Best practices in machine learning for chemistry

Statistical tools based on machine learning are becoming integrated into chemistry research workflows. We discuss the elements necessary to train reliable, repeatable and reproducible models, and recommend a set of guidelines for machine learning reports.

Nongnuch Artrith, Keith T. Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain and Aron Walsh

Chemistry has long-benefited from the use of models to interpret patterns in data, from the Eyring equation in chemical kinetics, the scales of electronegativity to describe chemical stability and reactivity, to the ligand-field approaches that connect molecular structure and spectroscopy. Such models are typically in the form of reproducible closed-form equations and remain relevant over the course of decades. However, the rules of chemistry are often limited to specific classes of systems (for example, electron counting for polyhedral boranes) and conditions (for example, thermodynamic equilibrium or a steady state).

Beyond the limits where simple analytical expressions are applicable or sophisticated numerical models can be computed, statistical modelling and analysis are becoming valuable research tools in chemistry. These present an opportunity to discover new or more generalized relationships that have previously escaped human intuition. Yet, practitioners of these techniques must follow careful protocols to achieve levels of validity, reproducibility, and longevity similar to those of established methods.

The purpose of this Comment is to suggest a standard of 'best practices' to ensure that the models developed through statistical learning are robust and observed effects are reproducible. We hope that the associated checklist (Fig. 1 and Supplementary Data 1) will be useful to authors, referees, and readers to guide the critical evaluation of, and provide a degree of standardization to, the training and reporting of machine learning models. We propose that publishers can create submission guidelines and reproducibility policy for machine-learning manuscripts assisted by the provided checklist. We hope that many scientists will spearhead this campaign and voluntarily provide a machine learning checklist to support their papers.

The growth of machine learning and making it FAIR

The application of statistical machine learning techniques in chemistry has a long history[1]. Algorithmic innovation, improved data availability, and increases in computer power have led to an unprecedented growth in the field[2,3]. Extending the previous generation of high-throughput methods, and building on the many extensive and curated databases available, the ability to map between the chemical structure of molecules and materials and their physical properties has been widely demonstrated using supervised learning for both regression (for example, reaction rate) and classification (for example, reaction outcome) problems. Notably, molecular modelling has benefited from interatomic potentials based on Gaussian processes[4] and artificial neural networks[5] that can reproduce structural transformations at a fraction of the cost required by standard first-principles simulation techniques. The research literature itself has become a valuable resource for mining latent knowledge using natural language processing, as recently applied to extract synthesis recipes for inorganic crystals[6]. Beyond data-mining, the efficient exploration of chemical hyperspace, including the solution of inverse-design problems, is becoming tractable through the application of autoencoders and generative models[7]. Unfortunately, the lack of transparency surrounding data-driven methods has led some scientists to question the validity of results and argue that the field faces a "reproducibility crisis"[8].

The transition to an open-science ecosystem that includes reproducible workflows and the publication of supporting data in machine-readable formats is ongoing within chemistry[9]. In computational chemistry, reproducibility and the implementation of mainstream methods, such as density functional theory, have been investigated[10]. This, and other studies[11], proposed open standards that are complemented by the availability of online databases. The same must be done for data-driven methods. Machine learning for chemistry represents a developing area where data is a vital commodity, but protocols and standards have not been firmly established. As with any scientific report, it is essential that sufficient information and data is made available for a machine learning study to be critically assessed and repeatable. As a community, we must work together to significantly improve the efficiency, effectiveness, and reproducibility of machine learning models and datasets by adhering to the FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data management and stewardship[12].

Below, we outline a set of guidelines to consider when building and applying machine learning models. These should assist in the development of robust models, providing clarity for manuscripts, and building the credibility needed for statistical tools to gain widespread acceptance and utility in chemistry.

Guidelines when using machine learning models

1. Data sources. The quality, quantity and diversity of available data impose an upper limit on the accuracy and generality of any derived model. The use of static datasets (for example, from established chemical databases) leads to a linear model construction process from data collection → model training. In contrast, dynamic datasets (for example, from guided experiments or calculations) lead to an iterative model-construction process that is sometimes referred to as active learning, with data collection → model training → use model to identify missing data → repeat. Care must be taken with data selection in both regimes.

Most data sources are biased. Bias can originate from the method used to generate or acquire the data, for example, an experimental technique that is more sensitive to heavier elements, or simulation-based datasets that favour materials with small crystallographic unit cells due to limits on the computational power available. Bias can also arise from the context of a dataset compiled for a specific purpose or by a specific sub-community,

Checklist for reporting and evaluating machine learning models	
1. Data sources	
1a. Are all data sources listed and publicly available?	
1b. If using an external database, is an access date or version number provided?	
1c. Are any potential biases in the source dataset reported and/or mitigated?	
2. Data cleaning	
2a. Are the data cleaning steps clearly and fully described, either in text or as a code pipeline?	
2b. Is an evaluation of the amount of removed source data presented?	
2c. Are instances of combining data from multiple sources clearly identified, and potential issues mitigated?	
3. Data representations	
3a. Are methods for representing data as features or descriptors clearly articulated, ideally with software implementations?	
3b. Are comparisons against standard feature sets provided?	
4. Model choice	
4a. Is a software implementation of the model provided such that it can be trained and tested with new data?	
4b. Are baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) provided?	
4c. Are baseline comparisons to current state-of-the-art provided?	
5. Model training and validation	
5a. Does the model clearly split data into different sets for training (model selection), validation (hyperparameter opimization), and testing (final evaluation)?	
5b. Is the method of data splitting (for example, random, cluster- or time-based splitting, forward cross-validation) clearly stated? Does it mimic anticipated real-world application?	
5c. Does the data splitting procedure avoid data leakage (for example, is the same composition present in training and test sets)?	
6. Code and reproducibility	
6a. Is the code or workflow available in a public repository?	
6b. Are scripts to reproduce the findings in the paper provided?	

Fig. 1 | A suggested author and reviewer checklist for reporting and evaluating machine learning models. This proposed checklist is also provided as Supplementary Data 1.

as recently explored for reagent choice and reaction conditions used in inorganic synthesis[13]. A classic example of the perils of a biased dataset came on 3 November 1948, when *The Chicago Tribune* headline declared 'Dewey Defeats Truman' based on projecting results from the previous day's U. S. presidential election. In truth, Truman defeated Dewey (303–189 in the Electoral College). The source of the error? The use of phone-based polls at a time when mostly wealthy (and Republican-leaning) citizens owned phones. One can imagine analogous sampling errors regarding chemical datasets, where particular classes of 'fashionable' compounds such as metal dichalcogenides or halide perovskites may feature widely, but do not represent the diversity of all materials.

It is important to identify and discuss the sources and limitations of a dataset. Bias may be intended and desirable, for example, in the construction of interatomic potentials from regions of a potential energy surface that are most relevant[14], but any bias, or attempts at its mitigation, should be discussed.

Databases often evolve over time, with new data added (continuously or by batch releases). For reasons of reproducibility, it is essential that these databases use some mechanism for version control (for example, release numbers, Git versioning, or timestamps) as part of the metadata and maintain long-term availability to previous versions of the database.

We recommend listing all data sources, documenting the strategy for data selection, and including access dates or version numbers. If data is protected or proprietary, a minimally reproducible example using a public dataset can be an alternative.

2. Data cleaning and curation. Raw datasets often contain errors, omissions, or outliers. It is common for databases to contain over 10% of erroneous data. Indeed, one study found that 14% of the data describing the elastic properties of crystals in the Materials Project is unphysical[15]. Cleaning steps include removing duplicates, entries with missing values, incoherent or unphysical values, or data type conversions. Data curation may also have been performed before publication of the original dataset. This cleaning of the data can also include normalization and homogenization, where several sources are combined. Attention should be given to the characterization of possible discrepancies between sources, and the impact of homogenization on derived machine learning models. The dramatic effect of data quality on model performance and the importance of careful data curation has been highlighted in the closely related field of cheminformatics[16,17]. One seminal study showed examples of how accumulation of database errors and incorrect processing of chemical structures

could lead to significant losses in the predictive ability of machine learning models[18]. When errors are identified in public databases, it is important to communicate these to the dataset maintainer as part of the research process.

The ability of a statistical model to be 'right for the wrong reasons' can occur when the true signal is correlated with a false one in the data. In one notable example, a high-accuracy model was trained to predict the performance of Buchwald–Hartwig cross-coupling[19]. The findings prompted the suggestion that almost the same accuracy could be achieved if all features in the dataset are replaced with random strings of digits[20].

We recommend describing all cleaning steps applied to the original data, while also providing an evaluation of the extent of data removed and modified through this process. As it is impossible to check large databases manually, the implementation and sharing of semi-automated workflows integrating data curation pipelines is crucial.

3. Data representation. The same type of chemical information can be represented in many ways. The choice of representation (or encoding) is critical in model building and can be as important for determining model performance as the choice of machine learning method. It is therefore essential to evaluate different representations when constructing a new model. For the representation of molecules and extended crystals, various approaches have been developed. Some capture the global features of the entire molecule or crystallographic unit cell, while others represent local features such as bonding environments or fragments, and some combine both aspects. Both hand-crafted descriptors, which make use of prior knowledge (and are often computationally efficient), and general learned descriptors (unbiased but usually computationally demanding) can be used. In chemistry, it is beneficial if the chosen representation obeys physical invariants of the system, such as symmetry[21]. While there is merit in developing new approaches, comparison with established methods (both in accuracy and cost) is advisable so that advantages and disadvantages are clear.

We recommend that the methods used for representing data are stated and compared with standard feature sets. It is advisable to draw from the experience of published chemical representation schemes, and their reference implementations in standard open libraries such as RDKit (https://www.rdkit.org), DScribe (https://singroup.github.io/dscribe), and Matminer (https://hackingmaterials.lbl.gov/matminer) before attempting to design new ones.

4. Model choice. Many flavours of machine learning exist, from classical algorithms such as the 'support-vector machines', ensemble methods like 'random forests', to deep learning methods involving complex neural network architectures. High accuracy in tasks involving chemical problems has been reported for graph-based neural networks designed to represent bonding interactions between elements[22,23]. Transfer-learning techniques make it possible to train superior models from the smaller datasets that are common in chemistry, with one success case being the retraining of a general-purpose interatomic potential based on a small dataset of high-quality quantum mechanical calculations[24].

However, the sophistication of a model is unrelated to the appropriateness for a given problem: higher complexity is not always better. In fact, model complexity often comes with the cost of reduced transparency and interpretability. The use of a six-layer neural network to predict earthquake aftershocks[25] was the subject of vigorous online debate, as well as a formal rebuttal[26] demonstrating that a single neuron with only two free parameters (as opposed to the 13,451 of the original model) could provide the same level of accuracy. This case highlights the importance of baselines that include selecting the most frequent class (classification), always predicting the mean (regression), or comparing results against a model with no extrapolative power, such as a 1-nearest-neighbour, which essentially 'looks up' the closest known data point when making a prediction. In cases where machine learning alternatives for conventional techniques are proposed, a comparison with the state-of-the-art is another important baseline test and a general measure of the success of the model.

We recommend justifying your model choice by including baseline comparisons to simpler — even trivial — models, as well as the current state-of-the-art. A software implementation should be provided so that the model can be trained and tested with new data.

5. Model training and validation. Training a robust model must balance underfitting and overfitting, which is important for both the model parameters (for example, weights in a neural network) and hyperparameters (for example, kernel parameters, activation functions, as well as the choice and settings of the training algorithm). Three datasets are involved in model construction and selection. A training set is used as an optimization target for models to learn from for a given choice of hyperparameters. An independent validation set is used to detect overfitting during training of the parameters. The model hyperparameters are optimized against the performance on the validation set. A test set of unseen data is then used to assess the accuracy of the final model and again to detect overfitting. These three sets can be formed from random splits of the original dataset, or by first clustering the data into similar types to ensure a diverse split is achieved. In estimating the training accuracy, the mean-squared errors are usually inspected and reported, but it should be confirmed that the accuracy is achieved uniformly over the whole dataset.

The computational intensiveness of the training process should also be reported as the utility of the approach to others will depend on the data and resource required. For example, sequence-based generative models are a powerful approach for molecular de novo design but training them using recurrent neural networks is currently only feasible if one has access to state-of-the-art graphics processing units and millions of training samples[27]. Following conventional terminology, the validation set is only used during training, whereas the independent test set is used for assessing a trained model prior to application. However, the accuracy of a trained model on an arbitrary test set is not a universal metric for evaluating performance.

The test set must be representative of the intended application range. For example, a model trained on solvation structures and energies under acidic conditions may be accurate on similar data, but not be transferable to basic conditions. Reliable measures of test accuracy can be difficult to formulate. One study assessed the accuracy of machine learning models trained to predict steel fatigue strength or critical temperature of superconductivity using random cross-validation or clustered by a diversity splitting strategy[28]. In the latter scenario, the model accuracies dropped substantially (2–4× performance reduction). The models were extremely fragile to the introduction of new and slightly different data, to the point of losing any predictive power.

Methods of validation that aim to test extrapolative (versus interpolative) performance are being developed either by excluding entire classes of compounds (known as leave-class-out selection or scaffold split) for testing[28], or by excluding the extreme values in the dataset for testing[29]. Another industry standard approach is time-split cross-validation[30], where a model is trained on historical data available at a certain date and tested on data that is generated later, simulating the process of prospective validation.

We recommend stating how the training, validation, and test sets were obtained, as well as the sensitivity of model performance with respect to the parameters of the training method, for example, when training is repeated with different random seeds or ordering of the dataset. Validation should be performed on data related to the intended application.

6. Code and reproducibility. There is a reproducibility crisis across all fields of research. If we set aside cases of outright misconduct and data fabrication, the selective reporting of positive results is widespread. Going deeper, data dredging (*p*-hacking) is a manipulation technique to find outcomes that can be presented as statistically significant, thus dramatically increasing the observed effect. 'Hypothesizing after the results are known' (HARKing) involves presenting a post-hoc hypothesis in a research report as if it were, in fact, an a priori hypothesis. To strengthen public trust in science and improve the reproducibility of published research, it is important for authors to make their data and code publicly available. This goes beyond purely computational studies and initiatives like the 'dark reactions project' to show the unique value of failed experiments that have never been reported in literature[31].

The first five steps require many choices to be made by researchers to train meaningful machine learning models. While the reasoning behind these choices should be reported, this alone is not sufficient to meet the burden of reproducibility[32]. Many variables that are not typically listed in the methods section of a publication can play a role in the final result – the devil is in the hyperparameters. Even software versions are important as default variables often change. For large developments, the report of a standalone code, for example in the *Journal of Open Source Software*, may be appropriate. It is desirable to report auxiliary software packages and versions required to run the reported workflows, which can be achieved by listing all dependencies, by exporting the software environment (for example, conda environments) or by providing standalone containers for running the code. Initiatives are being developed to support the reporting of reproducible workflows, including https://www.commonwl.org, https://www.researchobject.org and https://www.dlhub.org.

We recommend that the full code or workflow is made available in a public repository that guarantees long-term archiving (for example, an online repository archived with a permanent DOI). Providing the code not only allows the study to be exactly replicated by others, but to be challenged, critiqued and further improved. At the minimum, a script or electronic notebook should be provided that contains all parameters to reproduce the results reported.

Maintaining high digital standards

These new adventures in chemical research are only possible thanks to those who have contributed to the underpinning techniques, algorithms, codes, and packages. Developments in this field are supported by an open-source philosophy that includes the posting of preprints and making software openly and freely available. Future progress critically depends on these researchers being able to demonstrate the impact of their contributions. In all reports, remember to cite the methods and packages employed to ensure that the development community receives the recognition they deserve.

The suggestions put forward in this Comment have emerged from interactions with many researchers, and are in line with other perspectives on this topic[33,34]. While there is great power and potential in the application and development of machine learning for chemistry, it is up to us to establish and maintain a high standard of research and reporting. ❐

Editor's note: This article has been peer-reviewed.

Nongnuch Artrith[1,2]✉, Keith T. Butler[3]✉, François-Xavier Coudert[4]✉, Seungwu Han[5]✉, Olexandr Isayev[6,7]✉, Anubhav Jain[8]✉ and Aron Walsh[9,10]✉

[1]*Department of Chemical Engineering, Columbia University, New York, NY, USA.* [2]*Columbia Center for Computational Electrochemistry (CCCE), Columbia University, New York, NY, USA.* [3]*SciML, Scientific Computing Department, STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot, UK.* [4]*Chimie ParisTech, PSL University, CNRS, Institut de Recherche de Chimie Paris, Paris, France.* [5]*Department of Materials Science and Engineering, Seoul National University, Seoul, Korea.* [6]*Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, PA, USA.* [7]*Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, USA.* [8]*Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, California, USA.* [9]*Department of Materials, Imperial College London, London, UK.* [10]*Department of Materials Science and Engineering, Yonsei University, Seoul, Korea.*

Twitter: @nartrith; @keeeto2000; @fxcoudert; @olexandr; @jainpapers; @lonepair

✉*e-mail: na2782@columbia.edu; keith.butler@stfc.ac.uk; fx.coudert@chimieparistech.psl.eu; hansw@snu.ac.kr; olexandr@olexandrisayev.com; ajain@lbl.gov; a.walsh@imperial.ac.uk*

Published online: 31 May 2021
https://doi.org/10.1038/s41557-021-00716-z

References

1. Gasteiger, J. & Zupan, J. *Angew. Chem. Int. Ed.* **32**, 503–527 (1993).
2. Aspuru-Guzik, A. et al. *Nat. Chem.* **11**, 286–294 (2019).
3. Butler, K. T. et al. *Nature* **559**, 547–555 (2018).
4. Deringer, V. L. et al. *J. Phys. Chem. Lett.* **9**, 2879–2885 (2018).
5. Behler, J. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).
6. Kononova, O. et al. *Sci. Data* **6**, 203 (2019).
7. Sanchez-Lengeling, B. & Aspuru-Guzik, A. *Science* **361**, 360–365 (2018).
8. Hutson, M. *Science* **359**, 725–726 (2018).
9. Coudert, F. X. *Chem. Mater.* **29**, 2615–2617 (2017).
10. Lejaeghere, K. et al. *Science* **351**, aad3000 (2016).
11. Smith, D. G. A. et al. *WIREs Comp. Mater. Sci.* **11**, e1491 (2021).
12. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
13. Jia, X. et al. *Nature* **573**, 251–255 (2019).
14. Artrith, N. et al. *J. Chem. Phys.* **148**, 241711 (2018).
15. Chibani, S. & Coudert, F.-X. *Chem. Sci.* **10**, 8589–8599 (2019).
16. Tropsha, A. *Mol. Inform.* **29**, 476–488 (2010).
17. Gramatica, P. et al. *Mol. Inform.* **31**, 817–835 (2012).
18. Young, D., Martin, T., Venkatapathy, R. & Harten, P. *QSAR Comb. Sci.* **27**, 1337–1345 (2008).
19. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. *Science* **360**, 186–190 (2018).
20. Chuang, K. V. & Keiser, M. J. *Science* **362**, eaat8603 (2018).
21. Braams, B. J. & Bowman, J. M. *Int. Rev. Phys. Chem.* **28**, 577 (2009).
22. Chen, C. et al. *Chem. Mater.* **31**, 3564–3572 (2019).
23. Xie, T. & Grossman, J. C. *Phys. Rev. Lett.* **120**, 145301 (2018).
24. Smith, J. S. et al. *Nat. Commun.* **10**, 2903 (2019).
25. DeVries, P. M. R. et al. *Nature* **560**, 632–634 (2018).
26. Mignan, A. & Broccardo, M. *Nature* **574**, E1–E3 (2019).
27. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. *J. Cheminformatics* **9**, 48 (2017).
28. Meredig, B. et al. *Mol. Syst. Des. Eng.* **3**, 819–825 (2018).
29. Xiong, Z. et al. *Comp. Mater. Sci.* **171**, 109203 (2020).
30. Sheridan, R. P. *J. Chem. Inf. Model* **53**, 783–790 (2013).
31. Raccuglia, P. et al. *Nature* **533**, 73–76 (2016).
32. Reproducibility and replicability in science. *The National Academies of Sciences, Engineering, and Medicine* https://www.nationalacademies.org/our-work/reproducibility-and-replicability-in-science (accessed 13 May 2021).
33. Wang, A. Y.-T. et al. *Chem. Mater.* **32**, 4954–4965 (2020).
34. Riley, P. *Nature* **572**, 27–29 (2019).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41557-021-00716-z.

Peer review information *Nature Chemistry* thanks Joshua Schrier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.