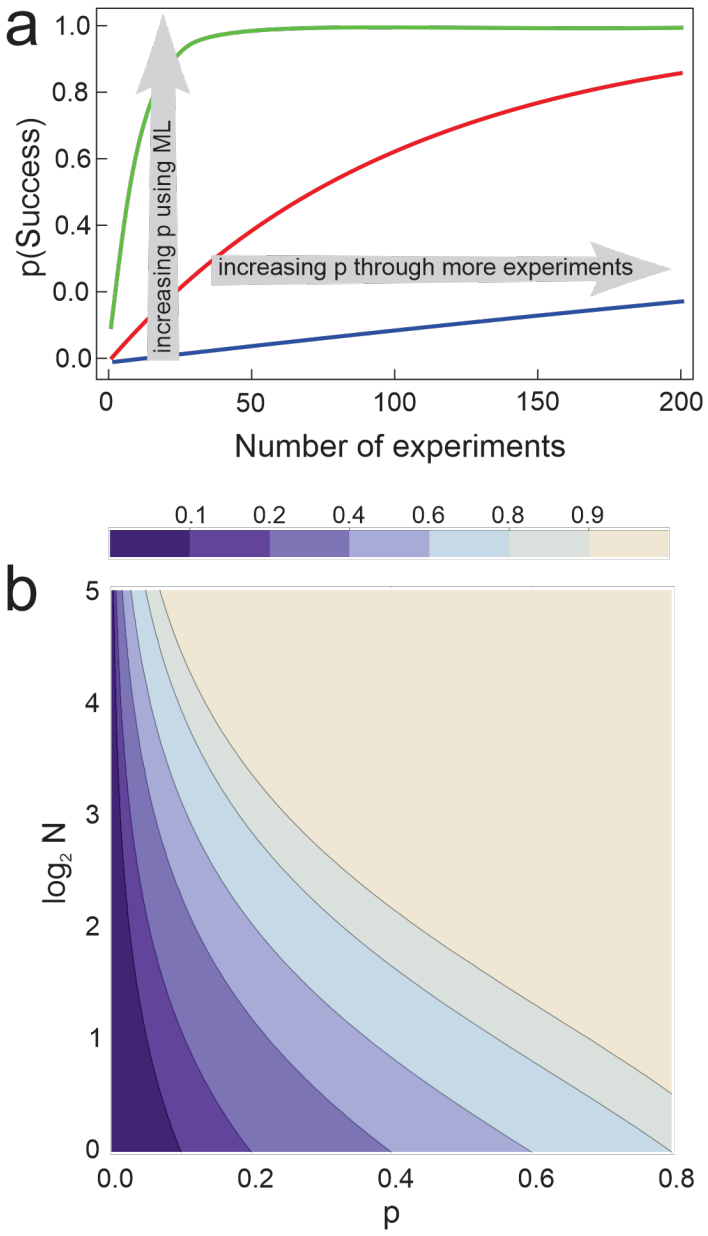

Figure 3: (a) The probability of at least one success in an experimental campaign can be increased by using ML to increase the probability that each experiment is successful, (increasing *p*), or by making more attempts (increasing *N*); (b) Contours showing the

probability of at least one success, as a function of changing p and N. Increasing p and N have a synergistic effect, but a large value of one can compensate for a smaller value of the other.

There are many synergies between ML and HTE, and one can gain some intuition from considering the materials discovery process with a simple statistical model, depicted schematically in Figure 3a. The probability that at least one successful material results from an ensemble of N independent trials, each of which has a success probability p, is $1-(1-p)^N$. HTE increases N and ML increase p. As depicted in Figure 3b, these have a complementary effect on the overall probability of success and one can compensate for a lower value of p by increasing N and vice versa. HTE generalizes this in several ways: First, p is no longer constant, but ideally increases as a function of time as new data is acquired to improve the model, i.e., dp/dt is positive. Second, removing intermediaries between the model and the process reduces the time delay between acquiring data for a model and using that improved model to acquire the next experiment; this is analogous to compounding interest more frequently. The advantages of active learning depend sensitively on the type of problem; at best it may require only a logarithm of the number of experiments required by random sampling but at worst may require the full number of sample points.[88] Empirical materials science studies have observed that poorly implemented active learning can decrease p.[89] In addition to increasing the chances of success, there are other benefits for ML model performance that result from increasing the volume and quality of experimental data. Data and metadata from automated experiments is "born digital" which facilitates its use, reporting, and sharing. By eliminating (unrecorded) human variations, automated processing can potentially improve the reproducibility of experiments, thus increasing the signal-to-noise ratio in the dataset. Such unrecorded unintentional variations in background conditions are also minimized

by performing more experiments in a smaller span of time. These are harder to quantify, but only increase the value of increasing both *N* and *p*, by use of automation and ML.

Although HTE is an essential enabling technology for discovering exceptional materials, it is not enough. A historical analogy is provided by combinatorial chemistry in drug discovery.[90] The lack of clinical successes from initial high-throughput synthesis in the 1980s, suggests that merely increasing *N* is insufficient. Incorporating modelling in the 1990s—specifically computational chemistry methods and informatics—increased *p* and ultimately lead to success. Ultimately, *p* depends on how the problem is framed and will determine where we look. In Section V we describe paths towards developing ML models that increase *p* for exceptional materials, rather than being limited to local optimizations.

III.D. Limitations of Pareto-front Multiobjective Optimization Strategies

Most efforts to apply ML for novel materials discovery have primarily focused on predicting a single property value or reactivity, but transformative applications require a balance of contraindicated responses. This has led researchers to pursue multiobjective optimization algorithms to find the best solution that balances multiple objectives, which may conflict with each other. In this type of non-dominated solution, the goal is not to find a single best solution but rather to identify a set of optimal solutions, The optimization of more than one mutually exclusive property is often described in terms of a *Pareto frontier*.[15,25] (& Ref. [20] pp. 177-184), depicted schematically in Figure 4. With each iteration of materials discovery, the Pareto front edges forward yielding materials with an optimal balance of often contraindicated properties. These tools have been applied recently across materials science ranging from solid-state battery electrolytes[91] to magnetic high-entropy alloys[3] to additive manufacturing[29] to polymer design.[92] (Also noted are

recent reviews of multiobjective optimization for organic molecules[93] and chemical reaction optimization.[70])

Pareto front optimization is a useful tool for decision-making. Methodologies like Chimera[28] have been proposed to optimize materials, even in the case of constrained design spaces with little information on the surface of the objective function. Active-learning Pareto front optimization methods, such as ε-PAL,[92] can be used to guide automated experimental procedure could optimally identify Pareto points in fewer evaluations. The Pareto front has been demonstrated to establish the optimal set of synthetic conditions that yield metallic films with excellent conductivity at lower processing temperatures.[79] A self-driving laboratory varied different reaction conditions for combustion synthesis (fuel source, fuel-to-oxidizer ratio, precursor solution concentration, and annealing temperature) to simultaneously maximize the film's conductivity and minimize the combustion temperature, by using a differential expected hypervolume improvement (qEHVI) algorithm.[94] The normalized hypervolumes increase smoothly to their maximum as the property response is explored, indicating stepwise advances. Pareto front-based strategies yield stepwise, continual advances of an objective that balance tradeoffs, which does not lend these methods to the necessary "leaps-and-bounds" advances required for transformational discoveries.

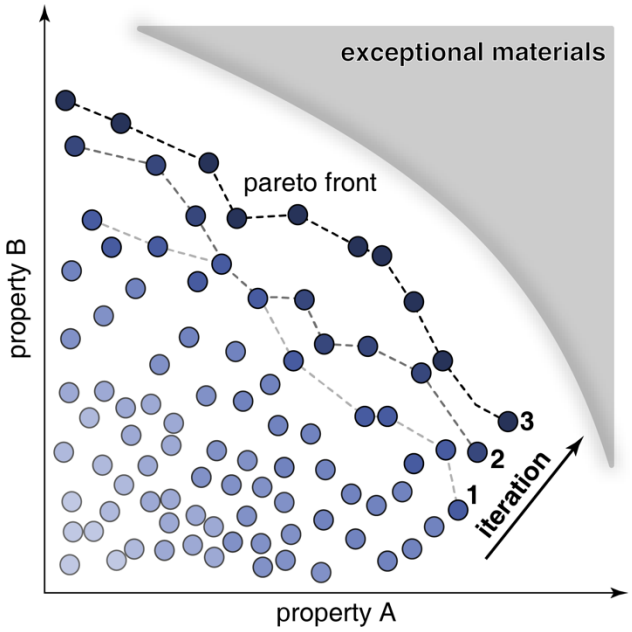

FIGURE 4: The Pareto front is a plot of optimal solutions for a multi-objective optimization problem. The *x*-axis represents one objective, the *y*-axis represents another objective. Points on the plot are Pareto-optimal solutions, with no other solutions better in both objectives. The Pareto front connects all these solutions and shows trade-offs between objectives. Points above the front are dominated by at least one other solution, while points on the front are non-dominated. The Pareto front moves forward with each iteration making it a useful tool for decision-making and optimization.

IV. Case Studies in the Discovery of Exceptional Materials

To illustrate how exceptional materials are discovered with and without using machine learning, and limitations of current ML tools, we explore two case studies: the discovery of High-T_c superconductors and superhard materials. In addition to allowing us to review applications of ML to these model cases of exceptional materials, it also provides concrete examples upon which to base our subsequent recommendations.

IV. A Case study: Serendipity and the Discovery of High-T_c Superconductors

Most exceptional materials were only found by serendipity, and only after obtaining the initial results by chance, is the usual scientific process unleashed to understand the underlying causes. In essence, the chance discovery of a new 'solution' provides a starting point for both the incremental process optimization techniques discussed above, as well as traditional hypothesis- and theory-driven science. To help understand this process, Yaqub's taxonomy classifies serendipitous discovers into four categories:[95] (i) Targeted search solves unexpected problem (Walpolian); (ii) Targeted search solves problem-in-hand via unexpected route (Mertonian); (iii) Untargeted search solves immediate problem (Bushian); (iv) Untargeted search solves a later problem (Stephanian).

The discovery of high-T_c superconductors illustrates the role of serendipity in the discovery of an exceptional material. Research in superconductivity, from its initial report in 1911[96] to 1986 was dominated by metallic systems.[97] The conventional wisdom suggested that superconductors should be metallic, have high symmetries and electronic density of states, and be structurally unstable or metastable. While most superconductivity work pre-1986 focused on metallic systems, specifically vanadium or niobium alloys, other systems for which BCS theory[98] did not work, such as intermetallics[99] and Chevrel phases[100] were being investigated. Despite the modest critical temperatures in these systems, they suggested that the conventional wisdom regarding superconductivity was too narrowly focused.

Two parallel efforts in the early 1980s were critically important in emergence of cuprate perovskites as high temperature superconductors. First, Raveau,[101,102] Poeppelmeier,[103,104] and Thomas[105,106] were developing a synthetic toolbox to control the oxygen stoichiometries in

perovskites, enabling mixed valencies that are critical to the existence of superconductivity in perovskites. Second, the broad investigations led to the observation of superconductivity in a series of non-metallics systems pre-1986 oxides,[107] including NbO[108] (which contains square planes, much like all cuprate superconductors), spinels ($LiTi_2O_4$)[109] a series of perovskite adjacent tungsten bronzes,[110] $Ba(Pb_{1-x}Bi_x)O_3$[111] and even the perovskite $SrTiO_3$.[112]

Reports of oxide superconductors intrigued Bednorz and Müller, coupled with both Raveau's synthetic advances and discovery of metallic conductivity in a copper-containing oxygen deficient perovskite,[113] inspired their discovery of superconductivity in the LaBaCuO systems (T_c = 28 K).[114] Their discovery is an example of Mertonian serendipity, as Bednorz and Müller were performing a targeted search in the system that solves a problem with unexpectedly high critical temperatures: Ceramics generally act as insulators, but this anomalous case provided a new solution to an established problem via an expected route. The initial announcement was followed by feverish activity for more than a decade that more closely followed a standard process optimization route and resulted in YBaCuO oxides with superconductivity at liquid nitrogen temperature[115] and the reigning HgBaCaCuO cuprate superconductor.[116]

Could ML models have assisted in this discovery? Undoubtedly, ML can help optimize materials once examples are known—for example, Pogue et al. recently used a ML model trained on >16,000 compounds in an iterative fashion to guide the synthesis of new superconductors, and found a new Zr-In-Ni superconductor with modest T_c = 9 K, as well as rediscovering a few known superconductors not in their training set.[51] But if one only had experimental knowledge of pre-1986 superconductors, would ML predict the existence of high T_c cuprates? The answer appears to be "no". In 1988, Villars and Phillips performed what would now be called feature selection and clustering using the known data of approximately 60 high-T_c materials (including YBaCuO);

however, their prediction (Figure 2 in their paper) does not predict BaCaCuO,[117] and it is unclear to what extent many other materials would have been predicted positive. Two decades later, Stanev et al. used the SuperCon database of over 16,000 compounds to train random forest models for predicting the T_c based solely on composition.[118] While they did not consider a time-separated holdout, Figure 4b in Ref. [118] shows that a model trained on low-T_c (primarily pre-1986) materials predicts all cuprates as erroneously low-T_c. (The failure to extrapolate could be a consequence of using a random forest model.) On the other hand, their results do suggest that once a few initial discoveries are made, the model can identify other examples—indeed, once cuprates are included in the training dataset, they comprise the vast majority of candidate superconductors. Meredig et al. observed that ML models trained without cuprate examples predict cuprates to be below-average superconductor[119] (See Figure 2 in Ref. [119]) Alternatively, Ling et al. used ML to quantify the uncertainty of T_c (rather than predict its value); iterative sampling materials guided by maximum uncertainty found high- T_c superconductors (including cuprates) in about a third of the experiments required by a random search.[120] These previous ML studies may have focused too narrowly on superconductivity; perhaps a broader study of metallic conductivity (rather than limiting to superconductivity), informed by earlier reports of metallic conductivity in LaSrCuO, would have served as the bridge from classical BCS superconductors to these new compounds.[121] Overall, this supports the claims made in Section III: existing ML approaches can be successful for materials optimization, but are not necessarily capable of finding new exceptional materials.

IV.B Case study: Machine Learning and the Discovery of Superhard Materials

There are similar strengths and limitations of current ML approaches for the discovery of superhard materials, defined as those having Vickers hardness (H_v) exceeding 40 GPa. Diamond

has long been considered the hardest known naturally occurring substance (a 10 out of 10 on the Mohs hardness scale and an $H_v \approx 100$ GPa) and significant efforts have gone into making synthetic diamonds. In 1954, a team of scientists at the General Electric (GE) Research Laboratory developed the first approach involving subjecting graphite to intense heat and pressure using a diamond press. The experiment yielded tiny, yellowish crystals that were confirmed to be diamonds. Over the next several decades, GE continued to refine the process through inventions in the 1970s like high-pressure, high-temperature (HPHT) synthesis, which involved subjecting a carbon source to extreme pressure and heat in the presence of a metal catalyst. HPHT synthesis allowed GE to create larger, higher-quality diamonds more efficiently than ever before. The company began selling synthetic diamonds for use in industrial applications, such as cutting tools and abrasives. At a similar time (1957), synthesis efforts were also focused on making the isostructural, isoelectronic cubic boron nitride (c-BN) using high-pressure, high-temperature synthesis. Superhard c-BN's unique properties, including its extreme hardness (HV ≈ 60 GPa) and thermal stability, make it an ideal material for use in ferrous cutting tools, grinding wheels, and other industrial applications. Today, c-BN is used extensively in the aerospace, automotive, and manufacturing industries. Given the tremendous application space, it is no surprise researchers have expended significant effort, with only moderate success, trying to emulate these properties.

However, it has not been easy. An analysis by Brgoch and co-workers provides insight on why so few superhard materials have been identified. They constructed a boosted machine learning regression model capable of predicting Vickers hardness. Using this model to predict the hardness of more than 60,000 inorganic compounds in Pearsons Crystal Dataset revealed that only 0.1% of known crystalline compounds surpass the superhard threshold at 0.5 N applied load, and only 0.01% meet this criterion at 5N applied load.[122] Not only is superhardness rare, the total dataset of

experimentally hardness measurements is relatively small (about 500 unique compositions).[122,123] Moreover, the dataset is biased to low hardness values and certain compositions (such as boron-containing compounds) are disproportionately present in this limited training data,[122] attributable to the same types of anthropogenic research biases observed in other experimental materials datasets.[54]

The limited experimental data might suggest that physics-based simulations could be a more appropriate path towards materials discovery. However, direct atomistic physical simulations by density functional theory (DFT) are unable to directly calculate the hardness as it is a property that involves multiple length-scales exceeding what can be achieved by direct simulation. One could instead use properties that are readily calculated by DFT (such as bulk and sheer moduli) as either initial selection criteria[123] or as inputs to semiempirical expressions for hardness.[124] Researchers have further paired these methods with crystal structure prediction algorithms (USPEX[125], CALYPSO[126], XtalOpt[127]) to predict new promising superhard compounds. ML can also be used to expand the search space enabled by DFT calculations.[123] These physically-motivated models provide some guidance but are generally worse at quantitative hardness predictions than direct ML methods.[122,124] ML models are accurate enough to be used to screen for interesting compounds in the Sc-Os-B phase space as a demonstration of their quantitively accuracy. The model was able to capture the change in hardness in a solid solution system ($Sc_{2-x}Y_x)OsB_6$ and the highly disordered borosilicide, $YB_{41.2}Si_{1.42}$. Additionally, Sc_2OsB_6 was determined to be nearly superhard ($H_v \approx 38$ GPa). Nevertheless, the hardness of these systems falls far from diamond or c-BN.

A more recent approach to the problem embraces the rarity of superhardness by treating it as an unsupervised anomaly detection problem.[128] In this work, an autoencoder model was trained

to find low-dimensional latent representations of crystal structure. Compounds with anomalous bonding motifs will be poorly described in this learned representation, and this can be used to identify anomalous structures for further investigation. While such structural anomalies do not directly indicate superhardness, the hypothesis is that these materials often contain unusual bonding motifs, which is substantiated by an empirical correlation between reconstruction error and superhardness. The methodology could be expanded to include a generative approach that can predict new crystal structures where the loss function (reconstruction error) is maximized and premised on the previous correlation, having a correspondingly higher hardness. Nevertheless, there is no guarantee that any combination of elements in any given crystal structure would surpass diamond as the hardest single-phase material.

V. Recommendations towards ML for Exceptional Materials

A close look at the history of science reveals that there is no single "scientific method" and that scientific advances often involve a rejection of established norms.[129] In that spirit of epistemological anarchism, we offer six maxims for guiding the research community, depicted schematically in Figure 5. There exists a variety of software and hardware needs that apply to laboratory automation in general, which will not be discussed, so as to focus on aspects specific to exceptional materials. While we have in mind primarily *experimental* discovery, many of these recommendations are equally applicable to autonomous computational discovery.[130] When possible, we have tried to illustrate these points by reference to specific applications in chemistry and materials science, but in many cases, we have drawn instead upon examples drawn from finance, oceanography, computer science, and evolutionary biology, among other fields.

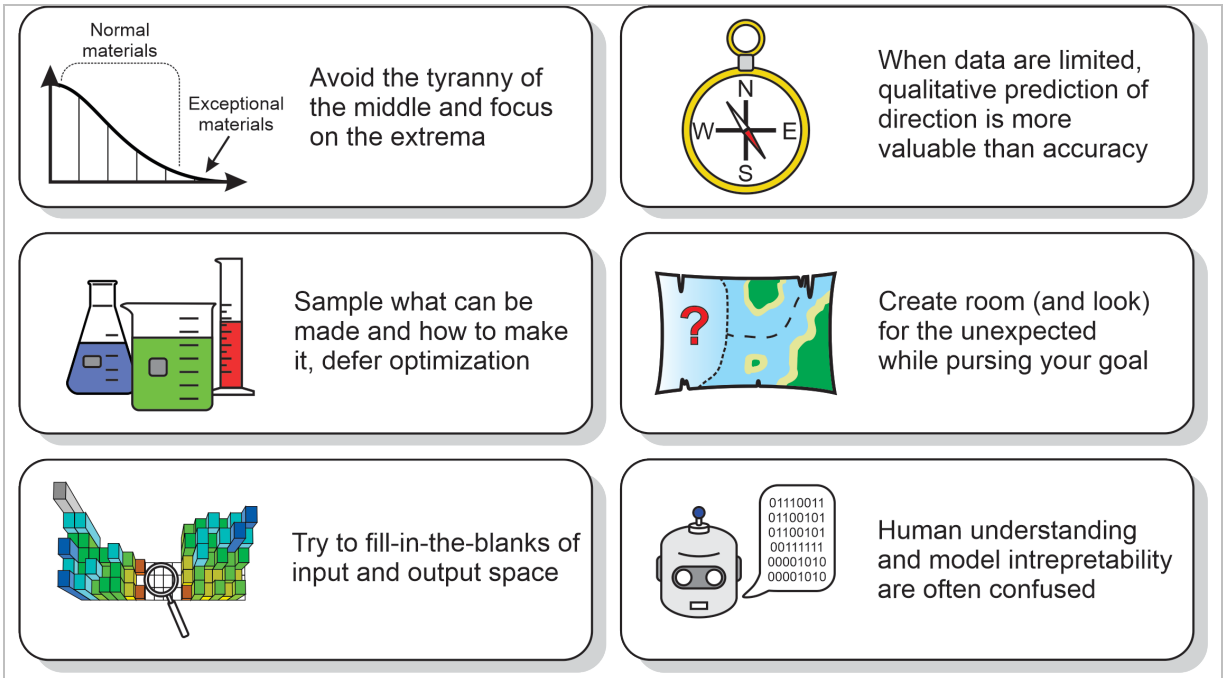

Figure 5: Six recommendations for research toward machine learning for exceptional materials.

Avoid the tyranny of the middle and focus on extrema. By definition, there is less training data at the extremes, resulting in greater model uncertainty associated with those regions. Typical metrics for ML training and evaluation emphasize performance on an average over the data, but this will be dominated by typical materials rather than the exceptional extrema. Common ML metrics (accuracy, R^2, etc.) do not express the intended goal when in the presence of such outcome imbalances,[131] and nor do they give a measure of the ability of an algorithm to guide iterative discovery.[57] Solving this problem may simply correspond to choosing different loss functions when training ML models. A possible analogy is to the use of Conditional Value at Risk (CVaR)—expected loss in the worst q% of cases—in portfolio optimization,[132] which corresponds to a 1-norm of q% largest magnitude entries.[133] Alternatively, it may require modifications of

existing algorithms. Typical reinforcement learning (RL) formulations do not correspond to the scientific discovery process, as they are aimed at maximizing a cumulative reward rather than the best possible result found.[134] Alternative formulations of the problem, such as the Max-*k*-arm bandit model,[135] better align with that goal, which has been demonstrated with *in silico* numerical experiments of exploring molecular SMILES strings to maximize the boiling point and other thermophysical properties described by an empirical proxy.[134] In the context of Bayesian optimization type strategies, an appropriate approach is the output-weighted optimal sampling introduced by Blanchard and Sapsis and co-workers,[136–138] which has been recently applied to extreme event discovery in epidemiological models, rogue waves, and structure mechanics.[139]

When data is limited, qualitative prediction of direction to the goal is more valuable than (interpolative) accuracy. If you are blindfolded, it is better to know the approximate *direction* to the goal than it is to know the exact *distance* to the goal. Focusing on accuracy in the early stages can be detrimental; for example, Random Forest models tuned to maximize only cross-validation accuracy may produce low-quality models.[119,140] But collecting just *any* data results in the "tyranny of the middle" problem discussed above. Rather, we want simple qualitative models that guide extrapolation (and data collection efforts) to collect relevant data, rather than quantitative interpolative accuracy. Don't build a perfect model with limited data—the important thing is to collect more data, and the right data.

The evolution of astronomy from Ptolomy to Kepler provides an insightful historical analogy. Kepler's model was neither more accurate nor significantly simpler than Ptolomy's.[141] Kepler himself noted in the introduction to *Astronomia Nova* "the [models] are for practical purposes equivalent to a hair's breadth, and produce the same results."[142] Mathematically, the

system is underdetermined, as the limited set of data can be fit by an arbitrarily complex model. The strength of Kepler's model was not necessarily that it *predicted* orbits more accurately, but rather that it enabled a piecemeal approach to *extract hypotheses* about the orbits of individual planets, enabling astronomers to develop observational tests.[143] For example, latent in Kepler's model was the information that Venus should have phases, and one could calculate its occurrence with sufficient predictive accuracy to enable Galileo's experimental observations which ruled out the Ptolemaic model.[144] (Ruling out the Tychonic model required improved instrumentation and data collection to enable the observation of stellar aberration by Bradley,[145] which provides a parallel to our argument for the role of HTE systems in exceptional materials discovery.)

The general strategy for underdetermined problems is to introduce *a priori* constraints. Classically, this was done by devising physical models in terms of the relevant variables and admissible functional forms of their interactions. Physics-based computer simulations serve a similar role,[130] although the examples above indicate their limits for exceptional materials.[53,124] We will focus purely on data-driven approaches. Strategies of *physics-informed machine learning*[146–148] are one approach for this problem. A recent application of this approach to determining the structure of oxide glasses is described by Bødker *et al.*[149] However, this is less applicable to exceptional materials which involve new physics precluded by using existing laws as constraints. *Feature selection* corresponds to an implied constraint that only a small subset of the input variables determines the system performance. The identified features are combined with simple models to make predictions. Some examples include the aforementioned synthesis of superhard materials,[123] but other examples include discovery of antimicrobial conjugated oligoelectrolytes[150] and perovskite crystal growth modifying additives.[151] Once hypothetically-relevant features correlated to the output are selected, then relatively simple models can be

constructed to make extrapolations. Even simple linear models can be quite effective for this purpose.[131] The features themselves need not have an interpretable relationship to the property being studied (*vide infra*)—they merely serve as a proxy for guiding the experiment selection. There is also no reason to restrict consideration to a predefined ML-model function type. More broadly, *symbolic regression* corresponds to the ansatz that a relatively simple combination of mathematical functions describes the behavior. There are a variety of applications of symbolic regression methods to problems in chemistry[152] and materials science.[153,154] In practice, symbolic regression is often combined with various feature selection methods, with examples including VS-SISSO[155] and transformer-based approaches for symbolic regression.[156]

Emphasizing qualitative direction has consequences for the design of HTE systems. For example, the experimental validation, especially at early stages might emphasize rapid (but potentially noisy) methods, rather than the types of rigorous methods used in subsequent stages of research in the interest of increasing coverage. This also suggests the need for appropriate data sharing and interoperability formats (such as the specification of experiments) to facilitate the hand-off between high- and low-throughput synthesis and characterization processes, especially when they occur in different laboratories. On the other hand, many historical examples of exceptional material discoveries resulted from comprehensive characterizations which were unnecessary to the immediate goals of the project, but which nonetheless revealed an unanticipated outcome (Walpolian or Stephanian serendipity[95]). For example, conductivity measurements in the LaBaCuO system revealed metallic behavior, which precipitated Bednorz and Müller's discovery. This suggests measuring as many different properties as possible, even if not directly related to the current research theme, and storing the results in public databases to allow for retroactive retrieval of surprises or the use in training ML models for different properties.

Sample what can be made and how to make it — defer optimization. As it is impossible to exhaustively enumerate all of the possibilities in these problems, one must instead sample the possibilities, which corresponds to the task of *generative* ML models; methods and applications of generative ML to chemical problems have very recently been reviewed in Ref. [157]. We advocate that these methods be used to cast a wide net. As noted by Herbert Simon, finding a global optimum to real-world problems often requires an intractable amount of time, effort, and computation, but finding a solution that *satisfices*—i.e., is feasible and meets or exceeds a baseline aspiration level—is often tractable.[158,159] This is marked in the case of combinatorial optimizations—like those involved in materials discovery—in which the number of possibilities grows exponentially in the problem variables, each of which must be checked. In these cases, we argue that merely sampling the solutions to find a satisficing solution should be our goal. Evolutionary theory suggests that introducing high levels of selection pressure restricts the scope and direction of exploration to a small neighborhood near high fitness individuals, and in turn delays or prevents innovation by inhibiting a series of slightly deleterious intermediate steps that are needed to find new optima.[160] For this reason, a collection of satisficing solutions can be more useful for our purpose than a few highly optimized examples.

To be more than a theoretical curiosity, it must be possible to synthesize the material. This may be subdivided into the question of whether the material *can* exist (i.e., fundamental thermodynamic constraints) and *how* it can be brought into existence (the sequence of practical operations and feasibility of required conditions). The former is partially addressed by the plethora of ML models for predicting ground state thermochemistry, along with a proper accounting for metastability.[35] The latter is partially addressed by ML approaches that use natural language

processing on the literature to extract experiment plans (for training) and then generate plans based on that data.[161] (A parallel discussion of these ideas as they apply to organic chemistry can be found in Ref. [157].)

More broadly, one can think of two extreme versions of this task. At one extreme, synthesizability is applied as a filter to a list of generated candidates. For example, using ML models to make predictions of superhardness, then applying a formation energy filter to identify the feasible compositions.[122]At the other extreme, synthesizability is imposed to generate candidates by enumerating (or defining) a state space of experimentally feasible composition and process conditions points and then allowing property prediction models to select within them. A more efficient approach would combine these extremes to avoid the need to evaluate candidates which are ultimately discarded by the subsequent process. This might range from including physics-based symmetry contraints,[162] directly incorporating a learned formation energy constraint into the generative process,[163] or by restricting the generating samples to obey compositional "grammatical" rules.[164] Fundamentally, the limits of synthesizability are defined in terms of the operational capabilities of the autonomous experiment system and what actually happens in the lab. Thus, an extreme version of the latter approach is simply to allow an algorithm to guide the HTE system directly. An example of this approach is a genetic algorithm optimization of gold nanoparticle synthesis experimental parameters to match a specified UV/vis spectra.[165]

The important thing is that the ML model leads to samples in the right neighborhood. One framework for thinking about this is provided by similarity-based kernel learning approaches, in which one can define a cost function associated with acquiring a desired (but difficult) data point versus several similar (but more easily acquired) data points, and then use a model trained on the local environment to infer the desired point.[166] The ease of acquisition can be computed by

combining materials, labor, and time constraints.[167] Another framework is provided by the Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) algorithm, an evolutionary algorithm used in reinforcement learning, which samples and stores multiple candidate solutions ("elites") on a grid to preserve a diverse set of characteristics for possible solution.[168] Zooming-based Bayesian optimizations have a similar alternation between global sampling and local optimization.[38] This is also reminiscent of Lévy flight models of animal foraging behavior, in which the search processes is characterized as a random walk with a heavy-tailed distribution of step sizes, and which in practice looks like local exploration in a region interspersed by large jumps to new regions.[169]

Create room (and look) for the unexpected while pursuing your goal. Scientists are trained to minimize the variance in their laboratory procedures. (There is even a new ACS journal, *Precision Chemistry*,[170] focused upon this goal.) In contrast, we advocate the opposite approach—Max Delbruck's *principle of limited sloppiness*: "If you are too sloppy, then you never get reproducible results, and then you never can draw any conclusions. But if you are just a little sloppy, then when you see something startling you … nail it down."[95] A theoretical justification is provided by Epsilon-greedy approaches in reinforcement learnings[171]—while one should mostly take what one believes to be the most profitable action, one should also allocate some fraction of effort to trying random new actions, just in case something good happens. This is synergistic with our previous recommendation to avoid premature optimization. Sloppiness can be engineered into the experiment process in many ways. Examples include adding randomness to materials experiment plans,[54] using an additional cost function to experiment generation that maximizes experiment diversity,[49] or taking advantage of uncontrolled changes in laboratory temperature and humidity as natural experiments.[172]

Despite advocating for deliberately "sloppy" reaction designs, we emphasize that this requires complete data capture of what actually transpired. HTE provides a natural synergy, as it enables complete, machine-readable data collection of meta-data and "failed" experiments which might not otherwise be recorded, but which are essential for ML training.[173] Furthermore, allowing for sloppier outcomes might simplify the design tolerances when constructing an HTE system.[14] Once the data is collected, ML methods for *anomaly detection* enable Walpolian and Mertonian serendipity (a targeted search that solves an unexpected problem, or a known problem in an unexpected way, respectively[95]). The role of structural anomaly detection in the discovery of superhard materials was discussed in Sect. IV.B,[128] and similar opportunities have been discussed for computer-vision-based scanning electron microscopy characterizations[174] and surface-enhanced Raman.[175] In their simplest form, this might take the form of detecting whether an unexpected change has occurred in one or more spectra is observed. This has been used to discover new organic synthesis reactions.[176,177] Coupling the observation of change in the spectra to neural network models of molecular structure has been used to steer the experimentation towards less predictable reactions.[178]

Data-reuse and sharing can also enable finding unexpected trends within and between labs by data sharing.[179] We are particularly interested in the model of Stephanian serendipity,[95] in which a prior solution (for example, a compound that was made and characterized for a different purpose) is found to solve a later problem because of some new insight. The best example we know is the identification of lead titanate as a stable photocathode for dye sensitized solar cells (DSSC), based on band structure similarity to known photocathodes.[180] Requirements by funders and publishers around FAIR (findable, accessible, interoperable, reusable)[181,182] and TRUE

(Transparent, Reproducible, Usable by others, Extensible)[183] data practices can help create such a resource for retroactive discovery.

Try to fill-in-the-blanks of input and output space. There is a great opportunity to develop ML methods that enable the other half of the taxonomy of serendipity, namely *untargeted search* (Bushian and Stephanian approaches).[95] Closely related is the importance of uncertainty quantification—fill in the portions of the map with the greatest uncertainty. The obvious way to frame this is in terms of the types of *inputs* (e.g., compositions and structure) that have not been observed before. Identifying where these gaps exist can be done by using databases, such as the identification of compositional gaps in the Materials Projects database discussed in Section III and Figure 3(b). Proactively, a strategy is to identify these unexplored compositions, use constraints (such as ML-based formation enthalpy estimators of stability) to determine which compositions are feasible, and then target experimental searches to fill in those blanks.[184] For example, the discovery of the many high-temperature cuprate superconductors was directly enabled by the Goldschmidt tolerance factor[185] which enables the determination of feasible compositions likely to result in the formation a perovskite. This, coupled with the solid-state literature and nascent Inorganic Crystal Structure Database,[186] resulted in a host of experiments targeted at potential novel materials that were both feasible and unreported. More recently, ML-based approaches have been applied to better explore the space of cuprate-like compounds.[187]

Another way to frame this problem is in terms of the types of *outcomes* that have not been observed before. The information entropy of the observed property distribution can be useful for identifying these dataset imbalances, and active learning used to prioritize new samples to correct these imbalances, recently demonstrated in the context of formation energy/structure biases of

intermetallic compounds.[36] In the context of exceptional materials, we are trying to understand how properties are coupled to one another. It might therefore be useful to fill in contraindicated regions with undesirable tradeoffs ("anti-exceptional materials") which might be equally rare. To achieve this one must learn general ways for getting to arbitrary output, which can serve as steppingstones to the desired solution. An extreme version of this approach explicitly rejects objective-based search, and focuses solely on output novelty.[188] Empirical evidence suggests that novelty-only strategies (which ignore any type of fitness objective function) can be highly effective in complex environments, such as video games.[189] *Random goal exploration* algorithms[190,191] select a random target defined in the space of possible outcomes and then infer the necessary inputs needed to achieve that goal. The process can be repeated iteratively until the target is reached, refining the model's knowledge of the input-output relationships. These methods have been demonstrated in the context of an HTE system for identifying novel protocell lipid formulations.[192] Blending the distinction between input- and outcome- oriented approaches, the so-called *diversity is all you need* strategy in reinforcement learning suggests optimizing for both novel outcomes and novel synthetic paths (inputs), again, without imposing other types of fitness objective functions.[193] Regardless of the specific optimization strategy, appropriate data sharing (vide supra) is a necessary ingredient for creating a synoptic data resource that can be used to identify the underexplored input and output spaces.

Do not confuse human understanding and model interpretability. The debate about whether deeper forms of knowledge exist and are more valuable than a merely true belief is one of the oldest discussions in epistemology, going back to Plato's *Meno.*[194] Suppose you have an oracle (perhaps even a GPT-3 based language model making chemical property predictions[68]) that

tells you where you can get lucky—does it matter how the prediction is made, provided you can verify it is true? Once the initial discovery is validated experimentally, the traditional scientific method can be unleashed to understand underlying causes systematically. We saw this pattern in the case study of high-T_c superconductors discussed previously, and there is no reason to hold ML-assisted discoveries to a higher standard. Essentially, we argue that the initial discovery stage should prioritize a form of *reliabilism* (defining knowledge as a reliably-formed true belief), with an emphasis placed on Ryle's *knowledge-how* (in contrast to *knowledge-that*). This neither requires an explanation of the workings of an arbitrary black-box ML model, nor is it recognizable as constituting a proper "scientific" explanation. (Whatever "explanation" means in practice.[129])

Leo Breiman famously contrasted *model culture,* which uses data to estimate the values of physically meaningful parameters, against *algorithm culture* (what we would now call ML) which views the algorithm as a black box whose parameters are meaningless apart from prediction quality.[77] The confusion between these two cultures leads to misapplication and misinterpretation about the scope of *explainable AI* (XAI) methods for communicating the inner workings of ML algorithms to humans. For a very recent review of trends in this field more broadly, see Ref.[195]; for reviews of interpretable and explainable methods applied specifically to materials science, see Refs. [196–200]. In practice common XAI methods can be misleading; a case study of the limitations of methods like SHAP and ensemble feature importance measures in experimental materials science problems has been discussed by Lei et al.[140] At best, XAI methods generate low-dimensional descriptions of how the model outputs behave based on changes to the inputs. That is, they only indicate features correlated to the model's output, with no claim of physical meaning or causality. Determining whether these features are meaningful requires human input. Practically speaking, XAI methods may be unnecessary for the initial discovery of exceptional materials.

Model explainability in these early stages is unnecessary because the models will be based on limited data, and thus prone to overfitting and oversimplification. Moreover, the most appropriate models for initial discovery—for both interpretability and extrapolation—may be the types of feature-selected linear models discussed above,[131] obviating the need for more sophisticated black-box model interpretability methods. In many cases, automatic identification of anomalies (*vide supra*) for review by a human operator suffices, so long as the anomalies are rare. The human scientist can then invoke their own reasoning, statistical evidence, or other forms of investigation to study the problem.

ML can certainly also play a role in building scientific understanding *after* initial discovery of an exceptional material. Most scientists adopt the epistemology advanced in Plato's *Phaedo*, but most popularly associated with Aristotle's *Physics*, which associates the inquiry into nature as a knowledge of causes.[201] Obtaining knowledge of causal relationships is useful, and can be automated by the use of modern causal influence methods[202] which have recently been applied to catalysis[203] and scanning probe microscopy.[204–206] Ultimately, causal explanations must go beyond merely the brute details of the experiments (such as the input settings on a particular instrument), and draw upon deeper semantic relationships underlying structure, property, processing, and characterization encoded in an explicit and machine-readable way. Ideally, this information is incorporated into interoperable knowledge graphs that would allow scientists (and automated inference engines) to operate on fully linked concepts and data instances.[207] In contrast to the statistical inference methods that constitute most of the applications of ML, semantic representations allow for logical inferences characteristic of symbolic AI (so called "good old fashioned AI") *en vogue* during the 1950s-1990s.[208] Progress towards semantic representations of chemistry and materials data and their applications are discussed in Refs. [207,209,210].

Provocatively, even without model interpretability or causal explanations, merely having access to "superhuman" AI improves human decision-making. This has been observed in the context of the board game Go.[211] In the 60 years of tournament data prior to AlphaGo, human decision quality remained roughly constant and human decision novelty decreased. The advent of AlphaGo caused both human decision quality and novelty to increase, which is attributed to the AI's use of optimal decisions free of historical biases, which in turn created new opportunities for human players to learn and innovate. In the context of scientific discovery, this suggests that merely employing the novelty-enhancing recommendations suggested above may be sufficient to improve human scientific understanding, even in the absence of model explainability, per se.

VI. Conclusion: Integrated Workflows for Exceptional Material Discovery

Traditional "manual" and autonomous materials discovery is based on a synthesize, characterize, learn, plan loop, depicted schematically in Figure 6a. (Similar process loops, with slightly different names, occur in a variety of scientific fields and the automation thereof, as discussed in Ref. [78].) Existing machine learning approaches accelerate this process by assisting in various optimization subtasks.[14] For example, the inevitable fine-tuning of synthesis and testing operations when dealing with new precursors can be delegated to algorithms, reducing the need for skilled labor in operating tools. The characterization process can be accelerated by using ML to automate spectral interpretation[199] and plan characterization campaigns that reduce the number of measurements required to more efficiently use available resources.[212] ML can also be used extract additional information from existing spectroscopy and microscopy methods.[213,214] As discussed in Section III.B, existing ML approaches excel at the variety of research-related optimization tasks.

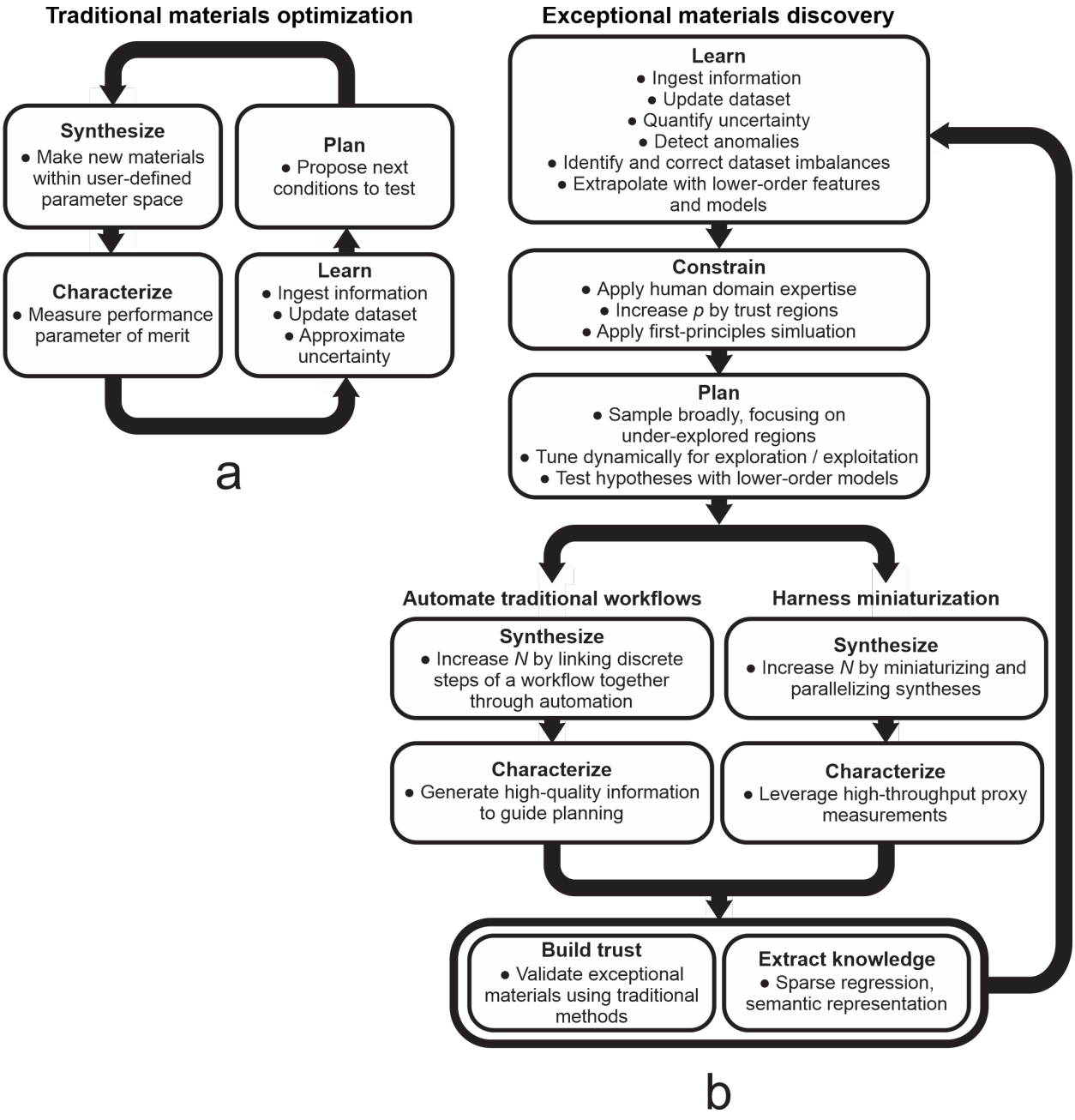

Figure 6: (a) Block diagram of typical autonomous workflow; (b) Block diagram of an autonomous workflow oriented towards exceptional materials.

How might the discovery workflow change to discover exceptional materials? A schematic is depicted in Fig 6b. As discussed in Section III.C, it is necessary to increase both *p* (corresponding to the *learn* and *plan* phases) and *N* (corresponding to *synthesize* and *characterize*). Section V presented examples and suggestions of how new types of ML can increase *p*. Given the low probability of exceptional materials, one might introduce an intermediate *constrain* phase to limit the possibilities. While this may include the types of thermodynamic and synthetic feasibility determination methods discussed in Section V, it is potentially broader in scope. For example, Liu et al. described how to merge human observation of sample quality into an ML acquisition function using soft constraints,[19] and Zubarev et al. recently described software to assist in eliciting and systematizing human subject matter expert advice about prioritization, level-of-knowledge, and risk assessments used as input to ML-assisted discovery of new photoacid generator for EUV lithography.[215]

Given the rarity of exceptional materials, it is also crucial to increase *N*, the number of unique material compositions tested per unit time by HTE methods, discussed in Section III.C. Broadly, this can be accomplished can either automating existing laboratory processes or developing new types of miniaturized processes. An extreme version of the former is a mobile robotic arm that uses the same synthesis and characterization equipment as a human chemist,[216] but it might consist of a dedicated "ChemPU" device[217,218] or a collection of modified-equipment capabilities orchestrated by a central sample management system.[219] While this has the advantage of using well-understood techniques, the opportunity for acceleration and scaling is limited by those existing synthesis and characterization techniques. An alternative is to embrace new types of miniaturized and high-throughput synthesis and characterization methods such as microfluidics systems,[220,221] direct writing from liquid precursors,[222,223] combinatorial deposition of sample

libraries,[224] and atomic scale dip-pen nanolithography.[225,226] (Given their novelty, the design of the devices themselves are a subject for traditional ML-based optimization.[227]) While these synthesis methods can potentially increase *N* by orders of magnitude, it introduces doubt whether the resulting products are representative bulk samples. These approaches make characterization the rate limiting step, requiring a shift to faster optical or electrical proxy measurements. The open challenge is to define the limits within which proxies are valid or fail, and how to dispatch the discovery process across these different types of modalities. As illustrated in Figure 6, this will require research efforts to demonstrate equivalence between standard and novel synthesis and characterization techniques. This would initially take the form of explicit trust building experiments conducted on both sets of instruments, but ideally could also be automated. At the level of understanding, this might take the form of knowledge-graph approaches to represent semantic relationships between the results of different types of methods applied to a sample.[207]

Finally, at the level of planning and coordination between these different types of modalities, the use of intelligent agents can be used to direct guide more purpose-driven planning and design tools, and automate aspects of the reasoning process across multiple facilities with different capabilities.[228,229] Putting these recommendations into the form of an integrated workflow should better enable the discovery of exceptional materials.

VI. Corresponding Authors

[1] Email: jschrier@fordham.edu
ORCID: 0000-0002-2071-1657

[2] Email: anorquis@haverford.edu
ORCID: 0000-0003-1040-5880

[3] Email: buonassi@mit.edu
ORCID: 0000-0001-8345-4937

[4] Email: jbrgoch@uh.edu
ORCID: 0000-0002-1406-1352

Notes

The authors declare no competing financial interest.

Acknowledgements

We thank Elliott Risch for discussions about semantic data representations, A. Gilad Kusne, Anubhav Jain, and Valentin Stanev for discussions about machine learning for novel superconductors, and Kenneth Poeppelmeier for discussions on the history of cuprate superconductors. This study is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001118C0036. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA. JS acknowledges resources of the MERCURY consortium (http://mercuryconsortium.org/) under NSF Grant No. CNS-2018427. TB acknowledges a research grant from First Solar. JB thanks the National Science Foundation (DMR-1847701).

Graphical Table of Contents Image

References

(1) Bhatti, S.; Manzoor, H. U.; Michel, B.; Bonilla, R. S.; Abrams, R.; Zoha, A.; Hussain, S.; Ghannam, R. Machine Learning for Accelerating the Discovery of High Performance Low-Cost Solar Cells: A Systematic Review. arXiv December 26, 2022. https://doi.org/10.48550/arXiv.2212.13893.

(2) Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics. *npj Comput Mater* **2022**, *8* (1), 1–9. https://doi.org/10.1038/s41524-022-00765-z.

(3) Huang, E.-W.; Lee, W.-J.; Singh, S. S.; Kumar, P.; Lee, C.-Y.; Lam, T.-N.; Chin, H.-H.; Lin, B.-H.; Liaw, P. K. Machine-Learning and High-Throughput Studies for High-Entropy Materials. *Materials Science and Engineering: R: Reports* **2022**, *147*, 100645. https://doi.org/10.1016/j.mser.2021.100645.

(4) Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. Applications of Machine Learning in Drug Discovery and Development. *Nat Rev Drug Discov* **2019**, *18* (6), 463–477. https://doi.org/10.1038/s41573-019-0024-5.

(5) Pandey, M.; Fernandez, M.; Gentile, F.; Isayev, O.; Tropsha, A.; Stern, A. C.; Cherkasov, A. The Transformational Role of GPU Computing and Deep Learning in Drug Discovery. *Nat Mach Intell* **2022**, *4* (3), 211–221. https://doi.org/10.1038/s42256-022-00463-x.

(6) Bannigan, P.; Aldeghi, M.; Bao, Z.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine Learning Directed Drug Formulation Development. *Advanced Drug Delivery Reviews* **2021**, *175*, 113806. https://doi.org/10.1016/j.addr.2021.05.016.

(7) Xie, S. Perspectives on Development of Biomedical Polymer Materials in Artificial Intelligence Age. *J Biomater Appl* **2023**, 08853282231151822. https://doi.org/10.1177/08853282231151822.

(8) Janet, J. P.; Kulik, H. J. *Machine Learning in Chemistry*; ACS In Focus; American Chemical Society: Washington, DC, USA, 2020. https://doi.org/10.1021/acs.infocus.7e4001.

(9) Butler, K. T.; Oviedo, F.; Canepa, P. *Machine Learning in Materials Science*; ACS In Focus; American Chemical Society: Washington, DC, USA, 2022. https://doi.org/10.1021/acsinfocus.7e5033.

(10) Ozin, G. A.; Loh, J. Y. Y. Chapter 9. Human–Machine Interface. In *Energy Materials Discovery*; Royal Society of Chemistry: Cambridge, 2022; pp 358–408. https://doi.org/10.1039/9781839163838-00358.

(11) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2.

(12) Wang, A. Y.-T.; Murdock, R. J.; Kauwe, S. K.; Oliynyk, A. O.; Gurlo, A.; Brgoch, J.; Persson, K. A.; Sparks, T. D. Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **2020**, *32* (12), 4954–4965. https://doi.org/10.1021/acs.chemmater.0c01907.

(13) Morgan, D.; Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 71–103. https://doi.org/10.1146/annurev-matsci-070218-010015.
(14) Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M. The Case for Data Science in Experimental Chemistry: Examples and Recommendations. *Nat Rev Chem* **2022**, *6*, 357–370. https://doi.org/10.1038/s41570-022-00382-w.
(15) Maruyama, B.; Hattrick-Simpers, J.; Musinski, W.; Graham-Brady, L.; Li, K.; Hollenbach, J.; Singh, A.; Taheri, M. L. Artificial Intelligence for Materials Research at Extremes. *MRS Bulletin* **2023**, *47*, 1154–1164. https://doi.org/10.1557/s43577-022-00466-4.
(16) Kumar, J. N.; Li, Q.; Tang, K. Y. T.; Buonassisi, T.; Gonzalez-Oyarce, A. L.; Ye, J. Machine Learning Enables Polymer Cloud-Point Engineering via Inverse Design. *npj Comput Mater* **2019**, *5* (1), 1–6. https://doi.org/10.1038/s41524-019-0209-9.
(17) Mekki-Berrada, F.; Ren, Z.; Huang, T.; Wong, W. K.; Zheng, F.; Xie, J.; Tian, I. P. S.; Jayavelu, S.; Mahfoud, Z.; Bash, D.; Hippalgaonkar, K.; Khan, S.; Buonassisi, T.; Li, Q.; Wang, X. Two-Step Machine Learning Enables Optimized Nanoparticle Synthesis. *npj Comput Mater* **2021**, *7* (1), 1–10. https://doi.org/10.1038/s41524-021-00520-w.
(18) Cao, L.; Russo, D.; Felton, K.; Salley, D.; Sharma, A.; Keenan, G.; Mauer, W.; Gao, H.; Cronin, L.; Lapkin, A. A. Optimization of Formulations Using Robotic Experiments Driven by Machine Learning DoE. *Cell Reports Physical Science* **2021**, *2* (1), 100295. https://doi.org/10.1016/j.xcrp.2020.100295.
(19) Liu, Z.; Rolston, N.; Flick, A. C.; Colburn, T. W.; Ren, Z.; Dauskardt, R. H.; Buonassisi, T. Machine Learning with Knowledge Constraints for Process Optimization of Open-Air Perovskite Solar Cell Manufacturing. *Joule* **2022**, *6* (4), 834–849. https://doi.org/10.1016/j.joule.2022.03.003.
(20) Boyd, S. P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK ; New York, 2004.
(21) Stadler, P. F. Landscapes and Their Correlation Functions. *J Math Chem* **1996**, *20* (1), 1–45. https://doi.org/10.1007/BF01165154.
(22) Kuhn, T. S. *The Structure of Scientific Revolutions*, Fourth edition.; The University of Chicago Press: Chicago ; London, 2012.
(23) Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*, 1st ed.; Random House: New York, 2007.
(24) Nugent, W. A. "Black Swan Events" in Organic Synthesis. *Angew. Chem. Int. Ed.* **2012**, *51* (36), 8936–8949. https://doi.org/10.1002/anie.201202348.
(25) Zunger, A. Inverse Design in Search of Materials with Target Functionalities. *Nat Rev Chem* **2018**, *2* (4), 0121. https://doi.org/10.1038/s41570-018-0121.
(26) Kohnert, A. A.; Wirth, B. D.; Wolverton, C.; Balachandran, P. V.; Capolungo, L. Modeling Materials under Coupled Extremes: Enabling Better Predictions of Performance. *MRS Bulletin* **2023**. https://doi.org/10.1557/s43577-022-00455-7.
(27) Barnes, T.; Schelhas, L.; Hansen, C.; Steinman, L.; Jain, A. *Durable Module Materials (DuraMAT) Consortium Final Technical Report*; NREL/TP-5K00-85569; 2023; p NREL/TP-5K00-85569, 1962808, MainId:86342. https://doi.org/10.2172/1962808.

(28) Häse, F.; M. Roch, L.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chemical Science* **2018**, *9* (39), 7642–7655. https://doi.org/10.1039/C8SC02239A.
(29) Erps, T.; Foshey, M.; Luković, M. K.; Shou, W.; Goetzke, H. H.; Dietsch, H.; Stoll, K.; von Vacano, B.; Matusik, W. Accelerated Discovery of 3D Printing Materials Using Data-Driven Multiobjective Optimization. *Science Advances* **2021**, *7* (42), eabf7435. https://doi.org/10.1126/sciadv.abf7435.
(30) Low, A.; Lim, Y.-F.; Hippalgaonkar, K.; Vissol-Gaudin, E. Mapping Pareto Fronts for Efficient Multi-Objective Materials Discovery. *AISY Interactive Papers* **2022**. https://doi.org/10.22541/au.167083158.87890589/v1.
(31) Hippalgaonkar, K.; Li, Q.; Wang, X.; Fisher, J. W.; Kirkpatrick, J.; Buonassisi, T. Knowledge-Integrated Machine Learning for Materials: Lessons from Gameplaying and Robotics. *Nat Rev Mater* **2023**, 1–20. https://doi.org/10.1038/s41578-022-00513-1.
(32) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational Screening of All Stoichiometric Inorganic Materials. *Chem* **2016**, *1* (4), 617–627. https://doi.org/10.1016/j.chempr.2016.09.010.
(33) Restrepo, G. Chemical Space: Limits, Evolution and Modelling of an Object Bigger than Our Universal Library. *Digital Discovery* **2022**, *1* (5), 568–585. https://doi.org/10.1039/D2DD00030J.
(34) Aykol, M.; Hegde, V. I.; Hung, L.; Suram, S.; Herring, P.; Wolverton, C.; Hummelshøj, J. S. Network Analysis of Synthesizable Materials Discovery. *Nat Commun* **2019**, *10* (1), 2018. https://doi.org/10.1038/s41467-019-10030-5.
(35) Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W. D.; Gamst, A. C.; Persson, K. A.; Ceder, G. The Thermodynamic Scale of Inorganic Crystalline Metastability. *Science Advances* **2016**, *2* (11), e1600225. https://doi.org/10.1126/sciadv.1600225.
(36) Rondinelli, J. M.; Chen, W. ET-AL: Entropy-Targeted Active Learning for Bias Mitigation in Materials Data. *Applied Physics Reviews* **2023**, *10* (2), 021403. https://doi.org/10.1063/5.0138913.
(37) Barnes, J. W.; Dimova, B.; Dokov, S. P.; Solomon, A. The Theory of Elementary Landscapes. *Applied Mathematics Letters* **2003**, *16* (3), 337–343. https://doi.org/10.1016/S0893-9659(03)80054-X.
(38) Siemenn, A. E.; Ren, Z.; Li, Q.; Buonassisi, T. Fast Bayesian Optimization of Needle-in-a-Haystack Problems Using Zooming Memory-Based Initialization. arXiv 2022. https://doi.org/10.48550/ARXIV.2208.13771.
(39) Kim, Y.; Kim, E.; Antono, E.; Meredig, B.; Ling, J. Machine-Learned Metrics for Predicting the Likelihood of Success in Materials Discovery. *npj Comput Mater* **2020**, *6* (1), 1–9. https://doi.org/10.1038/s41524-020-00401-8.
(40) *Materials Genome Initiative*. https://www.mgi.gov/ (accessed 2023-02-02).
(41) Xie, J.; Su, Y.; Zhang, D.; Feng, Q. A Vision of Materials Genome Engineering in China. *Engineering* **2022**, *10*, 10–12. https://doi.org/10.1016/j.eng.2021.12.008.
(42) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 49–69. https://doi.org/10.1146/annurev-matsci-090319-010954.

(43) Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic Algorithms for Computational Materials Discovery Accelerated by Machine Learning. *npj Comput Mater* **2019**, *5* (1), 1–6. https://doi.org/10.1038/s41524-019-0181-4.

(44) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in Materials Research: A Case Study in Carbon Nanotube Growth. *npj Comput Mater* **2016**, *2* (1), 1–6. https://doi.org/10.1038/npjcompumats.2016.31.

(45) Bateni, F.; Epps, R. W.; Antami, K.; Dargis, R.; Bennett, J. A.; Reyes, K. G.; Abolhasani, M. Autonomous Nanocrystal Doping by Self-Driving Fluidic Micro-Processors. *Advanced Intelligent Systems* **2022**, *4* (5), 2270020. https://doi.org/10.1002/aisy.202270020.

(46) Gongora, A. E.; Xu, B.; Perry, W.; Okoye, C.; Riley, P.; Reyes, K. G.; Morgan, E. F.; Brown, K. A. A Bayesian Experimental Autonomous Researcher for Mechanical Design. *Sci. Adv.* **2020**, *6* (15), eaaz1708. https://doi.org/10.1126/sciadv.aaz1708.

(47) Duros, V.; Grizou, J.; Xuan, W.; Hosni, Z.; Long, D.-L.; Miras, H. N.; Cronin, L. Human versus Robots in the Discovery and Crystallization of Gigantic Polyoxometalates. *Angew. Chem. Int. Ed.* **2017**, *56* (36), 10815–10820. https://doi.org/10.1002/anie.201705721.

(48) Shekar, V.; Nicholas, G.; Najeeb, M. A.; Zeile, M.; Yu, V.; Wang, X.; Slack, D.; Li, Z.; Nega, P. W.; Chan, E. M.; Norquist, A. J.; Schrier, J.; Friedler, S. A. Active Meta-Learning for Predicting and Selecting Perovskite Crystallization Experiments. *J. Chem. Phys.* **2022**, *156* (6), 064108. https://doi.org/10.1063/5.0076636.

(49) Shekar, V.; Yu, V.; Garcia, B. J.; Gordon, D. B.; Moran, G. E.; Blei, D. M.; Roch, L. M.; García-Durán, A.; Ani Najeeb, M.; Zeile, M.; Nega, P. W.; Li, Z.; Kim, M. A.; Chan, E. M.; Norquist, A. J.; Friedler, S.; Schrier, J. Serendipity Based Recommender System for Perovskites Material Discovery: Balancing Exploration and Exploitation across Multiple Models. ChemRxiv July 19, 2022. https://doi.org/10.26434/chemrxiv-2022-l1wpf.

(50) Sun, S.; Tiihonen, A.; Oviedo, F.; Liu, Z.; Thapa, J.; Zhao, Y.; Hartono, N. T. P.; Goyal, A.; Heumueller, T.; Batali, C.; Encinas, A.; Yoo, J. J.; Li, R.; Ren, Z.; Peters, I. M.; Brabec, C. J.; Bawendi, M. G.; Stevanovic, V.; Fisher, J.; Buonassisi, T. A Data Fusion Approach to Optimize Compositional Stability of Halide Perovskites. *Matter* **2021**, *4* (4), 1305–1322. https://doi.org/10.1016/j.matt.2021.01.008.

(51) Pogue, E. A.; New, A.; McElroy, K.; Le, N. Q.; Pekala, M. J.; McCue, I.; Gienger, E.; Domenico, J.; Hedrick, E.; McQueen, T. M.; Wilfong, B.; Piatko, C. D.; Ratto, C. R.; Lennon, A.; Chung, C.; Montalbano, T.; Bassen, G.; Stiles, C. D. Closed-Loop Machine Learning for Discovery of Novel Superconductors. arXiv December 22, 2022. https://doi.org/10.48550/arXiv.2212.11855.

(52) Stach, E.; DeCost, B.; Kusne, A. G.; Hattrick-Simpers, J.; Brown, K. A.; Reyes, K. G.; Schrier, J.; Billinge, S.; Buonassisi, T.; Foster, I.; Gomes, C. P.; Gregoire, J. M.; Mehta, A.; Montoya, J.; Olivetti, E.; Park, C.; Rotenberg, E.; Saikin, S. K.; Smullin, S.; Stanev, V.; Maruyama, B. Autonomous Experimentation Systems for Materials Development: A Community Perspective. *Matter* **2021**, *4* (9), 2702–2726. https://doi.org/10.1016/j.matt.2021.06.036.

(53) Fujinuma, N.; DeCost, B.; Hattrick-Simpers, J.; Lofland, S. E. Why Big Data and Compute Are Not Necessarily the Path to Big Materials Science. *Commun Mater* **2022**, *3* (1), 1–9. https://doi.org/10.1038/s43246-022-00283-x.

(54) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573* (7773), 251–255. https://doi.org/10.1038/s41586-019-1540-5.

(55) Duros, V.; Grizou, J.; Sharma, A.; Mehr, S. H. M.; Bubliauskas, A.; Frei, P.; Miras, H. N.; Cronin, L. Intuition-Enabled Machine Learning Beats the Competition When Joint Human-Robot Teams Perform Inorganic Chemical Experiments. *J. Chem. Inf. Model.* **2019**, *59* (6), 2664–2671. https://doi.org/10.1021/acs.jcim.9b00304.

(56) Kauwe, S. K.; Graser, J.; Murdock, R.; Sparks, T. D. Can Machine Learning Find Extraordinary Materials? *Computational Materials Science* **2020**, *174*, 109498. https://doi.org/10.1016/j.commatsci.2019.109498.

(57) Borg, C. K. H.; Muckley, E. S.; Nyby, C.; Saal, J. E.; Ward, L.; Mehta, A.; Meredig, B. Quantifying the Performance of Machine Learning Models in Materials Discovery. *Digital Discovery* **2023**. https://doi.org/10.1039/D2DD00113F.

(58) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine Intelligence for Chemical Reaction Space. *WIREs Comput. Mol. Sci.* **2022**, *12* (5), 1604. https://doi.org/10.1002/wcms.1604.

(59) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.

(60) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem. Int. Ed.* **2020**, *59* (2), 725–730. https://doi.org/10.1002/anie.201912083.

(61) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial Intelligence and Automation in Computer Aided Synthesis Planning. *Reaction Chemistry & Engineering* **2021**, *6* (1), 27–51. https://doi.org/10.1039/D0RE00340A.

(62) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive Chemistry: Machine Learning for Reaction Deployment, Reaction Development, and Reaction Discovery. *Chem. Sci.* **2023**, *14* (2), 226–244. https://doi.org/10.1039/D2SC05089G.

(63) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Science Advances* **2021**, *7* (15), eabe4166. https://doi.org/10.1126/sciadv.abe4166.

(64) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. https://doi.org/10.1021/jacs.1c12005.

(65) Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; ACM: Virtual Event Canada, 2021; pp 610–623. https://doi.org/10.1145/3442188.3445922.

(66) Li, K.; Hopkins, A. K.; Bau, D.; Viégas, F.; Pfister, H.; Wattenberg, M. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. arXiv January 25, 2023. http://arxiv.org/abs/2210.13382 (accessed 2023-02-02).

(67) Delétang, G.; Ruoss, A.; Grau-Moya, J.; Genewein, T.; Wenliang, L. K.; Catt, E.; Cundy, C.; Hutter, M.; Legg, S.; Veness, J.; Ortega, P. A. Neural Networks and the Chomsky Hierarchy. arXiv October 6, 2022. https://doi.org/10.48550/arXiv.2207.02098.
(68) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 All You Need for Low-Data Discovery in Chemistry? ChemRxiv February 14, 2023. https://doi.org/10.26434/chemrxiv-2023-fw8n4.
(69) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-Loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling. *Science* **2022**, *378* (6618), 399–405. https://doi.org/10.1126/science.adc8743.
(70) Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. A Brief Introduction to Chemical Reaction Optimization. *Chem. Rev.* **2023**. https://doi.org/10.1021/acs.chemrev.2c00798.
(71) Xiang, X.-D.; Takeuchi, I. *Combinatorial Materials Synthesis*; CRC Press, 2003.
(72) Koinuma, H.; Takeuchi, I. Combinatorial Solid-State Chemistry of Inorganic Materials. *Nature Mater* **2004**, *3* (7), 429–438. https://doi.org/10.1038/nmat1157.
(73) Takeuchi, I.; Lippmaa, M.; Matsumoto, Y. Combinatorial Experimentation and Materials Informatics. *MRS Bulletin* **2006**, *31* (12), 999–1003. https://doi.org/10.1557/mrs2006.228.
(74) Green, M. L.; Choi, C. L.; Hattrick-Simpers, J. R.; Joshi, A. M.; Takeuchi, I.; Barron, S. C.; Campo, E.; Chiang, T.; Empedocles, S.; Gregoire, J. M.; Kusne, A. G.; Martin, J.; Mehta, A.; Persson, K.; Trautt, Z.; Van Duren, J.; Zakutayev, A. Fulfilling the Promise of the Materials Genome Initiative with High-Throughput Experimental Methodologies. *Applied Physics Reviews* **2017**, *4* (1), 011105. https://doi.org/10.1063/1.4977487.
(75) Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. *Acc. Chem. Res.* **2022**. https://doi.org/10.1021/acs.accounts.2c00220.
(76) Delgado-Licona, F.; Abolhasani, M. Research Acceleration in Self-Driving Labs: Technological Roadmap toward Accelerated Materials and Molecular Discovery. *Advanced Intelligent Systems n/a* (n/a), 2200331. https://doi.org/10.1002/aisy.202200331.
(77) Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. *Nat. Synth* **2023**, 1–10. https://doi.org/10.1038/s44160-022-00231-0.
(78) Leins, D. A.; Haase, S. B.; Eslami, M.; Schrier, J.; Freeman, J. T. Collaborative Methods to Enhance Reproducibility and Accelerate Discovery. *Digital Discovery* **2023**, *2*, 12–27. https://doi.org/10.1039/D2DD00061J.
(79) MacLeod, B. P.; Parlane, F. G. L.; Rupnow, C. C.; Dettelbach, K. E.; Elliott, M. S.; Morrissey, T. D.; Haley, T. H.; Proskurin, O.; Rooney, M. B.; Taherimakhsousi, N.; Dvorak, D. J.; Chiu, H. N.; Waizenegger, C. E. B.; Ocean, K.; Mokhtari, M.; Berlinguette, C. P. A Self-Driving Laboratory Advances the Pareto Front for Material Properties. *Nat Commun* **2022**, *13* (1), 995. https://doi.org/10.1038/s41467-022-28580-6.
(80) Sarker, S.; Tang-Kong, R.; Schoeppner, R.; Ward, L.; Hasan, N. A.; Van Campen, D. G.; Takeuchi, I.; Hattrick-Simpers, J.; Zakutayev, A.; Packard, C. E.; Mehta, A. Discovering Exceptionally Hard and Wear-Resistant Metallic Glasses by Combining Machine-Learning

with High Throughput Experimentation. *Applied Physics Reviews* **2022**, *9* (1), 011403. https://doi.org/10.1063/5.0068207.

(81) Wu, T. C.; Aguilar-Granda, A.; Hotta, K.; Yazdani, S. A.; Pollice, R.; Vestfrid, J.; Hao, H.; Lavigne, C.; Seifrid, M.; Angello, N.; Bencheikh, F.; Hein, J. E.; Burke, M.; Adachi, C.; Aspuru-Guzik, A. A Materials Acceleration Platform for Organic Laser Discovery. *Advanced Materials n/a* (n/a), 2207070. https://doi.org/10.1002/adma.202207070.

(82) Muckley, E. S.; Vasudevan, R.; Sumpter, B. G.; Advincula, R. C.; Ivanov, I. N. Machine Intelligence-Centered System for Automated Characterization of Functional Materials and Interfaces. *ACS Appl. Mater. Interfaces* **2023**, *15* (1), 2329–2340. https://doi.org/10.1021/acsami.2c16088.

(83) Herck, J. V.; Abeysekera, I.; Buckinx, A.-L.; Cai, K.; Hooker, J.; Thakur, K.; Reydt, E. V. de; Voorter, P.-J.; Wyers, D.; Junkers, T. Operator-Independent High-Throughput Polymerization Screening Based on Automated Inline NMR and Online SEC. *Digital Discovery* **2022**, *1* (4), 519–526. https://doi.org/10.1039/D2DD00035K.

(84) Montoya, J. H.; Aykol, M.; Anapolsky, A.; Gopal, C. B.; Herring, P. K.; Hummelshøj, J. S.; Hung, L.; Kwon, H.-K.; Schweigert, D.; Sun, S.; Suram, S. K.; Torrisi, S. B.; Trewartha, A.; Storey, B. D. Toward Autonomous Materials Research: Recent Progress and Future Challenges. *Appl. Phys. Rev.* **2022**, *9* (1), 011405. https://doi.org/10.1063/5.0076324.

(85) Li, Y.; Xia, L.; Fan, Y.; Wang, Q.; Hu, M. Recent Advances in Autonomous Synthesis of Materials. *ChemPhysMater* **2022**, *1* (2), 77–85. https://doi.org/10.1016/j.chphma.2021.10.002.

(86) Ishizuki, N.; Shimizu, R.; Hitosugi, T. Autonomous Experimental Systems in Materials Science. *Sci. Technol. Adv. Mater.: Methods* **2023**, 2197519. https://doi.org/10.1080/27660400.2023.2197519.

(87) Clayton, A. D.; Pyzer-Knapp, E. O.; Purdie, M.; Jones, M. F.; Barthelme, A.; Pavey, J.; Kapur, N.; Chamberlain, T. W.; Blacker, A. J.; Bourne, R. A. Bayesian Self-Optimization for Telescoped Continuous Flow Synthesis. *Angew Chem Int Ed* **2023**, *62* (3). https://doi.org/10.1002/anie.202214511.

(88) Dasgupta, S. Analysis of a Greedy Active Learning Strategy. In *Advances in Neural Information Processing Systems*; MIT Press, 2004; Vol. 17.

(89) Rohr, B.; S. Stein, H.; Guevarra, D.; Wang, Y.; A. Haber, J.; Aykol, M.; K. Suram, S.; M. Gregoire, J. Benchmarking the Acceleration of Materials Discovery by Sequential Learning. *Chemical Science* **2020**, *11* (10), 2696–2706. https://doi.org/10.1039/C9SC05999G.

(90) Liu, R.; Li, X.; Lam, K. S. Combinatorial Chemistry in Drug Discovery. *Curr Opin Chem Biol* **2017**, *38*, 117–126. https://doi.org/10.1016/j.cbpa.2017.03.017.

(91) Harada, M.; Takeda, H.; Suzuki, S.; Nakano, K.; Tanibata, N.; Nakayama, M.; Karasuyama, M.; Takeuchi, I. Bayesian-Optimization-Guided Experimental Search of NASICON-Type Solid Electrolytes for All-Solid-State Li-Ion Batteries. *Journal of Materials Chemistry A* **2020**, *8* (30), 15103–15109. https://doi.org/10.1039/D0TA04441E.

(92) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias Free Multiobjective Active Learning for Materials Design and Discovery. *Nat Commun* **2021**, *12* (1), 2312. https://doi.org/10.1038/s41467-021-22437-0.

(93) Fromer, J. C.; Coley, C. W. Computer-Aided Multi-Objective Optimization in Small Molecule Discovery. arXiv October 13, 2022. https://doi.org/10.48550/arXiv.2210.07209.
(94) Daulton, S.; Balandat, M.; Bakshy, E. Differentiable Expected Hypervolume Improvement for Parallel Multi-Objective Bayesian Optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*; NIPS'20; Curran Associates Inc.: Red Hook, NY, USA, 2020.
(95) Yaqub, O. Serendipity: Towards a Taxonomy and a Theory. *Research Policy* **2018**, *47* (1), 169–179. https://doi.org/10.1016/j.respol.2017.10.007.
(96) van Delft, D.; Kes, P. The Discovery of Superconductivity. *Physics Today* **2010**, *63* (9), 38–43. https://doi.org/10.1063/1.3490499.
(97) Cava, R. J. Oxide Superconductors. *J. Am. Ceramic Soc.* **2000**, *83* (1), 5–28. https://doi.org/10.1111/j.1151-2916.2000.tb01142.x.
(98) Bardeen, J.; Cooper, L. N.; Schrieffer, J. R. Microscopic Theory of Superconductivity. *Phys. Rev.* **1957**, *106* (1), 162–164. https://doi.org/10.1103/PhysRev.106.162.
(99) Brandt, N. B.; Moshchalkov, V. V. Concentrated Kondo Systems. *Advances in Physics* **1984**, *33* (5), 373–467. https://doi.org/10.1080/00018738400101681.
(100) Chevrel, R.; Hirrien, M.; Sergent, M. Superconducting Chevrel Phases: Prospects and Perspectives. *Polyhedron* **1986**, *5* (1–2), 87–94. https://doi.org/10.1016/S0277-5387(00)84891-3.
(101) Nguyen, N.; Er-Rakho, L.; Michel, C.; Choisnet, J.; Raveau, B. Intercroissance de feuillets "perovskites lacunaires" et de feuillets type chlorure de sodium: Les oxydes. *Materials Research Bulletin* **1980**, *15* (7), 891–897. https://doi.org/10.1016/0025-5408(80)90212-3.
(102) Er-Rakho, L.; Michel, C.; Provost, J.; Raveau, B. A Series of Oxygen-Defect Perovskites Containing Cu^{II} and Cu^{III}: The Oxides $La_{3-x}Ln_xBa_3$ $[Cu^{II}_{5-2y}\ Cu^{III}_{1+2y}]$ O_{14+y}. *J. Solid State Chem.* **1981**, *37* (2), 151–156. https://doi.org/10.1016/0022-4596(81)90080-3.
(103) Poeppelmeier, K. R.; Leonowicz, M. E.; Longo, J. M. $CaMnO_{2.5}$ and $Ca_{2MnO}3.5</sub>$: New Oxygen-Defect Perovskite-Type Oxides. *J. Solid State Chem.* **1982**, *44* (1), 89–98. https://doi.org/10.1016/0022-4596(82)90404-2.
(104) Reller, A.; Jefferson, D. A.; Thomas, J. M.; Beyerlein, R. A.; Poeppelmeier, K. R. Three New Ordering Schemes for Oxygen Vacancies in $CaMnO_{3-x}$ Superlattices Based on Square-Pyramidal Co-Ordination of Mn^{3+}. *J. Chem. Soc., Chem. Commun.* **1982**, *0* (24), 1378–1380. https://doi.org/10.1039/C39820001378.
(105) Reller, A.; Jefferson, D. A.; Thomas, J. M.; Uppal, M. K. Oxygen Vacancy Ordering in a Selective Oxidation Catalyst: Calcium Manganese Oxide ($CaMnO_{2.8}$). *J. Phys. Chem.* **1983**, *87* (6), 913–914. https://doi.org/10.1021/j100229a005.
(106) Reller, A.; Thomas, J. M.; Jefferson, D. A.; Uppal, M. K. Superstructures Formed by the Ordering of Vacancies in a Selective Oxidation Catalyst: Grossly Defective $CaMnO_3$. *Proc. R. Soc. Lond. A* **1984**, *394* (1807), 223–241. https://doi.org/10.1098/rspa.1984.0077.
(107) Mott, N. *Metal-Insulator Transitions*; CRC Press, 2004. https://doi.org/10.1201/b12795.
(108) Hulm, J. K.; Jones, C. K.; Hein, R. A.; Gibson, J. W. Superconductivity in the TiO and NbO Systems. *J Low Temp Phys* **1972**, *7* (3–4), 291–307. https://doi.org/10.1007/BF00660068.
(109) Johnston, D. C.; Prakash, H.; Zachariasen, W. H.; Viswanathan, R. High Temperature Superconductivity in the Li Ti O Ternary System. *Materials Research Bulletin* **1973**, *8* (7), 777–784. https://doi.org/10.1016/0025-5408(73)90183-9.

(110) Ostenson, J. E.; Shanks, H. R.; Finnemore, D. K. Superconductivity in the Tungsten Bronzes. *Journal of the Less Common Metals* **1978**, *62*, 149–153. https://doi.org/10.1016/0022-5088(78)90024-3.
(111) Sleight, A. W.; Gillson, J. L.; Bierstedt, P. E. High-Temperature Superconductivity in the $BaPb_{1-x}Bi_xO_3$ Systems. *Solid State Commun.* **1975**, *17* (1), 27–28. https://doi.org/10.1016/0038-1098(75)90327-0.
(112) Pfeiffer, E. R.; Schooley, J. F. Effect of Stress on the Superconducting Transition Temperature of $SrTiO_3$. *J Low Temp Phys* **1970**, *2* (3–4), 333–352. https://doi.org/10.1007/BF00652506.
(113) Michel, C.; Rakho, L. E.; Hervieu, M.; Pannetier, J.; Raveau, B. $BaLa_4Cu_5O_{13+\delta}$, an Oxygen-Deficient Perovskite Built up from Corner-Sharing CuO_6 Octahedra and CuO_5 Pyramids. *J. Solid State Chem.* **1987**, *68* (1), 143–152. https://doi.org/10.1016/0022-4596(87)90297-0.
(114) Bednorz, J. G.; Müller, K. A. Possible High Tc Superconductivity in the Ba-La-Cu-O System. *Z. Physik B - Condensed Matter* **1986**, *64* (2), 189–193. https://doi.org/10.1007/BF01303701.
(115) Wu, M. K.; Ashburn, J. R.; Torng, C. J.; Hor, P. H.; Meng, R. L.; Gao, L.; Huang, Z. J.; Wang, Y. Q.; Chu, C. W. Superconductivity at 93 K in a New Mixed-Phase Y-Ba-Cu-O Compound System at Ambient Pressure. In *Ten Years of Superconductivity: 1980–1990*; Ott, H. R., Ed.; Perspectives in Condensed Matter Physics; Springer Netherlands: Dordrecht, 1993; Vol. 7, pp 281–283. https://doi.org/10.1007/978-94-011-1622-0_36.
(116) Schilling, A.; Cantoni, M.; Guo, J. D.; Ott, H. R. Superconductivity above 130 K in the Hg–Ba–Ca–Cu–O System. *Nature* **1993**, *363* (6424), 56–58. https://doi.org/10.1038/363056a0.
(117) Villars, P.; Phillips, J. C. Quantum Structural Diagrams and High-T_c Superconductivity. *Phys. Rev. B* **1988**, *37* (4), 2345–2348. https://doi.org/10.1103/PhysRevB.37.2345.
(118) Stanev, V.; Oses, C.; Kusne, A. G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine Learning Modeling of Superconducting Critical Temperature. *npj Comput Mater* **2018**, *4* (1), 1–14. https://doi.org/10.1038/s41524-018-0085-8.
(119) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can Machine Learning Identify the next High-Temperature Superconductor? Examining Extrapolation Performance for Materials Discovery. *Molecular Systems Design & Engineering* **2018**, *3* (5), 819–825. https://doi.org/10.1039/C8ME00012C.
(120) Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; Meredig, B. High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates. *Integr Mater Manuf Innov* **2017**, *6* (3), 207–217. https://doi.org/10.1007/s40192-017-0098-z.
(121) Tokura, Y.; Torrance, J. B.; Nazzal, A. I.; Huang, T. C.; Ortiz, C. Discovery of a New, Metallic (but Not Superconducting) Compound in the Lanthanum-Strontium-Copper-Oxygen System: La5SrCu6O15. *J. Am. Chem. Soc.* **1987**, *109* (24), 7555–7557. https://doi.org/10.1021/ja00258a064.
(122) Zhang, Z.; Mansouri Tehrani, A.; Oliynyk, A. O.; Day, B.; Brgoch, J. Finding the Next Superhard Material through Ensemble Learning. *Adv. Mater.* **2021**, *33* (5), 2005112. https://doi.org/10.1002/adma.202005112.

(123) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140* (31), 9844–9853. https://doi.org/10.1021/jacs.8b02717.
(124) Hickey, J. C.; Brgoch, J. The Limits of Proxy-Guided Superhard Materials Screening. *Chem. Mater.* **2022**, *34* (22), 10003–10010. https://doi.org/10.1021/acs.chemmater.2c02390.
(125) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary Crystal Structure Prediction. *Computer Physics Communications* **2006**, *175* (11–12), 713–720. https://doi.org/10.1016/j.cpc.2006.07.020.
(126) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A Method for Crystal Structure Prediction. *Computer Physics Communications* **2012**, *183* (10), 2063–2070. https://doi.org/10.1016/j.cpc.2012.05.008.
(127) Lonie, D. C.; Zurek, E. XtalOpt: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Computer Physics Communications* **2011**, *182* (2), 372–387. https://doi.org/10.1016/j.cpc.2010.07.048.
(128) Zhang, Z.; Brgoch, J. Treating Superhard Materials as Anomalies. *J. Am. Chem. Soc.* **2022**, *144* (39), 18075–18080. https://doi.org/10.1021/jacs.2c07957.
(129) Feyerabend, P. *Against Method: Outline of an Anarchistic Theory of Knowledge*; NLB ; Humanities Press: London : Atlantic Highlands, 1975.
(130) Mroz, A. M.; Posligua, V.; Tarzia, A.; Wolpert, E. H.; Jelfs, K. E. Into the Unknown: How Computation Can Help Explore Uncharted Material Space. *J. Am. Chem. Soc.* **2022**, *144* (41), 18730–18743. https://doi.org/10.1021/jacs.2c06833.
(131) Muckley, E. S.; Saal, J. E.; Meredig, B.; Roper, C. S.; Martin, J. H. Interpretable Models for Extrapolation in Scientific Machine Learning. arXiv December 16, 2022. http://arxiv.org/abs/2212.10283 (accessed 2022-12-24).
(132) Rockafellar, R. T.; Uryasev, S. Optimization of Conditional Value-at-Risk. *J. Risk* **2000**, *2* (3), 21–41. https://doi.org/10.21314/JOR.2000.038.
(133) Grimmer, B. *Printable Topics in Mathematics*. A Collection of CVaR Norm Balls: Norms and Measures of Magnitude Focusing on the Most Extreme Values in a Data Set. https://www.ams.jhu.edu/~grimmer/CVaR.pdf (accessed 2023-02-04).
(134) Kikkawa, N.; Ohno, H. Materials Discovery Using Max K-Armed Bandit. arXiv December 15, 2022. http://arxiv.org/abs/2212.08225 (accessed 2023-02-02).
(135) Cicirello, V. A.; Smith, S. F. The Max K-Armed Bandit: A New Model of Exploration Applied to Search Heuristic Selection. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*; 2005; Vol. 3, pp 1355–1361.
(136) Mohamad, M. A.; Sapsis, T. P. Sequential Sampling Strategy for Extreme Event Statistics in Nonlinear Dynamical Systems. *Proceedings of the National Academy of Sciences* **2018**, *115* (44), 11138–11143. https://doi.org/10.1073/pnas.1813263115.
(137) Blanchard, A.; Sapsis, T. Bayesian Optimization with Output-Weighted Optimal Sampling. *Journal of Computational Physics* **2021**, *425*, 109901. https://doi.org/10.1016/j.jcp.2020.109901.

(138) Blanchard, A.; Sapsis, T. Output-Weighted Optimal Sampling for Bayesian Experimental Design and Uncertainty Quantification. *SIAM/ASA J. Uncertainty Quantification* **2021**, *9* (2), 564–592. https://doi.org/10.1137/20M1347486.
(139) Pickering, E.; Guth, S.; Karniadakis, G. E.; Sapsis, T. P. Discovering and Forecasting Extreme Events via Active Learning in Neural Operators. *Nat Comput Sci* **2022**, *2* (12), 823–833. https://doi.org/10.1038/s43588-022-00376-0.
(140) Lei, K.; Joress, H.; Persson, N.; Hattrick-Simpers, J. R.; DeCost, B. Aggressively Optimizing Validation Statistics Can Degrade Interpretability of Data-Driven Materials Models. *J. Chem. Phys.* **2021**, *155* (5), 054105. https://doi.org/10.1063/5.0050885.
(141) Kuhn, T. S. *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*; Harvard University Press: Cambridge (Mass.), 1957.
(142) Kepler, J. Astronomia Nova (1609). In *New Astronomy*; Donahue, W. H., Translator; Cambridge University Press: Cambridge, 1992; pp 47–48.
(143) Miyake, T. Underdetermination and Decomposition in Kepler's Astronomia Nova. *Studies in History and Philosophy of Science Part A* **2015**, *50*, 20–27. https://doi.org/10.1016/j.shpsa.2014.09.008.
(144) Drake, S. Galileo, Kepler, and Phases of Venus. *J. Hist. Astron.* **1984**, *15* (3), 198–208. https://doi.org/10.1177/002182868401500304.
(145) Bradley, J. An Account of a New Discovered Motion of the Fix'd Stars. *Philos. Trans. Royal Soc.* **1728**, *35* (406), 637–661. https://doi.org/10.1098/rstl.1727.0064.
(146) Wagner, N.; Rondinelli, J. M. Theory-Guided Machine Learning in Materials Science. *Front. Mater.* **2016**, *3*. https://doi.org/10.3389/fmats.2016.00028.
(147) Karniadakis, G. E.; Kevrekidis, I. G.; Lu, L.; Perdikaris, P.; Wang, S.; Yang, L. Physics-Informed Machine Learning. *Nat Rev Phys* **2021**, *3* (6), 422–440. https://doi.org/10.1038/s42254-021-00314-5.
(148) Cuomo, S.; di Cola, V. S.; Giampaolo, F.; Rozza, G.; Raissi, M.; Piccialli, F. Scientific Machine Learning through Physics-Informed Neural Networks: Where We Are and What's Next. arXiv June 7, 2022. http://arxiv.org/abs/2201.05624 (accessed 2023-02-02).
(149) Bødker, M. L.; Bauchy, M.; Du, T.; Mauro, J. C.; Smedskjaer, M. M. Predicting Glass Structure by Physics-Informed Machine Learning. *npj Comput Mater* **2022**, *8* (1), 1–9. https://doi.org/10.1038/s41524-022-00882-9.
(150) Tiihonen, A.; Cox-Vazquez, S. J.; Liang, Q.; Ragab, M.; Ren, Z.; Hartono, N. T. P.; Liu, Z.; Sun, S.; Zhou, C.; Incandela, N. C.; Limwongyut, J.; Moreland, A. S.; Jayavelu, S.; Bazan, G. C.; Buonassisi, T. Predicting Antimicrobial Activity of Conjugated Oligoelectrolyte Molecules via Machine Learning. *J. Am. Chem. Soc.* **2021**, *143* (45), 18917–18931. https://doi.org/10.1021/jacs.1c05055.
(151) Hartono, N. T. P.; Ani Najeeb, M.; Li, Z.; Nega, P. W.; Fleming, C. A.; Sun, X.; Chan, E. M.; Abate, A.; Norquist, A. J.; Schrier, J.; Buonassisi, T. Principled Exploration of Bipyridine and Terpyridine Additives to Promote Methylammonium Lead Iodide Perovskite Crystallization. *Crystal Growth & Design* **2022**. https://doi.org/10.1021/acs.cgd.2c00522.
(152) Neumann, P.; Cao, L.; Russo, D.; Vassiliadis, V. S.; Lapkin, A. A. A New Formulation for Symbolic Regression to Identify Physico-Chemical Laws from Experimental Data. *Chemical Engineering Journal* **2020**, *387*, 123412. https://doi.org/10.1016/j.cej.2019.123412.

(153) Wang, Y.; Wagner, N.; Rondinelli, J. M. Symbolic Regression in Materials Science. *MRS Communications* **2019**, *9* (3), 793–805. https://doi.org/10.1557/mrc.2019.85.
(154) Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W.-J. Simple Descriptor Derived from Symbolic Regression Accelerating the Discovery of New Perovskite Catalysts. *Nat Commun* **2020**, *11* (1), 3513. https://doi.org/10.1038/s41467-020-17263-9.
(155) Guo, Z.; Hu, S.; Han, Z.-K.; Ouyang, R. Improving Symbolic Regression for Predicting Materials Properties with Iterative Variable Selection. *J. Chem. Theory Comput.* **2022**, *18* (8), 4945–4951. https://doi.org/10.1021/acs.jctc.2c00281.
(156) Kamienny, P.-A.; d'Ascoli, S.; Lample, G.; Charton, F. End-to-End Symbolic Regression with Transformers. arXiv April 22, 2022. http://arxiv.org/abs/2204.10532 (accessed 2023-01-11).
(157) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, jacs.2c13467. https://doi.org/10.1021/jacs.2c13467.
(158) Brown, R. Consideration of the Origin of Herbert Simon's Theory of "Satisficing" (1933-1947). *Management Decision* **2004**, *42* (10), 1240–1256. https://doi.org/10.1108/00251740410568944.
(159) Artinger, F. M.; Gigerenzer, G.; Jacobs, P. Satisficing: Integrating Two Traditions. *Journal of Economic Literature* **2022**, *60* (2), 598–635. https://doi.org/10.1257/jel.20201396.
(160) Miconi, T. Evolution and Complexity: The Double-Edged Sword. *Artif Life* **2008**, *14* (3), 325–344. https://doi.org/10.1162/artl.2008.14.3.14307.
(161) He, T.; Huo, H.; Bartel, C. J.; Wang, Z.; Cruse, K.; Ceder, G. Inorganic Synthesis Recommendation by Machine Learning Materials Similarity from Scientific Literature. arXiv February 4, 2023. https://doi.org/10.48550/arXiv.2302.02303.
(162) Zhao, Y.; Siriwardane, E. M. D.; Wu, Z.; Fu, N.; Al-Fahdi, M.; Hu, M.; Hu, J. Physics Guided Deep Learning for Generative Design of Crystal Materials with Symmetry Constraints. *npj Comput Mater* **2023**, *9* (1), 1–12. https://doi.org/10.1038/s41524-023-00987-9.
(163) Long, T.; Fortunato, N. M.; Opahle, I.; Zhang, Y.; Samathrakis, I.; Shen, C.; Gutfleisch, O.; Zhang, H. Constrained Crystals Deep Convolutional Generative Adversarial Network for the Inverse Design of Crystal Structures. *npj Comput Mater* **2021**, *7* (1), 1–7. https://doi.org/10.1038/s41524-021-00526-4.
(164) Margraf, J. T.; Ulissi, Z. W.; Jung, Y.; Reuter, K. Heterogeneous Catalysis in Grammar School. *J. Phys. Chem. C* **2022**, *126* (6), 2931–2936. https://doi.org/10.1021/acs.jpcc.1c10285.
(165) Salley, D.; Keenan, G.; Grizou, J.; Sharma, A.; Martín, S.; Cronin, L. A Nanomaterials Discovery Robot for the Darwinian Evolution of Shape Programmable Gold Nanoparticles. *Nat Commun* **2020**, *11* (1), 2771. https://doi.org/10.1038/s41467-020-16501-4.
(166) Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Improved Decision Making with Similarity Based Machine Learning. arXiv November 23, 2022. https://doi.org/10.48550/arXiv.2205.05633.
(167) Seifrid, M.; Hickman, R. J.; Aguilar-Granda, A.; Lavigne, C.; Vestfrid, J.; Wu, T. C.; Gaudin, T.; Hopkins, E. J.; Aspuru-Guzik, A. Routescore: Punching the Ticket to More Efficient Materials Development. *ACS Cent. Sci.* **2022**, *8* (1), 122–131. https://doi.org/10.1021/acscentsci.1c01002.

(168) Mouret, J.-B.; Clune, J. Illuminating Search Spaces by Mapping Elites. arXiv April 19, 2015. https://doi.org/10.48550/arXiv.1504.04909.
(169) Viswanathan, G. M.; Afanasyev, V.; Buldyrev, S. V.; Havlin, S.; da Luz, M. G. E.; Raposo, E. P.; Stanley, H. E. Lévy Flights in Random Searches. *Physica A: Statistical Mechanics and its Applications* **2000**, *282* (1), 1–12. https://doi.org/10.1016/S0378-4371(00)00071-6.
(170) *Precision Chemistry*. https://pubs.acs.org/journal/pcrhej (accessed 2023-02-02).
(171) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*, Second edition.; Adaptive computation and machine learning series; The MIT Press: Cambridge, Massachusetts, 2018.
(172) Nega, P. W.; Li, Z.; Ghosh, V.; Thapa, J.; Sun, S.; Hartono, N. T. P.; Nellikkal, M. A. N.; Norquist, A. J.; Buonassisi, T.; Chan, E. M.; Schrier, J. Using Automated Serendipity to Discover How Trace Water Promotes and Inhibits Lead Halide Perovskite Crystal Formation. *Appl. Phys. Lett.* **2021**, *119* (4), 041903. https://doi.org/10.1063/5.0059767.
(173) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533* (7601), 73–76. https://doi.org/10.1038/nature17439.
(174) Napoletano, P.; Piccoli, F.; Schettini, R. Anomaly Detection in Nanofibrous Materials by CNN-Based Self-Similarity. *Sensors* **2018**, *18* (1), 209. https://doi.org/10.3390/s18010209.
(175) Mozaffari, M. H.; Tay, L.-L. Anomaly Detection Using 1D Convolutional Neural Networks for Surface Enhanced Raman Scattering. In *SPIE Future Sensing Technologies*; SPIE, 2020; Vol. 11525, pp 162–168. https://doi.org/10.1117/12.2576447.
(176) McNally, A.; Prier, C. K.; MacMillan, D. W. C. Discovery of an α-Amino C–H Arylation Reaction Using the Strategy of Accelerated Serendipity. *Science* **2011**, *334* (6059), 1114–1117. https://doi.org/10.1126/science.1213920.
(177) Dragone, V.; Sans, V.; Henson, A. B.; Granda, J. M.; Cronin, L. An Autonomous Organic Reaction Search Engine for Chemical Reactivity. *Nat Commun* **2017**, *8* (1), 15733. https://doi.org/10.1038/ncomms15733.
(178) Caramelli, D.; Granda, J. M.; Mehr, S. H. M.; Cambié, D.; Henson, A. B.; Cronin, L. Discovering New Chemistry with an Autonomous Robotic Platform Driven by a Reactivity-Seeking Neural Network. *ACS Cent. Sci.* **2021**, *7* (11), 1821–1830. https://doi.org/10.1021/acscentsci.1c00435.
(179) G. Pelkie, B.; D. Pozzo, L. The Laboratory of Babel: Highlighting Community Needs for Integrated Materials Data Management. *Digital Discovery* **2023**. https://doi.org/10.1039/D3DD00022B.
(180) Moot, T.; Isayev, O.; Call, R. W.; McCullough, S. M.; Zemaitis, M.; Lopez, R.; Cahoon, J. F.; Tropsha, A. Material Informatics Driven Design and Experimental Validation of Lead Titanate as an Aqueous Solar Photocathode. *Materials Discovery* **2016**, *6*, 9–16. https://doi.org/10.1016/j.md.2017.04.001.
(181) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.;

Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data* **2016**, *3* (1), 160018. https://doi.org/10.1038/sdata.2016.18.

(182) Scheffler, M.; Aeschlimann, M.; Albrecht, M.; Bereau, T.; Bungartz, H.-J.; Felser, C.; Greiner, M.; Groß, A.; Koch, C. T.; Kremer, K.; Nagel, W. E.; Scheidgen, M.; Wöll, C.; Draxl, C. FAIR Data Enabling New Horizons for Materials Research. *Nature* **2022**, *604* (7907), 635–642. https://doi.org/10.1038/s41586-022-04501-x.

(183) Thompson, M. W.; Gilmer, J. B.; Matsumoto, R. A.; Quach, C. D.; Shamaprasad, P.; Yang, A. H.; Iacovella, C. R.; McCabe, C.; Cummings, P. T. Towards Molecular Simulations That Are Transparent, Reproducible, Usable by Others, and Extensible (TRUE). *Molecular Physics* **2020**, *118* (9–10), e1742938. https://doi.org/10.1080/00268976.2020.1742938.

(184) Lotfi, S.; Zhang, Z.; Viswanathan, G.; Fortenberry, K.; Mansouri Tehrani, A.; Brgoch, J. Targeting Productive Composition Space through Machine-Learning-Directed Inorganic Synthesis. *Matter* **2020**, *3* (1), 261–272. https://doi.org/10.1016/j.matt.2020.05.002.

(185) Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **1926**, *14* (21), 477–485. https://doi.org/10.1007/BF01507527.

(186) Levin, I. NIST Inorganic Crystal Structure Database (ICSD), 2020. https://doi.org/10.18434/M32147.

(187) Seegmiller, C. C.; Baird, S. G.; Sayeed, H. M.; Sparks, T. D. Discovering Chemically Novel, High-Temperature Superconductors. ChemRxiv March 13, 2023. https://doi.org/10.26434/chemrxiv-2023-8t8kt.

(188) Lehman, J.; Stanley, K. O. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* **2011**, *19* (2), 189–223. https://doi.org/10.1162/EVCO_a_00025.

(189) Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; Efros, A. A. Large-Scale Study of Curiosity-Driven Learning. arXiv August 13, 2018. https://doi.org/10.48550/arXiv.1808.04355.

(190) Baranes, A.; Oudeyer, P.-Y. Active Learning of Inverse Models with Intrinsically Motivated Goal Exploration in Robots. *Robotics and Autonomous Systems* **2013**, *61* (1), 49–73. https://doi.org/10.1016/j.robot.2012.05.008.

(191) Péré, A.; Forestier, S.; Sigaud, O.; Oudeyer, P.-Y. Unsupervised Learning of Goal Spaces for Intrinsically Motivated Goal Exploration. arXiv October 9, 2018. https://doi.org/10.48550/arXiv.1803.00781.

(192) Grizou, J.; Points, L. J.; Sharma, A.; Cronin, L. A Curious Formulation Robot Enables the Discovery of a Novel Protocell Behavior. *Sci. Adv.* **2020**, *6* (5), eaay4237. https://doi.org/10.1126/sciadv.aay4237.

(193) Eysenbach, B.; Gupta, A.; Ibarz, J.; Levine, S. Diversity Is All You Need: Learning Skills without a Reward Function. arXiv October 9, 2018. https://doi.org/10.48550/arXiv.1802.06070.

(194) Pritchard, D.; Turri, J.; Carter, J. A. The Value of Knowledge. In *The Stanford Encyclopedia of Philosophy*; Zalta, E. N., Nodelman, U., Eds.; Metaphysics Research Lab, Stanford University, 2022.

(195) Jacovi, A. Trends in Explainable AI (XAI) Literature. arXiv January 13, 2023. https://doi.org/10.48550/arXiv.2301.05433.
(196) Dybowski, R. Interpretable Machine Learning as a Tool for Scientific Discovery in Chemistry. *New J. Chem.* **2020**, *44* (48), 20914–20920. https://doi.org/10.1039/D0NJ02592E.
(197) Pilania, G. Machine Learning in Materials Science: From Explainable Predictions to Autonomous Design. *Computational Materials Science* **2021**, *193*, 110360. https://doi.org/10.1016/j.commatsci.2021.110360.
(198) Ghiringhelli, L. M. Interpretability of Machine-Learning Models in Physical Sciences. arXiv April 21, 2021. http://arxiv.org/abs/2104.10443 (accessed 2023-01-20).
(199) Oviedo, F.; Ferres, J. L.; Buonassisi, T.; Butler, K. T. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Acc. Mater. Res.* **2022**, *3* (6), 597–607. https://doi.org/10.1021/accountsmr.1c00244.
(200) Zhong, X.; Gallagher, B.; Liu, S.; Kailkhura, B.; Hiszpanski, A.; Han, T. Y.-J. Explainable Machine Learning in Materials Science. *npj Comput Mater* **2022**, *8* (1), 1–19. https://doi.org/10.1038/s41524-022-00884-7.
(201) Falcon, A. Aristotle on Causality. In *The Stanford Encyclopedia of Philosophy*; Zalta, E. N., Ed.; Metaphysics Research Lab, Stanford University, 2022.
(202) Pearl, J.; Glymour, M.; Jewell, N. P. *Causal Inference in Statistics: A Primer*; Wiley: Chichester, West Sussex, 2016.
(203) Ting, J. Y.; Barnard, A. S. Data-Driven Causal Inference of Process-Structure Relationships in Nanocatalysis. *Current Opinion in Chemical Engineering* **2022**, *36*, 100818. https://doi.org/10.1016/j.coche.2022.100818.
(204) Vasudevan, R. K.; Ziatdinov, M.; Vlcek, L.; Kalinin, S. V. Off-the-Shelf Deep Learning Is Not Enough, and Requires Parsimony, Bayesianity, and Causality. *npj Comput Mater* **2021**, *7* (1), 16. https://doi.org/10.1038/s41524-020-00487-0.
(205) Kalinin, S. V.; Liu, Y.; Vasudevan, R.; Kelley, K.; Ziatdinov, M. From Event Detection to Physical Hypothesis Learning via Automated and Autonomous Microscopy. *Microscopy and Microanalysis* **2022**, *28* (S1), 2062–2063. https://doi.org/10.1017/S143192762200798X.
(206) Liu, Y.; Ziatdinov, M.; Kalinin, S. V. Exploring Causal Physical Mechanisms via Non-Gaussian Linear Models and Deep Kernel Learning: Applications for Ferroelectric Domain Structures. *ACS Nano* **2022**, *16* (1), 1250–1259. https://doi.org/10.1021/acsnano.1c09059.
(207) Bai, J.; Cao, L.; Mosbach, S.; Akroyd, J.; Lapkin, A. A.; Kraft, M. From Platform to Knowledge Graph: Evolution of Laboratory Automation. *JACS Au* **2022**, *2* (2), 292–309. https://doi.org/10.1021/jacsau.1c00438.
(208) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, Fourth edition.; Pearson series in artificial intelligence; Pearson: Hoboken, 2021.
(209) Menon, A.; Krdzavac, N. B.; Kraft, M. From Database to Knowledge Graph — Using Data in Chemistry. *Current Opinion in Chemical Engineering* **2019**, *26*, 33–37. https://doi.org/10.1016/j.coche.2019.08.004.
(210) Chalk, Stuart J. AI4SD Video: SciData: Semantic Representation of Scientific Data and Applications in Chemistry. **2022**. https://doi.org/10.5258/SOTON/AI3SD0237.
(211) Shin, M.; Kim, J.; van Opheusden, B.; Griffiths, T. L. Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty. *Proceedings of the National*

Academy of Sciences **2023**, *120* (12), e2214840120. https://doi.org/10.1073/pnas.2214840120.
(212) Kusne, A. G.; Yu, H.; Wu, C.; Zhang, H.; Hattrick-Simpers, J.; DeCost, B.; Sarker, S.; Oses, C.; Toher, C.; Curtarolo, S.; Davydov, A. V.; Agarwal, R.; Bendersky, L. A.; Li, M.; Mehta, A.; Takeuchi, I. On-the-Fly Closed-Loop Materials Discovery via Bayesian Active Learning. *Nat Commun* **2020**, *11* (1), 1–11. https://doi.org/10.1038/s41467-020-19597-w.
(213) Kalinin, S. V.; Ophus, C.; Voyles, P. M.; Erni, R.; Kepaptsoglou, D.; Grillo, V.; Lupini, A. R.; Oxley, M. P.; Schwenker, E.; Chan, M. K. Y.; Etheridge, J.; Li, X.; Han, G. G. D.; Ziatdinov, M.; Shibata, N.; Pennycook, S. J. Machine Learning in Scanning Transmission Electron Microscopy. *Nat Rev Methods Primers* **2022**, *2* (1), 1–28. https://doi.org/10.1038/s43586-022-00095-w.
(214) Ziatdinov, M.; Ghosh, A.; Wong, C. Y. (Tommy); Kalinin, S. V. AtomAI Framework for Deep Learning Analysis of Image and Spectroscopy Data in Electron and Scanning Probe Microscopy. *Nat Mach Intell* **2022**, *4* (12), 1101–1112. https://doi.org/10.1038/s42256-022-00555-8.
(215) Zubarev, D.; Mendes, C. R.; Brazil, E. V.; Cerqueira, R.; Schmidt, K.; Segura, V.; Ferreira, J. J.; Sanders, D. Toward Human-AI Co-Creation to Accelerate Material Discovery. arXiv November 5, 2022. https://doi.org/10.48550/arXiv.2211.04257.
(216) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583* (7815), 237–241. https://doi.org/10.1038/s41586-020-2442-2.
(217) Hammer, A. J. S.; Leonov, A. I.; Bell, N. L.; Cronin, L. Chemputation and the Standardization of Chemical Informatics. *JACS Au* **2021**, *1* (10), 1572–1587. https://doi.org/10.1021/jacsau.1c00303.
(218) Manzano, J. S.; Hou, W.; Zalesskiy, S. S.; Frei, P.; Wang, H.; Kitson, P. J.; Cronin, L. An Autonomous Portable Platform for Universal Chemical Synthesis. *Nat. Chem.* **2022**, *14* (11), 1311–1318. https://doi.org/10.1038/s41557-022-01016-w.
(219) Koscher, B.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S.-C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F. Autonomous, Multi-Property-Driven Molecular Discovery: From Predictions to Measurements and Back. ChemRxiv April 13, 2023. https://doi.org/10.26434/chemrxiv-2023-r7b01.
(220) Breen, C. P.; Nambiar, A. M. K.; Jamison, T. F.; Jensen, K. F. Ready, Set, Flow! Automated Continuous Synthesis and Optimization. *Trends in Chemistry* **2021**, *3* (5), 373–386. https://doi.org/10.1016/j.trechm.2021.02.005.
(221) Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: Autonomous Discovery and Optimization of Multi-Step Chemistry Using a Self-Driven Fluidic Lab Guided by Reinforcement Learning. *Nat Commun* **2023**, *14* (1), 1403. https://doi.org/10.1038/s41467-023-37139-y.
(222) McGee, S.; Lei, Y.; Goff, J.; Wilkinson, C. J.; Nova, N. N.; Kindle, C. M.; Zhang, F.; Fujisawa, K.; Dimitrov, E.; Sinnott, S. B.; Dabo, I.; Terrones, M.; Zarzar, L. D. Single-Step Direct Laser Writing of Multimetal Oxygen Evolution Catalysts from Liquid Precursors. *ACS Nano* **2021**, *15* (6), 9796–9807. https://doi.org/10.1021/acsnano.1c00650.

(223) McGee, S.; Fest, A.; Chandler, C.; Nova, N. N.; Lei, Y.; Goff, J.; Sinnott, S. B.; Dabo, I.; Terrones, M.; Zarzar, L. D. Direct Laser Writing of Multimetal Bifunctional Catalysts for Overall Water Splitting. *ACS Appl. Energy Mater.* **2023**, *6* (7), 3756–3768. https://doi.org/10.1021/acsaem.2c03973.
(224) Gregoire, J. M.; Zhou, L.; Haber, J. A. Combinatorial Synthesis for AI-Driven Materials Discovery. *Nat. Synth* **2023**, 1–12. https://doi.org/10.1038/s44160-023-00251-4.
(225) Liu, G.; Petrosko, S. H.; Zheng, Z.; Mirkin, C. A. Evolution of Dip-Pen Nanolithography (DPN): From Molecular Patterning to Materials Discovery. *Chem. Rev.* **2020**, *120* (13), 6009–6047. https://doi.org/10.1021/acs.chemrev.9b00725.
(226) Liu, G.; Rong, M.; Hu, H.; Chen, L.; Xie, Z.; Zheng, Z. 3D Dip-Pen Nanolithography. *Advanced Materials Technologies* **2022**, *7* (8), 2101493. https://doi.org/10.1002/admt.202101493.
(227) McIntyre, D.; Lashkaripour, A.; Fordyce, P.; Densmore, D. Machine Learning for Microfluidic Design and Control. *Lab on a Chip* **2022**, *22* (16), 2925–2937. https://doi.org/10.1039/D2LC00254J.
(228) Kusne, A. G.; McDannald, A. Scalable Multi-Agent Lab Framework for Lab Optimization. *Matter* **2023**, *0* (0). https://doi.org/10.1016/j.matt.2023.03.022.
(229) Boiko, D. A.; MacKnight, R.; Gomes, G. Emergent Autonomous Scientific Research Capabilities of Large Language Models. arXiv April 11, 2023. https://doi.org/10.48550/arXiv.2304.05332.