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 7, 2026

Abstract

Machine learning promises to accelerate materials discovery by uncovering complex structure-property relationships—but what if our models succeed for the wrong reasons? Here, I show that across five materials domains—metal-organic frameworks, perovskite solar cells, batteries, and organic electronics—models achieve competitive performance by exploiting bibliographic metadata rather than learning meaningful chemistry. This is the materials science equivalent of the 'Clever Hans' effect: the horse that appeared to calculate but merely read its trainer's cues. Models trained on chemical descriptors predict publication metadata (authors, journals, years) with surprising accuracy, and models using only this predicted metadata often match the performance of conventional structure-property approaches. In predicting the top 10% most thermally stable metal-organic frameworks, metadata shortcuts prove indistinguishable from descriptor-based models under certain metrics. We have optimized performance without testing why our models work. The path forward requires rigorous evaluation frameworks that falsify competing hypotheses, diverse datasets designed to resist spurious correlations, and honesty about whether we seek chemical understanding or statistical utility. Both have value—but only systematic hypothesis testing, asking not just whether models work but why, elevates pattern matching to science.

1 Introduction

Machine learning holds both promise and peril for materials discovery. The promise lies in its capacity to uncover structure-property relationships too subtle or complex for human recognition.[1] The peril emerges when models excel by exploiting spurious patterns that collapse under new conditions.[2]

This phenomenon—impressive demonstration performance masking fundamental brittleness—represents a classic failure mode in pattern recognition, epitomized by the horse "Clever Hans".[3] Clever Hans appeared to perform arithmetic calculations, fooling audiences and experts alike, until careful investigation revealed he was simply reading unconscious cues from his questioners.

Modern machine learning exhibits analogous vulnerabilities. Computer vision models exploit spurious correlations—skin color in medical diagnosis,[4] background textures in animal classification[5]—achieving high accuracy through irrelevant shortcuts.[6,7] Recent reports from Leash Biosciences suggest similar risks in chemical property prediction: models achieve surprising accuracy at predicting compound provenance, potentially using authorship as a proxy for bioactivity rather than learning meaningful chemistry.[8]

Materials science offers abundant opportunities for such proxy learning (Figure 1). Research groups develop specialized expertise—some focus on solar cell stability optimization, others on MOF synthesis strategies. Laboratory names often appear in framework designations (UiO-66, MIL-101), creating direct author-material associations. Fields evolve through paradigm shifts—self-assembled monolayers transforming perovskite cells, post-synthetic modifications revolutionizing MOF properties[9]—embedding temporal signatures that models might exploit rather than learning fundamental chemistry.

This work systematically investigates whether commonly used materials datasets are vulnerable to such proxy learning. I augment established property prediction datasets with bibliographic metadata—author names, publication years, journal venues—then test whether models can predict this information from chemical descriptors alone. The results are sobering: models predict metadata with surprising accuracy, and models using only these predicted "bibliographic fingerprints" often match the performance of conventional structure-property approaches.

These findings expose how easily we can be misled about what our models actually learn. More importantly, they point toward concrete strategies for building more robust datasets and developing validation frameworks that actively test competing hypotheses about model performance.

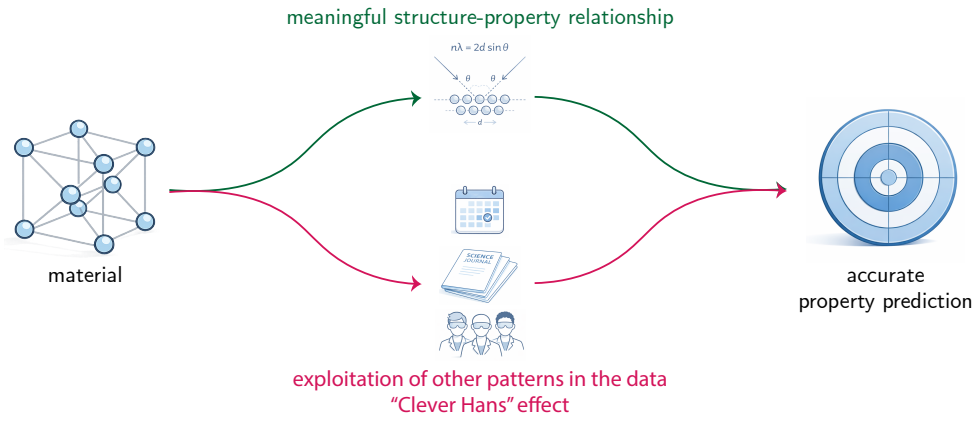

Figure 1: In machine learning, we often do not test competing hypotheses of how a model might obtain its answers. Machine learning models in materials science are trained to map material (descriptors) to property predictions. Models have much flexibility in how they learn this map from data. Ideally, 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 arrows). 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.

2 Results

I systematically assessed Clever Hans vulnerabilities across five materials domains: metal-organic frameworks, perovskite solar cells, battery materials, and organic electronics. Each case study tests whether models can achieve competitive property prediction performance using only bibliographic proxy signals rather than chemical understanding. Because proxy signal detectability depends critically on prediction task formulation and evaluation metrics, I analyze both regression and classification variants where appropriate, with additional results detailed in the appendix.

2.1 MOF Thermal Stability

Thermal stability represents a critical constraint for MOF applications across gas storage, catalysis, and separation processes.[10] Nandy *et al.*[11] systematically extracted decomposition temperatures from thermogravimetric analyses reported in the literature, creating opportunities to model thermal stability as either a continuous property (regression) or a discrete performance class (classification).[12]

Figure 2 demonstrates measurable proxy learning: structural descriptors predict bibliographic information significantly above baseline, enabling top-10% thermal stability classification with an accuracy of 0.901, approaching that of conventional structure-property models (0.926). However, Figure 9 shows that the measured effect size depends on the metric one uses to compare models.

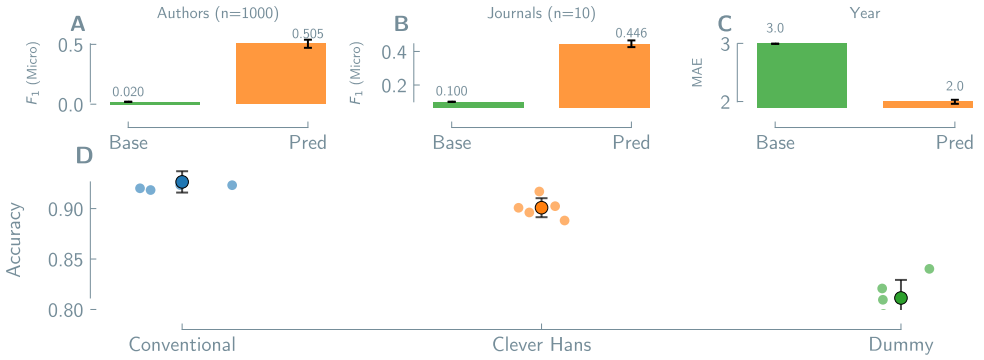

Figure 2: For the classification task of membership in the top-10% of thermally stable MOFs, one can be fooled (by Clever Hans effects). The model can predict the bibliographic information with high accuracy. **a** The model predicts the authors of the associated paper with high accuracy, much better 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, 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.

2.2 MOF Solvent Stability

Solvent removal stability poses an equally critical challenge for MOF deployment. Many frameworks collapse when synthesis solvents are removed during activation,

limiting their practical utility. Using the same text-mined dataset and descriptors,[11] I tested whether proxy signals could predict solvent stability outcomes.

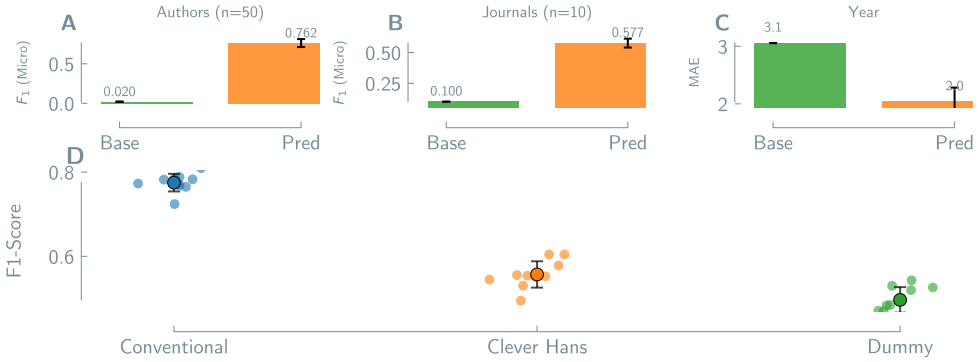

Figure 3: Measuring performance using accuracy, "Clever Hans" models can achieve a surprisingly good performance in predicting solvent removal stability of MOFs. a MOF descriptors can be used to predict author information better than a random baseline, but not with very high information. **b** The journal in which a given MOF stability result has been published can also be predicted with non-trivial accuracy. **c** The publication year can be predicted with a mean average error of two years. **d** A model trained on predicted bibliographic information can achieve non-trivial accuracy in correctly predicting the solvent stability of MOFs. Section A.1 shows the performance measured with other metrics.

Figure 3 shows that MOF structural descriptors predict bibliographic metadata with moderate accuracy. The proxy model achieves classification accuracy of 0.660, above baseline but below direct structure-property approaches, indicating partial Clever Hans susceptibility that varies across MOF stability properties.

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]

Figure 4 demonstrates that perovskite composition descriptors predict bibliographic information with measurable accuracy. The model achieves a micro-averaged F_1-score of 0.318 for the 10 most prolific authors, indicating detectable authorship signatures in the chemical data. Similar performance occurs for journal and publication year prediction.

Figure 4: Clever Hans models can achieve a high accuracy in predicting if a perovskite material is within the top-10% most efficient absorbers. a Composition description can predict which of the ten most prolific authors has been reporting a composition with a micro F_1 score of 0.318. **b** The journal in which a device has been published can be predicted with a similar performance. **c** The publication year can be predicted meaningfully better than with a mean baseline. **d** A model trained on this predicted bibliometric information can achieve an accuracy in classifying if an aborber belongs to the top-10% most efficient ones with an accuracy indistinguishable from the model directly trained on composition features.

The proxy model achieves classification accuracy of 0.900, comparable to direct composition-property models (0.899) for identifying top-10% efficient devices. This suggests potential reliance on author expertise patterns or temporal trends rather than composition-efficiency relationships.

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 battery composition descriptors exhibit limited proxy learning capability. Author prediction achieves an F_1-score of 0.165, while publication year prediction performs moderately above baseline. The resulting proxy model does not perform better than simply always predicting the mean.

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 wave-

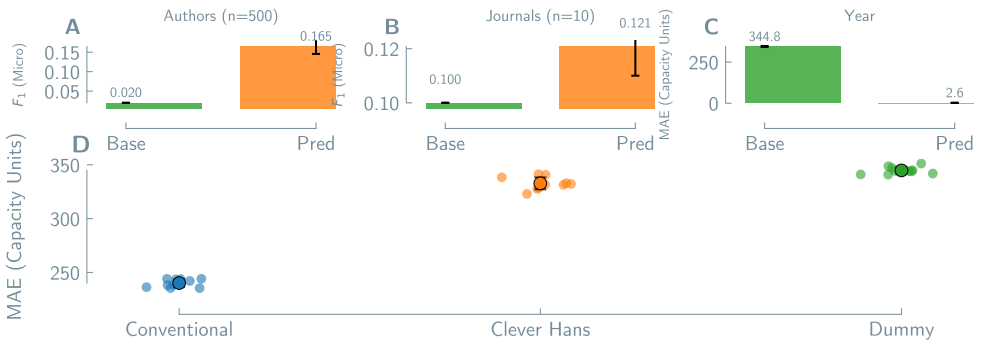

Figure 5: "Clever Hans" models do not achieve a good performance in predicting battery capacity. **a** Composition descriptors can be used to determine which of the 500 most prolific authors reported this material with a F_1 score of 0.165. **b** The journal in which a material has been reported can be predicted with a similar performance. **c** The publication year can be predicted better than with the mean baseline. **d** "Clever Hans" models are not distinguishable from mean predictions for the prediction of the capacity of battery materials.

length is one important performance metric that is optimized for these materials. Huang & Cole[18] text-mined using ChemDataExtractor.[19,20]

Molecular descriptors and fingerprints serve as features for both conventional and proxy models predicting maximum emission wavelength.

Figure 6 demonstrates that TADF molecular descriptors enable moderate bibliographic prediction. The proxy model achieves an MAE for maximum emission wavelength prediction between conventional and baseline approaches, indicating limited but detectable Clever Hans effects.

2.6 Overall Effects

Clever Hans effects vary significantly across material domains and prediction tasks. Proxy models achieved competitive performance in perovskite efficiency classification and MOF thermal stability, moderate effects in TADF wavelength prediction and MOF solvent stability, and negligible effects in battery capacity prediction. Effect detectability depends critically on evaluation metrics and baseline selection.

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

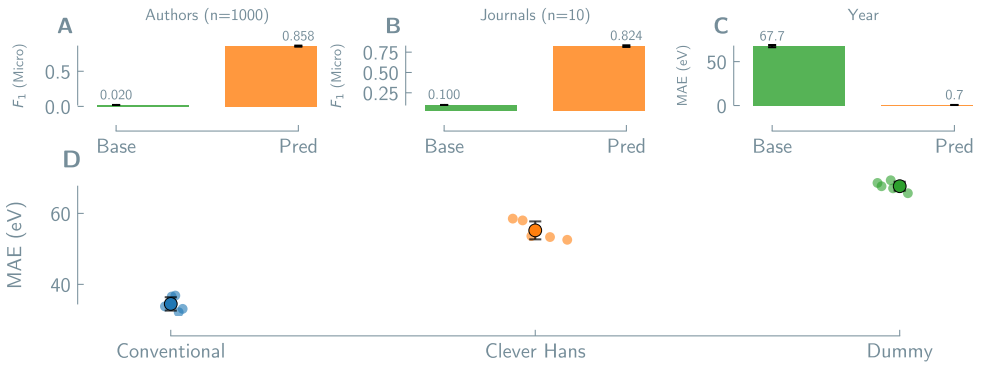

Figure 6: "Clever Hans" models perform worse than "conventional" models but better than simple baselines in predicting the maximum emission wavelengths of TADF molecules. **a** Using composition descriptors, the model can predict the the authors of the paper describing a material with an accuracy of 0.685 among the 1000 most profilic authors. **b** The model achieves an even higher performance in predicting in which of the ten most common journals a given entry has been published. **c** The publication year, too, can be predicted with a high performance just based on composition descriptors. **d** Using predicted bibliometric information, the authors achieve a mean absolute error between the "Conventional" and naïve baseline models.

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.

However, intellectual honesty demands acknowledging the heterogeneity of these effects. The Clever Hans phenomenon does not manifest uniformly across all metrics or datasets—in some cases, proxy models performed worse than random baselines on certain metrics while showing improvements on others. This variability suggests that the presence and magnitude of these effects depend on complex interactions between dataset construction, target properties, and evaluation approaches.

Nevertheless, the ability to predict bibliometric information—author identity, publication venue, and temporal patterns—with surprising accuracy across multiple domains reveals that datasets contain hidden signals we likely do not want influencing our models. Particularly telling are cases like TADF energy gap prediction and MOF solvent stability classification, where Clever Hans models achieved meaningfully better performance than baselines even in well-balanced datasets, suggesting these effects extend beyond simple artifacts of class imbalance.

Importantly, these results were obtained with minimal effort toward optimizing the proxy models, suggesting that more sophisticated exploitation of these signals could yield even stronger effects. While the magnitudes observed here may be smaller than those reported in other domains,[Geirhos2020, McCoy2019] they remain sufficient to warrant serious consideration of alternative hypotheses for model performance.

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.

Indeed, self-deception in models—learning proxy signals rather than intended mechanisms—is not always detrimental. If the underlying correlations remain stable and our goal is practical prediction rather than mechanistic insight, such models can provide substantial value. The key is transparency about what we are optimizing for and the limitations this imposes on model interpretability and generalizability. Acknowledging these trade-offs allows us to make informed decisions about when proxy signals are acceptable and when deeper understanding is essential.

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.

5.3 Cross-Validation Protocol

I used 10-fold cross-validation with random shuffling unless otherwise mentioned. Individual dots in swarm plots indicate the performance on the individual folds. For each of the 10 cross-validation folds, I maintain strict separation between training and testing phases: In the training phase, three models are trained simultaneously on the same training data: (1) the conventional model learns to map chemical descriptors directly to target properties, (2) the meta-prediction model learns to predict bibliographic information (authors, journals, publication years) from chemical descriptors, and (3) the proxy model learns to map predicted bibliographic information to target properties. In the testing phase, the conventional model predicts properties from descriptors on the held-out fold. The meta-prediction model generates bibliographic predictions for the test materials, and the proxy model uses the predicted bibliographic data (not ground-truth metadata) to predict properties. This protocol ensures that proxy models cannot access ground-truth bibliographic information during testing, only their own uncertain predictions. The comparison thus reflects realistic deployment scenarios where future materials would lack known authorship or publication context.

5.4 Datasets and Feature Engineering

5.4.1 Battery Dataset

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

The battery dataset contained 34669 entries with 273 chemical descriptors.

5.4.2 Perovskite Dataset

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

The perovskite dataset contained 4753 entries with 273 descriptors.

5.4.3 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]

The MOF thermal stability dataset contained 3128 entries with 174 structural and chemical descriptors. The MOF solvent stability dataset contained 2173 entries with 174 descriptors. I used the MOF descriptors provided by Nandy *et al.*[11].

5.4.4 TADF Dataset

I obtained the TADF dataset from Huang & Cole[18]. The TADF dataset contained 2089 entries with 2265 molecular descriptors. As inputs for the models, I used compositional descriptors.

5.5 Chemical Descriptor Generation

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

5.5.1 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.

5.5.2 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.

5.5.3 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.

5.6 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).

5.7 Data Processing

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

Data and Code Availability

To ensure reproducibility, this manuscript was generated using the framework.[54] The code to rebuild the paper (including code for all figures and numbers next to which there is a GitHub icon) can be found at . To facilitate reproduction, some intermediate analysis results are cached at `http://dx.doi.org/10.5072/zenodo.34706`.

Acknowledgement

This work was supported by the Carl Zeiss Stiftung. The author is a 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. 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).
5. 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).
6. 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).
7. 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).

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).

54. Luger, R. *et al.* Mapping stellar surfaces III: An Efficient, Scalable, and Open-Source Doppler Imaging Model. *arXiv preprint arXiv:2110.06271* (2021).

A Detailed Results

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

A.1 MOF Solvent Removal Stability

Section A.1 shows the performance for MOF solvent removal stability classification measured with different metrics. Section A.1 shows the performance of "Clever Hans" models as a function of the type of bibliometric information information that is used as model input.

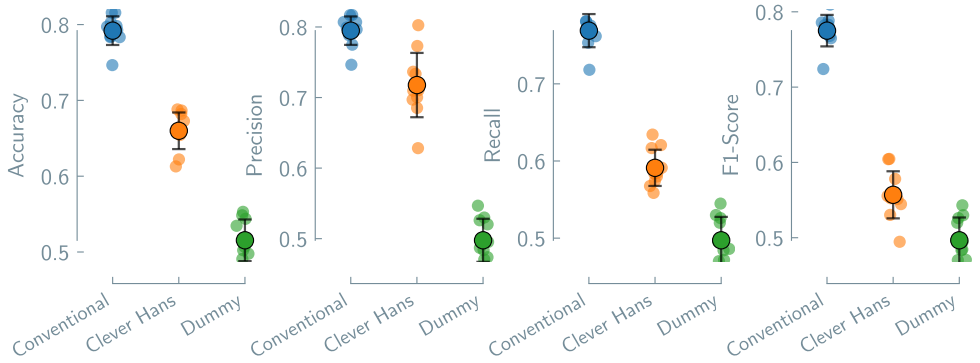

Figure 7: Performance of MOF solvent removal stability classification measured with different metrics. In all metrics, "Clever Hans" models outperform simple baselines. In some metrics, such as precision, "Clever Hans" models come close in performance to models directly trained on MOF descriptors ("Conventional").

Figure 8: Performance of "Clever Hans" models as a function of the type of predicted bibliometric information used in their feature set.

A.2 MOF Thermal Stability

Figure 9 shows that the measured difference in performance between models highly depends on the chosen metric.

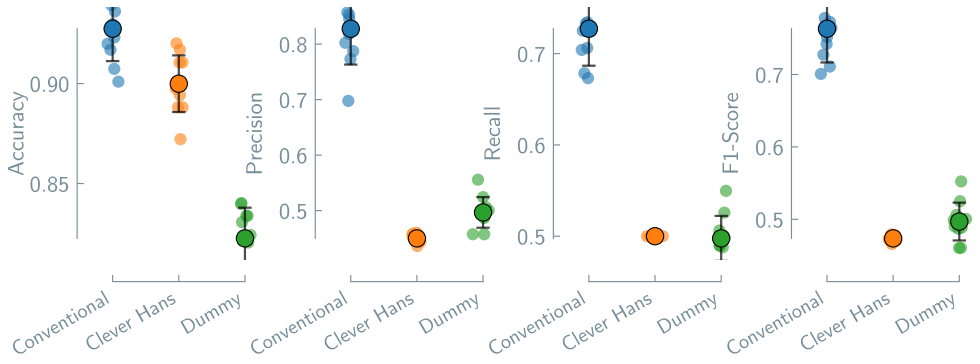

Figure 9: MOF stability prediction performance under different metrics. The measured difference between models highly depends on the chosen metric.

A.3 Perovskite Regression

Figure 10 shows the "Clever Hans" behavior in the regression case. Figure 11 shows the predictive performance as a function of the number of bibliometric features.

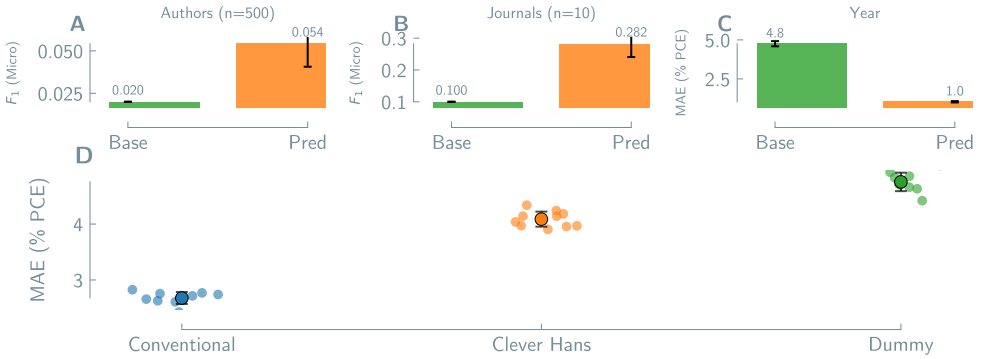

Figure 10: Performance in predicting bibliometric information and in predicting photoconversion efficiencies.

Figure 11: Impact of the number of bibliometric features on the performance of "Clever Hans" models in predicting PCE.