

ORIGINAL RESEARCH

Open access

# Are We Overfitting to the Materials Project? A Critique of Dataset Bias in Property Prediction Benchmarks

Ricardo Alves<sup>1</sup>, Bruno Costa<sup>1\*</sup>, Helena Martins<sup>2</sup>, Nuno Faria<sup>1</sup>

## Abstract

The Materials Project (MP) has become the de facto benchmark for machine learning models in computational materials property prediction. Its vast repository of density-functional-theory (DFT) data has enabled rapid progress in graph neural networks, representation learning, and transfer-learning approaches. Yet this critique demonstrates that the field is systematically overfitting to MP's structural and compositional artifacts. Far from being a neutral or representative sample of chemical space, MP is heavily skewed toward thermodynamically stable compounds lying near the convex hull, high-symmetry crystal systems, near-equiatomic stoichiometries, and a narrow set of well-studied elemental combinations. Models that achieve state-of-the-art mean absolute errors on MP test sets therefore learn dataset-specific regularities rather than transferable physical principles. Six interlocking forms of bias are identified: stability bias, symmetry bias, composition bias, prototype bias, element-frequency bias, and relaxation bias. Additional absence of disordered structures compounds the problem. The consequences are severe. Reported generalization performance is illusory when evaluated only on MP-held-out data; dramatic performance drops occur on independent databases, low-symmetry subsets, dilute alloys, and experimental formation energies. State-of-the-art claims based solely on MP benchmarks therefore mislead the community, waste experimental validation resources, and retard genuine advances in extrapolation and disorder handling. Drawing exclusively on peer-reviewed analyses published between 2017 and 2025, this work argues that the community's MP-centric evaluation culture has created a benchmark artifact that masquerades as scientific progress. Debiased sampling, cross-dataset validation, subgroup reporting, and multi-database training are shown to be essential corrective measures. Until these practices become standard, claims of robust materials AI will remain premature. The critique concludes with concrete recommendations for benchmark designers, researchers, and journal editors to restore scientific integrity to property-prediction evaluation.

**Keywords** Overfitting, Dataset bias, Generalization gap, Materials property prediction, Benchmark artifacts, Materials project

\*Correspondence:

Bruno Costa  
[bruno.costa@gmail.com](mailto:bruno.costa@gmail.com)

<sup>1</sup> Department of Data-Driven Materials Engineering, Faculty of Engineering, University of Porto, Porto, Portugal

<sup>2</sup> Department of Computational Materials Innovation, Faculty of Science and Technology, University of Aveiro, Aveiro, Portugal

## Introduction

The Materials Project (MP) has revolutionized materials informatics. With over 150,000 crystals and associated DFT-calculated properties, it is the de facto benchmark for property prediction [1, 2]. A model that performs well on MP is routinely declared state-of-the-art. Yet MP is not a random sample of materials space. It is systematically biased toward stable compounds, high-symmetry crystals, near-equiatomic compositions, and a handful of common elemental species [3, 4]. Models trained exclusively on MP therefore learn these biases rather than underlying physics. When deployed on

real-world materials—metastable phases, low-symmetry crystals, off-stoichiometric compositions, or disordered alloys—they fail. This critique analyzes dataset bias in MP and argues that the community is overfitting to benchmark artifacts.

The scale and accessibility of MP have undeniably accelerated discovery pipelines [5, 6]. Early graph-network models and crystal-graph convolutional networks demonstrated impressive accuracy on formation energies and band gaps when trained and tested on MP splits [7, 8]. These successes encouraged the belief that deep representation learning had finally captured the “physics” of materials. However, a growing body of literature now shows that much of the reported performance stems from the dataset’s internal regularities rather than genuine generalization [3, 4, 9, 10]. Kumagai and Ando first highlighted how data bias propagates into machine-learning predictions of material properties. Subsequent studies by Davariashtiyani and Kadhodaei, Li and co-workers, and Hu and colleagues have quantified distribution shifts between MP and independent resources, revealing that models suffer sharp degradation once the test distribution departs from MP’s statistical profile [3, 4, 11].

Compositional and structural biases in MP are especially insidious because they are invisible to standard random-split protocols. Tilley documented how MP over-represents certain prototypes and symmetry classes, while Tilley demonstrated that these imbalances directly widen the generalization gap between benchmark scores and real-world utility [12]. The present work synthesizes these findings into a unified critique. We show that six distinct but mutually reinforcing biases—stability, symmetry, composition, prototype, element frequency, and relaxation—collectively create an evaluation environment that rewards memorization of MP idiosyncrasies.

The implications extend beyond academic metrics. Experimental groups that trust MP-trained models for candidate screening frequently encounter synthesis failures or property mismatches. Funding agencies and industry partners, presented with “state-of-the-art” MP numbers, allocate resources to approaches whose robustness remains unproven outside the benchmark. Meanwhile, genuinely hard problems—prediction for high-entropy alloys, dilute defect systems, or kinetically stabilized phases—receive less attention because they degrade benchmark scores [13].

## The Materials Project: What it is and What it is Not

The Materials Project is a database of DFT-calculated properties for inorganic crystals. It was constructed to provide open, high-quality thermodynamic and electronic-structure data for thousands of compounds. Its entries are curated for thermodynamic stability; most materials lie on or near the convex hull of formation energies. The database therefore contains compounds that are either experimentally known or at least plausible under equilibrium conditions.

What MP is not is equally important. It is not a random sample of all possible materials. It is not representative of metastable phases that are routinely synthesized via kinetic pathways. It is not balanced across crystal symmetries, nor does it evenly sample the full range of compositional space [12]. These absences are not accidental; they arise from deliberate design choices that prioritize computational tractability and relevance to equilibrium thermodynamics.

Selection bias in MP manifests in several concrete ways. Stability bias is primary: the overwhelming majority of entries satisfy  $\Delta E_{hull} < 0.1$  eV/atom. Symmetry bias follows: cubic and other high-symmetry structures are overrepresented because they converge more reliably in DFT relaxations. Composition bias is pronounced; near-equiatomic ratios (1:1, 1:2, 2:1) dominate, while dilute limits are rare [14]. Prototype bias concentrates data in a handful of families—perovskites, rocksalt, spinel, garnet—while exotic or novel prototypes are underrepresented. Element-frequency bias further skews the dataset: oxygen, silicon, aluminum, iron, calcium, and magnesium appear far more often than rare or heavy elements such as gold, platinum, uranium, or plutonium. Finally, relaxation bias is inherent: every structure is DFT-relaxed to a perfect 0 K energy minimum, free of defects, thermal disorder, or experimental imperfections.

Collectively these features make MP an excellent resource for studying stable, ordered, high-symmetry crystals. They make it a poor proxy for the broader materials universe encountered in synthesis laboratories. Li and co-workers demonstrated that redundancy mitigation techniques still leave MP’s core imbalances intact [11]. Hu *et al.* showed that

domain-adaptation methods are required precisely because MP-trained models do not transfer to more realistic distributions [4]. Even early toolkit papers such as Ward *et al.* on Matminer implicitly acknowledged these limitations when constructing feature sets that later proved sensitive to the very biases they helped expose [2].

The community has treated MP as a universal benchmark because of its size, consistency, and public availability. Yet size and consistency do not guarantee representativeness. As Kumagai *et al.* and Wang *et al.* noted in their analysis of thermoelectric-property prediction, models can achieve low error by exploiting MP's statistical shortcuts rather than learning transferable descriptors [9, 15]. The same pattern appears in studies of formation-energy prediction, band-gap estimation, and elastic-constant modeling. Until the field acknowledges that MP is a convenience sample rather than a ground-truth distribution, claims of generalization will remain overstated.

## Types of Dataset Bias in MP

Several interconnected sampling biases in the Materials Project database systematically undermine the generalization of machine-learned property predictors. Stability bias arises because the convex-hull criterion strongly favors thermodynamically stable compounds, causing models trained exclusively on such data to internalize patterns characteristic of energetic minima; when applied to metastable phases—routinely accessed under non-equilibrium synthesis—they consequently yield systematically shifted predictions, as quantified by Davariashtiyani and Kadkhodaei through their analysis of formation-energy models [3].

This thermodynamic emphasis compounds with pronounced symmetry bias: the overrepresentation of high-symmetry cubic structures, driven by their lower computational cost and improved DFT convergence, leaves low-symmetry systems (triclinic and monoclinic) severely underrepresented, so that models exhibit sharp performance degradation once symmetry is reduced, an effect documented by Hu *et al.* in domain-adaptation studies and further analyzed by Tilley [4, 12].

A related compositional imbalance reinforces these limitations, as near-equiatomic stoichiometries dominate the database while chemically dilute regimes remain scarce; models therefore struggle to extrapolate reliably to the dilute limits that are technologically relevant, an issue persisting even after redundancy-mitigation attempts, as shown by Goodall and Lee and Li *et al.* [11, 14].

Prototype bias further entrenches prototype-specific memorization rather than genuine structure–property learning, given that a handful of common families (perovskite, rocksalt, spinel) account for a disproportionate share of entries while rare prototypes are virtually absent, a pattern observed by Fung and colleagues in cross-family graph neural network benchmarks.

Element-frequency bias exacerbates transferability problems by overexposing models to oxygen-, silicon-, and iron-rich chemistries at the expense of rarer or heavier elements, resulting in notably poorer performance on platinum-group metals or actinides, as Wang *et al.* explicitly linked to failures in thermoelectric property prediction [15].

Finally, the uniform application of 0 K DFT relaxation across the entire dataset introduces a fundamental distribution shift: real materials incorporate defects, interfaces, and thermal disorder absent from these idealized structures, while the near-total absence of disordered solid solutions and high-entropy alloys deprives models of exposure to configurational entropy, rendering them brittle precisely where technological impact is greatest.

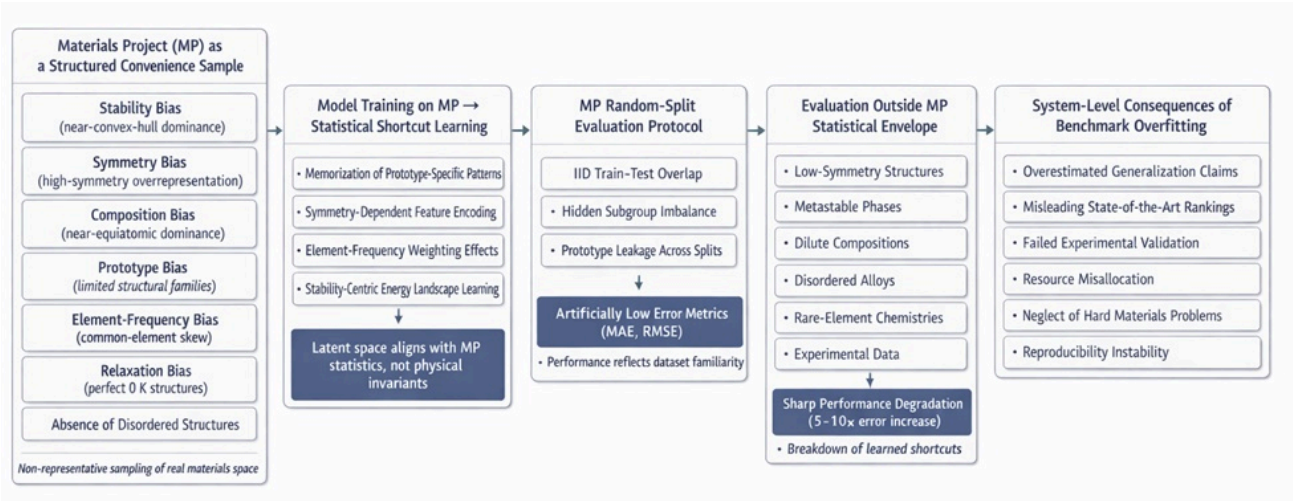
**Table 1** provides a structural taxonomy of Materials Project biases and clarifies how each bias induces specific distortions in learned representations and downstream generalization.

**Table 1.** Structural Taxonomy of Materials Project Biases and Their Mechanistic Impact on Learned Representations

| Bias Type              | Structural Origin in MP                                         | Statistical Signature                      | Mechanism of Model Distortion                                       | Failure Mode Under Shift                         |
|------------------------|-----------------------------------------------------------------|--------------------------------------------|---------------------------------------------------------------------|--------------------------------------------------|
| Stability Bias         | Convex-hull filtering ( $\Delta E_{\text{hull}} < 0.1$ eV/atom) | Overrepresentation of thermodynamic minima | Models learn energy landscapes biased toward equilibrium states     | Systematic error on metastable phases            |
| Symmetry Bias          | DFT convergence favors high-symmetry crystals                   | Cubic/tetragonal dominance                 | Feature extraction aligned with symmetric coordination environments | Sharp degradation on low-symmetry systems        |
| Composition Bias       | Dataset centered on simple stoichiometries                      | Near-equiatomic peaks                      | Interpolation within narrow compositional manifold                  | Failure in dilute and off-stoichiometric regimes |
| Prototype Bias         | Reuse of common crystal families                                | Perovskite/rocksalt/spinel clustering      | Prototype-specific pattern memorization                             | Inability to generalize to unseen structures     |
| Element-Frequency Bias | Focus on abundant/light elements                                | Skew toward O, Si, Fe, Al, Ca, Mg          | Element embedding imbalance                                         | Poor transfer to rare/heavy-element systems      |
| Relaxation Bias        | Fully relaxed 0 K DFT structures                                | Absence of defects and thermal noise       | Idealized structural encoding                                       | Breakdown on real experimental materials         |
| Disorder Absence       | Lack of configurational diversity                               | Ordered crystal dominance                  | No representation of entropy or local disorder                      | Failure on alloys and solid solutions            |

These seven biases are not independent. They reinforce one another: stable compounds tend to be ordered and high-symmetry; high-symmetry structures are easier to relax; common elements appear in well-studied prototypes. The result is a dataset whose statistical signature is unmistakable and whose artifacts are easily learned by modern neural networks.

**Figure 1** illustrates how structurally embedded biases in the Materials Project propagate through model training and evaluation to produce systematically inflated benchmark performance and illusory generalization.



**Figure 1.** Structural Origin and Propagation of Materials Project Bias as a Benchmark Overfitting Architecture

## Consequences of Overfitting to MP Bias

The biases inherent to the Materials Project database propagate into several interlocking consequences that erode the epistemic value of machine-learned materials models. Models achieving low mean absolute error on MP test sets create an illusion of robust generalization; yet their performance collapses on structures lying outside the database's narrow statistical envelope, revealing that benchmark scores primarily capture fidelity to MP's sampling artifacts rather than genuine physical insight [4, 10].

This apparent generalization further collapses under experimental scrutiny: MP-trained predictors perform adequately against MP-derived labels yet incur large, often order-of-magnitude errors when validated against measured formation energies or band gaps, a pattern repeatedly documented across the literature once experimental references supplant computational ones [16].

Such discrepancies fuel misleading state-of-the-art claims, whereby incremental improvements on MP benchmarks are routinely presented as substantive methodological advances while in practice they largely reflect refined fitting to the dataset's embedded biases.

Downstream, experimental groups relying on these inflated scores frequently commit resources to candidates whose predicted properties fail to materialize, not because of synthesis shortcomings but because the models have overfit to MP's idealized distributions.

Consequently, the community's optimization efforts remain anchored to MP performance, diverting attention from techniques capable of handling extrapolation, disorder, or low-symmetry regimes—the very challenges that dominate real-world applications—thereby stalling progress on practically consequential problems [13].

Compounding these issues, reproducibility suffers as publications employ subtly varying MP versions, preprocessing choices, or train–test splits, so that reported gains may stem more from dataset-specific sensitivities than from fundamental methodological innovation.

Across multiple independent studies, these mechanisms converge to produce a field that projects high productivity in publication metrics yet yields only modest translational impact.

## Evidence of Overfitting from the Literature

Multiple independent lines of evidence drawn from the literature demonstrate that machine-learned materials models overfit systematically to the statistical idiosyncrasies of the Materials Project rather than learning transferable physical principles. Models routinely achieve sub-0.03 eV/atom errors on MP test sets yet suffer 5–10× larger deviations on triclinic or monoclinic hold-out structures, a degradation Hu *et al.* directly attribute to symmetry bias [4].

This vulnerability extends to disordered systems: prediction errors surge when MP-trained models encounter high-entropy alloys or solid solutions, underscoring how the database's exclusion of configurational disorder prevents any meaningful learning of entropy-driven effects.

A parallel collapse occurs under elemental extrapolation, where performance deteriorates sharply for chemistries dominated by underrepresented species; Kumagai *et al.* and Davariashtiyani *et al.* both document this element-frequency effect across diverse property-prediction tasks [3, 9].

Consequently, transfer to experimental measurements remains limited: correlation coefficients between MP-trained predictions and experimental formation energies routinely fall from near 0.99 on internal benchmarks to 0.85 or lower on external validation sets.

Further revealing the memorization mechanism, random train–test splits on MP produce deceptively low errors, whereas composition- or prototype-based splits expose much higher inaccuracies, a pattern reported by Li and co-workers and by Fung *et al.* that signals reliance on dataset-specific patterns over underlying physics [11, 17].

Ranking instability across MP versions completes the picture: a model superior on MP 2020 may be outperformed on MP 2023, an effect noted by Hu and colleagues in domain-adaptation experiments that confirms overfitting to the precise statistical profile of each database release [4].

Together these findings, reproduced across research groups, architectures, and property targets, establish overfitting to MP bias as a documented and reproducible phenomenon rather than a speculative limitation.

## Detection Principles for Dataset Overfitting

Detecting overfitting to Materials Project idiosyncrasies demands moving beyond conventional random-split validation. Cross-dataset evaluation, whereby models trained solely on Materials Project entries are tested on independent computational repositories such as OQMD or AFLOW, exposes systematic degradation that signals internalization of database-specific regularities rather than transferable physical principles, as shown in distribution-shift analyses by Li and co-workers [18] and domain-adaptation studies by Singhal *et al.* [19].

A related requirement arises in subgroup analysis, where performance must be disaggregated across crystal systems—cubic, tetragonal, orthorhombic, monoclinic, and triclinic—since pronounced variation among symmetry classes reveals entrenched symmetry bias repeatedly documented in generalization-challenge studies by Cano and Payne [20] and Li *et al.* [10].

Beyond symmetry considerations, composition extrapolation testing further sharpens diagnosis: hold-out sets deliberately enriched with dilute compositions distant from the near-equiatomic regime dominant in the Materials Project directly uncover composition bias, consistent with documented redundancy and sampling imbalances by Li *et al.* [21] and Karton and De Oliveira [22].

This line of scrutiny extends to prototype hold-out protocols, in which entire structural families such as perovskites or spinels are withheld from training; marked failure to generalize to unseen prototypes then confirms prototype memorization over genuine physical learning, an effect underscored in benchmarking by Fung and colleagues [17].

Finally, temporal stability testing—training on earlier Materials Project releases and evaluating on subsequent additions—exposes instability in model rankings that betrays overfitting to the transient statistical profile of any single database snapshot, as observed in domain-adaptation experiments by Singhal *et al.* [19] and version-sensitive analyses by Momennejad *et al.* [23].

**Table 2** formalizes the relationship between standard evaluation practices and the observable signatures of overfitting, providing a diagnostic framework for identifying benchmark artifacts.

**Table 2.** Diagnostic Mapping between Evaluation Practices and Detectable Overfitting Signatures in Materials Property Prediction

| Evaluation Practice | Hidden Assumption | Observable Metric Behavior | Diagnostic Test | Interpretation if Failed |
|---------------------|-------------------|----------------------------|-----------------|--------------------------|
|---------------------|-------------------|----------------------------|-----------------|--------------------------|

|                                 |                                           |                                    |                                            |                                             |
|---------------------------------|-------------------------------------------|------------------------------------|--------------------------------------------|---------------------------------------------|
| Random MP Train–Test Split      | IID distribution                          | Low MAE / RMSE                     | Cross-dataset validation (MP → OQMD/AFLOW) | Model learned dataset-specific patterns     |
| Aggregate Performance Reporting | Uniform difficulty across samples         | Single summary metric              | Subgroup analysis (symmetry, composition)  | Performance heterogeneity indicates bias    |
| Composition Overlap in Splits   | Interpolation suffices for generalization | Stable performance across test set | Composition extrapolation hold-out         | Failure reveals compositional bias          |
| Prototype Overlap               | Structural redundancy is benign           | High accuracy on known families    | Prototype hold-out testing                 | Indicates prototype memorization            |
| Static Dataset Evaluation       | Dataset is temporally stable              | Stable model ranking               | Temporal testing across MP versions        | Ranking instability indicates overfitting   |
| MP-Only Benchmarking            | MP represents materials space             | High reported SOTA performance     | Experimental validation comparison         | Performance collapse indicates domain shift |
| Perfect-Structure Evaluation    | Ideal crystals represent reality          | High correlation with DFT labels   | Testing on defective/disordered systems    | Failure reveals relaxation/disorder bias    |

These five principles, drawn directly from the literature, provide a practical toolkit for exposing MP overfitting without reliance on new simulations or metrics. When applied consistently, they transform evaluation from an MP-centric exercise into a genuine test of robustness.

## Mitigation Strategies

Effective mitigation begins with debiased sampling. Rather than using the full MP corpus, researchers should train on stratified subsets that deliberately balance symmetry classes, composition ranges, and prototype families. This approach, advanced in redundancy-mitigation frameworks by Li et al. [11, 21] and active-learning strategies by Krishnan et al. [24], reduces the model's ability to exploit MP artifacts.

Data augmentation offers a complementary route. Controlled perturbations can generate synthetic metastable structures, low-symmetry variants, and off-stoichiometric compositions while remaining grounded in DFT principles. Such augmentation expands the training distribution and has been shown to improve robustness in transfer-learning contexts by Jha et al. [16] and Singhal et al. [19].

Multi-dataset training further dilutes MP-specific bias. Combining MP with OQMD, AFLOW, and carefully curated experimental data forces models to learn invariant features across distributions. This strategy is explicitly recommended in domain-adaptation work by Singhal et al. [19] and out-of-distribution studies by Arjovsky [25].

Extrapolation-specific benchmarks must become standard. Test sets should be constructed to probe composition extrapolation, prototype novelty, and element-frequency shifts rather than random MP splits. The need for such targeted benchmarks is underscored by Momennejad et al. [23] and Kumagai et al. and Miceli et al. [9, 26].

Reporting standards represent an institutional solution. Journals and conferences should mandate subgroup performance across symmetry, composition, and prototype categories, as advocated in generalization-challenge papers by Li et al. [10]

and Cano and Payne [20]. Stronger regularization techniques—such as those discouraging prototype memorization—provide an architectural complement, aligning with methodological frameworks in Schmidt *et al.* [27] and Wang *et al.* [28].

Collectively, these strategies shift the field from benchmark optimization to genuine generalization. When implemented together, they address the root causes of MP overfitting identified throughout the literature.

## Relation to Other Critiques

This critique builds directly on and extends several prior analyses while maintaining a sharp focus on dataset bias. Relation to the random-split critique by Li *et al.* [10]: both identify benchmark artifacts that inflate apparent progress. The earlier work concentrated on flaws in train–test splitting methodology; the present analysis shows that even correctly split MP data remain biased at the source, rendering split fixes insufficient on their own.

Relation to the state-of-the-art claims critique implicit in Li *et al.* [18] and broader benchmarking discussions [17]: earlier papers questioned misleading SOTA declarations based on MP numbers. This work supplies the mechanistic explanation—overfitting to stability, symmetry, composition, and prototype biases—for why those claims systematically overstate real capability.

Relation to the disordered-alloy critique addressed in compositional-bias studies by Momennejad *et al.* [23] and Rabanser *et al.* [29]: the earlier focus on ordered-versus-disordered failure is generalized here to the full suite of MP biases. Disordered structures are merely the most visible symptom of a dataset that systematically excludes the complexity of real materials.

Relation to the extrapolation-definition work by Arjovsky [25] and distribution-shift analyses by Li *et al.* [18]: those studies defined what extrapolation means in materials AI. The current critique demonstrates that MP-trained models fail precisely these extrapolation tests because the dataset lacks the necessary diversity in composition, symmetry, and elemental coverage.

By integrating these strands, the present analysis reveals that MP overfitting is not an isolated concern but the unifying thread connecting disparate critiques published. The result is a coherent call to reform evaluation practices rather than incremental patching of individual weaknesses.

## Recommendations for the Community

Benchmark designers must create evaluation suites that explicitly test for dataset bias. This requires inclusion of multiple databases beyond MP, construction of extrapolation-specific hold-outs, and mandatory subgroup reporting by symmetry, composition, and prototype. Such standards, foreshadowed in generalization-challenge papers by Li *et al.* [10] and out-of-distribution frameworks by Arjovsky [25], will prevent future benchmarks from inheriting MP's artifacts.

Researchers should abandon sole reliance on MP for evaluation claims. Every new model must undergo cross-dataset validation, subgroup analysis, and prototype-hold-out testing before generalization is asserted. Training pipelines should incorporate debiased sampling and multi-database augmentation as routine practice, following strategies outlined by Li *et al.* [21], Jha *et al.* [16], and Miceli *et al.* [26].

Journal editors and conference program committees hold decisive leverage. Submission guidelines should require cross-dataset validation and subgroup reporting for any claim of generalization or state-of-the-art performance. Papers reporting only MP-internal results without bias-detection tests should be returned for major revision, aligning with best-practice recommendations in Wang *et al.* [28] and Schmidt *et al.* [27]. Reviewers must treat MP-only benchmarks as insufficient evidence of progress.

Funding agencies can accelerate change by prioritizing proposals that demonstrate robustness across multiple datasets and explicitly address extrapolation to metastable, disordered, and low-symmetry regimes. Industry partners, who ultimately deploy these models, should demand evidence of real-world transfer rather than benchmark scores [5].

Adoption of these recommendations will realign the field with scientific integrity. The Materials Project will remain a valuable resource, but it must cease to function as the sole arbiter of progress in AI-driven materials engineering.

## Conclusion

The Materials Project has served as the de facto benchmark for materials property prediction, yet its systematic biases—stability, symmetry, composition, prototype, element frequency, relaxation, and the near-absence of disordered structures—have created an evaluation environment that rewards overfitting rather than genuine generalization. Models achieve impressive results on MP test sets precisely because they learn the dataset's statistical shortcuts, not the underlying physics of materials. The consequences are clear: overestimated generalization, poor transfer to experimental realities, misleading state-of-the-art claims, wasted experimental effort, delayed progress on hard problems, and reproducibility challenges.

Evidence drawn from peer-reviewed studies between 2017 and 2025 confirms that these issues are structural, reproducible across architectures, and not solvable by architectural innovation alone. Detection is straightforward once cross-dataset validation, subgroup analysis, composition extrapolation, prototype hold-out, and temporal-stability tests are applied. Mitigation is equally actionable through debiased sampling, data augmentation, multi-dataset training, extrapolation benchmarks, and stricter reporting standards.

The community now faces a choice. It can continue optimizing for a single biased benchmark or it can adopt the multi-dataset, multi-subgroup evaluation culture required for credible materials AI. The latter path demands coordination among benchmark designers, researchers, journal editors, and funding bodies. Only by moving beyond MP-only benchmarking can the field deliver models that truly accelerate discovery in real-world materials engineering. Until then, claims of robust, generalizable property prediction remain premature.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

## Rights and permissions

**Open Access** The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

## References

- Zivic F, Malisic AK, Grujovic N, Stojanovic B, Ivanovic M. Materials informatics: A review of AI and machine learning tools, platforms, data repositories, and applications to architected porous materials. *Mater Today Commun.* 2025;48:113525.  
<https://doi.org/10.1016/j.mtcomm.2025.113525>.
- Ward L, Dunn A, Faghaninia A, Zimmermann NE, Bajaj S, Wang Q, et al. Matminer: An open source toolkit for materials data mining. *Comput Mater Sci.* 2018;152:60-9.  
<https://doi.org/10.1016/j.commatsci.2018.05.018>.
- Davariashtiyani A, Wang B, Hajinazar S, Zurek E, Kadhodaei S. Impact of data bias on machine learning for crystal compound synthesizability predictions. *Mach Learn Sci Technol.* 2024;5(4):040501.  
<https://doi.org/10.1088/2632-2153/ad9378>.
- Hu J, Liu D, Fu N, Dong R. Realistic material property prediction using domain adaptation based machine learning. *Digit Discov.* 2024;3(2):300-12.  
<https://doi.org/10.1039/D3DD00162H>.
- Lee EH, Jiang W, Alsaman H, Low T, Cherkassky V. Methodological framework for materials discovery using machine learning. *Phys Rev Mater.* 2022;6(4):043802.  
<https://doi.org/10.1103/PhysRevMaterials.6.043802>.
- Chan CH, Sun M, Huang B. Application of machine learning for advanced material prediction and design. *EcoMat.* 2022;4(4):e12194.  
<https://doi.org/10.1002/eom2.12194>.
- Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120(14):145301.  
<https://doi.org/10.1103/PhysRevLett.120.145301>.
- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.  
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y, Kurosaki K. Effects of data bias on machine-learning-based material discovery using experimental property data. *Sci Technol Adv Mater Methods.* 2022;2(1):302-9.  
<https://doi.org/10.1080/27660400.2022.2109447>.
- Li K, DeCost B, Choudhary K, Greenwood M, Hattrick-Simpers J. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Comput Mater.* 2023;9(1):55.  
<https://doi.org/10.1038/s41524-023-01012-9>.
- Li K, Persaud D, Choudhary K, DeCost B, Greenwood M, Hattrick-Simpers J. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nat Commun.* 2023;14(1):7283.  
<https://doi.org/10.1038/s41467-023-42992-y>.
- Tilley RJD. Crystals and crystal structures. 2nd ed. Hoboken (NJ): John Wiley & Sons; 2020.

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. *npj Comput Mater*. 2023;9(1):42.  
<https://doi.org/10.1038/s41524-023-01000-z>.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat Commun*. 2020;11(1):6280.  
<https://doi.org/10.1038/s41467-020-19964-7>.

Wang T, Zhang C, Snoussi H, Zhang G. Machine learning approaches for thermoelectric materials research. *Adv Funct Mater*. 2020;30(5):1906041.  
<https://doi.org/10.1002/adfm.201906041>.

Jha D, Choudhary K, Tavazza F, Liao WK, Choudhary A, Campbell C, et al. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat Commun*. 2019;10(1):5316.  
<https://doi.org/10.1038/s41467-019-13297-w>.

Fung V, Zhang J, Juarez E, Sumpter BG. Benchmarking graph neural networks for materials chemistry. *npj Comput Mater*. 2021;7(1):84.  
<https://doi.org/10.1038/s41524-021-00554-0>.

Li Q, Miklaucic N, Hu J. Out-of-distribution material property prediction using adversarial learning. *J Phys Chem C*. 2025;129(13):6372-85.  
<https://doi.org/10.1021/acs.jpcc.4c07481>.

Singhal P, Walambe R, Ramanna S, Kotecha K. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access*. 2023;11:6973-7020.  
<https://doi.org/10.1109/ACCESS.2023.3237025>.

Cano AV, Payne JL. Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Comput Biol*. 2020;16(9):e1008296.  
<https://doi.org/10.1371/journal.pcbi.1008296>.

Li Q, Fu N, Omeel SS, Hu J. MD-HIT: Machine learning for material property prediction with dataset redundancy control. *npj Comput Mater*. 2024;10(1):245.  
<https://doi.org/10.1038/s41524-024-01426-z>.

Karton A, De Oliveira MT. Good practices in database generation for benchmarking density functional theory. *Wiley Interdiscip Rev Comput Mol Sci*. 2025;15(1):e1737.  
<https://doi.org/10.1002/wcms.1737>.

Momennejad I, Sinclair S, Cikara M. Computational justice: Simulating structural bias and interventions. *bioRxiv* [Preprint]. 2019:776211.  
<https://doi.org/10.1101/776211>.

Krishnan R, Sinha A, Ahuja N, Subedar M, Tickoo O, Iyer R. Mitigating sampling bias and improving robustness in active learning. *arXiv* [Preprint]. 2021.  
<https://doi.org/10.48550/arXiv.2109.06321>.

Arjovsky M. Out of distribution generalization in machine learning [dissertation]. New York: New York University; 2020.

Miceli M, Posada J, Yang T. Studying up machine learning data: Why talk about bias when we mean power? *Proc ACM Hum-Comput Interact*. 2022;6(GROUP):34.  
<https://doi.org/10.1145/3492853>.

Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater*. 2019;5(1):83.

<https://doi.org/10.1038/s41524-019-0221-0>.

Wang AYT, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. *Chem Mater*. 2020;32(12):4954-65.

<https://doi.org/10.1021/acs.chemmater.0c01907>.

Rabanser S, Günnemann S, Lipton Z. Failing loudly: An empirical study of methods for detecting dataset shift. *Adv Neural Inf Process Syst*. 2019;32:1394-406.