JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Perspective

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

Joshua Schrier,* Alexander J. Norquist,* Tonio Buonassisi,* and Jakoah Brgoch*

Cite This: *J. Am. Chem. Soc.* 2023, 145, 21699−21716
Read Online
ACCESS | Metrics & More | Article Recommendations

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 the realization of 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 oxide 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 are 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, which 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 (and scientific discovery more generally[16]). However, these applications have been demonstrated mainly in the context of incremental improvements and optimization. *Incremental* does not mean *easy*, and ML optimizations are often in high-dimensional spaces that would have otherwise required months or years of traditional experimentation to achieve the same results.[17−20] However, transformative discoveries have seldom been achieved by this approach.

In this perspective, we suggest that there are fundamental limitations hindering the application of ML to the discovery of *exceptional materials* that shift the research paradigm (in the Kuhnian sense[21]). 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 how existing ML methods have contributed to these efforts. We then provide six recommended research directions for ML that address this challenge. Finally, we conclude with a vision of a future materials research process implementing these research directions.

II. THE CHALLENGE OF THE EXCEPTIONAL

II.A. What is Exceptional? We define an *exceptional material* or *molecule* as one that enables new scientific exploration because it is out of the prior distribution of properties, composition, or application. These are *black swan events*—unpredicted surprises that have a significant effect on the field but are only rationalized only after the first observation.[22] For example, the discovery of high-T_c cuprate superconductors rejected the conventional wisdom of condensed matter physics; even though the initial T_c was not exceptionally high, it achieved modest values with an unusual composition and had a synthetic toolbox that allowed for further exploration which would rapidly transform 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.[23]

Published: September 27, 2023

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

Many technologies require materials that can withstand coupled extremes, such as simultaneous mechanical, thermal, radiation, and corrosive attack for next-generation nuclear reactors[15,24] or simultaneous high photoconversion efficiency and mechanical durability for photovoltaics.[25] Exceptionality may thus comprise not just one, but a constellation of potentially mutually exclusive properties.[15,26] These trade-offs can be purely empirical trends observed within a materials class, for example, expressed as Ashby plots (solid mechanics) or Robeson plots (in membrane separation materials); in this case, exceptionality is merely a novel observation. Alternatively, the trade-offs may be first-order approximations to the underlying rigorous theoretical relationships, such as the Wiedemann−Franz proportionality of electrical and thermal conductivity. An exceptional thermoelectric material requires simultaneously high electrical conductivity and low thermal conductivity and thus violating the underlying physical assumptions that contraindicate these relationships. This is a distinct problem from multiobjective optimization,[27−29] discussed in Section III.C.

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 electronegativity constraints, there are approximately 10^{10} compositions,[32] greatly exceeding 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),[30] and thermodynamic stability network calculations indicating an increasing trend in the discovery rate of new materials.[34] "Interesting" materials are often not distributed evenly across parameter space. For example, an investigation of possible high-entropy alloys found wide disparities in compound space group (Figure 1b).[31] Even knowledge of what materials *can* be made is limited, as many observed materials are metastable,[35] and computational thermochemistry data sets have biased distributions of formation energies for different structure types.[36] Furthermore, nearly all experimental and computational data consider low-pressure systems, yet chemical bonding and periodic trends are radically different at high pressures relevant to materials under extreme conditions.[37] 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 to the hole with (ideally) a few attempts by taking advantage of the landscape and "reading" its many properties. Similarly, research problems are easy if there exists a clear gradient toward the goal (by analogy, this may arise from the inherent topography which causes balls to roll toward the hole or the golfer's mental map of the course). It is harder if there are many traps where gradient-based heuristics fail. More formally, mathematicians have devised many ways to characterize 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 end point).[38] 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).[39] 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.[38,40] Practical algorithms can efficiently find solutions even when the response surface only approximately obeys these criteria.

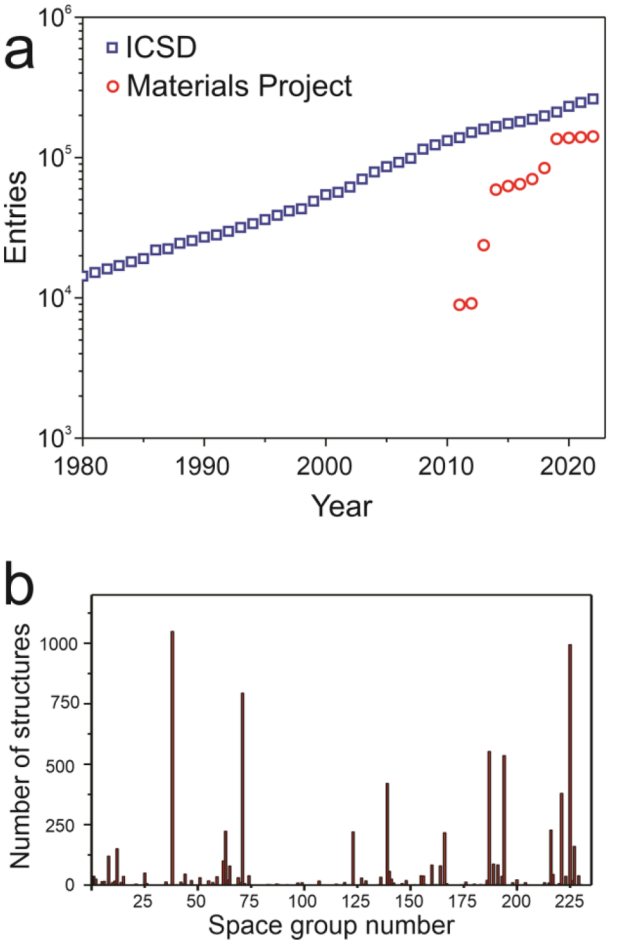

Figure 1. (a) Structure entries in the ICSD and Materials Project as a function of time, plotted on a logarithmic scale, adapted from ref 30. (b) Space group frequency for high-entropy alloys, adapted from ref 31.

Empirically, many successful materials ML problems are approximately smooth and convex response surfaces, with a broad basin of attraction toward a few local optima,[41] like the schematic example plotted in Figure 2a. Thus, it is unsurprising that ML-based approaches for representing the landscape can be successful, and iterative optimization is an efficient strategy. In contrast, exceptional materials are often comprised of much harder "needle in a haystack" problems,[42] 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 for the behavior of new candidates. A mathematician would precondition the problem into a more suitable form, but experimental scientists

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

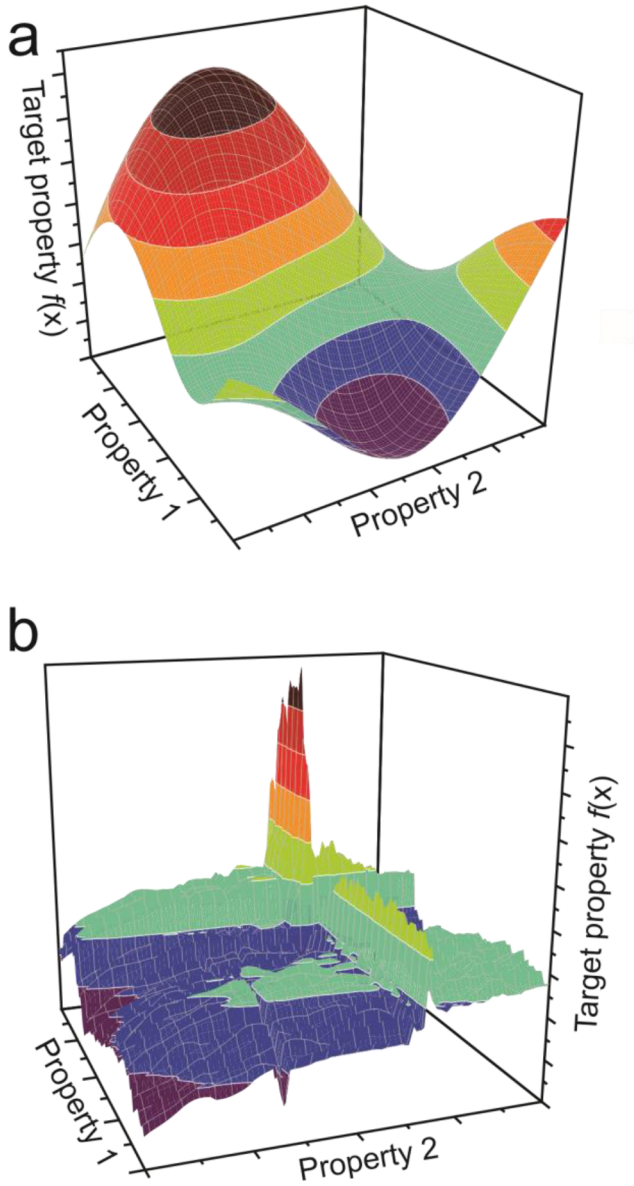

Figure 2. Iterative versus exceptional materials. (a) Previous work has focused on optimizations on smooth, convex response surfaces; (b) Exceptional material properties are often characterized by very sharp discontinuities as composition and reaction conditions are changed. Adapted from ref 41.

typically lack prior knowledge about the nature of the response function needed to apply an *a priori* transformation. 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.

III. THE STATE OF CURRENT MACHINE LEARNING APPROACHES

III.A. Computational Property Screening and Early ML. Multiagency funding efforts like the Materials Genome Initiative (MGI),[6] and similar efforts worldwide,[44] 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 early efforts used high-throughput density functional theory (DFT) calculations to create databases (e.g., Materials Project, AFLOW, OQMD, etc.). To expedite this process, researchers began using ML to perform preliminary regressions or make classification predictions to screen known crystal structure databases for materials with superb properties.[45] The efforts have since expanded dramatically using experimental data sets, autonomous experimentation, multitask, and transfer learning, among numerous other approaches. New tools, such as large-language model (LLM)-based code generators, eliminate technical barriers for nonexperts to perform routine data analysis tasks.[46] Surprisingly, general purpose LLMs can even be used to directly predict molecular and material properties using small amounts of example data.[47] But while predictions are numerous, experimental validation is quite rare.[45] Nevertheless, in nearly every example, the ML 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 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 these data to accelerate the screening process, thereby reducing the time it takes to map the response surface and achieve the desired result.[48] Incorporating model uncertainty with the predicted outcome enables an algorithmic guide to achieve the 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.[49] A random forest model trained on a small initial data set served as a proxy for the dependence of observed nanotube growth rate as a function of laser heating and the partial pressures of four gases. Active learning methods were used to sample uncertain new experimental conditions, and the algorithm was then employed to select the optimal set of input parameters to rapidly achieve the desired growth rate. Other illustrative ML-enhanced materials optimization examples include nanocrystal growth and optical properties in a microfluidic system,[50] mechanical properties of 3d-printed structures,[51] crystal growth conditions,[52−54] halide alloy stability,[20,55] and superconductivity.[56] See refs 45 and 57 for more comprehensive reviews.

Limitations of data-driven strategies have been noted in the literature, with the need for more data or higher-quality data being stressed.[58] Algorithmic performance can also depend on the initial data set (the "cold start" problem), and available data sets often exhibit sampling biases.[59] This problem can be partially mitigated by adding additional constraints to maximize the explored input space[54] or by incorporating human expertise in the loop.[60] While previous research articles have benchmarked computational methods and metrics for this task,[61,62] and a recent perspective discussed types of machine-learning guided iterative experimentation toward 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.[2,63] There is tremendous power (and computational complexity) associated with selecting a sequence of known reactions into a new arrangement. This presents an

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

immense combinatorial challenge where ML-derived heuristics can make the problem tractable,[64,65] with recent reviews discussing these efforts.[66,67] Although it has been suggested that deep-learning-based template-free methods can propose genuinely inventive new reactions,[68] performance can be poor outside the training set (even for undergraduate textbook reactions[69]) and predictions often reflect the most common reactions in the training set rather than optimal reactions.[70] Increasingly these make use of LLMs, which some AI researchers characterize as "stochastic parrots" because of a 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.[71] However, empirical evidence suggests that suitably trained LLMs can learn meaningful internal representations of a variety of problems.[72,73] 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.[68] Yet the problem of optimizing the reaction conditions and stoichiometries remains. 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[74] and reviewed more comprehensively in ref 75. 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.[76−79] Increasingly, this takes the form of closed-loop autonomous research systems or "self-driving laboratories[57,80−82] like the ARES system[49] mentioned above. Whether fully autonomous or not, each step in the workflow of design, synthesis, characterization, and optimization (also referred as design/build/test/learn) can be accelerated by incorporating ML tools.[83]

The synergies between ML and HTE are illustrated 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 increases 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. Autonomy 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. The improvement is problem-dependent; at best, iterative active learning requires only a logarithm of the number of experiments required by random sampling, but at worst may require the full number of sample points.[84] Empirical materials science studies have observed that poorly implemented active learning can decrease p.[85] The need for model updates in active learning can also interfere with parallelization of N. Second, autonomy reduces the delay between data acquisition and use of the improved model to acquire the next experiment; this is analogous to compounding interest more frequently. Third, these systems increase the volume and quality of experimental (meta)data, which facilitates its use for ML. By eliminating (unrecorded) human variations, automated processing can potentially improve the reproducibility of experiments, thus increasing the signal-to-noise ratio in the data set. Such unrecorded unintentional variations in background conditions

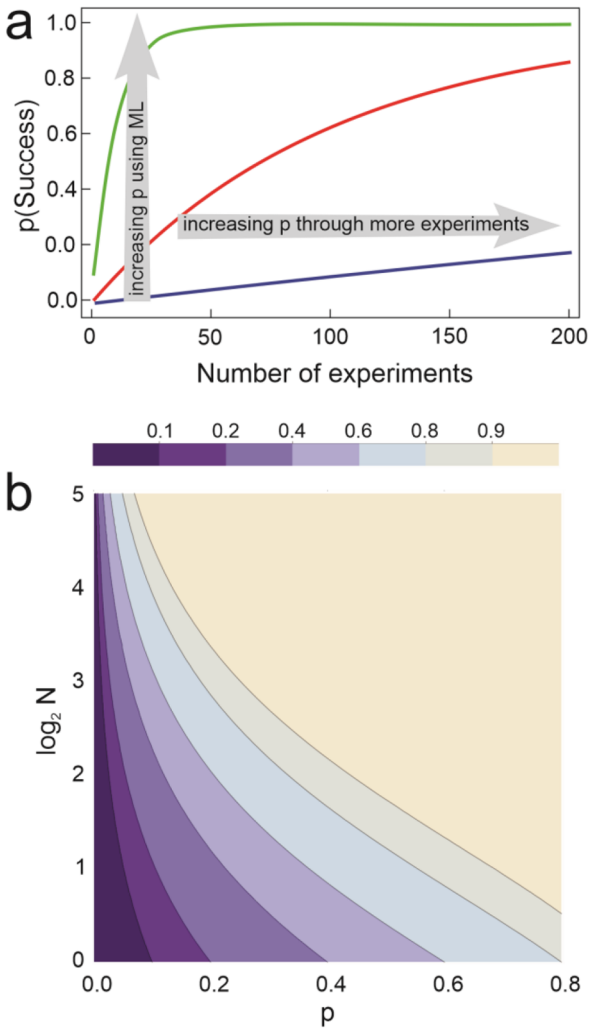

Figure 3. (a) The probability of at least one success in an experimental campaign can be increased by using ML to increase the probability that an experiment is successful, (increasing p), or by making more attempts (increasing N); (b) Contours showing the probability of at least one success, as a function of changing p and N. Increasing p and N has a synergistic effect, but a large value of one can compensate for a smaller value of the other.

are also minimized by performing more experiments in a shorter span of time.

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.[86] The lack of clinical successes from initial high-throughput synthesis in the 1980s suggests that merely increasing the N is insufficient. The incorporation of computational chemistry methods and informatics modeling in the 1990s increased p needed for success. Ultimately, p depends on how the problem is framed and determines where we look. Section V describes paths toward ML models that increase p for exceptional materials, rather than being limited to local optimizations.

III.D. Limitations of Pareto-front Multiobjective Optimization Strategies. For many applications, exceptionality requires not just a single property but a balance of multiple, possibly conflicting, objectives. When the objectives do not have an intrinsic priority, there is no single "best" solution but rather a

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

set of optimal solutions, described in terms of a *Pareto frontier*[15,26] (and ref 38 pp. 177−184), depicted schematically in Figure 4. Multiobjective optimization methods attempt to move the Pareto front forward, and have been applied recently across materials science ranging from solid-state battery electrolytes[87] to magnetic high-entropy alloys[3] to additive manufacturing[28] to polymer design,[88] and recent reviews discuss multiobjective optimization for organic molecules[89] and chemical reaction optimization.[75]

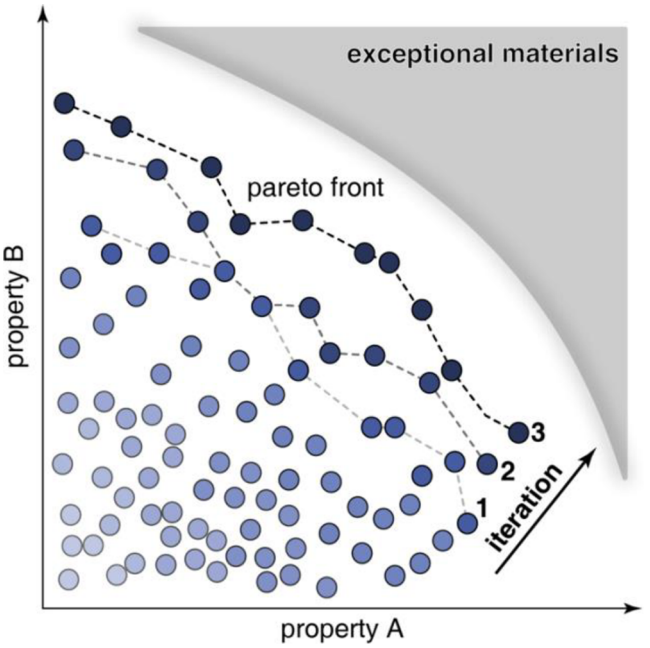

Figure 4. Pareto front is a plot of optimal solutions for a multiobjective optimization problem. The *x*-axis represents one objective, and the *y*-axis represents another objective. Points on the plot are Pareto-optimal solutions, with no other solutions better in both objectives. The Pareto front connects 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 nondominated. The Pareto front moves forward with each iteration making it a useful tool for decision-making and optimization, but not for identifying exceptional cases.

A wide variety of methodologies focus on different aspects of the problem; for example Chimera[27] handles constrained design spaces, and ε-PA[88] uses active learning to identify Pareto points in fewer evaluations. The strengths and limitations are nicely illustrated by a recent study which optimized the combustion synthesis (fuel source, fuel-to-oxidizer ratio, precursor solution concentration, and annealing temperature) of metal thin films to simultaneously maximize the film's conductivity and minimize the combustion temperature using a self-driving laboratory.[90] The differential expected hypervolume improvement (qEHVI) algorithm[91] employed allows monitoring of the exploration progress; in this study the normalized hypervolumes increase smoothly as the property response is explored, indicating stepwise advances. This type of stepwise, continual advances of an objective that balance trade-offs does not lend itself to the necessary "leaps-and-bounds" advances required for transformational discoveries and may have difficulty scaling.

IV. CASE STUDIES IN THE DISCOVERY OF EXCEPTIONAL MATERIALS

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

IV.A. Case Study: Serendipity and the Discovery of High-T_c Superconductors. Unexpected outcomes or applications are often the first step in scientific discovery.[92] The history of high-T_c superconductors illustrates the role of serendipity in exceptional material discovery. Research in superconductivity, from its initial report in 1911[93] to 1986, was dominated by metallic systems.[94] Conventional wisdom suggested that superconductors should be metallic, have high symmetries and electronic densities of state, and be structurally unstable or metastable, leading to a focus on vanadium and niobium alloys. However, investigation of systems for which BCS theory[95] did not work, such as intermetallics[96] and Chevrel phases,[97] motivated wider exploration, even if their critical temperatures were modest.

Parallel efforts in the early 1980s were critically important for the emergence of cuprate perovskite high temperature superconductors. First, Raveau,[98,99] Poeppelmeier,[100,101] and Thomas[102,103] were developing a synthetic toolbox to control 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 nonmetallic systems pre-1986 oxides,[104] including NbO[105] (which contains square planes, much like all cuprate superconductors), spinels ($LiTi_2O_4$),[106] a series of perovskite adjacent tungsten bronzes,[107] $Ba(Pb_{1-x}Bi_x)O_3$,[108] and even the perovskite $SrTiO_3$.[109]

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,[110] inspired their discovery of superconductivity in the LaBaCuO systems (T_c = 28 K).[111] Ceramics are generally insulators, but this anomalous case provided a new solution to an established problem via an unexpected route. The initial announcement motivated a decade of incremental optimization, during which superconductivity above the temperature of liquid nitrogen was quickly achieved in an YBaCuO oxide,[112] as well as the discovery of the reigning HgBaCaCuO cuprate superconductor.[113]

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 an 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.[56] 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 analysis (Figure 2 in their paper) does not predict BaCaCuO,[114] and it is unclear to what extent many other materials would be false positives. Three decades later, Stanev et al. used the SuperCon database of

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

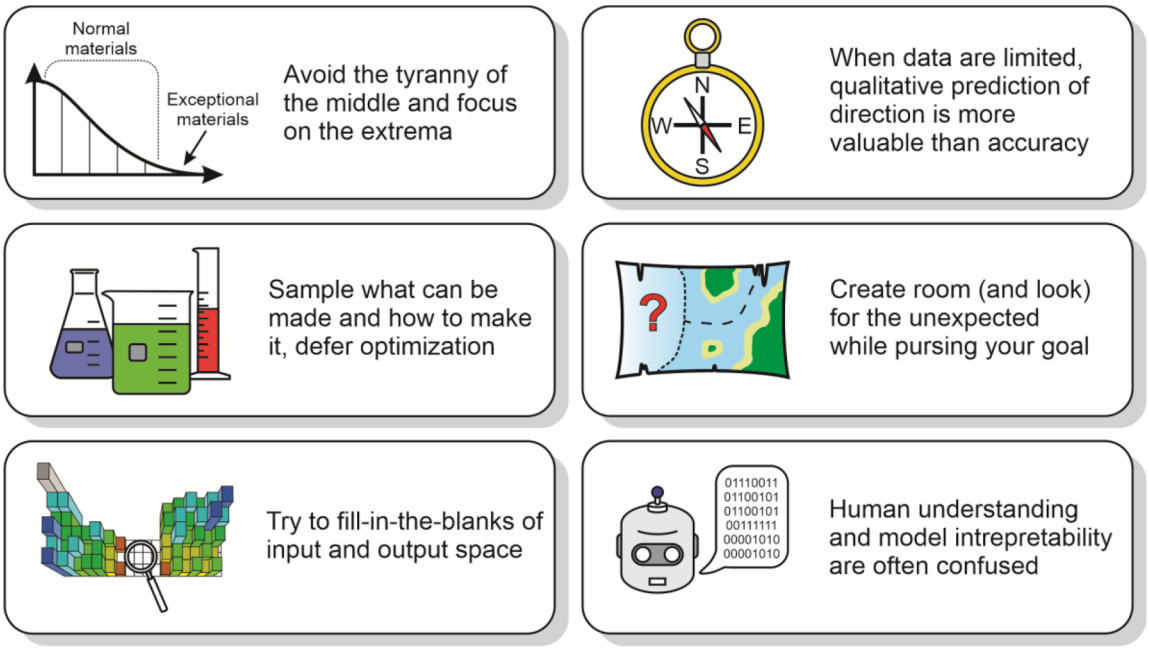

Figure 5. Six recommendations for research toward machine learning for exceptional materials.

over 16,000 compounds to train random forest models for predicting the T_c based solely on composition.[115] While they did not consider a time-separated holdout, Figure 4b in ref 115 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 suggest that a few initial discoveries suffice for ML to identify other examples; indeed, once cuprates are included in the training data set, they comprise the vast majority of candidate superconductors. Meredig et al. also observed that ML models trained without cuprate examples predict cuprates to be below-average superconductor[116] (see Figure 2 in ref 116). 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.[117] 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.[118] Overall, this supports the claims made in Section III: existing ML approaches can assist materials optimization but do not identify new exceptional materials.

IV.B. Case Study: Machine Learning and the Discovery of Superhard Materials. Similar strengths and limitations of the current ML apply to the discovery of superhard materials, defined as those having Vickers hardness (H_v) exceeding 40 GPa. Diamond is the hardest known naturally occurring substance ($H_v \approx 100$ GPa), and significant efforts have gone into making synthetic diamonds. In 1954, scientists at General Electric (GE) Research Laboratory developed the first approach involving the subjecting of graphite to intense heat and pressure using a diamond press. GE continued to refine the process through inventions in the 1970s like high-pressure, high-temperature (HPHT) synthesis, which subject 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 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 aerospace, automotive, and manufacturing industries. Given the tremendous application space, it is no surprise that 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 coworkers provides insight into 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 Pearson's Crystal Data set 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.[119] Not only is superhardness rare, the total data set of experimentally hardness measurements is relatively small (about 500 unique compositions).[119,120] Moreover, the data set is biased to low hardness values and certain compositions (such as boron-containing compounds) are disproportionately present in this limited training data,[119] attributable to the same types of anthropogenic research biases observed in other experimental materials data sets.[59]

The limited experimental data might suggest that physics-based simulations could be a more appropriate path toward materials discovery. However, atomistic DFT is 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

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

calculated by DFT (such as bulk and sheer moduli) as either initial selection criteria[120] or as inputs to semiempirical expressions for hardness.[121] Researchers have further paired these methods with crystal structure prediction algorithms (USPEX,[122] CALYPSO,[123] XtalOpt[124]) to predict new promising superhard compounds. ML can also be used to expand the search space enabled by DFT calculations.[120] These physically motivated models provide some guidance but are generally worse at quantitative hardness predictions than direct ML methods.[119,121] 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 captured hardness changes 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 values of these systems fall 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.[125] In this work, an autoencoder model was trained to find low-dimensional latent representations of the 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 TOWARD ML FOR EXCEPTIONAL MATERIALS

There is no single "scientific method", and scientific advances often involve rejecting established norms.[126] In that spirit of epistemological anarchism, we offer six maxims for guiding the research community, depicted schematically in Figure 5. While these recommendations focus primarily on *experimental* discovery, many are equally applicable to autonomous computational discovery.[127] When possible, we illustrate these points with applications in chemistry and materials science, but in many cases, we draw instead upon examples from finance, oceanography, computer science, and evolutionary biology.

V.A. Avoid the tyranny of the middle and focus on extrema. By definition, there are 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 exceptional extrema. Common ML metrics (accuracy, R^2, etc.) do not express the intended goal when in the presence of such outcome imbalances,[128] nor do they measure an algorithm's ability to guide iterative discovery.[62] 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,[129] which corresponds to a 1-norm of q% largest magnitude entries.[130] Alternatively, it may require modifications to existing algorithms. Typical reinforcement learning (RL) formulations maximize cumulative reward rather than the best possible result found.[131] Alternative problem formulations, such as the Max-k-arm bandit model,[132] better align with the goals of scientific discovery, as 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.[131] 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,[133−135] which has been recently applied to extreme event discovery in epidemiological models, rogue waves, and structure mechanics.[136]

V.B. When data are 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 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.[116,137] 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. Do not build a perfect model with limited data—the important thing is to collect more data, and the right data. An example of the importance of direction over accuracy exists in the initial reports of YBaCuO. Chu and co-worker's initial report of superconductivity at 93 K contained neither the formula nor structure of the new phase.[112,138] This report did, however, provide more than enough direction for solid-state chemists and ceramists to quickly identify the phase in question and initiate a decade of intense research.

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.[139] 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."[140] 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 the ability to *extract qualitative hypotheses* about individual planets and sufficient quantitative accuracy to guide observation.[141] Specifically, Kepler's model contained a latent hypothesis that Venus should have phases, and enabled sufficiently accurate calculations to direct Galileo's experimental observations disproving the Ptolemaic model.[142] Ruling out the Tychonic model required improved instrumentation and data collection to enable the observation of stellar aberration,[143] analogous to HTE for 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 the admissible functional forms of their interactions. Physics-based computer simulations serve a similar role,[127] although the examples above indicate their limits for exceptional materials.[58,121] We focus purely on data-driven approaches. Strategies of *physics-informed machine learning*[144−146] 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.[147] However, this is less applicable to exceptional materials which involve new physics precluded by using existing models as constraints (e.g., using BCS theory[95] to inform your

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

ML model will hinder discovering cuprate superconductors). *Feature selection* corresponds to an implied constraint that only a small subset of the input variables determine the system performance. The identified features are combined with simple models to make predictions. Some examples include the aforementioned synthesis of superhard materials,[120] but other examples include discovery of antimicrobial conjugated oligoelectrolytes[148] and perovskite crystal growth modifying additives.[149] Once hypothetically relevant features correlated to the output are selected, relatively simple models can be constructed to make extrapolations. Even simple linear models can be quite effective for this purpose.[128] 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[150] and materials science.[151,152] In practice, symbolic regression is often combined with various feature selection methods, with examples including VS-SISSO[153] and transformer-based approaches for symbolic regression.[154]

Emphasizing the qualitative direction has consequences for the design of HTE systems. Early stage validation might emphasize rapid (but potentially noisy) experimental 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 handoff 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.[92] For example, conductivity measurements in the LaBaCuO system revealed metallic behavior, foreshadowing 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.

V.C. 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; methods and applications of generative ML to chemical problems have very recently been reviewed in ref 155. 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.[156,157] 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 satisfactory 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.[158] For this reason, a collection of satisfactory 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.[159] (A parallel discussion of these ideas as they apply to organic and medicinal chemistry can be found in refs 155 and 160.)

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.[119] 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 that are ultimately discarded by the subsequent process. This might range from including physics-based symmetry contraints,[161] directly incorporating a learned formation energy constraint into the generative process,[162] or by restricting the generating samples to obey compositional "grammatical" rules.[163] The combination of empirical synthetic accessibility metrics, fragment- and synthesis-based constraints, and forward and reverse synthesis prediction to constrain generative models for drug design[160] can serve as a model for materials chemists. 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 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.[164]

The important thing is that the ML model leads to samples in the right neighborhood. One framework is similarity-based kernel learning, 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.[165] The ease of acquisition can be computed by combining materials, labor, and time constraints.[166] Another framework is provided by the Multidimensional 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.[167] Zooming-based Bayesian optimizations have a similar alternation between global sampling and local optimization.[41] This is also reminiscent of Lévy flight models of animal foraging

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

behavior, in which the search process 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.[168]

V.D. Create room (and look) for the unexpected while pursuing your goal. Scientists are trained to minimize variance in their laboratory procedures. There is even a new ACS journal, *Precision Chemistry*,[169] 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."[92] Epsilon-greedy approaches in reinforcement learning provide a theoretical justification[170]—one should mostly take the putative most profitable action, but also allocate some fraction of effort to random new actions in case they are better. This is synergistic with our previous recommendation to avoid premature optimization. Sloppiness can be active (e.g., adding randomness to materials experiment plans[59] or using an additional cost function to experiment generation that maximizes experiment diversity[54]) or passive (e.g., taking advantage of uncontrolled changes in laboratory temperature and humidity as natural experiments[171]). Variations in parameter values ("micro-sloppiness") are more easily achieved, but less likely to lead to large improvements; variations in reagent identity or steps ("macro-sloppiness") typically must be deliberately programmed.

Despite advocating for deliberately "sloppy" reaction designs, we emphasize that this requires the complete data capture of what actually transpired. HTE provides a natural synergy, as it enables complete, machine-readable data collection of metadata and "failed" experiments which might not otherwise be recorded, but which are essential for ML training.[172] Furthermore, allowing for sloppier outcomes might simplify the design tolerances when constructing an HTE system.[14] Once the data are collected, ML methods for *anomaly detection* enable automated serendipity. The role of structural anomaly detection in the discovery of superhard materials was discussed in Section IV.B,[125] and similar opportunities have been discussed for computer-vision-based scanning electron microscopy characterizations[173] and surface-enhanced Raman.[174] The simplest form may be detecting whether an unexpected change has occurred in one or more spectra, a strategy used to discover new organic synthesis reactions.[175,176] Coupling the observation of change in the spectra to neural network models of molecular structure has been used to steer the experimentation toward less predictable reactions.[177]

Data reuse and sharing can also enable finding unexpected trends within and between laboratories by data sharing.[178] There are many examples of scientific serendipity, 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.[92] For example, lead titanate was proposed as a stable photocathode for dye sensitized solar cells based on band structure similarity to known photocathodes.[179] Requirements by funders and publishers around FAIR (findable, accessible, interoperable, reusable)[180,181] and TRUE (Transparent, Reproducible, Usable by others, Extensible)[182] data practices can help create such a resource for retroactive discovery.

V.E. Try to fill-in-the-blanks of input and output space. There is a great opportunity to develop ML methods that enable *untargeted search*.[92] 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 3b. 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.[183] For example, the synthesis of many high-temperature cuprate superconductors was guided by the Goldschmidt tolerance factor[184] which enables the determination of feasible compositions likely to result in the formation of a perovskite. This, coupled with the solid-state literature and nascent Inorganic Crystal Structure Database,[185] 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.[186,187] However, even simple database queries and linear regressions (combined with DFT estimates of stability) suffice to identify potential compounds that fill-in the gaps in the distribution of observed apical and in-plane Cu−O distance distributions for this class of compounds.[188]

Alternatively, one can focus on previously unobserved *outcomes*. The information entropy of the observed property distribution can be useful for identifying outcome 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] To understand how properties are coupled to one another, it might even be useful to fill in equally rare contraindicated regions with undesirable trade-offs ("anti-exceptional materials"). Learning general ways to reach arbitrary outputs can serve as waypoints to the desired solution. An extreme version of this approach explicitly rejects objective-based search, and focuses solely on output novelty.[189] 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.[190] *Random goal exploration* algorithms[191,192] 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 identifying novel protocell lipid formulations.[193] Blending the distinction between input- and outcome-oriented approaches, the *diversity is all you need* reinforcement learning strategy suggests simultaneously optimizing for novel outcomes and synthetic paths (inputs), without imposing other types of fitness objective functions.[194] Regardless of the specific optimization strategy, appropriate data sharing (*vide supra*) is a prerequisite for identifying the underexplored input and output spaces.

V.F. Do not confuse human understanding and model interpretability. Is knowledge more than a merely true belief?[195] Suppose you have an oracle (e.g., an LLM-based chemical property predictor[47]) that tells you where you can get lucky—does it matter how the prediction is made, provided you can verify the outcome? Once the initial discovery is validated experimentally, the traditional scientific method can be

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

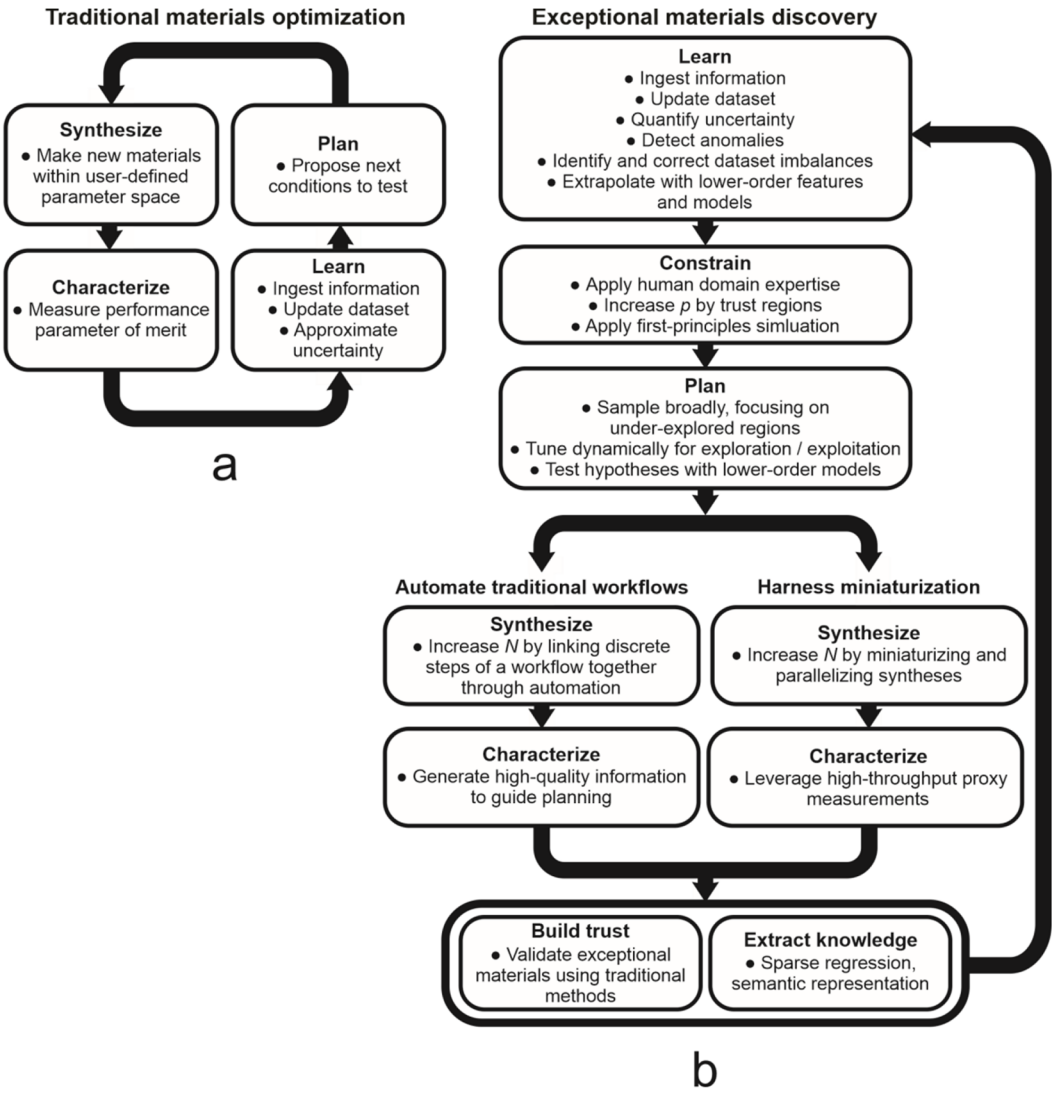

Figure 6. (a) Block diagram of typical autonomous workflow; (b) Block diagram of an autonomous workflow oriented toward exceptional materials.

unleashed to understand the 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 *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.[126]

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 model parameters as 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 196; for reviews of interpretable and explainable methods applied specifically to materials science, see refs 197−201. XAI is motivated by the desire to have the right answer justified by acceptable reasons, but in practice common XAI methods can be misleading; Lei et al. provide a case study of the limitations of SHAP and ensemble feature importance measures in experimental materials science problems.[137] At best, XAI methods generate low-dimensional descriptions of how the model outputs behave based on changes to the inputs. That is, they indicate only 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,[128] obviating the need for more sophisticated black-box model interpretability methods. In fact, empirical studies have found XAI detrimental in uncertain environments, as humans are more likely to reject helpful recommendations because of overconfidence in their trouble-shooting abilities.[202] In many cases, automatic identification of

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

anomalies (*vide supra*) for review by a human operator suffices as 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* the initial discovery of an exceptional material. Most scientists associate science with a knowledge of causes.[203] This can be automated by modern causal influence methods[204] which have recently been applied to catalysis[205] and scanning probe microscopy.[206−208] 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.[209] 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.[210] Progress toward semantic representations of chemistry and materials data and their applications are discussed in refs 209, 211, and 212.

Even without model interpretability or causal explanations, merely having access to "superhuman" AI improves human decision-making. The board game Go provides a case study: Human decision quality remained roughly constant and human decision novelty decreased in the 60 years of tournament data preceding AlphaGo, but access to AlphaGo increased *human* decision quality and novelty by inspiring players to depart from traditional strategies.[213] This suggests the novelty-enhancing recommendations suggested above may suffice to improve human scientific understanding, even without 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 83.) Existing ML approaches accelerate this process by assisting in various optimization subtasks,[14] such as fine-tuning of synthesis and testing operations when dealing with new precursors, reducing the need for skilled labor in operating tools. For example, characterization can be accelerated by automating spectral interpretation[200] and efficiently planning sample characterization campaigns.[214] ML can extract additional information from existing spectroscopy and microscopy methods.[215,216] As discussed in Section III.B, existing ML approaches excel at a variety of research-related optimization tasks.

How might the workflow change to discover exceptional materials? A schematic is depicted in Figure 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 presents 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,[20] and Zubarev et al. recently described software to assist in eliciting human expertise about prioritization, level-of-knowledge, and risk assessments used as input to ML-assisted discovery of new photoacid generator for EUV lithography.[217]

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, as discussed in Section III.C. Broadly, this can be accomplished by 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 equipment as a human chemist,[218] but it might consist of a dedicated "ChemPU" device[219,220] or a collection of modified equipment orchestrated by a central sample management system.[221] This has the advantage of using well-understood synthesis and characterization techniques but limits opportunities for acceleration and scaling. Alternatively, new types of miniaturized and high-throughput synthesis and characterization methods—e.g., microfluidics systems,[222,223] direct writing from liquid precursors,[224,225] combinatorial deposition of sample libraries,[226] and atomic scale dip-pen nanolithography[227,228]—can potentially increase N by orders of magnitude, but introduce doubt whether the resulting products are representative bulk samples. (Given their novelty, the design of the devices themselves is a subject for traditional ML-based optimization.[229]) Faster synthesis makes characterization the rate-limiting step, requiring a shift to faster optical or electrical proxy measurements. For example, computer-imaging-based methods can be used as a fast proxy for indenter-based hardness measurements, within certain bounds of materials composition and accuracy.[230] 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. Research efforts demonstrating equivalence between standard and novel synthesis and characterization techniques 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.[209]

Finally, at the level of planning and coordination between these different types of modalities, the use of LLM-based 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.[231,232] Putting these recommendations into the form of an integrated workflow should better enable the discovery of exceptional materials.

AUTHOR INFORMATION

Corresponding Authors

Joshua Schrier − *Department of Chemistry, Fordham University, The Bronx, New York 10458, United States;* orcid.org/0000-0002-2071-1657; Email: jschrier@fordham.edu

Alexander J. Norquist − *Department of Chemistry, Haverford College, Haverford, Pennsylvania 19041, United States;* orcid.org/0000-0003-1040-5880; Email: anorquis@haverford.edu

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716

Tonio Buonassisi − *Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-8345-4937; Email: buonassi@mit.edu

Jakoah Brgoch − *Department of Chemistry and Texas Center for Superconductivity, University of Houston, Houston, Texas 77204, United States;* orcid.org/0000-0002-1406-1352; Email: jbrgoch@uh.edu

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c04783

Notes

The authors declare the following competing financial interest(s): J.S. and A.J.N. own equity and serve as scientific advisors for Atinary Technologies, Inc., which develops data management and machine learning software for experiment planning. T.B. owns equity and serves as scientific advisor for Xinterra Pte. Ltd., which applies machine learning and high-throughput experiment methods to accelerate novel materials development.

■ ACKNOWLEDGMENTS

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). Craiyon.com was used to generate the graphical table of contents image using the prompt: "scientist looking at far away mountains with Erlenmeyer flasks floating in the sky".

■ 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. DOI: 10.48550/arXiv.2212.13893 (accessed 2023-09-15).

(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*, 84.

(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*, No. 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.

(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.

(6) Bannigan, P.; Aldeghi, M.; Bao, Z.; Häse, F.; Aspuru-Guzik, A.; Allen, C. Machine Learning Directed Drug Formulation Development. *Adv. Drug Delivery Rev.* **2021**, *175*, No. 113806.

(7) Xie, S. Perspectives on Development of Biomedical Polymer Materials in Artificial Intelligence Age. *J. Biomater Appl.* **2023**, *37* (8), 1355−1375.

(8) Janet, J. P.; Kulik, H. J. *Machine Learning in Chemistry*; ACS In Focus; American Chemical Society: Washington, DC, USA, 2020. DOI: 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. DOI: 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. DOI: 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.

(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.

(13) Morgan, D.; Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Annu. Rev. Mater. Res.* **2020**, *50* (1), 71−103.

(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.

(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 Bull.* **2022**, *47*, 1154−1164.

(16) Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; Anandkumar, A.; Bergen, K.; Gomes, C. P.; Ho, S.; Kohli, P.; Lasenby, J.; Leskovec, J.; Liu, T.-Y.; Manrai, A.; Marks, D.; Ramsundar, B.; Song, L.; Sun, J.; Tang, J.; Veličković, P.; Welling, M.; Zhang, L.; Coley, C. W.; Bengio, Y.; Zitnik, M. Scientific Discovery in the Age of Artificial Intelligence. *Nature* **2023**, *620* (7972), 47−60.

(17) 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.

(18) 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.

(19) 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), No. 100295.

(20) 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.

(21) Kuhn, T. S. *The Structure of Scientific Revolutions*, 4th ed.; The University of Chicago Press: Chicago; London, 2012.

(22) Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*, 1st ed.; Random House: New York, 2007.

(23) Nugent, W. A. Black Swan Events" in Organic Synthesis. *Angew. Chem., Int. Ed.* **2012**, *51* (36), 8936−8949.

(24) Kohnert, A. A.; Wirth, B. D.; Wolverton, C.; Balachandran, P. V.; Capolungo, L. Modeling Materials under Coupled Extremes: Enabling Better Predictions of Performance. *MRS Bull.* **2022**, *47*, 1120.

(25) 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. DOI: 10.2172/1962808.

(26) Zunger, A. Inverse Design in Search of Materials with Target Functionalities. *Nat. Rev. Chem.* **2018**, *2* (4), 0121.
(27) Hase, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chemical Science* **2018**, *9* (39), 7642−7655.
(28) 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. *Sci. Adv.* **2021**, *7* (42), No. eabf7435.
(29) Low, A.; Lim, Y.-F.; Hippalgaonkar, K.; Vissol-Gaudin, E. Mapping Pareto Fronts for Efficient Multi-Objective Materials Discovery. *AISY Interactive Papers* **2022**, DOI: 10.22541/au.167083158.87890589/v1.
(30) 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**, *8*, 241.
(31) Li, K.; DeCost, B.; Choudhary, K.; Greenwood, M.; Hattrick-Simpers, J. A Critical Examination of Robustness and Generalizability of Machine Learning Prediction of Materials Properties. *npj Comput. Mater.* **2023**, *9* (1), 55.
(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.
(33) Restrepo, G. Chemical Space: Limits, Evolution and Modelling of an Object Bigger than Our Universal Library. *Digital Discovery* **2022**, *1* (5), 568−585.
(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.
(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. *Sci. Adv.* **2016**, *2* (11), No. e1600225.
(36) Rondinelli, J. M.; Chen, W. ET-AL: Entropy-Targeted Active Learning for Bias Mitigation in Materials Data. *Applied Physics Reviews* **2023**, *10* (2), No. 021403.
(37) Rahm, M.; Cammi, R.; Ashcroft, N. W.; Hoffmann, R. Squeezing All Elements in the Periodic Table: Electron Configuration and Electronegativity of the Atoms under Compression. *J. Am. Chem. Soc.* **2019**, *141* (26), 10253−10271.
(38) Boyd, S. P.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK; New York, 2004.
(39) Stadler, P. F. Landscapes and Their Correlation Functions. *J. Math. Chem.* **1996**, *20* (1), 1−45.
(40) Barnes, J. W.; Dimova, B.; Dokov, S. P.; Solomon, A. The Theory of Elementary Landscapes. *Appl. Math. Lett.* **2003**, *16* (3), 337−343.
(41) Siemenn, A. E.; Ren, Z.; Li, Q.; Buonassisi, T. Fast Bayesian Optimization of Needle-in-a-Haystack Problems Using Zooming Memory-Based Initialization. *npj Comput. Mater.* **2023**, *9*, 79.
(42) 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.
(43) Materials Genome Initiative. https://www.mgi.gov/ (accessed 2023-02-02).
(44) Xie, J.; Su, Y.; Zhang, D.; Feng, Q. A Vision of Materials Genome Engineering in China. *Engineering* **2022**, *10*, 10−12.
(45) 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.
(46) Tu, X.; Zou, J.; Su, W. J.; Zhang, L. What Should Data Science Education Do with Large Language Models? arXiv July 7, 2023. DOI: 10.48550/arXiv.2307.02792. (accessed 2023−09−15).
(47) Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. Is GPT-3 All You Need for Low-Data Discovery in Chemistry? ChemRxiv May 16, 2023. DOI: 10.26434/chemrxiv-2023-fw8n4-v2 (accessed 2023-09-15).
(48) 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.
(49) 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.
(50) 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), No. 2270020.
(51) 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), No. eaaz1708.
(52) 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.
(53) 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), No. 064108.
(54) 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. DOI: 10.26434/chemrxiv-2022-l1wpf (accessed 2023-09-15).
(55) 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.
(56) 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. DOI: 10.48550/arXiv.2212.11855 (accessed 2023-09-15).
(57) 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.
(58) 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.
(59) 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.
(60) 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.
(61) Kauwe, S. K.; Graser, J.; Murdock, R.; Sparks, T. D. Can Machine Learning Find Extraordinary Materials? *Comput. Mater. Sci.* **2020**, *174*, No. 109498.
(62) 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**, *2*, 327−338.
(63) 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.

(64) 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.
(65) 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.
(66) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial Intelligence and Automation in Computer Aided Synthesis Planning. *React. Chem. Eng.* **2021**, *6* (1), 27−51.
(67) 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.
(68) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of Organic Chemistry Grammar from Unsupervised Learning of Chemical Reactions. *Sci. Adv.* **2021**, *7* (15), No. eabe4166.
(69) Borrelli, W.; Schrier, J. Evaluating the Performance of a Transformer-Based Organic Reaction Prediction Model; ChemRxiv, 29 July, 2021. DOI: 10.26434/chemrxiv-2021-3nqv9 (accessed 2023-09-15).
(70) 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.
(71) Bender, E. M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. In *On the Dangers of Stochastic Parrots: Can. Language Models Be Too Big?* 🦜, Proceedingsof the 2021 ACM Conference on Fairness, Accountability, and Transparency; ACM: Virtual Event Canada, 2021; pp 610−623. DOI: 10.1145/3442188.3445922.
(72) 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; 2023.
(73) Mirchandani, S.; Xia, F.; Florence, P.; Ichter, B.; Driess, D.; Arenas, M. G.; Rao, K.; Sadigh, D.; Zeng, A. Large Language Models as General Pattern Machines. arXiv July 10, 2023. DOI: 10.48550/arXiv.2307.04721 (accessed 2023-09-15).
(74) 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.
(75) 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**, *123* (6), 3089−3126.
(76) Xiang, X.-D.; Takeuchi, I. *Combinatorial Materials Synthesis*; CRC Press: 2003.
(77) Koinuma, H.; Takeuchi, I. Combinatorial Solid-State Chemistry of Inorganic Materials. *Nat. Mater.* **2004**, *3* (7), 429−438.
(78) Takeuchi, I.; Lippmaa, M.; Matsumoto, Y. Combinatorial Experimentation and Materials Informatics. *MRS Bull.* **2006**, *31* (12), 999−1003.
(79) 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. *Appl. Phys. Rev.* **2017**, *4* (1), No. 011105.
(80) 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**, *55* (17), 2454−2466.
(81) Delgado-Licona, F.; Abolhasani, M. Research Acceleration in Self-Driving Labs: Technological Roadmap toward Accelerated Materials and Molecular Discovery. *Advanced Intelligent Systems 5* (4), 2200331. DOI: 10.1002/aisy.202200331.
(82) Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. *Nat. Synth* **2023**, *2*, 483−492.
(83) 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.
(84) Dasgupta, S. Analysis of a Greedy Active Learning Strategy. In *Advances in Neural Information Processing Systems*; MIT Press, 2004; Vol. *17*.
(85) Rohr, B.; Stein, H. S.; Guevarra, D.; Wang, Y.; Haber, J. A.; Aykol, M.; Suram, S. K.; Gregoire, J. M. Benchmarking the Acceleration of Materials Discovery by Sequential Learning. *Chemical Science* **2020**, *11* (10), 2696−2706.
(86) Liu, R.; Li, X.; Lam, K. S. Combinatorial Chemistry in Drug Discovery. *Curr. Opin Chem. Biol.* **2017**, *38*, 117−126.
(87) 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.
(88) 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.
(89) Fromer, J. C.; Coley, C. W. Computer-Aided Multi-Objective Optimization in Small Molecule Discovery. *Patterns* **2023**, *4* (2), No. 100678.
(90) 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.
(91) 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.
(92) Yaqub, O. Serendipity: Towards a Taxonomy and a Theory. *Research Policy* **2018**, *47* (1), 169−179.
(93) van Delft, D.; Kes, P. The Discovery of Superconductivity. *Phys. Today* **2010**, *63* (9), 38−43.
(94) Cava, R. J. Oxide Superconductors. *J. Am. Ceram. Soc.* **2000**, *83* (1), 5−28.
(95) Bardeen, J.; Cooper, L. N.; Schrieffer, J. R. Microscopic Theory of Superconductivity. *Phys. Rev.* **1957**, *106* (1), 162−164.
(96) Brandt, N. B.; Moshchalkov, V. V. Concentrated Kondo Systems. *Adv. Phys.* **1984**, *33* (5), 373−467.
(97) Chevrel, R.; Hirrien, M.; Sergent, M. Superconducting Chevrel Phases: Prospects and Perspectives. *Polyhedron* **1986**, *5* (1−2), 87−94.
(98) 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. *Mater. Res. Bull.* **1980**, *15* (7), 891−897.
(99) 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.
(100) Poeppelmeier, K. R.; Leonowicz, M. E.; Longo, J. M. $CaMnO_{2.5}$ and $Ca_2MnO_{3.5}$: New Oxygen-Defect Perovskite-Type Oxides. *J. Solid State Chem.* **1982**, *44* (1), 89−98.
(101) 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.
(102) 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.
(103) 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. London A* **1984**, *394* (1807), 223−241.
(104) Mott, N. *Metal-Insulator Transitions*; CRC Press, 2004. DOI: 10.1201/b12795.
(105) 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.
(106) Johnston, D. C.; Prakash, H.; Zachariasen, W. H.; Viswanathan, R. High Temperature Superconductivity in the Li Ti O Ternary System. *Mater. Res. Bull.* **1973**, *8* (7), 777−784.
(107) Ostenson, J. E.; Shanks, H. R.; Finnemore, D. K. Superconductivity in the Tungsten Bronzes. *J. Less Common Metals* **1978**, *62*, 149−153.
(108) 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.
(109) 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.
(110) 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.
(111) 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.
(112) 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. DOI: 10.1007/978-94-011-1622-0_36.
(113) 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.
(114) Villars, P.; Phillips, J. C. Quantum Structural Diagrams and High-T_c Superconductivity. *Phys. Rev. B* **1988**, *37* (4), 2345−2348.
(115) 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.
(116) 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.
(117) 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.
(118) 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.
(119) 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), No. 2005112.
(120) 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.
(121) Hickey, J. C.; Brgoch, J. The Limits of Proxy-Guided Superhard Materials Screening. *Chem. Mater.* **2022**, *34* (22), 10003−10010.
(122) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary Crystal Structure Prediction. *Comput. Phys. Commun.* **2006**, *175* (11−12), 713−720.
(123) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A Method for Crystal Structure Prediction. *Comput. Phys. Commun.* **2012**, *183* (10), 2063−2070.
(124) Lonie, D. C.; Zurek, E. XtalOpt: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction. *Comput. Phys. Commun.* **2011**, *182* (2), 372−387.
(125) Zhang, Z.; Brgoch, J. Treating Superhard Materials as Anomalies. *J. Am. Chem. Soc.* **2022**, *144* (39), 18075−18080.
(126) Feyerabend, P. *Against Method: Outline of an Anarchistic Theory of Knowledge*; NLB; Humanities Press: London : Atlantic Highlands, 1975.
(127) 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.
(128) Muckley, E. S.; Saal, J. E.; Meredig, B.; Roper, C. S.; Martin, J. H. Interpretable Models for Extrapolation in Scientific Machine Learning. *Digital Discovery* 2023 (in press). DOI: 10.1039/D3DD00082F.
(129) Rockafellar, R. T.; Uryasev, S. Optimization of Conditional Value-at-Risk. *J. Risk* **2000**, *2* (3), 21−41.
(130) 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).
(131) Kikkawa, N.; Ohno, H. Materials Discovery Using Max K-Armed Bandit. arXiv December 15, 2022. DOI: 10.48550/arXiv.2212.08225 (accessed 2023-02-02).
(132) Cicirello, V. A.; Smith, S. F. The Max K-Armed Bandit: A New Model of Exploration Applied to Search Heuristic Selection. *Proceedings of the Twentieth National Conference on Artificial Intelligence* **2005**, *3*, 1355−1361.
(133) Mohamad, M. A.; Sapsis, T. P. Sequential Sampling Strategy for Extreme Event Statistics in Nonlinear Dynamical Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115* (44), 11138−11143.
(134) Blanchard, A.; Sapsis, T. Bayesian Optimization with Output-Weighted Optimal Sampling. *J. Comput. Phys.* **2021**, *425*, No. 109901.
(135) 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.
(136) 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.
(137) 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), No. 054105.
(138) 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. *Phys. Rev. Lett.* **1987**, *58* (9), 908−910.
(139) Kuhn, T. S. *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*; Harvard University Press: Cambridge, MA, 1957.
(140) Kepler, J. Astronomia Nova (1609). In *New Astronomy*; Donahue, W. H., Translator; Cambridge University Press: Cambridge, 1992; pp 47−48.
(141) Miyake, T. Underdetermination and Decomposition in Kepler's Astronomia Nova. *Studies in History and Philosophy of Science Part A* **2015**, *50*, 20−27.
(142) Drake, S. Galileo, Kepler, and Phases of Venus. *J. Hist. Astron.* **1984**, *15* (3), 198−208.
(143) Bradley, J. An Account of a New Discovered Motion of the Fix'd Stars. *Philos. Trans. Royal Soc.* **1728**, *35* (406), 637−661.
(144) Wagner, N.; Rondinelli, J. M. Theory-Guided Machine Learning in Materials Science. *Front. Mater.* **2016**, *3*. DOI: 10.3389/fmats.2016.00028.
(145) 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.

(146) 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. DOI: 10.48550/arXiv.2201.05624 (accessed 2023-02-02).
(147) 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.
(148) 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.
(149) 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. *Cryst. Growth Des.* **2022**, *22* (9), 5424−5431.
(150) 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. *Chem. Eng. J.* **2020**, *387*, No. 123412.
(151) Wang, Y.; Wagner, N.; Rondinelli, J. M. Symbolic Regression in Materials Science. *MRS Commun.* **2019**, *9* (3), 793−805.
(152) 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.
(153) 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.
(154) Kamienny, P.-A.; d'Ascoli, S.; Lample, G.; Charton, F. End-to-End Symbolic Regression with Transformers. arXiv April 22, 2022. DOI: 10.48550/arXiv.2204.10532 (accessed 2023-01-11).
(155) Anstine, D. M.; Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.* **2023**, *145* (16), 8736−8750.
(156) Brown, R. Consideration of the Origin of Herbert Simon's Theory of "Satisficing" (1933−1947). *Management Decision* **2004**, *42* (10), 1240−1256.
(157) Artinger, F. M.; Gigerenzer, G.; Jacobs, P. Satisficing: Integrating Two Traditions. *J. Econ. Lit.* **2022**, *60* (2), 598−635.
(158) Miconi, T. Evolution and Complexity: The Double-Edged Sword. *Artif Life* **2008**, *14* (3), 325−344.
(159) He, T.; Huo, H.; Bartel, C. J.; Wang, Z.; Cruse, K.; Ceder, G. Inorganic Synthesis Recommendation by Machine Learning Materials Similarity from Scientific Literature. *Sci. Adv.* **2023**, *9* (23), No. eadg8180.
(160) Stanley, M.; Segler, M. Fake It until You Make It? Generative de Novo Design and Virtual Screening of Synthesizable Molecules. *Curr. Opin. Struct. Biol.* **2023**, *82*, No. 102658.
(161) 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), 38.
(162) 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*, 66.
(163) Margraf, J. T.; Ulissi, Z. W.; Jung, Y.; Reuter, K. Heterogeneous Catalysis in Grammar School. *J. Phys. Chem. C* **2022**, *126* (6), 2931−2936.
(164) 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.
(165) Lemm, D.; von Rudorff, G. F.; von Lilienfeld, O. A. Improved Decision Making with Similarity Based Machine Learning. arXiv July 9, 2023. DOI: 10.48550/arXiv.2205.05633. (accessed 2023-09-15).
(166) 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.
(167) Mouret, J.-B.; Clune, J. Illuminating Search Spaces by Mapping Elites. arXiv April 19, 2015. DOI: 10.48550/arXiv.1504.04909. (accessed 2023-09-15).
(168) 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. *Phys. A: Stat. Mech. Appl.* **2000**, *282* (1), 1−12.
(169) Precision Chemistry. https://pubs.acs.org/journal/pcrhej (accessed 2023-02-02).
(170) Sutton, R. S.; Barto, A. G. *Reinforcement Learning: An Introduction*, Second ed..; MIT Press: Cambridge, MA, 2018.
(171) 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), No. 041903.
(172) 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.
(173) Napoletano, P.; Piccoli, F.; Schettini, R. Anomaly Detection in Nanofibrous Materials by CNN-Based Self-Similarity. *Sensors* **2018**, *18* (1), 209.
(174) 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. DOI: 10.1117/12.2576447.
(175) 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.
(176) 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.
(177) 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.
(178) Pelkie, B. G.; Pozzo, L. D. The Laboratory of Babel: Highlighting Community Needs for Integrated Materials Data Management. *Digital Discovery* **2023**, *2*, 544−556.
(179) 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.
(180) 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), No. 160018.
(181) 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.
(182) 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 Trans-

parent, Reproducible, Usable by Others, and Extensible (TRUE). *Mol. Phys.* **2020**, *118* (9−10), No. e1742938.
(183) 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.
(184) Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **1926**, *14* (21), 477−485.
(185) Levin, I. NIST Inorganic Crystal Structure Database (ICSD), 2020. DOI: 10.18434/M32147.
(186) Seegmiller, C. C.; Baird, S. G.; Sayeed, H. M.; Sparks, T. D. Discovering Chemically Novel, High-Temperature Superconductors. ChemRxiv March 13, 2023. DOI: 10.26434/chemrxiv-2023-8t8kt (accessed 2023-09-15).
(187) Wines, D.; Xie, T.; Choudhary, K. Inverse Design of Next-Generation Superconductors Using Data-Driven Deep Generative Models. *J. Phys. Chem. Lett.* **2023**, *14* (29), 6630−6638.
(188) Goodall, R. E. A.; Zhu, B.; MacManus-Driscoll, J. L.; Lee, A. A. Materials Informatics Reveals Unexplored Structure Space in Cuprate Superconductors. *Adv. Funct. Mater.* **2021**, *31* (52), No. 2104696.
(189) Lehman, J.; Stanley, K. O. Abandoning Objectives: Evolution Through the Search for Novelty Alone. *Evolutionary Computation* **2011**, *19* (2), 189−223.
(190) Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; Efros, A. A. Large-Scale Study of Curiosity-Driven Learning. arXiv August 13, 2018. DOI: 10.48550/arXiv.1808.04355 (accessed 2023-09-15).
(191) 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.
(192) Péré, A.; Forestier, S.; Sigaud, O.; Oudeyer, P.-Y. Unsupervised Learning of Goal Spaces for Intrinsically Motivated Goal Exploration. arXiv October 9, 2018. DOI: 10.48550/arXiv.1803.00781 (accessed 2023-09-15).
(193) 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), No. eaay4237.
(194) Eysenbach, B.; Gupta, A.; Ibarz, J.; Levine, S. Diversity Is All You Need: Learning Skills without a Reward Function. arXiv October 9, 2018. DOI: 10.48550/arXiv.1802.06070 (accessed 2023-09-15).
(195) 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; https://plato.stanford.edu/Archives/win2022/entries/knowledge-value/ (accessed 2023-09-15).
(196) Jacovi, A. Trends in Explainable AI (XAI) Literature. arXiv January 13, 2023. DOI: 10.48550/arXiv.2301.05433 (accessed 2023-09-15).
(197) Dybowski, R. Interpretable Machine Learning as a Tool for Scientific Discovery in Chemistry. *New J. Chem.* **2020**, *44* (48), 20914−20920.
(198) Pilania, G. Machine Learning in Materials Science: From Explainable Predictions to Autonomous Design. *Comput. Mater. Sci.* **2021**, *193*, No. 110360.
(199) Ghiringhelli, L. M. Interpretability of Machine-Learning Models in Physical Sciences. arXiv April 21, 2021. DOI: 10.48550/arXiv.2104.10443 (accessed 2023-01-20).
(200) 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.
(201) 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.
(202) DeStefano, T.; Kellogg, K.; Menietti, M.; Vendraminelli, L. Why Providing Humans with Interpretable Algorithms May, Counter-intuitively, Lead to Lower Decision-Making Performance. *SSRN Journal* **2022**, DOI: 10.2139/ssrn.4246077.
(203) Falcon, A. Aristotle on Causality. In *The Stanford Encyclopedia of Philosophy*; Zalta, E. N., Ed.; Metaphysics Research Lab, Stanford University: https://plato.stanford.edu/archives/spr2022/entries/aristotle-causality/, 2022.
(204) Pearl, J.; Glymour, M.; Jewell, N. P. *Causal Inference in Statistics: A Primer*; Wiley: Chichester, West Sussex, 2016.
(205) Ting, J. Y.; Barnard, A. S. Data-Driven Causal Inference of Process-Structure Relationships in Nanocatalysis. *Curr. Opin. Chem. Eng.* **2022**, *36*, No. 100818.
(206) 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.
(207) 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.
(208) 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.
(209) 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.
(210) Russell, S. J.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed..; Pearson: Hoboken, NJ, 2021.
(211) 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.
(212) Chalk, S. J. *AI4SD Video: SciData: Semantic Representation of Scientific Data and Applications in Chemistry* **2022**, DOI: 10.5258/SOTON/AI3SD0237.
(213) Shin, M.; Kim, J.; van Opheusden, B.; Griffiths, T. L. Superhuman Artificial Intelligence Can Improve Human Decision-Making by Increasing Novelty. *Proc. Natl. Acad. Sci. U. S. A.* **2023**, *120* (12), No. e2214840120.
(214) 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.
(215) 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.
(216) Ziatdinov, M.; Ghosh, A.; Wong, C. Y.; 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.
(217) 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. DOI: 10.48550/arXiv.2211.04257 (accessed 2023-09-15).
(218) 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.
(219) 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.
(220) 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.
(221) 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. DOI: 10.26434/chemrxiv-2023-r7b01 (accessed 2023-09-15).

(222) 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.
(223) 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.
(224) 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.
(225) 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.
(226) Gregoire, J. M.; Zhou, L.; Haber, J. A. Combinatorial Synthesis for AI-Driven Materials Discovery. *Nat. Synth* **2023**, *2*, 493.
(227) 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.
(228) Liu, G.; Rong, M.; Hu, H.; Chen, L.; Xie, Z.; Zheng, Z. 3D Dip-Pen Nanolithography. *Adv. Mater. Technol.* **2022**, *7* (8), No. 2101493.
(229) McIntyre, D.; Lashkaripour, A.; Fordyce, P.; Densmore, D. Machine Learning for Microfluidic Design and Control. *Lab Chip* **2022**, *22* (16), 2925−2937.
(230) Lew, A. J.; Stifler, C. A.; Cantamessa, A.; Tits, A.; Ruffoni, D.; Gilbert, P. U. P. A.; Buehler, M. J. Deep Learning Virtual Indenter Maps Nanoscale Hardness Rapidly and Non-Destructively, Revealing Mechanism and Enhancing Bioinspired Design. *Matter* **2023**, *6* (6), 1975−1991.
(231) Kusne, A. G.; McDannald, A. Scalable Multi-Agent Lab Framework for Lab Optimization. *Matter* **2023**, *6* (6), 1880−1893.
(232) Boiko, D. A.; MacKnight, R.; Gomes, G. Emergent Autonomous Scientific Research Capabilities of Large Language Models. arXiv April 11, 2023. DOI: 10.48550/arXiv.2304.05332 (accessed 2023-09-15).

https://doi.org/10.1021/jacs.3c04783
J. Am. Chem. Soc. 2023, 145, 21699−21716