ACS central science

FIRST REACTIONS

Learning More, with Less

Benjamín Sánchez-Lengeling and Alán Aspuru-Guzik

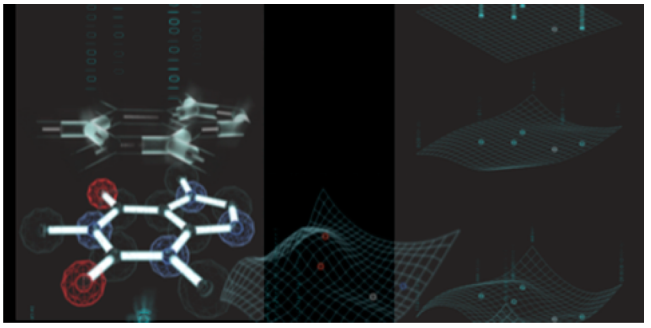
Department of Chemistry, Chemical Biology, 12 Oxford Street, Harvard University, Cambridge, Massachusetts 02138, United States

Teaching computer algorithms "chemical intuition" can improve their predictions.

Practicing chemists solve problems via "chemical intuition", a quality that lets them skip intermediate details and get to the essential result, even if the outcome is counterintuitive to the uninitiated. There is no human shortcut to building this intuition; chemists hone their skills through years of experience of learning and memorizing patterns of molecular structure and reactivity. It is in this spirit that Vijay Pande and co-workers propose in "Low Data Drug Discovery with One-Shot Learning" in this issue of *ACS Central Science*[1] a computational approach for chemical prediction by learning from a low number of examples. The paper touches on many central themes that are relevant to the intersection of the three main components of computation in chemistry: molecular representation, chemical space exploration, and machine learning to accelerate computation.

For discovering new molecules, the enormity of chemical space cannot be understated; the number of "small" to "medium" sized molecules is estimated to be in the range of 10^{60} to 10^{180},[2] a number that is a hundred orders of magnitude larger than the number of atoms in the visible universe. With just a considerably small number of examples, chemists are able to distinguish and assess the potential function of a molecule for a given task. For example, we recently created a "Molecular Tinder" application that helped us in the design of molecules for organic displays.[3] In analogy to the dating application, Molecular Tinder was a voting system that allowed us to harvest information from experimentalists who voted "Yes", "No", or "Maybe" on the synthesizability of molecules. Voting results allowed us to design algorithms that preferentially generated molecules with practical synthesic routes that were eventually synthesized and tested in real devices.[3]

Another very important aspect of human intuition is "transferability", which enables the generalization of knowledge learned in a particular domain to untested domains. Everyone who has passed an undergraduate organic chemistry test had to show that their brain is able to generalize from one domain to the other. This is a much more challenging task for a computer.



Han Altae-Tran

For example, we recently created a "Molecular Tinder" application that helped us in the design of molecules for organic displays.

We are sometimes able to predict with varying degrees of success these properties using quantum chemistry calculations, but when these simulations are involved, supralinear computational scaling laws hinder the application of most common algorithms to complex molecules. Therefore, to cover chemical space efficiently, we cannot go unaided by intuition if we ever hope to explore it for successful molecular design.

It is often thought in the artificial intelligence (AI) community that any human decision that can be done in a matter of a few seconds, can be in theory, learned and automated by a computer. There have been many recent examples where deep learning is solving increasingly complex tasks and getting closer to the performance of humans, even surpassing it in certain tasks such as the game Go with AlphaGo.[4] This progress has been propelled mainly by two factors: broader availability of data and cheaper

Published: April 18, 2017

DOI: 10.1021/acscentsci.7b00153
ACS Cent. Sci. 2017, 3, 275−277

computation. In part because we now automatically collect vast amounts of data on just about anything that can be digitized: photos, text, sound, voice, health records, GPS locations, and of course, molecules. With larger data sets there is much more potential to develop automated algorithms that turn this data into information and eventually into insight.

Therefore, to cover chemical space efficiently, we cannot go unaided by intuition if we ever hope to explore it for successful molecular design.

But what can be done when data is sparse? The algorithm of Pande and co-workers[1] is the first application of "one-shot learning" to chemistry. There are three key ingredients for the success of one shot-learning. First, one-shot learning overcomes the sparsity of the training data set by learning a similarity metric between molecules. To make this similarity transferrable, the second ingredient is making it a metric that is also related to their performance over several tasks. Finally, one-shot learning requires a flexible and meaningful data representation. They demonstrate this principle in a very challenging setting, using up to 10 positive and 10 negative molecules, rated based on their performance in a particular property of interest (activity/inactivity, etc.). Using data sets Tox21, MUV, and SIDER which relate to drug side effects, they show remarkably that the models are able to generalize. Models that Pande and co-workers trained on similarly related data sets are shown to be transferable to a certain degree, outperforming common ensemble methods such as random forests.

It will be interesting to see in the future how data sets as small as tens of molecules to large data sets of up to millions of data points are leveraged for prediction. The field of transfer learning also may enable the eventual use pretrained models on a variety of applications for which the original training was not directly intended.

One-shot learning employs aspects of a general class of machine learning algorithms called "attention mechanism" algorithms. These algorithms allow the mapping between chemical compounds into a continuous space. In this space, a metric between molecules can be tuned to a particular task. Recently, it was pointed out that one way of interpreting attention mechanisms is to relate them to the general concept of memory-augmented neural networks. By attending or focusing on certain parts of the data, the network is choosing what to observe from memory.[5]

It will be interesting to see in the future how data sets as small as tens of molecules to large data sets of up to millions of data points are leveraged for prediction.

Looking into the future, memory-augmented neural networks is one frontier of AI. By using the concept of memory, neural networks are able to crack previously unsolved complex and structured tasks.[6] It is reasonable to hypothesize that to solve hard chemical problems, we inherently need to store important examples or features for later recall. Hence memory-augmented neural networks, differentiable neural computers, neural Turing machines,[6] and other related algorithms will push the frontiers of prediction in chemistry.

Pande and co-workers employ graph convolutional networks (GCN)[7,8] in matching networks for molecular features which also opens the door to solving chemical problems in new ways. Molecular representation is still an active area of research. A good universal representation of a molecule should contain many of the symmetries on which its properties are invariant, typically permutation and isometry invariance for energetic properties. A further complication is the consideration of stereochemistry, several conformers, and overall compactness of the representation. To encourage these properties, most existing work has used a combination of topological features that encode molecular subgraph environments (fingerprint-type methods such as Morgan fingerprints[9]) and geometrical features such as bonds, angles, and physical interactions (Coulomb matrices, bag of bonds,[10] etc.). GCNs are able to encode information in the edges and nodes of each graph, holding topological, geometrical, and other chemically specific information, which ultimately might lead to a flexible, compressed, and optimized representation suited for each problem domain.

The future will keep both the chemistry and machine learning communities busy. There is still work to be done on the interpretability of GCNs. Together with our recent use of autoencoders[11] to optimize molecular properties in a generalizable setting, the continuous-space representation of molecules is an exciting direction for chemistry.

Another important frontier is the interaction and control of experiments with ML tools. Recent work by Raccuglia et al.[12] with the dark reaction projects shows how AI might be used in a chemist's toolbox, improving how we execute our science and collect our data. We look forward to the day where AI is blended in most aspects of chemical research.

DOI: 10.1021/acscentsci.7b00153
ACS Cent. Sci. 2017, 3, 275−277

As a final note, kudos are due to Pande and co-workers for releasing their code and training data sets as open source, as well as posting their manuscripts in preprint servers. The authors believe that all card-carrying modern theoretical researchers in the field should do the same. To preempt Twitter wars, we acknowledge that not all data sets, e.g., pharmaceutical or materials-related, can be made public due to IP considerations.

ACKNOWLEDGMENTS

We acknowledge the generous support of Dr. Anders G. Frøseth for our work on machine learning.

Author information

E-mail: aspuru@chemistry.harvard.edu.
Twitter: @A_Aspuru_Guzik.

ORCID

Alán Aspuru-Guzik: 0000-0002-8277-4434

REFERENCES

(1) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, DOI: 10.1021/acscentsci.6b00367.
(2) Kirkpatrick, P.; Ellis, C. *Nature* **2004**, *432*, 823.
(3) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Aspuru-Guzik, A.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15* (10), 1120−1127.
(4) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521* (7553), 436−444.
(5) Olah, C.; Carter, S. Attention and Augmented Recurrent Neural Networks. *Distill*, *1*(9), **2016**, 10.23915/distill.00001
(6) Graves, A.; Wayne, G.; Reynolds, M.; Harley, T.; Danihelka, I.; Grabska-Barwińska, A.; Hassabis, D.; et al. Hybrid computing using a neural network with dynamic external memory. *Nature* **2016**, *538* (7626), 471−476.
(7) Kipf, T. N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks, 2016. Retrieved from http://arxiv.org/abs/1609.02907.
(8) Gómez-Bombarelli, R.; Duvenaud, D.; Hernández-Lobato, J. M.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. Preprint: http://arxiv.org/abs/1610.02415.
(9) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742−754.
(10) Huang, B.; von Lilienfeld, O. A. Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
(11) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 2215−2223.
(12) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533* (7601), 73−6.

DOI: 10.1021/acscentsci.7b00153
ACS Cent. Sci. 2017, 3, 275−277