Article category: **Expert Recommendation**

The Case for Data Science in Experimental Chemistry: Examples and Recommendations

Junko Yano[a], Kelly J. Gaffney[b,c], John Gregoire[d], Linda Hung[e], Abbas Ourmazd[f], Joshua Schrier[g], James A. Sethian[h,i], Francesca M. Toma[j,k]

[a] Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[b] SLAC National Accelerator Laboratory, Menlo Park, CA, USA.
[c] PULSE Institute, SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA
[d] Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA, USA.
[e] Accelerated Materials Design and Discovery, Toyota Research Institute, Los Altos, CA, USA.
[f] University of Wisconsin, Milwaukee, WI, USA.
[g] Fordham University, Department of Chemistry, The Bronx, NY, USA.
[h] Department of Mathematics, University of California, Berkeley, CA, USA.
[i] Center for Advanced Mathematics for Energy Research Applications (CAMERA), , Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[j] Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Corresponding Authors: JYano@lbl.gov, kgaffney@slac.stanford.edu, gregoire@caltech.edu, linda.hung@tri.global, ourmazd@uwm.edu, jschrier@fordham.edu, sethian@berkeley.edu, fmtoma@lbl.gov

Abstract

The physical sciences community is increasingly taking advantage of the possibilities offered by modern data science to help meet challenges in experimental chemistry and potentially change the way we design, conduct, and understand results from experiments. Successfully exploiting these opportunities involves significant challenges. In this Expert Recommendation, we provide examples of how data science is changing the way we conduct experiments, and outline opportunities for further integrating data science and experimental chemistry to advance both of these fields. Our roadmap includes establishing stronger links between chemists and data scientists, developing chemistry-specific data science methods, integrating algorithms, software, and hardware to "co-design" chemistry experiments from inception, combining diverse and disparate data sources into a data network for chemistry research.

1. Introduction

Data-driven techniques, such as machine-learning (ML) and artificial-intelligence (AI), are rapidly becoming indispensable tools for scientific research,[1] and have been the topic of national[2] and international[3] reports, recent review and perspective articles,[4-6] and tutorial guides.[7,8] With some exceptions,[9] most work has focused on ML approaches trained on synthetic datasets, and used to accelerate computer simulations. However, emerging data-driven approaches for synthesis, spectroscopic interpretation, and optimal experimental design now highlight the potential to advance experimental chemistry with data-driven methods.[10,11] [add cites to: Cernak & Doyle & co.
10.1038/ s43586-021-00022-5 & Nichols10.1016/bs.pmch.2021.01.003 (also used later)] For example, combining such data analytical methods with automation or laboratory robotics could enable quasi autonomous research with minimal human input.[12,13] Improved data analytics and data sharing/reuse in experimental chemistry offer the opportunity to increase the rate and lower the cost of scientific discovery, and grow research productivity.

Parallel advances in data science and in experimental chemistry have rapidly expanded the opportunity to integrate these fields . Given the diversity of experimental methods, data acquisition techniques, and their assembly into experimental workflows,[14] the realm of possible workflows and methods for designing experiments far exceeds those realized by human researchers to-date. Data science methods are poised to aid workflow design and active steering of experiments to broaden the reach of experimental chemistry and enhance the rate and efficacy with which chemists explore the often daunting parameter spaces of experiment and synthesis. Capitalizing on these opportunities will require fundamental advances in both chemistry and data science, as well as changing how we conduct experiments, especially the development of technologies to facilitate large-scale data collection, sharing, and analysis. At the same time, validating outcomes of data science-based interpretation and prediction will be essential.

Here, we include key highlights from "At the Tipping Point: A Future of Fused Chemical and Data Science", a September 2020 workshop supported by the Council on Chemical Sciences, Geosciences, and Biosciences and sponsored by the Chemical Sciences, Geosciences, and Biosciences (CSGB) Division, Office of Basic Energy Sciences Office of Science U.S. Department of Energy. Participants from academia, industry, and national laboratories assessed opportunities and key research needs for the use of data science for new experimental approaches in chemistry and biochemistry, at experimental scales ranging from single-PI laboratories to large user facilities. We focus on experimental chemistry, and discuss how data science is changing the way we conduct experiments using case studies, and summarize what is required to take advantage of the significant advances in both fields.

2. A Broad Perspective of Data science

Science has always been driven by the interplay of data and theory. Data, which can come from observations, simulations, or experiment, aid in hypothesis and theory development. Theories codify understanding, offer predictions which can often help extrapolate into experimentally unexplored domains, and provide conceptual frameworks for suggesting new experiments and regions of possible interest. This interplay is central to scientific understanding.

The challenges and opportunities offered by this interplay have been accelerated by technological advances in detectors, computation, and algorithms, which have significantly changed data acquisition rates and the range of tools available to classify, analyze, and interpret data. In some experiments, acquiring many types of experimental data is no longer "expensive," and vast amounts can be easily accumulated. In other areas, the equipment and the experiments themselves are so expensive or over-subscribed that one must carefully choose which experiments to perform. The growing field of "data science" offers possibilities to combine opportunities enabled by advances in algorithms, hardware, and vast data sources. Further advances in chemical sciences will require exploiting and developing these efforts, augmenting the traditional approach of theory to selectively guide investigations with new approaches that can handle both large amounts of data, and the vast landscape of possibilities.

One important component of data-driven science is the perspective that data itself can shed light on processes and mechanisms, without requiring accompanying theories and models. Analyzing data without a theory-based roadmap is critical to making sense of the ever-increasing influx of data. This sounds more radical than it really is: relying on observations to frame (and sometimes justify) expectations has often outpaced theory and models. Data science embraces the importance of classification and identification of robust correlations in large, complex datasets that historically have been a pillar of theoretical advances, but now require new methods to deal with vast new quantities of data and accelerating acquisition rates.

The need for advanced techniques able to interpret and categorize data is an increasingly critical part of the scientific process. Advances in mathematical algorithms, broadly defined to include core mathematical ideas such as approximation theory, linear algebra, and differential equations, as well as statistics, signal and image processing, machine learning (ML) and artificial intelligence (AI), have been instrumental in extracting knowledge from data, and accelerating scientific progress in the data-experiment-theory interplay. As experiments become more complex, and instruments and detectors faster and more resolved, these needs will become increasingly prevalent.

Whether data science interpretations become an incremental step towards traditional model-based scientific understanding, or ultimately stand on an equal footing, and in some arenas, surpass model-based understanding remains unclear. Even in the absence of a data science *revolution*, data science will lead to *evolution* in how we generate and interpret scientific data. The challenge is to have some reliable way of saying whether one has enough experiments, or enough data, or enough observations, to justify making predictions with quantified uncertainty. While there is no single route to estimating the uncertainty (error) in the outcome of AI/ML approaches, good-practice methods range from the simple (and transparent) to the sophisticated (and generally

less transparent). Some of the best approaches rely on independently known "ground-truths" to estimate the error in the outcome of data-driven analysis in comparison with the available ground truths. Such estimates assess, in essence, the interpolation error. The assessment of predictions outside the training range entails additional complexities. At the end of the day, the current state of the art is such that one extrapolates beyond the training domain at one's own risk.

In the most radical interpretation , AI and ML techniques suggest that one need not have any preconceived notion of what experiments to perform, what variables to observe, and what weights to put on gathered information. Of course, ML/AI algorithms rely on hidden assumptions and biases, including, for example, definitions of closeness, similarities, and structures. Nonetheless, the idea and promise of these approaches are that the algorithms themselves will detect the important relationships, even if these relationships are not revealed in the standard form of analytical models, communicable principles, and foundational theories.

While there are many challenges associated with ML (**Box 1**) and no clear path on how to simultaneously address them, the opportunities are hard to ignore: an increasing amount of data is available, and better ways to use it will provide new insights. Three modalities by which data science could transform experimental chemistry are listed in **Box 2**. The hope and expectation are that data science methods can learn important relationships at previously unachieved speed and scale, and that those relationships can then be exploited to accelerate scientific progress.

In the following sections, we provide some chemical sciences case studies of advances and potential of interaction of experiments and data sciences, and discuss the challenges for moving the path forward.

3. Data science and chemical sciences

ML proponents promise profound advances within chemical sciences in arenas such as extracting collective coordinates, reaction paths, energy landscapes, and dynamics from lots of heterogeneous observations. Broadly speaking, it is expected to bring at least three important objectives within reach (**Box 2**). In the chemical sciences, there have been remarkable steps toward meeting these objectives, and the promise and potential is significant.[9,15-18] At the same time, there are limitations and pitfalls—We try to give examples in multiple different fields. Of course, these objectives stem from a continuum, rather than a discrete spectrum of possibilities, but it is helpful to independently address each objective.

3.1 ML-guided discovery

Experiments are traditionally either steered by intuition or by schemes in which a measurement plan is selected and implemented in advance, independent of the measurement results. Neither is efficient: the intuitive approach demands constant attention by a highly trained expert, and the exhaustive approach wastes instrument time by collecting a large amount of possibly redundant data.

As experiments become more complex, these approaches become even more problematic. Rather than simply being a question of efficiency, the central issue is that the combinatorics of high-dimensional parameter spaces yield a set of possible configurations that is too large to systematically explore with pre-arranged strategies.

Goal - Autonomous, Self-Guiding Laboratories. Instead, imagine a process by which a set of previously performed experiments is used to suggest what to try next. These suggestions may, for example, come from surrogate models, which represent lower-dimensional approximations to the landscape of collected data from sparsely sampled high-dimensional parameter space. Taking as input the available experimental data, both from the current experiment and available literature, as well as previously established scientific information, these models can then suggest experiments able to accomplish different or multiple goals, including:

- Aim new experiments at under-explored parts of the high-dimensional parameter space. These new experiments would configure the experimental parameters to examine under-sampled possibilities. The goal is to make sure that a full range of scientific results across the parameter space is collected efficiently.
- As experiments are performed and analyzed, focus new experiments on profitable configurations that are yielding, as experiments are performed and analyzed, insight into particularly desirable results.

One important goal is to couple this autonomous steering to advanced simulations and feedback metrics to allow experiments to discover regions in high dimensional configuration space that zero in on optimal parameters, such as those required to achieve desired results. A recent two-part review of autonomous discovery in the chemical sciences can be found in the literature,[19,20] as well as targeted reviews on autonomous materials science [cite:doi:10.1016/j.matt.2021.06.036], organic synthesis planning and optimization [cite: Cernak & Doyle & co.10.1038/ s43586-021-00022-5], medicinal chemistry [cite: Nichols 10.1016/bs.pmch.2021.01.003] and formulations[cite: 10.1002/aic.17248]. While this is often caricatured as "getting humans out of the process", hybrid approaches offer a valuable path forward. For example, combined human-algorithm teams can more efficiently identify crystallization and self-assembly conditions for inorganic synthesis compared to human-only or algorithm-only approaches.[20]

What is needed. Taking full advantage of these possibilities requires multiple advances, including: (a) configuring the data as it is collected so that it can be easily interpreted; (b) fast techniques for building representative surrogate models on-the-fly as data are collected; (c) examining these models to determine and suggest new experimental measurements; and (d) laboratory automation software and hardware that allow suggestions to become physical experiments (**FIG. 1**).

A pivotal role for ML and AI. Advances in ML and AI offer opportunities to achieve these goals. First, given the output of an experiment, these techniques can assess the collected data in the

context of other experiments and simulation results. As an example, suppose an experiment under a given set of input parameters yields a particular scattering pattern, spectrum measurement, or chemical signature. A robust and accurate machine learning algorithm can interpret these results in the context of known available data, detecting similarities and patterns which can evaluate the outcome. For example, models trained on crystallographic data can be used to predict crystallographic dimensionality and space group from thin-film x-ray diffraction.[cite: 10.1038/s41524-019-0196-x] Second, given the analyzed output of an experiment, emerging data science techniques can be used to efficiently build surrogate models. Suitably designed, they can take the analyzed output data and quickly estimate results that can steer the experiment.

This ability to automatically evaluate data as it is collected, and then suggest new directions, has applications across experimental science, including:

(1) Query and steer multi-dimensional processes;

(2) Suggest placement of sensors and data collection, determining which locations give the newest information;

(3) Efficiently construct surrogate models, especially when collecting information across multiple modalities, such as through combining imaging with chemical and materials databases. Considerable information can be gleaned by querying high-dimensional state space with many different techniques, such as tomography, mass spectrometry, and high-resolution IR imaging. Broad dissemination of such approaches can be utilized at multiple scales, from operation of single instruments to collections of instruments in individual labs and large-scale facilities. For example, successful demonstrations to date span autonomous benchtop chemical synthesis to the synchrotron experiment case study discussed below.

Case study 1: Autonomous experiments in traditional laboratories. Within a single laboratory, autonomy can couple control and measurement, delivering purpose-built experiments. Examples include microfluidic systems for synthesis and characterization of colloidal nanoparticles coupled to machine learning-based optimization of the optoelectronic properties,[21-23] and computer-controlled test stands for creating and electrochemically characterizing arbitrary liquid electrolyte solutions coupled with online optimization.[24] Autonomous organic synthesis optimizations in flow-based reactors have been demonstrated for a number of different systems,[cite: 10.1039/C9RE00096H & 10.1021/acs.joc.8b01821] and software has been developed to autonomously steer commercially available equipment in performing organic synthesis optimizations.[cite: Lapkin 10.1002/cmtd.202000044] Even when commercially available equipment does not exist, it is possible to combine existing equipment with only minimal modification. In one recent example, an autonomous system for optimizing Suzuki-Miyaura coupling reactions was created by combining commercial liquid handling and high performance liquid chromatography (HPLC) systems; the only hardware modification needed was to install an HPLC valve on the robot deck and incorporate relay switches to trigger the chromatographic equipment.[cite:10.1038/s42004-021-00550-x] A more wide-reaching approach exploits general purpose robots that interact with existing laboratory equipment[2]: in one configuration, a robot

synthesized 688 photocatalysts over 8 days using a Bayesian optimization scheme without human intervention, leading to a 6-fold increase in desirability compared to the initial compounds. Even with limitations in how existing knowledge, theory, and physical models are implemented in the autonomous search, such examples illustrate a time-efficient cost-effective use of available resources, shortening a project from months and years to a week. Ideally, in the future, the combination of advances in knowledge, theory and models) will enable optimized synthesis of novel compounds with targeted properties. However, even the development of autonomous processes for individual analytical subtasks within a research project—such as determining the solubility [cite: 10.1016/j.isci.2021.102176] and determining kinetics models by HPLC experiments[cite: 10.1039/C8RE00345A]—can be useful both for accelerating research progress and as building blocks for future systems.

Case study 2: Autonomous steering at synchrotron light sources. One current example of autonomous steering is provided by the gpCAM mathematical, algorithmic, and software framework[26] which has been used for a wide variety of experiments across the United States and abroad (**FIG. 2**). First, measurements of an autonomous experiment are chosen based on previous measurements. Next, surrogate model functions are computed by machine learning-based Gaussian process prediction, which can be constrained by domain-knowledge. Hybrid optimization methods are then used to locate the next best measurements. Finally, choices for optimal measurements are determined as a function of the surrogate model, its uncertainty, and the costs of a measurement. Using this approach and software framework, beam utilization was increased at Brookhaven's Center for Functional Nanomaterials and National Synchrotron Light Source-II from 15 percent to over 80 percent[26,27] with a five-fold decrease in the number of required experiments to obtain the same results. At the Berkeley Synchrotron Infrared Structural Biology beamline at the Lawrence Berkeley Laboratory's Advanced Light Source, the required amount of biological spectroscopy data that needed be collected was reduced as much as 50-fold.[28] At neutron sources at the Institut Laue-Langevin, experiment durations have been reduced from days to one night.

3.2. Harnessing complexity with data science

One well-traveled road in chemical experimental science has focused on optimizing control over the sample and the experimental apparatus. These efforts have emphasized the control of a limited set of critical parameters, which, in turn, imposes limits on the analysis by highlighting a few outputs with high signal to noise ratio to enhance interpretability. This constrains experimental methods to maximize control and homogeneity and minimize noise, fluctuations, and heterogeneity.

The scientific usefulness of this framework derives directly from how successfully the critical properties of experiment can be controlled. While this approach has generated many impressive successes, the inevitable limitations in sample and experimental control present

significant limitations to experimental design. Data science approaches can augment and expand the scope of experimental science by both accelerating the analysis and interpretation of experiments and enabling experiments to be performed successfully where control is impractical or risks undesirable alteration of the phenomena under study. As an example, current data science techniques applied to structure and image reconstructions are able to extract information from measurements recorded with far more noise and uncertainty than previously possible, greatly increasing the set of "viable" and productive experiments.

Clear cases where data science approach would be valuable include, but are not limited to, experiments using stochastic or noisy instrumentation like X-ray free electron lasers (see case study below), and field studies where natural variations in the environment provide an alternative means of determining how chemical systems respond to changing environmental conditions. In these examples, control of the relevant experimental parameter space cannot or should not be exercised ; the parameter space must be fully measured and correlated with the relevant experimental observables. This approach to experimentation greatly increases both the data volume and the challenges in identifying correlations between the measured, rather than controlled, variables with the experimental observables. The payoff is that information can be extracted that would be lost to traditional techniques of averaging over uncontrolled fluctuations or left unexplored by an experimenter with full control of the sampling of parameter space.

Goal - Relax requirements for experimental control and *a priori* design. The adoption of data-science methods in experimental planning and analysis enables scientists to reimagine the way we design and perform experiments by shifting the focus away from controlling the critical parameter to measuring fluctuations within the critical parameter space. Measuring, rather than controlling, the critical parameter space of the experiment shifts the emphasis of experimental design to data-intensive diagnostics that must be integrated into the experiment. This also requires significant changes in analysis, since the absence of control can generate significantly larger datasets with more complex correlations between the chemical properties of the sample being measured in the experiment and the instrument sampling of parameter space measured with diagnostics. An example, designed to mitigate the significant shot-to-shot variation of X-ray free electron lasers, can be found in **FIG. 3**. This approach may be a product of necessity instruments with delicate stability regimes, but it also presents the opportunity to identify unexpected correlations, since natural fluctuations in the experimental apparatus may generate experimental results a scientist may be biased to avoid. By providing real-time sampling of a complex experimental parameter space, pre-planned experiments are replaced with on-the-fly adaptive methods that reduce the time to acquire a signal and reduce problems of data redundancy. Furthermore, instead of relying upon a single high signal-to-noise output, alternative approaches might rely on many more weak (but easy to collect) signals to make chemical measurements.[29] FIG. 3 shows how the integration of fast ML/AI enabled analysis can lead to data-driven autonomous experimental workflows.

What is needed. To take full advantage of the above possibilities requires multiple advances, including: (a) development and implementation of diagnostics for measuring all experimental parameters that would have traditionally been controlled; (b) integration of these diagnostic signals with the experimental observables; and (c) fast analysis techniques for correlating the data from the diagnostics with the experiment to enable real-time assessment of experimental progress building representative surrogate models on-the-fly as data is collected.

A pivotal role for ML and AI. ML and AI techniques offer powerful chances to identify unexpected or hidden correlations revealed by the fluctuations in the experimental parameter space. Two examples include:

Case study 3: Identifying natural experiments from laboratory metadata. Chemical reactions can be highly sensitive to environmental conditions such as humidity. The typical experimental control strategy is to perform reactions in a glovebox, but this presents operational challenges. An alternative, demonstrated recently in the context of halide perovskite crystal growth, is to capture comprehensive electronic records of laboratory conditions associated with each experiment over an extended period of time.[cite: doi:10.1063/5.0059767] Using a dataset of 8470 experiments captured over a 20 month period, it was possible to identify statistical anomalies in reaction outcome that were correlated with laboratory humidity. The researchers confirmed this hypothesis by performing deliberate interventional experiments, and in the process, discovered systems in which water interfered with inverse temperature crystallization, contrary to previously hypothesized mechanisms.

Case study 4: X-ray Free Electron Lasers (XFELs). XFELs have transformed x-ray science by producing the world's brightest x-ray beams. The lasing process that generates these beams also leads to significantly larger fluctuations in key experimental parameters, particularly compared to synchrotron-based x-ray sources. Attempts to control key beam properties like pulse spectra, intensity, and duration have only led to partial success to date. As an alternative, one could instead measure large fluctuations in pulse properties on every shot, and then use data science methods to deconvolve the influence of pulse fluctuations for the observed experimental signal. In addition to reducing the experimental requirements for XFEL performance, this approach has the benefit of using every photon and thus giving an automatic brightness upgrade; for an XFEL this is a 100x improvement. Furthermore, this has the benefit of improving temporal resolution.

The above opportunities come with challenges. The inability to control the experimental apparatus necessitates the performance of two parallel measurements: one on the x-ray beam and the other on the sample being interrogated by the x-ray beam. Additionally, the success of the experiment requires high fidelity diagnostics and analysis methods to ensure x-ray beam fluctuations can be robustly differentiated from variations in the sample properties being investigated. Furthermore, adopting a supervised learning approach would initially require conducting parallel experiments using a traditional apparatus so as to build an appropriate training

set; as a result, cost savings would not be immediately realized, but would come when this information is applied to future sites. The planning stages of this type of work would require deep involvement of data science and modeling experts to assure stakeholders that algorithms will be able to perform this task robustly and reproducibly.[30]

Another advancement made by the XFELs is the application of X-ray sciences in the chemical phenomena in the femtosecond time regime. Ever since the launch of "femtochemistry" by Zewail and others, the ultrafast interactions initiated by the absorption of a photon have driven a quest to understand, and ultimately control the ultrafast structural dynamics of photoactivated molecular systems [see, e.g., 10.1146/annurev-physchem-032210-103522]. This quest has made it imperative to deal with noisy, incomplete, and fleeting signals recorded with substantial timing uncertainty. While experimental attempts to deal with such signals will continue to advance, recent AI/ML approaches have brought the greatest rewards [see, e.g., 10.1038/nature17627, https://doi.org/10.1101/2020.11.13.382218].

The measurement of dynamics is an important case in point. Since the celebrated work of Takens,[31,32] it has been recognized mathematically that the evolution of a wide range of dynamical systems is tightly constrained. As such, much less data is needed to recover dynamical information than currently thought necessary for proper experimental analysis. Takens showed that a series of snapshots, each representing a subset of the system variables, suffice to determine the behavior of dynamical systems as though all system variables had been measured. The ML-based realizations of this remarkable possibility are now being applied to ultrafast chemistry data previously thought too noisy, too incomplete, and too imprecise to be useful.[18] Extensions of this approach have been used to estimate the gestational age of fetuses with unprecedented accuracy,[33] indicating the generality of the algorithmic methods.

3.3. Data-driven experimental discovery

Not all important challenges in science conform to easily testable hypotheses. Research in chemistry often targets critical metrics, such as a specific photovoltaic energy conversion efficiency or a specific selectivity for a catalytic reaction. These metrics require materials to achieve performance beyond what has been demonstrated previously, so interpolation is not an effective strategy. Extrapolating from known materials and known phenomena may prove insufficient to hit a challenging performance target and motivate exploration off the well-beaten path. Hypothesis-driven research, which is generally derived from prior knowledge, and relies on testing a postulated outcome, may restrict inquiry and exploration.[34]

Goal - Automated Serendipity. In the absence of a hypothesis, trial and error becomes intractable as the search space increases. Efforts in lab automation can reduce the time needed for synthesis, characterization, and data-interpretation, thus increasing the rate at which new trials can be performed (this builds upon the lab automation efforts discussed in Section 3.1.). More broadly, data science approaches can be used to automate the process of extracting new "ideas" to try based

on collected datasets.[11] [also add relevant cites to: doi:10.1063/5.0059767 --cited above & 10.1038/s41524-019-0196-x --cited below] Comprehensive data management (discussed in Section 4) facilitates the process of identifying unexpected variations that can suggest directions for more deliberate inquiry. For this type of application, prediction accuracy is less critical because it suffices to be wrong less often, so as to focus on a more tractable portion of the available parameter space for experimental validation.

What is needed. Enhancing metric-driven research requires: (a) efficient and unbiased search and analysis tools, or at least tools whose bias is clearly delineated and transparent, (b) implementation of ML methods to identify unexpected or hidden correlations revealed by the fluctuations in the experimental parameter space, and (c) autonomous direction of search based on prior findings.

A pivotal role for ML and AI. Instead of performing a few experiments carefully selected by the chemist, this approach favors performing larger-scale combinatorial experiments to explore a broader and less biased search space. A short-term goal is merely to perform more experiments over the broadest possible search space, which is the goal of "classical" high-throughput experimentation or combinatorial chemistry.[35] More long-term goals use ML and AI to accelerate the characterization process and optimal selection of new experiments. Finally, there is a need for machine learning interpretability, and explainable AI (XAI) to inform humans: this may necessitate chemistry-specific interpretable machine learning methods.[36] Some early realizations of this in experimental chemistry include extracting hypotheses about organic molecular structure determinants of energy levels and solubility[cite: 10.1088/2632-2153] and human-algorithm teaming for synthesis of polyoxometalates.[cite: 10.1021/acs.jcim.9b00304]

Case Study 5 - Serendipity-driven reaction discovery. This type of non-selective "automated serendipity" has been successful in discovering organic reactions for photoredox catalyzed C-H arylation,[37] and palladium catalyzed C-N cross coupling.[38,39] For a general review, see Ref. [40]. Each of these have relied upon experimental hardware developments to perform synthesis and characterization with greater parallelism and smaller quantities of reagents. Data interpretation is accelerated by using data-science methods to identify when a reaction has occurred. In its simplest form this can entail looking for differences in product and reactant spectra and using this to prioritized subsequent experimental rounds,[41] with the understanding that this can provide only a preliminary investigation, and that subsequent human reinvestigations may be necessary to confirm the spectral interpretations.[42] A more sophisticated approach would use this data to construct empirical relationships between catalyst and substrate structures and the catalytic efficiency;[43] the resulting structure-property models can then serve to prioritize subsequent experimentation. Finally, a higher level goal is to perform autonomous optimization of the catalyst, substrate, and reaction condition designs using automated experimentation and planning algorithms.[44] In materials science, a similar progression from high-throughput synthesis and characterization, to increasing automated interpretation, to autonomy has also been reviewed,[12];

again, this is enabled by increased adoption of machine learning methods throughout the discovery lifecycle.[8,45]

Data-science approaches can help facilitate this serendipitous discovery process by reducing the need to "know" what one is looking for ahead of time. An example comes from the development of rare-earth-free permanent magnet materials.[46] A wide variety of Fe-Co-X alloys are synthesized combinatorially, resulting (in some cases) in one or more phases: many of these are unknown. Using non-negative matrix factorization methods, diffraction spectra are decomposed into estimates of the pure material spectra (which may never be previously observed, or can be matched against known databases) and estimates of the relative contribution of those phases. While the goal is to produce a phase diagram of different compositions, building a complete map over the compositions requires too much instrument time. Instead, a further improvement uses active learning approaches and Bayesian optimization methods to prioritize the (automated) acquisition of new experimental data points.[47] Reducing the number of diffraction measurements that must be acquired by several orders of magnitude reduces the amount of beamtime required or even enables the using of a single-PI scale diffractometer instead of a beamline source.

3.4. Data management and networking

Realizing new experimental paradigms for chemistry requires human and AI researchers to access a broad range of chemical information. Optimally, such information would include a variety of process and characterization data, as well as the metadata providing context for the experiments. We refer to this as a "data network" to invoke the imagery of a network wherein nodes are data from chemistry experiments and connections between nodes encode how the data are related. Scientific knowledge emerges from the relationships between material observations and interpretation, and data science can help shed light on these relationships.[48] Data networks leverage the scale and variety of modern chemistry data to enhance the utility of data-driven methods in chemistry experiments: here, we describe some important experimental and data science efforts needed to enable key efforts such as building a repository of knowledge by networking data, encoding the current state of a scientific field, and facilitating adoption of data science methods in chemistry experiments.

Goal - Repositories of knowledge. The primary goal of networking data is to share accumulated results to enable humans and machines to derive new knowledge from old data. Such an environment will allow scientists to directly explore and visualize the state of the field from such repositories and obtain faster access to details essential to research projects (as a complement to traditional literature search).

Traditionally, chemistry knowledge repositories are aggregated by a single organization and take the form of licensed datasets, reference volumes, or reference websites; some widely-used examples are the Powder Diffraction File,[49] the CRC Handbook, and the NIST Chemistry WebBook.[50] In some sense, these repositories reflect refined chemical knowledge. As an

example, consider the trajectory of experimental data from raw data acquisition to contextualization, analysis, interpretation, and validation through additional experiments. Repositories understandably have focused on only the final outcome of this data funnel. Instead, managing and cataloging data throughout these phases of knowledge refinement can help address issues of data scarcity that arise in the adoption of data science.

Given the volume of data now being generated by chemistry experiments, and the desire to accelerate the research workflow, there has been an increasing number of crowd-sourced efforts to build knowledge repositories at the same pace as research. Many such repositories in the chemistry domain are just getting started, however, one especially successful past example from the biology domain is the Protein Data Bank (PDB), a database that contains up-to-date structural data for large molecules.[51] Deposition into the PDB is a requirement for publication, resulting in at least $12 billion worth of research data contributed to the database over the past 40 years, and this a central repository produces $2.5 billion worth of increased research productivity annually (as of 2017).[52] Furthermore, this accumulated data enabled the development of the recent AlphaFold[cite: 10.1038/s41586-021-03819-2] and RoseTTAFold[cite: 10.1126/science.abj8754] models for predicting the three-dimensional structure of proteins based solely on amino acid sequence. Beyond data management, the biological and pharmaceutical fields have successfully created data networks and knowledge graphs that when coupled with rapidly evolving graph learning methods enable learning of new biological features, drug properties, etc.[53] Analogous advancements with experimental chemistry data would be a watershed advancement for incorporation of data science. To date, the most successful repositories of experimental chemistry data are structural databases such as the Cambridge Structure Database and ICSD, and spectral databases, such as the NMRShiftDB.[cite:https://nmrshiftdb.nmr.uni-koeln.de][cite: 10.1002/mrc.4263] An IUPAC project on "Development of a standard for FAIR data management of spectroscopic data" (FAIRSpec) was founded in 2019,[cite: https://iupac.org/project/2019-031-1-024] and progress is described in a recent report.[cite:https://doi.org/10.1255/sew.2021.a9] Databases of organic synthesis (such as Reaxys) are proprietary, and do not allow for free contribution and use. Comprehensive community repositories of chemical processes do not yet exist, but may emerge from nascent efforts at developing schema for representing laboratory actions such as XDL,[cite: 10.1126/science.abc2986] IBM/RXN,[cite:10.1038/s41467-020-17266-6] Autoprotocol,[cite: https://autoprotocol.org] and the ESCALATE materials/action specification[cite: ESCALATE paper], may serve as the basis for these types of projects. The advent of the Department of Energy PuRe Data Resources embodies an important step in this direction.[cite: https://www.energy.gov/science/office-science-pure-data-resources]

Once data networks are available, they can be used to accelerate the generation and testing of hypotheses via AI-driven encapsulation of existing knowledge. For instance, a network based on high-throughput density functional theory calculations can be explored by humans through web-based visualizations as illustrated in **FIG. 4**,[54] while its network metrics can also be used in a machine learning model to predict (or hypothesize) the synthesizability of new inorganic compounds.[55] This mode of hypothesis testing, which builds upon the concepts of Section 3.3, is

significantly different from the cycle of first proposing a hypothesis and then designing and completing experiments before any validation takes place. Instead, with a network of data, one can identify existing knowledge that accelerates hypothesis testing, adding value to data that were collected for a different purpose.

Data networks can also enhance the development of accurate predictive models to the benefit of the autonomous experimentation described in Section 3.1. While a single lab may possess insufficient data for training surrogate models, data networks may contain auxiliary data to augment the single lab's data. Of course, this has its own challenges: training models using data from multiple sources is non-trivial, and developing techniques for utilizing and linking heterogeneous from various sources is a major undertaking.

A final data science challenge that can be addressed with data networks relates to the frequent need for predictive models to extrapolate beyond the existing corpus of chemistry knowledge, as discussed in Section 3.3. While true extrapolations may be wildly inaccurate, data networks can be constructed with the appropriate connections, so that apparent extrapolations may in reality lie comfortably within the domain of validity of existing models. For instance, a given property of a given chemical may not have been measured by a given technique, but previous experiments that share the same property, chemical, or method may be of value; this assumes a shared framework for expressing the relationships upon which data science methods can be built.

What is needed. Realizing data networks and their benefits will require a variety of cultural and technical advancements. Many of the relationships among chemical experiments lay in their metadata, which includes details of the instruments and their settings, including essentially any knowledge required to reproduce the data. Agreed-upon software formats for recording experiment parameters, as opposed to manual setting of multiple knobs whose data record is limited to written notes, will greatly facilitate consistent tracking of experiment metadata. Data management programs such as ESCALATE[cite] and ESAMP[cite: 10.26434/chemrxiv.14583258.v1] are exemplars of this approach for chemistry and materials data, and further…" the data stewardship by making the data and metadata inseparable; an example of the rich types of interactive experiment description reporting that this enables can be found in the electronic Supporting Information of Ref. [57]

The chemical and analysis provenance of data is also critical. From lab notes to publications, chemicals are often labelled by what they are intended to be, and data annotations such as "background-subtracted" are often aspirational. From a data science perspective, the chemical under investigation in an experiment is most well-defined by the sequence of prior processes and experiments that produced the chemical. Assessing this provenance from literature data is often difficult if not impossible, motivating a re-thinking of how experimental data should be recorded and tracked.

Regarding data processing and interpretation, there are complementary challenges. Expert decisions during data analysis, for example in identifying which portion of a spectrum to analyze or what data artifacts may be present, are based on experience-based knowledge. Tracking the

provenance of data analysis will facilitate removal of human basis and uncover valuable information from raw data. On the other hand, application of expert prior knowledge may be necessary to gain traction in data analysis, and encoding this knowledge in data science algorithms is a major yet crucial challenge. Ultimately, artificial intelligent algorithms will have their own experience-based chemical knowledge, but only if we can provide the same quantity and quality of data, metadata, and provenance that underlies the knowledge progression of expert scientists.

We note that there are numerous practical challenges related to the ingestion and management of metadata and data provenance, additional imperfections of the data itself, as well as less technical considerations such as intellectual property and incentivization schemes. We refer to a recent DOE report "BES Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning" for recommendations on technical aspects of the data pipeline and network,[58] and Ref. [59] for a survey of motivations for building a data network.

A pivotal role for ML and AI. To establish data networks that enable scientists to aggregate and search relevant chemical knowledge, data science must be incorporated in data management to learn the relevance of metadata, provenance, and domain knowledge so that they can be appropriately modeled in data networks. Networking data should commence with models of relationships encoded in existing theories, as was recently demonstrated by PropNet,[60] which is built on the foundation of symbolic equations from materials physics. Bringing this network concept to express interrelationships of experimental data is a new paradigm in data management for chemical sciences.

Case study 6: X-ray absorption spectroscopy: XAS is a ubiquitous chemistry experiment technique: data science methods could facilitate faster data analysis and chemical feature recognition in measured elemental patterns. Machine learning models have been developed that can predict chemical features from XAS patterns collected under the same conditions as a relatively large training set, which for this application has been the computational XAS.[61-64] Expanding the scope of these models to experimental spectra could be enabled by aggregating XAS data from dozens of beamlines worldwide that collectively have acquired many thousands or perhaps millions of spectra to-date. However, variants of the technique rapidly complicate the problem, from fluorescence to electron detection modes, and from hard X-ray-open-atmosphere to soft X-ray-vacuum, and to various in situ and operando measurements of chemicals/materials in chemical reactors or other actively controlled conditions. As a result, beyond the challenge of aggregating the data itself, defining and representing the context of every XAS measurement is quite difficult and must begin with well-tracked and machine-readable metadata. Nevertheless, recent progress has been made on this front,[65] which is a critical step on the path to an XAS data network.

4. Recommendations

Regardless of the potential and limitations of ML and AI, there are still some uncertainties about how these approaches are transforming chemistry research. Just as Bayesian inference has impacted our inferential understanding, ML is rapidly changing the meaning of experimental knowledge. Such wide-ranging transformations provide a rich environment that should be part of every scientist's toolbox. To be sure, the limitations of ML are manifest. These include: whether an algorithm trained on one dataset can be used to produce reliable answers about a different dataset; whether a particular algorithm is robust against noise or attempts to deceive it; what the basis is for the answers an algorithm provides; and whether these answers are free of bias.

Transformation of experimental chemistry by ML requires active engagement in utilizing existing tools, injecting domain-specific knowledge into their design, and co-opting rather than avoiding their increasingly powerful impact (**FIG. 5**). We hope that these recommendations contribute to facilitating this engagement.

- **Develop data science methods for chemistry.** Chemists are increasingly incorporating data science techniques into their research. Many early applications used off-the-shelf methods to achieve dramatic advances, but there is a critical need to understand the limitations of existing algorithms for chemical datasets and develop specific ML tools for chemical problems that require new approaches. Methods are needed that incorporate relevant physical laws and other constraints to produce physically reasonable solutions, provide internal consistency, and capture experimental uncertainty. This may include representations that incorporate the appropriate symmetry behavior of structures and physical interactions [cite: 10.1021/acs.chemrev.1c00021] (such as invariance/equivariance[cite: review 10.1016/j.trechm.2020.10.006] and isometry[arXiv:2108.07233]), and periodic relationships of elements[cite: 10.1063/1.5108803] Such methods can form the basis for new modes of experiment, such as the case in Section 3.2 where relaxing experiment control enables the acquisition of a larger information throughput.

  Recommendation: Develop new ML/AI representations and techniques specific to chemistry by partnering together with data scientists, and help train a complementary workforce of interdisciplinary experts that can leverage the methods in experiment design and analysis. As an example of the value of these interdisciplinary approaches, a recent breakthrough, called "DeePMD-kit" (https://ieeexplore.ieee.org/document/9355242/) combines ab initio modeling, high performance computing and machine learning to tackle "first principles" molecular dynamics simulations by approximating ab initio data with deep neural networks, allowing far more extensive calculations and offering a bridge between machine learning and physical modeling. Similar types of combinations of ab initio results with data science methods and autonomous experimentation have been used to accelerate chemical optimization tasks.[cite: 10.1016/j.matt.2021.01.008] Building such bridges for experimental chemistry data will enhance the interpretability of data science

models and enable their deployment with smaller datasets since models are well conditioned by the incorporated chemistry.

- **Extend the reach and applicability of data-driven approaches in the chemical sciences.** Data-driven approaches are by nature interpolative, and typically obtain results by capitalizing on a library of dense, nearby, known solutions. With large enough datasets, this is often sufficient for solving many scientific problems. However, purely interpolative methods fail when one needs to extend predictions into new and unexplored regions of parameter space, or when dramatic changes happen between sparse elements. There is an interpolative power of data science that can be used to direct future research outside the bounds of current measurement and observation.[cite: supporting example "Can machine learning find extraordinary materials?" 10.1016/j.commatsci.2019.109498] As illustrated in Section 3.3, research in this direction can potentially be applied to accelerate discovery.

  Recommendation: Develop methods that can work with sparse representations in high-dimensional parameter spaces, providing guideposts for understanding the accuracy of interpolative measurements, and the applicability of extrapolative methods.

- **Transform research workflows by integrating measurement and observation tools, robotics, data-pipelines, and compute resources.** Data science methods can accelerate decision-making. To take advantage of this possibility, we need integrated laboratory automation systems that give algorithms and workflow a way of enacting processes in the laboratory, monitoring the results, and depositing the resulting data into shared repositories. As described in Section 3.1, accelerating the experiment cycle is especially valuable in shared facilities (e.g., synchrotrons), but equally needed in single-PI laboratories. Taken together, these integrated systems have the potential to unleash a virtuous cycle—experiments conducted by automated systems or robots are "born digital", reducing barriers to data sharing and reuse, and facilitating the development of better data science methods—but there are significant technical barriers. In addition to depositing data and software in FAIR repositories, open hardware should be encouraged, with relevant CAD files and control code deposited. Currently this type of data often appears in supporting information, but could also be the primary topic of articles in journals such as HardwareX [cite: https://www.journals.elsevier.com/hardwarex] exist for creating citable records for these types of efforts.

  Recommendation: Encourage a co-design approach to hardware, software, and algorithm development. Interdisciplinary teams can often reimagine the entire range of experimental workflow to embrace a new accelerated approach that integrates measurements, data, algorithms, and computing. Develop both modular and complete solutions, with an emphasis on interoperable and open hardware and software.

- **Integrate diverse data sources.** Chemical data are diverse, consisting of spectroscopic observations, structural information, processes descriptions, and many other types of measurements. Combining different types of data sources provides stronger evidence than any single data type. Often, crucial details are present only in unpublished "failures", calibrations, or metadata. While specific types of chemical data have been aggregated (e.g., crystallographic data), there are currently only limited automated mechanisms by which individual experiments consisting of diverse elements can contribute to a broader whole. Human researchers excel at placing a new piece of data in the context of the data and knowledge of their field, but their underlying reasoning about prior knowledge to make these assessments suffers from being slow, costly, biased, and inconsistent. AI methods for contextualizing data should be developed, requiring establishment of a foundation for automatic management of relationships in chemistry data, in order to achieve the goal of a network of data described in Section 3.4.

  Recommendation: Develop better ways of representing networks of data that encode the relationships between evidence in a machine-readable way. Create incentives for comprehensive data-sharing and reduce technical and social barriers to data deposition and access, through the creation of shared repositories and other mechanisms.

Acknowledgements:
This article evolved from presentations and discussions at the workshop "At the Tipping Point: A Future of Fused Chemical and Data Science" held in September 2020, sponsored by the Council on Chemical Sciences, Geosciences, and Biosciences of the US Department of Energy, Office of Science, Office of Basic Energy Sciences. The authors thank the members of the Council for their encouragement and assistance in developing this workshop. In addition, the authors are indebted to the agencies responsible for funding their individual research efforts, without which this work would not have been possible.

Author contributions:
K.J.G, J.G., L.H., A.O., J.Sc., J.Se., F.M. T., and J.Y. equally contributed and wrote this article of Expert Recommendation.

Figures:

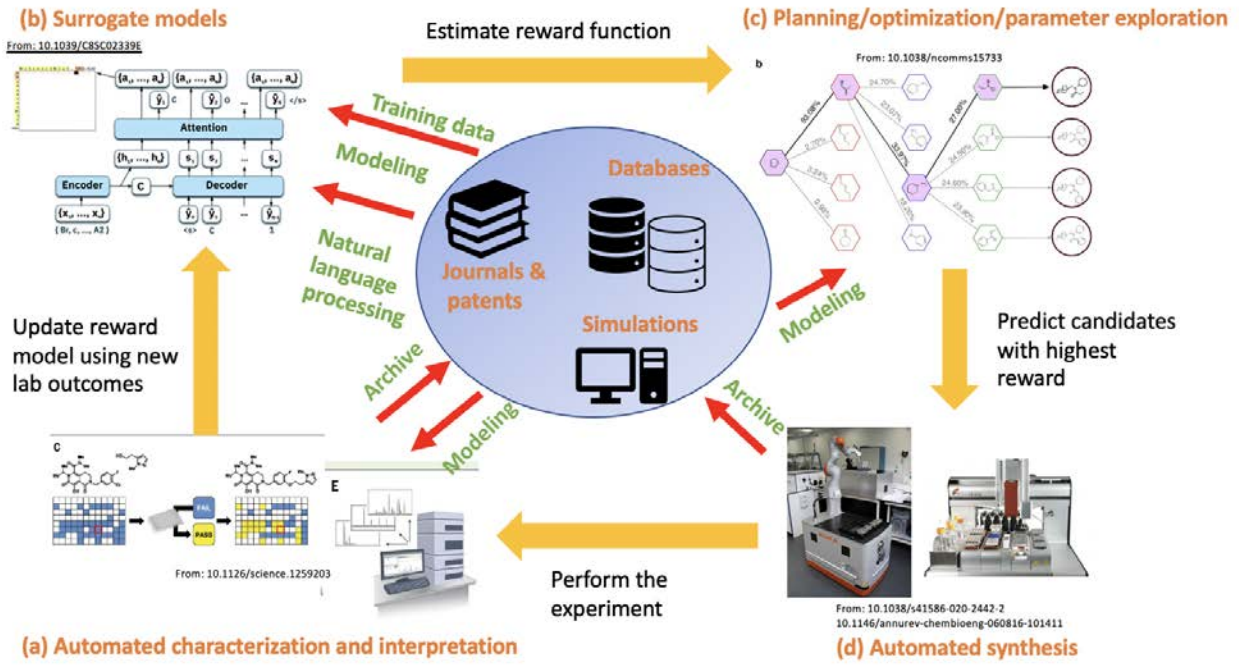

FIG. 1: Role of Data Sciences in Experimental Processes. Data science can play many roles in experimental processes, such as the autonomous synthesis and characterization (Section 3.1). To accomplish the experimental tasks (yellow arrows), several technologies (a)-(d) are required, necessitating data flows (red arrows) to and from repositories (Section 3.4).

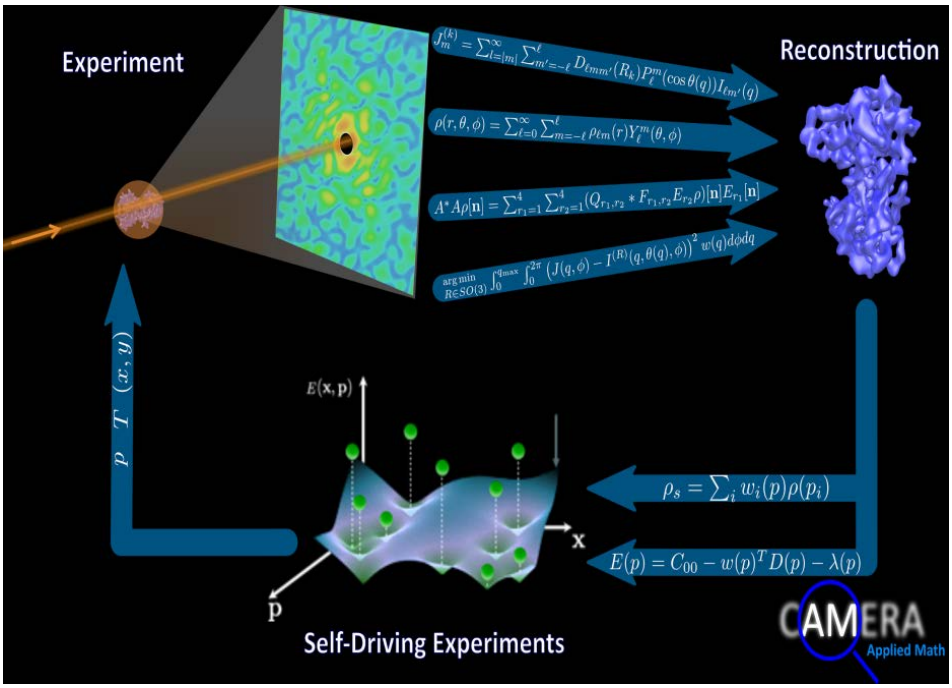

FIG. 2: AI/ML to accelerate, autonomously control, and understand experiments, using state-of-the-art mathematics coupled to advances in data science. Multi-tiered iterative projections (M-TIP) accurately interpret scattering images from light source experimental data, and Gaussian processes (gpCAM) suggest and drive new experiments --- working together in an autonomous loop, they optimizing the use of complex equipment (Figure made by J. Donatelli, M. Noack and J.A. Sethian (Univ. of California, Berkeley and Lawrence Berkeley National Laboratory).

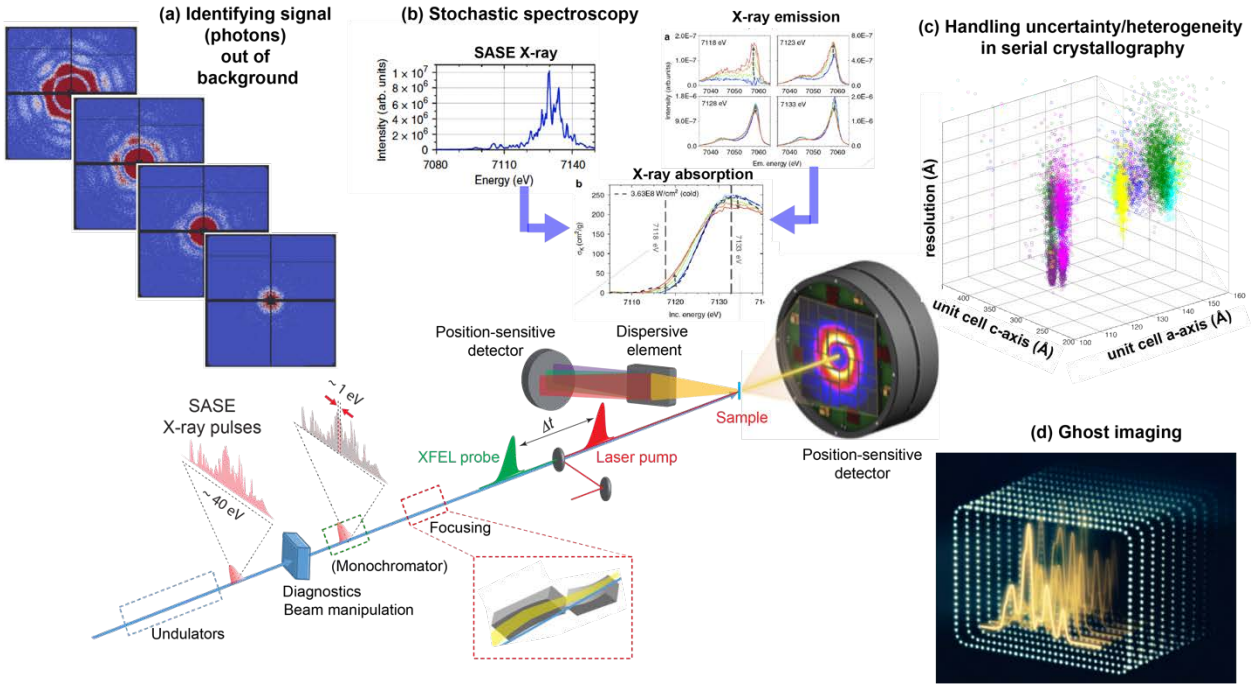

FIG. 3: **Application of ML to carry out new types of experiments at X-ray free electron laser facilities.** (a) Detecting small number of photons (signal) from a large instrument background in a single snapshot for imaging[66]; (b) X-ray spectroscopy that uses a stochastic nature of the XFELs, taking advantage of the random spikes of each XFEL pulse as a unique fingerprint, and correlating with outgoing emission signals from the system to construct spectra[67]; (c) An example of the heterogeneity in the unit cell distribution of Photosystem II crystals and its diffraction quality in serial crystallography where each XFEL pulse gives one diffraction image and in total about 3M data are plotted. From this visualization, one can learn that there are 5 different possible crystal isoforms. The authors of Ref. [68] later identified that dehydration was a critical parameter for shifting the isoform population. This on-line analysis of data is used to provide immediate feedback to determine subsequent sample preparation conditions for the best resolution. (d) Pump-probe ghost imaging. Similar to (b), this approach uses random spikes of XFEL pulses, and studies its interaction with matter. It can be used to map the full evolution of the system over time.[69] Image courtesy of Greg Stewart, SLAC National Accelerator Laboratory (https://www.energy.gov/science/articles/ghostly-images-could-ease-tracking-fleeting-reactions).

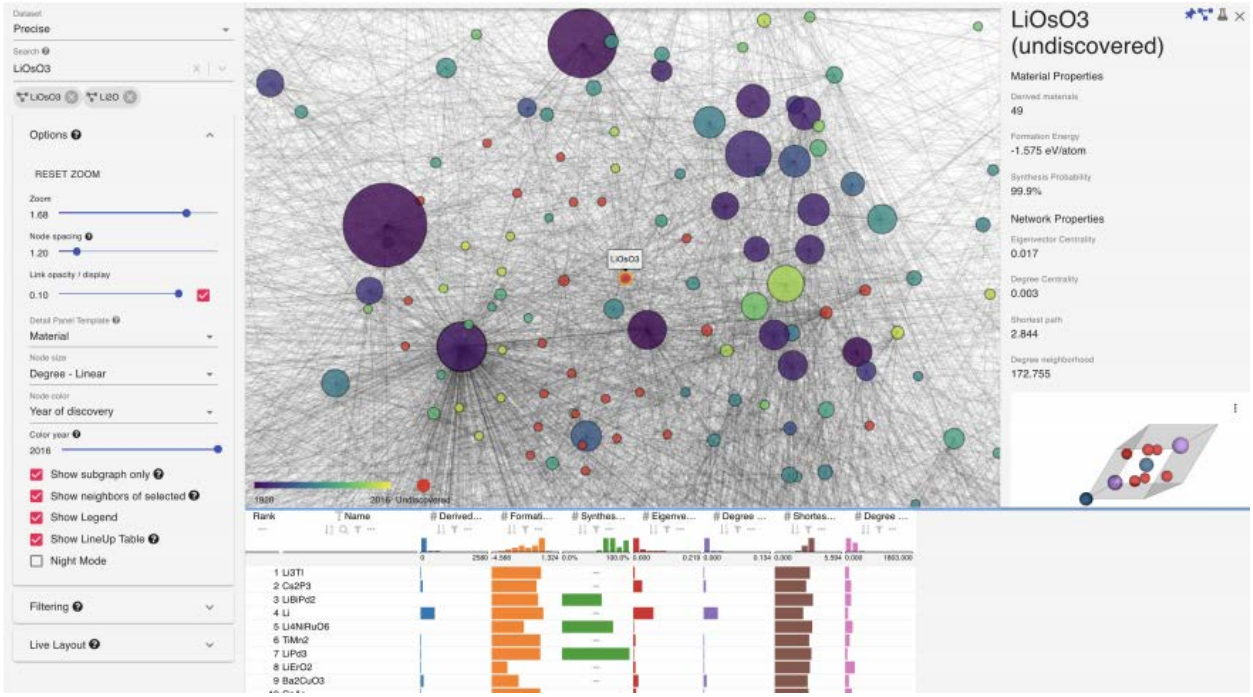

FIG. 4: MaterialNet - Materials Similarity Network, an early demonstration of the many interrelationships that exist among materials and chemicals. Networks can capture more relationships than a human could comprehend, and data science tools can learn from these relationships. From [54] (maps.matr.io). Reproduced under CC-BY license.

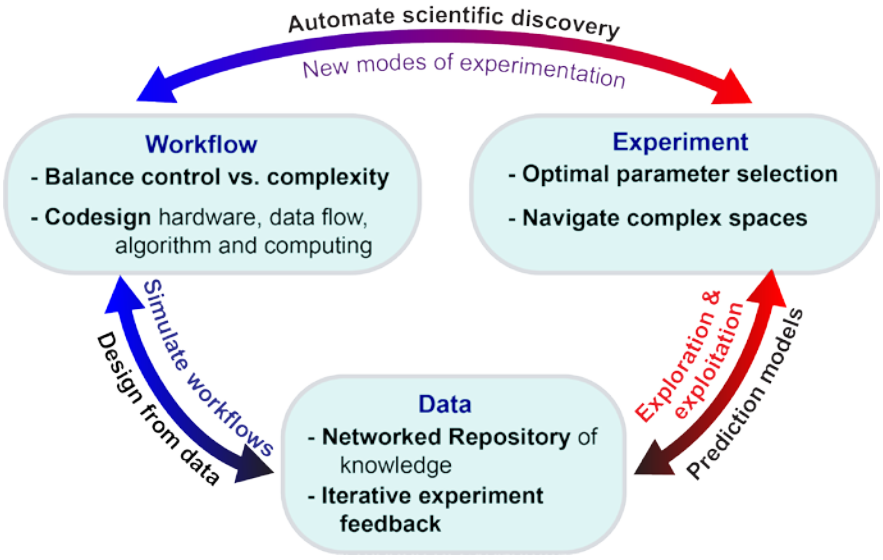

FIG. 5: Interplay of experiments, workflow, and data.

Box 1: Challenges associated with ML/AI.

ML has been typically applied to use-cases where the price of being wrong is small. In science—as in other fields—this is not always the case. With this in mind, important questions to critically evaluate the suitability of ML methods for application in scientific or other domains include:

- What criteria should be used to trust the output of a ML/AI analysis? That is, what level of verification is necessary and to what extent does that compromise the utility of the ML/AI approach?
- What evidence underlies how these methods make predictions? When is it reasonable or necessary to ask this question?
- Can AI/ML be used to predict, with quantifiable confidence, phenomena outside the domain used for constructing the algorithm? Currently, AI/ML are inherently designed for interpolation — given a big enough library of inputs matched with outputs, these algorithms can take a new input and combine information at nearby inputs to predict a possibly viable output. Scientific discovery, however, inherently involves investigation of new spaces (extrapolation or prediction), which stands in contrast to the primary focus of ML algorithm development to-date.
- An oft-stated virtue of these methods is the idea that they are transferable: predictive schemes in one field may be applied in other fields that appear to be unrelated. How can one know if and when predictions are transferable between fields?

Box 2: Three modalities by which data science could transform experimental chemistry.

Extract more information from existing, imperfect experimental data

In the most straightforward settings, data conforms to simple statistical expectations, with each snapshot representing an instance of noise added to a measurement of all relevant system variables. Such data rarely exist.

In reality, each snapshot represents an incomplete, noise-limited measurement of a subset of system variables. Real data are also often inhomogeneous, in the sense that each snapshot pertains to an unknown set of unintentionally changed system variables.

In other words, real data are incomplete (not all relevant system parameters measured), inhomogeneous (the snapshots emanate from differing values of one or more often unknown variables), and noisy (non-Gaussian pixel noise, inaccurate timestamps). Standard approaches to data analysis often successively reject "outliers" in order to obtain a sufficiently homogeneous dataset amenable to traditional analysis by averaging.

ML approaches, in contrast, attempt to "learn" the space spanned by the data, such as in identifying reaction coordinates ("collective variables") at work during the experiment, and use the information content of the entire dataset to reconstruct the system at any point in the space of reaction coordinates.[9,15-18] This offers a noise-robust approach to extracting far more information from the data than possible with traditional methods.

Optimally design experiments and workflow

Complex experiments with many input parameters generate sample points in high-dimensional spaces: the challenge of systematically navigating these spaces is rapidly outpacing human capabilities. Data-driven approaches can learn and exercise optimal control of experiments in real time, incorporating prior knowledge to efficiently find under-resolved regions and/or regions of interest. Such "on-the-fly" data methods can help experiments efficiently cover the landscapes on which the system of interest undergoes important, functionally relevant changes.[9,17]

Offer entirely new experimental modalities

The new generation of high-throughput instruments combined with the algorithmic ability to rapidly analyze very large datasets offers entirely new experimental modalities. As an example, chemical reaction events often take place via rarely sighted transition states. Up to now, one has resorted to complex time-resolved experiments to obtain snapshots of a system as it is driven over a transition state. In equilibrium, however, a collection of snapshots includes all states of the system, including those at high energies, albeit with exponentially diminishing probability.[9] A "sufficiently large" dataset of snapshots will thus include high-energy conformations. States at energies comparable with that released by ATP hydrolysis, for example, begin to appear in datasets with ~ 10^9 single-particle snapshots from an equilibrium ensemble of molecules. This offers the possibility to investigate important chemical processes without having to "track" each process in time. The key is the ability to collect and analyze billion-strong collections of single-particle snapshots, as dictated by the underlying statistical mechanics.

References:

1 Ourmazd, A. Science in the age of machine learning. *Nat. Rev. Phys.* **2**, 342-343, (2020).

2 NSF Workshop on 'Framing the Role of Big Data and Modern Data Science in Chemistry'. ((Arlington, VA, 17-19 April 2017); CHE-1733626, 2018).

3 Mission Innovation "Energy Materials Innovation Workshop". (Mexico City, 11-14 September 2017, (2018).

4 *X-Ray Free Electron Lasers: Applications in Materials, Chemistry and Biology*. (eds. Bergmann, U.; Yachandra, V. & Yano, J.) (Royal Society of Chemistry, 2017).

5 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547-555, (2018).

6 Morgan, D. & Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Ann. Rev. Mat. Res.* **50**, 71-103, (2020).

7 in *ACS In Focus* -1 (American Chemical Society, 2020).

8 Wang, A. Y.-T. *et al.* Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chem. Mater.* **32**, 4954-4965, (2020).

9 Dashti, A. *et al.* Retrieving functional pathways of biomolecules from single-particle snapshots. *Nat. Commun.* **11**, 4734, (2020).

10 Selvaratnam, B. & Koodali, R. T. Machine learning in experimental materials chemistry. *Catal. Today*, (2020).

11 Shi, Y., Prieto, P. L., Zepel, T., Grunert, S. & Hein, J. E. Automated Experimentation Powers Data Science in Chemistry. *Accounts Chem. Res.* **54**, 546-555, (2021).

12 Stein, H. S. & Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **10**, 9640-9649, (2019).

13 Flores-Leonar, M. M. *et al.* Materials Acceleration Platforms: On the way to autonomous experimentation. *Curr. Opin. Green Sustain. Chem.* **25**, 100370, (2020).

14 Experimental workflows inherently involve a sequence of physical tasks coupled to analysis of results: in the chemistry research setting workflows additionally include quality control and decision making interleaved throughout the sequence of tasks.

15 Dashti, A. *et al.* Trajectories of the ribosome as a Brownian nanomachine. *P. Natl. Acad. Sci. U.S.A* **111**, 17492, (2014).

16 Hosseinizadeh, A. *et al.* Conformational landscape of a virus by single-particle X-ray scattering. *Nat. Methods* **14**, 877-881, (2017).

17 Ourmazd, A. Cryo-EM, XFELs and the structure conundrum in structural biology. *Nat. Methods* **16**, 941-944, (2019).

18 Fung, R. *et al.* Dynamics from noisy data with extreme timing uncertainty. *Nature* **532**, 471-475, (2016).

19 Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem. Int. Edit.* **59**, 22858-22893, (2020).

20 Coley, C. W., Eyke, N. S. & Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem. Int. Edit.* **59**, 23414-23436, (2020).

21 Epps, R. W. *et al.* Artificial Chemist: An Autonomous Quantum Dot Synthesis Bot. *Adv. Mater.* **32**, 2001626, (2020).

22 Volk, A. A., Epps, R. W. & Abolhasani, M. Accelerated Development of Colloidal Nanomaterials Enabled by Modular Microfluidic Reactors: Toward Autonomous Robotic Experimentation. *Adv. Mater.* **33**, 2004495, (2021).

23 Abdel-Latif, K., Bateni, F., Crouse, S. & Abolhasani, M. Flow Synthesis of Metal Halide Perovskite Quantum Dots: From Rapid Parameter Space Mapping to AI-Guided Modular Manufacturing. *Matter* **3**, 1053-1086, (2020).

24 Whitacre, J. F. *et al.* An Autonomous Electrochemical Test Stand for Machine Learning Informed Electrolyte Optimization. *J. Electrochem. Soc.* **166**, A4181-A4187, (2019).

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

26 Noack, M. M. *et al.* A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering. *Sci. Rep-UK* **9**, 11809, (2019).

27 Noack, M. M., Doerk, G. S., Li, R., Fukuto, M. & Yager, K. G. Advances in Kriging-Based Autonomous X-Ray Scattering Experiments. *Sci. Rep-UK* **10**, 1325, (2020).

28 Sethian, J. A. Autonomous Data Acquisition for Large Scale Facilities. *Nature Review Physics*, (under review).

29 Cho, S.-Y. *et al.* Finding Hidden Signals in Chemical Sensors Using Deep Learning. *Anal. Chem.* **92**, 6529-6537, (2020).

30 Fagnan, K. *et al.* Data and Models: A Framework for Advancing AI in Science. *Report of the Office of Science Roundtable on Data for AI*, (2019).

31 Takens, F. in *Dynamical Systems and Turbulence, Warwick 1980.* (eds Rand, D. & Young, L. S.) 366-381 (Springer Berlin Heidelberg).

32 Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a Time Series. *Phys. Rev. Lett.* **45**, 712-716, (1980).

33 Fung, R. *et al.* Achieving accurate estimates of fetal gestational age and personalised predictions of fetal growth based on data from an international prospective cohort study: a population-based machine learning study. *The Lancet Digital Health* **2**, e368-e375, (2020).

34 Jia, X. *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251-255, (2019).

35 Krska, S. W., DiRocco, D. A., Dreher, S. D. & Shevlin, M. The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Accounts Chem. Res.* **50**, 2976-2985, (2017).

36 Dybowski, R. Interpretable machine learning as a tool for scientific discovery in chemistry. *New J. Chem.* **44**, 20914-20920, (2020).

37 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* **334**, 1114, (2011).

38 Buitrago Santanilla, A. *et al.* Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49, (2015).

39 Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236, (2018).

40 Selekman, J. A. *et al.* High-Throughput Automation in Chemical Process Development. *Annu. Rev. Chem. Biomol.* **8**, 525-547, (2017).

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

42 Sader, J. K. & Wulff, J. E. Reinvestigation of a robotically revealed reaction. *Nature* **570**, E54-E59, (2019).

43 Milo, A., Neel, A. J., Toste, F. D. & Sigman, M. S. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737, (2015).

44 Melodie, C. *et al.* Data-science driven autonomous process optimization. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.13146404.v2 (2020).

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

46 Kusne, A. G. *et al.* On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep-UK* **4**, 6367, (2014).

47 Kusne, A. G. *et al.* On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966, (2020).

48 Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Soc. Networks* **43**, 73-85, (2015).

49 Gates-Rector, S. & Blanton, T. The Powder Diffraction File: a quality materials characterization database. *Powder Diffr.* **34**, 352-360, (2019).

50 *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*, https://webbook.nist.gov/chemistry/ (2018).

51 RSCB, Protein Data Bank, https://www.rcsb.org/

52 Sullivan, K. P., Brennan-Tonetta, P. & Marxen, L. J. Economic Impacts of the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank. (Rutgers Office of Research Analytics, 2017).

53 Alshahrani, M. *et al.* Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* **33**, 2723-2730, (2017).

54 Choudhury, R., Aykol, M., Gratzl, S., Montoya, J. & Hummelshøj, J. S. MaterialNet: A web-based graph explorer for materials science data. *J. Open Source Software* **5**, 2105, (2020).

55 Aykol, M. *et al.* Network analysis of synthesizable materials discovery. *Nat. Commun.* **10**, 2018, (2019).

56 Pendleton, I. M. *et al.* Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* **9**, 846-859, (2019).

57 Li, Z. *et al.* Robot-Accelerated Perovskite Investigation and Discovery. *Chem. Mater.* **32**, 5650-5663, (2020).

58 Ratner, D. *et al.* BES Roundtable on Producing and Managing Large Scientific Data with Artificial Intelligence and Machine Learning. (2019).

59 Kwon, H.-K., Gopal, C. B., Kirschner, J., Caicedo, S. & Storey, B. D. A user-centered approach to designing an experimental laboratory data platform. *arXiv:2007.14443*, (2020).

60 Mrdjenovich, D. *et al.* propnet: A Knowledge Graph for Materials Science. *Matter* **2**, 464-480, (2020).

61 Carbone, M. R., Yoo, S., Topsakal, M. & Lu, D. Classification of local chemical environments from x-ray absorption spectra using supervised machine learning. *Phys. Rev. Mater.* **3**, 033604, (2019).

62 Zheng, C., Chen, C., Chen, Y. & Ong, S. P. Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure. *Patterns* **1**, 100013, (2020).

63 Torrisi, S. B. *et al.* Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comp, Mater.* **6**, 109, (2020).

64 Carbone, M. R., Topsakal, M., Lu, D. & Yoo, S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys. Rev. Lett.* **124**, 156401, (2020).

65 Cibin, G. *et al.* An open access, integrated XAS data repository at Diamond Light Source. *Radiat. Phys. Chem.* **175**, 108479, (2020).

66 Ayyer, K. *et al.* Low-signal limit of X-ray single particle diffractive imaging. *Opt. Express* **27**, 37816-37833, (2019).

67 Kayser, Y. *et al.* Core-level nonlinear spectroscopy triggered by stochastic X-ray pulses. *Nat. Commun.* **10**, 4761, (2019).

68 Young, I. D. *et al.* Structure of photosystem II and substrate binding at room temperature. *Nature* **540**, 453-457, (2016).

69 Ratner, D., Cryan, J. P., Lane, T. J., Li, S. & Stupakov, G. Pump-Probe Ghost Imaging with SASE FELs. *Phys. Rev. X* **9**, 011045, (2019).