CHEMISTRY OF MATERIALS
Cite This: Chem. Mater. 2019, 31, 9579−9581
Editorial
pubs.acs.org/cm

Five High-Impact Research Areas in Machine Learning for Materials Science

Over the past several years, the field of materials informatics has grown dramatically.[1] Applications of machine learning (ML) and artificial intelligence (AI) to materials science are now commonplace. As materials informatics has matured from a niche area of research into an established discipline, distinct frontiers of this discipline have come into focus, and best practices for applying ML to materials are emerging.[2] The purpose of this editorial is to outline five broad categories of research that, in my view, represent particularly high-impact opportunities in materials informatics today:

- *Validation by experiment or physics-based simulation.* One of the most common applications of ML in materials science involves training models to predict materials properties, typically with the goal of discovering new materials. With the availability of user-friendly, open-source ML packages such as `scikit-learn`,[3] `keras`,[4] and `pytorch`,[5] the process of training a model on a materials data set—which requires only a few lines of python code—has become completely commoditized. Thus, standard practice in designing materials with ML should include some form of validation, ideally by experiment[6−8] or, in some cases, by physics-based simulation.[9,10] Of particular interest are cases in which researchers use ML to identify materials whose properties are superior to those of *any* material in the initial training set;[11] such extraordinary results remain scarce.
- *ML approaches tailored for materials data and applications.* This category encapsulates a diverse set of method development activities that make ML more applicable to and effective for a wider range of materials problems. Materials science as a field is characterized by small, sparse, noisy, multiscale, and heterogeneous multidimensional (e.g., a blend of scalar property estimates, curves, images, time series, etc.) data sets. At the same time, we are often interested in exploring very large, high-dimensional chemistry and processing design spaces. Some method development examples to address these challenges include new approaches for uncertainty quantification (UQ),[12] extrapolation detection,[13] multi-property optimization,[14] descriptor development (i.e., the design of new materials representations for ML),[15−17] materials-specific cross-validation,[18,19] ML-oriented data standards,[20,21] and generative models for materials design.[22]
- *High-throughput data acquisition capabilities.* ML is notoriously data-hungry. Given the typically very high cost of acquiring materials data, both in terms of time and money, the materials informatics field is well-served by research that accelerates and democratizes our ability to synthesize, characterize, and simulate materials. Examples include high-throughput density functional theory calculations of materials properties,[23−25] applications of robotics, automation, and operations research to materials science,[26−30] and natural language processing (NLP) to extract materials data from text corpora.[31,32]
- *ML that makes us better scientists.* A popular refrain in the materials informatics community is that "ML will not replace scientists, but scientists who use ML will replace those who do not." This *bon mot* suggests that ML has the potential to make scientists more effective and enable them to do more interesting and impactful work. We are still in the nascent stages of creating true ML-based copilots for scientists, but research areas such as ML model explainability and interpretability[33,34] represent a valuable early step. Another example is the application of ML to accelerate or simplify materials characterization. Researchers have used deep learning to efficiently post-process and understand images generated via existing characterization methods such as scanning transmission electron microscopy (STEM)[35] and position averaged convergent beam electron diffraction (PACBED).[36]
- *Integration of physics within ML, and ML with physics-based simulations.* The paucity of data in many materials applications is a strong motivator for formally integrating known physics into ML models. One approach to embedding physics within ML is to develop methods that guarantee certain desirable properties by construction, such as respecting the invariances present in a physical system.[37] Another strategy is to use ML to model the difference between simulation outputs and experimental results. For example, Google and collaborators created TossingBot, a robotic system that learned to throw objects into bins with the aid of a ballistics simulation.[38] The researchers found that a physics-aware ML approach, wherein ML learned and corrected for the *discrepancy* between the simulations and real-world observations, dramatically outperformed a pure trial-and-error ML training strategy. In a similar vein, ML can enable us to derive more value from existing physics-based simulations. For example, ML-based interatomic potentials[39−41] represent a means of capturing some of the physics of first-principles simulations in a much more computationally efficient model that can simulate orders of magnitude more atoms. ML can also serve as "glue" to link physics-based models operating at various fidelities and length scales.[42]

As ML becomes more widely used in materials research, I expect that efforts addressing one or more of these five themes will have an outsized impact on both the materials informatics discipline and materials science more broadly.

Bryce Meredig*

Citrine Informatics, Redwood City, California 94063, United States

Published: December 10, 2019

DOI: 10.1021/acs.chemmater.9b04078
Chem. Mater. 2019, 31, 9579−9581

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: bryce@citrine.io.

Notes

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

■ ACKNOWLEDGMENTS

I would like to thank C. Borg, E. Kim, J. Ling, J. E. Saal, and J. Tappan for helpful comments on this manuscript.

■ REFERENCES

(1) Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **2016**, *41*, 399−409.

(2) Riley, P. Three pitfalls to avoid in machine learning. *Nature* **2019**, *572*, 27−29.

(3) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825−2830.

(4) Chollet, F. keras. https://github.com/fchollet/keras, 2015.

(5) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In *Neural Information Processing Systems Autodiff Workshop*, 2017.

(6) Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **2016**, *28*, 7324−7331.

(7) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine learning directed search for ultraincompressible, superhard materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844−9853.

(8) Wu, S.; Kondo, Y.; Kakimoto, M.-a.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Computational Materials* **2019**, *5*, 66.

(9) Mannodi-Kanakkithodi, A.; Pilania, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine learning strategy for accelerated design of polymer dielectrics. *Sci. Rep.* **2016**, *6*, 20952.

(10) Sendek, A. D.; Cubuk, E. D.; Antoniuk, E. R.; Cheon, G.; Cui, Y.; Reed, E. J. Machine learning-assisted discovery of solid Li-ion conducting materials. *Chem. Mater.* **2019**, *31*, 342−352.

(11) Rickman, J.; Chan, H.; Harmer, M.; Smeltzer, J.; Marvel, C.; Roy, A.; Balasubramanian, G. Materials informatics for the screening of multi-principal elements and high-entropy alloys. *Nat. Commun.* **2019**, *10*, 2618.

(12) Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; Meredig, B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integrating Materials and Manufacturing Innovation* **2017**, *6*, 207−217.

(13) Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. ChemRxiv, DOI: 10.26434/chemrxiv.7900277.v1, 2019.

(14) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chemical science* **2018**, *9*, 7642−7655.

(15) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

(16) Huo, H.; Rupp, M. Unified representation for machine learning of molecules and crystals. arXiv preprint arXiv:1704.06439, 2017.

(17) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2018**, *2*, 083802.

(18) Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering* **2018**, *3*, 819−825.

(19) Lu, H.-J.; Zou, N.; Jacobs, R.; Afflerbach, B.; Lu, X.-G.; Morgan, D. Error assessment and optimal cross-validation approaches in machine learning applied to impurity diffusion. *Comput. Mater. Sci.* **2019**, *169*, 109075.

(20) Michel, K.; Meredig, B. Beyond bulk single crystals: a data format for all materials structure-property-processing relationships. *MRS Bull.* **2016**, *41*, 617.

(21) Lin, T.-S.; Coley, C. W.; Mochigase, H.; Beech, H. K.; Wang, W.; Wang, Z.; Woods, E.; Craig, S. L.; Johnson, J. A.; Kalow, J. A. BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **2019**, *5*, 1523.

(22) 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*, 268−276.

(23) Toher, C.; Plata, J. J.; Levy, O.; De Jong, M.; Asta, M.; Nardelli, M. B.; Curtarolo, S. High-throughput computational screening of thermal conductivity, Debye temperature, and Grüneisen parameter using a quasiharmonic Debye model. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *90*, 174107.

(24) De Jong, M.; Chen, W.; Angsten, T.; Jain, A.; Notestine, R.; Gamst, A.; Sluiter, M.; Ande, C. K.; Van Der Zwaag, S.; Plata, J. J.; et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2015**, *2*, 150009.

(25) Emery, A. A.; Wolverton, C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites. *Sci. Data* **2017**, *4*, 170153.

(26) Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in materials research: a case study in carbon nanotube growth. *npj Computational Materials* **2016**, *2*, 16031.

(27) Sun, S.; Hartono, N. T.; Ren, Z. D.; Oviedo, F.; Buscemi, A. M.; Layurova, M.; Chen, D. X.; Ogunfunmi, T.; Thapa, J.; Ramasamy, S. et al. Accelerating Photovoltaic Materials Development via High-Throughput Experiments and Machine-Learning-Assisted Diagnosis. arXiv preprint arXiv:1812.01025v1, 2018.

(28) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; et al. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **2019**, *363*, eaav2211.

(29) Boyce, B. L.; Uchic, M. D. Progress toward autonomous experimental systems for alloy development. *MRS Bull.* **2019**, *44*, 273−280.

(30) Ortiz, B. R.; Adamczyk, J. M.; Gordiz, K.; Braden, T.; Toberer, E. S. Towards the high-throughput synthesis of bulk materials: thermoelectric PbTe-PbSe-SnTe-SnSe alloys. *Molecular Systems Design & Engineering* **2019**, *4*, 407−420.

(31) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **2017**, *29*, 9436−9444.

(32) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95.

(33) Ribeiro, M. T.; Singh, S.; Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*; 2016; pp 1135−1144.

(34) Ling, J.; Hutchinson, M.; Antono, E.; DeCost, B.; Holm, E. A.; Meredig, B. Building data-driven models with microstructural images:

DOI: 10.1021/acs.chemmater.9b04078

Generalization and interpretability. *Materials Discovery* **2017**, *10*, 19−28.

(35) Ziatdinov, M.; Dyck, O.; Maksov, A.; Li, X.; Sang, X.; Xiao, K.; Unocic, R. R.; Vasudevan, R.; Jesse, S.; Kalinin, S. V. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **2017**, *11*, 12742−12752.

(36) Xu, W.; LeBeau, J. M. A deep convolutional neural network to analyze position averaged convergent beam electron diffraction patterns. *Ultramicroscopy* **2018**, *188*, 59−69.

(37) Ling, J.; Jones, R.; Templeton, J. Machine learning strategies for systems with invariance properties. *J. Comput. Phys.* **2016**, *318*, 22−35.

(38) Zeng, A.; Song, S.; Lee, J.; Rodriguez, A.; Funkhouser, T. TossingBot: Learning to Throw Arbitrary Objects with Residual Physics. arXiv preprint arXiv:1903.11239, 2019.

(39) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(40) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(41) Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Science Advances* **2017**, *3*, e1603015.

(42) Hutchinson, M. L.; Antono, E.; Gibbons, B. M.; Paradiso, S.; Ling, J.; Meredig, B. Overcoming data scarcity with transfer learning. arXiv preprint arXiv:1711.05099, 2017.