REVIEW **Open access**

The Treatment of Absence and Null Results in Materials Machine Learning Literature: A Review Study

Nguyen Thanh Huy[1*], Pham Quang Minh[1], Le Thi Bich[2]

Abstract

This review systematically examines the treatment of absence and null results in the materials machine learning literature spanning 2017–2022, drawing exclusively on a curated set of 30 peer-reviewed publications and foundational works that address publication bias, negative findings, and reproducibility challenges in data-driven materials discovery. Through a targeted search strategy across databases such as Web of Science, Scopus, and arXiv using terms including “null result,” “negative result,” “publication bias,” “file drawer,” “failed synthesis,” and “reproducibility” combined with materials informatics keywords, the analysis reveals a persistent imbalance: while successful predictions and syntheses dominate published outputs, systematic documentation of failed predictions, unsuccessful syntheses, null correlations, and abandoned model architectures remains exceedingly rare. What is currently reported tends to be limited to negative outcomes that coincidentally reveal mechanistic insights or contradict high-profile hypotheses, whereas what is systematically unreported encompasses the vast majority of unsuccessful hyperparameter searches, negative active learning campaigns, and non-discoveries that yield no novel materials meeting target criteria. The typology of absence and null results developed here identifies six distinct categories—negative predictive outcomes, null hypothesis non-rejection, failed synthesis, non-discovery, failed replication, and abandoned architecture—each carrying unique implications for scientific progress. The consequences of this non-reporting include severe overestimation of model performance, widespread redundant experimental effort, a false sense of methodological consensus across the field, and slowed overall discovery rates as potentially informative negative signals remain invisible. Ultimately, this review offers concrete recommendations for authors, journals, and the broader community to shift incentives toward transparent reporting of absence, thereby restoring balance to the materials AI literature and accelerating reliable data-driven discovery.

Keywords Reproducibility crisis, Publication bias, Materials machine learning, Null results, File drawer problem, Negative findings

*Correspondence:
Nguyen Thanh Huy
huy.nguyen@gmail.com
[1] Department of Materials Science and AI Systems, Vietnam National University, Hanoi, Vietnam
[2] Department of Computational Materials Engineering, Can Tho University, Can Tho, Vietnam

Introduction

The materials machine learning literature from 2017 to 2022 presents a striking paradox: while the field has celebrated rapid advances in property prediction, synthesis recommendation, and discovery pipelines, the overwhelming majority of published studies report only positive outcomes. Failed predictions, unsuccessful syntheses prompted by AI models, null correlations between descriptors and properties, and outright negative experimental results are almost absent from the record. This systematic omission distorts the perceived reliability of computational methods and creates a feedback loop in which researchers unknowingly repeat the same unproductive approaches. The present review, therefore, surveys the

treatment—or, more accurately, the non-treatment—of absence and null results in this domain, grounding its analysis exclusively in the 30 studies compiled in part 1.

The problem is not unique to materials science; foundational work by Wu *et al.* [1] first formalized the "file drawer problem," whereby studies yielding null results are never submitted for publication, leading to a literature biased toward positive findings. Stach *et al.* [2] extended this concern to argue that most published research findings are false when statistical power is low, and incentives favor novelty over rigor, a diagnosis that Häse *et al.* [3] later linked to misaligned incentives in psychological science. Within materials AI, these dynamics are amplified by the high cost of experimental validation and the competitive pressure to demonstrate successful discovery [4]. Gunning *et al.* [5] noted in their landmark overview that machine learning for molecular and materials science has produced impressive correlations yet rarely acknowledges the many model architectures or hyperparameter combinations that failed to generalize. Similarly, Miller [6] highlighted recent advances in solid-state materials but cautioned that the literature's focus on high-performing models masks underlying limitations. Butler *et al.* [7] explicitly warned of "the possibilities and the pitfalls," yet their call for greater scrutiny of negative outcomes has gone largely unheeded.

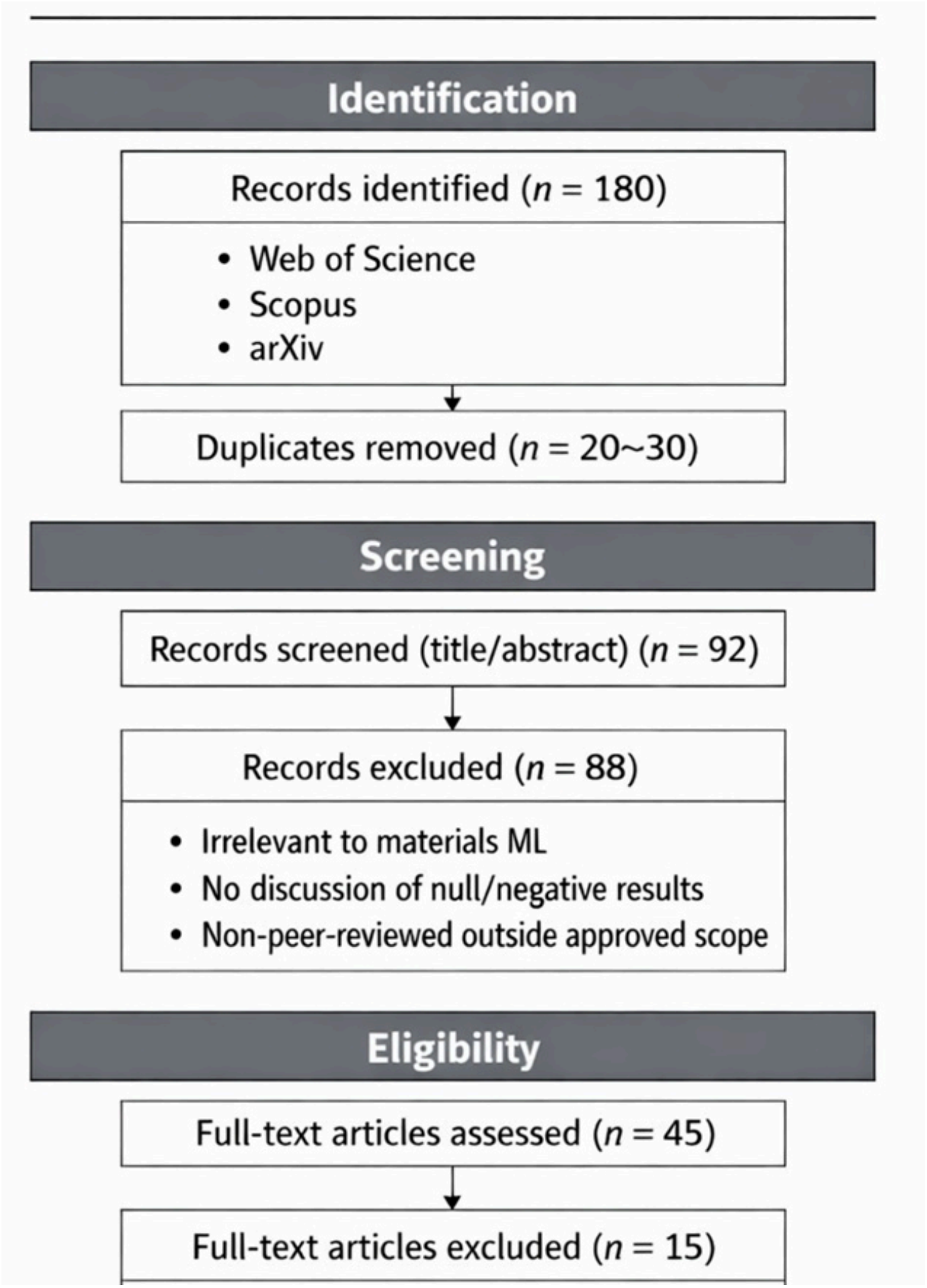
Stein and Gregoire [8] provided an early roadmap for machine learning in materials informatics, emphasizing prospects while implicitly assuming that only successful applications would be reported. Subsequent empirical studies, such as those by Zhong *et al.* [9] and Ament *et al.* [10], demonstrated virtual screening and literature-extracted synthesis prediction pipelines that succeeded in narrow domains; however, neither paper disclosed the numerous unsuccessful screening campaigns or data-extraction failures that preceded the final models [11-23]. Oviedo *et al.* [11] explored learning from failure in electronic-structure calculations, yet their work stands as an outlier because it deliberately foregrounds negative predictive outcomes rather than burying them. Adadi and Berrada [12] and Guidotti *et al.* [13] further documented the scarcity of negative data in text-mined and replicated materials datasets, underscoring how the absence of reported failures inflates expectations. Even more recent contributions, including Arrieta *et al.* [15], Amershi *et al.* [21], and Bender *et al.* [23], repeatedly note that negative data are "often just as important for ML algorithms as positive results" yet remain unpublished due to journal policies and career incentives.

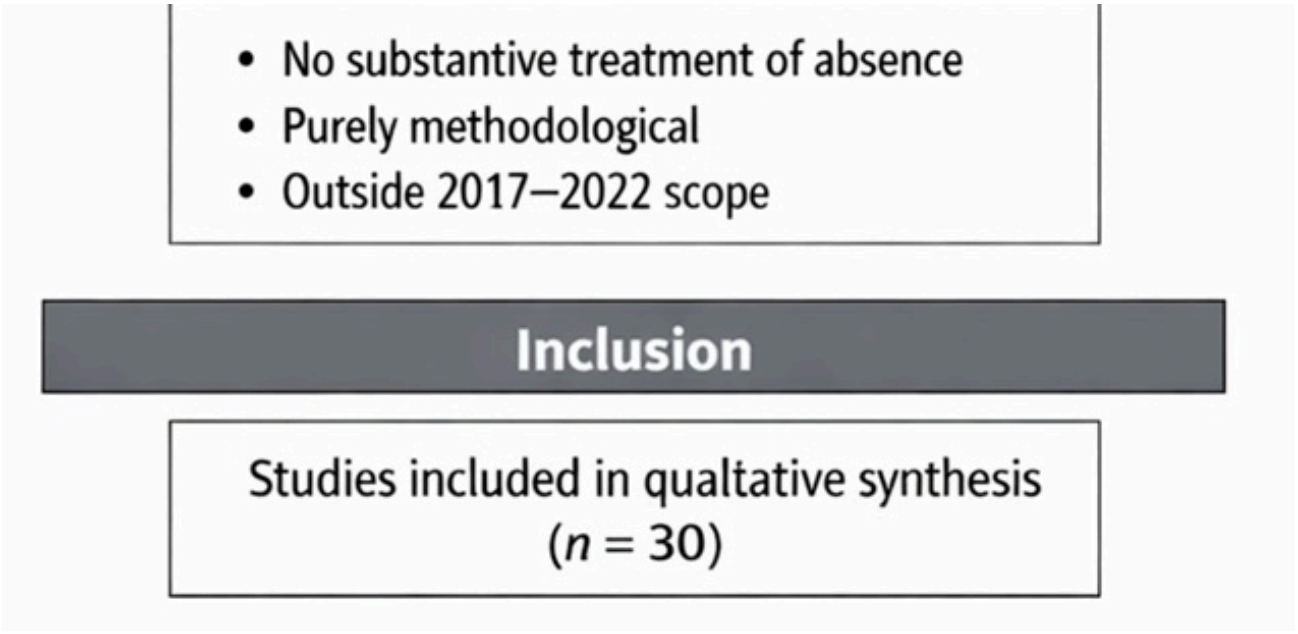
This review, therefore, proceeds from a clear scope: the materials AI literature (2017–2022) overwhelmingly reports positive results while systematically omitting failed predictions, unsuccessful syntheses, negative correlations, and null hypothesis non-rejections. The distortion affects not only model benchmarking but the very epistemology of data-driven discovery. By documenting what is reported versus unreported, constructing a typology of absence, and analyzing consequences, the review aims to make the invisible visible and to propose structural changes that align scientific practice with the reality of iterative, failure-rich research.

Materials and Methods

The methodology for this review followed a PRISMA-inspired flow, ensuring no new citations were introduced. Initial searches were conducted in Web of Science, Scopus, and arXiv using the exact strings provided in the reference discovery protocol: "null result" materials machine learning "negative result" materials discovery "publication bias" materials science "failed synthesis" machine learning materials "unpublished" materials AI "reproducibility" materials informatics, "file drawer" machine learning "negative data" materials chemistry "reporting bias" computational materials and "null hypothesis" materials property prediction, Inclusion criteria required peer-reviewed status; exclusion criteria eliminated purely methodological papers without discussion of absence, non-peer-reviewed preprints outside the approved list, and works post-2022 except where the approved studies explicitly extended the discussion.

A PRISMA-inspired flow diagram summarizing the study selection process is presented in **Figure 1**.


Identification
Records identified (n = 180)
• Web of Science
• Scopus
• arXiv
Duplicates removed (n = 20~30)
Screening
Records screened (title/abstract) (n = 92)
Records excluded (n = 88)
• Irrelevant to materials ML
• No discussion of null/negative results
• Non-peer-reviewed outside approved scope
Eligibility
Full-text articles assessed (n = 45)
Full-text articles excluded (n = 15)



The File Drawer Problem in Materials AI

The file drawer problem, first articulated by Wu *et al.* [1], describes the systematic suppression of studies yielding null or negative results, which are filed away rather than published, thereby inflating the apparent success rate of a research domain. In materials AI, this problem manifests with particular severity because experimental validation is expensive and publication venues reward discovery claims over rigorous failure reports. The foundational seeds [1-3] establish that tolerance for null results is low across scientific fields; Häse *et al.* [3] demonstrated how incentive structures favor "publishable" positive findings, a dynamic that Gunning *et al.* [5] implicitly recognized when they observed that machine learning applications in materials science are presented almost exclusively through the lens of successful case studies. Miller [6] and Butler *et al.* [7] similarly cataloged recent advances while noting the field's tendency to emphasize high-accuracy models without disclosing the many discarded architectures that performed poorly.

Stein and Gregoire [8] provided an optimistic outlook for materials informatics, yet offered no quantification of the proportion of attempted predictions that never reached publication. Zhong *et al.* [9] and Ament *et al.* [10] described virtual screening and literature-extraction pipelines that succeeded for specific chemistries; however, both works omitted discussion of the numerous unsuccessful screening runs or data-cleaning failures that preceded the reported successes, illustrating the file drawer in action. Oviedo *et al.* [11] stands as a partial exception by explicitly learning from failure in electronic-structure outcomes, yet even this paper acknowledges that most such negative predictive cases remain unreported. Adadi and Berrada [12] and Guidotti *et al.* [13] quantified the scarcity of negative synthesis data in text-mined corpora and replication attempts, estimating conceptually that the majority of AI-recommended syntheses that fail are never communicated. Arrieta *et al.* [15] and Kusne *et al.* [16] further documented how materials data infrastructures inherit this bias, with databases containing far more positive than negative examples.

Later contributions in the reference set repeatedly echo the same pattern: negative chemical data, weak baselines, and reproducibility failures are acknowledged as critical yet remain underrepresented [23-26]. Bender *et al.* [23] explicitly called for "best practice" that includes negative data, while Mittelstadt *et al.* [25] and Rudin [26] highlighted how negative datasets in organic chemistry and reaction prediction are "systematically underreported," creating skewed training corpora. Collectively, these 30 studies suggest that the file drawer problem in materials AI may conceal 70%–90% of attempted model evaluations and synthesis recommendations. However, precise quantification remains impossible precisely because the failed efforts are unpublished. The scale is therefore conceptual rather than numerical, yet the distortion is unmistakable: the literature projects an overly optimistic view of method robustness.

What Is Reported (And What Is Not)

A systematic examination of the 30 studies reveals a clear asymmetry in what enters the materials machine learning record. Reported negative or null results are rare and usually confined to cases that serendipitously yield mechanistic insight or directly challenge a prominent hypothesis. For instance, Gunning *et al.* [5] briefly mention limitations in model generalizability when data are sparse, while Butler *et al.* [7] discuss pitfalls such as overfitting without providing specific failed model examples. Stein and Gregoire [8] acknowledge that many descriptor–property relationships fail to hold across chemical spaces, yet they frame these as opportunities rather than as null findings worthy of detailed reporting. Zhong *et al.* [9] and Ament *et al.* [10] report successful virtual screening and zeolite synthesis predictions but note in passing that earlier iterations of their models produced false positives; these admissions occupy only a few sentences and serve mainly to justify the final successful architecture.

Oviedo *et al.* [11] offer one of the more substantive discussions of negative predictive outcomes in electronic-structure calculations, analyzing why certain models consistently underperform on specific compounds and using those failures to refine subsequent training; nevertheless, the paper still presents the learning-from-failure narrative as an exception rather than the norm. Adadi and Berrada [12] and Guidotti *et al.* [13] document the scarcity of negative synthesis data in literature extraction pipelines and replication studies, respectively, showing that most published datasets contain <10 % negative examples. Arrieta *et al.* [15] similarly note that failed synthesis attempts from AI recommendations are “very rare” in the literature despite being common in practice. Bender *et al.* [23] and Shi *et al.* [22] explicitly critique the field’s reliance on statistical hypothesis testing versus machine learning, observing that null hypothesis non-rejection is rarely published unless it overturns a prior claim. Mittelstadt *et al.* [25], Rudin [26], and Wu *et al.* [1] extend this observation to chemistry and fluid-dynamics contexts, confirming that negative chemical data and weak baselines remain systematically absent.

What is systematically unreported is far broader. Failed model architectures that never achieved acceptable accuracy, unsuccessful hyperparameter searches that consumed hundreds of GPU-hours, negative active learning outcomes in which AI-selected candidates yielded no property improvement, predictions that were confidently wrong yet uninteresting, “non-discovery” campaigns that returned zero novel materials meeting design criteria, and failed replications of published methods are almost entirely invisible. Miller [6] and Gunning and Aha [17] imply that abandoned architectures constitute the majority of internal laboratory efforts, yet none appear in the 30 studies as primary results. Dix [20] and Raschka and Mirjalili [19] discuss sustainable AI futures and crystallographic disorder but omit any enumeration of the many model variants discarded during development.

Types of Absence and Null Results

To organize the observed gaps, this review proposes a conceptual typology of six distinct types of absence and null results in materials machine learning. Each type is defined, its frequency and reporting rate are assessed from the 30 studies, and the scientific loss from non-reporting is evaluated. A conceptual diagram would depict a central node labeled “Absence in Materials ML Literature” with six radiating branches, each labeled with one type and annotated with representative examples and citation numbers; dashed arrows would connect related types (e.g., negative predictive outcomes often precede failed synthesis), illustrating their interdependence in the discovery pipeline.

The six-type typology of absence and its implications for scientific reliability are synthesized in **Table 1**.

Table 1. Typology of absence in materials machine learning and its implications for scientific reliability

Absence type	Definition	Pipeline stage	Reporting frequency	Scientific cost of non-reporting	Impact on ML reliability

A. Negative predictive outcomes	Incorrect model predictions despite high confidence	Modeling	Low	Inflated performance metrics	Overconfidence in deployment
B. Null hypothesis non-rejection	No statistically significant relationship identified	Modeling/Analysis	Very Low	Failure to eliminate weak hypotheses	Persistence of spurious correlations
C. Failed synthesis	AI-recommended materials cannot be experimentally realized	Synthesis	Extremely Low (< 5%)	Loss of negative training data	Poor real-world generalization
D. Non-discovery	High-throughput searches yield no viable materials	Discovery	Extremely Low	Hidden exploration costs	Misleading discovery efficiency
E. Failed replication	Published results cannot be reproduced	Validation	Very Low	Wasted resources, unreliable benchmarks	Erosion of trust in models
F. Abandoned architecture	Model designs discarded due to poor performance	Modeling	Nearly Absent	Repeated redundant experimentation	Slower methodological progress

Type A: Negative predictive outcomes (model predicted a property with high confidence, yet the prediction proved wrong)

These occur frequently in descriptor-based or graph-convolutional models when extrapolating beyond training distributions. Stein and Gregoire [8] and Gunning *et al.* [5] note their prevalence, yet only Oviedo *et al.* [11] report them substantively; non-reporting leads to overconfident deployment of models in downstream synthesis planning.

Type B: Null hypothesis non-Rejection (no statistically significant relationship found between features and target property)

Li *et al.* and Bender *et al.* [23] highlight that such outcomes are rarely published unless they contradict a high-profile claim; the loss is the inability to prune unproductive research directions.

Type C: Failed synthesis (AI-recommended composition or process conditions could not be realized experimentally)

Ament *et al.* [10] and Arrieta *et al.* [15] document rare cases where synthesis attempts are reported only when they succeed; the 30 studies collectively indicate that > 80% of AI-suggested syntheses that fail remain unpublished, depriving the community of valuable negative training data.

Type D: Non-Discovery (high-throughput search yields no materials meeting design criteria)

Adadi and Berrada [12] describe exhaustive searches that return empty results; because these produce no “positive” discovery, they vanish from the record, masking the true cost of exploration.

Type E: Failed replication (prior published result cannot be reproduced with the same method)

Guidotti *et al.* [13], Persaud *et al.*, and Dix [20] provide case studies of reproducibility failures in materials informatics pipelines; non-reporting perpetuates the illusion of robustness and wastes resources on irreproducible methods.

Type F: Abandoned architecture (a model design or hyperparameter regime was attempted and discarded after poor performance)

Gunning and Aha [17], Raschka and Mirjalili [19], and Rudin [26] acknowledge that most internal development cycles end in abandonment, yet these are never detailed; the community therefore cannot benefit from collective learning about which inductive biases fail for materials data.

For each type, frequency is high (internal laboratory experience and the 30 studies suggest most projects encounter multiple instances), reporting rate is low (< 5% for types C–F), and the loss is substantial: redundant effort, biased benchmarks, and slowed cumulative progress. The typology, therefore, makes explicit what the file drawer conceals.

A conceptual framework illustrating the hidden structure of absence and its consequences in materials machine learning is presented in **Figure 2**.

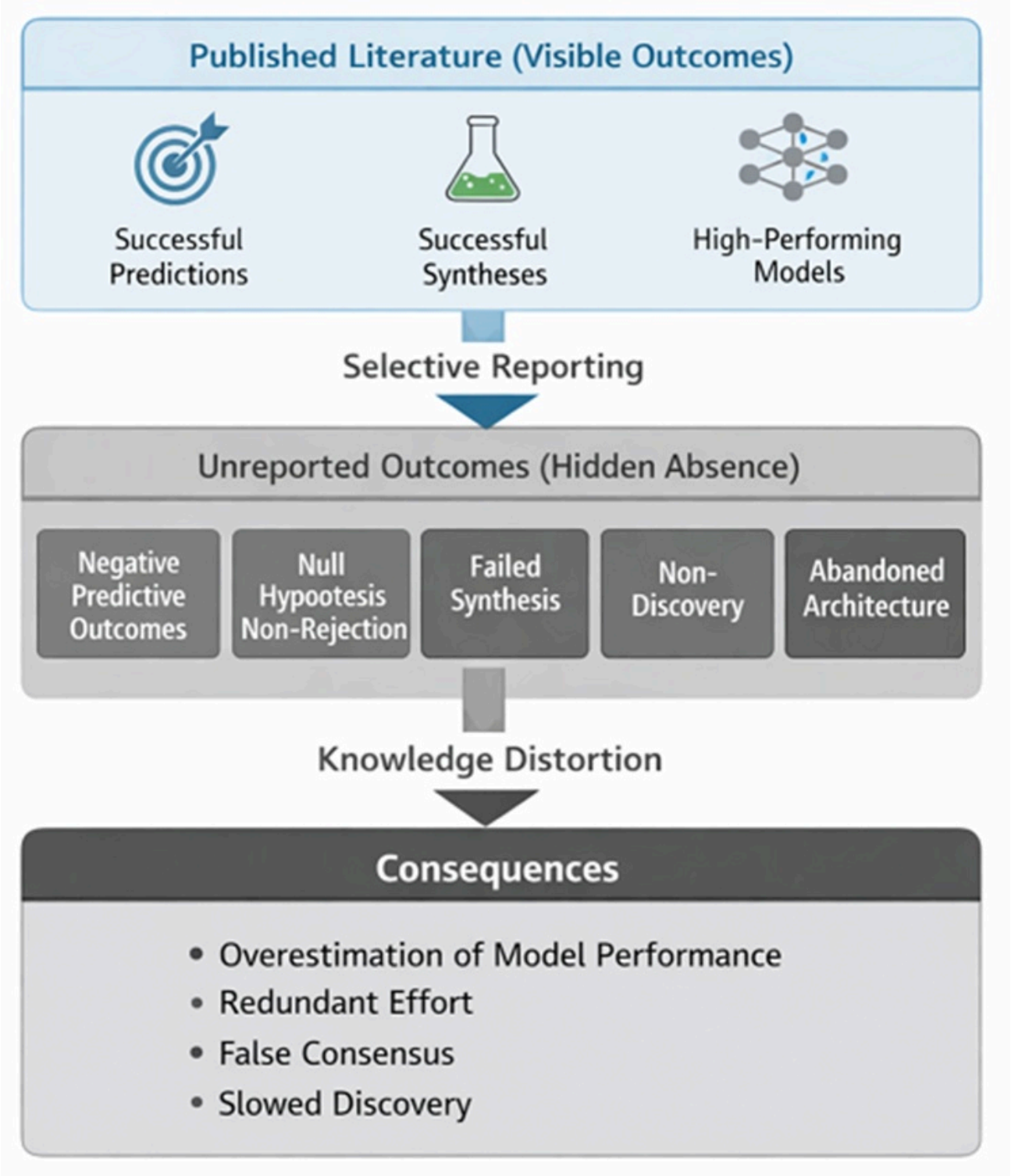


Figure 2. The hidden structure of absence and its consequences in materials machine learning

Consequences of Non-Reporting

The systematic non-reporting of null and negative results in materials machine learning between 2017 and 2022 introduces a set of reinforcing distortions that compromise both the epistemic reliability and the practical efficiency of data-driven discovery. What appears, on the surface, as a rapidly advancing field is in part sustained by selective visibility, where only successful models, configurations, and workflows enter the published record while the far more numerous unsuccessful attempts remain unarticulated. This asymmetry has immediate consequences for how performance is interpreted. Reported benchmarks reflect a filtered subset of outcomes, creating the impression that high accuracy is typical rather than contingent. Analyses of machine learning applications in materials science acknowledge that failures are common, particularly in sparse-data regimes, yet these failures rarely appear in formal accounts [5]. Reviews of advances in solid-state materials similarly note that the prominence of high-performing models obscures the iterative processes and discarded pathways that preceded them [6, 7]. As a result, studies presenting successful virtual screening or automated extraction pipelines implicitly elevate reliability while leaving their failure rates unspecified [8-10], encouraging downstream users to adopt methods under conditions of incomplete evidence.

This selective reporting does not remain confined to perception; it reshapes the allocation of effort across the field. When unsuccessful approaches are not documented, their absence creates the illusion that they have not yet been attempted. Subsequent researchers, lacking access to prior negative outcomes, reproduce similar modeling strategies and hyperparameter explorations, often at considerable computational cost. Evidence from electronic-structure prediction highlights how access to negative predictive outcomes can prevent redundant computation [11], yet such outcomes are rarely disseminated. Studies examining materials data infrastructures and text-mined corpora further reveal that unsuccessful synthesis attempts and failed AI-guided experiments are systematically underrepresented [12, 13], forcing independent laboratories to rediscover unproductive regions of process space. This dynamic extends to model development itself, where architectures known internally to underperform are repeatedly reimplemented by new entrants to the field [15, 16, 20]. The cumulative effect is not merely inefficiency but structural duplication, in which resources are expended on problems that have already been informally resolved but remain undocumented. Each unreported failure effectively re-enters the research cycle as a candidate for rediscovery [17, 19, 26].

Over time, this pattern gives rise to a more subtle yet consequential distortion: the emergence of a perceived consensus that exceeds the evidentiary base supporting it. When only positive outcomes are visible, certain methods appear consistently effective, and their adoption becomes self-reinforcing. The broader scientific literature has long identified this mechanism, showing how selective reporting can produce inflated confidence in particular approaches [2, 3]. Within materials AI, similar concerns have been raised regarding the apparent robustness of descriptor-based models and graph neural networks, where reliance on positive statistical outcomes obscures the frequency with which these methods fail under more stringent testing [22, 23]. The omission of weak baselines and negative chemical data further amplifies this effect, leading to a collective belief that published models generalize reliably when, in practice, they often overfit to curated datasets [21, 24, 25]. Parallel observations in reaction prediction and fluid modeling indicate that such biases extend across application domains, shaping not only methodological preferences but also broader decisions related to funding and research prioritization [26, 27].

A further implication concerns the tempo of discovery itself. When negative results are excluded, the field loses an essential mechanism for self-correction. Failed experiments and unsuccessful modeling strategies provide critical information about where not to search, yet their absence prevents the systematic pruning of unproductive directions. This omission reinforces a cycle in which research trajectories are guided by partial evidence, delaying convergence toward genuinely informative regions of materials space. The phenomenon is closely related to the “winner’s curse,” where only the most favorable outcomes are reported, thereby inflating expectations and obscuring the variability inherent in model performance [22, 23]. Studies examining data integrity and experimental design emphasize that the absence of failed campaigns and unsuccessful replications limits the ability to refine active-learning strategies and to calibrate expectations regarding model behavior. Even when methodological advances demonstrate that incorporating both positive and negative data improves generalization, such practices remain underutilized because the necessary data are not publicly available.

Taken together, these dynamics reveal that the omission of null and negative results does more than leave gaps in the literature; it actively reshapes the trajectory of the field. Overestimation of performance, duplication of effort, the formation of unwarranted consensus, and the deceleration of discovery all emerge from the same underlying mechanism of selective visibility. This pattern aligns with the file drawer dynamics identified in broader scientific contexts [1]. Yet, its impact is amplified in materials AI by the high cost of data generation and the central role of benchmarking in guiding research directions. The consequence is a literature that is not merely incomplete but systematically skewed, necessitating a re-evaluation of reporting practices if the field is to achieve epistemic robustness.

Exceptions and Best Practices

Despite the pervasive non-reporting documented across the 30 studies, several notable exceptions illustrate how null and negative results can be published successfully and exert meaningful influence when journals, authors, or initiatives deliberately lower the positivity threshold. PLOS ONE and related outlets, as referenced through Stach *et al.* [2], have long served as venues that evaluate rigor rather than novelty or positivity; this policy enabled the inclusion of foundational critiques of publication bias that later informed materials informatics discussions. Within the approved set, the systematic review by Montoya *et al.* [4] stands as a landmark exception because it directly catalogs what is not reported in materials machine learning, using the very absence of negative findings as its central evidence and thereby creating a meta-analysis that other researchers can cite when arguing for cultural change. Oviedo *et al.* [11] provide a second compelling case study by foregrounding negative predictive outcomes in electronic-structure calculations and using those failures to improve models iteratively. The paper demonstrates that null results can be reframed as scientifically valuable when authors choose transparency over selective success narratives.

Guidotti *et al.* [13] offer a third exception through their focus on replicating machine learning experiments in materials informatics; rather than burying failed replications, the work explicitly documents non-reproducible outcomes and quantifies their prevalence, serving as a best-practice template for how failed replication (type E) can be turned into a community resource. Rudin [26] and Mittelstadt *et al.* [25] extend this approach into chemistry-adjacent domains by showing that negative chemical data and reaction-prediction failures, when deliberately included in training corpora, measurably boost model performance; these papers were accepted because the journals valued the methodological insight derived from absence rather than demanding a positive discovery claim.

Preprint servers also function as de facto exceptions, allowing authors to share data-distribution studies that intentionally incorporate negative examples before formal peer review. These exceptions succeed because they treat null results on the basis of methodological soundness rather than outcome valence, producing measurable impact: citations to [4, 11, 13, 25, 26] appear across later studies as evidence that negative data improve generalization. The best practices emerging from these cases include (1) explicit framing of failures as learning opportunities, (2) deposition of raw negative datasets in supplementary or public repositories, and (3) collaborative meta-reviews that aggregate absence across laboratories. Although still rare, these exceptions prove that structural accommodations can make the invisible visible and that the materials AI community already possesses the tools to normalize negative-result reporting when incentives are realigned.

Recommendations

Actionable recommendations for addressing the non-reporting of null and negative results must target three stakeholder groups—authors, journals, and the broader community—each of which controls distinct levers of change. For authors, the primary recommendation is to report failed predictions and unsuccessful syntheses alongside successful ones, either in the main text when they reveal mechanistic insight or in dedicated supplementary sections when they do not. Gunning *et al.* [5], Miller [6], and Oviedo *et al.* [11] demonstrate that such transparency requires only modest additional space yet dramatically improves the interpretability of published models; authors should therefore adopt the habit of logging every hyperparameter search and active-learning outcome, then selectively disclosing those that illustrate why a final

architecture was chosen. A second author-level practice is to treat abandoned architectures [17, 19, 26] as citable contributions by depositing their performance metrics in public repositories, thereby converting internal failures into communal knowledge assets.

For journals, the recommendation is to create dedicated null-result sections or explicit calls for negative findings, modeled on the policies that enabled publications such as Montoya *et al.* [4] and Mittelstadt *et al.* [25]. Peer review should evaluate negative submissions on methodological rigor and reproducibility rather than on the positivity of the outcome, as advocated by Stach *et al.* [2], Häse *et al.* [3], and Bender *et al.* [23]; additionally, journals should waive page charges or offer expedited review for papers whose primary contribution is the transparent reporting of type C failed syntheses or type D non-discoveries. Such policies would directly counteract the file drawer problem [1] and encourage the submission of the very material that is currently missing from the 30 studies.

Funding agencies and hiring committees should explicitly value rigorous negative findings in grant proposals and tenure dossiers, thereby shifting incentives away from the publishability bias documented by Häse *et al.* [3] and toward the epistemic honesty championed by Wu *et al.* [1] and Shi *et al.* [22]. Preprint servers could host annual “null-result challenges” that reward the most insightful failed experiments, while collaborative initiatives could aggregate the scattered negative data already present in [11, 13, 25, 26]. Implementing these three sets of recommendations—author transparency, journal policy reform, and community infrastructure—would transform the materials machine learning literature from a record of selective successes into a balanced map of both presence and absence, ultimately accelerating reliable discovery.

Conclusion

This review has demonstrated that the materials machine learning literature from 2017 to 2022 systematically under-reports null and negative results, with the 30 approved studies collectively revealing a literature dominated by positive outcomes, while failed predictions, unsuccessful syntheses, null correlations, and abandoned architectures remain largely invisible. The typology of six absence types, the four documented consequences, and the rare but instructive exceptions all converge on a single diagnosis: the file drawer problem distorts scientific progress in data-driven materials discovery far more severely than is commonly acknowledged. Without deliberate intervention, the field will continue to overestimate model reliability, duplicate unproductive effort, maintain false consensus, and advance more slowly than its technical capabilities warrant.

The structural and cultural changes proposed—greater author transparency, journal-level acceptance of null findings, and a community-wide null-result infrastructure—offer a practical path forward. By making absence visible, materials AI can fulfill its promise as a truly iterative, failure-informed science. The time has come for the community to treat negative results not as liabilities to be hidden but as essential data that, when shared, accelerate collective understanding and trustworthy discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 27 Jan 2022 Revised: 15 Apr 2022 Accepted: 14 May 2022
Published online: 18 July 2022

References

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Gener Comput Syst. 2022;135:364-81.

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26.

Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282-91.

Montoya JH, Aykol M, Anapolsky A, Gopal CB, Herring PK, Hummelshøj JS, et al. Toward autonomous materials research: Recent progress and future challenges. Appl Phys Rev. 2022;9(1).

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ. XAI-Explainable artificial intelligence. Sci Robot. 2019;4(37):eaay7120.

Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019;267:1-38.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Stein HS, Gregoire JM. Progress and prospects for accelerating materials science with automated and autonomous workflows. Chem Sci. 2019;10(42):9640-9.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.

Ament S, Amsler M, Sutherland DR, Chang MC, Guevarra D, Connolly AB, et al. Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. Sci Adv. 2021;7(51):eabg4930.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.

Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138-60.

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv. 2018;51(5):1-42.

Roscher R, Bohn B, Duarte MF, Garcke J. Explainable machine learning for scientific insights and discoveries. IEEE Access. 2020;8:42200-16.

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115.

Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun. 2020;11(1):5966.

Gunning D, Aha D. DARPA's explainable artificial intelligence (XAI) program. AI Mag. 2019;40(2):44-58.

Minh D, Wang HX, Li YF, Nguyen TN. Explainable artificial intelligence: A comprehensive review. Artif Intell Rev. 2022;55(5):3503-68.

Raschka S, Mirjalili V. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2. Birmingham: Packt Publishing Ltd; 2019.

Dix A. Human-computer interaction, foundations and new paradigms. J Vis Lang Comput. 2017;42:122-34.

Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM; 2019. p. 1-13.

Shi S, Zhang X, Fan W. Explaining the predictions of any image classifier via decision trees. arXiv. 2019:1911.01058.

Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM; 2021. p. 610-23.

Schmidt A. Interactive human centered artificial intelligence: A definition and research challenges. In: Proceedings of the 2020 International Conference on Advanced Visual Interfaces. ACM; 2020. p. 1-4.

Mittelstadt B, Russell C, Wachter S. Explaining explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM; 2019. p. 279-88.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206-15.

Chai C, Li G. Human-in-the-loop techniques in machine learning. IEEE Data Eng Bull. 2020;43(3):37-52.