

ORIGINAL RESEARCH

Open access

Perspective: Why Materials AI Needs More Failure Reporting — A Position on Negative Results as Community Infrastructure

Anna Novak^{1*}, Tomas Hruby¹

Abstract

Materials AI stands at a critical juncture where rapid advances in machine learning and data-driven discovery are reshaping materials engineering. Yet the field is severely constrained by a deep-rooted publication bias that systematically favors positive results while suppressing negative outcomes. This position paper asserts that Materials AI urgently requires widespread failure reporting and that negative results must be recognized and maintained as essential community infrastructure. In the prevailing culture, only successful predictions, high-accuracy models, and newly discovered materials reach publication, while failed models, insufficient datasets, misleading benchmarks, and unsuccessful synthesis attempts remain hidden. The consequences are profound: researchers repeatedly traverse the same unproductive paths, the scientific literature presents a distorted view of reality, meta-analysis becomes impossible, and overall progress slows dramatically. Failure reporting functions as shared infrastructure—much like open databases or standardized protocols—by mapping the boundaries of what does not work and thereby enabling more efficient exploration by the entire community. Four core types of failures demand routine documentation: model failures, data failures, benchmark failures, and synthesis failures, with additional emphasis on transfer and extrapolation failures. By treating negative results as valuable contributions rather than career liabilities, the community can accelerate discovery, enhance reproducibility, and build a more trustworthy knowledge base. This paper articulates the position, analyzes the publication bias problem, defines failure reporting as infrastructure, classifies failure types, and identifies barriers before offering concrete recommendations. Embracing negative results will transform Materials AI from a literature of selective successes into a robust, transparent, and collectively intelligent discipline capable of delivering sustainable materials innovation at scale.

Keywords Materials AI, Failure reporting, Negative results, Reproducibility crisis, Publication bias, Community infrastructure

*Correspondence:

Anna Novak
anna.novak@gmail.com

¹ Department of Data-Driven Materials Engineering, Faculty of Engineering, Czech Technical University, Prague, Czech Republic

Introduction

The position

Materials AI suffers from a publication bias toward positive results. Papers that report successful predictions, accurate models, or discovered materials are published. Papers that report failures—models that didn't work, data that was insufficient, benchmarks that misled—are suppressed. This is bad science. Failure reporting is not a sign of weakness; it is community infrastructure. Knowing what doesn't work saves others from repeating the same mistakes. This position

paper argues that materials AI needs more failure reporting, and proposes that negative results be treated as valuable community resources [1, 2].

The field of computational and data-driven materials engineering has matured dramatically over the past decade [3, 4]. Sophisticated machine-learning architectures, large-scale materials databases, and automated discovery pipelines now routinely generate predictions that once required years of experimental effort [5, 6]. Yet this very success has exposed a structural weakness: the literature contains almost no systematic record of the countless dead ends that have been explored [7, 8]. Without failure reporting, the community operates in an information vacuum where the same unproductive approaches—flawed model architectures, biased datasets, or overly optimistic benchmarks—are reinvented by new researchers and new projects [9, 10].

This position matters now because the field has accumulated sufficient experience to recognize recurring patterns of failure [11, 12]. High-profile claims of “state-of-the-art” performance frequently collapse under closer scrutiny or when applied to new material classes [13, 14]. The absence of negative data means that lessons learned the hard way remain private, forcing each laboratory to rediscover limitations that could have been shared openly [15, 16]. In clinical research and drug discovery, regulatory frameworks already mandate the reporting of negative results precisely because hidden failures endanger patients and waste resources [17]. Materials AI, which aims to accelerate the discovery of sustainable energy materials, quantum materials, and next-generation alloys, cannot afford a similar hidden-cost structure [18, 19].

Treating negative results as community infrastructure reframes failure from a personal setback into a public good [20].

Figure 1 shows that failure reporting should be understood not as an optional supplement to successful Materials AI research, but as the core infrastructure required to convert hidden negative outcomes into cumulative community knowledge.

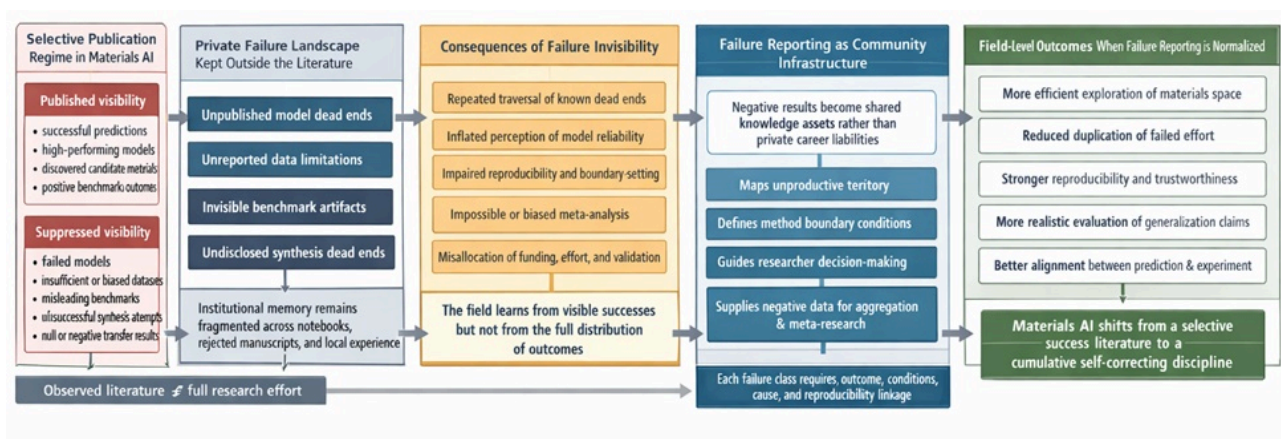


Figure 1. Conceptual architecture showing why failure reporting functions as core community infrastructure in Materials AI.

A shared failure registry would function like a map of explored but unproductive territory, guiding future expeditions away from known pitfalls and toward genuinely novel ground [1, 21]. It would provide boundary conditions for methods, reveal dataset biases, expose benchmark artifacts, and document synthesis routes that consistently fail [8, 22]. Most importantly, it would restore balance to the scientific record, enabling rigorous meta-analysis and preventing the systematic overestimation of model reliability that currently plagues the literature [23, 24].

The time has come for Materials AI to adopt a new norm: every major project should report both its successes and its failures with equal rigor. Only then can the community build the cumulative, self-correcting knowledge base required for

the ambitious materials challenges of the twenty-first century [4, 5]. This position paper develops that argument in detail, demonstrating why failure reporting is not optional but foundational to the future health and productivity of the field.

The Publication Bias Problem

What reaches publication in Materials AI today is overwhelmingly positive [13]. Journals feature new architectures that achieve “state-of-the-art” performance on standard benchmarks, successful discovery campaigns that yield promising candidate materials, and benchmarks that appear to validate the latest methods [3, 7]. Positive correlations between descriptors and properties are highlighted, while null or inverse relationships are omitted [2]. Reviewers and editors reward novelty and impact, metrics that positive outcomes satisfy far more readily than carefully documented failures [1, 15].

What remains unpublished is equally important but systematically invisible. Models that failed to generalize beyond their training distribution, synthesis attempts that never produced the target phase, benchmarks that produced misleadingly high scores for all methods, and null results showing no statistically meaningful correlation are almost never reported [14, 25]. The literature therefore presents a filtered, optimistic view of reality that does not reflect the true distribution of research effort [24, 26].

Several interlocking incentives sustain this bias. Journal editors and reviewers often view negative results as “not novel” or “lacking impact,” implicitly requiring success as a precondition for publication [2, 17]. Early-career researchers fear that documenting failures will damage their publication record, citation counts, and prospects for funding or tenure [15, 26]. Funding agencies, under pressure to demonstrate return on investment, implicitly reward proposals and reports that emphasize positive outcomes [5]. The result is a self-reinforcing cycle in which only polished success stories enter the public domain while the raw, messy reality of scientific exploration stays hidden [10, 16].

The consequences are severe and cumulative. First, the literature becomes systematically misleading: readers and subsequent researchers inherit an inflated sense of method reliability that does not survive real-world deployment [8, 22]. Second, time and resources are wasted as independent groups unknowingly repeat failed experiments or model trainings [9, 27]. Third, the field lacks any institutional memory of what does not work, forcing each new cohort to rediscover the same limitations [1, 21]. Fourth, meta-analysis—the statistical synthesis of evidence across studies—becomes impossible because negative and null findings are missing from the record [23, 24].

Publication bias also distorts resource allocation. Funding follows the most visible successes, leaving promising but failure-prone directions under-explored [5]. Reproducibility suffers because the conditions under which models or methods fail are rarely documented, making it difficult for others to verify claims or understand limitations [8, 10]. In extreme cases, the absence of negative data can lead to overconfidence in deployed materials AI tools, with downstream consequences for experimental validation and technology transfer [4, 7].

The bias is not unique to Materials AI, yet its impact is particularly damaging here. Materials discovery is an inherently high-risk, high-dimensional search problem; the vast majority of computational predictions do not translate into viable materials [3, 19]. Without a shared record of those failures, the community cannot learn efficiently from the enormous search space it has already sampled [14, 18]. The problem is therefore not merely academic—it directly impedes the societal goal of rapid, sustainable materials innovation [11, 12].

Why Failure Reporting is Community Infrastructure

Failure reporting is the systematic documentation and public sharing of experimental, computational, or modeling attempts that did not achieve the desired outcomes [17, 20]. It is not a confession of error but a deliberate contribution to collective knowledge. When researchers publish that a particular architecture underperformed on a target property, that a

dataset proved insufficient for extrapolation, or that a synthesis route consistently failed, they create a public good that benefits the entire field [9, 16].

Failure reporting supplies four critical resources. It generates a map of explored but unproductive territory, preventing others from repeating the same experiments [21]. It establishes boundary conditions for methods, clearly delineating where architectures or algorithms break down [14, 25]. It offers practical guidance to new researchers, shortening the learning curve by highlighting pitfalls that experienced groups have already encountered [8, 10]. Finally, it supplies the negative and null data required for rigorous meta-analysis, enabling the community to estimate true effect sizes and quantify heterogeneity across studies [23, 24].

The infrastructure analogy is apt. Just as roads, bridges, and power grids enable physical mobility and economic activity, failure databases and negative-result repositories enable efficient scientific mobility [16]. Without them, every research group must blaze its own trail through the same thickets of unproductive ideas. With them, the community can navigate the materials design space with greater speed and precision [3, 5]. Other disciplines have already recognized this value. Clinical trials require prospective registration and reporting of all outcomes, positive or negative, to protect patients and avoid duplication [17]. Drug-discovery consortia maintain public databases of failed compounds to accelerate collective learning [18]. In machine learning more broadly, platforms such as OpenML and PapersWithCode increasingly document both successes and failures to support reproducible research [9, 20].

Materials AI currently forfeits these advantages. The community lacks accessible knowledge that a particular model architecture fails systematically on rank-3 tensor properties, that a widely used dataset introduces extrapolation bias for high-entropy alloys, or that a popular benchmark artifact inflates performance across multiple methods [8, 14]. Without failure reporting, synthesis groups waste resources attempting routes that computational teams have already shown to be unreliable, and model developers cannot calibrate expectations about generalization limits [7, 22]. The knowledge that “Model X fails on property Y” or “Dataset D is unsuitable for task T” remains locked in laboratory notebooks and rejected manuscripts [1, 15].

By contrast, treating failure reporting as infrastructure creates a positive feedback loop. Documented failures improve the design of future studies, reduce wasted effort, and increase the credibility of positive claims because the community can see the full distribution of outcomes [2, 10]. Reproducibility improves because boundary conditions are explicit rather than tacit [8, 27]. New researchers enter the field with realistic expectations and practical guidance rather than an idealized literature of unbroken success [4, 11]. Ultimately, the shared infrastructure of negative results accelerates the very progress that positive-result-only publishing claims to deliver [3, 5].

Types of Failures to Report

To make failure reporting actionable, six distinct but interrelated types must be recognized and documented.

Table 1 consolidates the six failure classes into an infrastructure-oriented analytical schema by showing what each class reveals, what evidence must be reported, and how individual negative results become reusable community assets.

Table 1. Failure classes in Materials AI as infrastructure-grade knowledge assets: analytical role, minimum reporting fields, and collective reuse value.

Failure class	Analytical definition	What this failure reveals about the knowledge system	Minimum reportable evidence	Immediate local value	Collective infrastructure value

Model failure	A model architecture, representation choice, or training strategy performs poorly for a defined materials task under specified conditions	Methodological boundary conditions; mismatch between inductive bias and target property structure	Task definition; architecture/training details; quantitative outcome; baseline comparison; conditions; hypothesized mechanism	Prevents repeated investment in incompatible architectures	Builds a searchable map of method limits across tasks and material classes
Data failure	A dataset is insufficient, biased, incomplete, or distributionally misaligned with the intended predictive or inferential goal	Limits of dataset portability; hidden sampling bias; incompleteness of training support	Dataset identity/version; sampling scope; target task; failure metric; evidence of distributional mismatch; conditions	Improves future data selection and curation decisions	Enables community-wide identification of recurring data bottlenecks and under-covered regions of materials space
Benchmark failure	An evaluation protocol produces misleading estimates of performance because of artifacts in split design, leakage, saturation, or metric choice	Distortion in what the field recognizes as “progress”; inflation of apparent generalization	Benchmark description; split logic; competing results under alternative protocol; inflation estimate; artifact diagnosis	Protects individual studies from overclaiming	Supports reform of field-wide evaluation standards and leaderboard validity
Synthesis failure	Computationally predicted or hypothesized materials cannot be realized, stabilized, or reproduced under attempted experimental conditions	Breakdown between computational promise and experimental realizability	Material target; prediction basis; synthesis conditions; number of attempts; observed phases/outcomes; failure threshold	Saves experimental time and resources locally	Provides a shared record of nonviable routes and strengthens theory–experiment calibration
Transfer failure	Knowledge transferred from one domain, dataset, or pretraining regime degrades performance on the target problem	Boundary of cross-domain portability; hidden incompatibility between source and target spaces	Source domain; target domain; transfer setup; no-transfer baseline; performance degradation; conditions	Prevents uncritical reuse of pretrained models	Clarifies where transfer learning is robust, fragile, or counterproductive across materials domains

Extrapolation failure	A model performs acceptably in-distribution but collapses outside the training support in composition, structure, property, or process space	Limits of generalization claims; unsafe extension beyond observed chemical or structural regions	Training-domain definition; extrapolation domain; error increase; uncertainty behavior; failure threshold; conditions	Reduces overconfident deployment	Establishes operational guardrails for model use in unexplored materials space
-----------------------	--	--	---	----------------------------------	--

Model failures arise when architectures or training strategies prove incompatible with specific materials tasks, as when invariant graph neural networks produce near-zero outputs for piezoelectric tensor prediction and achieve less than 10 % of baseline performance on standard test sets [7, 20]. Reporting these outcomes with quantitative metrics, training curves, and ablation studies prevents redundant exploration of unsuitable approaches for tensorial properties. A related implication concerns data failures, in which available datasets introduce fundamental limitations, such as when the Materials Project database generates extrapolation biases for high-entropy alloys distant from binary and ternary distributions [8, 14]. Documenting such constraints equips researchers to select or curate data more effectively rather than assuming completeness confers universal applicability [21, 25].

This concern extends to benchmark failures, where standard evaluation protocols embed artifacts that distort assessment, for instance when random splits on materials datasets allow chemically similar compounds across train and test sets and inflate performance by tens of percent [8, 10]. Explicit identification of these artifacts compels the development of more rigorous standards and reduces dependence on contaminated leaderboards [9, 27]. Beyond immediate modeling concerns, synthesis failures record experimental realizations of computationally predicted materials that did not succeed, as in attempts to synthesize material M with composition C predicted by model X to exhibit property P, where all attempts under conditions S failed to yield the target phase [3, 5, 7]. Such accounts close the prediction-reality gap and spare downstream efforts.

Transfer failures further reveal boundaries when pre-training on one domain degrades performance on another, decreasing accuracy by more than 20 % when moving from large oxide datasets to sulfides or halides in certain property spaces [14, 18]. Similarly, extrapolation failures expose limits as models trained solely on binary alloys exhibit errors exceeding 50 % on ternary or higher-order systems [8, 14]. Each case, reported with quantitative measures, hypothesized mechanisms, and links to data and code, transforms the literature from selective highlights into a reliable guide for materials discovery [1, 16, 20].

Barriers to Failure Reporting

Despite the clear value of failure reporting, several entrenched barriers prevent its widespread adoption.

Table 2 shows that the persistence of publication bias in Materials AI is sustained by identifiable institutional mechanisms, each of which can be countered through a specific governance intervention that converts hidden failures into usable infrastructure.

Table 2. From publication bias to reporting infrastructure: barrier mechanism, epistemic cost, and governance intervention in Materials AI.

Barrier to failure reporting	Underlying mechanism	Epistemic cost to the field	Institutional actor with primary leverage	Governance intervention	Expected system effect if adopted

Perceived career risk	Researchers interpret visible failure as reputational damage rather than scientific contribution	Negative evidence stays private; the literature becomes selectively optimistic	Universities, hiring committees, funders, senior investigators	Recognize rigorously documented failures in evaluation, promotion, and grant reporting	Reclassifies failure disclosure from liability to credible scholarly contribution
Journal preference for positive novelty	Editorial and reviewer norms equate impact with success rather than with boundary-setting knowledge	Valuable negative findings are filtered out before entering the record	Journals, editors, reviewers	Create dedicated negative-results tracks and evaluate rigor, not only success	Increases the formal publication pathway for high-quality failure reports
Lack of citation credit	Incentive systems reward visible positive claims more than infrastructural clarification	Researchers rationally suppress work with lower expected bibliometric return	Journals, societies, indexing ecosystems	Introduce citable failure formats, registries, and curated thematic collections	Gives failure reports stable scholarly visibility and reference value
High effort with low reward	Failure reporting requires the same documentation labor as positive publication but yields less recognition	Underreporting persists even when failures are well understood	Funders, institutions, project leads	Budget reporting time explicitly and require structured failure deposition at project close	Reduces friction and normalizes reporting as part of standard research workflow
Fear of being scooped or strategically exposed	Groups assume revealing failed directions confers advantage to competitors	Nonproductive paths remain duplicated across laboratories	Research groups, consortia, funders	Encourage embargoed or sanitized registry entries where needed, followed by timed release	Preserves strategic flexibility while still contributing to collective learning
Lack of standardized format	Failure accounts are heterogeneous, difficult to search, and impossible to aggregate	No meta-analysis, no interoperability, no infrastructure effect	Community bodies, repositories, journals	Adopt common metadata fields across journals, preprints, and registries	Converts scattered narratives into machine-readable and comparable knowledge objects

Perceived career risk constitutes the most immediate barrier, as early-career researchers particularly worry that documenting failures will signal incompetence and undermine grant, hiring, or promotion prospects [15, 26]. In practice, such well-characterized failures reflect intellectual honesty and methodological rigor, yet prevailing incentive structures continue to favor polished success narratives [2]. This dynamic is reinforced by journal policies in materials science and machine learning that explicitly or implicitly discourage negative results, prioritizing novelty and positive impact while offering few dedicated tracks for rigorously characterized failures, often leading reviewers to dismiss them as insufficiently advancing the field [1, 7, 16, 17].

The problem deepens through limited credit and publication challenges. Failure reports typically attract fewer citations than positive counterparts, diminishing the perceived return on equivalent investments in design, analysis, and writing [2, 9, 15, 16]. Without standardized formats or recognition mechanisms, the process feels disproportionately burdensome, while fear of being scooped prompts some groups to withhold negative data despite the community-wide benefit of revealing unproductive directions [5, 8, 10, 20, 26]. Finally, the absence of agreed-upon metadata fields leaves failure reports scattered and difficult to discover or synthesize, undermining the very infrastructure needed to realize their scientific value [1, 10, 21, 27].

Overcoming these interlocking barriers demands coordinated action across journals, funders, institutions, and the research community itself.

Objections and Responses

A common objection is that “Negative results are not novel” [1, 2]. This view equates novelty exclusively with positive discovery and overlooks the deeper value of negative findings. Negative results save the community substantial resources by documenting unproductive paths that would otherwise be retraced [15, 16]. In a maturing field such as Materials AI, where thousands of model variants and dataset combinations have already been explored, the absence of novelty in a failure does not diminish its utility as shared infrastructure [17, 20].

Another frequent concern is “My failure might be due to my error, not a general limitation” [10, 26]. Researchers worry that reporting an unsuccessful attempt could expose methodological flaws unique to their implementation. The response is straightforward: report it anyway. Even if the failure stems from implementation details, the documentation allows others to judge context, replicate conditions, and distinguish between user error and fundamental limits [8, 9]. Transparent reporting of potential errors actually strengthens community trust and accelerates collective debugging [22, 27].

The objection “No journal will publish it” is frequently raised [1, 17]. While many traditional venues remain reluctant, emerging outlets such as Digital Discovery, PLOS ONE, and Patterns explicitly welcome negative results or null findings [8, 23]. Preprint servers and community repositories further lower the barrier, enabling rapid dissemination before formal publication [15, 20]. The landscape is shifting precisely because the reproducibility crisis has made the cost of hidden failures untenable [9, 10].

Some researchers claim “It will hurt my citation count” [2, 26]. Short-term citation metrics may indeed favor positive stories, yet long-term community respect accrues to those who demonstrate intellectual honesty [15, 16]. As failure reporting becomes normalized, papers that rigorously document both successes and failures will be cited as foundational boundary-setting works rather than isolated successes [4, 5]. The incentive structure itself must evolve through collective action [1].

Finally, the objection “Industry won’t allow sharing failures” applies primarily to proprietary work [5, 7]. Academic and publicly funded research, however, carries a responsibility to share negative outcomes for the public good [17, 21]. Industry partners increasingly recognize that open failure registries accelerate pre-competitive discovery, reducing duplicated effort across the ecosystem [3, 19]. Where proprietary constraints exist, sanitized or aggregated failure reports can still be shared without compromising intellectual property [14, 18].

Addressing these objections directly shows that the barriers are surmountable and that the scientific case for failure reporting outweighs individual concerns [2, 20].

Relation to Other Positions

Failure reporting is not an isolated proposal but is deeply intertwined with several established positions in the literature. Its strongest link is to the reproducibility agenda in machine learning and materials informatics [8, 10]. Reproducibility studies repeatedly demonstrate that published positive results are difficult to replicate precisely because the conditions under which methods fail are rarely documented [9, 27]. Failure reporting supplies the missing negative data that makes true reproducibility possible by defining the exact boundaries of claimed performance [7, 22].

The position also advances open science principles [17, 20]. Open science advocates argue that all results—not merely selected positive outcomes—should be shared to reduce bias and increase transparency [15, 26]. Failure reporting is the practical embodiment of this ideal: it transforms the scientific record from a curated highlight reel into a complete, unbiased dataset suitable for meta-research [23, 24]. Without negative results, open-science initiatives remain incomplete [1, 21].

Meta-research on publication bias further underscores the necessity of this position [2, 26]. Studies across disciplines show that the systematic suppression of negative findings distorts knowledge and wastes resources [15, 16]. Materials AI is especially vulnerable because its high-dimensional search space amplifies the cost of hidden failures [3, 5]. Failure reporting directly counters this bias by institutionalizing the publication of null and negative outcomes as standard practice [17, 20].

The position also builds on the emerging scholarship of failure mode analysis in computational materials [7, 16]. Recent works have begun cataloging specific model shortcomings, benchmark artifacts, and synthesis dead ends [8, 14]. This paper generalizes those isolated efforts into a systematic framework that treats failure reporting as core community infrastructure rather than occasional commentary [18, 25]. By classifying six failure types and proposing a standardized format, the position converts ad-hoc failure discussions into structured, searchable knowledge assets [9, 10].

Collectively, these relations demonstrate that failure reporting is not a radical departure but the logical next step in the ongoing maturation of Materials AI [4, 11]. It complements reproducibility, fulfills open-science commitments, rectifies publication bias, and elevates failure mode analysis from anecdote to infrastructure [1, 2]. The field now has the opportunity to integrate these threads into a coherent norm that values negative results as highly as positive discoveries [3, 5].

Recommendations for the Community

Concrete action is required across stakeholders to embed failure reporting as standard practice.

For journals, the recommendation is clear: create dedicated “Negative Results” sections and accept failure reports as full-length papers [1, 8]. Review criteria must shift from “novelty of success” to “rigor of documentation and clarity of implications” [2, 17]. Digital Discovery, Patterns, and npj Computational Materials are ideally positioned to lead by example, offering expedited review tracks for well-documented failures [15, 16].

For funders, the mandate should be to require failure reporting as a condition of grant closure [5, 21]. Funding calls can explicitly allocate resources for negative-result repositories and incentivize principal investigators to deposit failures alongside positive outcomes [10, 26]. Agencies such as the NSF, DOE, and EU Horizon programs could pilot “failure supplements” that reward transparent documentation of dead ends [9, 20].

For individual researchers, the call is to publish failures alongside successes whenever a project reaches meaningful scale [4, 7]. Preprints should be used for rapid dissemination of negative findings, allowing the community to benefit immediately [22, 27]. Early-career researchers in particular can build credibility by contributing to a shared failure registry, demonstrating both technical skill and scientific integrity [3, 11].

For the community as a whole, the priority is to establish a centralized “Materials AI Failure Registry” as an open, searchable database [1,16]. This registry would mirror successful models from clinical trials and drug discovery but tailored to computational and experimental materials workflows [17, 18]. A standardized metadata schema—detailed in the next section—must be adopted to ensure interoperability and machine-readability [8, 14]. Conferences and workshops should dedicate sessions to failure reports, normalizing their presentation and discussion [10, 28].

Implementation can begin modestly: existing repositories such as Zenodo or Figshare can host failure collections while a dedicated platform is developed [21, 25]. Professional societies in materials science and AI can endorse the registry and encourage its use in tenure and promotion dossiers [4, 5]. Within five years, the goal is for failure reporting to constitute 15–20 % of published Materials AI literature, providing the balanced dataset the field currently lacks [2, 15].

These recommendations are actionable, mutually reinforcing, and essential for transforming negative results from hidden liabilities into valued community infrastructure [1, 20].

Proposed Failure Reporting Format

To enable systematic aggregation and reuse, failure reports must follow a standardized template. The required fields are:

Failure type: one of Model, Data, Benchmark, Synthesis, Transfer, or Extrapolation. Goal: concise statement of what was attempted. Method: model architecture, dataset, benchmark, or synthesis protocol used. Outcome: quantitative description of the failure (error rates, success percentages, statistical measures). Why it failed: hypothesized mechanism or observed limitation (if known). Conditions: experimental or computational conditions under which the failure was observed. Related successes: any conditions or subsets where partial success occurred. Data/Code: link to reproducibility package (repository, DOI, or Zenodo archive).

This format ensures reports are concise yet sufficiently detailed for meta-analysis and reuse [8, 29].

Example Failure type: Model Goal: Predict piezoelectric tensor components for inorganic crystals. Method: Invariant graph neural network (CGCNN variant) trained on Materials Project dataset. Outcome: Model predicted all tensor components near zero, achieving <10 % of DFT baseline accuracy across 500 test structures. Why it failed: Imposed rotational invariance forces zero output for odd-rank tensors such as rank-3 piezoelectricity. Conditions: All tested materials with non-centrosymmetric structures; training on 10,000+ entries. Related successes: Model performed adequately on scalar properties (e.g., formation energy) under identical training. Data/Code: github.com/example/piezo-failure (includes trained weights, test splits, and analysis scripts) [7, 14].

Adopting this template across journals, preprints, and repositories will make failures machine-searchable and aggregatable [9, 21]. Authors should submit the structured report as supplementary material or as a standalone short communication [16, 17]. Over time, the registry can automatically ingest these fields, enabling dashboards that visualize common failure modes by material class or method type [1, 20]. The format deliberately avoids narrative length while preserving scientific traceability, striking the balance required for infrastructure-scale adoption [10, 27].

Conclusion

Materials AI needs more failure reporting. The positive-result-only literature is systematically misleading, wastes community resources, and slows the very progress it claims to accelerate. Failure reporting must be recognized and maintained as essential community infrastructure: it maps unproductive territory, defines methodological boundaries, and supplies the negative data required for trustworthy meta-analysis.

This position paper has classified six types of failures—model, data, benchmark, synthesis, transfer, and extrapolation—that demand routine documentation. It has identified the entrenched barriers of career risk, journal policies, lack of credit,

publication difficulty, fear of scooping, and absent standardization. It has answered common objections, situated failure reporting within the broader movements for reproducibility and open science, and provided concrete recommendations for journals, funders, researchers, and the community at large. Finally, it has proposed a standardized reporting format that makes negative results immediately useful and aggregatable.

The call to action is urgent and collective. Journals must open dedicated sections. Funders must require and reward failure reporting. Researchers must publish failures with the same rigor as successes. The community must build and maintain the Materials AI Failure Registry as shared infrastructure.

Only by treating negative results as valuable contributions rather than career liabilities can Materials AI mature into a self-correcting, efficient, and trustworthy discipline. The field has already generated enough positive claims; what it now desperately needs is an honest record of what does not work. Embracing failure reporting is not a retreat from excellence—it is the foundation for sustainable acceleration of materials discovery that society requires. The time to act is now.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 Mar 2025

Revised: 30 Jun 2025

Accepted: 17 Oct 2025

Published online: 18 January 2026

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

- Mlinarić A, Horvat M, Šupak Smolčić V. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochem Med (Zagreb)*. 2017;27(3):030201.
<https://doi.org/10.11613/BM.2017.030201>.

Heesen R, Bright LK. Publication bias is bad for science if not necessarily scientists. *R Soc Open Sci.* 2025;12(4):240688.
<https://doi.org/10.1098/rsos.240688>.

Otyepka M, Pykal M, Otyepka M. Advancing materials discovery through artificial intelligence. *Appl Mater Today.* 2025;47:102981.
<https://doi.org/10.1016/j.apmt.2025.102981>.

Jain A. Machine learning in materials research: Developments over the last decade and challenges for the future. *Curr Opin Solid State Mater Sci.* 2024;33(1):101189.
<https://doi.org/10.1016/j.cossms.2024.101189>.

DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific AI in materials science: A path to a sustainable and scalable paradigm. *Mach Learn Sci Technol.* 2020;1(3):033001.

Olivetti EA, Cole JM, Kim E, Kononova O, Ceder G, Han TYJ, et al. Data-driven materials research enabled by natural language processing and information extraction. *Appl Phys Rev.* 2020;7(4):041317.
<https://doi.org/10.1063/5.0021106>.

Boyce B, Dingreville R, Desai S, Walker E, Shilt T, Bassett KL, et al. Machine learning for materials science: Barriers to broader adoption. *Matter.* 2023;6(5):1320-3.
<https://doi.org/10.1016/j.matt.2023.03.028>.

Persaud D, Ward L, Hattrick-Simpers J. Reproducibility in materials informatics: Lessons from 'A general-purpose machine learning framework for predicting properties of inorganic materials'. *Digit Discov.* 2024;3(2):281-6.
<https://doi.org/10.1039/D3DD00199G>.

Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, et al. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *J Mach Learn Res.* 2021;22(164):1-20.

Semmelrock H, Ross-Hellauer T, Kopeinik S, Theiler D, Haberl A, Thalmann S, et al. Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Mag.* 2025;46(2):e70002.
<https://doi.org/10.1002/aaai.70002>.

Yelgel ÖC, Yelgel C. A review of machine learning approaches for the discovery of thermoelectric materials. *Adv Phys X.* 2025;10(1):2536269.
<https://doi.org/10.1080/23746149.2025.2536269>.

Huang L, Zhang X, Li S, Xie R. Text mining-assisted machine learning prediction and experimental validation of emission wavelengths. *NPJ Comput Mater.* 2026;12:98.
<https://doi.org/10.1038/s41524-026-01967-5>.

Zhou ZH. Machine learning. Singapore: Springer Singapore; 2021.
<https://doi.org/10.1007/978-981-15-1967-3>.

Wu Y, Liu L, Deng W, Jiang H, Li R, Hu M, et al. Data distribution matters: Enhancing model generalization in data-driven materials discovery by intentionally exploiting both positive and negative data. *Acta Physico-Chim Sin.* 2026:100291.
<https://doi.org/10.1016/j.actphy.2026.100291>.

Brazil R. Illuminating 'the ugly side of science': Fresh incentives for reporting negative results. *Nature.* 2024.
<https://doi.org/10.1038/d41586-024-01389-7>.

Cranford S. Want for nothing, need for null, useful output from negative results. *Matter.* 2024;7(5):1679-83.
<https://doi.org/10.1016/j.matt.2024.04.005>.

Bespalov A, Steckler T, Skolnick P. Be positive about negatives—recommendations for the publication of negative (or null) results. *Eur Neuropsychopharmacol.* 2019;29(12):1312-20.

<https://doi.org/10.1016/j.euroneuro.2019.10.007>.

Toniato A, Vaucher AC, Laino T, Graziani M. Negative chemical data boosts language models in reaction outcome prediction. *Sci Adv.* 2025;11(24):eadt5578.

<https://doi.org/10.1126/sciadv.adt5578>.

Yin X, Ma J, Zhang S, Cui G. Machine learning drives a new paradigm in inorganic solid-state electrolytes research. *Comput Mater Today.* 2026;10:100051.

<https://doi.org/10.1016/j.commt.2026.100051>.

Karl F, Kemeter LM, Dax G, Sierak P. Position: Embracing negative results in machine learning. *arXiv [Preprint]*. 2024;arXiv:2406.03980.

<https://doi.org/10.48550/arXiv.2406.03980>.

Zhuang Y, Yang X, Zhang C, Jia X, Zhang D, Li M, et al. Materials databases: Foundations of modern digital materials. *Precis Chem.* 2026.

<https://doi.org/10.1021/prechem.5c00449>.

Lau ML, Burleigh A, Terry J, Long M. Materials characterization: Can artificial intelligence be used to address reproducibility challenges? *J Vac Sci Technol A.* 2023;41(6):060801.

<https://doi.org/10.1116/6.0002809>.

Bruckner T, Wieschowski S, Heider M, Deutsch S, Drude N, Tölch U, et al. Measurement challenges and causes of incomplete results reporting of biomedical animal studies: Results from an interview study. *PLoS One.* 2022;17(8):e0271976.

<https://doi.org/10.1371/journal.pone.0271976>.

Colombo M, Duev G, Nuijten MB, Sprenger J. Statistical reporting inconsistencies in experimental philosophy. *PLoS One.* 2018;13(4):e0194360.

<https://doi.org/10.1371/journal.pone.0194360>.

Maloney MP, Coley CW, Genheden S, Carson N, Helquist P, Norrby PO, et al. Negative data in data sets for machine learning training. *Org Lett.* 2023;25(17):2945-7.

<https://doi.org/10.1021/acs.orglett.3c01282>.

Salvador RB, Cavallari DC, Rands D, Tomotani BM. Publication practice in taxonomy: Global inequalities and potential bias against negative results. *PLoS One.* 2022;17(6):e0269246.

<https://doi.org/10.1371/journal.pone.0269246>.

Alpaydin E. *Machine learning*. Rev updated ed. Cambridge (MA): MIT Press; 2021. 280 p.

Reau M, Lagarde N, Zagury JF, Montes M. Nuclear receptors database including negative data (NR-DBIND): A database dedicated to nuclear receptors binding data including negative data and pharmacological profile: Miniperspective. *J Med Chem.* 2019;62(6):2894-904.

<https://doi.org/10.1021/acs.jmedchem.8b01105>.

Murty S, Russell RR. Bad outputs. In: Ray SC, Chambers RG, Kumbhakar SC, editors. *Handbook of production economics*. Singapore: Springer; 2022. p. 483-535.

https://doi.org/10.1007/978-981-10-3455-8_3.