Reconciling deep learning and firstprinciple modelling for the investigation of transport phenomena in chemical engineering

Reconciling deep learning and first-principle modelling for the investigation of transport phenomena in chemical engineering

Agnese Marcato[1], Daniele Marchisio*[1], Gianluca Boccardo[1]

[1]Department of Applied Science and Technology, Institute of Chemical Engineering, Politecnico di Torino, Torino, Italy

Correspondence

*Daniele Marchisio Email: daniele.marchisio@polito.it

Summary

The use of machine learning in chemical engineering has the potential to greatly improve the design and analysis of complex systems. However, there are also risks associated with its adoption, such as the potential for bias in algorithms and the need for careful oversight to ensure the safety and reliability of machine learning-powered systems. This paper explores the opportunities and risks of using machine learning in chemical engineering, and provides a perspective on how it may be integrated into engineering practices in a responsible and effective manner.

We generated the text of this abstract with GPT-3, OpenAI's large-scale language-generation model. Upon generating the draft, we ensured that the language was to our liking and we take ultimate responsibility for the content of this publication.

Keywords: artificial intelligence; machine learning; deep learning; computational modelling; CFD; MD

1 Introduction

Science and engineering, being both human activities, are not immune to fashion trends. Indeed, the use of machine learning, and artificial intelligence in general, is part of this trend. It is now extremely fashionable to work in this area and the pressure on researchers to include some of these elements in their research proposals, daily research activities, and papers is tremendous. Different reactions are observed within the community, ranging between two extremes which we have labelled as: (1) the conformist and (2) the obstinate opposer. The conformist will adopt machine learning in their work, even if not needed, necessary, or useful, just because it is fashionable, and the conformist likes to be fashionable. They want to do exactly what everybody else is doing. The obstinate opposer will NOT adopt machine learning in their work, even if needed, necessary, or useful, just because it is fashionable, and the obstinate opposer hates to be fashionable. They have more important things to do than simply what the others are doing.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/cjce.24838

And then, of course, all the shades of grey in between.

It is evident that working in this area is now challenging for many reasons, among which is a certain bias which, depending on the individuals involved, ranges from naïve enthusiasm for anything that has something to do with machine learning to nonconstructive criticism for the very same reasons. We want to contribute to the very polarized ongoing discussion on machine learning in chemical engineering by sharing with our scientific community our perspective on this controversial topic. We want to do it not because we think we have something to teach, but because we feel it could be useful to share our path, to explain why we are using machine learning in our daily job, and to confide what we think would be exciting to do in the near future.

2 Computational modelling

Computational modelling is huge in chemical engineering. In the investigation of (molecular) transport phenomena, a plethora of methods and tools are involved ranging from ab initio quantum chemistry methods, classical full atom and coarse-grained molecular dynamics methods, mesoscopic methods, and continuum models, among which are computational fluid dynamics, chemical reaction engineering models, and process models. Each of these tools is characterized by a certain degree of empiricism, which increases when moving from ab initio calculations to process modelling. The computational models we deal with now are way more complex than what we used to have in the past, but we can still distinguish between first-principle models and data-driven models, to which we have recently added hybrid models. In first-principle models, when we observe disagreement with experimental data, we improve the theory. In data-driven models, when we observe disagreement, we improve the "fitting", or maybe we should say we improve the regression. Indeed, also in first-principle models, there are unknown parameters that are identified through fitting, but still the difference persists.

In the ongoing discussion, we think it is also useful to reflect on why we use models. One reason might be because by building the model we gain deeper understanding. At the end we might not even use the model anymore, because we now have understood everything and we can throw the model away. In this first case, usually first-principle models are employed. Another reason might be because we want quantitative reliable predictions for designing a process, for optimizing it, or for scaling it up and down. In this context, we do not care about the deeper understanding; we just care about the predictions. In this second case, a data-driven model can indeed be employed. In many cases, one does computational modelling for both reasons.

3 Why machine learning in computational modelling?

One problem with machine learning is that you often need "big data". In many contexts, "big data" is available; for example, for social media and search engines, which collect individual

data from our devices. In chemical engineering, this “big data” is not always there, except if the data is produced already within a computer, as is the case with computer simulations obtained from computational models. This is very easy to realize now, thanks to high performance computing, software orchestrators, and high throughput workflows, which allow the production of a huge amount of simulation data in a short time and at little cost.

High performance computing represents a good reason why machine learning is now so popular. Especially with graphical processing unit architectures, a large amount of data can be collected, stored, and employed to train data-driven models. Another reason for the recent increase in popularity of machine learning is that we have now libraries that are easy to use and integrated into complex workflows. Even though the fundamentals of the deep learning theory were introduced before the 90s, only with the advent of the CUDA platform in 2007 has the training of those algorithms become computationally feasible for the practitioners of machine learning (i.e., data and computer scientists). Then, when in 2015 and 2016 respectively Tensorflow and Pytorch were released as open-source Python libraries by Google and Facebook, deep learning tools became concretely available to a wider community, including chemical engineers, by greatly reducing their barrier to entry.

The availability of these libraries together with an outstanding open community online and, above all, training on the theory of deep learning has given to traditional modelling researchers, who are familiar with statistics and coding, the opportunity to move quickly through the learning curve and take an active role in machine learning research. Open-source physics-based toolboxes (like OpenFOAM for computational fluid dynamics, LAMMPS for molecular dynamics, and the different codes for discrete element methods available) made chemical engineers developers, rather than simple users. Nowadays in machine learning their contribution can go beyond being the creators of the dataset or passive end-users of the simplest codes.

The objective of the deep learning model has to be clear and bounded; in fact, the dataset creation and repartition, the training strategy, and the model architecture must be set in order to target the level of generalization required. Researchers should be aware of the level of generalization achieved by their model, which constitutes the basis of a study about reliability of the model, when it is employed in end-applications such as optimization or multiscale modelling. It has been shown for some type of machine learning tools, such as neural networks, that tailoring the model in terms of architecture, loss function, and input features to the prediction objective can improve their generalization efficacy and reduce significantly the amount of data required during training. This activity can be performed knowing in depth the physics of the problem, and a computer scientist cannot easily do that. We think that the research process cannot be cleanly split into chemical engineering competence and data science competence.

4 Some good reasons why you should give it a try

We list a few examples in which we think machine learning can have an impact on chemical engineering. These are also exciting areas to be working in the next decades.

- **Molecular dynamics force fields.** Ab initio quantum chemistry methods are extremely expensive, but very accurate. Moreover, in their different formulations, the codes implementing them do not scale as efficiently as classical molecular dynamics codes. It does, therefore, make sense to run an incredibly large number of small ab initio quantum chemistry calculations to train an artificial neural network and generate interatomic potentials. This allows the investigation of (very) large systems, for (very) long times, with classical molecular dynamics simulations, characterized however by quantum accuracy. These calculations would be inaccessible by ab initio methods. The total computational costs of producing the dataset for training and of the training itself are completely paid off by the very efficient scaling of classical molecular dynamics codes.[1, 2]
- **Multiphase CFD model closures.** Similarly to what was just discussed, in the simulation of polydisperse multiphase systems with computational fluid dynamics codes, direct numerical simulations methods, with fully-resolved interfaces, are very accurate, but extremely computationally expensive. On the contrary, modelling approaches in which the interface is not resolved, either by using point-particles and Lagrangian tracking or Eulerian multi-fluid models, are extremely cheap and can be employed to investigate (very) large systems for (very) long times. Also in this case, direct numerical simulations can be used to formulate closures for less detailed simulations. One typical closure is for the drag force, for dense polydisperse multiphase systems, for which elements of the disperse phase (i.e., particles, droplets, or bubbles) interact indirectly with each other via three-way coupling. In a context in which it turns out to be very challenging to come up with an accurate analytical closure (i.e., for the drag force), even with available data from direct numerical simulations, the use of machine learning, to do the fitting, can be extremely promising.[3, 4]
- **Multiscale modelling.** Most of the systems of interest in chemical engineering have an intrinsic multiscale nature. It is necessary to rely on constitutive equations to consider lower-scale phenomena. Constitutive equations are simplifications of the real microscale behaviour, and they fail outside of the range of their validity hypotheses. Detailed microscale simulations could be called by macroscale simulations in order to consider the actual complexity among scales; however, this approach can be unfeasible when physics-based models are computationally expensive. In this case, machine learning models can be trained on a dataset of microscale simulations to convey the information to a higher scale. This approach has been proposed for complex rheologies [5] and porous media problems.[6–8] In microscale simulations of porous media, the actual geometry of the porous structure is the boundary for the solution of the transport equations. Conversely, in macroscale models, geometrical parameters and source terms appear in the transport equations.

Those terms can be evaluated by constitutive equations, but when geometries become complex,[9] they fail, so microscale simulations are necessary for their evaluation. Since microscale simulations cannot be integrated directly into macroscale solvers, they can be employed to train a surrogate model, which becomes a feasible bridge between scales.

- **Identifying physics-based model parameters.** In many cases in continuum models, the transport equations include terms not directly representing the physical reality of the phenomenon at the molecular scale, but an appropriately calibrated model: for example, in the case of population balance equations,[10] this happens for the parameters of the aggregation or breakage kernels. Some of the parameters are often only obtainable by repeated runs of CFD simulations and an iterative comparison with available experimental data. In this case, if one is able to train a reasonably accurate data-driven model which is able to give a fast response, it is possible to conduct this parameter optimization much faster, and create a finely-tuned CFD model which is coherent with the experimental ground truth.
- **Process optimization and control.** The idea of "optimization by means of a fast and accurate model" is not exhausted in the application just mentioned of *optimizing unknown parameters* so to obtain a physically coherent model. When an accurate data-driven model of the process is available (such as a neural network surrogating the full simulation results,[11]) it is possible to use it to *optimize the design* of a process,[12, 13] be it its geometrical definition or the operating conditions at which the process performance is maximum. Moreover, the fast response model that has been used upstream in the design phase for the process can also be used downstream when the process is in place: the instantaneous predictions opening the way for large parameter-space optimization during the design of the process will also prove invaluable for its in-line control, where an advanced model predictive control system can then be developed to smoothly and accurately keep a process inside the desired parameters of operation.
- **Simulation on the loop: Building a data-driven model with sparse experimental data.** One of the key features of machine learning is the necessity of a dataset with which to train the model. Experimental points are usually limited to small campaigns of experiments and the operating conditions are usually explored in a rigid structured way. In this case, the dataset can be enriched by physics-based simulations, validated on the available experimental points. This validated computational tool allows for a flexible operating conditions exploration and the possibility to obtain a dataset wide enough for the generalization purposes of the machine learning model. Finally, the obtained data-driven model can be easily called to address new experimental campaigns.

- **Artificial intelligence in the laboratory.** Machine learning models can find employment in laboratory routines when there is a lack of fast models for decision-making. In this case, computationally expensive simulations can hardly be used, so fast response models can find application. The first works in this perspective are found in the energy storage field.[14] In the production process of electrodes for batteries, researchers and technicians must follow many steps, each of them characterized by different operating conditions to be set. Knowing in advance the quantitative effect of operating conditions on a final desired property can decrease the trial-and-error phase and the cost of the experiments.[15]

5 Putting it all together: How to (carefully) move forward

Given all of these examples, we cannot but think that the place where machine learning and chemical engineering modelling meet constitutes fertile ground for new and exciting research. This is true both for industrial research but also, increasingly, in the field of research for informing public policy about the management of collective spaces and resources: examples of the latter range from sustainable management of natural resources (dealing with groundwater availability, irrigation policies, forest management, etc.), [16, 17] resulting in policies for infrastructure funding allocation, to models dealing with predictions of the wider effects of global climate change.[18]

That is not to say, however, that this new playground comes without pitfalls and risks—especially for the very fast uptake that this kind of modelling is seeing both in the private and the public spheres. Science suffers, and has suffered for quite some time, from what has been called a *reproducibility crisis*. This crisis is due to many factors, two chief among these. First, a proliferation of the published scientific production which led to, and is caused by, an increased focus on producing research with new positive results with correspondingly less emphasis on reproducing already published results. Secondly, a researcher wanting to reproduce results is sometimes faced with an actual inability to do that: this is most often not due to these results being wrong or incorrectly presented, but instead it is in many cases due to the inaccessibility of the tools used to obtain these results. One example of this in computational modelling is the use of commercial or closed-source software, which is being partly counteracted by the now widespread effort in using open-source alternatives, together with the push for FAIR data practices (Findable, Accessible, Interoperable, Reusable).

In this context, there is the possibility of data-driven modelling adding another (and, in our opinion, more subtle) layer of risk. While first-principles models, as implemented in current computer simulations, are able to trace a clear and explainable line from the solved equations to the results they produce, machine learning models are instead able to function as black boxes in which the relationships between the user inputs and the resulting outputs are less clear. Beyond the problem of interpretability and explainability of these (seemingly, as we will see) black box models, there is the even greater risk of **bias** in artificial intelligence algorithms.

Specifically, while wrong scientific models have always been dangerous, their ailment was almost always *technical* in nature, to be cured with technical tools (e.g.: careful statistical analysis, better models), whereas the additional layer of complexity in the construction of machine learning models opens the way for mistakes in the *human conceptualization* of the investigated problem, coming especially with the way the starting data is collected—scientific data that by its very nature gives the appearance of being, and which should be, objective.

Since the new resurgence of the use of AI in the 2010s,[19] problems related to bias have been at the forefront of both the public discussion and the technical discussion around AI.[20–22] Misplaced analyses or an incorrect model construction resulting in a final model whose skewed results touch the dimensions of gender, language, race, and so on can have hugely damaging societal impacts in crystallizing or reinforcing existing biases. These failures have (most often) been promptly identified and pointed out, resulting in quick corrections. This is due both to the (highly justified) great level of global scrutiny around these topics, and also because biased models resulting in skewed, politically biased, or otherwise socially unpalatable results *fail in an apparent manner*. Or, to put it in another way, they fail in a way for which it is simpler for the layman to recognize they are looking at something wrong, making these kind of biases very serious but also somewhat easier to recognize. Herein lies the issue: we are, as a society and as individuals, more equipped to recognize when an algorithm violates our intuitive vision of equality, than when it would give us an equally damaging answer in the form of a very technical, but very impactful, public steering policy.

That being said, it is our opinion that this is not reason enough to not engage with this exceedingly fascinating (and still evolving) technology: the use of machine learning in the physical sciences comes with a wealth of new possibilities, and as with all disruptive changes, they have to be managed with great responsibility, of which we should be conscious both in our work and in the evaluation of the work of others. Several steps in this direction have already been made, dealing for example with the problem of interpretability, revealing modern machine learning models to be not-so-black boxes after all: classical machine learning models have been explored,[23] as well as the results from convolutional neural networks,[24, 25] also building on frameworks developed in other disciplines, such as game theory.[26, 27] Then, careful data provenance has to be of course an ongoing effort on part of each research team, which has to be able to answer to the question of why their data looks the way it does, why it was collected that way, and what its limitations are.

Chemical and materials engineers, if willing to contaminate their expertise with experience in computer science and in the art of artificial intelligence thinking, may surely be (and maybe should be) the stewards of an aware and responsible transition to an ever more *digital* practice and understanding of the scientific process. This will be no different than what we did as a community when scientists first delving in numerical simulation had to arm themselves with a deep knowledge of machine numerical analysis in addition to transport phenomena: the field of numerical verification flourished,[28, 29] and numerical modelling is now an invaluable tool at our disposal, notwithstanding its pitfalls and the risks in a careless use of its results.

We should approach deep learning and artificial intelligence much in the same way: with openness to new ideas, conscious care, and great enthusiasm.

References

[1] Kristof T. Schütt, Huziel E. Sauceda, P.J. Kindermans, Alexandre Tkatchenko, K. R. Müller, *The Journal of Chemical Physics* **2018**, *148* (24), 241722.

[2] Stefan Chmiela, Huziel E Sauceda, Klaus-Robert Müller, Alexandre Tkatchenko, *Nature communications* **2018**, *9* (1), 1–10.

[3] Bhargav Siddani, S Balachandar, William C Moore, Yunchao Yang, Ruogu Fang, *Theoretical and Computational Fluid Dynamics* **2021**, *35* (6), 807–830.

[4] S. Balachandar, W.C. Moore, G. Akiki, K. Liu, *Theoretical and Computational Fluid Dynamics* **2020**, *34* (4), 401–428.

[5] Lifei Zhao, Zhen Li, Bruce Caswell, Jie Ouyang, George Em Karniadakis, *Journal of Computational Physics* **2018**, *363*, 116–127.

[6] Agnese Marcato, Gianluca Boccardo, Daniele Marchisio, *Industrial & Engineering Chemistry Research* **2022**, *61* (24), 8530–8541.

[7] Allen K. Ting, Javier E. Santos, Eric Guiltinan, *Energies* **2022**, *15* (23), 8871.

[8] Agnese Marcato, Gianluca Boccardo, Daniele Marchisio, *Chemical Engineering Journal* **2021**, *417*, 128936.

[9] Enrico Agostini, Gianluca Boccardo, Daniele Marchisio, *Chemical Engineering Science* **2022**, *255*, 117583.

[10] Graziano Frungieri, Gianluca Boccardo, Antonio Buffo, Hossein Ali Karimi-Varzaneh, Marco Vanni, *Chemical Engineering Science* **2022**, *260*, 117859.

[11] Agnese Marcato, Javier E Santos, Gianluca Boccardo, Hari Viswanathan, Daniele Marchisio, Maša Prodanović, *Chemical Engineering Journal* **2022**.

[12] Donggeun Park, Jemyung Cha, Moonjeong Kim, Jeung Sang Go, *Engineering Applications of Computational Fluid Mechanics* **2020**, *14* (1), 180–201.

[13] Marc Duquesnoy, Chaoyue Liu, Diana Zapata Dominguez, Vishank Kumar, Elixabete Ayerbe, Alejandro A Franco, *Energy Storage Materials* **2023**, *56*, 50–61.

[14] Teo Lombardo, Marc Duquesnoy, Hassna El-Bouysidy, Fabian Årén, Alfonso Gallo-Bueno, Peter Bjørn Jørgensen, Arghya Bhowmik, Arnaud Demortière, Elixabete Ayerbe, Francisco Alcaide, Marine Reynaud, Javier Carrasco, Alexis Grimaud, Chao Zhang, Tejs Vegge, Patrik Johansson, Alejandro A. Franco, *Chemical Reviews* **2021**, *122* (12), 10899–10969.

[15] Ricardo Pinto Cunha, Teo Lombardo, Emiliano N Primo, Alejandro A Franco, *Batteries & Supercaps* **2020**, *3* (1), 60–67.

[16] S Sahoo, TA Russo, J Elliott, I Foster, *Water Resources Research* **2017**, *53* (5), 3878–3895.

[17] Sina Ardabili, Amir Mosavi, Majid Dehghani, Annamária R Várkonyi-Kóczy, in *International conference on global research and education*, Springer, **2020**, pp. 52–62.

[18] Chris Huntingford, Elizabeth S Jeffers, Michael B Bonsall, Hannah M Christensen, Thomas Lees, Hui Yang, *Environmental Research Letters* **2019**, *14* (12), 124007.

[19] Gary Marcus, Ernest Davis, *Rebooting AI: Building artificial intelligence we can trust*, Vintage, **2019**.

[20] The Guardian, *Microsoft 'deeply sorry' for racist and sexist tweets by AI chatbot*, **2016**, accessed 2023/1/9. https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot.

[21] CNET, *Meta Trained an AI on 48M Science Papers. It Was Shut Down After 2 Days*, **2022**, accessed 2023/1/9. https://www.cnet.com/science/meta-trained-an-ai-on-48-million-science-papers-it-was-shut-down-after-two-days/.

[22] NPR, *Facebook Apologizes After Its AI Labels Black Men As 'Primates'*, **2021**, accessed 2023/1/9. https://www.npr.org/2021/09/04/1034368231/facebook-apologizes-ai-labels-black-men-primates-racial-bias.

[23] Osbert Bastani, Carolyn Kim, Hamsa Bastani, *arXiv preprint* **2017**, https://arxiv.org/abs/1705.08504, accessed 2023/1/9.

[24] Shawn Xu, Subhashini Venugopalan, Mukund Sundararajan, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, **2020**, pp. 9680–9689.

[25] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, Tolga Bolukbasi, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, **2021**, pp. 5050–5058.

[26] Scott M Lundberg, Su-In Lee, in *Advances in Neural Information Processing Systems*, (Eds: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett), Curran Associates, Inc., **2017**.

[27] Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin, *arXiv preprint* **2016**, http://arxiv.org/abs/1602.04938, accessed 2023/1/9.

[28] Patrick J Roache, *Verification and validation in computational science and engineering*, *Vol. 895*, Hermosa Albuquerque, NM, **1998**.

[29] William L Oberkampf, Christopher J Roy, *Verification and validation in scientific computing*, Cambridge University Press, **2010**.