Bringing Atomistic Deep Learning to Prime Time

Nathan C. Frey
MIT
ncfrey@mit.edu

Siddharth Samsi
MIT
ssamsi@mit.edu

Bharath Ramsundar
Deep Forest Sciences
bharath@deepforestsci.com

Connor W. Coley
MIT
ccoley@mit.edu

Vijay Gadepally *
MIT
vijayg@mit.edu

Abstract

Artificial intelligence has not yet revolutionized the design of materials and molecules. In this perspective, we identify four barriers preventing the integration of atomistic deep learning, molecular science, and high-performance computing. We outline focused research efforts to address the opportunities presented by these challenges.

1 Bottlenecks

Atomistic deep learning encompasses neural networks that learn representations of matter - from 0D molecules to 1D nanowires, 2D surfaces, 3D proteins, polymers, and crystalline materials [1]. Unlike other canonical application areas of DL like speech and images, data for atomistic systems is expensive to acquire and heterogeneous, with a practically infinite space of emergent properties of interest. AI revolutionized natural language processing and computer vision, but we are still largely waiting for AI to revolutionize the design of matter. Atomistic systems underpin every physical and digital technology, and the potential impact of AI-enabled matter design cannot be overstated. In this perspective, we identify four key challenges (Figure 1) to the widespread adoption of atomistic DL and outline promising paths forward.

The first three challenges are practical difficulties related to robustness and scaling methods to be "production ready." They correspond to the three resources that DL methods traditionally leverage at scale to achieve superhuman success: data, model size, and compute. The fourth challenge is a more fundamental issue related to AI's place in the toolkit for natural scientists.

Data openness In many application areas of atomistic DL, practitioners are data starved. Data remains proprietary to the companies that generate it, or not easily accessible to research groups that did not generate the data. Physical experiments and high-fidelity computational simulations can be expensive and time-consuming, so acquiring more labeled data is often not practical.

*DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

NeurIPS 2021 AI for Science Workshop.

There are a number of promising trends addressing this challenge. Adoption of FAIR (findable, accessible, interoperable, reusable) data principles within the research community [2, 3] makes it easier for researchers to leverage data generated outside their own labs. Databases like the Materials Project [4] and automated experimental validation promise to alleviate the data bottleneck in many domains [5, 6, 7]. Active learning and Bayesian optimization strategies for intelligent sampling are shifting the paradigm from random or brute-force data collection to targeted data acquisition [8, 9]. However, it remains important to continue unfocused, high-throughput approaches that generate datasets to train general purpose models and to explore new applications in an open-ended fashion. Datasets generated through intentional experimental design strategies may only be useful to the groups that generated them (and that have the capabilities to generate them).

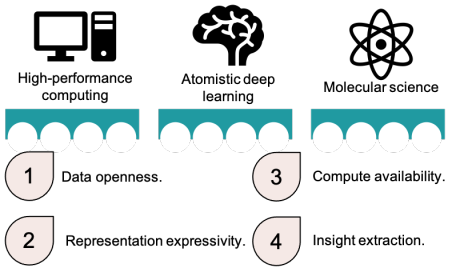

Figure 1: Key challenges for bridging high-performance computing, atomistic deep learning, and molecular science.

Representation expressivity Geometric priors and therefore geometric DL are a natural choice for atomistic systems [10, 11], which can often be represented as graphs of atoms (nodes) and chemical bonds (edges). However, graph neural networks often struggle to compete against human-crafted featurizations like chemical fingerprint vectors and much simpler ML methods like random forests and logistic regression in low-data limits [12, 13]. To achieve widespread adoption, atomistic DL methods and molecular representations must be expressive enough to capture the truly staggering diversity of emergent behavior exhibited by matter [14], while also being robust and easy to use for non-DL experts.

Following the surprising performance of GPT-3 [15], one path forward is to leverage large language models trained on string-based representations of matter [16]. A natural extension is to combine the self-attention mechanism of transformers with the equivariance inherent to molecular graphs [17]. Identifying optimal pre-training tasks is also an ongoing and fruitful area of research [16, 18]. Scaling studies [19] are needed to investigate whether these architectures and representations are already sufficient to handle generic molecular ML tasks, or if further methodological advances are needed.

Compute availability A common thread in headline-grabbing AI achievements is the need for massive compute resources, which are often only available to large corporations. Recent reviews of large language models have argued that non-profit research organizations should also be investigating models at scale [20]. Most atomistic DL methods are not tested beyond small benchmark datasets on a single GPU [18]. Even doing a thorough hyperparameter optimization for a moderately-sized dataset can be a burden. As the business value of atomistic DL grows, large tech companies will become increasingly active in the field. To take two recent examples, industry AI research groups have made exciting contributions to catalyst discovery with the Open Catalyst Project [21] and protein folding with AlphaFold [22]. There is an opportunity to anticipate this trend in atomistic DL and prepare for it, to avoid the paradigm seen in fields like NLP, where fundamental methodological advances are largely left to university research groups and large-scale studies are done by companies.

An emphasis on open science [23] and closer integration between academic research groups and high-performance computing centers, cloud providers, and industry AI groups can alleviate this blocker. Engineering advances like PyTorch Distributed [24] lower the barriers for operating at scale, and HPC centers like MIT Supercloud [25] and NERSC are increasingly prioritizing GPU resources, GPU acceleration, and interactive ML workflows. As scalable model training becomes easier, researchers have the opportunity to investigate emergent behavior in larger models and critically examine where new methods are needed and where scaling existing methods is sufficient.

Insight extraction Scientists and engineers need to *trust* atomistic DL methods to drive further adoption. Human practitioners don't have an intuition about the failure modes and so there is a resistance to conduct physical experiments based solely on model predictions. Trust is built at a practical level when methods are useful, and at a more basic level when methods provide new, surprising, and relevant scientific insights. Black-box models are useful when they automate a tedious task [26] and the bounds of the model's applicability are well-understood. A loftier goal is for

atomistic DL to provide novel insights - discovering new functional forms of matter, uncovering causal or mechanistic processes underpinning empirical observations, etc.

Atomistic DL that discovers not only unexplored, but also synthetically accessible [27, 28, 29, 30], matter will build trust outside the core community of DL-practitioners. Uncertainty quantification of model predictions adds another necessary dimension that will clarify when models are operating outside of their domain of applicability, and uncertainty is a familiar metric for experimentalists to use to assess the trustworthiness of new methods [9, 31].

Conclusion The increasing availability of molecular modeling data and the heterogeneity of datasets and tasks present an exciting opportunity for real-world impact by integrating atomistic DL, high-performance computing, and molecular science. Investigating the basic structure of atomistic DL models and their scaling behavior will accelerate the widespread adoption of these methods and uncover what, if anything, learned representations of matter can tell us about matter itself.

References

[1] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559, 2018. ISSN 14764687. doi: 10.1038/s41586-018-0337-2.

[2] Claudia Draxl and Matthias Scheffler. Nomad: The fair concept for big-data-driven materials science, 2018.

[3] Casper W. Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J. Conduit, Shyam Dwaraknath, Matthew L. Evans, Ádám Fekete, Abhijith Gopakumar, Saulius Gražulis, Andrius Merkys, and et al. Optimade, an api for exchanging materials data. *Scientific Data*, 8(1), Aug 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00974-z. URL http://dx.doi.org/10.1038/s41597-021-00974-z.

[4] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323. URL http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi.

[5] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials*, 4(5):053208, 2016. doi: 10.1063/1.4946894. URL https://aip.scitation.org/doi/abs/10.1063/1.4946894.

[6] Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke. Data-driven materials science: Status, challenges, and perspectives. *Advanced Science*, 6(21):1900808, 2019. doi: https://doi.org/10.1002/advs.201900808. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.201900808.

[7] Connor W. Coley, Dale A. Thomas, Justin A.M. Lummiss, Jonathan N. Jaworski, Christopher P. Breen, Victor Schultz, Travis Hart, Joshua S. Fishman, Luke Rogers, Hanyu Gao, Robert W. Hicklin, Pieter P. Plehiers, Joshua Byington, John S. Piotti, William H. Green, A. John Hart, Timothy F. Jamison, and Klavs F. Jensen. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365, 2019. ISSN 10959203. doi: 10.1126/science.aax1566.

[8] David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science*, 12(22): 7866–7881, 2021. ISSN 2041-6539. doi: 10.1039/d0sc06805e. URL http://dx.doi.org/10.1039/d0sc06805e.

[9] Florian Häse, Matteo Aldeghi, Riley J. Hickman, Loïc M. Roch, and Alán Aspuru-Guzik. Gryffin: An algorithm for bayesian optimization of categorical variables informed by expert knowledge. *Applied Physics Reviews*, 8(3):031406, Sep 2021. ISSN 1931-9401. doi: 10.1063/5.0048164. URL http://dx.doi.org/10.1063/5.0048164.

[10] Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608, Aug 2016. ISSN 1573-4951. doi: 10.1007/s10822-016-9938-8. URL http://dx.doi.org/10.1007/s10822-016-9938-8.

[11] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017. doi: 10.1109/MSP.2017.2693418.

[12] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci.*, 9:5441–5451, 2018. doi: 10.1039/C8SC00148K. URL http://dx.doi.org/10.1039/C8SC00148K.

[13] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1), Sep 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL http://dx.doi.org/10.1038/s41524-020-00406-3.

[14] Bharath Ramsundar, Steven Kearnes, Patrick Riley, Dale Webster, David Konerding, and Vijay Pande. Massively multitask networks for drug discovery, 2015.

[15] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[16] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[17] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020.

[18] Aneesh Pappu and Brooks Paige. Making graph neural networks worth it for low-data molecular machine learning, 2020.

[19] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020.

[20] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia

Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2021.

[21] C. Lawrence Zitnick, Lowik Chanussot, Abhishek Das, Siddharth Goyal, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Thibaut Lavril, Aini Palizhati, Morgane Riviere, Muhammed Shuaibi, Anuroop Sriram, Kevin Tran, Brandon Wood, Junwoong Yoon, Devi Parikh, and Zachary Ulissi. An introduction to electrocatalyst design using machine learning for renewable energy storage, 2020.

[22] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 2021. ISSN 14764687. doi: 10.1038/s41586-021-03819-2.

[23] Michael Woelfle, Piero Olliaro, and Matthew H. Todd. Open science is a research accelerator. *Nature Chemistry*, 3, 2011. ISSN 17554330. doi: 10.1038/nchem.1149.

[24] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. Pytorch distributed: Experiences on accelerating data parallel training, 2020.

[25] Albert Reuther, Jeremy Kepner, Chansup Byun, Siddharth Samsi, William Arcand, David Bestor, Bill Bergeron, Vijay Gadepally, Michael Houle, Matthew Hubbell, Michael Jones, Anna Klein, Lauren Milechin, Julia Mullen, Andrew Prout, Antonio Rosa, Charles Yee, and Peter Michaleas. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pages 1–6. IEEE, 2018.

[26] Rafael Gómez-Bombarelli. Reaction: The near future of artificial intelligence in materials discovery. *Chem*, 4(6):1189–1190, 2018. ISSN 2451-9294. doi: https://doi.org/10.1016/j.chempr.2018.05.021. URL https://www.sciencedirect.com/science/article/pii/S2451929418302304.

[27] Wenhao Gao and Connor W. Coley. The synthesizability of molecules proposed by generative models, 2020.

[28] Nathan C. Frey, Jin Wang, Gabriel Iván Vega Bellido, Babak Anasori, Yury Gogotsi, and Vivek B. Shenoy. Prediction of synthesis of 2d metal carbides and nitrides (mxenes) and their precursors with positive and unlabeled machine learning. *ACS Nano*, 13:3031–3041, 2019. ISSN 1936086X. doi: 10.1021/acsnano.8b08014. URL http://pubs.acs.org/doi/10.1021/acsnano.8b08014.

[29] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, Georgios Markopoulos, Soonok Jeon, Hosuk Kang, Hiroshi Miyazaki, Masaki Numata, Sunghan Kim, Wenliang Huang, Seong Ik Hong, Marc Baldo, Ryan P Adams, and Alán Aspuru-Guzik. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials*, 15:1120, 2016. doi: 10.1038/nmat4717https://www.nature.com/articles/nmat4717#supplementary-information. URL http://dx.doi.org/10.1038/nmat4717.

[30] Martin Seifrid, Riley J. Hickman, Andrés Aguilar-Granda, Cyrille Lavigne, Jenya Vestfrid, Tony C. Wu, Théophile Gaudin, Emily J. Hopkins, and Alán Aspuru-Guzik. Routescore: Punching the ticket to more efficient materials development. *ChemRxiv*, 2021. doi: 10.33774/chemrxiv-2021-k0qx5.

[31] Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning, 2019.