https://doi.org/10.1038/s44286-023-00010-4

Data-driven molecular design and simulation in modern chemical engineering

Thomas E. Gartner III, Andrew L. Ferguson & Pablo G. Debenedetti

Opportunities and challenges in data-driven chemical engineering thermodynamics, statistical mechanics and molecular simulation are discussed, and new possibilities offered by machine learning in these areas are assessed. Examples suggest how integration of data science and molecular simulation can prove impactful for the future of chemical engineering.

Statistical mechanics and molecular simulation have a long and fruitful relationship with chemical engineering. Beginning with their first applications to condensed matter in the 1950s, molecular simulations have deepened our understanding of the microscopic basis underlying processes and key topics of interest to chemical engineers, including phase equilibria, fluid-phase properties and mixture thermodynamics. The modern discipline of chemical engineering[1] has broadened to include biomolecular engineering, cellular and tissue engineering, advanced materials synthesis and processing, and sustainability, among other topics, resulting in a rejuvenated and essential profession, well suited to contribute to the solution of some of the most pressing global challenges facing humanity today. Computational approaches, like the chemical engineering discipline itself, have similarly expanded in power and scope. One driver of continued advances in computational chemical engineering has been the introduction of data-driven approaches to model and methods development, data analysis, and to connecting simulation to experiment. While interest in this area has grown dramatically, the application of machine learning (ML) and data science in modern chemical engineering is still in its relative infancy. In this Comment, we will discuss opportunities and challenges in data-driven chemical engineering thermodynamics, statistical mechanics and molecular simulation. We aim to provide an assessment of new opportunities offered by ML in the abovementioned areas, identifying especially promising opportunities and examples. We will frame our discussion with examples drawn from our own research areas in which ML has played an enabling and indispensable role, namely, the properties and metastable phase behavior of water, and data-driven protein design. We conclude with a forward-looking perspective on how the integration of data science and molecular simulation could prove impactful for the future of chemical engineering.

Data-driven protein design

Proteins are molecular machines that underpin the functions of biology. Ever since Max Perutz's and John Kendrew's Nobel Prize-winning work to solve the structure of hemoglobin and myoglobin, the design of proteins has been of central interest to chemical and biological engineers with applications ranging from reactor engineering to clean energy to public health. Directed evolution (DE) is a powerful protein engineering strategy pioneered by the 2018 Chemistry Nobel Prize laureate, Frances Arnold[2], that introduces random mutations, identifies the

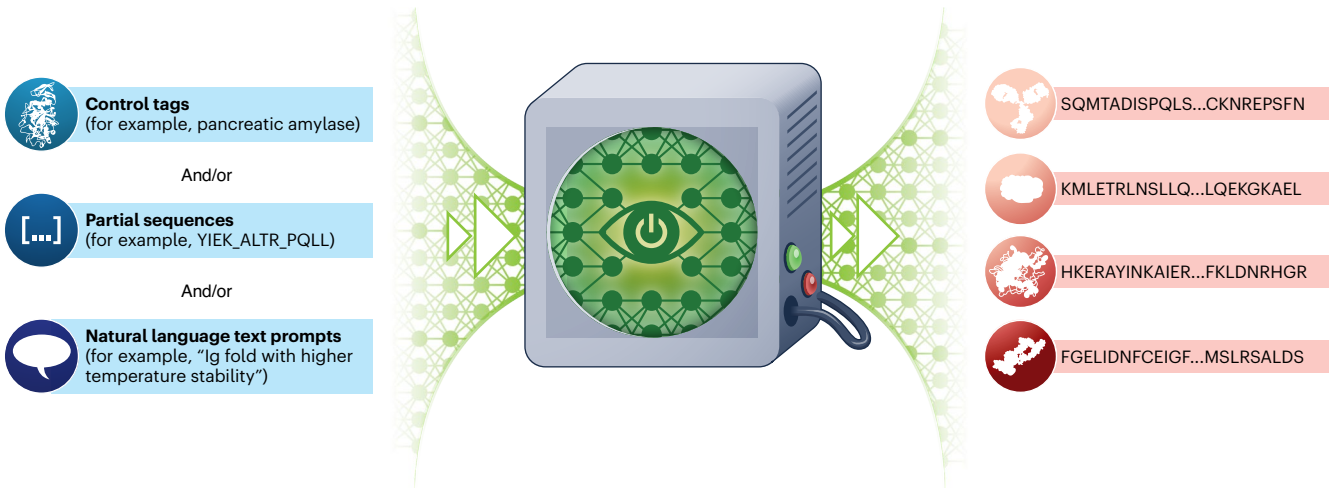

Fig. 1 | Protein language models with text-guided conditioning for data-driven protein design. Schematic illustration of the generic structure of conditional deep generative protein language models. A trained protein language model can be conditioned on control tags, partial sequences and/or natural language text prompts to guide the generation of synthetic protein sequences with desired structure and/or function.

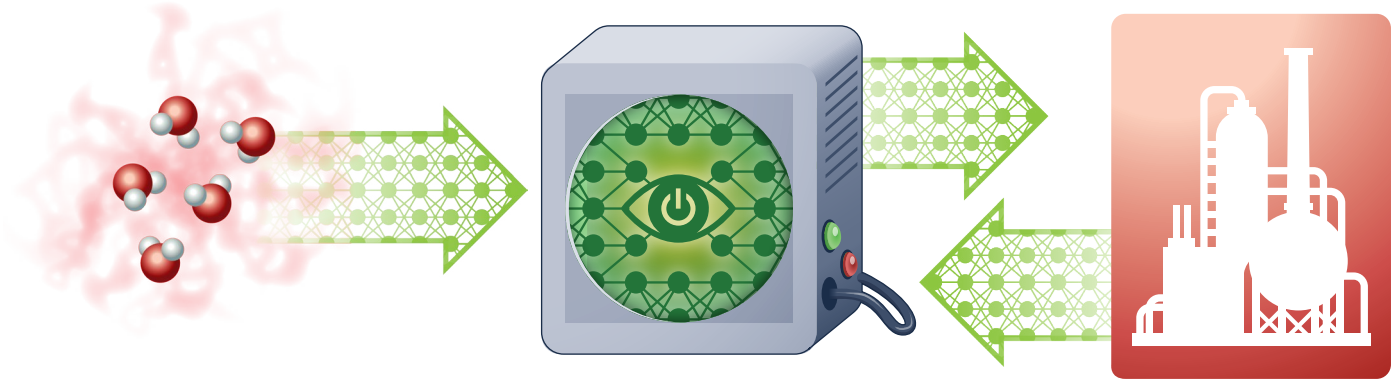

Fig. 2 | Linking molecular and plant scales. ML-enabled force fields have the power to translate ab initio-level information at the atomic scale (left) into large-scale computer simulations (center) that could be directly incorporated into chemical process design and control paradigms of the future (right).

top performing sequences using a functional assay, and recycles the best candidates for subsequent rounds of mutation and selection. A drawback of the approach lies in the random nature of the mutational search that produces a large fraction of non-functional mutants and tends to restrict the mutational search to the local vicinity of the parent. In recent years, there has been an explosion of interest in machine-learning-guided directed evolution (MLDE)[3] in an effort to ameliorate these deficiencies. These approaches typically employ deep neural network models to learn 'nature's blueprint' for protein design from natural sequence databases, and then sample from this distribution to design synthetic proteins with desired functional characteristics. The success of such techniques seems to lie in their capacity to efficiently learn complex, many-body and non-intuitive design rules, and to generatively design novel sequences adhering to these learned rules. The most promising approaches tend to be founded on integrated synthesis of biophysics, statistical mechanics and artificial intelligence, and chemical and biological engineers have played a leading role in the development of this new field.

One particularly compelling recent development in MLDE is the integration of self-supervised protein language models with text-guided conditioning. In a sense, this can be viewed as the integration of the learned syntax and grammar of proteins with that of human language, wherein the latter is used to program desired structural and functional characteristics into sequences designed by the former. Three recent examples illustrate the potential of this approach. The ProGen model of Madani et al.[4] employs a conditional transformer decoder that accepts a keyword or taxonomic 'control tag' to produce sequences within particular classes along with, optionally, the first few amino acids of the sequence (Fig. 1). The trained model was used to generate 1 million synthetic lysozyme sequences, of which 90 were selected for experimental synthesis, 66 of which reported functional activity in digesting the cell wall of *Micrococcus lysodeikticus*. The Chroma model of Ingraham et al.[5] uses a denoising diffusion model to convert random polymer sequences into all-atom protein structures and guides structure generation by conditioning it on previously learned functional relationships between protein structure and function. The model was used to generate novel protein structures using text prompts for fold class (for example, "Ig fold" and "beta barrel") and natural language captions (for example, "Protein with CHAD domain"). The ProteinDT model of Liu et al.[6] uses multimodal learning (that is, learning over multiple modalities or domains) to align the sequence representation of a protein with an attendant natural language annotation. User-inputted text descriptions can then be mapped to corresponding protein sequences possessing desired characteristics. In a campaign to elevate the stability of the villin HP35 subdomain, the trained model was supplied with the prompts "This protein has higher stability" and "This protein has fewer intrinsically disordered regions". These three examples are emblematic of the tantalizing potential of natural-language-directed protein design within MLDE campaigns and represent an exciting new frontier in chemical and biological engineering.

Machine-learning-enabled force fields

A central methodological choice in any molecular simulation study is the model used to describe the interatomic interactions. Progress in intermolecular potential development has been closely intertwined with the emergence of molecular simulations in chemical engineering research. Typically, the level of chemical realism included in a particular model is strongly coupled with its computational expense. The computational investigation of the molecular basis underlying problems of relevance in chemical engineering, including phase separation and transport phenomena (for example, separations processes), materials structure–property relationships (for example, ion transport in polymer electrolytes) and biomolecular processes (for example, protein aggregation), requires simulations sampling time and length scales that are only attainable with semiempirical classical atomistic models or with larger-scale coarse-grained models. By contrast, ab initio approaches, which provide additional chemical realism necessary to model reactivity, polarizability and many-body effects, were previously out of reach for studies such as those alluded to above, owing to their considerable computational expense.

A new class of intermolecular potential models built on data-driven methods may mitigate the need for such a sharp trade off. In this recent approach, the interatomic potential energy and force calculation is performed by a surrogate ML model trained to reproduce the results of an expensive reference technique (for example, density functional theory) for a subset of configurations, enabling ab initio-level predictions over length and time scales orders of magnitude larger than the previous state-of-the-art[7]. This advance is thus endowing ab initio simulations with direct relevance to chemical engineering problems. While exciting, these approaches are still in

their infancy, and important aspects of their development and use remain active areas of research.

We illustrate the power of ML-derived potentials with a topic that has long been of interest to us: the metastable phase behavior of water. Water is a widely studied material, which, from our perspective, translates into dozens of available water models of varying complexity. Crucially, there exists abundant experimental data, thus we have a strong understanding of which water models succeed or fail in predicting a particular property. However, one area in which experimental data are comparatively sparse is the metastable phase behavior of supercooled water. Simulations and experiments have suggested that water may undergo a metastable liquid–liquid transition under deeply supercooled conditions, separating into high- and low-density liquids[8]. Most previous simulation evidence for the liquid–liquid transition was obtained using semiempirical models parameterized to match available experiments, thus their use for deeply supercooled liquid water represents somewhat of an extrapolation. Furthermore, some empirical water models predict the existence of the liquid–liquid transition, while others do not. Recently, two of us have applied ML potentials to provide purely predictive (that is, non-empirical) evidence for the liquid–liquid transition in water, strengthening the available computational evidence in support of this phenomenon[9]. Owing to the long relaxation times of the supercooled liquid and the high degree of statistical certainty needed, ML potentials were a vital enabler of this work, facilitating the use of ab initio techniques in this difficult-to-study regime.

We stress, however, that ML potentials must be used with care, and further development and dissemination of best practices are needed to solidify their widespread use for molecular modeling. In particular, assembling the model training data is non-trivial and requires both intuition and persistence. In our experience, the addition of new training configurations to an existing dataset can have unexpected impacts on the accuracy and stability of the model — more data is not always better. One must also consider the benefits and/or drawbacks of generating general purpose ML potential models versus targeted models specifically tailored to the phenomenon of interest. Active learning and on-the-fly learning approaches can partially automate the model building process, but current algorithms still require non-trivial user input. Further, most systems will not have the abundance of benchmarking data that are available for water, so users must think carefully about how to evaluate the quality of their ML potential. Given the large number of competing ML algorithms, identifying the best ML method for a particular task is a challenge. Efforts are underway to identify the leading approaches and to develop best practices for model development and benchmarking[10], but further work is needed towards open data sharing and knowledge dissemination for molecular dynamics applications. While the potential power of ML-enabled force fields is easily appreciated, these algorithms are still relatively expensive (10–100 times slower than classical atomistic models, even on state-of-the-art hardware[11]); further advances in computational efficiency and user-friendly implementation[12] will help push these methods to the forefront of molecular simulations. If such strides are made, there is great potential for ab initio-level simulations to play a central role in the future of chemical engineering research and industrial practice (Fig. 2).

Outlook

The examples that we have chosen to highlight in this Comment reflect our own research interests. Looking more broadly, we have identified three frontier areas presenting especially promising opportunities for enriching the research and practical dimensions of chemical engineering. They are ML-guided molecular discovery and optimization[13], ML-aided computational catalysis and reaction engineering[14], and the integration, via ML, of ab initio-level molecular information into plant-level process control[15] (Fig. 2).

The promising future of ML in chemical engineering notwithstanding, considerable challenges remain to be addressed. The fruitful application of ML requires that researchers and practitioners be well educated in the relevant fundamentals and praxis of these techniques and the integration of these tools with domain expertise. Incorporating statistics, data science and artificial intelligence into the chemical engineering undergraduate curriculum is essential. The challenge is that chemical engineering is already at the high end of required credit hours, compared with other engineering disciplines. Thus, the academic chemical engineering community will eventually have to confront difficult choices to provide the tools necessary to fully unlock ML's potential in our discipline. At the very least, this will require retrofitting existing courses through appropriate examples and assignments, but it is not clear to us that this is sufficient.

The availability of adequate computational resources is often a limiting factor in ML research. Generally, this means having access to local, small clusters of graphics processing units (GPUs) for development and testing, and to large, shared GPU clusters, either on premises or in the cloud, for training and deployment of large models. Cost, space and cooling requirements are limiting factors that are already testing institutional capabilities, and may well require multi-institutional consortia, creative new partnerships with industry, or massive new government investments at the federal and state levels.

We believe that ML has a bright future in molecular simulation and design, and more broadly in chemical engineering, both as an increasingly indispensable component of fundamental research and as an enabling tool in industrial practice. A sampling of ongoing work referenced in this Comment illustrates the extent to which ML is already enabling the generation and application of new knowledge, in ways that would have been unthinkable a few years ago.

Thomas E. Gartner III[1], Andrew L. Ferguson[2] & Pablo G. Debenedetti[3]

[1]Department of Chemical and Biomolecular Engineering, Lehigh University, Bethlehem, PA, USA. [2]Prtizker School of Molecular Engineering, University of Chicago, Chicago, IL, USA. [3]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA.

e-mail: pdebene@princeton.edu

Published online: 11 January 2024

References

1. *New Directions for Chemical Engineering: The National Academies Consensus Study Report* (The National Academies Press, 2022).
2. Arnold, F. H. *Angew. Chem. Int. Ed.* **57**, 4143–4148 (2018).
3. Ferguson, A. L. & Ranganathan, R. *ACS Macro Lett.* **10**, 327–340 (2021).
4. Madani, A. et al. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
5. Ingraham, J. et al. *Nature* **623**, 1070–1078 (2023).
6. Liu, S. et al. Preprint at https://doi.org/10.48550/arxiv.2302.04611 (2023).
7. Behler, J. *J. Chem. Phys.* **145**, 170901 (2016).
8. Palmer, J. C., Poole, P. H., Sciortino, F. & Debenedetti, P. G. *Chem. Rev.* **118**, 9129–9151 (2018).
9. Gartner, T. E. III, Piaggi, P. M., Car, R., Panagiotopoulos, A. Z. & Debenedetti, P. G. *Phys. Rev. Lett.* **129**, 255702 (2022).
10. Fu, X. et al. Preprint at https://doi.org/10.48550/arXiv.2210.07237 (2023).
11. Lu, D. et al. *Comput. Phys. Commun.* **259**, 107624 (2021).
12. Zeng, J. et al. *J. Chem. Phys.* **159**, 054801 (2023).

13. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. *Nature* **559**, 547–555 (2018).
14. Tran, R. et al. *ACS Catal.* **13**, 3066–3084 (2023).
15. Adjiman, C. S., Galindo, A. & Jackson, G. *Comput. Aided Chem. Eng.* **34**, 55–64 (2014).

Competing interests

A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Applications 16/887,710 and 17/642,582, US Provisional Patent Applications 62/853,919, 62/900,420, 63/314,898, 63/479,378, and 63/521,617, and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466. T.E.G. and P.G.D. declare no competing interests.

Additional information

Peer review information *Nature Chemical Engineering* thanks the anonymous reviewers for their contribution to the peer review of this work.