EXPERT
RECOMMENDATION



# The case for data science in experimental chemistry: examples and recommendations

*Junko Yano*[1]✉, *Kelly J. Gaffney*[2,3]✉, *John Gregoire*[4]✉, *Linda Hung*[5]✉, *Abbas Ourmazd*[6]✉, *Joshua Schrier*[7]✉, *James A. Sethian*[8,9]✉ *and Francesca M. Toma*[10]✉

Abstract | The physical sciences community is increasingly taking advantage of the possibilities offered by modern data science to solve problems in experimental chemistry and potentially to change the way we design, conduct and understand results from experiments. Successfully exploiting these opportunities involves considerable challenges. In this Expert Recommendation, we focus on experimental co-design and its importance to experimental chemistry. We provide examples of how data science is changing the way we conduct experiments, and we outline opportunities for further integration of data science and experimental chemistry to advance these fields. Our recommendations include 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; and combining diverse and disparate data sources into a data network for chemistry research.

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[4,5] and tutorial guides[6,7]. With some exceptions[8], 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[9–12]. For example, combining such data analytical methods with automation or laboratory robotics could enable quasi-autonomous research with minimal human input[13,14]. Improved data analytics and data sharing and reuse in experimental chemistry offer the opportunity to increase the rate and lower the cost of scientific discovery, fostering growth in 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 approaches to their assembly into experimental workflows (defined as a sequence of physical tasks coupled to the analysis of results), the number of possible workflows and methods for designing experiments far exceeds those realized by human researchers so far. Data science methods are poised to aid workflow design and the active steering of experiments to broaden the reach of experimental chemistry and to increase the rate and efficacy with which chemists explore the often daunting parameter spaces of experiments and syntheses. Capitalizing on these opportunities will require fundamental advances in both chemistry and data science, as well as changes in how we conduct experiments, especially the development of technologies to facilitate large-scale data collection, sharing and analysis. At the same time, validating the outcomes of data-science-based interpretation and prediction will be essential.

In this Expert Recommendation, we include key highlights from 'At the tipping point: a future of fused chemical and data science', a workshop held in September 2020, and sponsored by the Council on Chemical Sciences, Geosciences, and Biosciences (CSGB) Division, Office of Basic Energy Sciences, Office of Science, US Department of Energy. Participants from academia, industry and national laboratories assessed opportunities and key research needs for the use of data science in new experimental approaches in chemistry and biochemistry, at experimental scales ranging from single-PI laboratories to large user facilities. With a focus on experimental chemistry, we discuss how data science is changing the way we conduct experiments, using

✉e-mail: 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

https://doi.org/10.1038/s41570-022-00382-w

case studies to highlight important developments, and summarize what is required to take advantage of the advances in both fields.

## 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 the development of hypotheses and theories. Theories codify understanding, offer predictions that can often enable extrapolation 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 considerably increased data acquisition rates and widened the range of tools available to classify, analyse and interpret data. In some experiments, the acquisition of many types of experimental data is no longer 'expensive', and vast amounts can easily be accumulated. One example is high-throughput data collection at synchrotrons. Investments at these large facilities have reduced the experimental cost for single investigators and increased the size of data. 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 myriad possibilities to combine advances in algorithms, hardware and high-throughput data acquisition modalities. Further advances in the chemical sciences will require the systematic exploitation and development of these efforts, augmenting the traditional theoretical approaches to selectively guide 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 provide insight into processes and mechanisms, without requiring accompanying theories and models. Analysing data without a theory-based roadmap is key 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 emerged before theories and models. Data science embraces the importance of classification and the 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 increasing quantities of data and accelerating data acquisition rates.

The need for advanced techniques that are able to interpret and categorize data is an increasingly crucial 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, and ML and 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. Two of the major areas requiring new AI and ML algorithms are, first, techniques to analyse and steer experiments as data are produced, and, second, post-processing of ever-larger datasets. In the first area, it might not be possible to conduct formal mathematical reconstructions and analyses fast enough under vastly increased spatial and temporal resolution, generated at faster and faster rates. In such cases, algorithms augmented by AI and ML will be needed to sort quickly through results to determine whether an experiment is headed in the right direction. In the second area, it might be possible to extract more understanding from collected data than previously thought, and this understanding, which may be buried in the data, could be revealed with these new techniques.

Whether data science interpretations will become an incremental step towards traditional model-based scientific understanding or will ultimately stand on an equal footing with (and, in some arenas, surpass) model-based understanding remains unclear. Even in the absence of a data science revolution, data science will cause the ways we generate and interpret scientific data to evolve. The challenge is to have a reliable way of determining whether one has enough experiments, enough data or enough observations to justify making predictions with quantified uncertainty. Although there is no single route to estimating the uncertainty (error) in the outcome of AI and ML approaches, 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. Such estimates assess, in essence, the interpolation error. The assessment of predictions outside the training range entails additional complexities. Ultimately, 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 a preconceived notion of what experiments to perform, what variables to observe and what weights to put on gathered information. Of course, AI and ML algorithms rely on hidden assumptions and biases, including 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

**Author addresses**

[1]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[2]SLAC National Accelerator Laboratory, Menlo Park, CA, USA.
[3]PULSE Institute, SLAC National Accelerator Laboratory, Stanford University, Stanford, CA, USA.
[4]Division of Engineering and Applied Science, California Institute of Technology, Pasadena, CA, USA.
[5]Accelerated Materials Design and Discovery, Toyota Research Institute, Los Altos, CA, USA.
[6]University of Wisconsin, Milwaukee, WI, USA.
[7]Fordham University, Department of Chemistry, The Bronx, NY, USA.
[8]Department of Mathematics, University of California, Berkeley, CA, USA.
[9]Center for Advanced Mathematics for Energy Research Applications (CAMERA), Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[10]Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Box 1 | **Challenges associated with ML and AI**

Machine learning (ML) has typically been applied to use-cases in which 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 or artificial intelligence (AI) analysis? That is, what level of verification is necessary and to what extent does that compromise the utility of the ML or AI approach?
- What evidence underlies how these methods make predictions? When is it reasonable or necessary to ask this question?
- Can AI and ML be used to predict, with quantifiable confidence, phenomena outside the domain used for constructing the algorithm? Currently, AI and ML approaches 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 possible viable output. Scientific discovery, however, inherently involves investigation of new spaces (extrapolation or prediction), which contrasts with 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 can be applied in other fields that appear to be unrelated. How can one know if and when predictions are transferable between fields?

if these relationships are not revealed in the standard form of analytical models, communicable principles or foundational theories.

Although there are many challenges associated with ML (BOX 1) and no clear path 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 case studies from the chemical sciences that highlight advances and the potential of the interaction between experiments and data science, followed by a discussion of the challenges ahead.

## Data science and chemical sciences

Proponents of ML techniques promise profound advances within chemical sciences in areas such as the extraction of collective coordinates, reaction paths, energy landscapes and dynamics from many heterogeneous observations. Broadly, data science methods are expected to bring at least three important objectives within reach (BOX 2). In the chemical sciences, there have been remarkable steps towards meeting these objectives, and the potential is substantial[8,15–18]. At the same time, there are limitations and pitfalls, and in the following, we give examples from multiple fields. Of course, these objectives fall on a continuum rather than a discrete spectrum of possibilities, but it is helpful to independently address each objective.

### *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.*** Imagine a process by which a set of previously performed experiments is used to suggest what to try next. These suggestions might, 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 suggest experiments to accomplish different or multiple goals. For example, new experiments could be aimed at underexplored parts of the high-dimensional parameter space. These experiments would configure the experimental parameters to examine under-sampled possibilities. The goal is to ensure that a full range of scientific results across the parameter space is efficiently collected. In another example, as experiments are performed and analysed, they could be focused on configurations that yield insight into particularly desirable results.

An important goal is to couple this autonomous steering to advanced simulations and feedback metrics to enable experiments to discover regions in high-dimensional configuration space that have optimal parameters, such as those required to achieve desired results. For further information on autonomous discovery in the chemical sciences, we refer readers to REFS[19,20], as well as to targeted reviews on autonomous materials science[21], organic synthesis planning and optimization[11], medicinal chemistry[12] and formulations[22]. Although autonomous experimentation is often caricatured as removing humans from 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 with human-only or algorithm-only approaches[20].

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

***A pivotal role for ML and AI.*** Advances in ML and AI offer opportunities of achieving 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 ML algorithm can interpret these results in the context of known available data, detecting similarities and patterns that can be used to evaluate the outcome. For example, models trained on crystallographic data can be used to predict crystallographic dimensionality and the space group from thin-film X-ray diffraction patterns[23]. Second, given the analysed output of an experiment, emerging data science techniques can be used to efficiently build surrogate models. Suitably designed, these models can take the analysed output data and quickly estimate results that can be used to steer the experiment.

This ability to automatically evaluate data as it is collected, and then suggest new directions, has applications across experimental science. This approach can be used to query and steer multi-dimensional processes and to inform the placement of sensors and data collection, determining which locations give the newest information. The construction of surrogate models is particularly efficient when information is collected 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 different techniques, such as tomography, mass spectrometry and high-resolution infrared imaging. Such approaches can be used at multiple scales, from the operation of single instruments to collections of instruments in individual laboratories and large-scale facilities. For example, successful demonstrations to date span autonomous benchtop chemical synthesis to the synchrotron experiment discussed in case study 2.

***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 the synthesis and characterization of colloidal nanoparticles coupled to ML-based optimization of the optoelectronic properties[24–26], and computer-controlled test stands for creating and electrochemically characterizing arbitrary liquid electrolyte solutions coupled with online optimization[27,28]. Autonomous optimization of organic synthesis in flow-based reactors has been demonstrated for several systems[29,30], and software has been developed that can autonomously steer commercially available equipment in performing such optimizations[31]. 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 to incorporate relay switches to trigger the chromatographic equipment[32]. A more wide-reaching approach exploits general-purpose robots that interact with existing laboratory equipment[33]: in one configuration, a robot synthesized 688 photocatalysts over 8 days using a Bayesian optimization scheme without human intervention, leading to a six-fold increase in the photocatalytic performance compared with the initial compounds. Even with limitations on how existing knowledge, theory and physical models are implemented in the autonomous search, such examples illustrate the time-efficient and cost-effective use of available resources, shortening a project from months and years to a week. Ideally, future advances in knowledge, theory and models will enable the optimized synthesis of new compounds with targeted properties. However, even the development of autonomous processes for individual analytical subtasks within a research project, such as solubility screening[34] and determining kinetic models by HPLC experiments[35], can be useful both for accelerating research progress and as building blocks for future systems.

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 conform 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 (such as non-Gaussian pixel noise and inaccurate timestamps). Standard approaches to data analysis often successively reject 'outliers' to obtain a sufficiently homogeneous dataset amenable to traditional analysis by averaging.

Machine learning approaches, by contrast, attempt to 'learn' the space spanned by the data, such as 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[8,15,17,18,45]. This offers a noise-robust approach to extracting more information from the data than is possible with traditional methods.

**Optimally design experiments and workflows**

Complex experiments with many input parameters generate sample points in high-dimensional spaces, and 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 to efficiently cover the landscapes in which the system of interest undergoes important, functionally relevant changes[8,17].

**Offer new experimental modalities**

The new generation of high-throughput instruments combined with the algorithmic ability to rapidly analyse very large datasets offers new experimental modalities. As an example, chemical reaction events often occur via rarely sighted transition states. Until now, complex time-resolved experiments have been required in order 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[8]. A 'sufficiently large' dataset of snapshots will thus include high-energy conformations. For example, states at energies comparable with that released by ATP hydrolysis begin to appear in datasets comprising of the order of $10^9$ single-particle snapshots from an equilibrium ensemble of molecules. Such large datasets offer the opportunity to investigate important chemical processes without having to track each process in time. The key is the ability to collect and analyse several billion single-particle snapshots, as dictated by the underlying statistical mechanics.

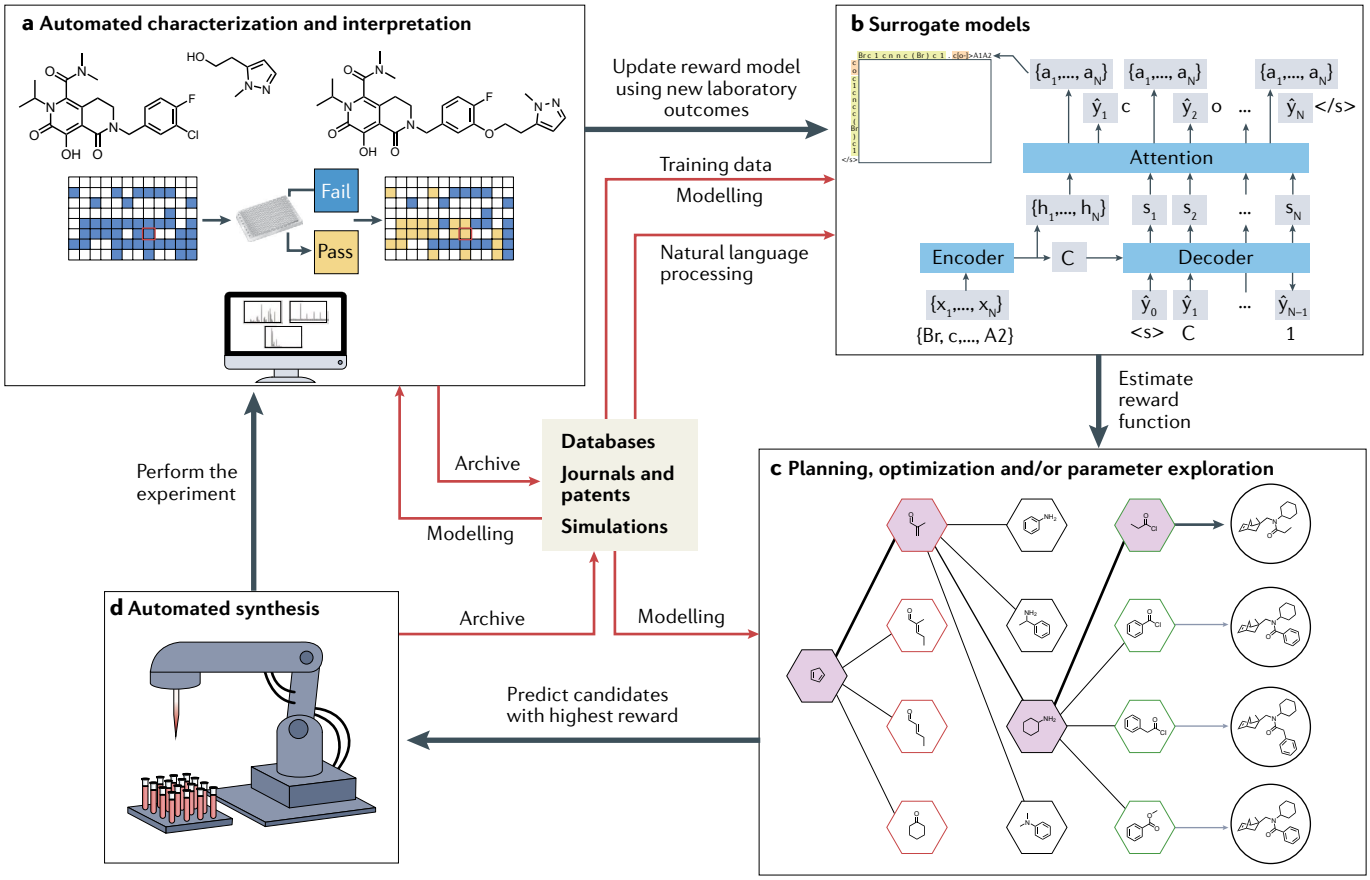


Fig. 1 | **Role of data science in experimental processes.** Data science can have many roles in experimental processes, such as in autonomous synthesis and characterization. **a**–**d** | To accomplish the experimental tasks (grey arrows), several technologies are required, necessitating data flows (red arrows) to and from repositories. Part **a** adapted with permission from REF.[58], AAAS. Part **b** reprinted with permission from REF.[106], Royal Society of Chemistry. Part **c** reprinted from REF.[61], CC BY 4.0.

***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[36–38], which has been used for a wide variety of experiments across the USA and elsewhere (FIG. 2). First, the measurements to take in an autonomous experiment are chosen on the basis of previous measurements. Next, surrogate model functions are computed by ML-based Gaussian process prediction, which can be constrained by domain knowledge. Hybrid optimization methods are then used to identify the next-best measurements to take. Finally, choices for the optimal measurements are determined as a function of the surrogate model, its uncertainty and the costs of a measurement. Using the gpCAM approach and software framework, beam utilization was increased at Brookhaven National Laboratory's Center for Functional Nanomaterials and the National Synchrotron Light Source II (USA) from 15% to more than 80%[36–38] with a five-fold decrease in the number of experiments required to obtain the same information as from previous methods. At the Berkeley Synchrotron Infrared Structural Biology beamline at the Lawrence Berkeley Laboratory's Advanced Light Source (USA), the required amount of biological spectroscopic data that needed to be collected was reduced by as much as 50-fold[38]. At neutron sources at the Institut Laue–Langevin (France), experiment durations have been reduced from days to one night[38].

### *Harnessing complexity with data science*

One well-travelled road in chemical experimental science is the optimization of 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. The analysis process constrains experimental methods to maximize control and homogeneity and to minimize noise, fluctuations and heterogeneity.

The scientific usefulness of the above framework derives directly from how successfully the critical properties of an experiment can be controlled. Although this traditional approach has generated many impressive successes, the inevitable limitations in sample and experimental control present considerable limitations to experimental design. Data science approaches can augment and expand the scope of experimental science both by accelerating the analysis and interpretation of experiments and by enabling experiments to be

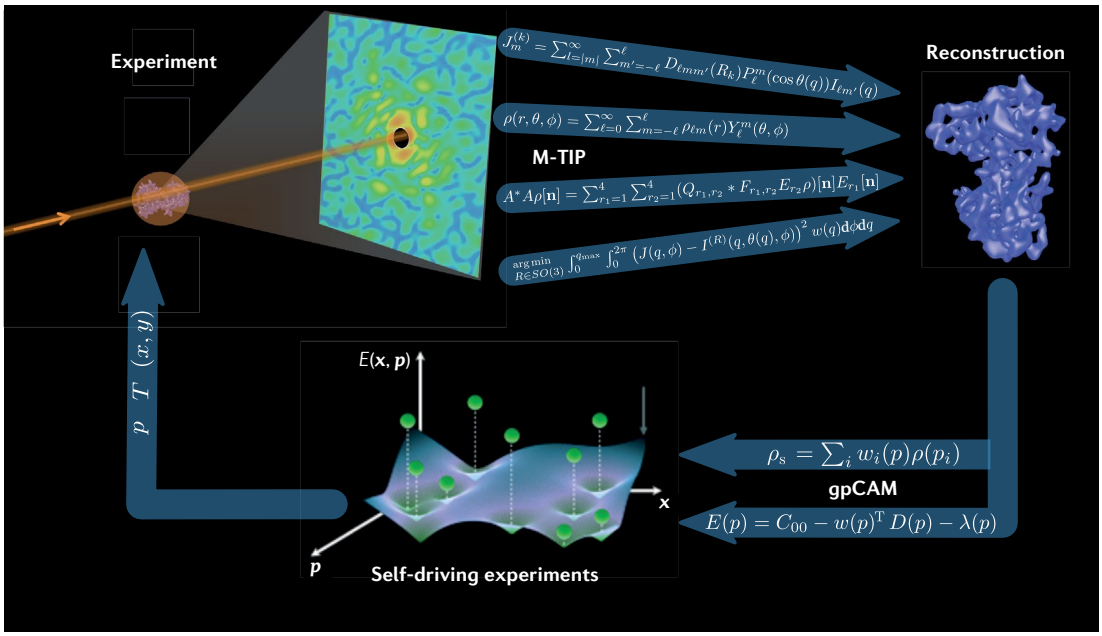


Fig. 2 | **Artificial intelligence and machine learning deployed to accelerate, autonomously control and understand experiments, using state-of-the-art mathematics coupled to advances in data science.** The Center for Advanced Mathematics for Energy Research Applications (CAMERA) has developed multi-tiered iterative projections (M-TIP) to accurately interpret scattering images from light-source experimental data, and Gaussian processes (gpCAM) to suggest and drive new experiments. Working together in an autonomous loop, they optimize the use of complex equipment. Image courtesy of J. Donatelli, M. Noack and J. A. Sethian.

performed successfully when control is impractical or risks undesirable alteration of the phenomena under study. For example, current data science techniques applied to structure and image reconstructions can extract information from measurements recorded with more noise and uncertainty than has previously been possible, greatly increasing the set of 'viable' and productive experiments.

Clear cases in which a data science approach would be valuable include, but are not limited to, experiments that use stochastic or noisy instrumentation, such as X-ray free-electron lasers (XFELs; see case study 4 below), and field studies in which 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 otherwise 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 that 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 approach also requires changes in analysis, because the absence of control can generate much larger datasets with more complex correlations between the chemical properties of the sample being measured in the experiment and the instrument sampling of the parameter space being measured with diagnostics. As an example, such approaches have been designed to mitigate the shot-to-shot variation of XFELs (FIG. 3), which affect the outcome of X-ray diffraction (FIG. 3b,d,e) and X-ray spectroscopy (FIG. 3c) measurements. Such an approach might be a product of necessity for instruments with delicate stability regimes, but it also presents the opportunity to identify unexpected correlations, because natural fluctuations in the experimental apparatus might generate experimental results that 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 needed to acquire a signal and to reduce problems of data redundancy. Furthermore, instead of relying on a single high signal-to-noise output, alternative approaches might rely on many more weak (but easy to collect) signals to make chemical measurements[39]. The integration of fast ML-enabled and AI-enabled analysis can lead to data-driven autonomous experimental workflows.

***What is needed.*** The advances required to capitalize on the above possibilities include the development and implementation of diagnostics for measuring experimental parameters that would have traditionally been controlled; integration of these diagnostic signals with the experimental observables; and fast analysis techniques for correlating the data from the diagnostics with the experiment to enable real-time assessment of experimental progress and on-the-fly construction of representative surrogate models as data are collected.

***A pivotal role for ML and AI.*** ML and AI techniques offer powerful opportunities to identify unexpected or hidden correlations revealed by the fluctuations in the experimental parameter space. Two examples are illustrated in case studies 3 and 4.

***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

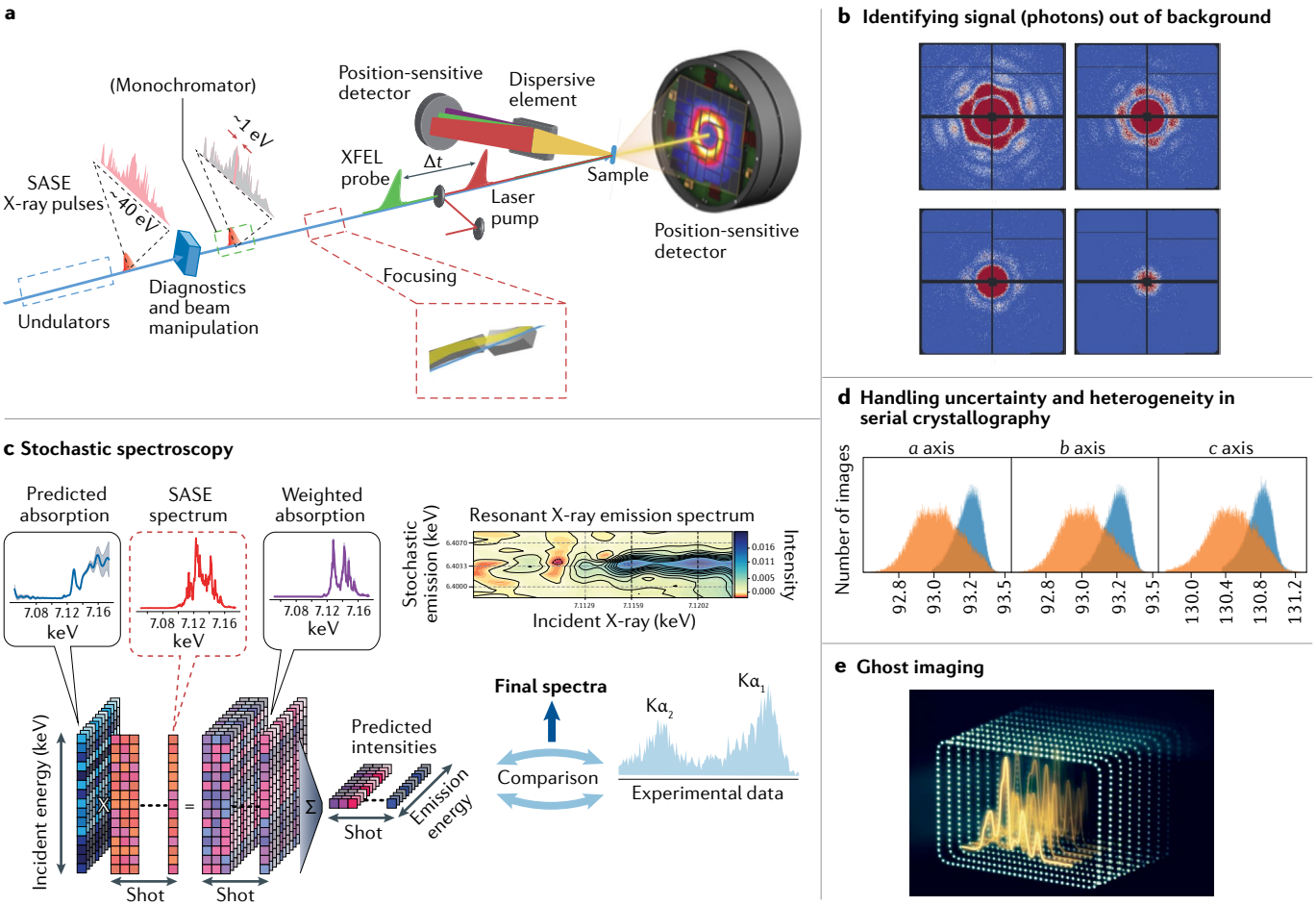


Fig. 3 | **Application of machine learning to conduct new types of experiments at XFEL facilities. a** | Scheme showing an X-ray free-electron laser (XFEL) and the general experimental design for spectroscopy and diffraction and/or scattering experiments[107]. Coherent X-rays are generated using relativistic electrons from a linear accelerator propagating through an undulator. At saturation, the X-ray pulses emitted from this self-amplified spontaneous emission (SASE) process have a relative bandwidth of ~0.2% (for example, ~40 eV at 9 keV of X-ray energy), with a pulse duration of a few to several tens of femtoseconds. $\Delta t$ is the time interval between the pump and the probe. **b** | Detection of a small number of photons (a weak signal) from a large instrument background in a single snapshot for imaging[108]. This ability is crucial for enabling X-ray single-particle imaging of sub-10-nm-sized biomolecules, for which very few photons from particle scattering are expected to be measured in a single snapshot, in comparison to the large instrument background signals. Machine learning (ML) methods are required to extract the weak signal from particles. **c** | Fe 1s2*p* resonant inelastic X-ray spectroscopy (RIXS) that uses the stochastic nature of the polychromatic XFEL SASE beam, taking advantage of the random spikes of each XFEL pulse as a unique fingerprint. These fingerprints are correlated with outgoing emission signals from the system under study to construct spectra[42]. In principle, a monochromator can be used to determine accurately (with an ~1-eV bandwidth) the energy of the incoming X-rays that interact with the sample, but at the cost of a loss of two orders of magnitude in the X-ray flux. Alternatively, measuring the spectrum on every shot and using ML approaches to correlate the experimental signal with these spectra on every shot enables the spectral fluctuations to be disentangled from the experimental observable without the need for an X-ray monochromator. **d** | An example of the heterogeneity in the unit cell distribution of thermolysin crystals, showing that there are different crystal isoforms[109]. This online analysis of data is used to provide immediate feedback to determine the subsequent sample preparation conditions for the best resolution[110]. **e** | Pump–probe ghost imaging. Similar to the method shown in part **c**, this approach uses the random spikes of XFEL pulses to study the interaction of the pulses with matter. This method can be used to map the full evolution of a system over time[111]. Part **a** adapted from REF.[107], Springer Nature Limited. Part **b** reprinted from REF.[108], CC-BY.4.0. Part **c** adapted from REF.[42], CC BY 4.0. Part **d** reprinted with permission from REF.[109], Computational Crystallography Newsletter. Part **e**, image courtesy of Greg Stewart, SLAC National Accelerator Laboratory.

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 the laboratory conditions associated with each experiment over an extended period of time[40]. Using a dataset of 8,470 experiments captured over a 20-month period, it was possible to identify statistical anomalies in the 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 have transformed X-ray science by producing extremely bright X-ray beams. The lasing process that generates these beams also leads to much larger fluctuations in key experimental parameters, particularly compared with synchrotron-based X-ray sources. Attempts to control key beam properties, such as pulse spectra, intensity and duration, have so far only been partially successful. 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 on the observed experimental signal[41,42]. 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 100-fold improvement. Furthermore, such an approach has the benefit of improving the 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 that 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 parallel experiments to be conducted using 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 modelling experts to assure stakeholders that algorithms are able to perform this task robustly and reproducibly[43].

XFELs have also advanced the application of X-ray science to 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, for example, REF.[44]). This quest has made it imperative to deal with noisy, incomplete and fleeting signals recorded with substantial timing uncertainty. Although experimental attempts to deal with such signals will continue to advance, recent AI and ML approaches have brought the greatest rewards (see, for example, REFS[18,45]).

***Case study 5: theory for dynamics in chemistry.*** The measurement of dynamics is an important case in point, wherein AI and ML techniques can help alleviate long-standing experimental problems. Since the celebrated work of Takens[46] and Packard[47], 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 had been thought necessary for proper experimental analysis[48]. Takens showed that a series of snapshots, each representing a subset of the system variables, is sufficient to determine the behaviour of dynamical systems, as if 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[49], indicating the generality of the algorithmic methods.

Another example includes a recent breakthrough, namely the deep-learning package DeePMD-kit[50], which combines ab initio modelling, high-performance computing and ML to tackle 'first-principles' molecular dynamics simulations by approximating ab initio data with deep neural networks. This approach allows for calculations that can treat larger systems over longer timescales and offers a bridge between ML and physical modelling. Similar types of combinations of ab initio results with data science methods and autonomous experimentation have been used to accelerate chemical optimization tasks[51]. Building on established simulation methods and relating this to experimental chemistry data will increase the interpretability of data science models and enable their deployment with smaller datasets.

### *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 beaten path. Hypothesis-driven research, which is generally derived from prior knowledge and relies on testing a postulated outcome, might restrict inquiry and exploration[52].

***Goal: automated serendipity.*** In the absence of a hypothesis, trial and error becomes intractable as the search space increases. Efforts in laboratory 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 on the

laboratory automation efforts discussed above). More broadly, data science approaches can be used to automate the process of extracting new 'ideas' to try on the basis of collected datasets[10,23,40]. Comprehensive data management (discussed below) facilitates the process of identifying unexpected variations that can suggest directions for more deliberate inquiry. For this type of application, prediction accuracy is less crucial because it suffices to be wrong less often than an undirected search, 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 efficient and unbiased search and analysis tools, or at least tools with a bias that is clearly delineated and transparent; implementation of ML methods to identify unexpected or hidden correlations revealed by the fluctuations in the experimental parameter space; and an 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 favours performing larger-scale combinatorial experiments to explore a broader and less biased search space. A short-term goal is simply to perform more experiments over the broadest possible search space, which is the goal of 'classical' high-throughput experimentation or combinatorial chemistry[53]. More long-term goals use ML and AI to accelerate the characterization process and to optimize selection of new experiments. Finally, there is a need for ML interpretability and explainable AI (XAI) to inform humans; this may necessitate chemistry-specific interpretable ML methods[54]. Some early realizations of this approach in experimental chemistry include the extraction of hypotheses about organic molecular-structure determinants of energy levels and solubility[55] and human–algorithm teaming for the synthesis of polyoxometalates[56].

***Case study 6: serendipity-driven reaction discovery.*** This type of non-selective 'automated serendipity' has been successful in the discovery of organic reactions for photoredox catalysed C–H arylation[57] and Pd-catalysed C–N cross coupling[58,59]. For a general review of high-throughput automation in chemistry, see REF.[60]. These applications have relied on experimental hardware developments to perform synthesis and characterization with greater parallelism and smaller quantities of reagents. 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 information to prioritize subsequent experimental rounds[61], with the understanding that this can provide only a preliminary investigation and that subsequent human reinvestigations may be necessary to confirm the spectral interpretations[62]. A more sophisticated approach would use this data to construct empirical relationships between the catalyst and substrate structures and the catalytic efficiency[63]; the resulting structure–property models can then serve to prioritize subsequent experimentation. Finally, a higher-level goal is to perform autonomous optimization of the design of the catalyst, substrate and reaction conditions using automated experimentation and planning algorithms[64]. In materials science, there has been a similar progression from high-throughput synthesis and characterization, to increasing automated interpretation and autonomy[13]; again, this progress is being enabled by increased adoption of ML methods throughout the discovery lifecycle[7,65].

Data science approaches can help to facilitate this serendipitous discovery process by reducing the need to know what one is looking for ahead of time. An example is the development of rare-earth-free permanent magnets[66]. A wide variety of Fe–Co–X (where X is a transition metal element such as Mo, W, Ta, Zr, Hf or V) alloys were synthesized combinatorially, resulting (in some cases) in one or more phases, many of which were unknown. Using non-negative matrix factorization methods, diffraction spectra were decomposed into estimates of the pure material spectra (which had either not been previously observed or could be matched against known databases) and estimates of the relative concentrations of the different phases. The goal is to produce a phase diagram of different compositions; however, building a complete map over the compositions would require too much instrument time. Instead, a further improvement used active-learning approaches and Bayesian optimization methods to prioritize the (automated) acquisition of new experimental data points[67]. Reducing the number of diffraction measurements that must be acquired by several orders of magnitude decreases the amount of beamtime required or even enables the use of the types of diffractometers found in a typical single-PI laboratory instead of a beamline source.

### *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 that provide 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 to shed light on these relationships[68,69]. Data networks leverage the scale and variety of modern chemistry data to enhance the utility of data-driven methods in chemistry experiments. In this section, 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 the 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

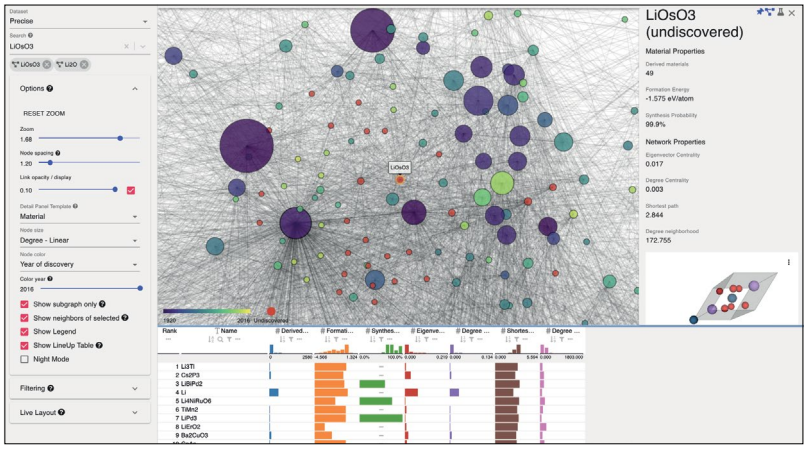


Fig. 4 | **Visualizing a data network.** MaterialNet is a web-based application that can be used to visually explore the relationships within a materials database such as the materials similarity network. The materials are represented by nodes and the links between them encode specific relationships, such as chemical similarity, demonstrating the many interrelationships that exist between materials and chemicals. Networks can capture more relationships than a human can comprehend, and data science tools can learn from these relationships[81]. Reprinted from REF.[81], CC BY-4.0.

repositories and to obtain faster access to details essential to research projects (as a complement to traditional literature searches).

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[70], the CRC Handbook of Chemistry and Physics, and the National Institute of Standards and Technology (NIST) Chemistry WebBook[71]. In one sense, these repositories contain highly refined chemical knowledge. As an example, consider the trajectory of experimental data from the acquisition of raw data to contextualization, analysis, interpretation and validation through additional experiments. Repositories have understandably focused on only the final outcome of this data funnel. Instead, managing and cataloguing 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. One especially successful example is the Protein Data Bank (PDB), which is an archive of experimentally determined protein structures (and is highlighted in case study 7 below)[72]. Similar types of machine-readable experimental chemistry databases would be a watershed in the incorporation of data science. To date, the most successful repositories of experimental chemistry data are structural databases, such as the Cambridge Structural Database (CSD) and the Inorganic Crystal Structure Database (ICSD), and spectral databases, such as NMRShiftDB[73]. An IUPAC project, 'Development of a standard for FAIR data management of spectroscopic data' (FAIRSpec) was founded in 2019 (REF.[74]), and progress is described in a recent report[75].

Most databases of organic synthesis (such as Reaxys) are proprietary and do not allow free contribution, although nascent efforts such as the Open Reaction Database[76] and Chemotion Repository have started to address this need, facilitated by integration with open-source electronic laboratory notebook (ELN) software such as ChemotionELN[77]. In parallel, there are several efforts aimed at developing schema for representing laboratory actions, such as XDL[78], IBM RXN[79], Autoprotocol and the ESCALATE (Experiment Specification, Capture and Laboratory Automation Technology) materials and action specification[80]. The advent of the US Department of Energy Office of Sciences' PuRe Data Resources embodies an important step in this direction.

Once data networks are available, they can be used to accelerate the generation and testing of hypotheses through AI-driven encapsulation of existing knowledge. For example, a network based on data from high-throughput density functional theory calculations can be explored by humans through web-based visualizations[81] using MaterialNet (FIG. 4), while its network metrics can be used in a ML model to predict (or hypothesize) the synthesizability of new inorganic compounds[82]. This mode of hypothesis testing, which builds on the concepts discussed above in the section on data-driven experimental discovery, is markedly 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 above. Although a single laboratory may possess insufficient data for training surrogate models, data networks might contain auxiliary data to augment the laboratory's data. Of course, this has its own challenges: training models to use data from multiple sources is non-trivial, and developing techniques for using and linking heterogeneous data 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 above. True extrapolations may be wildly inaccurate, but data networks can be constructed with the appropriate connections so that applications to new compounds and conditions will more often lie comfortably within the domain of validity of existing models. For example, a property of a given chemical may not have been measured by a certain technique, but previous experiments that share the same property, chemical or method may be used to infer missing values; 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 various cultural and technical advances. Many of the relationships between chemical experiments lie in their metadata, which include details of the

instruments and their settings, and other knowledge required to reproduce the data. Agreeing on software formats for recording experimental parameters, as opposed to manually setting multiple knobs whose data record is limited to written notes, will greatly facilitate consistent tracking of experimental metadata. Data management programmes such as ESCALATE[80] and Event-Sourced Architecture for Materials Provenance Management and Application to Accelerated Materials Discovery (ESAMP)[83] are examples of this approach to chemistry and materials data stewardship by making the data and metadata inseparable; examples of the rich types of interactive experiment reporting that this approach enables can be found in the supporting information of REF.[84].

The chemical and analysis provenance of data is also crucial. From laboratory notes to publications, chemicals are often labelled according to 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 best 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, and motivates a re-thinking of how experimental data should be recorded and tracked.

There are complementary challenges for data processing and interpretation. Expert decisions during data analysis, such as identifying which portion of a spectrum to analyse or what data artefacts might be present, are informed by experience-based knowledge. Tracking the provenance of data analysis will facilitate the removal of human bias and uncover valuable information from raw data. By contrast, the 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, AI 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, which are compounded by the imperfections of the data itself as well as hurdles introduced by less technical considerations such as intellectual property and incentivization schemes. We refer readers to a 2019 US Department of Energy report for recommendations on technical aspects of the data pipeline and network[85] and REF.[86] 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 into data management to learn the relevance of metadata provenance and domain knowledge so that they can be appropriately modelled in data networks. Networking data should commence with models of relationships encoded in existing theories, as was recently demonstrated by propnet (a knowledge graph for materials properties)[87], which is built on equations in which the variables are physical properties of materials. Using this network concept to express interrelationships of experimental data is a new paradigm in data management for chemical sciences.

***Case study 7: structural chemistry.*** Deposition of 3D structural data into the Protein Databank (PDB) is a requirement for the publication of protein structures, resulting in research data estimated to be worth at least US$12 billion having been contributed to the database over the past 40 years. Moreover, this central repository results in an increase in research productivity worth US$2.5 billion annually (as of 2017)[88]. The accumulated data within the PDB has enabled the development of the recent AlphaFold[89] and RoseTTAFold[90] models for predicting the 3D structure of proteins solely on the basis of their amino acid sequences. 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, for example, new biological features and drug properties[91].

***Case study 8: X-ray absorption spectroscopy.*** X-ray absorption spectroscopy (XAS) is a widely used experimental technique for characterizing the geometric and electronic structure of materials. Data science methods could facilitate faster data analysis and recognition of chemical features in measured elemental patterns. ML models have been developed that can predict chemical features from XAS patterns collected under the same conditions as those of a relatively large training set, which has been demonstrated using computed XAS spectra[92–95]. Expanding the scope of these models to experimental spectra could be enabled by aggregating XAS data from dozens of beamlines worldwide that have collectively acquired many thousands or perhaps millions of spectra to date. However, variants of the technique rapidly complicate the problem, ranging from fluorescence to electron detection modes, and from hard X-ray-open-atmosphere to soft X-ray-vacuum, as well as various in situ and operando measurements of chemicals or 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 difficult and must begin with well tracked and machine-readable metadata. Nevertheless, recent progress has been made on this front[96], which is a key step on the path to an XAS data network.

## Recommendations

ML and AI are rapidly changing the meaning of experimental knowledge. They provide rich information and analytical tools that should be part of every scientist's toolbox. Of course, ML techniques must be used with care: some of the challenges include trying to understand whether an algorithm trained on one dataset can be used to produce reliable answers about a different dataset; whether a particular algorithm is robust to noise or attempts to deceive it; what are the reasons for the answers an algorithm provides; and whether these answers are free of bias.

Nonetheless, even with these challenges, there are tremendous opportunities for ML and AI to transform

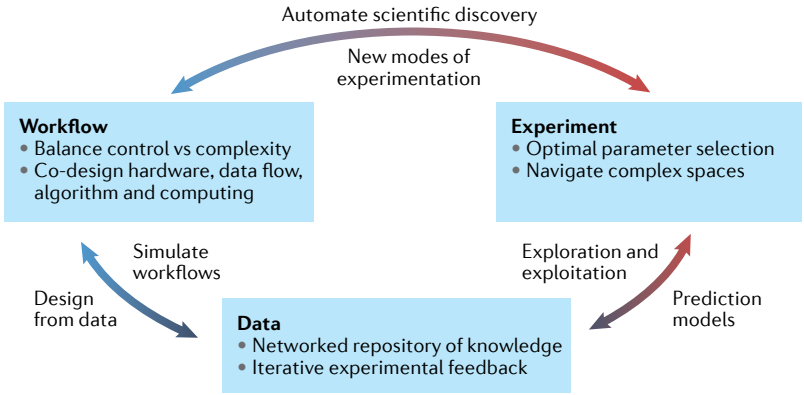


Fig. 5 | **Interplay of experiments, workflow and data.** Experiments are performed in a workflow with decisions based on prior data, producing new data that characterize the experiment and workflow as well as the materials and chemicals under investigation. We illustrate modes of interactions for accelerating and amplifying scientific discovery, which requires the active engagement of data scientists, experimentalists and theoreticians.

experimental chemistry. Capitalizing on these opportunities will require active engagement of both the data science and chemistry communities in using existing tools, injecting domain-specific knowledge into their design, and customizing and targeting these techniques (FIG. 5). In the following, we outline several recommendations that we hope will contribute to facilitating this engagement.

***Develop data science methods for chemistry***

Chemists are increasingly incorporating data science techniques into their research, and many early applications that used off-the-shelf methods have achieved notable advances. To move forward, it is crucial to understand the limitations of existing algorithms for chemical datasets and to 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. Ways to incorporate known physical laws might include representations that incorporate the appropriate symmetry behaviour of structures and physical interactions[97] (such as invariance and equivariance[98] and isometry[99]) and the periodic trends in the properties of elements[100]. Such methods can form the basis for new modes of experiment, including the relaxation of experimental control to enable the acquisition of larger information throughput.

***Recommendation.*** Develop new ML and AI representations and techniques specific to chemistry by partnering with chemical and data scientists and train a complementary workforce of interdisciplinary experts that can leverage the methods in experimental design and analysis. This training could take various forms, including the incorporation of ML into existing undergraduate laboratory experiments[101,102], workshops and bootcamps[103]; dedicated courses for ML and AI in chemistry[104]; and graduate specializations. Datasets and software reproducibility are important. Journals should strongly encourage or require that data and software are deposited in repositories that adhere to FAIR (Findable, Accessible, Interoperable and Reusable) principles. Peer review of data and code may be necessary in addition to traditional content reviewing. Data-centred journals, such as *Scientific Data, Data in Brief* and *Chemical Data Collections*, will also have an important role in disseminating citable datasets that have been created independently of typical hypothesis-driven research efforts.

***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 and known solutions. With datasets that are large enough, this interpolative approach is often sufficient for solving many scientific problems. We note that purely interpolative methods typically fail when one needs to extend predictions into new and unexplored regions of the parameter space or when dramatic changes occur between sparse elements. Nevertheless, even within the scope of pure interpolation, the power of data science can be used to direct future research outside the bounds of current measurements and observations[105]. As illustrated above, research in this direction can potentially be applied to accelerate discovery.

***Recommendation.*** Develop ML methods that work with sparse representations in high-dimensional parameter spaces, to provide 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 computational resources***

Data science methods can accelerate decision making. To exploit this possibility, we need integrated laboratory automation systems that enable algorithms and workflows to enact processes in the laboratory, monitor the results and deposit the resulting data into shared repositories. Accelerating the experimental cycle is especially valuable in shared facilities (such as synchrotrons) but is equally needed in single-PI laboratories. Together, these integrated systems have the potential to establish a virtuous cycle — experiments conducted by automated systems or robots that are 'born digital', which reduces barriers to data sharing and reuse, and facilitates the development of better data science methods — but there are considerable technical barriers. Open-source hardware should be encouraged, with relevant computer-aided design (CAD) files and control code deposited into appropriate repositories, such as the Open Hardware Repository. Currently, this type of data often appears in supporting information, but could also be the primary topic of articles in journals such as Reviews of Scientific Instruments and HardwareX, which create citable records for equipment development.

***Recommendation.*** Encourage a co-design approach to hardware, software and algorithm development. Interdisciplinary teams can often reimagine the entire range of experimental workflows to embrace an

accelerated approach that integrates measurements, data, algorithms and computing. This approach is enabled by developing both modular and complete solutions, with an emphasis on interoperable and open hardware and software. For successful collaboration, experts in different disciplines need to be connected through strong scientific scope, shared personnel, and frequent communication and group discussions. The groups should thus share a feeling of project ownership.

### *Integrate diverse data sources*

Chemical data are diverse, consisting of spectroscopic observations, structural information, process descriptions, and many other types of measurement. 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. Although specific types of chemical data have been aggregated (such as crystallographic data), there are currently only limited automated mechanisms by which individual experiments that comprise diverse elements can contribute to a broader whole. Human researchers excel at placing a new piece of data in the context of the prior data and knowledge of their field, but their reasoning that underpins these assessments suffers from being slow, costly, biased and inconsistent. AI methods for contextualizing data should be developed, which requires the establishment of a foundation for automatic management of relationships in chemistry data, in order to achieve the goal of a network of data.

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

Published online: 21 April 2022


1. Ourmazd, A. Science in the age of machine learning. *Nat. Rev. Phys.* **2**, 342–343 (2020).
2. National Science Foundation. Framing the Role of Big Data and Modern Data Science in Chemistry. *NSF* https://www.nsf.gov/mps/che/workshops/data_chemistry_workshop_report_03262018.pdf (2018).
3. *Mission Innovation* (Energy Materials Innovation, 2018); http://mission-innovation.net/wp-content/uploads/2018/01/Mission-Innovation-IC6-Report-Materials-Acceleration-Platform-Jan-2018.pdf.
4. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
5. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annu. Rev. Mater. Res.* **50**, 71–103 (2020).
6. Janet, J. P. & Kulik, H. J. *Machine Learning In Chemistry* (American Chemical Society, 2020).
7. Wang, A. Y.-T. et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **32**, 4954–4965 (2020).
8. Dashti, A. et al. Retrieving functional pathways of biomolecules from single-particle snapshots. *Nat. Commun.* **11**, 4734 (2020).
9. Selvaratnam, B. & Koodali, R. T. Machine learning in experimental materials chemistry. *Catal. Today* **371**, 77–84 (2021).
10. Shi, Y., Prieto, P. L., Zepel, T., Grunert, S. & Hein, J. E. Automated experimentation powers data science in chemistry. *Acc. Chem. Res.* **54**, 546–555 (2021).
11. Shen, Y. et al. Automation and computer-assisted planning for chemical synthesis. *Nat. Rev. Meth. Prim.* **1**, 23 (2021).
12. Nichols, P. L. Automated and enabling technologies for medicinal chemistry. *Progr. Med. Chem.* **60**, 191–272 (2021).
13. Stein, H. S. & Gregoire, J. M. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **10**, 9640–9649 (2019).
14. Flores-Leonar, M. M. et al. Materials acceleration platforms: on the way to autonomous experimentation. *Curr. Opin. Green. Sustain. Chem.* **25**, 100370 (2020).
15. Dashti, A. et al. Trajectories of the ribosome as a Brownian nanomachine. *Proc. Natl Acad. Sci. USA* **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. Ed.* **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. Ed.* **59**, 23414–23436 (2020).
21. Stach, E. et al. Autonomous experimentation systems for materials development: a community perspective. *Matter* **4**, 2702–2726 (2021).
22. Cao, L., Russo, D. & Lapkin, A. A. Automated robotic platforms in design and development of formulations. *AIChE J.* **67**, e17248 (2021).
23. Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mat.* **5**, 60 (2019).
24. Epps, R. W. et al. Artificial chemist: an autonomous quantum dot synthesis bot. *Adv. Mater.* **32**, 2001626 (2020).
25. 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).
26. 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).
27. Whitacre, J. F. et al. An autonomous electrochemical test stand for machine learning informed electrolyte optimization. *J. Electrochem. Soc.* **166**, A4181–A4187 (2019).
28. Dave, A. et al. Autonomous discovery of battery electrolytes with robotic experimentation and machine learning. *Cell Rep. Phys. Sci.* **1**, 100264 (2020).
29. Wimmer, E. et al. An autonomous self-optimizing flow machine for the synthesis of pyridine–oxazoline (PyOX) ligands. *React. Chem. Eng.* **4**, 1608–1615 (2019).
30. Cortés-Borda, D. et al. An autonomous self-optimizing flow reactor for the synthesis of natural product carpanone. *J. Org. Chem.* **83**, 14286–14299 (2018).
31. Jeraal, M. I., Sung, S. & Lapkin, A. A. A machine learning-enabled autonomous flow chemistry platform for process optimization of multiple reaction metrics. *Chem. Meth.* **1**, 71–77 (2021).
32. Christensen, M. et al. Data-science driven autonomous process optimization. *Commun. Chem.* **4**, 112 (2021).
33. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
34. Shiri, P. et al. Automated solubility screening platform using computer vision. *iScience* **24**, 102176 (2021).
35. Waldron, C. et al. An autonomous microreactor platform for the rapid identification of kinetic models. *React. Chem. Eng.* **4**, 1623–1636 (2019).
36. Noack, M. M. et al. A kriging-based approach to autonomous experimentation with applications to X-ray scattering. *Sci. Rep.* **9**, 11809 (2019).
37. Noack, M. M., Doerk, G. S., Li, R., Fukuto, M. & Yager, K. G. Advances in kriging-based autonomous X-ray scattering experiments. *Sci. Rep.* **10**, 1325 (2020).
38. Noack, M. M., Zwart, P. H. & Ushizima, D. M. et al. Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities. *Nat. Rev. Phys.* **3**, 685–697 (2021).
39. Cho, S.-Y. et al. Finding hidden signals in chemical sensors using deep learning. *Anal. Chem.* **92**, 6529–6537 (2020).
40. Nega, P. W. et al. Using automated serendipity to discover how trace water promotes and inhibits lead halide perovskite crystal formation. *Appl. Phys. Lett.* **119**, 041903 (2021).
41. Kayser, Y. et al. Core-level nonlinear spectroscopy triggered by stochastic X-ray pulses. *Nat. Commun.* **10**, 4761 (2019).
42. Fuller, F. D. et al. Resonant X-ray emission spectroscopy from broadband stochastic pulses at an X-ray free electron laser. *Commun. Chem.* **4**, 84 (2021).
43. Fagnan, K. et al. *Data and Models: A Framework for Advancing AI in Science* (OSTI, 2019).
44. Domcke, W. & Yarkony, D. R. Role of conical intersections in molecular spectroscopy and photoinduced chemical dynamics. *Annu. Rev. Phys. Chem.* **63**, 325–352 (2012).
45. Hosseinizadeh, A. et al. Single-femtosecond atomic-resolution observation of a protein traversing a conical intersection. *Nature* **599**, 697–701 (2021).
46. Takens, F. in *Dynamical Systems and Turbulence, Warwick 1980* (eds Rand, D. & Young, L.S.) 366–381 (Springer, 1981).
47. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716 (1980).
48. Hosseinizadeh, A. et al. Few-fs resolution of a photoactive protein traversing a conical intersection. *Nature* **599**, 697–701 (2021).
49. 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. *Lancet Dig. Health* **2**, e368–e375 (2020).
50. Jia, W. et al. in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* 1–14 (IEEE, 2020); https://dl.acm.org/doi/abs/10.5555/3433701.3433707.
51. Sun, S. et al. A data fusion approach to optimize compositional stability of halide perovskites. *Matter* **4**, 1305–1322 (2021).
52. Jia, X. et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 (2019).
53. 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. *Acc. Chem. Res.* **50**, 2976–2985 (2017).
54. Dybowski, R. Interpretable machine learning as a tool for scientific discovery in chemistry. *N. J. Chem.* **44**, 20914–20920 (2020).
55. Guan, W. et al. Quantum machine learning in high energy physics. *Mach. Learn. Sci. Technol.* **2**, 011003 (2021).

56. Duros, V. et al. Intuition-enabled machine learning beats the competition when joint human-robot teams perform inorganic chemical experiments. *J. Chem. Inf. Model.* **59**, 2664–2671 (2019).
57. 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).
58. Buitrago Santanilla, A. et al. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
59. Lin, S. et al. Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236 (2018).
60. Selekman, J. A. et al. High-throughput automation in chemical process development. *Annu. Rev. Chem. Biomol.* **8**, 525–547 (2017).
61. 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).
62. Sader, J. K. & Wulff, J. E. Reinvestigation of a robotically revealed reaction. *Nature* **570**, E54–E59 (2019).
63. Milo, A., Neel, A. J., Toste, F. D. & Sigman, M. S. Organic chemistry. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737–743 (2015).
64. Melodie, C. et al. Data-science driven autonomous process optimization. *Comm. Chem.* **4**, 112 (2021).
65. Li, J. et al. AI applications through the whole life cycle of material discovery. *Matter* **3**, 393–432 (2020).
66. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
67. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Commun.* **11**, 5966 (2020).
68. Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: dynamic network models of science's unfolding structure. *Soc. Netw.* **43**, 73–85 (2015).
69. Bai, J. et al. From platform to knowledge graph: evolution of laboratory automation. *J. Am. Chem. Soc. Au* **2**, 292–309 (2022).
70. Gates-Rector, S. & Blanton, T. The Powder Diffraction File: a quality materials characterization database. *Powder Diffr.* **34**, 352–360 (2019).
71. Linstrom, P. J. & Mallard, W. G. (eds) *NIST Chemistry WebBook, NIST Standard Reference Database Number 69* (National Institute of Standards and Technology, 2022).
72. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
73. Kuhn, S. & Schlörer, N. E. Facilitating quality control for spectra assignments of small organic molecules: nmrshiftdb2 — a free in-house NMR database with integrated LIMS for academic service laboratories. *Magn. Reson. Chem.* **53**, 582–589 (2015).
74. Hanson, R. et al. *Development Of A Standard For Fair Data Management Of Spectroscopic Data* (IUPAC, 2020).
75. Hanson, R. M. J. et al. FAIR enough? *Spectrosc. Eur. World* **33**, 25–31 (2021).
76. Kearnes, S. M. et al. The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
77. Tremouilhac, P. et al. Chemotion ELN: an open source electronic lab notebook for chemists in academia. *J. Cheminform.* **9**, 54 (2017).
78. Mehr, S. H. M., Craven, M., Leonov Artem, I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101–108 (2020).
79. Vaucher, A. C. et al. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).
80. 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).
81. Choudhury, R., Aykol, M., Gratzl, S., Montoya, J. & Hummelshøj, J. S. MaterialNet: a web-based graph explorer for materials science data. *J. Opn Src. Softw.* **5**, 2105 (2020).
82. Aykol, M. et al. Network analysis of synthesizable materials discovery. *Nat. Commun.* **10**, 2018 (2019).
83. Statt, M. R. et al. ESAMP: event-sourced architecture for materials provenance management and application to accelerated materials discovery. Preprint at *ChemRxiv* https://doi.org/10.26434/chemrxiv.14583258.v1 (2021).
84. Li, Z. et al. Robot-accelerated perovskite investigation and discovery. *Chem. Mater.* **32**, 5650–5663 (2020).
85. Ratner, D. et al. Office Of Basic Energy Sciences (BES) roundtable on producing and managing large scientific data with artificial intelligence and machine learning. *US DOE OSTI* https://doi.org/10.2172/1630823 (2019).
86. Kwon, H.-K., Gopal, C. B., Kirschner, J., Caicedo, S. & Storey, B. D. A user-centered approach to designing an experimental laboratory data platform. Preprint at *arXiv* https://arxiv.org/abs/2007.14443 (2020).
87. Mrdjenovich, D. et al. Propnet: a knowledge graph for materials science. *Matter* **2**, 464–480 (2020).
88. 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).
89. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
90. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
91. Alshahrani, M. et al. Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* **33**, 2723–2730 (2017).
92. 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).
93. 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).
94. Torrisi, S. B. et al. Random forest machine learning models for interpretable X-ray absorption near-edge structure spectrum-property relationships. *npj Comput. Mater.* **6**, 109 (2020).
95. Carbone, M. R., Topsakal, M., Lu, D. & Yoo, S. Machine-learning X-ray absorption spectra to quantitative accuracy. *Phys. Rev. Lett.* **124**, 156401 (2020).
96. Cibin, G. et al. An open access, integrated XAS data repository at diamond light source. *Radiat. Phys. Chem.* **175**, 108479 (2020).
97. Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 9759–9815 (2021).
98. Smidt, T. E. Euclidean symmetry and equivariance in machine learning. *Trends Chem.* **3**, 82–85 (2021).
99. Ropers, J., Mosca, M. M., Anosova, O., Kurlin, V. & Cooper, A. I. Fast predictions of lattice energies by continuous isometry invariants of crystal structures. Preprint at https://arxiv.org/abs/2108.07233 (2021).
100. Herr, J. E., Koh, K., Yao, K. & Parkhill, J. Compressing physics with an autoencoder: creating an atomic species representation to improve machine learning models in the chemical sciences. *J. Chem. Phys.* **151**, 084103 (2019).
101. Sharma, A. Laboratory glassware identification: supervised machine learning example for science students. *J. Comput. Sci. Ed.* **12**, 8–15 (2021).
102. Thrall, E. S., Lee, S. E., Schrier, J. & Zhao, Y. Machine learning for functional group identification in vibrational spectroscopy: a pedagogical lab for undergraduate chemistry students. *J. Chem. Educ.* **98**, 3269–3276 (2021).
103. Lafuente, D. et al. A gentle introduction to machine learning for chemists: an undergraduate workshop using python notebooks for visualization, data processing, analysis, modeling. *J. Chem. Ed.* **98**, 2892–2898 (2021).
104. Gressling, T. *Data Science in Chemistry: Artificial Intelligence, Big Data, Chemometrics and Quantum Computing with Jupyter* (Walter de Gruyter, 2020).
105. Kauwe, S. K., Graser, J., Murdock, R. & Sparks, T. D. Can machine learning find extraordinary materials? *Comput. Mat. Sci.* **174**, 109498 (2020).
106. Schwaller, P. et al. "Found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**, 6091–6098 (2018).
107. Bergmann, U. et al. Using X-ray free-electron lasers for spectroscopy of molecular catalysts and metalloenzymes. *Nat. Rev. Phys.* **3**, 264–282 (2021).
108. Ayyer, K. et al. Low-signal limit of X-ray single particle diffractive imaging. *Opt. Express* **27**, 37816–37833 (2019).
109. Brewster, A. et al. Processing serial crystallographic data from XFELs or synchrotrons using the cctbx.xfel GUI. *Comput. Crystallogr. Newsl.* **10**, 22–39 (2019).
110. Young, I. D. et al. Structure of photosystem II and substrate binding at room temperature. *Nature* **540**, 453–457 (2016).
111. 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).



### 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
All authors contributed equally to all aspects of the article.

### Competing interests
The authors declare no competing interests.


### Peer review information
*Nature Reviews Chemistry* thanks Martin Green, Venkatasubramanian Viswanathan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.


### Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### RELATED LINKS
**Autoprotocol:** https://autoprotocol.org/
**Cambridge Structural Database:** https://www.ccdc.cam.ac.uk/
**CAMERA:** https://camera.lbl.gov/
**Chemotion Repository:** https://www.chemotion-repository.net/welcome
**FAIR principles:** https://www.go-fair.org/fair-principles/
**HardwareX:** https://www.journals.elsevier.com/hardwarex
**IBM RXN:** https://rxn.res.ibm.com/
**Inorganic Crystal Structure Database:** https://www.psds.ac.uk/icsd
**MaterialNet:** https://maps.matr.io/
**NMRShiftDB:** https://nmrshiftdb.nmr.uni-koeln.de/
**Open Reaction Database:** http://open-reaction-database.org
**Protein Data Bank:** https://www.rcsb.org/
**PuRe Data Resources:** https://www.energy.gov/science/office-science-pure-data-resources
**Reaxys:** https://www.elsevier.com/solutions/reaxys