Clever Materials: When Machine Learning Models Fool Themselves

Kevin Maik Jablonka [1,2,3,4,✉]

[1]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany
[2]Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstrasse 12-14, 07743 Jena, Germany
[3]Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
[4]Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
[✉]mail@kjablonka.com

February 5, 2026

Abstract

...

1 Introduction

Learning from data is appealing but also scary. It is appealing because it promises to make predictions that are too subtle or complex for humans to recognize.[1] It is scary because some of these patterns might be meaningless and thus lead to complete failures of predictions.[2]

Exploiting meaningless patterns — which can lead to impressive performance in demonstrations but complete failures in other tests — is a known failure mode of pattern recognition. It is often discussed in the context of the horse "Clever Hans".[3] The horse "Clever Hans" was thought to be able to count. This was until it was found that "Clever Hans" could not count but rather relied on subtle clues from his owner in creating the answer.

In image recognition models, this Clever Hans effect has been demonstrated in breadth. Models rely on spurious correlations (also called "shortcuts")[4,5] — skin color on clinical images,[6] or the background on images of animals,[7]. Recently, the startup Leash Biosciences reported that such failure modes might also hamper bioactivity predictions.[8] Models are surprisingly good at predicting who produced a compound. This is an intent signal that models might exploit in making activity predictions.

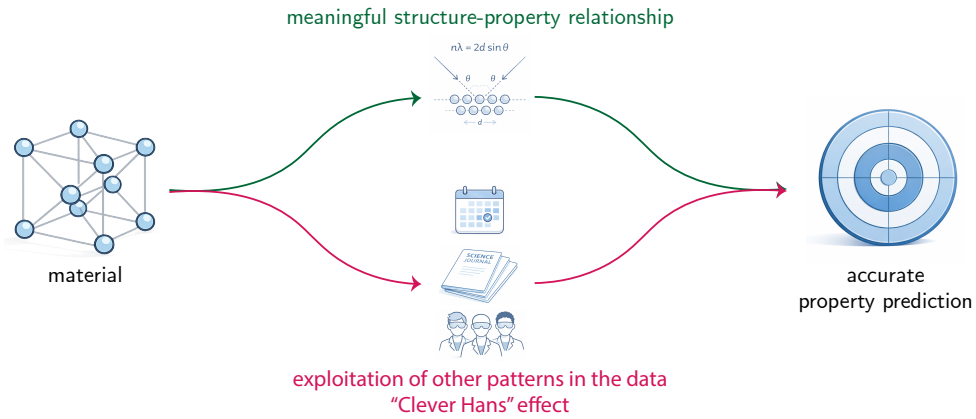

Figure 1: In machine learning, we often do not test competing hypotheses of how a model might obtain its answers. Machine learning models in materials science are trained to map material (descriptors) to property predictions. Models have much flexibility in how they learn this map from data. Ideally, they discover robust and meaningful structure-property relationships that also generalize in new settings. This, however, is not guaranteed. Models might also exploit other patterns in the data as a shortcut to a prediction ("purple"). For instance, it might be easy for the model to spot what researchers produced a given material or in what journal it has been published. Based on those inferences, it might deduce property "guesses" (as the knowledge of the research group or the publication time can be correlated to the property). The model thus might learn to make good predictions for the wrong reasons. This is known as the "Clever Hans" effect. The scientific method asks us to test if such alternative patterns can explain good model performance. This is seldom done. In this work, I do it for a few case studies.

This seems like a reasonable, but also worrying, failure mode. Also in other chemical and material domains one might expect a lot of potential for proxy signals a model could exploit that are not directly linked to any (causal) understanding of chemistry: We know that certain groups focus on optimizing stability of solar cells, the university or research group name is sometimes even part of the name of a metal-organic framework — and we know the research agendas of different groups (Figure 1). Similarly, we know that research fields evolve and performance tends to grow with discoveries of certain materials design features:[9] Be it self-assembled monolayers in perovskite solar cells or post-synthetic modification in MOFs.

In this work, I investigate how relevant such confounders are in commonly used materials datasets. For this, I enrich datasets that have been used for material property prediction tasks with meta information such as author names, publication years, or publication venue. I show that this metainformation can often be predicted with

surprising accuracy. Models trained on those predicted "metainformation fingerprints" show, in some cases, performance that is indistinguishable from models that are trained on "meaningful" descriptors.

These findings underline — again — how easy it is to fool oneself in machine learning (in the chemical sciences). But they also provide indications on how to improve data collection efforts and improve the validation of scientific machine learning models.

2 Results

I analyzed how severe the Clever Hans problem might be across five different materials case studies spanning organic electronics, energy storage, and porous materials. Since results depend not only on the dataset but also on the prediction task and the metrics one uses to analyze the results, I report additional variants of some of the case studies in the appendix. The code is reusable and can be easily applied to other case studies.

2.1 MOF Thermal Stability

Metal-organic frameworks (MOF) are being trialed for various applications.[10] In many of those applications, thermal stability is an essential criterion. To be able to model this, Nandy *et al.*[11] mined thermogravimetric analysis data from the literature and derived decomposition temperatures. Based on this, one can build models to predict the decomposition temperature (a regression problem)[12] or whether a given structure belongs to the most stable ones (a binary classification problem). Figure 2 shows how well one can predict meta-information based on materials descriptors and, based on those predictions, whether the MOF belongs to the top 10% most stable ones.

2.2 MOF Solvent Stability

Besides thermal stability, solvent stability is an essential requirement for MOF application. Solvent stability describes whether a structure collapses when bound solvent (e.g., from the synthesis) is removed under activation (e.g., using vacuum and heat). This data, too, was text-mined by Nandy *et al.*[11] and I used the same descriptors. And for this property, too, Nandy *et al.*[12] reported machine learning models. As for the thermal stability case study, I use similar features in my model.

Figure 3 shows that for this case, too, materials descriptors can predict bibliographic information with high accuracy. While the model trained to directly map material descriptors to solvent stability outperforms the one where we use the predicted bibliographic information to predict solvent stability, this "Clever Hans" model still meaningfully outperforms the baseline.

Figure 2: For the classification task of membership in the top-10% of thermally stable MOFs, one can be fooled (by Clever Hans effects). The model can predict the bibliographic information with high accuracy. **a** The model predicts the authors of the associated paper with high accuracy, much better than a random baseline. **b** This also holds for predicting in which journal the entry was published or the year in which the paper was published (**c**). Using the predicted bibliographic information, a model can also predict with high accuracy if the MOF belongs to the top-10% thermally stable ones. However, one needs to highlight that the effect is smaller — or not even there — if analyzed under a different metric or for a regression setting. The dummy baseline for classification is a stratified random sampling (using the empirical probabilities from the training dataset) and the mean prediction for the regression case.

Figure 3

2.3 Perovskite Solar Cell Efficiency

Perovskite solar cells are another way in which materials scientists aim to have a positive impact on the energy transition.[13] A metric of central importance here is the power conversion efficiency (PCE). It was mined by Jacobsson *et al.*[14] in a manual approach and by Shabih *et al.*[9] in an automated one with large language model-based data extraction.[15]

As a case study, I again analyze whether we can predict whether the device belongs to the top 10% efficient devices. For both the "Conventional" and "Clever Hans" models, I derive descriptors for the composition of the absorber.

2.4 Battery Capacity

For sustainability, energy does not only need to be converted. For this, batteries are important. Huang & Cole[16] text-mined battery materials alongside performance metrics.

Figure 5 shows that while the model can predict bibliographic information better than the random baseline based on the composition features, it still does so with relatively low performance. The prediction of the publication year shows higher performance.

Figure 4

Figure 5: Battery capacity

2.5 TADF Emitter Properties

Thermally activated delayed fluorescence (TADF) is one mechanism to improve the efficiency of organic light-emitting diodes (OLED)s.[17] The maximum emission wavelength is one important performance metric that is optimized for these materials. Huang & Cole[18] text-mined using ChemDataExtractor.[19,20]

As a feature set for the "Conventional" and "Clever Hans" models, I use a broad set of molecular descriptors and fingerprints. Both "Conventional" and "Clever Hans" models aim to predict the maximum emission wavelength.

3 Discussion

Machine learning has transformed materials discovery,[21–23] but the findings here highlight a critical gap: we often fail to rigorously test alternative hypotheses for why our models perform well. The scientific method demands that we actively seek to falsify our hypotheses,[24–26] yet in machine learning, we tend to focus on optimizing performance rather than exploring competing explanations.

The Clever Hans effect represents just one class of alternative hypotheses we should systematically explore. When we claim that models "learn meaningful chemistry," we must test whether simpler explanations — such as shortcuts via author identity or publication date — could account for the observed performance.[27] This requires a shift from asking "does this model work?" to "why does this model work, and what are all the ways it could be wrong?"

The space of potential confounders is vast and often non-obvious. Beyond the shortcuts via meta-information investigated in this work, models might exploit dataset construction artifacts, measurement biases, or many other spurious effects.[28,29] Systematically exploring these alternatives is computationally intensive but crucial for scientific rigor.

LLM-based agents might offer a promising approach to automate this exploration.[30,31] These systems could generate and test competing hypotheses in parallel, exploring the space of potential explanations more thoroughly than human researchers typically manage. Such agents could serve as "devil's advocates," systematically challenging our assumptions about why models succeed.

3.1 Toward Robust Materials Data Infrastructure

Another angle is to reconsider how we generate, curate, and share materials data. The field needs coordinated infrastructure that prioritizes diversity and robustness

over convenience.[32] Convenience and short-term reward are often too easy and compelling to optimize for due to collective action problems trapping an ecosystem in a suboptimal state, where every actor knows that changes would be needed, but no one wants to make the first move.[33]

Most automated screening approaches optimize specific objectives using limited building blocks, which creates exactly the kind of proxy signals that models learn to exploit. And also human researchers are biased in how they explore chemical space.[34] Organizations that can generate diverse data at scale—potentially focused on the lowest cost per reproducible data point rather than pushing particular research agendas—might help to address this problem.

But it is important to keep in mind that in some circumstances, we will never be able to acquire "enough" data. Thus, we also need renewed focus on how we evaluate models.[35] Instead of asking whether models work, we should ask why they work and systematically explore alternative explanations. This means actively trying to disprove our own — but also others' — claims about model performance. To enable others to do so, access to data and code is obviously a prerequisite. But one could also envision that some of these tests might require new experiments — which could be facilitated using infrastructure as a service or incentivized using "bug bounties" for research papers, models, or datasets. We need to accept that receiving feedback — even if it is pointing out a mistake in our own work — is a gift.

4 Conclusions

Model evaluation has always been challenging in materials science.[36] We have developed increasingly sophisticated techniques to address this: time-based splitting,[37,38] scaffold splits, leave-one-cluster-out cross-validation,[39,40] cluster-based splits,[41] and property-based splits.[42,43] In some domains, even challenges have been organized.[44–46] Each technique revealed new ways that models could fail to generalize, forcing us to be more rigorous in our evaluation practices.

This work highlights yet another layer of complexity. Models can achieve impressive performance not by learning meaningful chemistry, but by exploiting subtle biases in how our datasets are constructed. Across five different materials domains, I find that proxy signals—such as shortcut learning via publication meta-information—can provide substantial predictive power.

The simplest explanation for good model performance might often be shortcut learning, not meaningful understanding of chemistry. This is an uncomfortable truth, but the space of potential confounders extends far beyond what current evaluation techniques can catch.

Like the original Clever Hans, our models may be performing impressive feats—but for all the wrong reasons. They excel not because they understand chemistry,

but because they have learned to read the subtle clues inadvertently embedded in our data. Thus, we should not only ask whether our models can achieve good performance, but also whether we can trust what that performance actually means.

This is not necessarily a condemnation of all shortcut learning. Models that exploit proxy signals can still provide statistically reliable predictions and practical value—as long as the underlying patterns remain stable and as long as we only care about the average performance. The critical issue is transparency about what we are doing and why.

If our goal is scientific understanding and robust generalization to genuinely new materials, we must systematically explore alternative explanations and build models that resist spurious correlations. If our goal is simply a predictive tool that works well on average, we can accept some brittleness—but we should communicate openly that the model may fail in unpredictable ways when the hidden assumptions break down.

In the end, the Clever Hans problem forces us to confront a choice about scientific machine learning: Do we want tools that advance chemical understanding, or are we content with sophisticated pattern matchers that reflect various biases and artifacts? Both approaches can have merit, but honesty about which path we are taking will help to unlock real acceleration using machine learning.

5 Methods

5.1 Clever Hans Analysis Framework

I implemented a systematic framework to quantify Clever Hans effects in materials property prediction. For each dataset, I trained three types of models: (1) conventional models that predict material properties directly from chemical descriptors, (2) indirect models that first predict meta-information (author identity, journal, publication year) from the same descriptors and then use these predictions to estimate material properties, and (3) dummy baselines using stratified sampling for classification or mean prediction for regression.

The indirect prediction approach tests whether meta-information contains sufficient signal to achieve competitive prediction performance. If models can predict material properties as accurately using only proxy information as using chemical descriptors, this indicates potential Clever Hans effects in the dataset.

5.2 Model Architecture and Training

All models used gradient boosting, implemented with `LightGBM`[47] with default hyperparameters. I performed 10-fold cross-validation with random shuffling for all analyses. Each fold compared performance across the three model types using identical train/test splits to ensure paired comparisons.

5.3 Datasets and Feature Engineering

5.3.1 Battery Dataset

I obtained the battery dataset from Huang & Cole[16]

5.3.2 Perovskite Dataset

I obtained the perovskite dataset from Shabih *et al.*[9], which is based on Jacobsson *et al.*[14].

5.3.3 MOF Datasets

I obtained the MOF datasets from Nandy *et al.*[11]. The dataset already contains pre-computed features such as revised autocorrelation functions.[21]

5.3.4 TADF Dataset

I obtained the TADF dataset from Huang & Cole[18].

5.4 Chemical Descriptor Generation

For datasets containing molecular or compositional information, I generated comprehensive chemical descriptors to serve as baseline features for property prediction.

5.4.1 Molecular Descriptors from SMILES

For datasets with SMILES (Simplified Molecular-Input Line-Entry System) strings,[48] I computed molecular descriptors using RDKit[49]. The molecular feature set included:

- **2D descriptors**: All available RDKit molecular descriptors (∼200 features), including molecular weight, LogP, topological polar surface area, number of aromatic rings, hydrogen bond donors/acceptors, and rotatable bonds.
- **Fingerprints**: 2048-bit circular fingerprints with radius 2, capturing local chemical environments and structural motifs.

Molecules were parsed from SMILES strings, and invalid or unparseable structures were excluded.

5.4.2 Composition Descriptors

For datasets with chemical formulas (battery materials, perovskites), I computed composition-based descriptors using matminer[50]. The composition feature set included:

- **Element properties**: Elemental statistics (mean, standard deviation, range) for atomic properties including atomic radius, electronegativity, ionization energy, and electron affinity using the Magpie preset[51].
- **Stoichiometric features**: Composition statistics including element fractions, number of components, and chemical complexity metrics.
- **Meredig descriptors**: Extended element property statistics including orbital contributions and chemical bonding characteristics[52].

Chemical formulas were parsed using pymatgen[53], and compositions that could not be parsed were excluded from analysis.

5.4.3 Feature Processing

Generated descriptors were processed to handle missing values and ensure numerical stability for gradient boosting models. Features with excessive missing values ($>$50%) were excluded, and remaining missing values were imputed with feature medians. For XGBoost and LightGBM models, additional preprocessing included clipping extreme values to prevent numerical overflow and replacing infinite values with conservative bounds.

5.5 Meta-Information Extraction

I enriched the datasets with publication meta-information using the Crossref API to retrieve bibliographic data, including author names, journal titles, and publication years. I created binary features indicating the presence of the top-N most frequent authors and journals in each dataset, where N was varied across 10, 50, 100, and 500 (or maximum available).

5.6 Data Processing

All datasets were preprocessed to remove entries with missing target values or author information.

Data Availability

The datasets used in this study are available on Zenodo: [DOI placeholder].

Code Availability

All analysis code is available on GitHub: [URL placeholder].

Acknowledgement

This work was supported by the Carl Zeiss Stiftung. The author is member of the NFDI consortium FAIRmat - Deutsche Forschungsgemeinschaft (DFG) - Project 460197019.

Declaration of Generative AI and AI-assisted Technologies in the Research and Writing Process

I used Anthropic's Claude models as "copilot" in code development. I also used those models to improve language and readability. After using this service, I reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Hardt, M. & Recht, B. *Patterns, predictions, and actions: Foundations of machine learning* (Princeton University Press, 2022).
2. Lones, M. A. Avoiding common machine learning pitfalls. *Patterns* **5,** 101046. ISSN: 2666-3899. http://dx.doi.org/10.1016/j.patter.2024.101046 (Oct. 2024).
3. Lapuschkin, S. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications* **10.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-019-08987-4 (Mar. 2019).
4. Brown, A. *et al.* Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications* **14.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-023-39902-7 (July 2023).
5. Howard, F. M. *et al.* The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications* **12.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-021-24698-1 (July 2021).
6. Pooch, E. H. P., Ballester, P. L. & Barros, R. C. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv: 1909.01940* (2019).

7. Xiao, K. Y., Engstrom, L., Ilyas, A. & Mądry, A. Noise or Signal: The Role of Image Backgrounds in Object Recognition. *International Conference on Learning Representations* (2020).

8. Blevins, A. D. & Quigley, I. K. Clever Hans in Chemistry: Chemist Style Signals Confound Activity Prediction on Public Benchmarks. https://github.com/Leash-Labs/chemist-style-leaderboard/blob/trunk/clever_hans.pdf (2025).

9. Shabih, S. *et al.* An autonomous living database for perovskite photovoltaics. *arXiv preprint arXiv: 2601.17807* (2026).

10. Kalmutzki, M. J., Hanikel, N. & Yaghi, O. M. Secondary building units as the turning point in the development of the reticular chemistry of MOFs. *Science Advances* **4.** ISSN: 2375-2548. http://dx.doi.org/10.1126/sciadv.aat9180 (Oct. 2018).

11. Nandy, A. *et al.* MOFSimplify, machine learning models with extracted stability data of three thousand metal–organic frameworks. *Scientific Data* **9.** ISSN: 2052-4463. http://dx.doi.org/10.1038/s41597-022-01181-0 (Mar. 2022).

12. Nandy, A., Duan, C. & Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal–Organic Frameworks. *Journal of the American Chemical Society* **143,** 17535–17547. ISSN: 1520-5126. http://dx.doi.org/10.1021/jacs.1c07217 (Oct. 2021).

13. Correa-Baena, J.-P. *et al.* Promises and challenges of perovskite solar cells. *Science* **358,** 739–744. ISSN: 1095-9203. http://dx.doi.org/10.1126/science.aam6323 (Nov. 2017).

14. Jacobsson, T. J. *et al.* An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nature Energy* **7,** 107–115. ISSN: 2058-7546. http://dx.doi.org/10.1038/s41560-021-00941-3 (Dec. 2021).

15. Schilling-Wilhelmi, M. *et al.* From text to insight: large language models for chemical data extraction. *Chemical Society Reviews* **54,** 1125–1150. ISSN: 1460-4744. http://dx.doi.org/10.1039/D4CS00913D (2025).

16. Huang, S. & Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Scientific Data* **7.** ISSN: 2052-4463. http://dx.doi.org/10.1038/s41597-020-00602-2 (Aug. 2020).

17. Liu, Y., Li, C., Ren, Z., Yan, S. & Bryce, M. R. All-organic thermally activated delayed fluorescence materials for organic light-emitting diodes. *Nature Reviews Materials* **3.** ISSN: 2058-8437. http://dx.doi.org/10.1038/natrevmats.2018.20 (Apr. 2018).

18. Huang, D. & Cole, J. M. A database of thermally activated delayed fluorescent molecules auto-generated from scientific literature with ChemDataExtractor. *Scientific Data* **11.** ISSN: 2052-4463. http://dx.doi.org/10.1038/s41597-023-02897-3 (Jan. 2024).

19. Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling* **56,** 1894–1904. ISSN: 1549-960X. http://dx.doi.org/10.1021/acs.jcim.6b00207 (Oct. 2016).

20. Mavračić, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *Journal of Chemical Information and Modeling* **61,** 4280–4289. ISSN: 1549-960X. http://dx.doi.org/10.1021/acs.jcim.1c00446 (Sept. 2021).

21. Moosavi, S. M., Jablonka, K. M. & Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *Journal of the American Chemical Society* **142,** 20273–20287. ISSN: 1520-5126. http://dx.doi.org/10.1021/jacs.0c09105 (Nov. 2020).

22. Saal, J. E., Oliynyk, A. O. & Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annual Review of Materials Research* **50,** 49–69. ISSN: 1545-4118. http://dx.doi.org/10.1146/annurev-matsci-090319-010954 (July 2020).

23. Gubernatis, J. E. & Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Physical Review Materials* **2.** ISSN: 2475-9953. http://dx.doi.org/10.1103/PhysRevMaterials.2.120301 (Dec. 2018).

24. Platt, J. R. Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* **146,** 347–353. ISSN: 1095-9203. http://dx.doi.org/10.1126/science.146.3642.347 (Oct. 1964).

25. Popper, K. *The logic of scientific discovery* (Routledge, 2005).

26. Chamberlin, T. C. The Method of Multiple Working Hypotheses: With this method the dangers of parental affection for a favorite theory can be circumvented. *Science* **148,** 754–759. ISSN: 1095-9203. http://dx.doi.org/10.1126/science.148.3671.754 (May 1965).

27. Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362.** ISSN: 1095-9203. http://dx.doi.org/10.1126/science.aat8603 (Nov. 2018).

28. Zhou, W., Liu, F., Zheng, H. & Zhao, R. Mitigating data bias and ensuring reliable evaluation of AI models with shortcut hull learning. *Nature Communications* **16.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-025-60801-6 (July 2025).

29. Jones, C. *et al.* A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence* **6,** 138–146. ISSN: 2522-5839. http://dx.doi.org/10.1038/s42256-024-00797-8 (Feb. 2024).

30. Ramos, M. C., Collison, C. J. & White, A. D. A review of large language models and autonomous agents in chemistry. *Chemical Science* **16,** 2514–2572. ISSN: 2041-6539. http://dx.doi.org/10.1039/D4SC03921A (2025).

31. Alampara, N. *et al.* General purpose models for the chemical sciences. *arXiv e-prints,* arXiv–2507 (2025).

32. Krishnan, N. M. A. & Jablonka, K. M. Real AI advances require collaboration. *Nature Reviews Chemistry* **9,** 573–574. ISSN: 2397-3358. http://dx.doi.org/10.1038/s41570-025-00750-2 (Aug. 2025).

33. Nielsen, M. *Reinventing discovery* 2nd ed. en (Princeton University Press, Princeton, NJ, Apr. 2020).

34. Jia, X. *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573,** 251–255. ISSN: 1476-4687. http://dx.doi.org/10.1038/s41586-019-1540-5 (Sept. 2019).

35. Goldman, J. & Tsotsos, J. K. Statistical Challenges with Dataset Construction: Why You Will Never Have Enough Images. *arXiv preprint arXiv: 2408.11160* (2024).

36. Alampara, N., Schilling-Wilhelmi, M. & Jablonka, K. M. Lessons from the trenches on evaluating machine learning systems in materials science. *Computational Materials Science* **259,** 114041. ISSN: 0927-0256. http://dx.doi.org/10.1016/j.commatsci.2025.114041 (Sept. 2025).

37. Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of chemical information and modeling* **53,** 783–790 (2013).

38. Landrum, G. A. *et al.* SIMPD: an algorithm for generating simulated time splits for validating machine learning approaches. *Journal of Cheminformatics* **15.** ISSN: 1758-2946. http://dx.doi.org/10.1186/s13321-023-00787-9 (Dec. 2023).

39. Durdy, S., Gaultois, M. W., Gusev, V. V., Bollegala, D. & Rosseinsky, M. J. Random projections and kernelised leave one cluster out cross validation: universal baselines and evaluation tools for supervised machine learning of material properties. *Digital Discovery* **1,** 763–778. ISSN: 2635-098X. http://dx.doi.org/10.1039/D2DD00039C (2022).

40. Meredig, B. *et al.* Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering* **3,** 819–825. ISSN: 2058-9689. http://dx.doi.org/10.1039/C8ME00012C (2018).

41. Guo, Q., Hernandez-Hernandez, S. & Ballester, P. J. Scaffold Splits Overestimate Virtual Screening Performance. *arXiv preprint arXiv: 2406.00873* (2024).

42. Jablonka, K. M., Rosen, A. S., Krishnapriyan, A. S. & Smit, B. An Ecosystem for Digital Reticular Chemistry. *ACS Central Science* **9,** 563–581. ISSN: 2374-7951. http://dx.doi.org/10.1021/acscentsci.2c01177 (Mar. 2023).

43. Kunchapu, S. & Jablonka, K. M. PolyMetriX: an ecosystem for digital polymer chemistry. *npj Computational Materials* **11,** 312 (2025).

44. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Current Opinion in Structural Biology* **15,** 285–289. ISSN: 0959-440X. http://dx.doi.org/10.1016/j.sbi.2005.05.011 (June 2005).

45. Tetko, I. V. Tox24 Challenge. *Chemical Research in Toxicology* **37,** 825–826. ISSN: 1520-5010. http://dx.doi.org/10.1021/acs.chemrestox.4c00192 (May 2024).

46. Llinas, A. & Avdeef, A. Solubility Challenge Revisited after Ten Years, with Multilab Shake-Flask Data, Using Tight (SD ~ 0.17 log) and Loose (SD ~ 0.62 log) Test Sets. *Journal of Chemical Information and Modeling* **59,** 3036–3040. ISSN: 1549-960X. http://dx.doi.org/10.1021/acs.jcim.9b00345 (May 2019).

47. Shi, Y. *et al. lightgbm: Light Gradient Boosting Machine* R package version 4.6.0 (2025). https://github.com/Microsoft/LightGBM.

48. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **28,** 31–36. ISSN: 1520-5142. http://dx.doi.org/10.1021/ci00057a005 (Feb. 1988).

49. Landrum, G. RDKit: Open-Source Cheminformatics Software. https://github.com/rdkit/rdkit/ (2025).

50. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152,** 60–69. ISSN: 0927-0256. http://dx.doi.org/10.1016/j.commatsci.2018.05.018 (Sept. 2018).

51. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2,** 1–7 (2016).

52. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* **89,** 094104 (2014).

53. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68,** 314–319 (2013).

A Detailed Results

In this section, I show performance for metadata and property prediction in more detail.