01. Chemistry, Materials, and Nanoscience

The ability to design and refine materials and chemical compounds has always been key to the rapid advancement of society's technology and infrastructure. Today's complex technologies require a broad spectrum of needs when developing and optimizing materials and chemicals with desired performance [1–3], such as mechanical, electronic, optical, and magnetic properties (e.g., smartphones use up to 75 different elements compared to the twentieth-century version that had only ~30). This new level of technological complexity, combined with the need to search undiscovered areas of the chemical and materials landscape without clear theories or synthesis directions, [4] requires new paradigms that utilize artificial intelligence (AI).

AI will become an integral part of a scientist's arsenal, alongside pen and paper, and experimental and computational tools. It will accelerate the next scientific discoveries and the design and development of revolutionary technologies benefiting society. AI will identify both promising materials and chemicals, and the reaction pathways to make them [5]. Scientists will use AI to generate scientific data in a rational way, formulating new physical models and theoretical insights that drive new paths for rational design of materials and chemicals, and exploring atomic design spaces currently unimaginable.

1. State of the Art

Our ability to discover new materials and chemical reactions is driven by intuition, design rules, models, and theories derived from scientific data generated by experiments and simulation. The number of materials and chemical compounds that can be derived is astronomical, so finding the desired ones can be like looking for a needle in a haystack. Currently, various machine learning (ML) approaches are used to help scientists explore complex information and data sets with the goal of gaining new insights that lead to scientific discoveries. Future discoveries of advanced materials could be greatly accelerated through ML. Note, for example, the timeline from discovery of $LiMn_2O_4$ to nickel-manganese-cobalt (NMC) materials for batteries. Using known data, we could use ML to accelerate discovery of new material classes for batteries from 14 years to less than 5 years (Figure 1.1).

Nowadays, experimental characterization tools routinely provide picometer/picosecond resolved images at an ever-increasing rate, and, when coupled with a modern camera, are capable of providing several hundreds of frames per second. This pushes the data size into the several hundreds of terabytes (TB) per experiment for a single microscope [6]. Real-time analysis of this data, aided by AI, is

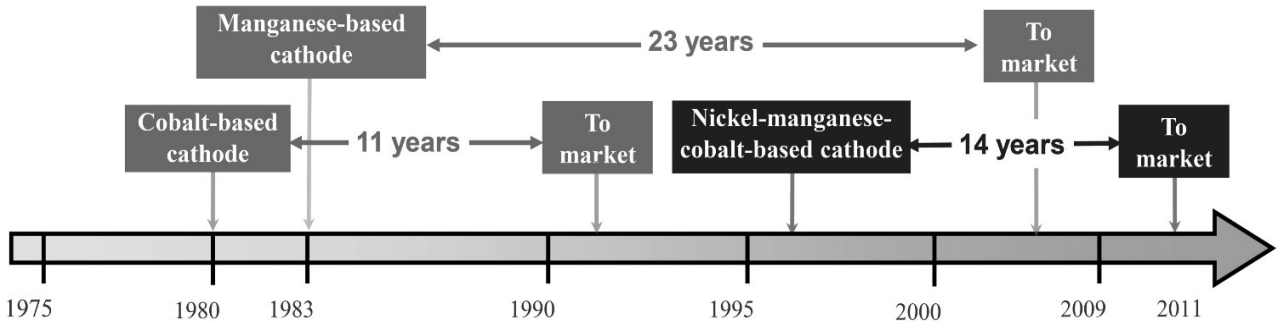

Figure 1.1 *Timeline from discovery of $LiMn_2O_4$ to NMC materials for batteries.*

needed to provide rapid feedback to and from models and simulations that can both inform and validate decisions. Such rapid feedback would also enable experimental adjustments on the fly. Progress has begun to address **two major gaps** in the current paradigm of materials design and discovery that typically proceeds via synthesis ⇒ characterization ⇒ theory.

First, continuous growth in high-performance computing (HPC) capabilities, combined with the development of efficient and scalable electronic structure calculation methods, is enabling scientists to virtually explore materials and chemical compounds. Large databases have come online containing the simulated properties of millions of relatively simple materials and chemical compounds. Deep learning (DL) approaches are being developed for various tasks, such as predicting properties or structure, but this barely scratches the surface of the full atomic design space available to us. Even more, the real world is far more complicated than the simple structures often studied by electronic structure calculations, and simulations investigating systems under device-relevant conditions are still prohibitively expensive. Advances are needed in reliable and precise computational techniques that accurately (and rapidly) address the increasingly complex functionalities required for today's technological applications.

Second, significant progress has been made toward fully exploiting all of the information contained in experimental and computational data to predict and understand new materials. An example is the automated image analysis and recognition based on DL networks that was successfully developed to identify and enumerate defects, and that created a library of (meta) stable defect configurations (Figure 1.2). The electronic properties of the sample surface were further explored by atomically resolved scanning tunneling microscopy (STM). Density functional theory (DFT) was used to estimate the STM signatures of the classified defects from the created library, allowing for the identification of several defect types across multiple imaging platforms. This approach now allows automatic creation of defect libraries in solids, explores the metastable configurations that are always present in real materials, and provides correlative studies with other atomically resolved techniques than can provide comprehensive insight into defect functionalities.

It is this integration and analysis of multiple, complex data sources combined with current state-of-the-art ML approaches that holds great promise for a drastic acceleration of materials and chemical compound discovery.

2. Major (Grand) Challenges

Finding new materials or chemical compounds that have unique properties needed for real-world applications—for example, batteries that hold **10x** the storage capacity compared to today's batteries, or materials that capture more solar energy at greater efficiency—is a grand challenge due to the nearly infinite chemical or atomic design space to which scientists have access. To date, our modern chemical and materials synthesis and discovery process incorporates a wide range of design rules and theories, alongside advanced characterization tools capable of observing synthesis processes on size and time scales at which they occur. At the same time, high-throughput screening via theory-driven approaches, per the materials genome, has provided guidance in identifying promising candidates optimized for particular properties. Early work in ML shows the potential for AI to start to provide guidance on the synthesis pathways to make a material or chemical. The underlying grand challenge as outlined by the Basic Energy Sciences Advisory Committee (BESAC) is how to design and perfect atom- and energy-efficient synthesis of revolutionary new forms of matter with tailored properties. This requires us to explore materials and chemical compounds compositions entirely unknown, driving questions such as, where in our atomic design space do we look? How do

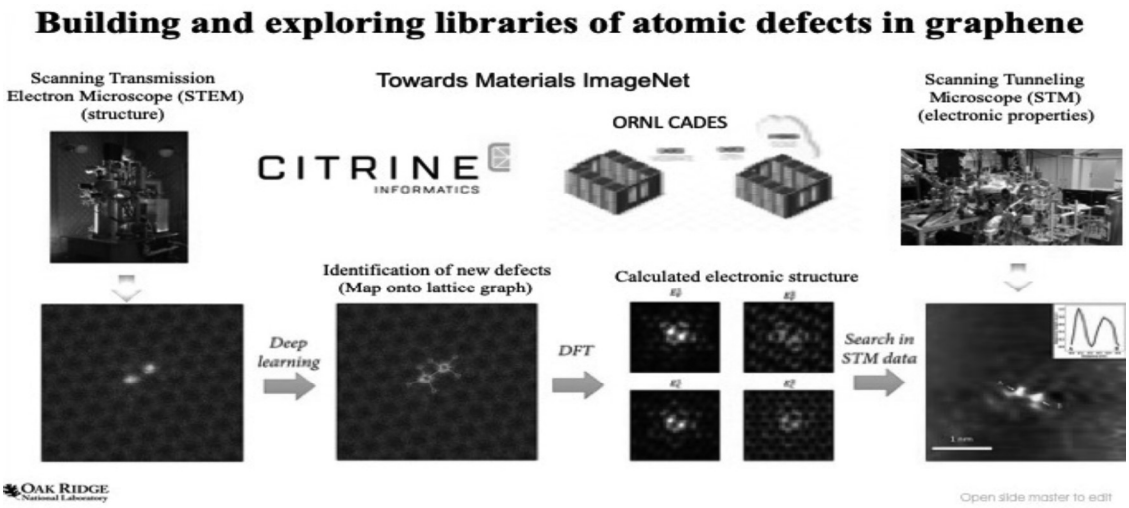

Figure 1.2 *A scanning transmission electron microscope (STEM) images materials where there are defects present or intentionally induced by the electron beam in the STEM. DL via convolutional neural networks is used to process the data to recognize and categorize defects. These data are populated into a database hosted by CITRINE Informatics. DFT calculations via HPC are used to predict STM images for the different defect classes, which then are used to train the DL in a similar fashion to the STEM, and then deposited into the database [7].*

we search the chosen space in the most efficient way or decide to move on to other areas? Can we develop new design rules? Aiding this would be the ability to understand the length- and time-scale evolution of functional chemical and materials systems.

The primary challenges are concisely described by BESAC's 2015 report, *Challenges at the Frontiers of Matter and Energy: Transformative Opportunities for Discovery Science*.

- Mastering Hierarchical Architectures and Beyond-Equilibrium Matter
- Beyond Ideal Materials and Systems: Understanding the Critical Roles of Heterogeneity, Interfaces, and Disorder
- Revolutionary Advances in Models, Mathematics, Algorithms, Data, and Computing
- Harnessing Coherence in Light and Matter
- Exploiting Transformative Advances in Imaging Capabilities across Multiple Scales

Specifically, gaps/challenges that need to be addressed by AI/ML are listed below.

Design metastable phases and materials that persist out of equilibrium. These materials enable access to a diversity of properties beyond the limits drawn by equilibrium thermodynamics. For example, optically driven processes of materials could provide more control over the chemical processes and lead to new materials, such as metastable phases or new low-dimensional materials with dynamics controlled by in-plane heterogeneity rather than layer stacking order. Another example is self-assembly, where transient (non-equilibrium) intermediate states frequently appear, and control of assembly pathways can enable improved structural control. Modern characterization systems such as electron and scanning probe microscopies may allow "bottom-up" fabrication of new structures that are metastable, which allows arrays, for example, of topological defects to be created with nanometer precision for desired properties. The challenge is to do this in an efficient and reproducible fashion; this requires in-line analytics and feedback of very high velocity and volume data streams.

In January 2018, the U.S. Department of Energy's (DOE's) Office of Advanced Scientific Computing Research (ASCR) hosted a Basic Research Needs workshop focused on ML for science. This workshop resulted in development of priority research directions (PRDs) for interpretability, domain awareness, robustness, and needed capabilities (Workshop report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence, https://www.osti.gov/servlets/purl/1478744). Although the workshop highlighted significant investment in ML for the analysis of big data, there has been less activity on the generation of such data sets—a critical need as DOE's major experimental facility upgrades begin commissioning. PRD-6 from the workshop, intelligent automation and decision-support, is highly relevant as timely advances in AI and ML will be critical to enable the full scientific potential. To make AI/ML successful for the large experimental and computational data from our facilities, there are challenges in terms of archiving metadata and preserving provenance, workflows to manage data transfer to and from instruments and integration with HPC facilities, development of software stacks (federated), and uncertainty quantification to identify regions of model validity.

Understand and control interfacial processes and properties. Controlling interfaces (liquid/liquid, gas/solid, etc.) often rely on precise control of atomic bonding and molecular interactions between two dissimilar phases. The ideal strategy to avoid performance-limiting defects in materials, for example, is to minimize perturbation of the atomic order at the interface by preserving a high degree of crystallographic order (e.g., epitaxy). However, atomic scale insights into grown structures present significant inverse problems that have been difficult to address. This may potentially be tackled using combined physics-ML methodologies (Figure 1.3). Additionally, chemical separations, an area which is fundamentally important to almost every aspect of our daily lives, from the energy we utilize to our medications to chemical purification, including water, can see transformative advances with AI in terms of refining and optimizing experimental approaches. The use of AI will aid the pursuit of grand challenges such as understanding complex hierarchical correlations, from molecular-scale interactions up to transport phenomena, and mapping energy landscapes for the chemical and materials transformations that occur during aging of separation materials/chemicals.

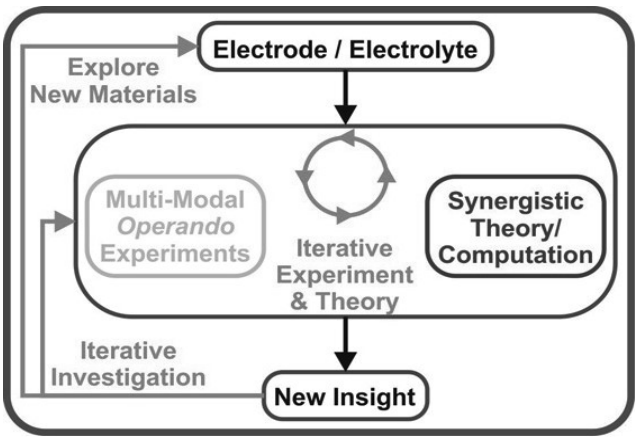

Figure 1.3 *An integrated approach for future design of materials interfaces tailored for performance. Key to this vision is inclusion of multi-modal operando experiments enabled by AI/ML.*

Design materials and molecules for quantum information sciences (QIS). Much of the transformative success of technologies underlying the information age was built on our ability to manipulate chemical composition and doping, and hence electronic band structure and electrochemical potential, within materials at tiny length scales, encode local electronic properties as the physical instantiation of information, and thus control the storage, flow, and processing of information. We now stand on the brink of a *quantum information revolution*. Here, breakthroughs will be driven by the ability to harness the interplay and evolution of quantum entangled and coherent ensembles as the physical representation and processing of information. This will provide radically new opportunities in computation, enabling exponentially higher speeds and efficiencies and the ability to solve problems that are currently intractable. As such, there is a desperate need to deliver systems for potential solid-state qubits, photon sources, and quantum sensing systems [BES Roundtable, Opportunities for Basic Research

for Next-Generation Quantum Systems, Oct. 30–31 (2017); Roundtable, Opportunities for Quantum Computing in Chemical and Materials Sciences, Oct. 31–Nov. 1 (2017)]. Promising advances at DOE facilities in layered materials stamping and a new pulsed laser deposition (PLD) system will generate rich structural, heterointerface, and functional property datasets that will require deep AI/ML analysis and real time control. This analysis/control will need to be done *in situ* and on the timeframe of the experiments to enable smart-steering of the synthesis processes toward successful quantum materials.

Understand the critical roles of heterogeneity in complex systems. Heterogeneities and interfaces underlie novel functionalities and drive dynamical processes, such as charge and exciton transport (e.g., along grain boundaries), charge separation (at Type II heterojunctions) and recombination (at Type I heterojunctions), spin evolution, and transport of ions or molecules through ordered and disordered systems (e.g., at battery interfaces or through metal organic frameworks). However, understanding transient and time-dependent processes in material and chemical systems is enormously challenging; examples include identifying chemical reaction pathways, visualizing electronic and optoelectronic processes at their native lengths (single atoms to many nanometers) and time scales (femto to nanoseconds and beyond) in heterogeneous materials, and studying exchange processes between excitations on various length scales. Progress can be made via high-throughput materials synthesis and automated atomic-scale/multimodal characterization. Here the aim is to broadly understand how population diversity influences growth and behavior, with the ultimate goal of creating a closed-loop materials property prediction, synthesis, and characterization loop. By understanding and controlling heterogeneity, it may be finally possible to design multifunctional and self-regenerating catalytic systems.

Understand and master energy and information with capabilities rivaling those of biological systems. Biological systems naturally transform and distribute energy through photosynthesis and subsequent decomposition of photosynthetic material. Conversion of energy to biomass can occur via various mechanisms, including photosynthetic and chemical pathways with oxygen (i.e., aerobic) and without oxygen (i.e., anaerobic). Greater insights are needed into the regulation of these pathways, the mechanisms responsible for the reactions, and environmental influences on the reactions. This improved understanding is a precursor to enabling changes in pathways that may uncover new or more efficient energy sources.

3. Advances in the Next Decade

In the next five to 10 years, AI will be an integral part of a scientist's discovery and design arsenal. Scientists will use AI to generate scientific data in a rational way, formulating new physical models and theoretical insights that drive new paths of rational design of materials and chemicals, exploring atomic design spaces currently unimaginable.

The ultimate form of AI for materials, chemistry, and nanoscience constitutes ***autonomous-smart experiments and simulations, including synthesis and automated discovery,*** that integrate all aspects of the materials and chemistry discovery loop—from preparation through characterization, to data interpretation and feedback—in order to minimize the experimental trials needed to achieve a desired property or set of properties. This could allow vastly more challenging materials and chemical compound problems to be tackled. However, such an autonomous process will still require *expert scientists in the loop* to ensure viability and success. Overall, the vision of "autonomous-smart experiments" is an as-yet unrealized grand challenge, as the parameter space is simply too large to manage in traditional ways. AI/ML can clearly be a

transformative key to bridge this gap, but it will require addressing a number of challenges (ranging from teaching the AI physical concepts and rational design decisions), making experimental instruments "smart," integrating experimental and simulation data, working with large and diverse sets of streaming data, and having precise control over the experiments. AI/ML can be transformative in terms of high-throughput screening, drastically accelerating simulation capabilities to achieve desired precision with very low computational cost and opening the door to virtually explore a much larger part of the available design space.

Efficient materials, chemical, and device characterization are critical elements in the scientific discovery workflow. As such, the characterization capabilities are constantly used for the determination of chemical composition, structure, physical properties, and overall functionality. In general, this involves (1) an analytical step to confirm that the target chemicals and/or materials are produced; (2) characterization of the physical properties, morphologies, defects, and interfaces of the functional materials and chemicals by multiple probes/techniques; (3) characterization of the functional properties, *in situ/operando*, in devices. This means it will require new analysis across all of these platforms, including registration of data from different instruments (e.g., pan sharpening) and scaling for structure-property mapping. It will be important to fully enable *in situ multimodal analysis with streaming data,* for example, implementing online analysis and active learning during an experiment when more than one type of probe is being used (as data will be streaming at potentially very high velocity and volume).

With AI and ML automation of model-building and decision-making in experimental loops, machine-guided synthesis, processing, and ultimately materials and chemistry discovery can be achieved, enabling discovery, synthesis, and control of novel processes and properties (Figure 1.4).

In the next decade, all the upgrades to DOE's light sources will be completed alongside the proton power upgrade at the neutron source. Thus, there will be significant advances and new information in the following areas.

New data sets/instruments online. There will be a continued increase in the capabilities in detectors/cameras alongside accelerators that will lead to a tremendous increase in potentially high-quality information from microscopes and light sources. Those instrument advances will provide extreme volumes and velocities of data that contain deep information regarding materials/chemistry processes alongside a modality that enables manipulation and control of the materials.

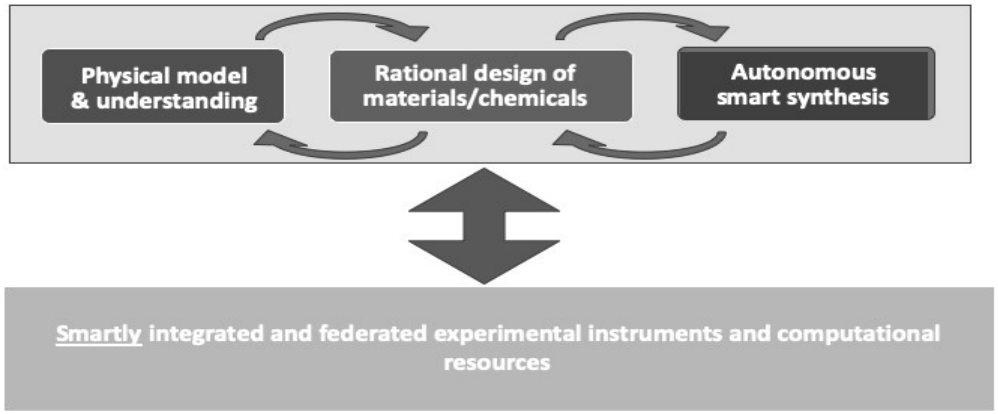

Figure 1.4 *Schematic illustration of the elements of experiments and computations that are required to enable autonomous-smart experiments for materials/chemical design/synthesis.*

Enhancement in big data and data curation. There must be a focused effort to link major facilities and capabilities, such as our leadership computing facilities and our microscopy, light, and neutron sources, to characterize and fully understand the new materials. We need a radical improvement on data sharing, analysis, and curation that will catalyze scientific discovery. *This requires the development of protocols, common data formats, and complete metadata to document and curate the full history and knowledge of the synthesized material.* Furthermore, workflows to integrate knowledge across multiple facilities, and the ability to create and draw on knowledge graphs to better inform modeling and propose new experiments, should be expected. Ultimately, a shared and curated source of data that is easily searchable and minable will be a fundamentally needed infrastructure. Progress is expected along the lines of new AI platforms that integrate diverse scientific data resources, including the literature, and respective mining engines, which will enable automatic development of training sets from heterogeneous experimental and simulated data (see Chapter 12, Data Lifecycle and Infrastructure).

Rare events detection and identification. Rare events are events that occur very infrequently, i.e., their frequency ranges from 0.1 percent to less than 10 percent. While these events are low probability, they can have high impact.

Events such as failure in materials under stress, or side reactions in gas phase chemistry that may occur on time scales too short for humans to observe, are very important to identify. Near-term adaptive control of some experiments—when implemented as real-time decision-making during an experiment—can identify regions of interest and save the relevant data. The introduction of AI into instrument control systems will allow detection when their alignment has drifted and then perform automated alignment and recalibration.

Computers and algorithms. There will continue to be major advances in computer capacity and mathematical algorithms which will further enhance the ability to perform in-line and real-time analysis of experimental and computational data.

Accelerated simulation. Continued advances in computing capacity and computational chemistry and materials methodologies, combined with ML network development, will provide new sets of data for AI/ML and decision making.

New AI/ML techniques. Advances are expected in reinforcement (algorithms that employ reward/punishment), active learning and neuromorphic computing that may be used "at the edge"—where people and things meet (AI/ML at the edge)—as well as in explainable and interpretable AI/ML (see Chapter 10, AI Foundations and Open Problems). Particularly important will be advances in AI/ML approaches that can deal effectively with sparse, unlabeled data.

4. Accelerating Development

To achieve the vision of autonomous-smart experiments/discovery, a number of technical challenges must be addressed. It will be critical to accelerate development in the following areas.

Advance edge computing and integrated experimental instruments. Computing at the experimental instrument(s) for on-the-fly analysis with feedback during an experiment will need to be implemented to maximize information gain and efficient control. This will be particularly important for multimodal experimental probes that require analysis across different platforms. Edge computing for automating aspects of experiments, such as for AI/ML-assisted tuning of the environment, importance sampling, next-experiment recommendation, etc., will be critical. Additionally, on-demand pipelines to HPC for automatic spawning of jobs directly related to

discoveries at the instrument are needed. This can be important for forming databases based on higher levels of ML models trained on simulated data, where the simulations would require an HPC environment. The goal is to provide fast *on-the-fly* analysis of "streaming" experimental data.

Enable *in situ* multimodal analysis. Characterization capabilities are constantly used for the determination of chemical composition, materials structure, physical properties, and how such properties correlate with functionality. In general, this involves (1) an analytical step to confirm that the target chemicals and/or materials are produced; (2) characterization of the physical properties, morphologies, defects, and interfaces of the functional materials, by multiple probes/techniques; (3) characterization of the multi-functional properties, *in situ*/*operando*, in devices (*in vacuo, in solute, in atmosphere*) across a broad frequency range. Achieving acceleration will require new *in situ* multimodal diagnostic approaches which incorporate all of these analytical platforms in one experiment. These include registration of data from different instruments/*in situ* probes and scaling (e.g., pan sharpening) for structure-property mapping, multimodal cross-correlation, and building of frameworks to integrate knowledge in a rigorous physics-based framework that incorporates uncertainty quantification meta data analytics. *In situ* data analytics, including cross modeling, will be approached on two levels, the first level at the point of experiment using edge computing and at the second level of HPC. The ML algorithms will be incorporated as a part of *in situ* multimodal analysis. It will lead to machine-guided decision-making algorithms for selection of optimal experimental condition, minimal number of experiments, and reduced model error.

Enable automated smart characterization. The use of active learning and Bayesian methodologies in combination with predictive modeling during experimental characterization can enable the efficient exploration of heterogeneities in materials and the delicate balance in chemical compounds and reactions. The goal is to minimize the uncertainty and to maximize physics knowledge gain.

Enable AI/ML approaches to represent physics. Dictated by the laws of physics, only discretized structures exist in nature. This "discreteness" needs to be represented properly in the encoding space to control erroneous predictions and misclassifications. New and novel mathematical approaches are needed to incorporate physical constraints and symmetries into the representation and encoding of chemical and materials data, feature detection, and the learning process itself. New kernels that can operate on hierarchical structured data for similarity quantification to enable the application of uncertainty-aware regression methods are also needed.

Enable big-fast data at the signal-noise edge. Use of ML models for characterization at the dose limited range is critical for autonomous experiments. Big-data-based techniques, such as four-dimensional scanning transmission electron microscopy (4D-STEM), are limited by how fast the data can be collected, with the bottlenecks arising from detector readout times and data transfer rates. This imposes constraints on the sample since it rules out beam-sensitive samples that will not be stable under the comparatively slower imaging conditions, and also dynamic *in situ* experiments. Fast detection is possible, but the data is noisier. Current state-of-the-art iterative analysis protocols are more susceptible to noise, and next-generation ML models trained on HPC-simulated datasets can be a way to bridge this gap between big data in microscopy and dynamic microscopy.

This includes integrating data efficiently from different characterization techniques to provide a more complete perspective on materials structure and function. Even with this promising progress, there is still tremendous need for work that can bridge a number of critical gaps,

including delivering a set of open-source petascale quantum simulation, data assimilation, and data analysis tools for functional materials design, within an approach that includes uncertainty quantification and experimental validation and verification of AI models (see Chapter 10, AI Foundations and Open Problems).

Develop a workforce that can work across domains. Existing and emerging training programs in chemistry and materials need to be expanded to ensure a workforce that understands AI approaches and how they can best benefit problems in chemistry and materials discovery.

5. Expected Outcomes

Success in achieving autonomous-smart experiments will lead to transformative advances in:

- The diversity of materials properties possible beyond the limits drawn by equilibrium thermodynamics or our imagination based on discovered design rules.
- The realization of multifunctional and self-regenerating catalytic systems.
- The control of interfaces optimized to perform desired functions.
- On-the-fly materials and (bio)chemical design and synthesis.
- The discovery of unknown synthesizable materials and complex chemical species **1000x faster** and with desired properties.

6. References

1. Riordan, M. & Hoddeson, L., Crystal Fire: The Invention of the Transistor and the Birth of the Information Age, W. W. Norton & Company, 1998.
2. Sze, S. M., *Physics of Semiconductor Devices,* 2nd Edition, John Wiley and Sons, New York, 1981.
3. Shockley, W., Electrons and Holes in Semiconductors: With Applications to Transistor Electronics, D. Van Nostrand Company, Inc., 1950.
4. Fuechsle, M. et al., A single-atom transistor. *Nat. Nanotechnol.* **7**, 242–246 (2012).
5. Sumpter, B. G., Vasudevan, R. K., Potok, T., Kalinin, S. V., A bridge for accelerating materials design. *npj Comp. Mat.* **1**: 15008 (2015). DOI: 10.1038/npjcompumats.2015.8
6. Kalinin, S. V., Sumpter, B. G., & Archibald, R. K., Big-deep-smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
7. M. Ziatdinov, et al., “Building and exploring libraries of atomic defects in graphene: Scanning transmission electron and scanning tunneling microscopy study,” Sci. Adv. 5:eaaw8989 (2019). DOI: 10.1126/sciadv.aaw8989.