Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2023, **2**, 1644

What is missing in autonomous discovery: open challenges for the community

Phillip M. Maffettone, *[ab] Pascal Friederich, *[ip] Sterling G. Baird,[c] Ben Blaiszik, [de] Keith A. Brown, [f] Stuart I. Campbell, [a] Orion A. Cohen,[g] Rebecca L. Davis, [h] Ian T. Foster,[de] Navid Haghmoradi, [i] Mark Hereld, [de] Howie Joress,[j] Nicole Jung,[k] Ha-Kyung Kwon, [l] Gabriella Pizzuto, [m] Jacob Rintamaki,[n] Casper Steinmann, [o] Luca Torresi[p] and Shijing Sun[lq]

Self-driving labs (SDLs) leverage combinations of artificial intelligence, automation, and advanced computing to accelerate scientific discovery. The promise of this field has given rise to a rich community of passionate scientists, engineers, and social scientists, as evidenced by the development of the Acceleration Consortium and recent Accelerate Conference. Despite its strengths, this rapidly developing field presents numerous opportunities for growth, challenges to overcome, and potential risks of which to remain aware. This community perspective builds on a discourse instantiated during the first Accelerate Conference, and looks to the future of self-driving labs with a tempered optimism. Incorporating input from academia, government, and industry, we briefly describe the current status of self-driving labs, then turn our attention to barriers, opportunities, and a vision for what is possible. Our field is delivering solutions in technology and infrastructure, artificial intelligence and knowledge generation, and education and workforce development. In the spirit of community, we intend for this work to foster discussion and drive best practices as our field grows.

Received 1st August 2023
Accepted 5th October 2023

DOI: 10.1039/d3dd00143a

rsc.li/digitaldiscovery

1 Introduction

Scientific experimentation and discovery is teetering on the precipice of a new industrial revolution. Acceleration of science by combining automation and artificial intelligence (AI) has begun to revolutionize the structure of scientific experiments across physics,[1] chemistry,[2–8] materials science,[9–13] and biology.[14] The integration of high-throughput experimentation, AI, data science, and multi-scale modeling have spawned great interest,[15] notable results,[16] and substantive expectations. These expectations include acceleration of experimental throughput, new discoveries, technological readiness, and industrial adoption. Such excitement has elicited a suite of conferences (from a 2017 North American workshop culminating in the first Mission Innovation report[17] coining the name Materials Acceleration Platform (MAP), to the most recent Accelerate Conference), dedicated publication platforms, and increased funding from governments and the private sector. Furthermore, as a link between algorithms and the real world, self-driving labs (SDLs) are a prerequisite for further advancements of autonomous and AI-driven research, as targeted by, for example, the Turing AI Scientist Grand Challenge as a forward-looking roadmapping effort.[18] While this rapid community advancement may not yet constitute a scientific revolution,[19] it does initiate a technical revolution that will likely change the pace at which we see scientific breakthroughs (Fig. 1).

A self-driving laboratory (SDL) can be described as a scientific system that performs autonomous experimentation (AE). That is, it uses automation and AI to operate and select each successive experiment, without requiring human intervention.

[a]*National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY 11973, USA. E-mail: pmaffetto@bnl.gov; pascal.friederich@kit.edu*
[b]*BigHat Biosciences, San Mateo, CA 94403, USA*
[c]*University of Utah, Salt Lake City, UT 84108, USA*
[d]*University of Chicago, Chicago, IL 60637, USA*
[e]*Argonne National Laboratory, Lemont, IL 60439, USA*
[f]*Boston University, Boston, MA 02215, USA*
[g]*Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[h]*University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada*
[i]*Karlsruhe Institute of Technology, Institute of Nanotechnology, Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, 76344, Germany*
[j]*Materials Measurement Science Divison, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA*
[k]*Karlsruhe Institute of Technology, Institute of Biological and Chemical Systems, Hermann-von-Helmholtz-Platz 1, Eggenstein-Leopoldshafen, 76344, Germany*
[l]*Toyota Research Institute, Los Altos, CA 94022, USA*
[m]*University of Liverpool, Liverpool, L69 3BX, UK*
[n]*Stanford University, Stanford, CA, 94305, USA*
[o]*Department of Chemistry and Bioscience, Aalborg University, 9220, Aalborg, Denmark*
[p]*Karlsruhe Institute of Technology, Institute of Theoretical Informatics, Engler-Bunte-Ring 8, 76131, Karlsruhe, Germany*
[q]*University of Washington, Seattle, WA, 98195, USA*

Fig. 1 A triptych of stable-diffusion generated images describing a self-driving lab for autonomous scientific discovery.[22]

Several other terms are commonly used in this domain which are worth disambiguation from SDLs. High-throughput experimentation (HTE) applies automation technology or engineering practices to increase data generation rates, but these experiments are often fully designed by human experts or predefined at the start of an experimental campaign.[16] Similarly, "Lab 4.0" refers to "intelligent automation systems and digitization in modern laboratory environments, and work practises"; however, these do not strictly imply fully autonomous experimentation. A common term used for SDLs in the materials science domain is materials acceleration platform (MAP).[7,17] The major differentiator between platforms is the degree of autonomy imbued in the system—and increasingly, the degree of human-AI collaboration. In the following, we will use the term SDL and focus broadly on (semi-) automated platforms that accomplish (high-throughput) experiments, process and analyse results autonomously, and use that analysis to guide future experiments. While an SDL can be typified by a closed loop of synthesis and analysis unit operations,[20] a single SDL can be incorporated as a unit operation inside a larger SDL, so long as it meets the criteria above.[21]

Herein, we describe a community perspective on the state of SDLs, focused on open challenges and concerns for the future. This perspective incorporates insights from academia, government, and industry, including both users and developers of SDLs. This work grew organically from discussions started at the first Accelerate Conference in September 2022, and is intended to foster an ongoing discourse to influence the field of autonomous experimentation. While we will first describe the current state of the field, this is by no means a comprehensive review, and we encourage the reader toward reviews and perspectives of SDLs and autonomous experimentation.[1–14,16,23–38] Following this, we will turn our attention to barriers and opportunities associated with data, hardware, knowledge generation, scaling, education, and ethics. As the field of autonomous experimentation grows and SDLs become more common, we hope to see rapid growth in scientific discovery. In the community approach to vision that follows, we look to ethically and equitably accelerate this growth and adoption.

2 Data, data, everywhere, and not a piece to parse

A core difference between research using SDLs and conventional research is the amount and structure of data that are generated. Compared to human-guided experimentation, SDLs can enable rapid experiments, with improved reproducibility, and with automatically generated metadata. This presents many opportunities to reduce the barrier and overheads to share, collect and use data globally. At the same time, it leads to challenges in terms of storage, analysis, and even interpretation of data. Perhaps one of the unique opportunities afforded by SDLs is that data could be shared in real time with the community while preserving its provenance. In the following, we will discuss best practices for data sharing, but also existing barriers and incentives to lower and overcome those barriers. Data are foundational to SDLs and leveraging the full potential of data-driven approaches in science remains an area of rich opportunity.

Best practices of data sharing and publication are summarised in the FAIR principles,[39,40] *i.e.*, data that are findable, accessible, interoperable, and reusable. Findable means that data should be easy to find and access, with clear and consistent metadata, and data identifiers. Accessible means that data should be openly available through a trustworthy and stable digital repository with a clear and simple access mechanism. Interoperable means that data should be structured with common standards, shared vocabularies, and standardized formats that enable integration and reuse. Reusable means that data should be well-described with provenance and licenses, allowing it to be used and cited in new research and for new purposes. FAIR methods are most effective when applied continuously during research, rather than only at the moment of data publication.[41] These principles enable the linking of many experiments and simulations, and enable SDLs to save resources, leverage existing knowledge, and utilize synergies. Despite the successful implementation of FAIR principles by some exemplary practitioners,[42,43] there remain numerous technological and cultural barriers to FAIR data sharing. It is thus important to be forthcoming about these barriers, as we look to understand what strategies or incentives are viable for improving the overall data publishing landscape.

Primary technical barriers to entry center around data acquisition and digitization. An early challenge is the movement of data from unit operations to data storage infrastructure, databases, or long-term repositories.[44–47] While electronic lab notebooks can help integrate persistent manual processes, data acquisition in SDLs is mostly challenged by lack of accessibility, as many scientific instruments still produce data in obfuscated proprietary or binary formats[46] and data often remains undiscoverable at the point of data generation. Developing open interfaces to hardware is an ongoing and increasingly solved task, which will enable open software to collect and manage the data as they are produced. Data – not only characterization data but also metadata describing sample processing conditions – should ideally be recorded as close to the source as possible. If manually recorded later, this will not only increase the effort of data collection but also potentially lead to data loss due to incompleteness and missing standardization. Again, electronic lab notebooks connected to electronic equipment are one way to digitize (process) data directly during generation, while removing the effort of additional book-keeping.

Once created, digital data can easily be stored on a hard disk as a series of named files, yet this approach does little to ensure that the data can be found, understood, and reused in the future by collaborators, the community, or machines, without insight from the original data creators.[48] This problem is further exacerbated by the fact that most researchers are accustomed to the file systems of their laptops, but lack the background, training, or incentive to use shared community databases, common metadata standards, unique data identifiers, and well-defined vocabularies. Furthermore, there are additional technical challenges connected to choosing appropriate databases (hierarchical, relational, non-relational) and data storage with appropriate data protection, safety, and maintenance (local or cloud). The development of research data management (RDM) tools, often driven by bottom-up initiatives within domain specific communities, help overcome these hurdles. RDM developments can facilitate FAIR data by providing appropriate shared metadata standards, domain-specific vocabularies and ontologies, and also software and storage solutions. An example RDM system that provides some of this functionality is contained in the Bluesky project, which implements a standard "Data API (application programming interface)" that is exposed to users and other systems rather than requiring knowledge of internal implementation details on where and how the data are stored.[44]

The last set of technical barriers relates to data interoperability, which also extends to the homogenization of heterogeneous data sources. Relating data from disparate equipment, including distributed instrumentation, multi-fidelity probes, or even simulations, requires some knowledge of the core metadata used to describe each relevant experiment. Unfortunately, constructing these relationships requires some degree of standardization and is highly discipline dependent. Attempts at universal materials formats[28,49–52] and ontologies[53,54] exist to address this challenge, with much work remaining. Crystallography provides an excellent example of standardizing data formats for interoperability (protein data bank and crystallographic information files).[55,56] While this standardization provides a means of comparing measurements, it does not provide rich metadata for samples. Data validation and schema[57] can be used to create interoperable sample data. The use of digitally structured protocols[58] and publishing peer-reviewed methods articles in journals such as may also aid in this challenge while providing appropriate incentives to researchers. Overcoming this challenge will take effort, open communication, and many revisions. Even in a single lab, lean engineering practices should be applied to a regularly updated and versioned data model (*e.g.*, ambient humidity or materials batch numbers may not be required data fields until a keen researcher notes them as important exogenous variables). We therefore encourage the community to develop, collaborate on, and publish data models, particularly in a version-controlled manner. Even highly specific data models can prove impactful and be improved upon.

An open challenge in many domains—or sub-domains—remains the creation and support of easy-to-use, open-access domain-specific repositories and databases. Similar to the domain-specific nature of sample data models, the possibilities are innumerable and can start small and iterate. It is important to note the difference between generic databases and repositories such as FigShare and GitHub, currently only domain-specific databases provide data with enough structure to be reused within the community.[59] To improve discoverability, there are a number of FAIR data databases and repositories in the materials science community that may prove useful to collecting SDL data and for the SDL community to build upon, including the Materials Project,[60] AFLOW,[61] the Materials Data Facility,[62,63] OQMD,[64] and NOMAD.[65] However, none have yet been closely integrated with SDL. As more databases become available, a new opportunity will be present in building infrastructure to interoperate between them or merge them under single data models. This again relates back to challenges with the ingestion of standardized data formats for various hardware vendors to seamlessly relate measurements.[66] Quality control and continuous integration (QC/CI) are challenging tasks even for data workflows from a single source;[67] as interoperation grows, so will the challenges for quality control and integration. This will be a welcome opportunity and a hallmark of community progress. It can be addressed by first building internal trust in data through QC pipelines that are incorporated in a data model. Cloud tools such as AWS and Azure, as well as workflow managers[68] can be used to build automation into the construction, maintenance, and QC of these databases.

Aside from the substantial technical challenges of building a robust data infrastructure, there are many cultural hurdles. Some are centered around sharing proprietary data that represents a material value to its creators. Others reflect researchers' fear of being "scooped" or yielding a competitive advantage. While technical issues can be solved with clever engineering and sufficient funding, cultural issues cannot. Instead, workshops, conferences, training programs, and higher education serve as the engines of cultural change. We suggest conceptualizing improved data management as a socio-technical transition pushing against both technical and cultural lock-in to existing practices.

2.1 Suggestions for data sharing incentives

We have identified three strategies for incentivizing the production of FAIR data, putting into focus the human researcher, rather than MAP technology:

- Reducing friction.
- Providing rewards.
- Demanding requirements.

Creating the technical means to share data in a simple way is a necessary condition to allow researchers to share data in a sustainable way. Reducing friction means making it simple and effortless for researchers to upload their data to the appropriate places. Whether this is the time it takes to create a DOI for data, or the effort needed to validate the data into a given database schema, people will be more likely to engage with a tool if it does not feel like work. We encourage the "customer development" required for creating an appropriate user interface for data tools.[69] This ties into the above

discussion of creating and maintaining field-specific databases, which are easily accessible and also integratable into SDLs to further enable the automation of data publishing pipelines. If these are accessible with a continually refined user interface, the community will be more likely to contribute their data.

Even if they are easy to use, technical tools are not enough to create a sustainable development toward openly sharing data. To encourage researchers to share data more broadly, we believe that data sharing should be met with recognition and rewards. The recognition could take the form of data citations, which are tracked by the statistics of access and downloads. Such recognition would directly create incentives to publish in databases that impose high standards to ensure high quality, reproducibility, and documentation of data, as such databases will be used and thus cited more by the community. Furthermore, this can incentivize the publication of "negative-data".[70] Broad data availability can further lead to citations by researchers without access to labs and HPC infrastructure. Citable data can then lead to the definition of new metrics. Consider, for instance, the social and professional impact of the h-index. While there are systematic challenges associated with over-reliance on a single metric, the h-index provides a more holistic—if flawed—means for measuring impact. We envision a complementary metric for data (*e.g.*, a d-index), that could be built off of the unique identifiers for data in a database, allowing data to be referenced digitally in publications. While new metrics bring new concerns, integrating these metrics into the traditional advancement criteria at research institutions would produce a dramatic cultural shift.

An alternative approach to creating incentives within the currently existing research reward system centered around publications and citations could include the use of automated papers that are published regularly, listing all recent contributors to a public database. This can even be transferred to track software impact, *e.g.*, publishing a citable list of contributions to given repositories. This idea is not entirely new, as large software tools regularly publish reports of new versions with all contributors as authors,[71] and some publishers have created venues for this type of content.[72] The main outcome of this strategy is converting data contributions into publications/citations and pivots the credit mechanism to an already imperfect set of metrics. Additionally, it may engender "over-publishing"[73] leading to an oversaturation of the primarily text-based publishing system. This can be stemmed in part by limiting the text-based content of the actual journal article and shifting the reviewers' focus to the quality and presentation of the data. While the career advancement for a scientific software engineer can be as closely tied to their publication record as their version history of their software, no such paradigm has been publicly recognized by an R1 University, especially not in the context of data provenance. Thus, we think creative solutions to provide micro-incentives to sharing data and exercising good data practices will be a critical need in making progress in this field and encourage social scientists to consider this research problem.

In continued circumstances where a reward infrastructure proves lacking, we encourage mandates through peer review and funding agencies. This policy can be enacted at the journal or editor level, or be enforced at the level of individual peer review, as FAIR data are a reasonable pre-requisite for reproducibility. Some journals already enforce code review as part of their process for scientific software, and it is sensible to enforce data review for papers that describe large datasets. With regards to funding, many government funding agencies require a data plan, albeit, enforcement is critical. Furthermore, data plans often only include positive results. Public access to data created using public funds is paramount and should be rigidly enforced by funding bodies. Data plans, however well-designed, are not useful if they are not used. There is also no reason why this must be done from scratch: for example, a data management framework could be developed and highlighted alongside a new dataset. Best practices will naturally differ between communities with different measurement techniques, instrumentation, and figures of merit. It is therefore advisable that data management frameworks be designed through active collaboration between scientific communities and funding agencies.

3 Integrating hardware into SDLs

The experimental apparatus that constitutes the physical embodiment of the SDL provides its own set of challenges. While data and software solutions are reusable across a large swath of the research landscape, hardware advancements must be capable of handling the specifics of the research problem at hand. Further, physical platforms are generally the most capital-intensive part of an SDL, including the costs of the scientific instrumentation required for a given experimental workflow, and often custom automation hardware. Therefore, we stress an objective that the hardware powering SDLs generalise across different experiments to make the investment reasonable.

There are three common approaches to hardware in current SDLs: building hardware from individual components (*e.g.*, motors, pumps, controllers, detectors); using workcells (*i.e.* integrated systems which bring together automation equipment, analysis tools, and software, to accomplish rigid predefined tasks); and integrating unit operations with an anthropomorphic robotic platform. The from-scratch approach is most common in specialist equipment, such as beamlines,[44,69] new microscopes,[74] or specialist synthesis and characterization approaches.[75,76] Much of this equipment came out of the older high-throughput community.[77] Recently, growing maker communities have driven some build-your-own workcells similar to 3D printer technology.[78] Workcells first came to use in high throughput biological applications,[79] and now have commercial providers across the physical sciences.[80] These are relatively rigid unit operations for a given experiment type, although it has been demonstrated that a robotic arm can be integrated internally which could provide more flexibility.[81] To add even more flexibility, the use of mobile robots for transporting sample vials and using equipment across different laboratory stations has been demonstrated,[20] in addition to the usage of heterogeneous robotic platforms depending on the laboratory tasks.[82] As such the current state of the art for

laboratory automation varies depending on the commonality of the task to automate, and the commercial demand for products; a generic liquid handler has a much broader market appeal than a domain-specific workflow.

Although building hardware is hard, it pales in comparison to the challenge of integrating hardware. This has amassed public[42,44,46,82,83] and private efforts. Developing a hardware approach for an SDL involves a combination of deciding on the best tools for the laboratory task and building common interfaces for those tools. As an example, we contrast a synthesis workflow that can be accomplished in a commercially available workcell or using custom equipment, an analysis workflow that depends on advanced detectors, and a sample management workflow that uses robotic arms. The synthesis workcell may include a software interface that cannot be rebuilt, whereas the custom synthesis equipment will need to choose an effective software interface. Advanced detectors on the other hand are commonly integrated using open software tools.[84,85] Lastly, robotic arms are traditionally driven using middleware that, through already available libraries, have in-built motion planning, low-level controllers, and perception.[83] While these are fundamental for autonomous robotic platforms, given that most synthesis or analysis workflows are carried out in open-loop, these functionalities have not been fully exploited. This then raises the question of how to best integrate these disparate systems without attempting the mammoth task of rebuilding them all using a single software tool.

We encourage the development and reuse of open-source, non-proprietary hardware communication, as well as interfaces between those common platforms. This would facilitate sample exchange across multiple commercial experimental tools, in conjunction with bespoke tooling, as well as improving knowledge transfer, reproducibility, and generalization to new SLDs. Interfaces should make use of industry-standard message bus technology, that will enable both local and cloud operation. Moreover, these interfaces are compatible with streaming data, that enable *in situ* and real-time measurements with automated data-processing pipelines. The materials community has called for investment in "the redesign of microscopes, synchrotron beamlines, and other sophisticated instrumentation to be compatible with robotic sample handling—akin to the multi-plate-handling robots in the bio-community."[16] Such innovation would be empowered by the adoption of open frameworks and message buses (*e.g.*, by integrating a Robotic Operating System (ROS) enabled robot, a Bluesky driven beamline, and propriety unit operation with a standard message).

The economy of public research requires that SDL platforms—or at least their components—be reusable beyond the scope of a single research project. One approach to this is to develop modular systems that can be added to over time to meet the new demands of new research questions. This would be supported by a common physical sample interchange environment. The details of sample interchange greatly depends on the form factor of the material being measured.

Samples can generally be clustered into three varieties: liquid, bulk solids (including thin films on substrates) and powders. Liquid handling is the most common in current SDL setups, having been pioneered by the bio-pharmaceutical sector. Samples can be physically handed between instruments in vessels, moved through the use of pipetting,[80] or directly pumped through piping.[76] Handling of bulk solids is also relatively straightforward. Robotic arms are able to move samples between processing and characterization tools.[86] The sample surface is also readily available in this form factor for characterization. Each sample may be homogeneous or contain multiple sub-samples.[87,88] Powders are perhaps the most challenging to deal with using automation. While precursor powders can be readily dispensed, powders as a product can be difficult to handle. While powders can be moved in a vial, this is often not a form factor that is amenable for characterization. For example, X-ray diffraction typically requires creating a flat surface or packing the powders into a capillary. Conversely, powders for catalysis or absorption of gasses typically require these powders to be packed into specialized columns for testing. Further complicating this process is the fact that powders can have a great variety of flow properties, requiring adaptive manipulation. We encourage robotics and mechatronics research in this area, as automation of powder handling would solve a particularly impactful tactile challenge.

3.1 Applying robotics to SDLs

Deploying anthropomorphic robotic systems in SDLs is a promising area of research in that it enables the use of existing or standard instrumentation, empowers human collaboration for non-automated tasks, and is generally more flexible.[89] Despite the increased cost and complexity, there have been several research efforts in this area, spanning applications from autonomous solubility screening,[81,82,90] photocatalysis,[20] and automated synthesis.[91] SDLs provide an exciting semi-structured environment where the robotics community can transfer their methods to novel applications. Robotics researchers have focused their attention on various applications—from household environments to extreme environments such as nuclear and space—that possess common underlying challenges with SDLs. For example, the challenges of assistive home robots related to grasping transparent glassware are also present in SDLs. Towards this goal, learning-based methods *e.g.* TranspareNet[92] and MVTrans[93] have been demonstrated for detecting laboratory glassware. In addition, there exists the need to have more task-specific grippers, such as grippers that are specifically designed for laboratory containers and well-plates.[94] While a universal gripper that would exhibit the dexterity of a human hand may seem ideal, we are still quite far from this and most laboratories have to either adapt their current grippers with 3D printed parts and/or use tool exchangers. Alternatively, laboratory tools *e.g.* pipettes can be made more robot-friendly.[95] Amongst others, these approaches will pave the way to having more robust laboratory robotics for experiments over long periods of time.

While there is an increasing interest towards using anthropomorphic robotic platforms, primarily due to their attractive nature of being deployed in human labs, there still exist a large number of challenges with having these platforms carry out

long-term experiments. As a robot can be described as any system of sensors and actuators that can perform a complex task, it is worth noting when a workcell or other automated unit operation can be used to substitute or supplement an anthropomorphic robot. To this end, modular systems that can—in principle—handle a wider range of experiments due to the closed nature of their subsystems have been demonstrated as alternatives.[96,97] To date, the state-of-the-art in using anthropomorphic robotic systems in SDLs have carried out experiments on open benches, which limits the generalisation of experiments towards materials that have a higher degree of toxicity.

Democratization of laboratory automation is a crucial path for the contemporary developments of individual lab groups to transcend beyond local, bespoke solutions. It has thus far been exceptionally difficult for the community to transfer knowledge between laboratories regarding their hardware development and integration. We are calling for a focus and investment in modularity[30] and open hardware,[32,98–104] that makes use of the aforementioned communication approaches. Publishing modular hardware components—either through new journals or maker spaces—will enable the community to take advantage of rapid prototyping and manufacturing. The combination of open hardware and open control software will have the accelerating effect of democratizing access to lab automation and ensuring our brilliant peers have this technology regardless of resources.

3.2 Miniaturization

Notably, we see multiple opportunities in miniaturization and reducing the footprint of hardware to both reduce the barrier to entry and reduce material consumption and waste production. Digital microfluid platforms have a high capability for miniaturization because they replace moving mechanical parts with a liquid that needs to be moved.[105] The HTE community in biology and chemistry has long used microwell plates to reduce sample volumes to the microliter scale.[106,107] In materials science and chemistry, miniaturization can have a high impact[108,109] but is to date underexplored. In addition to adopting microwell plates,[110] microfluidic reactors allow for samples in SDL to reach the nanoliter scale.[111] SDLs built entirely using flow chemistry can take advantage of established chemical engineering without anthropomorphic robots. Flow reactors are also easier to integrate with online characterization techniques.[2] Milli- or micro-fluidic reactors have the advantages of higher rates of heat, mass, and photon transfers as a result of the enhanced surface-to-volume ratio. Nonetheless, flow reactors suffer from the possibility of clogging when dealing with solid-state materials or precipitates.[112] These cases (including thin-film preparation, battery materials, or polymerization with precipitation of solid products or byproducts) are better suited for parallel batch reactors.[2] Modular microfluidic units could accelerate process optimization and formulation discovery; however, a standardized protocol in modular configuration for a targeted reactive system needs to be established.[113]

In addition to the application of microfluidic reactors, miniaturization has also been applied to devices and solids sample arrays. This has been approached using combinatorial synthesis in small areas (<1 mm^2),[114] multinary thin film synthesis,[115,116] and microdroplet array synthesis.[117–119] Recent work with scanning probes has shown that sub-femtoliter solutions can be patterned and combined on surfaces, providing further opportunities to miniaturize experiments.[120,121] Such samples can also serve as miniaturized reactors,[122] having the broad effects of reducing material consumption and increasing experimental throughput. In addition to miniaturizing samples, it is also powerful to minimize the analytical instrument to study these samples, as exemplified by scanning droplet cells to study corrosion[123] and adapting electrochemical characterization techniques.[124,125] While not all characterization techniques can be applied in miniaturized platforms,[2] we encourage further developments in this area that increase automated capacity while reducing material consumption.

3.3 Actions for the community

Advancements in hardware often demand tailored considerations, tightly aligned with the unique demands of specific research endeavors. Nonetheless, these physical platforms constitute a significant portion of an SDL's investment, encompassing scientific instrumentation and bespoke automation hardware. Therefore we emphasize a crucial objective: the hardware underpinning SDLs must demonstrate versatility and transferability, transcending individual experiment boundaries.

In navigating this challenge, fostering open-source hardware communication and interfaces proves paramount. Standardized interfaces that facilitate seamless sample exchange across a spectrum of commercial experimental tools and bespoke setups become imperative. Adopting industry-standard message bus technology, compatible with streaming data for real-time measurements and automated data-processing pipelines, holds promise to enable communication between platforms. We expect leveraging anthropomorphic robots to integrate bespoke equipment to be a crucial area of future research, and encourage the adoption of open solutions such as the Robotic Operating System to prevent vendor lock-in and enable technology transfer.

Furthermore, the democratization of laboratory automation stands as a crucial aspiration. The path forward advocates for a thoughtful interplay of modular hardware, open communication standards, strategic integration of robotics, and the pursuit of miniaturization. We expect this to be achieved in part by facilitating the publication of modular hardware components and fostering their integration with open-source software. This paves the way for sustainable, accessible, and efficient SDLs, wherein hardware seamlessly caters to the ever-evolving landscape of scientific exploration.

4 Algorithms

At present, many algorithmic approaches to governing self-driving laboratories are in the category of global optimizers.

These tools aim to find the optimal solution within a specified search space. Evolutionary and genetic algorithms, being prominent members of this framework, draw inspiration from the process of natural selection to iteratively evolve a population of candidate solutions, progressively converging towards optimal or near-optimal solutions. These algorithms excel in exploring vast and complex search spaces, making them particularly suited for optimization problems with high-dimensional or non-convex landscapes. Bayesian optimization, another pivotal component, addresses the trade-off between exploration and exploitation in the search for optimal solutions. By constructing a probabilistic surrogate model of the objective function, Bayesian optimization intelligently chooses points to evaluate, significantly reducing the number of function evaluations required for optimization. Reinforcement learning, on the other hand, mimics the process of learning through trial and error in an interactive environment. It allows an agent to optimize decisions by receiving rewards based on actions, ultimately learning an optimal policy. These algorithms have been deployed in domains like robotic control and games, and found recent adoption in SDLs.[126] The integration of global optimization techniques represents a powerful approach to solving a wide array of real-world problems by efficiently navigating complex search spaces and finding high-quality solutions. In the following, we will highlight where global optimizers have supplanted other experimental design techniques, and discuss where they have fallen short.

4.1 Autonomous decision making for optimization problems

The traditional design of experiments (DoE) becomes rapidly impractical for high dimensional problems due to the exponential growth of the number of required experiments. Incorporating ML in SDLs has emerged as an efficient way to explore the chemical space and to speed up experimentation. Taking advantage of the information generated during the optimization process itself, ML enables an iterative experimental design that maximizes the information gained per sample and that requires a smaller number of experiments with respect to traditional DoE.[112,127]

At present, a variety of ML approaches have been applied to SDLs. Genetic algorithms (GA) are a class of adaptive heuristic search algorithms inspired by the process of natural selection that can be used for solving both constrained and unconstrained optimization problems. GAs have been applied recently to optimize the conditions to produce gold nanoparticles.[128] Reinforcement learning (RL) has also been successfully applied in SDLs.[126,129,130] This is an ML paradigm that enables an agent to learn through trial and error in an interactive environment by taking actions and receiving rewards, with the goal to learn a generic approach that maximizes the total reward over time. The most widely used decision-making algorithm in SDLs is Bayesian optimization (BO), a method particularly suited to balance the trade-off between exploration and exploitation of the input parameter space. BO has been applied to SDLs both in the single objective setting[3,81,131] and, more recently, for the simultaneous optimization of multiple objectives.[132] There are two main strategies for the implementation of multi-objective BO: combining multiple objectives into one (*e.g.*, Chimera[133]) and identifying a Pareto front that trades off among the multiple objectives (*e.g.*, qNEHVI[134]). The second approach has the advantage of not requiring the experimentalist to select the trade-off among the different objectives *a priori*.

A challenge in SDLs is selecting suitable algorithms for each specific scenario. Open-source software packages for SDLs offer an easy-to-use starting point for non-experts in machine learning to begin autonomous experimentation (ChemOS,[127] EDBO+[3,135]). Other general-purpose libraries are becoming more user-friendly and are constantly updated with SOTA methods, such as BoTorch,[136] or Ax, an adaptive experimentation platform built on top of BoTorch. Often, off-the-shelf decision-making algorithms in general require further tuning, which can slow research and even undermine an experimentalist's intent in purpose of applying them. In this sense, the lack of open-access datasets for experimental campaigns is a current issue. Relating back to our discussions around data (Section 2), the availability of data would allow researchers to evaluate novel algorithms on multiple surrogate systems based on real experiments, while the absence of such datasets could considerably impede the development of dedicated algorithms and create obstacles for the development of autonomous platforms.[137,138]

An open challenge is the incorporation of generative models[139–141] for compositions,[142–146] crystals,[147–152] and molecules[142,153–156] in SDLs. While being very successful in finding hypothetical molecules with tailormade properties, *i.e.* solving the inverse problem of materials design, generative models frequently suggest molecules with complex or unknown synthesis routes,[157,158] which limits their real-world impact and prevents their application in automated labs. Including synthesizability or even synthesis planning in generative models for inverse design, as well as developing versatile multi-objective generative models is a promising path toward their integration in SDLs.[159,160]

A further open, yet more technical challenge in SDLs is the systematic exploitation of existing data and also process descriptions[161] that are published in the literature. Automatic extraction of that data and conversion from natural, *i.e.* informal language to computer-readable, *i.e.* formal language is an open challenge, currently requiring a large amount of manual work.[162] Large language models can potentially help in that task[163–166] but further research is required to reliably extract data and knowledge from scientific literature.

4.2 Knowledge generation

The ultimate goal of scientific experiments is typically not (only) the generation of data, but the generation of knowledge and understanding.[167] From that aspect, the main objective of SDLs should go beyond solving optimization problems in high dimensional spaces of materials and processing conditions. Rather, they should aim to interpret that data, link it with other data (potentially from other SDLs and databases), and help to

generate and test scientific hypotheses (potentially in a semi-autonomous or autonomous manner[168]). To come closer to the goal of autonomous generation of scientific knowledge and understanding, we propose multiple considerations for the future.

In order to leverage data generated across labs and use it to train models, further data analysis, and generate new scientific knowledge, we underscore the efforts highlighted in Section 2 to link data through shared data formats, metadata definitions, vocabularies and ontologies. This will move SDLs beyond the discovery of singular, interesting data points. From these innovations, data can be subsequently used for transfer learning models, multi-fidelity and multi-task models, as well as representation learning methods, which can learn from heterogeneous datasets.[169] Such pre-trained models enhance the decision-making process in SDLs. In particular, they not only learn from locally generated data but already have prior knowledge that enhances decision-making early in an experimental campaign. Lastly, we encourage the development of AI/ML methods for SLDs that reach beyond optimization problems. SDLs are very good at finding optima, which can act as sources of inspiration for scientific understanding, as defined in Krenn *et al.*[167] Furthermore, increasingly sophisticated explainable and self-explaining machine learning models for molecules,[170,171] materials science[172,173] and particularly SDLs[174,175] pave ways towards autonomous loops of hypothesis generation. Methods such as automated generation of counterfactuals[176] offer further opportunities for automated hypothesis testing, when combined with fully automated synthesis and characterization, or accurate predictive simulation workflows.

5 Scaling autonomous discovery

Taking the opportunities of data, hardware, and software in concert, we can turn our attention to economies of scale. In this section, we focus our discussion on scaling *via* interconnection of multiple distinct hardware modules and SDLs,[21] or *via* increasing the size, capacity, or extent of a given SDL. The scalability of SDLs supports a faster, more efficient exploration of experimental parameter space, as well as a larger volume output of manufacturing processes.

SDLs at the laboratory scale already empower HTE increasing throughput by orders of magnitude. Scaling an SDL beyond the single lab will increase experimental throughput proportionately. Moreover, moving from local autonomy to distributed autonomy will enable optimization and search over multiple length scales, characterization methods, and related systems. For example, a lab-scale SDL within a single confined system, such as glove box handling of liquids, may enable a higher experimental throughput. Coupling this with an SDL at a different scale, such as a synchrotron, may enable new discoveries by bridging techniques across multiple length scales. Scalable computational approaches can capitalize on a greater experimental throughput and the interplay between modular SDLs for a more efficient search through space. Designing SDLs that act in concert over multiple fidelities and length scales will enable materials verification and validation that is nearer to industrial requirements. This ties in with the "advanced manufacturing" movement that has received significant attention in the last decade.[177]

Building scalable SDLs requires a particular landscape of considerations. Broadly these include the manufacturing considerations and the transition from automation to local autonomy to distributed autonomy. There are challenges in scaling the volume of experiments, synchronizing data (Section 2), algorithms for handling multifidelity and multimodal data (Section 4), and software that enables distributed orchestration across platforms. As the number of experiments performed per unit of time increases, it is crucial that experimental platforms are found that minimize the amount of material required per experiment (Section 3). This reduces the costs of experiments, makes them more amenable to parallelization, and reduces the time required for some types of processing steps. However, this may change the relevance to manufacturing scale, so there is a case to be made for multi-scale automation that features high-throughput experiments at a highly miniaturized scale and lower-throughput experiments at larger scales.

Scaling beyond single-laboratory SDLs creates some open questions in cost analysis. How do we model the cost of implementing and operating such a large-scale high-throughput discovery architecture? Can we quantify the aspects of scale that result in cost efficiency? The capital expenses would include floor space, instrumentation (sample production, characterization, storage), and mechanical infrastructure (robotics, table space, resource garages). The operational expenses would include materials, power, maintenance, and replacement. There are broad considerations around architectural design, and efficiently accommodating the disparate instrumentation sets required by different workflows. We encourage such engineering and feasibility analyses by public research centers to be made broadly available, to inform continual improvement cycles.[178] Calling back to our discussion on modularity, we further highlight the economic analyses of "platform" (*i.e.* incremental) *vs.* "bespoke" (*i.e.* single-leap) development strategies for large projects, and resoundingly encourage the former.[179]

When considering the transition from automation to local autonomy to distributed autonomy, we are focusing first on the distinction between HTE and AE/SDLs, with a second distinction between an isolated SDL and a distributed SDL or network of distributed SDLs.[21] Mobility may be a key component to making the transition from local to distributed autonomy. As discussed in Section 3, this can be accomplished by fixed material transport systems (robotic rail systems, custom feeders) or by more generalized mobile robotic technology.

Large-scale autonomous experimentation also creates ample opportunities for software innovations in orchestration, communication, and algorithms. Scaling SDLs will require actional workflow systems—mediated perhaps by powerful workflow languages—designed to enable a description of the requirements for each sub-task in a workflow as well as the specifics needed to connect intermediate products from one sub-task to the input of subsequent sub-tasks.[28,34,44,180,181]

Intelligent search or orchestration that scales is a second opportunity for software innovation in an ongoing area of rich study. This requires a capacity for evaluating incoming results from experiments carried out in parallel on the work floor, and converting this new data into new experimental queries. This computational component connects the stream of outputs from in flight experiments back around to the process that injects new experiments into the available resource pool. It does so in light of all previously collected experimental data, models of the processes being studied, and a growing model of the abstract landscape defined by the problem goal specification.

Scaling global optimization algorithms is a critical endeavor to handle increasingly complex and data rich problems. Techniques like distributed computing, parallel processing, and hardware acceleration have been leveraged to process immense amounts of data and execute computationally intensive tasks more efficiently. Moreover, advancements in approximate and variational GPs have played a pivotal role in scaling Bayesian optimization, a key global optimization approach. For instance exact GPs scale as $\mathcal{O}(n^3)$ for computational complexity, but using sparse approximations (*e.g.*, inducing points) reduce that complexity to $\mathcal{O}(nm^2)$ or $\mathcal{O}(m^3)$ where m is the number of inducing points. Algorithms that enable uncertainty quantification at scale will be a core need of SDLs moving forward.

6 Education

A large span of educational backgrounds is currently required to drive the development of SDLs, ranging from traditional sciences, technology, engineering, and math (STEM) to humanities. This raises the question of how to approach educating the next generation of students and researchers so that they are prepared to develop and responsibly use automated and self-driving labs. With this section, we hope to promote discussion amongst educators as to how to integrate the ideas and techniques of SDLs into a curriculum and provide resources that will enable such development. This section is organized into three subsections, (1) the topics that researchers in this field should know, (2) mechanisms for teaching them, and (3) thoughts on how to assess success. On a larger scale, we hope to spur discussion as to what should be considered foundational knowledge in higher level education and how we enable future generations of scientists to contribute to and advance these growing sectors of research and discovery.

6.1 Fundamental knowledge underpinning AI-accelerated science

As we have already discussed in great detail, SDLs bridge a broad range of topics such as AI, computing, engineering and automation of experiments. Prospective scientists entering the field of acceleration will typically have one domain of expertise but maybe little to no experience in other domains. Thus, it may be challenging for an individual to master all skills necessary in addition to their experience in their application area. The multidisciplinary nature of the field poses challenges to educators in both deciding on the prerequisites and lecturing for an audience with diverse academic backgrounds and research interests.

The intellectual barrier to entry into AI-accelerated research involves fundamental theoretical knowledge as well as practical skills, both of which must be acquired through training and practice. Specifically, practitioners must understand the fundamental topics of:

- How and when to conceive of scientific research as an iterative workflow, involving the selection and performance of experiments together with the subsequent analysis of results.
- How one can go from an existing gap in knowledge to defining critical bottlenecks that limit the speed at which that gap can be filled.
- The mathematics and computer science underpinning SDLs, *e.g.* statistics, probability theory, linear algebra, programming, automated data analysis, basic machine learning and automated planning/decision-making.
- Lab automation including existing solutions, modular setup of SDLs, communication protocols of automated equipment.
- And the history of AI- and automation-accelerated research.

- Lab automation including existing solutions, modular setup of SDLs, communication protocols of automated equipment.
- And the history of AI- and automation-accelerated research.

In addition, practitioners should have the following skills:

- Data management and curation.
- Algorithmic data processing including scripting the extraction, analysis, and presentation of large datasets from possibly heterogeneous sources.
- Interdisciplinary teamwork combining software, hardware, and domain experts.
- And fluency across these disciplinary intersections for effective communication with diverse researchers.

6.2 Mechanisms for training in AI-accelerated science

There are a number of approaches to learning and teaching about AI-accelerated science that can appeal to a wide array of educational backgrounds, experience levels, and time commitments. We envision a multi-faceted approach to training new students and existing researchers in SDL-accelerated research. Here we provide a list of a few such facets that are organized from lowest barrier to entry to those that require more time, expertise, and resource commitment.

- Freely available videos are a great resource for beginners and the community would benefit from repositories of such videos that allow learners to sort and search to find topics of interest among curated or trusted videos. Such methods have been used systematically in engineering education.[182]
- For students excited about active participation, workshops at conferences are an excellent resource, especially those that are paired with large meetings. For instance, the MRS Data Science Tutorial Organizers hosted two machine learning competitions in recent years, one in fall 2021 focusing on active learning and another in fall 2022 on supervised machine learning.[183]

• More formal courses can be helpful for some learners and be longer and more substantial than workshops and tutorials. The collaborative nature of the AI-accelerated materials community raises the possibility of jointly developing a course that can have a presence at multiple universities and touch on many material or chemical domains. One example of how such a course may be developed collaboratively by many researchers comes from the area of computational chemistry, where a shared course has been developed between multiple institutions.

• As a powerful resource for either advanced users of instructors developing curricula, shared digital resources such as datasets and code can be directly shared. While these often require a certain degree of expertise to incorporate, they provide access to powerful techniques. However, it is necessary to efficiently share these with attribution. One avenue for doing this are repositories of materials informatics resources such as REMI: REsource for Materials Informatics.†

• Student internships in companies that develop robotics hardware and software as well as companies that have long-standing experience in advanced automation will help to transfer existing industry know-how into academic environments and apply it to advance SDL technology.

• As in many fields, there is no substitute for learning with hands-on experience. SDLs provide some unique opportunities for such hands-on learning. Suitable cost-effective SDLs can be adopted by instructors for teaching settings.[184] Alternatively, enthusiastic learners can directly leverage these resources to learn independently. Furthermore, students should be incorporated in existing and currently developed SDLs through thesis projects or research internships, to be exposed to SDL technology as early as possible, and also to transfer knowledge and know-how between labs.

A number of these examples show the value of openly sharing software, hardware, and data. As such, we view this as a strong encouragement to continue and expand the practice for education in AI-accelerated research.

6.3 Assessment of educational activities

It is important to evaluate the effectiveness of coordinated training and education efforts in the community. Moreover, the insight from these assessments should be shared to collectively improve practices. While there is no single metric that can address all facets of education activities, the following avenues stand out as promising processes to gather and act on feedback.

• Workshops and tutorials can be evaluated by quantifying attendance, soliciting feedback, and student outcomes. While attendance itself is not especially important as different venues lend themselves to different scales, the fluctuation of this quantity over time can speak to the impact and reputation of the event. Feedback on educational events should always be solicited following the event when thoughts are still fresh in the minds of the attendees. This can be through informal short surveys administered online. For classes with certificates, filling out feedback can be made a prerequisite for receiving the certificate. Actual learning outcomes are more challenging but more important to measure. The key question is whether the educational efforts led to research action. While open to confirmation bias, a starting point would be to interview successful students to learn what facets of their education were most effective.

• Open-source educational resources can be assessed through the degree to which they are accessed. Basic analytics can help determine how many new users are present and how long these users spend with the resources. Such results can help guide the refinement of further resources. That said, there is a difference between useful, popular, and correct, which means that raw user numbers do not tell the whole story. It would be useful to couple such metrics with expert-curated recommendations to highlight effective resources.

• One avenue that has already been impactful in machine learning and computer science is the use of competitions. In addition to generating excitement about the field, these can serve as an avenue for assessment in evaluating the results from typical participants. In addition to competitions with a broad scope, in-class competitions have several advantages,[185] including encouraging learning by doing and providing an easy-to-implement platform for evaluating student progress.

Just as self-driving labs represent an iterative cycle in which experiments are performed and the outcome is used to learn and choose subsequent experiments, it is important to view pedagogy along the same lines and use assessment to build on and improve past efforts. This reflects the dynamic and evolving nature of this field and our expectation that it will grow and evolve in the years to come.

7 Ethics and community

Research on and with autonomous experiments could have a powerful impact on society. A development and deployment process that does not include careful planning, broad consultation, competent execution and ongoing adaptation might create long-term harms that outweigh SDLs' benefits. Although anticipating all potential complications is impossible, exploring possible problems—as well as solutions and mitigations—early and frequently could reduce the expected cost of such issues. The space of ethical considerations relevant to SDLs is too broad to canvass comprehensively here, but this section highlights a few key categories.

First, by lowering the cost and increasing the accessibility of scientific R&D, SDLs could profilerate destructive capabilities as well as research progress. For example, these facilities might enable actors with malicious intent to develop hazardous materials of biological, chemical and nuclear origin. Research has already shown that relatively simple machine learning methods can generate novel toxins that are potentially more fatal than previously-known substances and that do not feature on chemical controls and watch lists, creating new governance challenges.[186] SDLs, if not managed carefully, could enable further experimentation along these lines and possibly the large-scale production of dangerous substances.[187] SDL

† https://pages.nist.gov/remi/.

governance will have to balance responsiveness to these concerns with sensitivity to researcher privacy. Physical and cyber security will also play a critical role, since poorly-secured labs, regardless of the soundness of their governance, could be vulnerable to hijacking by hostile actors.

Second, neglect and AI safety failures could lead to risks similar to those of malicious intent. Equipment malfunctions, insufficient cleaning and maintenance, poor storage practices and so forth might inadvertently create harmful substances, for instance by contaminating a procedure that would otherwise be safe. This issue is of course not unique to SDLs—is a general lab safety concern—but the absence of regular human supervision removes a critical auditing layer. Relatedly, the increased role of automated systems in SDLs raises the importance of addressing AI safety issues: a powerful, unaligned system prone to misinterpreting user requests or unfamiliar with a comprehensive range of lab safety practices, standards and risks could, given access to a well-stocked scientific facility, do tremendous damage by, for instance, mixing volatile substances or developing and dispersing toxins or pathogens[188,189] These are among the scenarios that most concern AI safety researchers.[190,191]

Third, SDLs, like prior automation, could have adverse social and political consequences. Historical parallels from the industrial revolution to the recent rise of the gig economy show that these costs can include unemployment and underemployment, reduced mental health, a sense of diminished community and security, and inequitable economic impacts. These problems in turn can trigger escalatory cycles of political backlash, and can result in regulation that slows technical progress. To minimize the likelihood and impact of such dangers, the SDL community should not only study technical aspects of the technology, but also investigate adjacent social systems and relevant historical precedents. To this end, SDL developers should partner with economists, historians, social activists and stakeholders likely to affected by the development and deployment of these technologies. However, given the impossibility of perfectly predicting complex social systems, individuals working on SDLs should also prioritize ongoing monitoring of and adaptation to unexpected developments.

Fourth, as a form of economic activity that involves industrial components and processes, SDLs could have negative environmental impacts. For instance, a rise in the use of heavy machinery as the costs of experimentation drop might raise carbon emissions, and increasing chemical R&D might damage local ecosystems. Guarding against these risks will, like addressing social and political consequences, require a combination of foresight (in this case, making the most of the growing environmental science literature) and responsiveness to unexpected developments. Unlike social and political risks, however, environmental issues stand to benefit fairly directly from the research that SDLs could enable.[192,193] We encourage the SDL community to make climate-related topics a top area of investigation.

Fifth, even if none of the above risks transpire, SDLs could cause harm *via* incurring inequitable impacts, for instance by concentrating economic gains, additional research prestige, *etc.* amongst privileged groups. We encourage building working environments that promote equity, support diversity and require inclusion. As a uniquely interdisciplinary community, we celebrate the strengths that derive from differences. In creating working environments where all are welcomed, valued, respected, and invited to participate fully, we will accelerate SDLs' positive impact. The community should incorporate equity and justice in the selection and implementation of education, research, development, policy, and commercialization. This includes openly distributing the results of early-stage research and development in line with FAIR practices, as well as continued commitment to ethical and reproducible research.

Identifying risks does not solve them, but it represents an important first step. Ideally, SDLs could create products and processes that actively counter these dangers, *e.g.* by enabling people to concentrate on the safest, most enjoyable aspects of discovery, contributing to climate change mitigation and adaptation, and making scientific knowledge and experimentation more equitable and accessible. However, this will not happen without active guidance; realizing this vision will require a concerted effort from the SDL community.

8 Conclusion

The field of AE and SDLs has the potential to power a new revolution in the pace and nature of scientific discovery. As with any revolution, the community shapes the process and outcome. At this pivotal nascent moment, we acknowledge that there are rich opportunities at every intersection of our community, from software to hardware to education and ethics. It is crucial that we take deliberate action to ensure that our collective progress as fruitful and positive as possible. Important considerations that must be addressed include how we acquire, store, manage, and share our data, as well as how we develop, disseminate, and scale our hardware and software solutions. These innovations do not happen in a vacuum and, as such, we have also highlighted the ethical implications of the field and future education and community needs. Born out of a diverse discourse and community feedback, we hope this perspective will provide guidelines, encouragement, and facilitate community building.

Disclaimer

These opinions, recommendations, findings, and conclusions do not necessarily reflect the views or policies of NIST or the United States Government. Certain equipment, instruments, software, or materials are identified in this paper for informational purposes. Such identification is not intended to imply recommendation or endorsement of any product or service by the authors or their respective institutions, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Data availability

As this is a [Review/Perspective article], no primary research results, data, software or code have been included.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We are grateful to our peers for their gracious and varied contributions to this work. In particular we would like to thank Tantum Collins for his insights around the ethics of AI. As many ideas and pieces were shared across various media, the organizing authors have been listed first for correspondence, followed by an alphabetical list of all other contributors. P. M. used resources of the National Synchrotron Light Source II, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Brookhaven National Laboratory under Contract No. DE-SC0012704, and resources of a BNL Laboratory Directed Research and Development (LDRD) project 23-039 "Extensible robotic beamline scientist for self-driving total scattering studies". P. F. and N. H. acknowledge support by the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21001B (German-Canadian Materials Acceleration Center). P. F. and L. T. acknowledge support by the Federal Ministry of Education and Research (BMBF) under Grant No. 01DM21002A (FLAIM). C. S. acknowledges financial support from VILLUM FONDEN (Grant No. 50405). B. B., I. T. F., and M. H. acknowledge financial support from Argonne National Laboratory, a U.S. Department of Energy Office of Science laboratory operated under Contract No. DE-AC02-06CH11357.

Notes and references

1 K. M. Roccapriore, S. V. Kalinin and M. Ziatdinov, *Advanced Science*, 2022, **9**, 2203422.
2 M. Abolhasani and E. Kumacheva, *Nat. Synth.*, 2023, 1–10.
3 B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams and A. G. Doyle, *Nature*, 2021, **590**, 89–96.
4 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends Chem.*, 2019, **1**, 282–291.
5 C. W. Coley, N. S. Eyke and K. F. Jensen, *Angew. Chem., Int. Ed.*, 2020, **59**, 22858–22893.
6 Y. Xie, K. Sattari, C. Zhang and J. Lin, *Prog. Mater. Sci.*, 2023, **132**, 101043.
7 M. M. Flores-Leonar, L. M. Mejía-Mendoza, A. Aguilar-Granda, B. Sanchez-Lengeling, H. Tribukait, C. Amador-Bedolla and A. Aspuru-Guzik, *Curr. Opin. Green Sustainable Chem.*, 2020, **25**, 100370.
8 B. Goldman, S. Kearnes, T. Kramer, P. Riley and W. P. Walters, *J. Med. Chem.*, 2022, **65**, 7073–7087.
9 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, et al., *Nat. Rev. Mater.*, 2018, **3**, 5–20.
10 J. H. Montoya, M. Aykol, A. Anapolsky, C. B. Gopal, P. K. Herring, J. S. Hummelshøj, L. Hung, H.-K. Kwon, D. Schweigert, S. Sun, S. K. Suram, S. B. Torrisi, A. Trewartha and B. D. Storey, *Applied Physics Reviews*, 2022, **9**, 011405.
11 X. Peng and X. Wang, *MRS Bull.*, 2023, 179–185.
12 M. L. Green, B. Maruyama and J. Schrier, *Applied Physics Reviews*, 2022, **9**, 030401.
13 A. Choudhury, *Arch. Comput. Methods Eng.*, 2021, **28**, 3361–3381.
14 H. Narayanan, F. Dingfelder, A. Butté, N. Lorenzen, M. Sokolov and P. Arosio, *Trends Pharmacol. Sci.*, 2021, **42**, 151–165.
15 M. Seifrid, J. Hattrick-Simpers, A. Aspuru-Guzik, T. Kalil and S. Cranford, *Matter*, 2022, **5**, 1972–1976.
16 E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, C. P. Gomes, J. M. Gregoire, A. Mehta, J. Montoya, E. Olivetti, C. Park, E. Rotenberg, S. K. Saikin, S. Smullin, V. Stanev and B. Maruyama, *Matter*, 2021, **4**, 2702–2726.
17 A.-G. Alán, K. Persson and H. Tribukait-Vasconcelos, *Materials Acceleration Platform: Accelerating Advanced Energy Materials Discovery by Integrating High-Throughput Methods with Artificial Intelligence*, Mission innovation technical report, 2018.
18 H. Kitano, *npj Syst. Biol. Appl.*, 2021, **7**, 29.
19 D. Shapere, *Philos. Rev.*, 1964, **73**, 383–394.
20 B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick and A. I. Cooper, *Nature*, 2020, **583**, 237–241.
21 P. M. Maffettone, S. Campbell, M. D. Hanwell, S. Wilkins and D. Olds, *Cell Rep. Phys. Sci.*, 2022, **3**, 101112.
22 R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, 2021.
23 M. Abolhasani and E. Kumacheva, *Nat. Synth.*, 2023, 1–10.
24 M. Abolhasani and K. A. Brown, *MRS Bull.*, 2023, 134–141.
25 C. Arnold, *Nature*, 2022, **606**, 612–613.
26 C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, L. Jesus, R. Berriel, T. M. Paixão, F. Mutz, L. de Paula Veronese, T. Oliveira-Santos and A. F. De Souza, *Expert Syst. Appl.*, 2021, **165**, 113816.
27 F. Delgado-Licona and M. Abolhasani, *Adv. Intell. Syst.*, 2022, 2200331.
28 C. J. Leong, K. Y. A. Low, J. Recatala-Gomez, P. Q. Velasco, E. Vissol-Gaudin, J. D. Tan, B. Ramalingam, R. I. Made, S. D. Pethe, S. Sebastian, Y.-F. Lim, Z. H. J. Khoo, Y. Bai, J. J. W. Cheng and K. Hippalgaonkar, *Matter*, 2022, **5**, 3124–3134.
29 D. Lowe, *The Downside of Chemistry Automation*, 2019, **https://www.science.org/content/blog-post/downside-chemistry-automation**.
30 B. P. MacLeod, F. G. L. Parlane, A. K. Brown, J. E. Hein and C. P. Berlinguette, *Nat. Mater.*, 2021, 722–726.
31 B. Maruyama, J. Hattrick-Simpers, W. Musinski, L. Graham-Brady, K. Li, J. Hollenbach, A. Singh and M. L. Taheri, *MRS Bull.*, 2022, **47**, 1154–1164.

32 M. May, *Nature*, 2019, **569**, 587–588.
33 J. M. Perkel, *Nature*, 2017, **542**, 125–126.
34 F. Rahmanian, J. Flowers, D. Guevarra, M. Richter, M. Fichtner, P. Donnely, J. M. Gregoire and H. S. Stein, *Adv. Mater. Interfaces*, 2022, **9**, 2101987.
35 M. Seifrid, R. Pollice, A. Aguilar-Granda, Z. Morgan Chan, K. Hotta, C. T. Ser, J. Vestfrid, T. C. Wu and A. Aspuru-Guzik, *Acc. Chem. Res.*, 2022, **55**(17), 2454–2466.
36 M. Seifrid, J. Hattrick-Simpers, A. Aspuru-Guzik, T. Kalil and S. Cranford, *Matter*, 2022, **5**, 1972–1976.
37 R. Vescovi, R. Chard, N. Saint, B. Blaiszik, J. Pruyne, T. Bicer, A. Lavens, Z. Liu, M. E. Papka, S. Narayanan, N. Schwarz, K. Chard and I. Foster, *Linking Scientific Instruments and HPC: Patterns, Technologies, Experiences*, 2022.
38 R. Vescovi, R. Chard, N. D. Saint, B. Blaiszik, J. Pruyne, T. Bicer, A. Lavens, Z. Liu, M. E. Papka, S. Narayanan, N. Schwarz, K. Chard and I. T. Foster, *Patterns*, 2022, **3**, 100606.
39 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al., *Sci. Data*, 2016, **3**, 160018.
40 L. C. Brinson, L. M. Bartolo, B. Blaiszik, D. Elbert, I. Foster, A. Strachan and P. W. Voorhees, *MRS Bull.*, 2023, **48**, 1–5.
41 W. Dempsey, I. Foster, S. Fraser and C. Kesselman, *Harvard Data Sci. Rev.*, 2022, **4**(3), DOI: **10.1162/99608f92.44d21b86**, **https://hdsr.mitpress.mit.edu/pub/qjpg8oik**.
42 I. M. Pendleton, G. Cattabriga, Z. Li, M. A. Najeeb, S. A. Friedler, A. J. Norquist, E. M. Chan and J. Schrier, *MRS Commun.*, 2019, **9**, 846–859.
43 B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard and I. Foster, *MRS Commun.*, 2019, **9**, 1125–1133.
44 D. Allan, T. Caswell, S. Campbell and M. Rakitin, *Synchrotron Radiat. News*, 2019, **32**, 19–22.
45 W. Mahnke, S.-H. Leitner and M. Damm, *OPC Unified Architecture*, Springer, Berlin, Germany, 2009th edn, 2009.
46 H. Bär, R. Hochstrasser and B. Papenfuß, *J. Lab. Autom.*, 2012, **17**, 86–95.
47 R. Vescovi, R. Chard, N. D. Saint, B. Blaiszik, J. Pruyne, T. Bicer, A. Lavens, Z. Liu, M. E. Papka, S. Narayanan, N. Schwarz, K. Chard and I. Foster, *Patterns*, 2022, **3**, 100606.
48 B. G. Pelkie and L. D. Pozzo, *Digital Discovery*, 2023, 544–556.
49 D. Srivastava, *Deepanshs/Csdmpy: V0.4.1*, Zenodo, 2021.
50 H. Gong, J. He, X. Zhang, L. Duan, Z. Tian, W. Zhao, F. Gong, T. Liu, Z. Wang, H. Zhao, W. Jia, L. Zhang, X. Jiang, W. Chen, S. Liu, H. Xiu, W. Yang and J. Wan, *Sci. Data*, 2022, **9**, 787.
51 D. J. Srivastava, T. Vosegaard, D. Massiot and P. J. Grandinetti, *PLoS One*, 2020, **15**, e0225953.
52 L. Wilbraham, S. H. M. Mehr and L. Cronin, *Acc. Chem. Res.*, 2021, **54**, 253–262.
53 R. Duke, V. Bhat and C. Risko, *Chem. Sci.*, 2022, **13**, 13646–13656.
54 The Minerals Metals & Materials Society (TMS), *Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering*, TMS, Pittsburgh, PA, 2017.
55 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
56 S. R. Hall, J. D. Westbrook, N. Spadaccini, I. D. Brown, H. J. Bernstein and B. McMahon, in *Specification of the Crystallographic Information File (CIF)*, ed. S. R. Hall and B. McMahon, Springer, Dordrecht, Netherlands, 2005, pp. 20–36.
57 E. Breck, M. Zinkevich, N. Polyzotis, S. Whang and S. Roy, *Proceedings of SysML*, 2019, 334–347.
58 L. Teytelman, A. Stoliartchouk, L. Kindler and B. L. Hurwitz, *PLoS Biol.*, 2016, **14**, e1002538.
59 P. Tremouilhac, C.-L. Lin, P.-C. Huang, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, R. Ulrich, B. Neumair, A. Streit, et al., *Angew. Chem., Int. Ed.*, 2020, **59**, 22771–22778.
60 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., *APL Mater.*, 2013, **1**, 011002.
61 R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. B. Nardelli and S. Curtarolo, *Comput. Mater. Sci.*, 2014, **93**, 178–192.
62 B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke and I. Foster, *Jom*, 2016, **68**, 2045–2052.
63 B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard and I. Foster, *MRS Commun.*, 2019, **9**, 1125–1133.
64 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *Jom*, 2013, **65**, 1501–1509.
65 C. Draxl and M. Scheffler, *J. Phys.: Mater.*, 2019, **2**, 036001.
66 P. Herring, C. B. Gopal, M. Aykol, J. H. Montoya, A. Anapolsky, P. M. Attia, W. Gent, J. S. Hummelshøj, L. Hung, H.-K. Kwon, et al., *SoftwareX*, 2020, **11**, 100506.
67 M. Apted and S. Murphy, *Choosing a CI/CD approach: AWS Services with BigHat Biosciences*, 2021, **https://aws.amazon.com/blogs/devops/choosing-ci-cd-aws-services-bighat-biosciences/**.
68 R. Chard, J. Pruyne, K. McKee, J. Bryan, B. Raumann, R. Ananthakrishnan, K. Chard and I. T. Foster, *Future Gener. Comput. Syst.*, 2023, **142**, 393–409.
69 P. Maffettone, D. Allan, S. I. Campbell, M. R. Carbone, T. Caswell, B. L. DeCost, D. Gavrilov, M. Hanwell, H. Joress, J. Lynch, B. Ravel, S. Wilkins, J. Wlodek and D. Olds, *AI for Accelerated Materials Design NeurIPS 2022 Workshop*, 2022.
70 P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73–76.
71 C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy,

W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, *Nature*, 2020, **585**, 357–362.
72 C. H. Ward, *Integr. Mater. Manuf. Innov.*, 2015, **4**, 190–191.
73 A. R. Akbashev and S. V. Kalinin, *Nat. Mater.*, 2023, **22**, 270–271.
74 J. Hohlbein, B. Diederich, B. Marsikova, E. G. Reynaud, S. Holden, W. Jahr, R. Haase and K. Prakash, *Nat. Methods*, 2022, **19**, 1020–1025.
75 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson and A. Aspuru-Guzik, *Nat. Rev. Mater.*, 2018, **3**, 5–20.
76 B. DeCost, H. Joress, S. Sarker, A. Mehta and J. Hattrick-Simpers, *Towards Automated Design of Corrosion Resistant Alloy Coatings with an Autonomous Scanning Droplet Cell*, 2022.
77 H. Joress, M. L. Green, I. Takeuchi and J. R. Hattrick-Simpers, *Encyclopedia of Materials: Metals and Alloys*, Elsevier, Oxford, 2022, pp. 353–371.
78 J. Vasquez, *Jubilee: A Toolchanging Homage To 3d Printer Hackers Everywhere*, 2019, **https://hackaday.com/2019/11/14/jubilee-a-toolchanging-homage-to-3d-printer-hackers-everywhere/**.
79 R. D. King, J. Rowland, S. G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L. N. Soldatova, A. Sparkes, K. E. Whelan and A. Clare, *Science*, 2009, **324**, 85–89.
80 P. A. Beaucage and T. B. Martin, *Chem. Mater.*, 2023, **35**, 846–852.
81 B. P. MacLeod, F. G. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. Yunker, M. B. Rooney, J. R. Deeth, et al., *Sci. Adv.*, 2020, **6**, eaaz8867.
82 H. Fakhruldeen, G. Pizzuto, J. Glawucki and A. I. Cooper, *IEEE Int. Conf. Robot. Autom.*, 2022, 6013–6019.
83 M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler and A. Y. Ng, *ICRA Workshop on Open Source Software*, 2009, p. 5.
84 L. R. Dalesio, A. J. Kozubal and M. R. Kraimer, *International conference on accelerator and large experimental physics control systems*, 1991.
85 S. Pinter and A. Yoaz, *Proceedings of the 29th Annual IEEE/ACM International Symposium on Microarchitecture. MICRO*, 1996, vol. 29, pp. 214–225.
86 B. P. MacLeod, F. G. L. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhsousi, D. J. Dvorak, H. N. Chiu, C. E. B. Waizenegger, K. Ocean, M. Mokhtari and C. P. Berlinguette, *Nat. Commun.*, 2022, **13**, 995.
87 H. Joress, B. L. DeCost, S. Sarker, T. M. Braun, S. Jilani, R. Smith, L. Ward, K. J. Laws, A. Mehta and J. R. Hattrick-Simpers, *ACS Comb. Sci.*, 2020, **22**, 330–338.
88 J. S. Weaver, A. L. Pintar, C. Beauchamp, H. Joress, K.-W. Moon and T. Q. Phan, *Mater. Des.*, 2021, **209**, 109969.
89 B. P. MacLeod, F. G. L. Parlane, A. K. Brown, J. E. Hein and C. P. Berlinguette, *Nat. Mater.*, 2022, **21**, 722–726.
90 G. Pizzuto, J. De Berardinis, L. Longley, H. Fakhruldeen and A. I. Cooper, *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–7.
91 J. X.-Y. Lim, D. Leow, Q.-C. Pham and C.-H. Tan, *IEEE Trans. Autom. Sci. Eng.*, 2021, **18**, 2185–2190.
92 H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti and A. Garg, *Seeing Glass: Joint Point Cloud and Depth Completion for Transparent Objects*, 2021.
93 Y. R. Wang, Y. Zhao, H. Xu, S. Eppel, A. Aspuru-Guzik, F. Shkurti and A. Garg, *MVTrans: Multi-View Perception of Transparent Objects*, 2023.
94 H. Zwirnmann, D. Knobbe, U. Culha and S. Haddadin, *Dual-Material 3D-Printed PaCoMe-Like Fingers for Flexible Biolaboratory Automation*, 2023, **https://arxiv.org/abs/2302.03644**.
95 N. Yoshikawa, K. Darvish, A. Garg and A. Aspuru-Guzik, *Digital pipette: Open hardware for liquid transfer in self-driving laboratories*, 2023.
96 S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone and L. Cronin, *Science*, 2019, **363**, eaav2211.
97 J. S. Manzano, W. Hou, S. S. Zalesskiy, P. Frei, H. Wang, P. J. Kitson and L. Cronin, *Nat. Chem.*, 2022, **14**, 1311–1318.
98 S. Eggert, P. Mieszczanek, C. Meinert and D. W. Hutmacher, *HardwareX*, 2020, **8**, e00152.
99 A. Faiña, B. Nejati and K. Stoy, *Appl. Sci.*, 2020, **10**, 814.
100 C. J. Forman, *PLoS Biol.*, 2020, **18**, e3000858.
101 R. Keesey, A. Tiihonen, A. E. Siemenn, T. W. Colburn, S. Sun, N. T. P. Hartono, J. Serdy, M. Zeile, K. He, C. A. Gurtner, A. C. Flick, C. Batali, A. Encinas, R. R. Naik, Z. Liu, F. Oviedo, I. M. Peters, J. Thapa, S. I. P. Tian, R. H. Dauskardt, A. J. Norquist and T. Buonassisi, *Digital Discovery*, 2023, **2**, 422–440.
102 R. Keesey, R. LeSuer and J. Schrier, *HardwareX*, 2022, **12**, e00319.
103 K. Laganovska, A. Zolotarjovs, M. Vázquez, K. M. Donnell, J. Liepins, H. Ben-Yoav, V. Karitans and K. Smits, *HardwareX*, 2020, **7**, e00108.
104 L. D. Pozzo, *J. Open Hardw.*, 2021, **5**, 6.
105 M. A. Soldatov, V. V. Butova, D. Pashkov, M. A. Butakova, P. V. Medvedev, A. V. Chernov and A. V. Soldatov, *Nanomaterials*, 2021, **11**, 619.
106 R. P. Hertzberg and A. J. Pope, *Curr. Opin. Chem. Biol.*, 2000, **4**, 445–451.
107 B. J. Battersby and M. Trau, *Trends Biotechnol.*, 2002, **20**, 167–173.
108 R. A. Potyrailo and E. J. Amis, *High-throughput analysis: a tool for combinatorial materials science*, Springer Science & Business Media, 2012.
109 V. Karthik, K. Kasiviswanathan and B. Raj, *Miniaturized testing of engineering materials*, CRC Press, 2016.
110 A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, et al., *Science*, 2015, **347**, 49–53.

111 R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Han, K. G. Reyes, A. Amassian and M. Abolhasani, *Adv. Mater.*, 2020, **32**, 2001626.
112 F. Delgado-Licona and M. Abolhasani, *Adv. Intell. Syst.*, 2023, 2200331.
113 A. A. Volk, Z. S. Campbell, M. Y. Ibrahim, J. A. Bennett and M. Abolhasani, *Annu. Rev. Chem. Biomol. Eng.*, 2022, **13**, 45–72.
114 I. Takeuchi, J. Lauterbach and M. J. Fasolka, *Mater. Today*, 2005, **8**, 18–26.
115 A. Ludwig, *npj Comput. Mater.*, 2019, **5**, 70.
116 P. Nikolaev, D. Hooper, N. Perea-Lopez, M. Terrones and B. Maruyama, *ACS Nano*, 2014, **8**, 10214–10222.
117 W. Feng, E. Ueda and P. A. Levkin, *Adv. Mater.*, 2018, **30**, 1706111.
118 A. Rosenfeld, C. Oelschlaeger, R. Thelen, S. Heissler and P. A. Levkin, *Mater. Today Bio*, 2020, **6**, 100053.
119 M. Seifermann, P. Reiser, P. Friederich and P. Levkin, *High-throughput synthesis and machine learning assisted design of photodegradable hydrogels*, 2023.
120 V. Saygin, B. Xu, S. B. Andersson and K. A. Brown, *ACS Appl. Mater. Interfaces*, 2021, **13**, 14710–14717.
121 K. A. Brown, *Matter*, 2022, **5**, 3112–3123.
122 T. Zech, P. Claus and D. Hönicke, *Chimia*, 2002, **56**, 611.
123 M. Lohrengel, A. Moehring and M. Pilaski, *Fresenius' J. Anal. Chem.*, 2000, **367**, 334–339.
124 K. J. Jenewein, G. D. Akkoc, A. Kormányos and S. Cherevko, *Chem Catal.*, 2022, **2**, 2778–2794.
125 W. Zhang, R. Wang, F. Luo, P. Wang and Z. Lin, *Chin. Chem. Lett.*, 2020, **31**, 589–600.
126 P. M. Maffettone, J. K. Lynch, T. A. Caswell, C. E. Cook, S. I. Campbell and D. Olds, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 025025.
127 L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. Yunker, J. E. Hein and A. Aspuru-Guzik, *PLoS One*, 2020, **15**, e0229862.
128 D. Salley, G. Keenan, J. Grizou, A. Sharma, S. Martín and L. Cronin, *Nat. Commun.*, 2020, **11**, 2771.
129 P. Rajak, A. Krishnamoorthy, A. Mishra, R. Kalia, A. Nakano and P. Vashishta, *npj Comput. Mater.*, 2021, **7**, 108.
130 J. Li, J. Li, R. Liu, Y. Tu, Y. Li, J. Cheng, T. He and X. Zhu, *Nat. Commun.*, 2020, **11**, 2046.
131 A. E. Gongora, B. Xu, W. Perry, C. Okoye, P. Riley, K. G. Reyes, E. F. Morgan and K. A. Brown, *Sci. Adv.*, 2020, **6**, eaaz1708.
132 B. P. MacLeod, F. G. Parlane, C. C. Rupnow, K. E. Dettelbach, M. S. Elliott, T. D. Morrissey, T. H. Haley, O. Proskurin, M. B. Rooney, N. Taherimakhsousi, et al., *Nat. Commun.*, 2022, **13**, 995.
133 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Chem. Sci.*, 2018, **9**, 7642–7655.
134 S. Daulton, M. Balandat and E. Bakshy, *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 2187–2200.
135 J. A. G. Torres, S. H. Lau, P. Anchuri, J. M. Stevens, J. E. Tabora, J. Li, A. Borovika, R. P. Adams and A. G. Doyle, *J. Am. Chem. Soc.*, 2022, **144**, 19999–20007.
136 M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson and E. Bakshy, *Advances in neural information processing systems*, 2020, vol. 33, pp. 21524–21538.
137 Q. Liang, A. E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J. R. Deneault, D. Bash, F. Mekki-Berrada, S. A. Khan, et al., *npj Comput. Mater.*, 2021, **7**, 188.
138 R. W. Epps, A. A. Volk, M. Y. Ibrahim and M. Abolhasani, *Chem*, 2021, **7**, 2541–2545.
139 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
140 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.
141 J. Noh, G. H. Gu, S. Kim and Y. Jung, *Chem. Sci.*, 2020, **11**, 4871–4881.
142 Q. Ai, A. J. Norquist and J. Schrier, *Digital Discovery*, 2022, **1**, 255–265.
143 Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu and J. Hu, *npj Comput. Mater.*, 2020, **6**, 1–7.
144 V. Korolev, A. Mitrofanov, A. Eliseev and V. Tkachenko, *Mater. Horiz.*, 2020, **7**, 2710–2718.
145 Y. Sawada, K. Morikawa and M. Fujii, *arXiv*, 2019, Preprint, arXiv:1910.11499, DOI: **10.48550/arXiv.1910.11499**.
146 R. Xin, E. M. D. Siriwardane, Y. Song, Y. Zhao, S.-Y. Louis, A. Nasiri and J. Hu, *J. Phys. Chem. C*, 2021, **125**, 16118–16128.
147 S. G. Baird, K. M. Jablonka, M. D. Alverson, H. M. Sayeed, M. F. Khan, C. Seegmiller, B. Smit and T. D. Sparks, *JOSS*, 2022, **7**, 4528.
148 S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik and Y. Jung, *ACS Cent. Sci.*, 2020, **6**, 1412–1420.
149 T. Long, N. M. Fortunato, I. Opahle, Y. Zhang, I. Samathrakis, C. Shen, O. Gutfleisch and H. Zhang, *npj Comput. Mater.*, 2021, **7**, 66.
150 J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, *Matter*, 2019, **1**, 1370–1384.
151 Z. Ren, S. I. P. Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun, X. Wang, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung and T. Buonassisi, *Matter*, 2022, **5**, 314–335.
152 T. Xie, X. Fu, O.-E. Ganea, R. Barzilay and T. Jaakkola, *arXiv*, 2022, Preprint, arXiv:2110.06197, DOI: **10.48550/arXiv.2110.06197**.
153 N. Anand and P. Huang, *Advances in Neural Information Processing Systems*, 2018.
154 R. R. Eguchi, C. A. Choe and P.-S. Huang, Ig-VAE: Generative Modeling of Protein Structure by Direct 3D Coordinate Generation, *PLOS Comp. Bio.*, 2022, **18**, 1–18.
155 N. C. Frey, V. Gadepally and B. Ramsundar, *arXiv*, 2022, Preprint, arXiv:2201.12419, DOI: **10.48550/arXiv.2201.12419**.

156 Z. Li, S. P. Nguyen, D. Xu and Y. Shang, *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, Boston, MA, 2017, pp. 1085–1091.
157 W. Gao and C. W. Coley, *J. Chem. Inf. Model.*, 2020, **60**, 5714–5723.
158 C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay and K. F. Jensen, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1608.
159 W. Gao, P. Raghavan and C. W. Coley, *Nat. Commun.*, 2022, **13**, 1075.
160 J. Bradshaw, B. Paige, M. J. Kusner, M. Segler and J. M. Hernández-Lobato, *Advances in Neural Information Processing Systems*, 2019.
161 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
162 Y. Luo, S. Bag, O. Zaremba, A. Cierpka, J. Andreo, S. Wuttke, P. Friederich and M. Tsotsalas, *Angew. Chem., Int. Ed.*, 2022, **61**, e202200242.
163 T. Gupta, M. Zaki and N. A. Krishnan, *npj Comput. Mater.*, 2022, **8**, 102.
164 A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson and A. Jain, *arXiv*, 2022, preprint, arXiv:2212.05238, DOI: **10.48550/arXiv.2212.05238**.
165 A. White, *Paper QA*, 2023.
166 A. M. Bran, S. Cox, A. D. White and P. Schwaller, *ChemCrow: Augmenting Large-Language Models with Chemistry Tools*, 2023.
167 M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich, A. Nigam, et al., *Nat. Rev. Phys.*, 2022, **4**, 761–769.
168 P. Friederich, M. Krenn, I. Tamblyn and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 025027.
169 G. A. Khoury, R. C. Baliban and C. A. Floudas, *Sci. Rep.*, 2011, **1**, 90.
170 R. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, *Adv. Neural. Inf. Process. Syst.*, 2019, **32**, 9240–9251.
171 J. Teufel, L. Torresi, P. Reiser and P. Friederich, *arXiv*, 2022, preprint, arXiv:2211.13236, DOI: **10.48550/arXiv.2211.13236**.
172 F. Oviedo, J. L. Ferres, T. Buonassisi and K. T. Butler, *Acc. Mater. Res.*, 2022, **3**, 597–607.
173 X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski and T. Y.-J. Han, *npj Comput. Mater.*, 2022, **8**, 204.
174 G. Pilania, *Comput. Mater. Sci.*, 2021, **193**, 110360.
175 B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski and T. Y.-J. Han, *npj Comput. Mater.*, 2019, **5**, 108.
176 G. P. Wellawatte, A. Seshadri and A. D. White, *Chem Sci.*, 2022, **13**, 3697–3705.
177 Y. Cheng, R. Matthiesen, S. Farooq, J. Johansen, H. Hu and L. Ma, *Int. J. Prod. Econ.*, 2018, **203**, 239–253.
178 W. E. Deming, *Quality, productivity, and competitive position*, Massachusetts Inst Technology, 1982.
179 A. Ansar and B. Flyvbjerg, *Oxf. Rev. Econ. Policy*, 2022, **38**, 338–368.
180 J. R. Deneault, J. Chang, J. Myung, D. Hooper, A. Armstrong, M. Pitt and B. Maruyama, *MRS Bull.*, 2021, **46**, 566–575.
181 S. Steiner, J. Wolf, S. Glatzel, A. Andreou, J. M. Granda, G. Keenan, T. Hinkley, G. Aragon-Camarasa, P. J. Kitson, D. Angelone and L. Cronin, *Science*, 2019, **363**, eaav2211.
182 E. D. Lindsay and J. R. Morgan, *Eur. J. Eng. Educ.*, 2021, **46**, 637–661.
183 S. Sun, K. Brown and A. G. Kusne, *Matter*, 2022, **5**, 1620–1622.
184 S. G. Baird and T. D. Sparks, *Matter*, 2022, **5**, 4170–4178.
185 M. Gamarra, A. Dominguez, J. Velazquez and H. Páez, *Comput. Appl. Eng. Educ.*, 2021, **30**, 472–482.
186 F. Urbina, F. Lentzos, C. Invernizzi and S. Ekins, *Nat. Mach. Intell.*, 2022, **4**, 189–191.
187 R. J. Hickman, P. Bannigan, Z. Bao, A. Aspuru-Guzik and C. Allen, *Matter*, 2023, **6**, 1071–1081.
188 A. Turchin and D. Denkenberger, *AI Soc.*, 2020, **35**, 147–163.
189 J. O'Brien and C. Nelson, *Health Secur.*, 2020, **18**, 219–227.
190 A. Koehler and B. Hilton, Preventing catastrophic pandemics, 2023, **https://80000hours.org/problem-profiles/preventing-catastrophic-pandemics/**, accessed: Sept 29, 2023.
191 B. Hilton, Preventing an AI-related catastrophe, 2022, **https://80000hours.org/problem-profiles/artificial-intelligence/**, accessed: Sept 29, 2023.
192 E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith and A. Curioni, *npj Comput. Mater.*, 2022, **8**, 84.
193 R. Giro, M. Elkaref, H. Hsu, N. Herr, G. De Mel and M. Steiner, *March Meeting*, 2023.