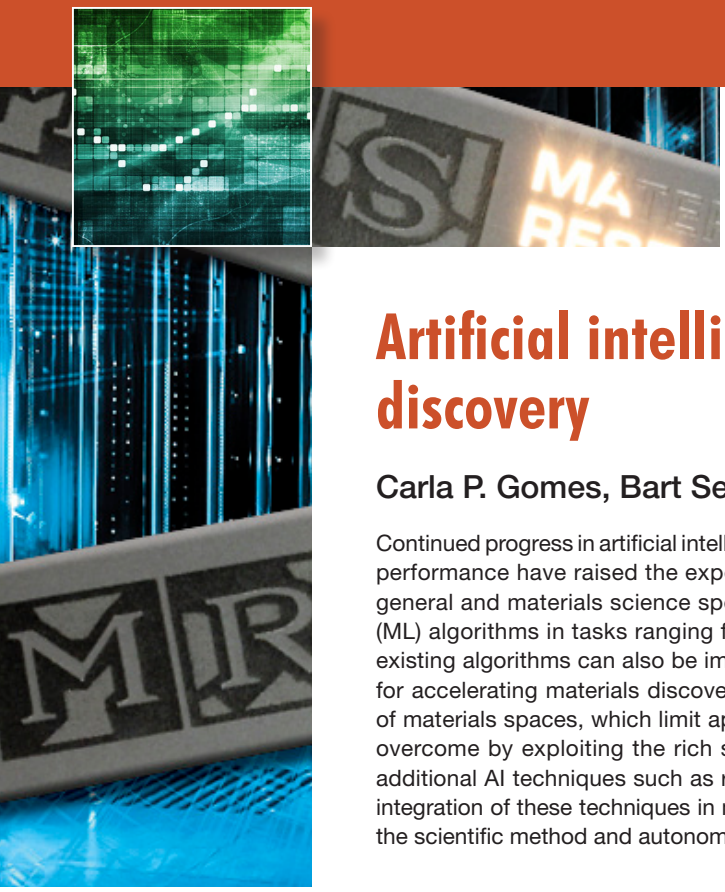

Artificial intelligence for materials discovery

Carla P. Gomes, Bart Selman, and John M. Gregoire

Continued progress in artificial intelligence (AI) and associated demonstrations of superhuman performance have raised the expectation that AI can revolutionize scientific discovery in general and materials science specifically. We illustrate the success of machine learning (ML) algorithms in tasks ranging from machine vision to game playing and describe how existing algorithms can also be impactful in materials science, while noting key limitations for accelerating materials discovery. Issues of data scarcity and the combinatorial nature of materials spaces, which limit application of ML techniques in materials science, can be overcome by exploiting the rich scientific knowledge from physics and chemistry using additional AI techniques such as reasoning, planning, and knowledge representation. The integration of these techniques in materials-intelligent systems will enable AI governance of the scientific method and autonomous scientific discovery.

Artificial intelligence and emergence of intelligent machines

Artificial intelligence (AI) is a subfield of computer science, with the ambitious goal of studying and designing intelligent systems. In recent years, we have witnessed rapid progress in AI—these systems are now reaching human-level and even superhuman-level performance for a range of tasks such as speech recognition, image interpretation, machine translation (Google translate), and gameplay (DeepBlue[1] and AlphaZero[2] for Chess, Watson[3] for Jeopardy!, and AlphaGo[4] and AlphaZero for Go). There is also a general belief that AI is poised to radically transform many components of our society and economy. One example is self-driving cars, which incorporate real-time image recognition and control, and are close to becoming a reality. This tremendous progress is leading to a radical shift in AI research, from a mainly academic endeavor to a much broader field with increasing industrial and governmental investments.

Given the tremendous advances in AI, the broader scientific community has taken note and is exploring the use of AI for scientific discovery.[5–13] In particular, the materials science community has started using AI techniques to accelerate materials discovery. The main current trend is focused on using machine learning (ML) techniques. This is understandable, in part due to that fact that much of the recent AI achievements, especially those concerning superhuman capabilities such as face, image, and speech recognition, are rooted in ML, and more specifically, deep learning (DL), which are described in further detail in the section on "Supervised learning and DL." Nevertheless, while the tremendous progress in DL is undeniable, in areas such as vision, speech recognition, language translation, and autonomous driving, its limitations are well recognized, in particular, given that DL approaches heavily depend on the availability of large amounts of examples or labeled data, which often are not available. The current state of the art of DL has been compared to what Kahneman terms "System 1," described in his book *Thinking, Fast and Slow*.[14] System 1 encompasses a human's routine data processing—it is fast, effortless, a type of pattern recognition that is automatic. In particular, perception, including vision and hearing, are part of System 1. Our perceptual abilities are highly developed with a good part of our cerebral cortex devoted to perception.

Humans also possess a more rational "System 2," which in contrast to System 1 is slow, requiring careful thinking and reasoning, and is needed for solving complex problems beyond reflexive responses. Pure ML and DL are not suitable for such complex tasks. Nevertheless, AI encompasses many other techniques such as search, reasoning, planning, and knowledge representation. These techniques have played various roles throughout the different developmental phases of the AI field. They are poised to become more relevant and play

Carla P. Gomes, Department of Computer Science, Cornell University, USA; gomes@cs.cornell.edu
Bart Selman, Department of Computer Science, Cornell University, USA; selman@cs.cornell.edu
John M. Gregoire, Joint Center for Artificial Photosynthesis, California Institute of Technology, USA; gregoire@caltech.edu
doi:10.1557/mrs.2019.158

an increasing role to complement pure ML approaches and further extend and propel AI research to tackle more complex problems. These problems include those concerning scientific discovery, the types of tasks that require carefully thinking and reasoning, and are performed by humans using System 2.

Reasoning is the process of combining pieces of information into newly inferred facts. Scientific reasoning is arguably our most advanced example of reasoning. For example, given Newton's Laws, we can derive the trajectories of thrown objects as well as those of planets and stars. In this context, "reasoning" is the process of transferring implicit information in nature into explicit information as scientific laws. More generally, knowledge can be captured in causal models. Reasoning often also encompasses search, which involves considering alternate possibilities. Another form of reasoning and search is planning. Given a set of possible actions one can take and a goal, find a sequence of actions that leads to the goal. When playing a game, such as chess, planning yields a strategy to win the game.

The dramatic recent success of AI in materials science, which is summarized in the "Illustrative examples of ML in materials science" section, has resulted from the incorporation of ML and DL in a variety of research tasks. To elevate the use of AI from a research tool to a new paradigm of discovery, we must integrate the key complementary techniques: learning, reasoning, and planning. This integration process is nontrivial and requires significant advances to be fully realized in the physical sciences. Opportunities, challenges, and successes to date are discussed next to highlight the emergence and promise of AI for materials discovery.

ML and data-driven approaches

Supervised learning and DL

Machine learning concerns models that "learn" from data. There are several settings for learning from data, the most developed of which is referred to as supervised learning from labeled data. Labeled data consist of a set of input–output pairs, where the training data output is referred to as "labels." For example, an input can be an image, with the ML task being to label the main object in the image, or the input can be a picture of a person's face for the ML task of labeling the age of the person. Each input–output pair is a training example for the ML algorithm, which the ML algorithm uses to discover a mapping from an input to an output label.

The key challenge for ML methods is that we want them to generalize well beyond the given input–output examples. That is, we want the learned function to provide the correct label for inputs it has not seen before. When the output is continuous, supervised learning is called regression, whereas when the output is discrete, it is called classification. An example of regression is learning a function that predicts the average temperature of a given location during the month of April from historical data. An example of classification is learning a function for predicting whether it will be a sunny or rainy day. There are different learning models or algorithms for supervised learning such as classical statistical models (e.g., generalized linear regression models) or ML models such as decision trees, kernel methods, artificial neural networks, and deep neural networks.

A key factor in determining the effectiveness of ML methods is how the input data are represented. Most traditional approaches represent the inputs not in their most basic form, rather they first compute several additional properties, also called "features" or "attributes," to serve as model inputs. Decision trees take as input a vector of attribute values and return an output via a sequence of tests: each node of the tree tests the value of one of the input attributes, and the branches from the node are labeled with the possible values or range of values of the attribute. A leaf node corresponds to a value or label to be returned for a given input vector of attributes. Decision trees, such as those found in instruction manuals, are intuitive to understand and are increasingly effective when used in "ensembles" that consist of a set of multiple decision trees, whose collective output is obtained via aggregation or voting. The intuitive modeling and relative ease of training have facilitated broad deployment of decision trees for supervised learning.

The ML revolution involves several advancements with many recent successes resulting from the development of DL, motivating further discussion of how DL is so radically different from other supervised learning approaches.

DL uses a computational model inspired by the neural networks found in the brain. More specifically, in DL, we train a large network of simple compute units. In such a deep neural network (DNN), each compute unit is a highly simplified model of a biological neuron. The artificial neurons propagate information from the input layer of neurons through a sequence of layers to eventually reach a final layer, which is called the output layer. For example, when the input is the pixel values of a digital image, the output layer has a unit for each of the possible labels. When the neural net is properly trained, the output unit corresponding to the correct label for the input image is activated via propagation of the input through the sequence of neural network layers. The computation performed by the network depends on the connection strengths or "weights" of the interlayer links, which conceptually mimic the synaptic connections between neurons. The training method is based on stochastic gradient descent (SGD), which is a basic multivariate function optimization technique. SGD allows for an incremental process of updating the weights. Decades of effort in the development of DL methods led to a 2012 breakthrough where the confluence of sufficient computational power and training data set size (1 million labeled images) enabled training of a DNN with an unprecedented combination of prediction speed and accuracy for classification and recognition of objects in images.[15]

It is useful to consider why DNNs work so well compared to alternative ML methods. The key issue is that DL avoids the need for so-called "feature engineering." The traditional approach to ML required the identification of a set of properties

or attributes of the input examples to use by the learning algorithm. For example, for image recognition, one would not use the raw pixel values of the image, but rather first process that data to identify higher-level features of the image such as lines, edges, regions, foreground, and background in the image. While these attempts to assist the model are well intended and motivated by so-called "image filters" found in the early layers of biological vision systems, the decisions made during feature engineering constrain the model.

DL essentially solves the feature engineering problem for ML. By examining large amounts of data, it was found that a DNN can uncover the right features to use for the overall ML task. These features are computed in the early set of layers in the DNN. It is often difficult or impossible to understand exactly what features are computed. More specifically, the class of functions that can be computed with a 20+ layer deep net is incredibly rich and can be characterized by what is computable using a finite depth Boolean circuit, which can capture high degrees of hidden nonlinearities in the data. It is still far from understood exactly how SGD can uncover (or learn) these features when given enough labeled data. The typical need for training data sets with millions of data points arises from the hundreds of thousands of weight parameters to be determined. Deep nets only start to shine when given large amounts of data. Image recognition networks are trained on more than 1 million labeled example images.

This brings us to the final component for making DL a success, namely graphics processing units (GPUs). Graphics processing cards were initially developed for fast rendering of images in video games. GPUs excel at the matrix-style computations that are central to deep net learning, enabling parallel computation and bolstering the dissemination of DL methods.

Unsupervised learning

Models have also been developed for tasks where training data are unavailable, but these unsupervised methods are not as mature as supervised learning. Clustering is a common technique used to identify clusters of input examples, according to some metric. There is also a variety of dimensionality reduction techniques, which exploit the redundancy of the input data set by finding a smaller set of new variables, each being a combination of the input variables, containing basically the same information as the input variables. Examples of dimensionality reduction techniques include principal component analysis, factor analysis, independent component analysis, and nonnegative matrix factorization (NMF).[16] NMF methods have the advantage of producing nonnegative feature vectors, which is desirable in many situations in which the feature vectors correspond to nonnegative components.

Another class of power techniques being developed concerns deep generative models that are suitable for unsupervised data. Essentially, deep generative models take advantage of the DL machinery to extract hidden structure in the data and learn an approximation of its distribution (e.g., images) in an unsupervised way. The two most popular models are variational autoencoders (VAEs) and generative adversarial networks (GANs). VAEs learn a low-dimensional latent representation of the training data, using a probabilistic graphical model, assuming a Gaussian prior distribution, to easily sample new data points from the learned distribution during inference time. GANs are based on a game theory approach, comprising two adversarial networks—a generator and discriminator. The goal is to reach the Nash equilibrium between the two networks, where each is optimal assuming the other is held constant. The generator produces samples from the data distribution (e.g., images) with the goal of fooling the discriminator into believing that it is a "real" image.

Illustrative examples of ML in materials science

Due to the need for large training sets, a key enabling capability for ML in materials science was the establishment of computational materials databases such as the Materials Project,[17] the Open Quantum Materials Database,[18] and Automatic Flow for materials discovery,[19] where a training data point (input and label) can be extracted from each of the 10^3–10^6 database entries. Training ML models from computational data has become fairly commonplace, and has led to the identification of experimentally verified new materials such as superhard materials.[20] With concerted efforts in data curation, large training data sets have also been constructed from experimental data, leading to the successful deployment of ML in areas such as labelling images,[21,22] predicting new metallic glasses,[23] and designing chemical synthesis.[24] Historically, computational prediction in materials science has involved the application of a principled method, most commonly, density functional theory (DFT) or other physics- and chemistry-based methods, to each candidate material or synthesis method. While this approach has led to numerous materials discoveries[25,26] as well establishment of the computational materials database previously noted, methods with sufficient predictive accuracy have substantial computational expense that limits the ability to screen "everything."

One vision for ML as a complementary computational prediction tool is that provided sufficient training data, a sufficiently accurate model will be learned that can then be evaluated for new inputs as a minute fraction of the computational expense compared to principled methods, alleviating requirements to be judicious in the selection of candidate inputs. Adventurous exploration of input space is conceptually ideal for identifying materials with unprecedented properties of interest, but a key limiting factor is that standard ML models are generally only predictive for inputs that are within the confines of the training data. Finding appropriate representations for materials to capitalize on the predictive power of ML is an active area of research.[27–29] AI approaches for addressing this critical limitation with reasoning techniques are discussed next.

Direct application of ML in materials science has also been successful as a research tool to accelerate specific tasks or boost the performance of an existing method. The ML of density functionals (to improve DFT methods) provided an early example of ML in materials science,[30] and the community has continued to develop ML methods to quantify uncertainty, correct errors, and prescreen inputs to improve the performance of principled computational methods.[31] These advances have undoubtedly been enabled by the computational prowess of computational materials scientists, but the promise of ML-accelerated materials discovery has garnered substantial interest in the broader community.

A commensurate development has been the democratization of ML methods via open-source software packages, and some computational materials application programming interfaces (APIs) now include standard ML algorithms that are readily used by nonexperts, fostering ML adoption in materials research.[27,32,33] These advancements are accompanied by a risk that physics and chemistry fundamentals are not appropriately utilized, a concern noted in a recent National Academies workshop.[34] While the lack of principled logic is a design feature of algorithms such as DNNs, generating scientific understanding of materials will be best served by remaining rooted in the fundamentals. This requires the development of techniques that are specific to materials science as opposed to direct application of methods developed for applications such as machine vision, which lack the structured logical frameworks that embody the physical sciences.

An intriguing materials science problem where scientific reasoning has been employed is phase mapping, wherein a phase diagram is an automated determination from a collection of x-ray diffraction patterns.[35] The computational complexity of the problem and the need to inject prior knowledge to obtain meaningful solutions has motivated the deployment of a variety of unsupervised learning techniques.[35–40] Unsupervised learning is also critical for new modes of materials design, as recently noted by Sanchez-Lengeling and Aspuru-Guzik,[41] who describe how deep generative models facilitate inverse design. This new mode of discovery is enabled by the generative aspect of such AI methods, which propose a functional material as opposed to computational screening of candidate materials. These examples are representative of the recent proliferation of materials-specific learning algorithms, which are the foundation for scientific reasoning AI systems and hold great promise for enabling discovery, as discussed below.

Learning and reasoning for scientific discovery

Incorporating scientific knowledge

Performing science research involves the discovery of new scientific knowledge from data. Herein, we discuss specific challenges this process represents for standard ML approaches, namely, the sparsity of rich annotated training data sets and the need for intricate scientific background knowledge and reasoning capabilities to derive new knowledge from data. While approaches for utilizing ML methods with typical materials data sets are being actively explored,[42] addressing the grand challenge of AI-driven discovery will require the integration of ML with other complementary and powerful AI techniques.[12,43–46]

First, as discussed in the previous section, the most successful ML methods require large amounts of input–output pairs of data (labeled data). Improvements in automated sensors, automated data-collection methods, and a variety of high-throughput instruments provide scientists with incredible amounts of data. Such data, however, are not labeled data, and therefore not suitable for supervised learning techniques. Examples of non-labeled data can be found in spectroscopy, time series analysis, and automatic text analysis. Such data require unsupervised learning methods, which can identify data patterns or extract key features of the data, and is a considerably less developed subfield of AI than supervised learning.

A second issue concerns the fact that standard ML methods are in essence a form of data interpolation: what we learn from labeled data about new unseen data essentially results from interpolating from previously seen examples. More technically, the test examples (unseen examples) are assumed to come from the same probability distribution as the training examples. However, in scientific discovery, we generally want to predict far outside the training distributions. Much of materials research aims to identify top-performing materials that are by definition beyond the confines of the available data. The vast dimensionality of materials space poses challenges in developing representations of materials and models that can competently explore such a large space.

A third issue is that scientific discovery is a process of combining new observations and data with existing scientific knowledge, an important strategy for lowering the demands on data. First-principles models, whose origins are traceable to the most fundamental aspects of the problem at hand, fully capitalize on existing scientific knowledge, while most current ML models do not have a principled way of incorporating prior scientific knowledge. Sparse training data and a lack of principled reasoning can severely limit the applicable scope of ML models, motivating a bridge from first-principles to ML to improve generalization of the models.

A fourth issue is that most scientific discoveries involve a process of active experimentation where experiments are specifically designed to test a specific hypothesis. Current ML methods are largely passive in their modeling of hidden data relationships. A promising area of ML is called active learning, in which the data collection is continuously adjusted and optimized in order to identify the most promising scientific hypotheses.

Finally, the search space of possible useful scientific hypotheses is generally vastly beyond the reach of any exhaustive or systematic search procedure. Mathematically, these spaces are highly "combinatorial" in nature, and consequently, the number of hypotheses to consider can easily exceed the number of particles in the universe. The illusion that many approaches can be scaled up by just increasing computational power,

faster hardware, and larger memories is a misconception that still prevails in many disciplines, including materials science, especially for those not aware of the real computational resource barriers to solve certain truly hard computational problems. Such problems often encounter exponential scaling in the compute resources required for their solutions. Computer scientists have identified thousands of such problems and demonstrated that progress on these tasks requires specialized methods that exploit hidden data structure and utilize deductive methods. The essence of such methods is to avoid brute force exploration of possibilities, which requires (typically nonexistent) large amounts of data. A successful automated scientific procedure *de facto* needs to harness the combinatorial explosion by using scientific knowledge to choose the most promising hypotheses to explore.

Integrating learning and reasoning

The previously listed challenges lead us well beyond what is feasible with current pure ML methods, including DL. However, the AI field over the last several decades has also developed a variety of techniques to deal with many aspects of these challenges that can be adapted, expanded, and combined with ML techniques to further enhance the power of AI for materials science. For example, reasoning and knowledge representation methods allow us to represent various forms of scientific knowledge. Probabilistic reasoning techniques allow us to select the most promising hypotheses. Planning methods allow us to design a sequence of steps to reach a goal given an initial state.

A good example of the effectiveness of creating a synergy of AI approaches that lies beyond pure ML approaches is AlphaGo, a program developed by Deep Mind that recently outperformed the best human players of the abstract strategy board game Go, an achievement that had not been expected to have occurred for at least another decade. A core component of AlphaGo is its DL network to evaluate its Go board positions. However, an equally important component is the use of randomized game tree search technique, Monte Carlo Tree Search, which was developed in 2006.[47] Another key factor in the AlphaGo success was deep reasoning via a so-called reinforcement learning technique, which combines DL with evaluation of the effect of different action sequences and the search for optimal moves under uncertainty in frameworks that model decision making, in particular Markov decision processes.[4] AlphaGo Zero, a successor to AlphaGo, recently achieved superhuman performance in the game of Go playing against itself using reinforcement learning.[48] A generalization of AlphaGo Zero, AlphaZero, taught itself how to play several games, starting from random play, defeating a world champion program in chess and shogi (Japanese chess), as well as Go.[2,4,49–51]

Some recent successes in materials and chemistry research are building on these innovative multi-technique approaches to achieve superhuman scientific capabilities. ML models for molecular systems have been tailored to directly incorporate underlying physical constraints, most notably, by encoding physical invariants into the ML models.[52–54] A similar concept was employed by Welborn et al.,[55] who utilized first-principles methods to the furthest practical extent, dovetailed with ML to predict more expensive properties. The corresponding high expense of the training data made use of first principles critical to effectively invoke ML with modest training data.

The chemical synthesis prediction of Segler, Preuss, and Waller,[24] where AI is used to identify the building blocks of target small organic molecules, is an excellent example of combining AI planning, symbolic reasoning, Monte Carlo Tree Search, and DL. The AI system named "CRYSTAL" was recently developed for the phase mapping problem previously described and identifies novel solutions by interweaving algorithmic bots to obtain multiple perspectives of the data and to enforce prior scientific knowledge.[56] Expanding this approach for synthesis and characterization of novel materials is the primary goal of our recent effort to develop the Scientific Autonomous Reasoning Agent (SARA), envisioned as a platform for autonomous components and for human–machine collaboration.[57] At a high level, this family of algorithms aims to realize AI governance of the scientific method, as illustrated in **Figure 1**.

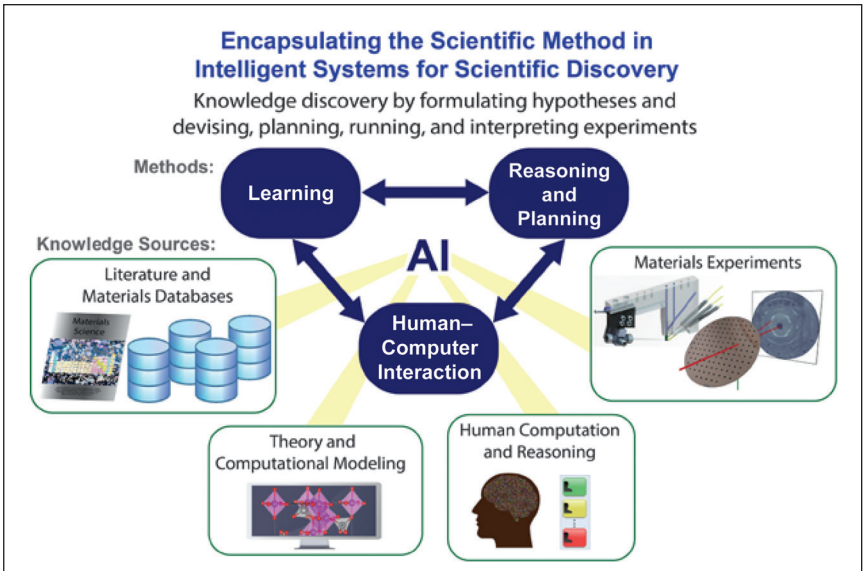

Figure 1. Amplifying human capabilities by using artificial intelligence (AI) to accelerate discovery using the scientific method. Scientific discovery requires an integration of a range of AI techniques, going beyond purely data-driven machine learning. The primary classes of AI techniques, learning, reasoning, and planning are combined with human–computer interaction, and this ensemble of techniques enables integration of multiple knowledge sources, as illustrated with examples including databases, theory, experiments, and human reasoning.

Autonomous experiments comprise a central component of AI-driven science and have been developed in several scientific disciplines areas[5,12,45,58] and recently demonstrated in materials science.[9,59] While these examples of autonomous experimentation, and more generally active learning,[60,61] are notable advancements, the experimental systems are engineered within experimental boundaries designed by humans based on perceived relevance to the problem at hand. Enabling exploration beyond these boundaries and pursuit of scientific knowledge beyond the task at hand is a level of AI in scientific discovery that we have yet to realize. The development of methods that integrate multiple AI techniques is pivotal for realizing the promise of AI for accelerating scientific discovery.

Conclusion

AI is transforming a variety of disciplines and is increasingly evident in our daily lives. AI is also impacting materials science, most commonly by employing ML algorithms to automate specific research tasks. Integration of multiple and complementary AI techniques, such as ML, search, reasoning, planning, and knowledge representation, will further accelerate knowledge discovery in materials science.

Acknowledgments

This work was supported by an award from the Toyota Research Institute; NSF Award Nos. CCF-1522054 and CNS-0832782 (Expeditions), CNS-1059284 (Infrastructure), and IIS-1344201 (INSPIRE); ARO Award No. W911-NF-14-1-0498; AFOSR Multidisciplinary University Research Initiatives (MURI) Program FA9550-18-1-0136; and US DOE Award No. DE-SC0004993.

References

1. M. Campbell, A.J. Hoane, F. Hsu, *Artif. Intell.* **134**, 57 (2002).
2. D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, *Science* **362**, 1140 (2018).
3. D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, C. Welty, *AI Mag.* **31**, 59 (2010).
4. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, *Nature* **529**, 484 (2016).
5. J. Bohannon, *Science* **357**, 16 (2017).
6. Y. Gil, M. Greaves, J. Hendler, H. Hirsh, *Science* **346**, 171 (2014).
7. P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, E. Sargent, *Nature* **552**, 23 (2017).
8. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj. Comput. Mater.* **3**, 54 (2017).
9. P. Nikolaev, D. Hooper, F. Webber, R. Rao, K. Decker, M. Krein, J. Poleski, R. Barto, B. Maruyama, *npj. Comput. Mater.* **2**, 16031 (2016).
10. E. Smalley, *Nat. Biotechnol.* **35**, 604 (2017).
11. E. Horvitz, S. Zilberstein, *Artif. Intell.* **126**, 1 (2001).
12. R.D. King, K.E. Whelan, F.M. Jones, P.G.K. Reiser, C.H. Bryant, S.H. Muggleton, D.B. Kell, S.G. Oliver, *Nature* **427**, 247 (2004).
13. J.R. Kitchin, *Nat. Catal.* **1**, 230 (2018).
14. D. Kahneman, *Thinking, Fast and Slow* (Farrar, Straus, and Giroux, New York, 2011).
15. A. Krizhevsky, I. Sutskever, G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.* **25**, F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger, Eds. (ACM Publications, New York, 2012), pp. 1097–1105, http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (accessed April 22, 2019).
16. D.D. Lee, H.S. Seung, *Nature* **401**, 788 (1999).
17. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**, 011002 (2013).
18. S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **1**, 15010 (2015).
19. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012).
20. A. Mansouri Tehrani, A.O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T.D. Sparks, J. Brgoch, *J. Am. Chem. Soc.* **140**, 9844 (2018).
21. B.L. DeCost, E.A. Holm, *Comput. Mater. Sci.* **110**, 126 (2015).
22. N. Borodinov, S. Neumayer, S.V. Kalinin, O.S. Ovchinnikova, R.K. Vasudevan, S. Jesse, *npj Comput. Mater.* **5**, 5 (2019).
23. F. Ren, L. Ward, T. Williams, K.J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, *Sci. Adv.* **4**, eaaq1566 (2018).
24. M.H.S. Segler, M. Preuss, M.P. Waller, *Nature* **555**, 604 (2018).
25. A. Jain, Y. Shin, K.A. Persson, *Nat. Rev. Mater.* **1**, 15004 (2016).
26. Q. Yan, J. Yu, S.K. Suram, L. Zhou, A. Shinde, P.F. Newhouse, W. Chen, G. Li, K.A. Persson, J.M. Gregoire, J.B. Neaton, *Proc. Natl. Acad. Sci. U.S.A.* **114**, 3040 (2017).
27. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2**, 16028 (2016).
28. O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **8**, 15679 (2017).
29. A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, I. Tanaka, *Phys. Rev. B* **95**, 144110 (2017).
30. J.C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **108**, 253002 (2012).
31. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **559**, 547 (2018).
32. E. Gossett, C. Toher, C. Oses, O. Isayev, F. Legrain, F. Rose, E. Zurek, J. Carrete, N. Mingo, A. Tropsha, S. Curtarolo, *Comput. Mater. Sci.* **152**, 134 (2018).
33. L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **152**, 60 (2018).
34. National Academies of Sciences and Medicine, *Data Science: Opportunities to Transform Chemical Sciences and Engineering: Proceedings of a Workshop—in Brief* (National Academies Press, Washington, DC, 2018).
35. J.R. Hattrick-Simpers, J.M. Gregoire, A.G. Kusne, *APL Mater.* **4**, 053211 (2016).
36. H.S. Stein, S. Jiao, A. Ludwig, *ACS Comb. Sci.* **19**, 1 (2017).
37. V. Stanev, V.V. Vesselinov, A.G. Kusne, G. Antoszewski, I. Takeuchi, B.S. Alexandrov, *npj Comput. Mater.* **4**, 43 (2018).
38. S. Ermon, R. Le Bras, S.K. Suram, J.M. Gregoire, C.P. Gomes, B. Selman, R.B. Van Dover, "Pattern Decomposition with Complex Combinatorial Constraints: Application to Materials Discovery," *Proc. 29th AAAI Conf. Artif. Intell.* (AAAI Press, Austin, TX, 2015).
39. S.K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R.B. van Dover, C.P. Gomes, J.M. Gregoire, *ACS Comb. Sci.* **19**, 37 (2017).
40. C.J. Long, D. Bunker, X. Li, V.L. Karen, I. Takeuchi, *Rev. Sci. Instrum.* **80**, 103902 (2009).
41. B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **361**, 360 (2018).
42. Y. Zhang, C. Ling, *npj Comput. Mater.* **4**, 25 (2018).
43. S.J. Gershman, E.J. Horvitz, J.B. Tenenbaum, *Science* **349**, 273 (2015).
44. M. Schmidt, H. Lipson, *Science* **324**, 81 (2009).
45. D. Waltz, B.G. Buchanan, *Science* **324**, 43 (2009).
46. S.H. Muggleton, *Nature* **440**, 409 (2006).
47. L. Kocsis, C. Szepesvári, "Bandit Based Monte-Carlo Planning," *Proc. Mach. Learn. ECML 2006*, J. Fürnkranz, T. Scheffer, M. Spiliopoulou, Eds. (Springer, Berlin, Germany, 2006), pp. 282–293.
48. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, *Nature* **550** (7676), 354 (2017).
49. L.P. Kaelbling, M.L. Littman, A.W. Moore, *J. Artif. Intell. Res.* (1996), https://arxiv.org/abs/cs/9605103v1 (accessed April 22, 2019).
50. R.S. Sutton, A.G. Barto, *Reinforcement Learning, Second Edition* (MIT Press, Cambridge, MA, 2018), https://mitpress.mit.edu/books/reinforcement-learning-second-edition (accessed April 22, 2019).
51. R.S. Sutton, A.G. Barto, *Introduction to Reinforcement Learning* (MIT Press, Cambridge, MA, 1998).
52. A. Grisafi, D.M. Wilkins, G. Csányi, M. Ceriotti, *Phys. Rev. Lett.* **120**, 036002 (2018).
53. J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, "Neural Message Passing for Quantum Chemistry," *Proc. 34th Int. Conf. Mach. Learn.* **70** (JMLR.org, 2017), pp. 1263–1272, http://dl.acm.org/citation.cfm?id=3305381.3305512 (accessed April 19, 2019).
54. S. Kearnes, K. McCloskey, M. Berndl, V. Pande, P. Riley, *J. Comput. Aided Mol. Des.* **30**, 595 (2016).

55. M. Welborn, L. Cheng, T.F. Miller, *J. Chem. Theory Comput.* **14**, 4772 (2018).
56. C. Gomes, J. Bai, Y. Xue, J. Bjorck, B. Rappazzo, S. Ament, R. Bernstein, S. Kong, S. Suram, R. van Dover, J. Gregoire, *MRS Commun.* (forthcoming).
57. https://community.apan.org/wg/afosr/w/researchareas/22949/scientific-autonomous-reasoning-agent-sara-integrating-materials-theory-experiment-and-computation.
58. L.M. Roch, F. Häse, C. Kreisbeck, T. Tamayo-Mendoza, L.P.E. Yunker, J.E. Hein, A. Aspuru-Guzik, *Sci. Robot.* **3**, eaat5559 (2018).
59. D.P. Tabor, L.M. Roch, S.K. Saikin, C. Kreisbeck, D. Sheberla, J.H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C.J. Brabec, B. Maruyama, K.A. Persson, A. Aspuru-Guzik, *Nat. Rev. Mater.* **3**, 5 (2018).
60. T. Lookman, P.V. Balachandran, D. Xue, R. Yuan, *npj Comput. Mater.* **5**, 25 (2019).
61. A.G. Kusne, T. Gao, A. Mehta, L. Ke, M.C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M.J. Kramer, C. Long, I. Takeuchi, *Sci. Rep.* **4**, 6367 (2014). □

Carla Gomes is a professor of computer science and the director of the Institute for Computational Sustainability at Cornell University. She received her PhD degree in computer science from The University of Edinburgh, UK. Her research interests include artificial intelligence and machine learning with a focus on scientific discovery and computational sustainability. She is a Fellow of the Association for the Advancement of Artificial Intelligence (AAAI), a Fellow of the Association for Computing Machinery (ACM), and a Fellow of the American Association for the Advancement of Science (AAAS). Gomes can be reached by email at gomes@cs.cornell.edu.

Bart Selman is the Joseph C. Ford Professor of Engineering and Computer Science at Cornell University and President-Elect of the Association for the Advancement of Artificial Intelligence. He received his PhD degree in computer science from the University of Toronto, Canada. His research interests include artificial intelligence, machine learning, knowledge representation, and computational sustainability. He is the recipient of the Inaugural IJCAI John McCarthy Research Award and a Fellow of the American Association for Artificial Intelligence (AAAI), the American Association for the Advancement of Science (AAAS), and the American Association for Computing Machinery (ACM). Selman can be reached by email at selman@cs.cornell.edu.

John Gregoire leads the High Throughput Experimentation Group at the California Institute of Technology, where he is also the thrust coordinator for Photoelectrocatalysis in the Joint Center for Artificial Photosynthesis, a US Department of Energy, Energy Innovation Hub. He received his PhD degree in physics from Cornell University. His research focuses on discovering and understanding energy-related materials via combinatorial and high-throughput experimental methods, and their integration with materials theory and artificial intelligence. Gregoire can be reached by email at gregoire@caltech.edu.