ACS central science

EDITORIAL

ACS Central Science Virtual Issue on Machine Learning

How was your morning? Perhaps you woke up, did a little online shopping while brewing your coffee, posted some pictures on social media over breakfast, glanced over the world news, drove to work, checked your email, picked up your mail, and opened up your latest issue of *ACS Central Science*. Pretty unremarkable, right? Maybe, but in the few hours that you have been awake you have most likely interacted with numerous instances of machine learning algorithms ticking away just below the surface of our everyday lives.

The term "machine learning" may be defined as algorithms that allow computers to learn to perform tasks, identify relationships, and discern patterns without the need for humans to provide the underlying instructions. Conventional algorithms operate by sequentially executing a preprogrammed set of rules to achieve a particular outcome. Machine learning algorithms, by contrast, are instead provided with a set of examples by the user and *train themselves to learn the rules from the data*. This powerful idea dates back to at least the 1950s, but has only been fully realized in recent years with the advent of sufficiently large digital data sets over which to perform training—for example, Google photo albums, Amazon shopping lists, Netflix viewing histories—and sufficiently powerful computer hardware and algorithms to perform the training—typically powerful graphics cards developed for the computer game industry that can be hijacked to conduct machine learning. This paradigm has revolutionized multiple domains of science and technology, with different variants of machine learning dominating, and in some cases enabling, multifarious applications such as retail recommendation engines, facial detection and recognition, language translation, autonomous and assisted driving, spam filtering, and character recognition. The success of these algorithms may be largely attributed to their enormous flexibility and power to extract patterns, correlations, and structure from data. These features can be nonintuitive and complicated functions that are difficult for humans to parse, or exist as weak signals that are only discernible from large, high-dimensional data sets that defy conventional analysis techniques.

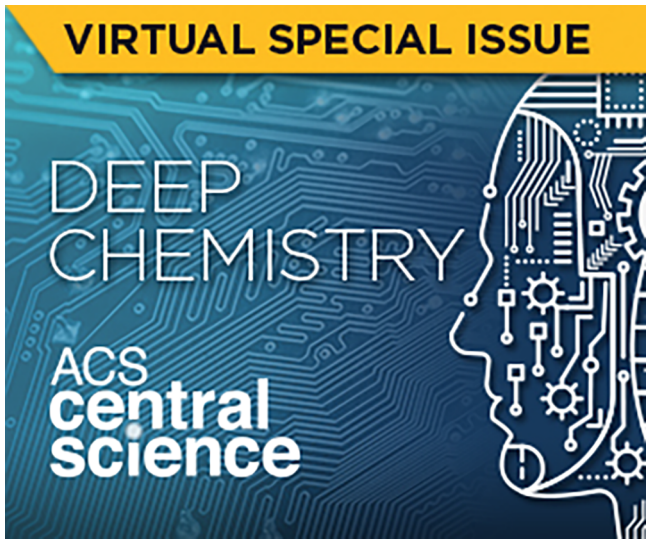
There remains a fundamental difference between artificial and human intelligence—no machine has yet exhibited generic human cognition, and for now, the Turing Test remains intact[1]—but machine performance in certain specific tasks is unequivocally superhuman. A prominent example is provided by Google's Go-playing computer program AlphaGo Zero. This program was provided only with the rules of the ancient board game and learned to play by playing games against itself in a form of reinforcement learning.[2] After just 3 days of training, AlphaGo Zero roundly defeated the best previous best algorithm (AlphaGo Lee) that had itself beaten the 18-time (human) world champion Lee Sedol 100 games to 0.[3] Remarkably, AlphaGo Zero employed previously unknown strategies of play that had never been discovered by human players over the 2500 year history of the game.

Machine learning is also advancing into many aspects of scientific inquiry, and the chemical sciences stand in the vanguard through the establishment of new tools and paradigms with which to engage important problems in molecular design, quantum chemistry, molecular structure prediction, and organic synthesis. The power and potential of these new techniques is hard to overestimate. In a twist on Eugene Wigner's famous 1960 paper *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*,[4] Alon Halevy, Peter Norvig, and Fernando Pereira assert that instead of relying exclusively on the development of ever more sophisticated and elegant theories we should "*embrace complexity and make use of the best ally that we have: the unreasonable effectiveness of data*".[5] All applications of machine learning in chemical science essentially engage this goal by learning to extract models, rules, and predictions from data, but one approach stands out for its remarkable power and flexibility in a diversity of problems—deep neural networks.

Artificial neural networks (ANNs) are a type of machine learning algorithm whose structure and function is loosely based on the architecture of the animal brain. Each artificial

Published: August 8, 2018

DOI: 10.1021/acscentsci.8b00528
ACS Cent. Sci. 2018, 4, 938−941



Featuring three Outlooks, 13 Research Articles and several pieces of editorial content, the Deep Chemistry Virtual Issue demonstrates the vibrant growth in deep and machine learning in chemistry.

neuron represents a mathematical unit that receives, aggregates, and operates on signals from a set of input neurons, and passes the resulting signal onto a group of output neurons. The connecting synapses between neurons amplify or dampen the signals through adjustable weights. Usually the neurons are arranged in layers, with the input layer accepting a representation of the data to be analyzed, a number of hidden layers performing the processing, and an output layer presenting the result. The ANN learns by adjusting the synapse weights to optimize its performance over a training data set provided by the user. Once an ANN is trained and its reliability confirmed on known but independent test data, it can then be employed to make predictions. The flexibility and power of ANNs can be traced to the universal approximation theorem,[6] which, loosely stated, asserts that ANNs with sufficiently many neurons can approximate essentially any mathematical relation between the input and output layers. An ANN is termed "deep" if it contains more than one hidden layer, providing the network with multiple hierarchical layers of abstraction within which to extract patterns and perform computation. The benefit of deep learning is the greater compactness and flexibility per neuron as well as the emergence of latent variables that can be manipulated by the network and sometimes interpreted by human operators. Deep learning has proven to be a powerful approach in a diversity of applications, and there is now a plethora of different deep neural network architectures—convolutional, autoencoding, recurrent, bidirectional, Siamese, and many more—each tailored to possess functionalities suited to particular tasks.

The present virtual issue presents a snapshot of some current applications of machine learning in chemical science with a focus on deep neural networks. The *Research Articles* collected here report exciting progress in a diversity of problems by combining domain expertise with machine learning tools. Swamidass and co-workers employ convolutional neural networks to predict molecular sites of biological reactivity[7] and epoxidation,[8] and introduce novel network architectures to model nonlocal quantum chemical features.[9] In the context of reaction prediction and engineering, Aspuru-Guzik and co-workers[10] and Green and Jensen and co-workers[11] use deep learning to predict the products of organic reactions, Pande and co-workers use recurrent neural networks for retrosynthetic reactant prediction,[12] and Zare and co-workers use deep reinforcement learning to optimize reaction conditions.[13] The problem of drug design is engaged by Waller and co-workers employing recurrent neural networks as generative models,[14] by Aspuru-Guzik and co-workers using encoder-decoder network architectures,[15] and by Pande and co-workers using a novel network architecture to perform one-shot learning.[16] Yang and Gao and co-workers employ Bayesian learning and variational optimization to determine the reaction coordinate for an in-water (retro-)Claisen rearrangement,[17] Pentelute and co-workers use random forest classifiers to predict cell-penetrating peptides to deliver therapeutics,[18] and Aspuru-Guzik and co-workers apply automatic differentiation to compute derivatives in quantum chemical calculations.[19] In *Center Stage*, Neil Savage interviews Alán Aspuru-Guzik about quantum computing, machine learning, and open access.[20] In *First Reactions* Sánchez-Lengeling and Aspuru-Guzik discuss how to train machines to possess chemical intuition.[21] In a triplet of *Outlook* articles, Aspuru-Guzik, Lindh, and Reiher consider the future of computer simulation in quantum chemistry,[22] Ley and co-workers consider technological advances in chemical synthesis,[23] and Cronin and co-workers consider new algorithms for robotic chemical discovery.[24]

The banner successes of machine learning in chemical science—high-throughput molecular screening, drug design, force-field development—are attracting ever more researchers to apply these tools to ever more areas at an ever quickening pace. What advances in this space might we anticipate in the coming years?

From a technical perspective, the immediate frontiers in machine learning likely lie in physics-aware artificial intelligence (PAI) and explainable artificial intelligence (XAI). As elegantly laid out in a recent DARPA announcement, the development of AI technology may be considered as a series of waves.[25] The first wave lies in the past and concerned the development of rule-based expert systems; the second wave

DOI: 10.1021/acscentsci.8b00528
ACS Cent. Sci. 2018, 4, 938−941

is our present deployment of machine learning to learn rules by statistical data analysis; the third wave is the future development of PAI technologies that learn through explanatory models with the relevant physics "baked in". These PAI technologies promise to deliver superior performance by constraining the model to adhere to physical laws (e.g., conservation equations, symmetries) and cope better with sparse and/or noisy data. XAI concerns the development of machine learning models that come equipped with human comprehensible explanations of their predictions and actions.[26] Accurate predictive performance and ease of interpretability frequently stand in conflict, and it is the goal of XAI to marry the interpretability of simple older models (e.g., multiple linear regression) with the power of more complex but less scrutable modern approaches (e.g., deep neural networks). Opaque high-performance models may be adequate for many applications, but increasing model complexity has given rise to an increasing need for the machine to tell us how it got to the answer it did. Providing this rationalization can be critical in ensuring that we do not erroneously overextrapolate and can trust and substantiate the model predictions. Comprehensible explanations can be absolutely critical for particular tasks to ensure that we are getting the right answer for the right reasons (e.g., medical diagnosis), and it is unlikely that machine learning tools will become an accepted tool in these domains until XAI becomes sufficiently mature. Understanding how the machines "think" may tell us how to better understand the system at hand and maybe even teach us something about human cognition, a position vociferously advocated for in Douglas Hofstadter's entreaty "*Why conquer a task if there's no insight to be had from the victory?*".[27] Engaging the goals of PAI and XAI will likely involve the establishment of fundamentally new machine learning models and architectures as well as substantial retrofitting of existing techniques, the development of novel model analysis protocols, and the hierarchical nesting of machine learning models of varying complexity.

From a cultural and educational standpoint, machine learning approaches will be democratized and made broadly available through cheaper and more powerful graphics processing unit (GPU) hardware, the development of user-friendly software, and access to larger and more freely available databases. Data science training will become more tightly integrated into disciplinary training at the undergraduate and graduate levels, and there will be a proliferation of master's degree programs focusing on data science and machine learning. Barriers will be broken down between chemical science and data science through these curricular changes, and also through workshops, conferences, and hackathons designed to bring these communities together. Ultimately, the boundary between disciplinary and data science will become blurred. These trends will conspire to make machine learning a ubiquitous and indispensable tool, with artificial intelligence working side-by-side with human practitioners akin to the role played by the slide rule, scientific calculator, and personal computer in their own ages. In their respective *Outlook* articles, Aspuru-Guzik, Lindh, and Reiher posit a "Chemical Turing Test" wherein communication with an artificial intelligence environment is indistinguishable from communicating with an expert chemist,[22] and Cronin and co-workers consider the potential for intelligent chemical robots with a real-time feedback loop between computational data analysis and automated experimentation.[24] Perhaps it is not such a jump to contemplate a future confluence of these advances to produce intelligent robotic lab assistants that can teach themselves particular aspects of chemistry to attain superhuman performance in the mold of AlphaGo Zero? Beyond the realm of chemical science, is it so far-fetched to think of deep learning technologies helping lawyers to argue, composers to score, philosophers to reason, and artists to create? The age of machine learning in chemical science is upon us and it will leave few areas of our discipline untouched. This special collection highlights just the tip of iceberg, and we can look forward to many exciting innovations and developments in the years to come.

Andrew L. Ferguson

Institute for Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States

Author Information

E-mail: andrewferguson@uchicago.edu.

ORCID

Andrew L. Ferguson: 0000-0002-8829-9726

Notes

Views expressed in this editorial are those of the author and not necessarily the views of the ACS.

REFERENCES

(1) Turing, A. M. Computing machinery and intelligence. In *Parsing the Turing Test*; Epstein, R., Roberts, G., Beber, G., Eds.; Springer, 2009; pp 23−65. DOI: 10.1007/978-1-4020-6710-5_3.

(2) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of Go without human knowledge. *Nature* **2017**, *550*, 354−359.

(3) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484−489.

DOI: 10.1021/acscentsci.8b00528
ACS Cent. Sci. 2018, 4, 938−941

(4) Wigner, E. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics* **1960**, *13* (1), 1−14.
(5) Halevy, A.; Norvig, P.; Pereira, F. The unreasonable effectiveness of data. *IEEE Intelligent Systems* **2009**, *24* (2), 8−12.
(6) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **1991**, *4* (2), 251−257.
(7) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. Modeling reactivity to biological macromolecules with a deep multitask network. *ACS Cent. Sci.* **2016**, *2* (8), 529−537.
(8) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS Cent. Sci.* **2015**, *1* (4), 168−180.
(9) Matlock, M. K.; Dang, N. L.; Swamidass, S. J. Learning a local-variable model of aromatic and conjugated systems. *ACS Cent. Sci.* **2018**, *4* (1), 52−62.
(10) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725−732.
(11) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **2017**, *3* (5), 434−443.
(12) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **2017**, *3* (10), 1103−1113.
(13) Zhou, Z.; Li, X.; Zare, R. N. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent. Sci.* **2017**, *3* (12), 1337−1344.
(14) Segler, M. H.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **2018**, *4* (1), 120−131.
(15) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez- Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(16) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **2017**, *3* (4), 283−293.
(17) Zhang, J.; Zhang, Z.; Yang, Y. I.; Liu, S.; Yang, L.; Gao, Y. Q. Rich dynamics underlying solution reactions revealed by sampling and data mining of reactive trajectories. *ACS Cent. Sci.* **2017**, *3* (5), 407−414.
(18) Wolfe, J. M.; Fadzen, C. M.; Choo, Z.-N.; Holden, R. L.; Yao, M.; Hanson, G. J.; Pentelute, B. L. Machine learning to predict cell-penetrating peptides for antisense delivery. *ACS Cent. Sci.* **2018**, *4* (4), 512−520.
(19) Tamayo-Mendoza, T.; Kreisbeck, C.; Lindh, R.; Aspuru-Guzik, A. Automatic differentiation in quantum chemistry with applications to fully variational Hartree−Fock. *ACS Cent. Sci.* **2018**, *4*, 559.
(20) Savage, N. *ACS Cent. Sci.* **2015**, *1*, 58.
(21) Sánchez-Lengeling, B. and Aspuru-Guzik. Learning more, with less. *ACS Cent. Sci.* **2017**, *3* (4), 275−277.
(22) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The matter simulation (r)evolution. *ACS Cent. Sci.* **2018**, *4* (2), 144−152.
(23) Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. Enabling technologies for the future of chemical synthesis. *ACS Cent. Sci.* **2016**, *2* (3), 131−138.
(24) Henson, A. B.; Gromski, P. S.; Cronin, L. Designing algorithms to aid discovery by chemical robots. *ACS Cent. Sci.* **2018**, *4*, 793.
(25) Disruption Opportunity Special Notice The Physics of Artificial Intelligence (PAI) DARPA-SN-18-65. https://www.fbo.gov/spg/ODA/DARPA/CMO/DARPA-SN-18-65/listing.html.
(26) DARPA Broad Agency Announcement Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53. https://www.fbo.gov/spg/ODA/DARPA/CMO/DARPA-BAA-16-53/listing.html.
(27) Somers, J. The Man Who Would Teach Machines to Think. *The Atlantic*, Nov, 2013.

DOI: 10.1021/acscentsci.8b00528
ACS Cent. Sci. 2018, 4, 938−941