Advancing data-driven chemistry by beating benchmarks

Helge S. Stein [1,2,*]

Enabled by data management and digitalization adoption in chemistry, machine learning (ML) is accelerating chemistry through automated data analysis, materials embeddings, property prediction, and active learning. Beyond existing demonstrations of ML in chemistry, there is a critical need for chemically driven benchmarks to make ML models fail in a constructive manner.

Through the proliferation of data management and digitalization, data science methods are attracting serious attention in chemistry and materials research. Whilst early demonstrations in the field showed the general applicability of data science in the field of chemistry, benchmarking [1–3] is becoming a key driving force towards autonomous materials discovery and upscaling [4].

Deployment of ML models has been very successful in performance-driven fields such as electrocatalysis, photovoltaics, and batteries, where demonstrations mostly revolved about finding models capable of predicting materials properties upon which materials embeddings, automatic data analysis, active learning, and explainable ML in chemistry emerged.

Data-driven chemistry requires ways to 'teach' a model ' what' a material is, to model the underlying physicochemical relationships. Arguably the simplest way to express a material is by describing/counting its constituents or structural fragments. This is necessary as data-driven models typically require some tensor-like input, which is mapped onto an output(s). Most data science practitioners, including myself, will likely have encountered that generating a giant one-hot representation, or 'embedding' [5], in which the composition is represented by a long vector that entails some statistics about what elements are in a material, will lead to somewhat performant models, but the sparsity (and curse of dimensionality) of the data inevitably lead to overall low performing models. Especially when extrapolation towards higher performance or new composition spaces, or high interpretability, is needed.

User-friendly implementations of ML models for accelerated chemistry [1–3] enable testing virtually any model on any materials science dataset. This allows practitioners in the design or testing phase to make their models break, help understand why models break, to eventually make them 'unbreakable' [6]. Even for a seemingly facile task like regression, one can define simple but hard to beat benchmarks, like k-fold cross forward validation by Xiong and colleagues [7] (i.e., sorting by the to-be-predicted performance to perform a 'sorted' k-fold cross-validation through splitting by percentile). They evaluate the predictive power of ML models, thus separating the interpolation from the extrapolation performance. Many models will exhibit low errors, depending on their complexity, as shown schematically in Figure 1A, but the key metric to seek is the true predictive power in extreme cases, as discussed in Figure 1B. There, the sorted k-fold cross-validation is schematically shown in comparison to a random holdout. What one typically finds is that upon interpolation, most models perform very well; but when tasked with extrapolating towards high or low performance, they tend to perform poorly. In my opinion, this or a similar applicable test should become a standard test for any ML study dealing with property prediction.

Accompanying a thorough assessment of a model's performance on extrapolation and testing its limits should be an unbiased assessment whether the model, data, or data representation is at fault. Naturally, most data science practitioners consider the models to be the faulty link in the data science pipeline, leading to great advancements in adapting the underlying regressors and classifiers for chemistry and physics by incorporating physically meaningful constraints.

An extreme example of needing to predict well into high-performing or unseen composition or parameter spaces is found in active learning [3]. Here, a model is trained on a few data points and is then tasked with suggesting the next optimal experiment that is poised to either perform better or reduce the model's uncertainty. For this task, different benchmarking frameworks and metrics exist, in which researchers can pit different optimizers against each other. Through the availability of benchmarks for active learning, the field is lifted to a new level of productivity where new creative uses of pooling and integration of accelerated instruments yield creative ways of accelerating chemistry. Active learning will penalize those models that cannot operate out of distribution in Figure 1C as any model that is over- or underfit, or has other issues will not accelerate the optimization task.

Even with state-of-the-art regression and classification models, there exist materials and challenges that are poorly described by simple embedding methods counting elements or structural fragments. These include inverse materials design for molecules and solid-state materials as well as explainable ML models for chemistry.

I currently identify three research tracks for embedding materials that can be loosely

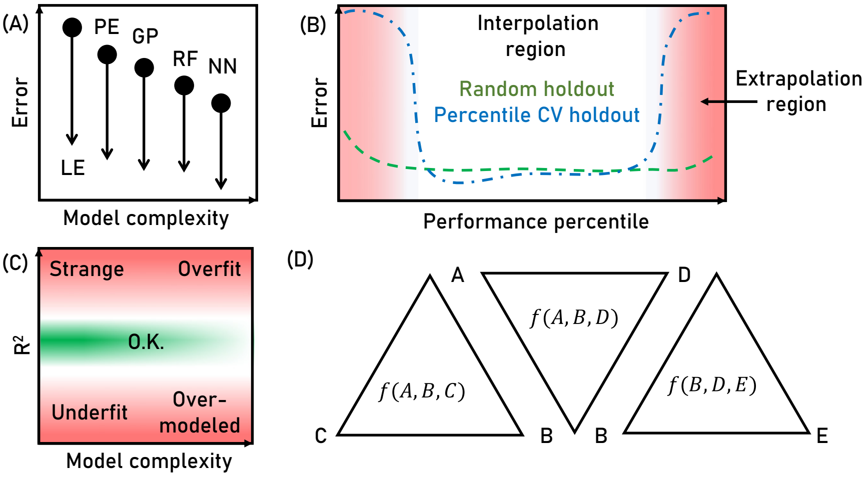

Figure 1. Challenges and prospects in deploying data-driven chemistry. (A) While more complex models [linear (LE) polynomial ensembles (PE), Gaussian processes (GP), random forests (RF), and neural nets (NN)] will allow for generally lower prediction errors, they come at the cost of being harder to interpret and tend to mostly perform subpar in benchmarks testing the extrapolation capabilities. This is shown in (B), where a hypothetical dataset is sorted by the performance and the top and bottom percentile regions are used as test and validation datasets. This loosely corresponds to a decoupling of the interpolation and extrapolation performance in a sorted k-fold cross-validation scheme and allows for a fair model assessment. (C) Different cases of issues when training machine learning models in chemistry are shown, with the general cautionary advice being that any overly well-performing model should be examined with great care. (D) Examples of adjacent ternary composition spaces of the elements A-B-C-D-E to visualize that even though these composition spaces should offer some similarities, the learning of the physicochemical interactions is complex.

categorized into either: (i) compositional/structural-fragment, (ii) partial or full-structural [6], or (iii) mixed-mode compositional-embeddings [8,9]. The motivation behind these is to be able to predict into new and unseen composition spaces or generate new molecules from knowledge, designed for a specific function. In this field, another seemingly simple but hard to beat benchmark was developed by Kong and colleagues [8]. Here, a model is tasked with predicting into an unseen composition space or into new composition spaces, including previously unseen elements. This is exemplarily shown in Figure 1D. Consider a materials property f to be predicted in a new composition space. When taught with a subset of possible other ternary A-B-C-D-E compositions, or even only with higher order combinations, the prediction into unseen composition spaces would certainly make most models falter [9]. This interesting formulation of a reasonably simple to formulate but hard to beat baseline inspired Kong and colleagues [8] to develop density of states-based elemental embeddings that allow models to perform significantly better than before. Inverse design is, however, not easily possible on compositions, such that other models that yield synthesizable structures are necessary.

To this end, great advancements in organic chemistry were demonstrated by several groups. Gómez-Bombarelli and colleagues [5] demonstrated the first continuous, differentiable, and invertible embeddings of molecular structures. These were then extended towards 3D crystal structures [10], allowing the inverse design of both molecules and extended solid-state crystals. For molecules there is now even the possibility to generate molecular embeddings with perfect invertibility [6], without, however, the possibility of exact 3D reconstruction. Underlying these advancements is again a simple benchmark that measures whether a structure is a valid, hypothetically possible, molecule (or SMILES representation thereof).

Implied with all the utilizations of benchmarking and the resulting advancements is the availability of data. Considering the aforementioned examples, most demonstrations originate from theoretical chemistry. This is likely due to the only recent deployments of large public experimental datasets in materials science and publicly searchable databases. More importantly, data from experiments needs to be analyzed, a task mostly performed manually. With the goal of 'closing the loop' of materials discovery to upscaling, I see a critical need to automate data analysis in the experimental sciences. There are still a great number of unsolved challenges in the preprocessing and analysis of experimental data. Emblematic for the necessity of preprocessing spectra is the statistical learning of background signals [11]. Spectroscopy is, however, only the tip of the iceberg, as common techniques in electrochemistry such as electrochemical impedance spectroscopy, Fourier-transform infrared spectroscopy (FTIR), and X-ray photoelectron spectroscopy (XPS) are still analyzed manually and exhibit similarities with many complex problems in computer science.

Concluding, I see great advancements to come to the field of data-driven materials science and chemistry though a broader adoption of chemically inspired benchmarks that allow a physically meaningful assessment of the quality of data quality, data representation, data visualization, and model quality. There is, however, a critical need to measure and assess the discovered explanations of ML models in the emergent field of explainable artificial intelligence in chemistry, tools to visualize high-dimensional chemical datasets to extract new fundamental knowledge faster, and machine and human interpretable languages [12] to describe experiments.

Acknowledgments

This work contributes to the research performed at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe) and was funded by the German Research Foundation (DFG) under Project ID 390874152 (POLiS Cluster of Excellence). This project received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 957189. This project received funding from BMBF in the framework of the BMBF-Kompetenzcluster InZePro for the projects DataBatt No 03XP0323D and InForm No 03XP0363A.

Declaration of interests

No interests are declared.

[1] Helmholtz Institute Ulm (HIU), Helmholtzstr. 11, 89081 Ulm, Germany

[2] Karlsruhe Institute of Technology (KIT), Institute of Physical Chemistry (IPC), Fritz-HaberWeg 2, 76131 Karlsruhe, Germany

*Correspondence:
helge.stein@kit.edu (H.S. Stein).

https://doi.org/10.1016/j.trechm.2022.05.003

References

1. Dunn, A. *et al.* (2020) Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *NPJ Comput. Mater.* 6, 138
2. Häse, F. *et al.* (2021) Olympus: a benchmarking framework for noisy optimization and experiment planning. *Mach. Learn.: Sci. Technol.* 2, 035021
3. Rohr, B. *et al.* (2020) Benchmarking the acceleration of materials discovery by sequential learning. *Chem. Sci.* 11, 2696–2706
4. Stein, H.S. and Gregoire, J.M. (2019) Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* 10, 9640–9649
5. Gómez-Bombarelli, R. *et al.* (2018) Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* 4, 268–276
6. Krenn, M. *et al.* (2022) SELFIES and the future of molecular string representations. *arXiv* Published online March 31, 2022. https://doi.org/10.48550/arXiv2204.00056
7. Xiong, Z. *et al.* (2020) Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput. Mater. Sci.* 171, 109203
8. Kong, S. *et al.* Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* 13, 949
9. Kong, S. *et al.* (2021) Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* 8, 021409
10. Noh, J. *et al.* (2019) Inverse design of solid-state materials via a continuous representation. *Matter* 1, 1370–1384
11. Ament, S.E. *et al.* (2019) Multi-component background learning automates signal detection for spectroscopic data. *NPJ Comput. Mater.* 5, 1–7
12. Steiner, S. *et al.* (2019) Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* 363, eaav2211