ORIGINAL RESEARCH **Open access**

Redefining "Reproducibility" in Data-Driven Materials Synthesis: Between Algorithmic and Experimental Replication

Fatima Al-Zahra[1*], Amina El Idrissi[1]

Abstract

Reproducibility is a cornerstone of scientific integrity, yet in the rapidly expanding field of data-driven materials synthesis, the term is used inconsistently and often misleadingly. Authors frequently claim “reproducibility” without specifying whether they refer to computational verification of models or successful laboratory synthesis of predicted materials. This boundary/Definitional article clarifies that “reproducibility” in this domain encompasses two fundamentally distinct and orthogonal concepts: algorithmic reproducibility and experimental replication. Algorithmic reproducibility is defined as the ability to obtain identical numerical results (within documented floating-point tolerances) when the same code is executed on the same data in the same computational environment. Experimental replication, by contrast, is the ability of an independent laboratory to synthesize the same material—within clearly defined characterization tolerances—by strictly following the published synthesis protocol. The article demonstrates that these two forms of reproducibility operate in separate ontological domains (computational versus chemical) and are frequently conflated in the literature, leading to overclaims, misdirected research effort, and erosion of community trust. Through a systematic boundary analysis, four-level assessment scales are established for each concept, a comparative framework (including an interaction matrix) is presented, and common gray zones and boundary cases are examined. The analysis shows that algorithmic reproducibility is necessary but insufficient for validating materials predictions, while experimental replication is necessary but insufficient for validating the underlying machine-learning pipeline. Only explicit reporting of both at defined levels constitutes a complete and trustworthy claim. This framework provides authors, reviewers, journals, and funders with a precise, operational vocabulary and a practical two-part Reproducibility Declaration standard. Its adoption will reduce ambiguity, strengthen the credibility of AI-driven materials discovery, and accelerate the reliable translation of computational predictions into verifiable laboratory outcomes. The distinctions introduced here are essential for maturing data-driven materials science into a robust, reproducible discipline.

Keywords Materials informatics, Machine learning, Reproducibility, Data-driven materials synthesis, Algorithmic reproducibility, Experimental replication

*Correspondence:
Fatima Al-Zahra
fatima.zahra@gmail.com
[1] Department of Materials Informatics and Engineering, Faculty of Sciences and Engineering, University of Casablanca, Casablanca, Morocco

Introduction

“Reproducibility is the foundation of science.” This statement appears, in one form or another, in the opening paragraphs of nearly every manuscript that applies machine learning to materials synthesis [1-6]. Authors assert that their models, workflows, or recommended synthesis routes are reproducible, and reviewers and editors expect such claims. Yet the term “reproducibility” itself remains remarkably elastic in the literature on data-driven materials engineering.

Does reproducibility mean that another researcher can download the code, load the identical dataset, and obtain the same numerical output to within floating-point precision? Or does it mean that an independent laboratory, following the written synthesis instructions, can produce a material whose X-ray diffraction pattern, scanning electron micrograph, and measured properties match those reported in the original paper? These are not rhetorical questions. They represent two entirely different scientific guarantees—one computational, the other chemical—and the field has treated them as interchangeable for too long [7-9].

The consequences of this conflation are already visible. Computational papers release GitHub repositories and celebrate "reproducibility" while the suggested synthesis routes fail when attempted elsewhere [10, 11]. Experimental papers provide detailed protocols and claim "reproducible synthesis" yet offer no public code or training data that would allow others to verify the underlying machine-learning predictions [12]. In both cases the label "reproducible" is applied, but the guarantee delivered to the community is partial at best.

The conceptual separation and hierarchical relationship between algorithmic reproducibility and experimental replication are illustrated in **Figure 1**.

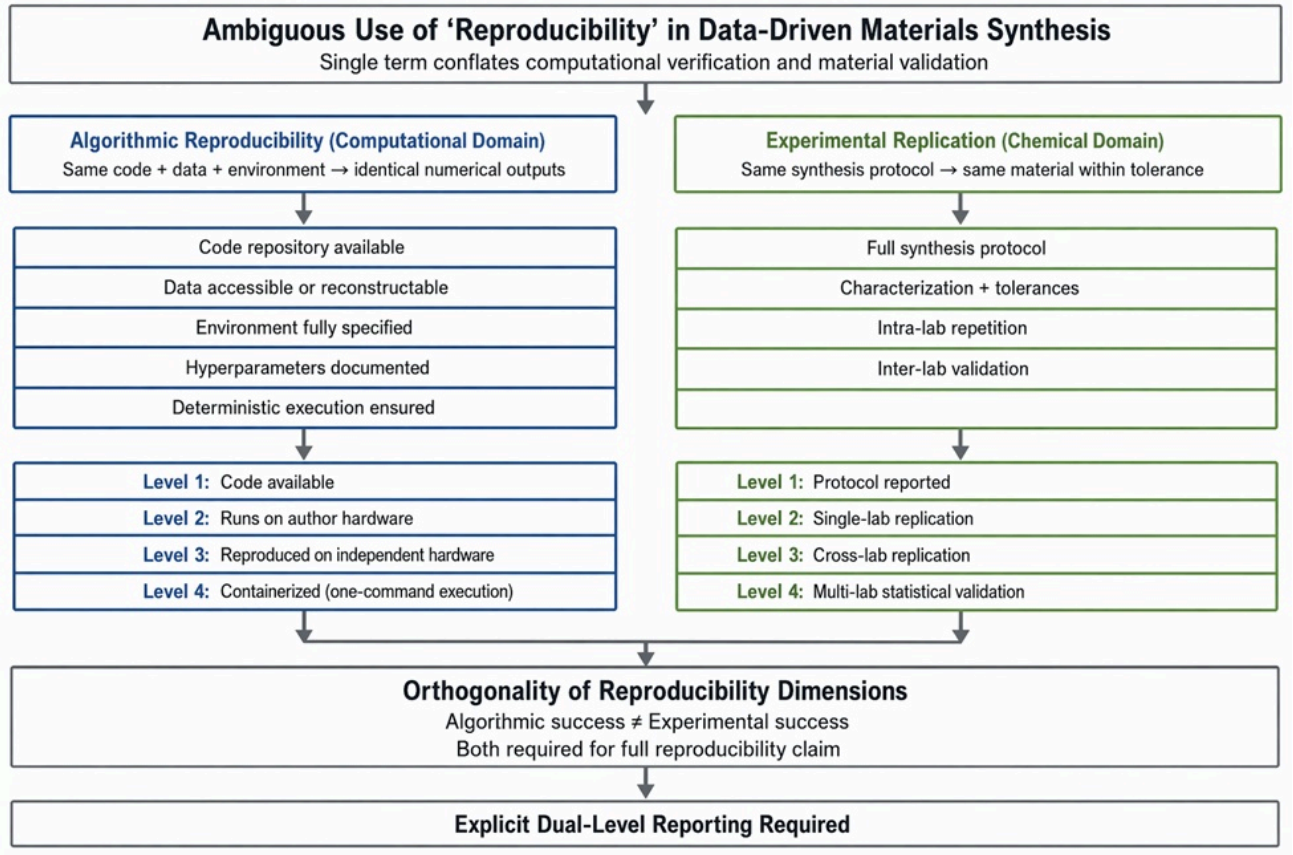

Figure 1. Hierarchical Separation of Algorithmic Reproducibility and Experimental Replication in Data-Driven Materials Synthesis

This article performs a boundary/Definitional analysis of reproducibility specifically for data-driven materials synthesis. It distinguishes algorithmic reproducibility—the strict computational notion of identical numerical results—from experimental replication—the chemical notion of obtaining the same material. The analysis is conceptual rather than empirical: no new datasets are introduced, no models are trained, and no synthesis experiments are performed. Instead, the paper maps the conceptual territory, proposes operational definitions, and establishes explicit boundary conditions for each type of claim.

Section 2 documents how the four most common usages of “reproducibility” in the current literature create systematic confusion. Sections 3 and 4 present the two core definitions—algorithmic reproducibility and experimental replication—together with the practical requirements and limitations of each. Section 5 examines the gap that opens when one form is achieved without the other and illustrates why both must be reported separately if the field is to move beyond overclaims. By the end of Part 1, the reader will possess a precise vocabulary and a set of testable criteria that can be applied immediately to any data-driven materials synthesis paper.

Current Usage and Confusions

The literature on machine learning for materials synthesis deploys the term “reproducible” in overlapping yet distinct senses, each carrying a different implicit guarantee that papers seldom make explicit. In computational contributions the word most often signals that code and training data have been deposited in a public repository so that the reported numerical results can, in principle, be regenerated [13, 14]. This usage secures algorithmic reproducibility—identical inputs yield identical outputs within floating-point limits—yet offers no assurance that the predicted synthesis route will succeed in the laboratory.

A related but separate claim appears in experimental and hybrid studies, where authors assert that “our synthesis procedure is reproducible” after tabulating precursor concentrations, temperatures, stirring times, and atmospheres [10, 12]. Here the guarantee, when realized, is experimental replication: an independent laboratory can obtain the target material within acceptable characterization tolerances. In practice such assertions typically rest on single-laboratory validation and rarely incorporate interlaboratory verification.

When attention shifts to the model itself, “our model is reproducible” usually denotes that retraining the neural network or random forest on the same data and hyperparameters produces statistically comparable performance metrics [7, 13]. This statistical reproducibility operates within an often-unspecified tolerance and reveals nothing about whether the model’s predictions will translate into viable syntheses [15]. The most ambiguous formulation—“results are reproducible”—can invoke any or all of the preceding meanings, leaving readers to infer the author’s intent [9, 16].

These semantic overlaps generate substantive misunderstandings. Computational papers that release code and claim reproducibility [13] are frequently read by experimentalists as validating the recommended syntheses, while experimental protocols [12] are taken by computationalists as evidence that the underlying models are openly verifiable. Neither inference is warranted. Perfect algorithmic reproducibility can coexist with consistent laboratory failure, just as robust experimental replication can rest on unverifiable machine-learning foundations.

Recent reviews have begun to document these inconsistencies [9, 17, 18], yet the community still lacks a shared operational vocabulary. Absent explicit definitions and boundary conditions, the label “reproducible” functions more as rhetorical currency than as a reliable scientific guarantee.

To clarify the systematic ambiguity in current terminology, **Table 1** dissects the four dominant uses of ‘reproducibility’ and their implicit guarantees.

Table 1. Disambiguation of Reproducibility Terminology in Current Literature

Term Used in Literature	Actual Meaning	Scientific Domain	Implicit Guarantee	Common Misinterpretation	Resulting Risk
“Code is reproducible”	Algorithmic reproducibility	Computational	Identical numerical outputs	Assumed synthesis validity	Failed lab replication

“Synthesis is reproducible”	Experimental replication	Chemical	Same material produced	Assumed model validity	Unverified predictions
“Model is reproducible”	Statistical reproducibility	Computational	Similar performance metrics	Assumed physical correctness	Overgeneralization
“Results are reproducible”	Ambiguous	Mixed	Undefined	Assumed full reproducibility	Conceptual confusion

The following two sections supply the missing definitions and the criteria needed to evaluate claims rigorously.

Algorithmic Reproducibility

Algorithmic reproducibility is defined here as the ability to obtain identical numerical results (within documented floating-point tolerance) when the same code is executed on the same data with the same computational environment.

Achieving this form of reproducibility requires five concrete elements: (1) the complete source code must be publicly available in a version-controlled repository; (2) the exact training and test datasets must be available or described with sufficient precision that they can be reconstructed; (3) the computational environment—operating system, library versions, CUDA driver, random seeds—must be fully specified; (4) all hyperparameters must be listed; and (5) all non-deterministic operations (for example, GPU-based stochasticity) must be replaced by deterministic equivalents or explicitly documented [19-24].

When these conditions are met, algorithmic reproducibility guarantees that other researchers can verify the computational results and can build directly on the published numbers. It does not guarantee that the model generalizes to new compositions, nor does it guarantee that the suggested synthesis procedure will produce the target material in the laboratory.

Experimental Replication

Experimental replication is defined here as the ability to synthesize the same material—within clearly stated characterization tolerances—when an independent laboratory follows the identical written synthesis procedure.

The definition requires four elements: (1) a detailed, step-by-step protocol that specifies precursors, concentrations, temperatures, times, atmospheres, and equipment settings; (2) explicit characterization methods and acceptance tolerances (for example, XRD peak positions ±0.05°, SEM particle-size distribution within 10 %); (3) statistical validation from multiple independent trials within each laboratory; and (4) interlaboratory validation involving at least two independent research groups.

When these conditions are satisfied, experimental replication guarantees that the material can be made by others and that the synthesis route is sufficiently robust for practical use. It does not guarantee that the original machine-learning model predicted the outcome correctly, nor does it guarantee that the material possesses every target property claimed in the computational study.

Four practical levels mirror those defined for algorithmic reproducibility: • Level 1 (protocol reported): synthesis conditions are described in the paper or supplementary information. • Level 2 (single-lab replication): the original laboratory has successfully repeated the synthesis and confirmed the material identity. • Level 3 (cross-lab replication): at least one independent laboratory has reproduced the material within the stated tolerances. • Level 4 (multi-lab statistical): three or more laboratories have performed the synthesis, and statistical measures of variability have been reported.

Current practice in data-driven materials papers overwhelmingly stops at Level 1 [10, 12]. Detailed protocols are published, but independent replication—especially cross-laboratory replication—is reported only occasionally. The gap between Level 1 and Level 3 is particularly problematic because a protocol that works perfectly in one laboratory may fail elsewhere owing to subtle differences in humidity, precursor purity, or equipment calibration. Only Level 3 or 4 provides the chemical trustworthiness the community requires.

Experimental replication is therefore a necessary but never sufficient condition for a complete reproducibility claim. It certifies the robustness of the synthesis route; it cannot certify the correctness of the underlying computational model.

The structural differences between the two reproducibility dimensions are formalized in **Table 2**, highlighting their orthogonality.

Table 2. Comparative Structural Requirements for Algorithmic Reproducibility and Experimental Replication

Dimension	Algorithmic Reproducibility	Experimental Replication
Ontological domain	Computational (data, code, execution)	Chemical (materials, reactions, synthesis)
Core objective	Numerical identity	Material equivalence
Primary inputs	Code, dataset, environment	Protocol, precursors, equipment
Validation criterion	Bitwise or tolerance-level output match	Characterization within tolerance
Sources of variability	Hardware, randomness, software versions	Environment, purity, equipment calibration
Failure mode	Non-runnable or inconsistent outputs	Irreproducible synthesis outcome
Verification method	Independent execution	Independent laboratory replication
Sufficiency for full reproducibility	Necessary but insufficient	Necessary but insufficient

The Gap between Algorithmic and Experimental Reproducibility

Algorithmic reproducibility and experimental replication are conceptually orthogonal. One concerns bits and floating-point numbers; the other concerns atoms and reaction kinetics. Perfect algorithmic reproducibility can coexist with complete failure of experimental replication, and vice versa.

Consider a paper that releases fully containerized code (algorithmic Level 4) and predicts a novel superconductor composition. Every reviewer can regenerate the exact prediction numbers, yet when five independent laboratories attempt the suggested synthesis, none obtains the target phase [9, 11]. Algorithmic reproducibility has been achieved; experimental replication has not. The computational claim is valid, but the materials claim is not.

Conversely, consider a paper that supplies an exhaustive synthesis protocol for a metal-organic framework and demonstrates successful replication across three laboratories (experimental Level 3). The material is reproducibly made, yet the authors withhold the training data and model weights. Experimental replication has been achieved; algorithmic reproducibility has not. The chemical claim is valid, but the computational claim cannot be verified.

These examples are not hypothetical edge cases; they reflect patterns documented across the literature [7, 10, 12]. The dangerous consequence is that readers treat the single word “reproducible” as a blanket guarantee. A computational

group assumes the synthesis will work; an experimental group assumes the model can be trusted. When reality contradicts the assumption, trust in the entire data-driven pipeline erodes.

The interaction between algorithmic and experimental levels—and its implications for scientific validity—is synthesized in **Table 3**.

Table 3. Reproducibility Level Interaction Matrix and Interpretation of Scientific Claims

Algorithmic Level	Experimental Level	Interpretation of Claim	Scientific Strength	Practical Reliability
Level 4	Level 1	Fully verifiable computation, unvalidated synthesis	Moderate	Low
Level 3	Level 3	Verified computation and cross-lab validated synthesis	High	High
Level 1	Level 3	Robust material synthesis, unverifiable model	Moderate	High
Level 2	Level 2	Partial verification in both domains	Low–Moderate	Moderate
Level 4	Level 4	Fully reproducible pipeline (ideal standard)	Very High	Very High

The field therefore requires both forms of reproducibility for any paper that claims to advance materials synthesis via machine learning.

Algorithmic reproducibility ensures that the prediction engine is verifiable. Experimental replication ensures that the predicted material can actually be made. Only when both are reported at explicit levels can a paper be considered fully reproducible in the data-driven materials context.

The remaining sections of the manuscript will explore boundary cases, relations to adjacent concepts, practical implications, a proposed reporting standard, and a call for community adoption.

Boundary Cases and Gray Zones

Even with clear definitions, boundary cases expose the persistent friction between algorithmic reproducibility and experimental replication, revealing why a single label “reproducible” remains fundamentally insufficient. When code is fully containerized at the highest algorithmic level yet training data stay private or proprietary [13, 14], exact regeneration of computational results becomes possible across machines, but the absence of data prevents model retraining or inspection, rendering the algorithmic claim hollow even if an experimental protocol appears sufficient for replication.

This tension sharpens when detailed synthesis protocols omit critical micro-scale parameters such as exact stirring speeds, ramp rates, or precursor lot numbers [10, 12]. Although the protocol meets nominal experimental standards on paper, independent laboratories often fail to obtain the target phase, demonstrating that flawless algorithmic reproducibility of the underlying model can coexist with irreproducible chemical outcomes.

Compounding the difficulty is the intrinsic stochasticity of nucleation and growth processes, which precludes exact atomic-level replication and necessitates domain-specific tolerances—such as phase purity above 95 % by XRD and particle size distributions within ±15 % [11, 12]. Without such agreed thresholds, identical protocols can yield divergent declarations of success or failure across laboratories.

Further complications arise when equipment calibration or precursor purity varies between sites [10], allowing the originating laboratory to achieve consistent results while others require minor, often undocumented adjustments whose legitimacy remains undefined absent community consensus on acceptable deviations.

In the most common practical scenario, synthesis batches exhibit partial success under nominally identical conditions [12], rendering experimental replication inherently statistical rather than deterministic and compelling a shift from absolute claims to precise reporting of success rates and variability.

These boundary conditions underscore that reproducibility in data-driven materials synthesis is rarely binary. Authors must therefore specify which dimensions are achieved while openly acknowledging inherent gray zones, rather than defaulting to the ambiguous umbrella term “reproducible.”

Relation to Other Concepts

The proposed framework connects directly to several adjacent concepts in the literature without collapsing them.

Relation to open science: Butler and colleagues [17] called for setting standards for data-driven materials science and emphasized open workflows. This article builds on that call by distinguishing workflow reproducibility (algorithmic) from materials reproducibility (experimental). Open code and data satisfy the former; open, validated protocols satisfy the latter [25-27]. Both are required for genuine open science in this domain.

Relation to model cards and reporting standards: Kapoor *et al.* [16] introduced REFORMS recommendations for machine-learning-based science. The present framework extends that work by requiring explicit reproducibility-level declarations rather than generic model cards [28]. A model card can document algorithmic Level 4 while simultaneously stating that experimental replication remains at Level 1.

Relation to dataset bias and leakage: Kapoor and Narayanan [9] documented how leakage creates a reproducibility crisis in machine-learning-based science. Biased or non-public datasets undermine algorithmic reproducibility because results cannot be independently regenerated [13]. Even perfect code fails the algorithmic test if data are missing. Experimental replication, however, can still be achieved if the protocol is chemically robust, illustrating once again that the two concepts are orthogonal.

Relation to failure reporting: Recent position pieces [11] have urged the community to publish negative results and failed replications. The boundary definitions supplied here provide the precise language needed: authors can now state “algorithmic reproducibility achieved at Level 3; experimental replication failed at Level 1 despite multiple attempts.” Such transparent reporting converts failures into structured knowledge rather than hidden embarrassment.

By anchoring these relations in explicit definitions and levels, the framework prevents the vague umbrella term “reproducibility” from absorbing every adjacent good practice. It forces specificity while remaining compatible with existing open-science and machine-learning standards.

Implications for Practice

The boundary definitions carry immediate consequences for every stakeholder.

For computational researchers: State explicitly which reproducibility is being claimed and at what level. Do not write “our results are reproducible” without adding “algorithmic reproducibility: Level 3; experimental replication: Level 1.” Achieve at least Level 3 for algorithmic claims by ensuring independent execution produces identical numbers. Containerization (Level 4) should become the default for new submissions [13, 14].

For experimental researchers: Treat protocol reporting as the minimum (Level 1) rather than the finish line. Perform and document single-lab replication (Level 2) before publication. Where feasible, collaborate on cross-lab replication (Level 3). Report statistical variability rather than single successful batches [10, 12].

For journals and editors: Replace the generic “reproducibility” checkbox with a required two-part declaration. Introduce separate badges: one for algorithmic reproducibility (green for Level 3–4) and one for experimental replication (green for Level 3–4). Reject manuscripts that use the unqualified word “reproducible” without level specification. Digital Discovery and npj Computational Materials are well positioned to lead this change [17].

For funders: Require a reproducibility plan in every proposal that distinguishes the two forms. Prioritize funding for interlaboratory replication studies and for maintenance of public containers and protocols. Treat Level 1 claims as preliminary and demand evidence of higher levels for follow-on funding.

Collective adoption of these practices will reduce overclaims, minimize wasted synthesis attempts, and restore confidence that a data-driven materials paper actually delivers what it promises.

Proposed Reporting Standard

All data-driven materials synthesis papers should include a mandatory Reproducibility Declaration placed in either the methods or conclusions section. This declaration must explicitly separate two distinct dimensions—algorithmic reproducibility and experimental replication—and report the precise level achieved for each, supported by concrete evidence.

For algorithmic reproducibility, authors must specify the level achieved on a four-point scale and provide verifiable proof. This includes, at minimum, a link to a public repository containing code and data, details of the computational environment (such as a container image where applicable), and confirmation that the results can be independently executed. The levels are defined as follows: Level 1 indicates that code is available; Level 2 confirms that the code runs on the authors’ hardware; Level 3 demonstrates that identical numerical results can be reproduced on independent hardware; and Level 4 requires full containerization enabling one-command reproduction.

For experimental replication, authors must likewise report the level achieved using a parallel four-point scale, along with supporting documentation. Evidence should include the location of the full experimental protocol, whether replication has occurred within a single laboratory or across multiple independent laboratories, and appropriate statistical measures of variability where applicable. The levels are defined as: Level 1 for a fully reported protocol; Level 2 for successful replication within a single lab; Level 3 for replication across independent labs; and Level 4 for multi-laboratory validation with statistical rigor.

A typical declaration in a hybrid computational–experimental paper might read as follows:

“Algorithmic reproducibility: Level 3 (code and data available in a public repository; results independently reproduced on separate hardware within a tolerance of 1e-8). Experimental replication: Level 1 (complete synthesis protocol provided in supplementary materials; no independent laboratory replication yet conducted).”

This structured declaration compels authors to clearly identify strengths and limitations rather than obscuring them under a generic claim of “reproducibility.” It enables reviewers to quickly evaluate whether both computational and experimental dimensions are adequately addressed. Journals, in turn, gain a standardized and machine-readable signal of reproducibility quality. Over time, this clarity will support more accurate interpretation of contributions: for example, work achieving algorithmic Level 4 but experimental Level 1 may represent a mature computational advance but an early-stage experimental result, whereas achieving Level 3 in both dimensions signals a robust and well-validated contribution.

Adopting this standard removes the longstanding ambiguity surrounding the term “reproducible,” ensuring that its meaning is precise, comparable, and actionable across the materials synthesis community.

Conclusion

“Reproducibility” has become the most frequently claimed yet least precisely defined term in data-driven materials synthesis. This boundary/Definitional analysis demonstrates that the single word actually conflates two distinct concepts: algorithmic reproducibility (same code, same data, same environment → same numerical result) and experimental replication (same synthesis procedure → same material).

Operational definitions and four-level boundary scales now exist for both. Algorithmic reproducibility certifies computational integrity; experimental replication certifies chemical robustness. The two can—and frequently do—exist independently, creating a dangerous gap when readers assume both are present. Boundary cases and gray zones further illustrate that reproducibility is rarely binary and must be reported with explicit levels and tolerances.

The framework connects cleanly to open-science standards, machine-learning reporting recommendations, leakage analysis, and calls for negative-result publication. Practical implications extend to every actor: computational and experimental researchers must specify levels, journals must enforce separate declarations, and funders must support interlaboratory validation.

The proposed reporting standard supplies a concrete, low-burden mechanism: a two-row table that separates the two forms of reproducibility and states the exact level achieved. Its adoption will replace rhetorical claims with verifiable guarantees.

The field stands at a crossroads. Continued use of the ambiguous term “reproducible” will perpetuate overclaims, failed replications, and eroded trust. Precise language and explicit levels, by contrast, will accelerate reliable translation of machine-learning predictions into laboratory reality and restore reproducibility to its rightful place as the verifiable foundation of data-driven materials science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Mar 2025 Revised: 21 Jun 2025 Accepted: 14 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

Richard C, Pichon PA, Nováková J, Irannejad Najafabadi N, Sá J. Application of machine learning to nanomaterial design with silver nanoprisms. Discov Nano. 2026;21(1):65.
https://doi.org/10.1186/s11671-026-04503-y.

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.

Xu RHJ. Workflows and automated platforms for nanocrystal synthesis, purification, and characterization [dissertation]. Urbana (IL): University of Illinois Urbana-Champaign; 2025.

Callaghan S. Toward machine learning-enhanced high-throughput experimentation for chemistry. Patterns (N Y). 2021;2(3):100221.
https://doi.org/10.1016/j.patter.2021.100221.

Glasby LT, Moghadam PZ. Hydrogen storage in MOFs: Machine learning for finding a needle in a haystack. Patterns (N Y). 2021;2(7):100305.
https://doi.org/10.1016/j.patter.2021.100305.

Miran W, Imamura G, Okamoto A. Architecting data-driven microbial electrochemistry from scratch. Patterns (N Y). 2022;3(11):100637.
https://doi.org/10.1016/j.patter.2022.100637.

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.

Pouchard L, Lin Y, Van Dam H. Replicating machine learning experiments in materials science. In: Parallel computing: Technology trends. Amsterdam: IOS Press; 2020. p. 743-55.
https://doi.org/10.3233/APC200105.

Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns (N Y). 2023;4(9):100804.
https://doi.org/10.1016/j.patter.2023.100804.

Yu X, Dutta S, Andreo J, Wuttke S. Identifying synthetic variables influencing the reproducible microfluidic synthesis of ZIF nano- and micro-particles. Commun Mater. 2024;5(1):210.
https://doi.org/10.1038/s43246-024-00644-8.

Sun W, David N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. Faraday Discuss. 2025;256:614-38.
https://doi.org/10.1039/D4FD00112E.

Hemmer K, Boström HLB, Krause S, Lotsch BV, Fischer RA. Strategies to achieve reproducible synthesis of phase-pure Zr-porphyrin metal-organic frameworks. Commun Mater. 2024;5(1):264.
https://doi.org/10.1038/s43246-024-00690-2.

Li K, DeCost B, Choudhary K, Hattrick-Simpers J. A reproducibility study of atomistic line graph neural networks for materials property prediction. Digit Discov. 2024;3(6):1123-9.
https://doi.org/10.1039/D4DD00064A.

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.

Ho SY, Wong L, Goh WW. Avoid oversimplifications in machine learning: Going beyond the class-prediction accuracy. Patterns (N Y). 2020;1(2):100025.
https://doi.org/10.1016/j.patter.2020.100025.

Kapoor S, Cantrell EM, Peng K, Pham TH, Bail CA, Gundersen OE, et al. REFORMS: Consensus-based recommendations for machine-learning-based science. Sci Adv. 2024;10(18):eadk3452.
https://doi.org/10.1126/sciadv.adk3452.

Butler KT, Choudhary K, Csányi G, Ganose AM, Kalinin SV, Morgan D. Setting standards for data driven materials science. NPJ Comput Mater. 2024;10(1):231.
https://doi.org/10.1038/s41524-024-01411-6.

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.

Leipzig J, Nüst D, Hoyt CT, Soiland-Reyes S, Ram K, Greenberg J. The role of metadata in reproducible computational research. Patterns (N Y). 2021;2(9):100322.
https://doi.org/10.1016/j.patter.2021.100322.

Hartley M, Olsson TSG. dtoolAI: Reproducibility for deep learning. Patterns (N Y). 2020;1(5):100073.
https://doi.org/10.1016/j.patter.2020.100073.

Horton MK, Woods-Robinson R. Addressing the critical need for open experimental databases in materials science. Patterns (N Y). 2021;2(12):100411.
https://doi.org/10.1016/j.patter.2021.100411.

Kuckertz P, Fuchs B, Schönau J, Gardian H, Knosala K, Arellano Ruiz ES, et al. A framework for reproducibly managing coupled research software and data assets based on shared transformation functions. Patterns (N Y). 2026;7(5):101493.
https://doi.org/10.1016/j.patter.2026.101493.

Schackart KE 3rd, Imker HJ, Cook CE. Detailed implementation of a reproducible machine learning-enabled workflow. Data Sci J. 2024;23(1):23.
https://doi.org/10.5334/dsj-2024-023.

Aguilar-Bejarano E, Lea D, Sivakumar K, Mase JM, Omidvar R, Li R, et al. Helix 1.0: An open-source framework for reproducible and interpretable machine learning on tabular scientific data [Preprint]. arXiv:2507.17791. 2025.
https://doi.org/10.48550/arXiv.2507.17791.

Turkyilmaz-van der Velden Y, Dintzner N, Teperek M. Reproducibility starts from you today. Patterns (N Y). 2020;1(6):100099.
https://doi.org/10.1016/j.patter.2020.100099.

Vescovi R, Chard R, Saint ND, Blaiszik B, Pruyne J, Bicer T, et al. Linking scientific instruments and computation: Patterns, technologies, and experiences. Patterns (N Y). 2022;3(10):100606.
https://doi.org/10.1016/j.patter.2022.100606.

Pelser T, Weinand JM, Kuckertz P, Stolten D. ETHOS.REFLOW: An open-source workflow for reproducible renewable energy potential assessments. Patterns (N Y). 2025;6(2):101172.

https://doi.org/10.1016/j.patter.2025.101172.

Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132-5.
https://doi.org/10.1038/s41592-021-01256-7.