Best practices in machine learning for chemistry

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

Best practices in machine learning for chemistry

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

As statistical tools based on machine learning become integrated into chemistry research workflows, we discuss the elements necessary to train and report reliable, repeatable, and reproducible models.

Chemistry has long benefited from the power of applying models to interpret patterns in data. Standard relations range 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 (e.g. electron counting for polyhedral boranes) and conditions (e.g. 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 generalised relationships that have previously escaped human intuition. Yet, practitioners of these techniques must follow careful protocols to achieve similar levels of validity, reproducibility and longevity as 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 will be useful to authors, referees, and readers to guide the critical evaluation of, and provide a degree of standardisation to, the training and reporting of machine learning (ML) models. We propose that publishers can create ML manuscript submission guidelines and reproducibility policy along with the provided checklist. We hope that many scientists will spearhead this campaign and voluntarily provide an ML checklist to support their papers.

The application of statistical ML techniques to chemical systems has a long history[1], but increasing computer power has recently led to an unprecedented growth of 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 and physical properties of molecules and materials has been widely demonstrated

[1] J. Gasteiger and J. Zupan, Angew. Chem. Int. Ed. 32, 503 (1993); https://doi.org/10.1002/anie.199305031
[2] A. Aspuru-Guzik et al, Nature Chem. 11, 286 (2019); https://www.nature.com/articles/s41557-019-0236-7
[3] K. T. Butler et al, Nature 559, 547 (2018); https://doi.org/10.1038/s41586-018-0337-2

using supervised learning for both regression (e.g. reaction rate) and classification (e.g. reaction outcome) problems. Molecular modelling has benefited, for example, 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 of known facts, efficient exploration of chemical hyperspace including the solution of inverse design problems is becoming possible through the application of autoencoders and generative models[7].

Unfortunately, the lack of transparency surrounding data-driven methods lead scientists to question the validity of obtained results[8]. As even the findings of important experimental studies could not be replicated, some argue that science is in "reproducibility crisis"[9]. The transition to an open science ecosystem that includes reproducible research workflows and the publication of supporting data in machine-readable formats is ongoing within chemistry[10]. In computational chemistry, reproducibility and implementations of mainstream methods like density functional theory has been investigated[11]. This, and other studies[12], proposed for open standards that are complemented by the availability of online databases. The same must be done for data-driven methods. ML 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 an ML study to be critically assessed and repeatable. As a community, we must work together to significantly improve the efficiency, effectiveness and reproducibility of ML models and datasets by adhering to the FAIR (findable, accessible, interoperable, reusable) guiding principles for scientific data management and stewardship[13].

Below, we outline a set of guidelines to consider when building and applying ML models. These should assist in the development of robust models, in providing clarity for

[4] V. L. Deringer et al, J. Phys. Chem. Lett. 9, 2879 (2018); https://doi.org/10.1021/acs.jpclett.8b00902
[5] J. Behler, Angew. Chem. Int. Ed. 56, 12828 (2017); https://doi.org/10.1002/anie.201703114
[6] O. Kononova et al, Sci Data 6, (2019); https://www.nature.com/articles/s41597-019-0224-1
[7] B. Sanchez-Lengeling and A. Aspuru-Guzik, Science 361, 360 (2018); https://doi.org/10.1126/science.aat2663
[8] M. Hutson, Science 359, 725 (2018); https://doi.org/10.1126/science.359.6377.725
[9] D. Fanelli, PNAS **115**, 2628 (2018); https://doi.org/10.1073/pnas.1708272114
[10] F. X. Coudert, Chem. Mater. 29, 2615 (2017); https://pubs.acs.org/doi/10.1021/acs.chemmater.7b00799
[11] K. Lejaeghere et al, Science 351, aad3000 (2016); https://science.sciencemag.org/content/351/6280/aad3000
[12] D. G. A. Smith et al, WIREs Comp. Mater. Sci. 11, e1491 (2021); https://onlinelibrary.wiley.com/doi/10.1002/wcms.1491
[13] M. D. Wilkinson et al, Sci. Data 3, 160018 (2016); https://doi.org/10.1038/sdata.2016.18

manuscripts, and in building the credibility needed for statistical tools to gain widespread acceptance and utility in chemistry.

1. Data sources. The quality, quantity and diversity of available data set an upper limit on the accuracy and generality of any derived models. There are established chemical databases that contain structures and properties derived from combinations of measurements and/or simulations[14]. When the data is home-made, as in constructing a training set for ML interatomic potentials, the detailed condition of data generation should be provided for reproducibility. It is important to recognise that 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 computational power available. Bias can also arise from the context of a dataset compiled for a specific purpose or by a specific sub-community, as recently explored for reagent choice and reaction conditions used in inorganic synthesis[15]. A classic example of the perils of a biased dataset came on November 3, 1948, when The Chicago Tribune headline declared "Dewey Defeats Truman" based on projecting results from the previous day's presidential election. In truth, Truman handily 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 "hot" compounds such as metal dichalcogenides or halide perovskites may feature widely, but do not represent the diversity of all materials.

We distinguish between two cases: (i) static datasets (e.g. from published databases) lead to a linear model construction process from data collection → model training; (ii) dynamic datasets (e.g. from guided experiments or calculations) lead to an iterative model construction process that is sometimes referred to as active learning[16], with data collection → model training → use model to identify missing data → repeat. It is important to identify and discuss the source and limitations of the dataset including consequences of bias.[17] Bias may of course be intended and desirable, e.g. the construction of interatomic potentials from regions of a potential energy surface that are most relevant[18]. Databases often evolve over time, with new data added (continuously or by batch releases). For reasons of reproducibility, it is crucial that these databases use some mechanism for version control (e.g. release numbers, Git versioning, or timestamps) as part of the metadata and maintain long-term availability to previous versions of the database. *We*

[14] L. Glasser, J. Chem. Educ. 93, 542 (2016); https://doi.org/10.1021/acs.jchemed.5b00253
[15] X. Jia et al, Nature 573, 251 (2019); https://www.nature.com/articles/s41586-019-1540-5
[16] J. S. Smith et al, J. Chem. Phys. 148, 241733 (2018); https://doi.org/10.1063/1.5023802
[17] J. Sieg et al, J. Chem. Inf. Model. 59, 947 (2019); https://doi.org/10.1021/acs.jcim.8b00712
[18] N. Artrith et al, J. Chem. Phys. 148, 241711 (2018); https://doi.org/10.1063/1.5017661

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. A recent study of mechanical properties found that materials databases can contain 10 to 20% of unphysical data[19]. Cleaning steps include removing duplicates, entries with missing values, incoherent or unphysical values, or type conversions. Data curation may also have been performed before publication of the original dataset. This cleaning of the data can also include normalisation and homogenisation, where several sources are combined. Attention should be given to the characterisation of possible discrepancies between sources, and the impact of homogenisation on derived ML models. The importance of a careful data curation has been highlighted in the closely related field of cheminformatics[20,21]. One seminal study showed examples of how accumulation of database errors and incorrect processing of chemical structures could lead to significant losses of predictive ability of ML models[22]. 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 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 ML model was trained to predict the performance of Buchwald−Hartwig cross-coupling[23] The findings prompted a debate that almost the same accuracy could be achieved if all features in the dataset are replaced with random strings of digits[24]. *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 a critical choice in model building, which can be as important for determining model performance as the ML method itself. It is therefore essential to evaluate different methods when constructing a new model.

[19] S. Chibani and F.-X. Coudert, Chem. Sci. 10, 8589 (2019); https://doi.org/10.1039/C9SC01682A
[20] A. Tropsha, Mol. Inf. 29, 476 (2010); https://doi.org/10.1002/minf.201000061
[21] P. Gramatica et al, Mol. Inform. 31, 817 (2012); https://doi.org/10.1002/minf.201200075
[22] D. Young, T. Martin, R. Venkatapathy, and P. Harten, QSAR Comb. Sci. 27, 1337 (2008); https://onlinelibrary.wiley.com/doi/abs/10.1002/qsar.200810084
[23] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, and A. G. Doyle, Science 360, 186 (2018); https://science.sciencemag.org/content/360/6385/186
[24] K. V. Chuang and M. J. Keiser, Science 362, (2018); https://science.sciencemag.org/content/362/6416/eaat8603

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 that make use of prior knowledge (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[25]. 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 drawing 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 ML 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[26,27]. Transfer learning techniques make it possible to train more powerful 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 [28]. However, the sophistication of a model is unrelated to the appropriateness for a given problem: more complex is not always better! In fact, model complexity comes often comes with the cost of reduced transparency and interpretability. A report of using a 6-layer neural network to predict earthquake aftershocks[29] was the subject of vigorous online debate, as well as a formal rebuttal [30] demonstrating that a single neuron with only 2 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-neighbor, which essentially "looks up" the closest known data point when making a prediction. In cases where ML 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*

[25] B. J. Braams and J. M. Bowman, Int. Rev. Phys. Chem. 28, 577 (2009); https://doi.org/10.1080/01442350903234923
[26] C. Chen et al, Chem. Mater. 31, 3564 (2019); https://doi.org/10.1021/acs.chemmater.9b01294
[27] T. Xie and J. C. Grossman, Phys. Rev. Lett. 120, 145301 (2018); https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.120.145301
[28] J. S. Smith et al, Nature Comm. 10, 2903 (2019); https://www.nature.com/articles/s41467-019-10827-4
[29] P. M. R. DeVries et al, Nature 560, 632 (2018); https://www.nature.com/articles/s41586-018-0438-y
[30] A. Mignan and M. Broccardo, Nature 574, E1 (2019); https://www.nature.com/articles/s41586-019-1582-8

recommend justifying your model choice by including baseline comparisons to simpler, even trivial, models.

5. Model training and validation. Training a robust model must balance underfitting and overfitting, which is important for both the model parameters (e.g. weights in a neural network) and hyperparameters, such as kernel parameters, activation functions, and the choice and parameters of the training algorithm. Three datasets are involved in model construction and selection. A *training set* is used as an optimisation 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 optimised 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 data set, 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[31].

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 will unlikely transfer to basic conditions. Reliable measures of test accuracy can be difficult to formulate. One study assessed accuracy of ML models trained to predict steel fatigue strength or critical temperature of superconductivity using random cross-validation or clustered by diversity splitting strategy. In the later scenario, the model accuracies dropped substantially (2-4x performance reduction). The models were extremely fragile to the introduction of new yet 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 holding out entire classes of compounds (known as leave-class-

[31] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, J. Cheminformatics 9, 48 (2017); https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0235-x

out selection or scaffold split) for testing[32] or by holding out the extreme values in the dataset for testing[33]. Another industry standard approach is time-split cross-validation,[34] 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, e.g. 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 (or *p*-value 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 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" show the unique value of failed experiments that have never been reported in literature[35].

The first 5 steps require many choices to be made by researchers to train meaningful ML models. While the reasoning behind these choices should be reported, this is not sufficient to meet the burden of reproducibility[36]. 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 (e.g. Conda environment) or by providing standalone containers (e.g. Docker containers) for running the code. A number of 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, at the minimum, a script or electronic notebook should be provided that contains all parameters to reproduce the results, ideally in an*

[32] B. Meredig et al, Mol. Syst. Des. Eng., 3, 819 (2018); https://doi.org/10.1039/C8ME00012C
[33] Z. Xiong et al, Comp. Mater. Sci. 171, 109203 (2020); https://doi.org/10.1016/j.commatsci.2019.109203
[34] R. P. Sheridan, J. Chem. Inf. Model. 53, 783 (2013); https://doi.org/10.1021/ci400084k
[35] P. Raccuglia et al, Nature 533, 73 (2016); https://www.nature.com/articles/nature17439
[36] https://www.nationalacademies.org/our-work/reproducibility-and-replicability-in-science

online repository that guarantees long-term archiving (e.g. a repository archived with a permanent DOI).

These new adventures in chemical research are only possible thanks to those who have contributed to the underpinning ML 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 receive 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[37,38]. 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.

[37] A. Yu-Tung et al, Chem. Mater. 32, 4954 (2020); https://doi.org/10.1021/acs.chemmater.0c01907
[38] P. Riley, Nature 572, 27 (2019); https://www.nature.com/articles/d41586-019-02307-y

Checklist for reporting and evaluating machine learning models	
1. Data sources	
1a. Is the data used to train the model publicly available?	✓
1b. If using an external database, is a version / date provided?	✓
1c. Are any potential biases in the source data set 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 source data removed 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 (e.g. 1-nearest neighbor, 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 optimisation), and testing (final evaluation)?	✓
5b. Is the method of data split (e.g. random split, cluster-based splitting, time-based split, forward cross-validation) clearly stated? Does it accurately mimic anticipated real-world application?	✓
5c. Does the data splitting procedure avoid data leakage (e.g. 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?	✓

A suggested author and reviewer checklist for reporting and evaluating machine learning models.

Author Affiliations
[1]Department of Chemical Engineering, Columbia University, New York, NY 10027, USA. [2]Columbia Center for Computational Electrochemistry (CCCE), Columbia University, New York, NY 10027, USA. [3]SciML, Scientific Computing Department, STFC Rutherford Appleton Laboratory, Harwell Campus, Didcot OX11 0QX, UK [4]Chimie ParisTech, PSL Research University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France. [5]Department of Materials Science and Engineering, Seoul National University, Seoul 08826, Korea. [6]Computational Biology Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, PA 15213, USA. [7]Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [8]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, California

94720, USA. [9]Department of Materials, Imperial College London, London SW7 2AZ, UK [10]Department of Materials Science and Engineering, Yonsei University, Seoul 03722, Korea.