YANGZESHENG (ANDREW) SUN
sun00032@umn.edu | (+1) 612-666-3155
Department of Chemistry, University of Minnesota, Minneapolis MN 55455, USA

Description of Research

Machine Learning on Molecular Simulations for Nanoporous Materials Discovery

Note: This statement mainly pertains to my PhD research conducted at University of Minnesota. It gives an overview about the problem and its significance, briefly summarizes my previous work and related research, and proposes new research directions.

Chemical storage and separation are critical in solving various energy and environmental problems in the real world, and nanoporous materials are believed to be promising candidates for these applications.[1] For example, a material which efficiently adsorbs hydrogen can be made as the fuel tank of a hydrogen-powered vehicle,[2] and a material which separates alcohol from water will greatly benefit the production of biofuels.[3]

By virtue of large-scale computational screening on high-performance computing (HPC) resources, molecular simulations have dramatically accelerated the discovery of nanoporous materials for these crucial applications. The screening workflow contains multiple stages of increasingly accurate simulations on decreasing pools of candidates and has been employed on discovering nanoporous materials for hydrogen storage,[4] methane storage,[5] hydrocarbon separation,[6] and sulfur capture.[7] Typically, a complete screening process towards a target application consists of 10^4–10^6 simulation runs and consumes 10^5–10^7 CPU hours,[8] and may generate terabytes of simulation trajectory data. Therefore, it is of great interest to utilize machine learning (ML) methods to further improve such workflows. A sufficiently-trained ML model performs fast and accurate evaluations, thus facilitating prediction tasks otherwise intractable to pure molecular simulation. The following parts of the statement will describe the present and future research towards this objective, and are divided into two thrusts, *surrogate model* and *data-driven simulation*.

Surrogate model refers to a ML model equivalent to a molecular simulation in its input parameters and output physical properties. More generally, the "surrogate model" term originates from optimization theory where a more readily computable model is employed to approximate an intractable optimization problem.[9] For nanoporous materials discovery, the optimization problem is to maximize the application metric with respect to the material structure and/or thermodynamic conditions. Since molecular simulations are usually treated as black-box models, gradient-based optimization is in most cases impossible. However, with a neural network accurately approximating the simulation results, one can readily optimize the metric on the neural network and then perform a minimal amount of simulations to validate the prediction. Physics-informed inductive biases or learning algorithms are crucial to construct surrogate models because physical consistency can greatly improve the model's generalization and transferability to new systems. These inductive biases can be encoded in the network architecture or serve as regularization or the entire objective of ML models.[10–12] For predicting the adsorption of guest molecules in nanoporous materials, domain-specific models (adsorption isotherms) widely used in chemistry and engineering have very similar mathematical forms as logistic regression,[13] thus neural networks with sigmoid outputs naturally contain inductive biases to predict adsorption in nanoporous materials. As molecular simulations are based on statistical mechanics, it is also possible to derive a learning algorithm which minimizes the KL divergence between the "true" statistical mechanics distribution and the approximating distribution parametrized by the neural network. In complex adsorption systems where domain-specific models fail to predict the simulation results, a neural network trained in such manner was able to accurately approximate the simulations to the same magnitude as their precision.[14] Besides using physical principles, meta-learning can also be effective in directly learning the inductive biases using simulation data from multiple materials at multiple thermodynamic states. Through a meta-learning network, the parameters for the surrogate model of each nanoporous material were jointly learned among many materials, significantly improving its extrapolation ability in the state space. Meta-learning also enables few-shot prediction of a new material based on a limited number of simulations and is transferable from simulation data to experimental data.[15]

Data-driven simulation refers to a molecular simulation whose algorithms directly and actively involve ML models. In contrast to a completely physical simulation in which the simulation algorithm is independent to the data generated, a data-driven simulation becomes more accurate or computationally efficient with increasing amounts of data available through previous simulations. Performing high-quality simulations for complex systems and processes is an indispensable part of the vision to “make the world computable”.[16] Nevertheless, these simulations are especially challenging because of 1) the cost and poor scaling of quantum chemistry calculations in large systems involving quantum effects or chemical reactions[17] and 2) the difficulty in sampling microscopic configurations of complex molecules and materials with structural constraints and free energy barriers.[18] In the first case, neural network atomistic potentials are commonly employed to achieve quantum mechanical accuracy of the potential energy surface of the system with few orders of magnitude less computational cost,[19] thus data-driven simulations with neural network potentials can be performed by training the model on the fly during a first principles molecular simulation.[20,21] First principles Monte Carlo simulations have been an effective approach to understanding reactive adsorption in nanoporous materials.[22] Given the structural symmetry and spatial confinement of nanoporous materials, it is highly possible that repeated quantum chemistry calculations are performed for similar chemical environments, opening up great potentials for these simulations to be accelerated by machine learning. In the second case, machine-learned sampling methods and generative models for microscopic configurations can be a promising approach to address the problem. When the result structure for the molecular simulation is known, such as in protein folding, a neural energy function can be learned end-to-end which are free of energy barriers, and the predicted protein structure was obtained through a simulation using the learned energy function.[23] For other complex systems with thermodynamic properties of interest, including nanoporous materials, there is no ground truth configuration and the simulation results are obtained from sampling the equilibrium distribution of configurations. Generative models based on neural networks are known to be powerful in modeling complex distributions such as natural images,[24] therefore it is of great interest to leverage them for generating microscopic systems and simulation trajectories. One of the groundbreaking works in this aspect is the Boltzmann generator.[18] The Boltzmann generator is an invertible neural network transforming between the microscopic configurational distribution and a Gaussian latent distribution, so that new configurations for a system can be efficiently sampled from the latent distribution. In this way, data-driven simulations can be directly performed through dynamics or Monte Carlo sampling in the latent coordinates. Similar approaches also include using machine learning to find collective variables for molecular simulations.[25] The data-driven simulation methodology can greatly benefit simulations for nanoporous materials where phase changes of the adsorbed molecule exist.[26] These phase changes are important in studying the adsorption mechanisms of nanoporous materials, while enhanced sampling methods and special simulation setups are usually required.

References

1. Li, B., Wen, H. M., Zhou, W. & Chen, B. Porous metal-organic frameworks for gas storage and separation: What, how, and why? *Journal of Physical Chemistry Letters* vol. 5 3468–3479 (2014).
2. Broom, D. P. *et al.* Outlook and challenges for hydrogen storage in nanoporous materials. *Appl. Phys. A Mater. Sci. Process.* **122**, 1–21 (2016).
3. Sreekumar, S., Baer, Z. C., Pazhamalai, A., Gunbas, G., Grippo, A., Blanch, H. W., Clark, D. S. & Toste, F. D. Production of an acetone-butanol-ethanol mixture from Clostridium acetobutylicum and its conversion to high-value biofuels. *Nat. Protoc.* **10**, 528–537 (2015).
4. Colón, Y. J., Fairen-Jimenez, D., Wilmer, C. E. & Snurr, R. Q. High-throughput screening of porous crystalline materials for hydrogen storage capacity near room temperature. *J. Phys. Chem. C* **118**, 5383–5389 (2014).
5. Simon, C. M., Kim, J., Gomez-Gualdron, D. A., Camp, J. S., Chung, Y. G., Martin, R. L., Mercado, R., Deem, M. W., Gunter, D., Haranczyk, M., Sholl, D. S., Snurr, R. Q. & Smit, B. The materials genome in action: Identifying the performance limits for methane storage. *Energy Environ. Sci.* **8**, 1190–1199 (2015).
6. Chung, Y. G., Bai, P., Haranczyk, M., Leperi, K. T., Li, P., Zhang, H., Wang, T. C., Duerinck, T., You, F., Hupp, J. T., Farha, O. K., Siepmann, J. I. & Snurr, R. Q. Computational Screening of Nanoporous Materials for Hexane and Heptane Isomer Separation. *Chem. Mater.* **29**, 6315–6328 (2017).
7. Shah, M. S., Tsapatsis, M. & Siepmann, J. I. Identifying Optimal Zeolitic Sorbents for Sweetening of Highly Sour Natural Gas. *Angew. Chemie* **128**, 6042–6046 (2016).
8. Bai, P., Jeon, M. Y., Ren, L., Knight, C., Deem, M. W., Tsapatsis, M. & Siepmann, J. I. Discovery of optimal zeolites for challenging separations and chemical transformations using predictive materials modeling. *Nat. Commun.* **6**, (2015).
9. Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R. & Kevin Tucker, P. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences* vol. 41 1–28 (2005).
10. Greydanus, S., Dzamba, M. & Yosinski, J. Hamiltonian Neural Networks. in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)* (2019).
11. Chmiela, S., Tkatchenko, A., Sauceda, H. E., Poltavsky, I., Schütt, K. T. & Müller, K. R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, (2017).
12. Raissi, M., Perdikaris, P. & Karniadakis, G. E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **378**, 686–707 (2019).
13. Foo, K. Y. & Hameed, B. H. Insights into the modeling of adsorption isotherm systems. *Chemical Engineering Journal* vol. 156 2–10 (2010).
14. Sun, Y.-Z.-S., DeJaco, R. F. & Siepmann, J. I. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. *Chem. Sci.* **10**, (2019).
15. Sun, Y.-Z.-S., DeJaco, R. F. & Siepmann, J. I. Predicting hydrogen storage in nanoporous materials using meta-learning. in *Machine Learning and the Physical Sciences Workshop, NeurIPS 2019* (2019).
16. Eggimann, B. L., Sun, Y.-Z.-S., DeJaco, R. F., Singh, R., Ahsan, M., Josephson, T. R. & Siepmann, J. I. Assessing the Quality of Molecular Simulations for Vapor-Liquid Equilibria: An Analysis of the TraPPE Database. *J. Chem. Eng. Data* (2019) doi:10.1021/acs.jced.9b00756.
17. Fetisov, E. O., Kuo, I.-F. W., Knight, C., VandeVondele, J., Van Voorhis, T. & Siepmann, J. I. First-Principles Monte Carlo Simulations of Reaction Equilibria in Compressed Vapors. *ACS Cent. Sci.* **2**, 409–415 (2016).
18. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science.* **365**, 1001 (2019).
19. Behler, J. Neural network potential-energy surfaces in chemistry: A tool for large-scale

simulations. *Physical Chemistry Chemical Physics* vol. 13 17930–17955 (2011).

20. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, (2015).

21. Botu, V. & Ramprasad, R. Adaptive machine learning framework to accelerate *ab initio* molecular dynamics. *Int. J. Quantum Chem.* **115**, 1074–1083 (2015).

22. Fetisov, E. O., Shah, M. S., Knight, C., Tsapatsis, M. & Siepmann, J. I. Understanding the Reactive Adsorption of H_2S and CO_2 in Sodium-Exchanged Zeolites. *ChemPhysChem* **19**, 512–518 (2018).

23. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. Learning protein structure with a differentiable simulator. in *7th International Conference on Learning Representations (ICLR 2019)* (2019).

24. Brock, A., Donahue, J. & Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. in *7th International Conference on Learning Representations (ICLR 2019)* (2019).

25. Rogal, J., Schneider, E. & Tuckerman, M. E. Neural-Network-Based Path Collective Variables for Enhanced Sampling of Phase Transformations. *Phys. Rev. Lett.* **123**, (2019).

26. Dejaco, R. F., Elyassi, B., Dorneles De Mello, M., Mittal, N., Tsapatsis, M. & Siepmann, J. I. Understanding the unique sorption of alkane- α, ω -diols in silicalite-1. *J. Chem. Phys.* **149**, (2018).