Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2024, **3**, 23

Accelerated chemical science with AI

Seoin Back, *[a] Alán Aspuru-Guzik,†[bc] Michele Ceriotti, [d] Ganna Gryn'ova, [ef] Bartosz Grzybowski, [ghi] Geun Ho Gu,[j] Jason Hein, [k] Kedar Hippalgaonkar, [lm] Rodrigo Hormázabal, [n] Yousung Jung, †[op] Seonah Kim, [q] Woo Youn Kim, [r] Seyed Mohamad Moosavi,[s] Juhwan Noh,[t] Changyoung Park,[n] Joshua Schrier, [u] Philippe Schwaller,[v] Koji Tsuda, [wxy] Tejs Vegge, [z] O. Anatole von Lilienfeld †[caaab] and Aron Walsh [acad]

In light of the pressing need for practical materials and molecular solutions to renewable energy and health problems, to name just two examples, one wonders how to accelerate research and development in the chemical sciences, so as to address the time it takes to bring materials from initial discovery to commercialization. Artificial intelligence (AI)-based techniques, in particular, are having a transformative and accelerating impact on many if not most, technological domains. To shed light on these questions, the authors and participants gathered in person for the ASLLA Symposium on the theme of 'Accelerated Chemical Science with AI' at Gangneung, Republic of Korea. We present the findings, ideas, comments, and often contentious opinions expressed during four panel discussions related to the respective general topics: 'Data', 'New applications', 'Machine learning algorithms', and 'Education'. All discussions were recorded, transcribed into text using Open AI's Whisper, and summarized using LG AI Research's EXAONE LLM, followed by revision by all authors. For the broader benefit of current researchers, educators in higher education, and academic bodies such as associations, publishers, librarians, and companies, we provide chemistry-specific recommendations and summarize the resulting conclusions.

Received 25th October 2023
Accepted 6th December 2023

DOI: 10.1039/d3dd00213f

rsc.li/digitaldiscovery

[a]*Department of Chemical and Biomolecular Engineering, Institute of Emergent Materials, Sogang University, Seoul, Republic of Korea. E-mail: sback@sogang.ac.kr*
[b]*Departments of Chemistry, Computer Science, University of Toronto, St. George Campus, Toronto, ON, Canada*
[c]*Acceleration Consortium and Vector Institute for Artificial Intelligence, Toronto, ON, M5S 1M1, Canada*
[d]*Laboratory of Computational Science and Modeling (COSMO), École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
[e]*Heidelberg Institute for Theoretical Studies (HITS gGmbH), 69118, Heidelberg, Germany*
[f]*Interdisciplinary Center for Scientific Computing, Heidelberg University, 69120, Heidelberg, Germany*
[g]*Center for Algorithmic and Robotized Synthesis (CARS), Institute for Basic Science (IBS), Ulsan, Republic of Korea*
[h]*Institute of Organic Chemistry, Polish Academy of Sciences, Warsaw, Poland*
[i]*Department of Chemistry, Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea*
[j]*Department of Energy Engineering, Korea Institute of Energy Technology (KENTECH), Naju, 58330, Republic of Korea*
[k]*Department of Chemistry, University of British Columbia, Vancouver, BC, V6T 1Z1, Canada*
[l]*School of Materials Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore*
[m]*Institute of Materials Research and Engineering, Agency for Science Technology and Research, 2 Fusionopolis Way, 08-03, Singapore 138634, Singapore*
[n]*LG AI Research, Seoul, Republic of Korea*
[o]*Department of Chemical and Biomolecular Engineering, KAIST, Daejeon, Republic of Korea*
[p]*School of Chemical and Biological Engineering, Interdisciplinary Program in Artificial Intelligence, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea*
[q]*Department of Chemistry, Colorado State University, 1301 Center Avenue, Fort Collins, CO 80523, USA*
[r]*Department of Chemistry, KAIST, Daejeon, Republic of Korea*
[s]*Chemical Engineering & Applied Chemistry, University of Toronto, Toronto, Ontario M5S 3E5, Canada*
[t]*Chemical Data-Driven Research Center, Korea Research Institute of Chemical Technology, Daejeon, 34114, Republic of Korea*
[u]*Department of Chemistry, Fordham University, The Bronx, NY 10458, USA*
[v]*Laboratory of Artificial Chemical Intelligence (LIAC) & National Centre of Competence in Research (NCCR) Catalysis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*
[w]*Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba 277-8561, Japan*
[x]*Center for Basic Research on Materials, National Institute for Materials Science, Tsukuba, Ibaraki 305-0044, Japan*
[y]*RIKEN Center for Advanced Intelligence Project, Tokyo 103-0027, Japan*
[z]*Department of Energy Conversion and Storage, Technical University of Denmark, 301 Anker Engelunds vej, Kongens Lyngby, Copenhagen, 2800, Denmark*
[aa]*Departments of Chemistry, Materials Science and Engineering, and Physics, University of Toronto, St George Campus, Toronto, ON, Canada*
[ab]*Machine Learning Group, Technische Universität Berlin and Berlin Institute for the Foundations of Learning and Data, 10587, Berlin, Germany*
[ac]*Department of Materials, Imperial College London, London SW7 2AZ, UK*
[ad]*Department of Physics, Ewha Women's University, Seoul, Republic of Korea*
† The symposium was organized by Yousung Jung, Alán Aspuru-Guzik, and O. Anatole von Lilienfeld. The authors are listed in alphabetical order, except for the first author who took charge of organizing the initial draft written by all co-authors who contributed to different sections.

I. Introduction

With the unprecedented developments of artificial intelligence (AI) technology, chemical science is now entering a radically new era. High-performance computing and virtual screening techniques identify compounds to synthesize for target applications, while automated robotics perform synthesis and characterizations. Additionally, AI suggests new experiments based on the data collected by robotic platforms. In this autonomous laboratory workflow, data science plays a central role in accelerating discoveries in chemical science. The 15th ASLLA Symposium on 'Accelerated Chemical Science with AI' was held at the Korea Institute of Science and Technology (KIST) on the 25–28 September 2022 in Gangneung, Republic of Korea. The workshop brought together 45 participants from around the world to discuss machine learning and automation for the chemical sciences.

In addition to brief talks from the attendees, the conference placed emphasis on panel discussions on the themes of Data, New applications, Machine learning (ML) algorithms, and Education. This Perspective aims to effectively communicate the insights and discussions from these panels to the broader research community.

Numerous recent review and perspective articles have extensively explored the role of data science, ML and AI in various domains of experimental chemistry, including general chemistry,[1] synthetic chemistry and chemical reactions,[2–5] as well as theoretical topics such as chemical compound space exploration[6] and force-field development.[7,8] Additionally, recent reviews have addressed the application of autonomous research systems in materials science,[9–16] organic chemistry,[17–19] inorganic chemistry,[20] porous materials,[21] nanoscience,[22,23] drug formulation[24,25] and biomaterials.[26] Reviews also exist on the topic of self-driving laboratories[27,28] and their low-cost incarnations.[29] While previous recommendations have covered 'best practices' in machine learning for chemistry,[30] including uncertainty quantification,[31] our focus in this Perspective is to present specific recommendations derived from a very rich set of panel discussions by many active researchers in the field rather than reiterating those already discussed themes. We refer the reader to them to more in-depth conversations.

Continuing with the focus on AI, the Whisper program[32] was used to transcribe the panel discussions, and EXAONE[33] was used to generate automated summaries. These algorithmically generated summaries served as the initial drafts of the following works, which we subsequently edited and annotated to ensure clarity. Through this process, it became clear that the panel discussions encompassed overlapping topics, highlighting the shared challenges in the field of AI in chemical science. To underscore these critical challenges, we have reorganized the discussions into common themes: data, new applications, ML algorithms, and education.

II. Data

The quality and scale of data play a pivotal role in developing high-performance ML models. Thus, it is unsurprising that data consistently emerged as a focal point of discussion in all panel sessions. This section aims to offer a concise summary of the insightful discourse on database building, to facilitate the creation of robust and effective machine learning models.

Building better databases

Comprehending the diversity and richness of datasets is vital for developing generalizable ML models.[34] Employing metrics to assess novelty and methods to down-select datasets to eliminate redundant data can serve as remedies in certain cases. When confronted with limited data, hand-crafted descriptors, *e.g.*, coarse-grained descriptors, can be a pragmatic approach in low-data materials discovery tasks.

Furthermore, the availability of high/multi-fidelity benchmark datasets is essential.[35] The benefits of improved training data efficiency when using multi-level learning in chemical compound space have been demonstrated on multiple occasions.[36,37]

When dealing with high-cost, high-fidelity data acquisition, the development of automated workflows that incorporate uncertainty quantification,[38] encompassing both epistemic (model's inability to fit the data distribution) and aleatoric (noise in the data) uncertainties, along with active learning, can be beneficial. Moreover, delta learning methods and incorporation of physical rules as inductive bias within the machine learning algorithms have shown to reduce the size of required data.[39] Furthermore, sampling techniques such as entropic sampling and self-learning population annealing can serve as effective data acquisition strategies. These techniques enable effective weighting of the density of states of the final property in relation to input descriptors, facilitating a comprehensive understanding of different regions of the chemical space. In addition to forward models, observations have suggested that machine learning can also contribute to knowledge-augmented data generation within a discrete and sparse chemical space, particularly in the context of inverse generative design.[12,40,41]

Despite the significant emphasis on developing theoretical strategies for efficiently constructing databases with high-fidelity data, there is a need for additional efforts to ensure that these databases are also user-friendly for interdisciplinary research, *i.e.*, permit even non-domain-expert AI practitioners to interact with the data with minimal intervention. This accessibility is essential for facilitating the test of new algorithmic developments. For instance, when the first large quantum dataset with coordinates and multiple molecular properties for more than 100 000 small organic molecules, QM9, was published in 2014, the total energy was included alongside the free atomic energy. While experts can easily calculate derived properties from this information, such as reaction energies or atomization energies by respectively subtracting the total energies of constitutional isomers or subtracting the free atomic energy from the total energy for any given stoichiometry, this process can pose an unnecessary barrier for non-experts, requiring them to invest time and effort in understanding the underlying definitions of basic chemical properties. Hence, the

development of easy-to-use, web browser-based interfaces for predictive models is of great importance.[42] At the same time, the systematic management of meta-information remains important to ensure the reliability of the constructed database. For example, tools such as AiiDA[43] and NoMaD[44,45] record comprehensive data provenance for 'static' materials simulations.

Finally, it is important to distinguish between multiple datasets categories: smaller, more accurate, and computationally challenging ones that serve specific practical purposes, and datasets specifically designed for benchmarking ML models. This differentiation helps avoid situations where research solely focuses on improving model performance to surpass benchmarks without effectively translating those advancements into practical applications, (overfitting). In this context, dynamic management of databases within the relevant research community proves to be fruitful, as discussed below.

Dynamic community database

For ML algorithms to effectively capture the true complexity of the chemical and materials compound space, it is crucial to overcome biases present in existing databases. This requires a collaborative effort within the community to enable true discovery. To facilitate this goal, the successful implementation of the Common Task Framework (CTF) in the protein community, in conjunction with the Protein Data Bank, has served as a model. The following list outlines key components in datasets that could help to facilitate and foster collaborations between non-experts and experts in solving such problems:

(1) Tasks: clearly defined tasks with precise mathematical interpretation, physical meaning, and chemical purpose.

(2) Accessibility: availability of easily accessible gold-standard datasets in a standardized format, publicly accessible and ready for use.

(3) Metrics: specification of one or more proposed quantitative metrics for each task to measure success.

(4). Evaluation: continuously updated leaderboards that rank state-of-the-art methods and/or data-splits that allow us to better track the model improvements and generalization to out-of-domain (OOD).

(5) Discovery: ability to generate new data as needed, by "Augmenting with chemical knowledge."

Discussions specific to organic reactions databases

While significant progress has been made in the past decade with the emergence of deep learning, the effectiveness of purely data-driven approaches in organic synthesis planning remains to be determined.[46–51]

Large databases of reactions, such as USPTO,[52] Pistachio,[53] Reaxys,[54] and SciFinder,[55] do exist. However, the knowledge contained within these databases falls short regarding quality, diversity, and accessibility. For instance, while USPTO offers open access, its quality may be lower compared to the limited, paid access but higher-quality Reaxys. Reproducibility has also become a point of concern. Additionally, despite the vast number of experimental data available in these reaction databases, only a limited number of reaction types have sufficiently large numbers of examples, typically a few hundred or more, which hinders the development of practical/useful/general AI models.[56] Efforts such as the Open Reaction Database are notable for trying to address these limitations,[57] but remain populated with data from USPTO, with only a few hundred brand-new entries – this poses a question of how to best incentivize synthetic chemists to deposit their results (both positive and negative) into such databases.

Correspondingly, purely data-driven approaches in organic synthesis planning would greatly benefit from maximal training data efficiency when learning. Potential solutions to enhance efficiency include Delta-learning and transfer learning,[58] multi-level learning,[36,37] and few-shot learning techniques.[59] However, the challenge of sparse data becomes particularly pronounced when attempting to identify the scope of "impossible" reactions. If a certain reaction is not listed in a database, one often assumes it cannot happen. But this assumption is mostly true for the types of reactions that happen often. As mentioned earlier, such classes are relatively limited in number and occurrence.[60]

When high-quality datasets are lacking, an alternative, albeit more labor-intensive approach, is expert coding within programs like Chematica or AllChemy. These programs can perform advanced-level synthesis planning, even for complex natural products.[61]

One conclusion reached with broad consensus is the ever-increasing need for improved quality and open databases in all AI-related efforts, not only for reaction data but also for describing rules of chemical reactivity, or the properties of experimentally-available and virtual ligands to find new catalysts.[62,63] Moreover, new featurization schemes may be necessary, particularly ones that consider stereochemical, steric hindrance, and long-range interaction aspects of reactions on complex scaffolds.

Publisher's role

The consensus among many participants was that funding bodies and scientific journals should adopt stringent requirements to foster the open availability, completeness, curation, and standardized formatting of published data. However, determining the specific standards and formats for data remains an ongoing question.

Similarly, it was emphasized that the codes utilized to generate the data should be accessible, unless licensed, and well-documented. Such practices align with the increasing adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) policies in the scientific community.[64,65] Another related challenge is facilitating broader access to proprietary data and/or establishing new repositories where researchers can deposit results of both successful and, importantly, unsuccessful experiments they have conducted.

On the former issue, the panelists agreed that professional non-profit organizations, such as the American Chemical Society (ACS), should consider opening up their extensive repositories or, at the very least, enabling broader academic access. Currently, the SciFinder dataset contains approximately 100 million reactions, yet it remains completely inaccessible for

downloads, severely limiting systematic data analyses. Given its status as a non-profit organization, the ACS is seen to have an ethical obligation to share the datasets it accumulates. While the CAS Common Registry initiative[66] is appreciated, restricted licensing hinders research progress. Thinking more broadly, policies that require disseminating a complete set of data and code as a requirement for publication will help accelerate progress in this field. ACS has started defining research data policy recommendations to achieve this goal.[67] An excellent example of this is RSC's new journal *Digital Discovery*,[68] which has a dedicated data and code reviewer to assess submitted materials for documentation and reproducibility.

III. New applications

Non-equilibrium states

Particular emphasis should be given to developing benchmark training sets that extend beyond equilibrium structures.[69] Such sets, *e.g.*, Transition1x, should enable advancing methods capable of describing dynamics, activated processes, and chemical reaction networks/pathways.[69,70]

Utilizing experimental data

Computational data has played a significant role in AI-driven materials discovery. However, specific critical properties remain inaccessible to these computational approaches regarding real-world applications. To enhance the impact of computational discoveries, it becomes crucial to develop AI methods that can predict the synthesizability of materials.[71] The panel emphasized the importance of establishing an efficient two-way communication channel between theoreticians and experimentalists, as well as the need for integrated autonomous workflows that bridge both domains.[72–75]

Simultaneously, the experimental literature tends to exhibit bias towards "success stories" while failing experiments often go unreported.[76,77] This bias can arise from various factors, such as the superior performance or ease of synthesis and characterization of certain materials for unrelated applications. Consequently, the available data on chemical space for exploration with AI becomes limited, impeding the discovery of genuinely novel systems. From a modeling perspective, a data point perceived as a "failure" in experimental terms can be just as valuable for training models as a data point from a "successful" experiment. Although the concept of a "Journal of Failed Research" remains elusive, the panel suggested that well-documented and openly available metadata from experiments, regardless of outcomes, could address this limitation by providing theoreticians with more extensive and diverse training sets in terms of structure and composition. Moreover, it was highlighted that the context of an experiment matters in defining what constitutes a "failed experiment". For instance, a seemingly failed experiment in one context may actually lead to successful outcomes or the discovery of new compounds in a different context.

During the discussions, the topic of how AI empowers creativity in chemistry was addressed. It was acknowledged that AI is ultimately a tool that accelerates technological advancements and scientific discoveries. The progress made in this field has undeniably expedited the pace of invention. It can also be argued that AI enhances the occurrence of "eureka moments" by facilitating new insights and understanding. This aspect is intricately linked to the exploration of new concepts and the perception of reality. As a creative discipline, chemistry is driven by scientists motivated to uncover novel phenomena, unencumbered by pre-established physical laws. For example, this could involve stabilizing challenging structures, creating unconventional solvation environments, or discovering previously unknown and aesthetically pleasing spin states. Therefore, by leveraging AI to comprehend the existing knowledge and venture into unexplored territories, creative pursuits in chemistry can be truly enhanced. In particular, the question of what it entails for AI to gain scientific understanding based on data is a very relevant question due to the advent of large language models (LLMs) and their applications to chemistry.[78–81] In this context, philosophical and conceptual frameworks like the one proposed by Krenn, *et al.* are needed.[82]

Addressing the multi-scale nature of materials

An example discussed was the need to provide detailed descriptions of the operating conditions of functional materials at their relevant scales,[83–85] and under intended operating conditions.[86] This information is crucial for facilitating inverse design. Much of the work in the field currently follows a bottom-up approach, focusing on the development of machine learning potentials to extend the accessible time- and length scales in atomic-scale simulations. This is necessary to ensure sufficient statistical sampling for retaining predictive accuracy.[87,88] Different materials exhibit limiting processes and reactions at various scales. For instance, catalysts' activity and selectivity[89] and the performance of thermoelectric materials[90,91] are governed at the atomic scale, while durability and reliability involve processes at the meso- to micro-scale or beyond.

The concept of self-driving labs was also discussed,[9] with considerations given to the expenses associated with building, maintaining, and operating such facilities, especially when tailored for testing various optimization algorithms. The idea of "virtual labs" emerged as an alternative, where multi-level modeling is utilized to mimic real-world experiments. For example, in the context of batteries, simulations running on materials could be linked to single-cell and battery-pack configurations to understand the key influences from microstructure to system performance.

There is also a need to approach data dynamically. Building data in a multi-modal capacity to capture different scales or incorporating new experiments and calculations is critical for aiding chemical discovery. It is crucial to emphasize the importance of top-down approaches, starting from the meso/micro-scale phase-field[60] and seamlessly coupling them with ML potentials[92] for autonomous parameterization. Additionally, to enable more meaningful AI-driven discoveries, it is highly desirable to restrict the search to compounds that are easy to synthesize and provide synthesis recipes.

IV. Machine learning algorithms

Given these considerations, the natural question also arises: what other foundational AI advancements, explicitly addressing the needs of science datasets, are yet to be developed? What are the current and future needs? The following non-exhaustive list represents the open challenges discussed as areas of focus for the AI community when interacting with the sciences.

Encoding algorithms for science

Identifying AI algorithm development specific to the sciences (chemistry, physics, materials) that has been driven by clearly defined needs is an important consideration. One notable example is the effect of differentiation and the loss function in the case of organic molecules, as observed in the QM9 dataset. In this dataset, which provides quantum chemical properties for a comprehensive chemical space of small organic molecules, the use of different loss functions for the training and testing sets was necessary to discover new motifs with desired functionality.[35] This requirement arises due to the unique challenge of extrapolating from known molecules to identifying motifs and properties that differ from the original set encountered by the algorithm. This specific example highlights the demand for novel machine learning techniques tailored to the field of chemistry.

An additional example of algorithmic developments, partially inspired by chemical applications, involves the construction of models that incorporate physical symmetries into their structure. In the case of interatomic potentials, since the early stages of this field the crucial insight has been the requirement for models to be exactly invariant to rotations, translations, and atom index permutations.[93] More recently, these ideas have been expanded to create physics inspired models that build upon covariant features/representations, an extension motivated by the widespread presence of vectorial and tensorial targets in quantum chemistry.[94] It is noteworthy that these developments have progressed independently and in parallel with similar efforts in computer science,[95] albeit formulated using different terminology and with less mathematical generality.

During the panel discussions, intriguing questions were raised regarding the potential integration of data-centered and expert methods and the extent to which this integration could be achieved.[96,97] Hybrid approaches were proposed as a means to leverage the encoded knowledge of experts while maintaining the flexibility and adaptability of data-driven approaches. It was also observed that the raw reaction rules derived from either of these approaches can be significantly enhanced through further refinement using quantum mechanical (QM) or molecular mechanical (MM) calculations. For instance, MM methods can be employed to calculate strains and estimate the applicability of reaction rules to cyclization reactions.[98]

Another notable example of a hybrid approach involves breaking down the barriers between different methodologies. This includes merging electronic structure theory and machine learning[99] or creating a unified framework that combines simulations and experimental data.[100] Such models have the potential to learn by effectively integrating diverse sources of information.

Going beyond the interpolative nature of machine learning

In the pursuit of discovering crystals or molecules with new functionalities or improved properties, enhancing the extrapolative performance of machine learning models becomes crucial. However, due to the interpolative nature of ML models, accurately predicting data from domains outside the training data distribution remains a challenge.[101] One intriguing and challenging topic discussed was the development of AI techniques that consider the minimum amount of information necessary to learn everything from the system. Additionally, there was a significant focus on the necessity and development of multi-objective optimizations for new materials discovery.[102,103]

Considering these fundamental AI advancements for enabling chemical discovery, it was noted that most multi-objective, multi-fidelity constrained problems addressed in self-driving labs today tend to prioritize higher performance based on predefined objectives. However, to advance chemistry knowledge, algorithms need to be further tailored for interpretability, extrapolation to learn new science, and hypothesis testing, which fundamentally require different approaches. A recent example involves dedicated exploration of the Pareto front, allowing the extraction of local correlations with near-optimal performance to aid in result understanding.[104,105]

The subsequent topic of discussion revolved around using the acceleration and discovery of new molecules/materials successfully validated in the lab as metrics of success in applying machine learning in chemistry. However, going beyond the speed of material development, true discovery of new concepts,[82] such as topological materials, remains elusive. This led to the question of exploring deeper paths in AI to unlock such possibilities.[106] One potential avenue is considering an automatic system that generates novel questions, although formulating the problems is typically within the domain of human experts. In scientific discovery, anomalies or outliers often lead to new findings. Optimization algorithms are already designed to find regions of high uncertainty in the parameter space, which are often unexplored. Rewarding data points in those regions, even if only a small percentage results in actual discoveries, can lead to the real discovery of new phenomena. Additionally, digitizing existing knowledge in chemistry and creating a comprehensive corpus of our current understanding can help define a concept of "known unknowns" for AI, making the idea less vague and facilitating exploration beyond what is already known. An example was shared regarding an automated robotic system developed by David MacMillan's group at Princeton University, which achieved "accelerated serendipity" by assembling molecules with no known history of interactions and rewarding accidental reactivity.[107] This approach resulted in discovering new reactions or improved methods for existing reactions. Furthermore, emphasizing the uncertainty quantification of AI models was

highlighted as a critical step, as rewarding areas of large uncertainty in active learning frameworks necessitates the quantification and understanding of the epistemic and aleatoric uncertainty of the models,[38,108,109] and the errors at each step.

V. Education

All participants unanimously agreed on the importance of introducing machine learning, AI, and autonomous research throughout the chemistry curriculum, starting at the undergraduate level and potentially even earlier. While acknowledging the significance of specialized graduate education, these skills are deemed essential for all chemists. Both academia and industry increasingly seek applicants with a solid programming background. As a case study, Novo Nordisk, a major company in Northern Europe, is heavily investing in digital transformation and envisions a future where half of its chemists are computational ("non-wet") chemists. Universities play a vital role in developing such a workforce. Therefore, our discussions primarily focused on undergraduate education unless specified otherwise.

Various educational strategies were explored during the discussions. One extreme example is Nanyang Technological University in Singapore, where university students are mandated to have coursework in computational thinking, data science, and machine learning. Similarly, Imperial College London and the Denmark Technical University have university-wide initiatives to incorporate data and machine learning competencies within the undergraduate curriculum. Another approach involves offering dedicated single courses such as "Data Science for Chemistry" or "Autonomous Discovery" as upper-level electives.[110,111] Some participants shared experiences of incorporating aspects of ML/AI/data science into existing courses or pedagogical laboratory experiences.[112–116] Some of these adaptations were driven by the restrictions imposed by the COVID-19 pandemic. For instance, alternative machine-learning-oriented "computational labs" were developed as substitutes for traditional wet labs. Additionally, remote-control access to laboratory equipment[117] and mailing students Lego kits to build and operate their autonomous systems were also explored. A recent review of low-cost self-driving laboratories collects many of the above efforts in comprehensive categorizations.[118]

Curriculum

What should comprise this curriculum? At a minimum, this coursework should train all chemistry students in (i) elementary programming, (ii) data management best practices, (iii) statistics, (iv) elementary machine learning model construction and evaluation.

Many science and engineering degree programs already require computer programming or numerical computing courses. Historically, these courses were taught in FORTRAN or MATLAB, although the recent trend is to move towards Python, which has become the standard language for machine learning. There are both advantages and disadvantages to having this course taught by a computer science department, considering university politics and topical relevance to students. On the one hand, departments may be protective of their specific areas of study, and other departments may lack the staffing necessary to support the teaching of new classes. On the other hand, students often benefit from direct applications of programming to their primary coursework, which may be lacking in broader service courses. Regardless of how it is offered, it is crucial that students learn elementary programming as early as possible, as it serves as a foundational skill for the other topics covered in the curriculum. It also enables students to undertake projects in their final year focusing on automation or modeling. By adopting this approach, we can create a new generation of students proficient in coding.

Data management encompasses various aspects, including importing, visualization, and adhering to scientific practices such as FAIR data principles. It also involves the development of ontologies, schema, and understanding of intellectual property rights. Incorporating data management into the education of all chemists is crucial, as data generation is inherent to the field, and funding agencies as well as publishers require data management policies. One approach to instilling these practices is to have students create data management plans for projects or upload data from teaching labs to actual repositories. Emphasizing the importance of reporting every repetition of an experiment is essential. Comprehensive data management practices will greatly benefit students when preparing papers, and reinforcing these practices throughout their undergraduate and graduate education is highly valuable.

Statistics is a well-established field and requires no introduction. However, an ideal curriculum would place greater emphasis on computational approaches to statistics.[119]

New forms of education

A side conversation discussed the potential role of virtual reality (VR) in education.[120,121] One panelist highlighted the use of VR in classes to enhance students' understanding of internal structures and processes within battery cells, as well as assist in building crystal structures. The discussion also touched upon the application of VR in outreach programs. For instance, in 2021, chemistry was the theme of the "Explore Science" fair organized by the Klaus Tschira Foundation (Germany) to inspire enthusiasm for natural sciences in schoolchildren. Activities involved realistic and interactive VR explorations with underlying simulations to foster early intuitions about chemistry even before university. It was noted that students often lack these intuitions due to chemistry kits becoming less engaging over time, depriving them of experiences that previous generations had at their age. To address this, the panel suggested the dissemination of VR methods and low-cost laboratory automation kits[111,122–125] to high schools, which could improve student experiences while adhering to modern safety and liability restrictions. At the university level, digital twin simulations of laboratory processes could serve as pre-lab training opportunities, familiarizing students with equipment and lab procedures.

The literature also offers examples of mixed-reality enhancements in teaching microfluidics.[47]

Another potential application for training is using "body cam" footage or similar technologies to provide mentorship in the laboratory. The COVID-19 pandemic, with its need for remote work and limited laboratory occupancy, presented opportunities for pilot projects exploring augmented reality. In these projects, a trainer could supervise trainees from a remote location and provide relevant information directly into the trainee's field of view.

Challenges

The essential ideas in machine learning models for chemistry are continually evolving, with rapid advancements in important models and the rise of deep learning. However, certain core skills and best practices related to model construction, data leakage prevention, data augmentation, and model evaluation remain consistent. The rapid development and accessibility of machine learning software present their own challenges. It is easy to become overwhelmed and attempt to learn everything at once, leading to suboptimal understanding and application of concepts. Participants were cautious about suggesting specific topic selections due to the rapidly changing nature of the field.

When incorporating new computational material into coursework, trade-offs need to be made. Constructive overlaps can be found by substituting programming exercises for lengthy symbolic derivations or incorporating data analysis and sharing exercises instead of traditional laboratory report writing assignments. However, it is inevitable that some content will need to be removed. For instance, some institutions have chosen to reduce math components or replace manual experimental laboratory work with computer-based assignments, which has been well-received by students but has also caused tension within departments. Another approach could involve creating summer coding "bootcamps" that provide anywhere from 1–12 weeks of intensive coding experiences for undergraduate and graduate students, leveraging theory faculty members and inviting guest speakers. However, it is important to recognize that these extracurricular experiences may not engage all students and require faculty to donate their time. More case studies are needed to further explore these trade-offs, and the evolution of curriculum is expected to progress slowly.

Barriers and challenges exist in promoting the incorporation of machine learning into chemistry education. Many chemists outside of the subfield may not perceive it as essential and may lack the necessary skills to teach the material. However, there is value in providing a rigorous education based on fundamentals, and statistical data analysis may serve as a starting point that can act as a gateway to statistical learning methods.

Public perception

Chemistry faces an image problem compared to computer science. Enrollments in computer science programs are increasing while enrollments in physical sciences, including chemistry, are generally decreasing. One possible reason for this trend is the perception that software jobs are more prestigious than careers in science. Factors such as higher salaries, early exposure to computers compared to chemistry sets, and negative perceptions of chemistry as 'polluting' or 'bad' may contribute to this disparity. Despite being the architects of matter, chemists often remain in the background in many applications. The general public may need to be made aware of the significant role chemists and materials scientists play in scientific advancements, such as space exploration, where chemical expertise is essential for activities like analyzing samples and developing chemical processes. Promoting green chemistry can also enhance the appeal of chemistry by highlighting its potential to provide solutions rather than being perceived as a source of problems.[126,127]

The prospect of increased productivity through AI and autonomous research presents an opportunity to elevate the career value of chemists. However, changing perceptions about the role of AI in chemistry and securing investments in autonomous laboratories remain challenges. To attract attention and support, it is not enough to have robots in laboratories; the robots should engage in groundbreaking chemistry and contribute to discoveries that would otherwise be impossible.

On the brighter side

The rise of data-driven approaches in chemistry education may alleviate challenges by reducing the emphasis on memorization and increasing focus on generally applicable concepts and approaches. With the availability of databases and computational tools, students no longer need to rely solely on memorizing vast amounts of information.[128] Instead, they can learn to access and utilize information effectively. This aligns with the changing perspectives of today's students, who view knowledge as the ability to access and apply information rather than simply remembering facts. The evolving nature of assessment also supports this shift. Many instructors adopted "open book" examinations during the COVID-19 pandemic, realizing that online resources easily overcome traditional memorization-based assessments.[129,130] As a result, assessments now require higher-order thinking and problem-solving skills. By incorporating AI-related topics and emphasizing data analysis and decision-making, the curriculum can foster long-term learning and focus on critical concepts while equipping students with practical problem-solving skills.

VI. Conclusions

In conclusion, the 15th ASLLA Symposium offered valuable insights into the use of AI in accelerating chemical science. The discussions on Data, New Applications, Machine Learning Algorithms, and Education highlighted the pivotal role of AI-based techniques in driving the rapid advancement of research and development in the field, as well as the importance of incorporating ML and data science into the curriculum to educate future generations. Key future directions included fostering data transparency, exploring novel applications of AI in chemistry, refining machine learning algorithms for more accurate predictions, and integrating AI-based learning into

chemical education. The importance of cooperation among researchers, educators, associations, publishers, and companies was emphasized in all panel discussions to facilitate AI in chemical science. The authors anticipate the continuation of efforts from various fields, expecting that such endeavors will eventually lead to critical innovations in the field of chemistry.

Data availability

This Perspective is derived from the ASLLA symposium panel discussion, and as such, no data or codes are available for sharing.

Author contributions

Writing – original draft, Alán Aspuru-Guzik, Seoin Back, Michele Ceriotti, Ganna Gryn'ova, Bartosz Grzybowski, Geun Ho Gu, Jason Hein, Kedar Hippalgaonkar, Rodrigo Hormázabal, Yousung Jung, Seonah Kim, Woo Youn Kim, Seyed Mohamad Moosavi, Juhwan Noh, Changyoung Park, Joshua Schrier, Philippe Schwaller, Koji Tsuda, Tejs Vegge, O. Anatole von Lilienfeld, and Aron Walsh. Writing – review & editing, Seoin Back. Funding acquisitions, Yousung Jung, Alán Aspuru-Guzik, and O. Anatole von Lilienfeld.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The symposium organizers (YJ, AAG, and AVL) are grateful to KIST for generous financial support to organize the symposium. YJ acknowledges support from IITP Korea (No. 2021-0-01343, Artificial Intelligence Graduate School Program for Seoul National University & No. 2021-0-02068, Artificial Intelligence Innovation Hub) and NRF of Korea funded by Ministry of Science and ICT (RS-2023-00283902). PS acknowledges support from the NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation. A. A.-G. acknowledges support from the Acceleration Consortium, a Canada First Research Excellence Fund at the University of Toronto as well as Anders G. Frøseth.

References

1 J. Yano, K. J. Gaffney, J. Gregoire, L. Hung, A. Ourmazd, J. Schrier, J. A. Sethian and F. M. Toma, *Nat. Rev. Chem.*, 2022, **6**, 357–370.
2 K. Jorner, A. Tomberg, C. Bauer, C. Sköld and P.-O. Norrby, *Nat. Rev. Chem.*, 2021, **5**, 240–255.
3 M. Meuwly, *Chem. Rev.*, 2021, **121**, 10218–10239.
4 P. Schwaller, A. C. Vaucher, R. Laplaza, C. Bunne, A. Krause, C. Corminboeuf and T. Laino, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1604.
5 F. Strieth-Kalthoff, F. Sandfort, M. H. Segler and F. Glorius, *Chem. Soc. Rev.*, 2020, **49**, 6154–6168.
6 B. Huang and O. A. Von Lilienfeld, *Chem. Rev.*, 2021, **121**, 10001–10036.
7 I. Poltavsky and A. Tkatchenko, *J. Phys. Chem. Lett.*, 2021, **12**, 6551–6564.
8 T. Zubatiuk and O. Isayev, *Acc. Chem. Res.*, 2021, **54**, 1575–1585.
9 E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi and I. Foster, *Matter*, 2021, **4**, 2702–2726.
10 J. H. Montoya, M. Aykol, A. Anapolsky, C. B. Gopal, P. K. Herring, J. S. Hummelshøj, L. Hung, H.-K. Kwon, D. Schweigert and S. Sun, *Appl. Phys. Rev.*, 2022, **9**, 011405.
11 K. Hippalgaonkar, Q. Li, X. Wang, J. W. Fisher III, J. Kirkpatrick and T. Buonassisi, *Nat. Rev. Mater.*, 2023, **8**, 241–260.
12 B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.
13 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz and H. Tribukait, *Nat. Rev. Mater.*, 2018, **3**, 5–20.
14 Z. Yao, Y. Lum, A. Johnston, L. M. Mejia-Mendoza, X. Zhou, Y. Wen, A. Aspuru-Guzik, E. H. Sargent and Z. W. Seh, *Nat. Rev. Mater.*, 2023, **8**, 202–215.
15 R. Pollice, G. dos Passos Gomes, M. Aldeghi, R. J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C. T. Ser and Z. Yao, *Acc. Chem. Res.*, 2021, **54**, 849–860.
16 Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha and R. Q. Snurr, *Nat. Mach. Intell.*, 2021, **3**, 76–86.
17 N. S. Eyke, B. A. Koscher and K. F. Jensen, *Trends Chem.*, 2021, **3**, 120–132.
18 B. A. Grzybowski, T. Badowski, K. Molga and S. Szymkuć, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2023, **13**, e1630.
19 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**, 2454–2466.
20 N. J. Szymanski, Y. Zeng, H. Huo, C. J. Bartel, H. Kim and G. Ceder, *Mater. Horiz.*, 2021, **8**, 2169–2198.
21 S. M. Moosavi, K. M. Jablonka and B. Smit, *J. Am. Chem. Soc.*, 2020, **142**, 20273–20287.
22 J. A. Bennett and M. Abolhasani, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100831.
23 H. Tao, T. Wu, M. Aldeghi, T. C. Wu, A. Aspuru-Guzik and E. Kumacheva, *Nat. Rev. Mater.*, 2021, **6**, 701–716.
24 Z. Bao, J. Bufton, R. J. Hickman, A. Aspuru-Guzik, P. Bannigan and C. Allen, *Adv. Drug Delivery Rev.*, 2023, 115108.
25 Y. Ivanenkov, B. Zagribelnyy, A. Malyshev, S. Evteev, V. Terentiev, P. Kamya, D. Bezrukov, A. Aliper, F. Ren and A. Zhavoronkov, *ACS Med. Chem. Lett.*, 2023, **14**, 901–915.
26 A. L. Ferguson and K. A. Brown, *Annu. Rev. Chem. Biomol. Eng.*, 2022, **13**, 25–44.
27 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends Chem.*, 2019, **1**, 282–291.
28 R. J. Hickman, P. Bannigan, Z. Bao, A. Aspuru-Guzik and C. Allen, *Matter*, 2023, **6**, 1071–1081.

29 S. Lo, S. Baird, J. Schrier, B. Blaiszik, S. Kalinin, H. Tran, T. Sparks and A. Aspuru-Guzik, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv-2023-6z9mq.
30 N. Artrith, K. T. Butler, F.-X. Coudert, S. Han, O. Isayev, A. Jain and A. Walsh, *Nat. Chem.*, 2021, **13**, 505–508.
31 G. Vishwakarma, A. Sonpal and J. Hachmann, *Trends Chem.*, 2021, **3**, 146–156.
32 A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, arXiv preprint arXiv:2212.04356, 2022.
33 EXAONE, https://www.lgresearch.ai/exaone, accessed 19th Sep, 2023.
34 S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit and H. J. Kulik, *Nat. Commun.*, 2020, **11**, 1–10.
35 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. Von Lilienfeld, *Sci. Data*, 2014, **1**, 1–7.
36 P. Zaspel, B. Huang, H. Harbrecht and O. A. von Lilienfeld, *J. Chem. Theory Comput.*, 2018, **15**, 1546–1559.
37 R. Batra, G. Pilania, B. P. Uberuaga and R. Ramprasad, *ACS Appl. Mater. Interfaces*, 2019, **11**, 24906–24918.
38 J. Busk, P. B. Jørgensen, A. Bhowmik, M. N. Schmidt, O. Winther and T. Vegge, *Mach. Learn.: Sci. Technol.*, 2021, **3**, 015012.
39 S. M. Moosavi, B. Á. Novotny, D. Ongari, E. Moubarak, M. Asgari, Ö. Kadioglu, C. Charalambous, A. Ortega-Guerrero, A. H. Farmahini and L. Sarkisov, *Nat. Mater.*, 2022, **21**, 1419–1425.
40 S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik and Y. Jung, *ACS Cent. Sci.*, 2020, **6**, 1412–1420.
41 J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik and Y. Jung, *Matter*, 2019, **1**, 1370–1384.
42 P. C. St. John, Y. Guan, Y. Kim, S. Kim and R. S. Paton, *Nat. Commun.*, 2020, **11**, 2328.
43 M. Uhrin, S. P. Huber, J. Yu, N. Marzari and G. Pizzi, *Comput. Mater. Sci.*, 2021, **187**, 110086.
44 C. Draxl and M. Scheffler, *MRS Bull.*, 2018, **43**, 676–682.
45 C. Draxl and M. Scheffler, *J. Phys.: Mater.*, 2019, **2**, 036001.
46 M. H. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604–610.
47 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers and H. Gao, *Science*, 2019, **365**, eaax1566.
48 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.
49 P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano and T. Laino, *Chem. Sci.*, 2020, **11**, 3316–3325.
50 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *J. Cheminf.*, 2020, **12**, 70.
51 S. Chen and Y. Jung, *JACS Au*, 2021, **1**, 1612–1620.
52 D. Lowe, Chemical reactions from US patents (1976-Sep2016), https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, accessed 19th Sep, 2023.
53 J. Mayfield, D. Lowe and R. Sayle, Pistachio, https://www.nextmovesoftware.com/pistachio.html, accessed 19th Sep, 2023.
54 Reaxys, https://www.reaxys.com, accessed 19th Sep, 2023.
55 SciFinder, https://scifinder.cas.org, accessed 19th Sep, 2023.
56 S. Szymkuć, T. Badowski and B. A. Grzybowski, *Angew. Chem.*, 2021, **133**, 26430–26436.
57 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.
58 G. Pesciullesi, P. Schwaller, T. Laino and J.-L. Reymond, *Nat. Commun.*, 2020, **11**, 4874.
59 P. Seidl, P. Renz, N. Dyubankova, P. Neves, J. Verhoeven, J. K. Wegner, M. Segler, S. Hochreiter and G. Klambauer, *J. Chem. Inf. Model.*, 2022, **62**, 2111–2120.
60 D. P. Kovács, W. McCorkindale and A. A. Lee, *Nat. Commun.*, 2021, **12**, 1695.
61 B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska and W. Beker, *Nature*, 2020, **588**, 83–88.
62 M. Busch, M. D. Wodrich and C. Corminboeuf, *Chem. Sci.*, 2015, **6**, 6754–6761.
63 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario and M. S. Sigman, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
64 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos and P. E. Bourne, *Sci. Data*, 2016, **3**, 1–9.
65 *Open Research Data and Data Management Plans*, 2022, https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf.
66 CAS Common Chemistry, https://commonchemistry.cas.org/, accessed 22nd July, 2023.
67 ACS Research Data Guidelines, https://publish.acs.org/publish/data_guidelines, accessed 22nd July, 2023.
68 Digital Discovery, https://www.rsc.org/journals-books-databases/about-journals/digital-discovery/, accessed 22nd July, 2023.
69 M. Schreiner, A. Bhowmik, T. Vegge, J. Busk and O. Winther, *Sci. Data*, 2022, **9**, 779.
70 M. Schreiner, A. Bhowmik, T. Vegge, P. B. Jørgensen and O. Winther, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045022.
71 J. Jang, G. H. Gu, J. Noh, J. Kim and Y. Jung, *J. Am. Chem. Soc.*, 2020, **142**, 18836–18843.
72 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.
73 L. M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L. P. Yunker, J. E. Hein and A. Aspuru-Guzik, *Sci. Robot.*, 2018, **3**, eaat5559.
74 M. Sim, M. G. Vakili, F. Strieth-Kalthoff, H. Hao, R. Hickman, S. Miret, S. Pablo-García and A. Aspuru-

Guzik, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv-2023-v2khf.
75 M. Vogler, J. Busk, H. Hajiyani, P. B. Jørgensen, N. Safaei, I. E. Castelli, F. F. Ramirez, J. Carlsson, G. Pizzi and S. Clark, *Matter*, 2023, **6**, 2647–2665.
76 S. M. Moosavi, A. Chidambaram, L. Talirz, M. Haranczyk, K. C. Stylianou and B. Smit, *Nat. Commun.*, 2019, **10**, 539.
77 X. Jia, A. Lynch, Y. Huang, M. Danielson, I. Lang'at, A. Milder, A. E. Ruby, H. Wang, S. A. Friedler and A. J. Norquist, *Nature*, 2019, **573**, 251–255.
78 M. Skreta, N. Yoshikawa, S. Arellano-Rubach, Z. Ji, L. B. Kristensen, K. Darvish, A. Aspuru-Guzik, F. Shkurti and A. Garg, *arXiv*, 2023, preprint arXiv:2303.14100, DOI: 10.48550/arXiv.2303.14100.
79 A. M. Bran, S. Cox, A. D. White and P. Schwaller, *arXiv*, 2023, preprint arXiv:2304.05376, DOI: 10.48550/arXiv.2304.05376.
80 D. A. Boiko, R. MacKnight and G. Gomes, *arXiv*, 2023, preprint arXiv:2304.05332, DOI: 10.48550/arXiv.2304.05332.
81 G. M. Hocky and A. D. White, *Digital Discovery*, 2022, **1**, 79–83.
82 M. Krenn, R. Pollice, S. Y. Guo, M. Aldeghi, A. Cervera-Lierta, P. Friederich, G. dos Passos Gomes, F. Häse, A. Jinich and A. Nigam, *Nat. Rev. Phys.*, 2022, **4**, 761–769.
83 V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold and S. R. Elliott, *Nature*, 2021, **589**, 59–64.
84 S. Han, G. Barcaro, A. Fortunelli, S. Lysgaard, T. Vegge and H. A. Hansen, *npj Comput. Mater.*, 2022, **8**, 121.
85 G. H. Gu, J. Lim, C. Wan, T. Cheng, H. Pu, S. Kim, J. Noh, C. Choi, J. Kim and W. A. Goddard III, *J. Am. Chem. Soc.*, 2021, **143**, 5355–5363.
86 S. Han, S. Lysgaard, T. Vegge and H. A. Hansen, *npj Comput. Mater.*, 2023, **9**, 139.
87 M. Ceriotti, C. Clementi and O. Anatole von Lilienfeld, *Chem. Rev.*, 2021, **121**, 9719–9721.
88 A. E. Mikkelsen, H. H. Kristoffersen, J. Schiøtz, T. Vegge, H. A. Hansen and K. W. Jacobsen, *Phys. Chem. Chem. Phys.*, 2022, **24**, 9885–9890.
89 Q. Wang, J. Pan, J. Guo, H. A. Hansen, H. Xie, L. Jiang, L. Hua, H. Li, Y. Guan and P. Wang, *Nat. Catal.*, 2021, **4**, 959–967.
90 Y. Zhang, Y. Zheng, K. Rui, H. H. Hng, K. Hippalgaonkar, J. Xu, W. Sun, J. Zhu, Q. Yan and W. Huang, *Small*, 2017, **13**, 1700661.
91 D. Bash, Y. Cai, V. Chellappan, S. L. Wong, X. Yang, P. Kumar, J. D. Tan, A. Abutaha, J. J. Cheng and Y. F. Lim, *Adv. Funct. Mater.*, 2021, **31**, 2102606.
92 P. Friederich, F. Häse, J. Proppe and A. Aspuru-Guzik, *Nat. Mater.*, 2021, **20**, 750–761.
93 F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Chem. Rev.*, 2021, **121**, 9759–9815.
94 A. Grisafi, D. M. Wilkins, G. Csányi and M. Ceriotti, *Phys. Rev. Lett.*, 2018, **120**, 036002.
95 T. Cohen and M. Welling, Group Equivariant Convolutional Networks, *Proceedings of The 33rd International Conference on Machine Learning*, 2016, https://proceedings.mlr.press/v48/cohenc16.html.
96 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha and T. Wu, *Nat. Mater.*, 2016, **15**, 1120–1127.
97 S. Nagasawa, E. Al-Naamani and A. Saeki, *J. Phys. Chem. Lett.*, 2018, **9**, 2639–2646.
98 K. Molga, E. P. Gajewska, S. Szymkuć and B. A. Grzybowski, *React. Chem. Eng.*, 2019, **4**, 1506–1521.
99 M. Ceriotti, *MRS Bull.*, 2022, **47**, 1045–1053.
100 J. Weinreich, N. J. Browning and O. A. von Lilienfeld, *J. Chem. Phys.*, 2021, **154**, 134113.
101 J. Schrier, A. J. Norquist, T. Buonassisi and J. Brgoch, *J. Am. Chem. Soc.*, 2023, **145**, 21699–21716.
102 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Chem. Sci.*, 2018, **9**, 7642–7655.
103 R. Hickman, M. Sim, S. Pablo-García, I. Woolhouse, H. Hao, Z. Bao, P. Bannigan, C. Allen, M. Aldeghi and A. Aspuru-Guzik, *ChemRxiv*, 2023, DOI: 10.26434/chemrxiv-2023-8nrxx.
104 C. J. Taylor, A. Pomberger, K. C. Felton, R. Grainger, M. Barecka, T. W. Chamberlain, R. A. Bourne, C. N. Johnson and A. A. Lapkin, *Chem. Rev.*, 2023, **123**, 3089–3126.
105 J. C. Fromer and C. W. Coley, *Patterns*, 2023, **4**, 100678.
106 J. Schrier, A. J. Norquist, T. Buonassisi and J. Brgoch, *J. Am. Chem. Soc.*, 2023, **145**, 21699–21716.
107 A. McNally, C. K. Prier and D. W. MacMillan, *Science*, 2011, **334**, 1114–1117.
108 J. Busk, M. Schmidt, O. Winther, T. Vegge and P. B. Jørgensen, *Phys. Chem. Chem. Phys.*, 2023, **25**, 25828–25837.
109 S. Chen and Y. Jung, *Nat. Mach. Intell.*, 2022, **4**, 772–780.
110 S. Vargas, S. Zamirpour, S. Menon, A. Rothman, F. Häse, T. Tamayo-Mendoza, J. Romero, S. Sim, T. Menke and A. Aspuru-Guzik, *J. Chem. Educ.*, 2020, **97**, 689–694.
111 L. Saar, H. Liang, A. Wang, A. McDannald, E. Rodriguez, I. Takeuchi and A. G. Kusne, *MRS Bull.*, 2022, **47**, 881–885.
112 A. K. Sharma, *J. Comput. Sci. Educ.*, 2021, **12**, 8–15.
113 E. S. Thrall, S. E. Lee, J. Schrier and Y. Zhao, *J. Chem. Educ.*, 2021, **98**, 3269–3276.
114 D. Revignas and V. Amendola, *J. Chem. Educ.*, 2022, **99**, 2112–2120.
115 D. Lafuente, B. Cohen, G. Fiorini, A. A. García, M. Bringas, E. Morzan and D. Onna, *J. Chem. Educ.*, 2021, **98**, 2892–2898.
116 A. G. St James, L. Hand, T. Mills, L. Song, A. S. J. Brunt, P. E. Bergstrom Mann, A. F. Worrall, M. I. Stewart and C. Vallance, *J. Chem. Educ.*, 2023, **100**, 1343–1350.
117 R. C. Cachichi, G. Girotto Junior, E. Galembeck, J. A. M. Schewinsky Junior, D. Ferreira Gomes and J. d. A. Simoni, *J. Chem. Educ.*, 2020, **97**, 3667–3672.
118 S. Lo, S. Baird, J. Schrier, B. Blaiszik, S. Kalinin, H. Tran, T. Sparks and A. Aspuru-Guzik, 2023, DOI: DOI: 10.26434/chemrxiv-2023-6z9mq-v2.
119 J. Vanderplas, *Statistics for hackers*, Portland, Oregon, 2016.

120 M. Abdinejad, B. Talaie, H. S. Qorbani and S. Dalili, *J. Sci. Educ. Technol.*, 2021, **30**, 87–96.
121 R. van Dinther, L. de Putter and B. Pepin, *J. Chem. Educ.*, 2023, **100**, 1537–1546.
122 S. G. Baird and T. D. Sparks, *Matter*, 2022, **5**, 4170–4178.
123 R. Keesey, R. LeSuer and J. Schrier, *HardwareX*, 2022, **12**, e00319.
124 L. C. Gerber, A. Calasanz-Kaiser, L. Hyman, K. Voitiuk, U. Patil and I. H. Riedel-Kruse, *PLoS Biol.*, 2017, **15**, e2001413.
125 E. Li, A. T. Lam, T. Fuhrmann, L. Erikson, M. Wirth, M. L. Miller, P. Blikstein and I. H. Riedel-Kruse, *PLoS One*, 2022, **17**, e0275688.
126 L. B. Armstrong, M. C. Rivas, Z. Zhou, L. M. Irie, G. A. Kerstiens, M. T. Robak, M. C. Douskey and A. M. Baranger, *J. Chem. Educ.*, 2019, **96**, 2410–2419.
127 Y. Liu, *J. Chem. Educ.*, 2022, **99**, 2588–2596.
128 G. N. Quam, *J. Chem. Educ.*, 1940, **17**, 363.
129 R. M. Baker, M. E. Leonard and B. H. Milosavljevic, *J. Chem. Educ.*, 2020, **97**, 3097–3101.
130 J. G. Nguyen, K. J. Keuseman and J. J. Humston, *J. Chem. Educ.*, 2020, **97**, 3429–3435.