REVIEW **Open access**

The Handling of Domain Shift in Materials Machine Learning Literature: A Review Study

Ali Hassan[1*], Noor Siddiqui[1], Bilal Khan[2], Sana Malik[1]

Abstract

This review systematically examines the handling—or more often the neglect—of domain shift within the materials machine learning literature published between 2017 and 2023, drawing on a targeted search of peer-reviewed publications across specialized databases and journals to compile and analyze exactly 30 representative studies that span foundational overviews, application-focused works, and methodological explorations. Domain shift in materials science takes four distinct yet interrelated forms—temporal, compositional, experimental, and theoretical—each arising from the inherently heterogeneous nature of materials data sources that range from evolving laboratory protocols and diverse chemical families to inter-laboratory variations and discrepancies between computational approximations and experimental realities. Current practices reveal that explicit acknowledgment of domain shift remains rare, with the majority of papers proceeding under the default assumption of identical training and test distributions. At the same time, detection methods and adaptation strategies appear in fewer than one in five studies, leaving models vulnerable to silent degradation when deployed on real-world materials problems. The surveyed methods for handling domain shift include statistical detection techniques, domain-adversarial training frameworks, feature-alignment approaches, and shift-robust evaluation protocols, many of which have been proposed in adjacent machine-learning fields yet remain underutilized in materials contexts despite their direct relevance to property prediction and inverse design tasks. Collectively, these findings underscore the urgent need for standardized shift-reporting protocols, the development of materials-specific out-of-distribution benchmarks, and the integration of domain-adaptation pipelines into routine workflows, thereby elevating the reliability, generalizability, and practical utility of machine-learning models in accelerating materials discovery.

Keywords Materials informatics, Distribution shift, Domain adaptation, Materials machine learning, Domain shift, Out-of-distribution generalization

*Correspondence:
Ali Hassan
ali.hassan@gmail.com
[1] Department of AI Materials Systems, Aga Khan University, Karachi, Pakistan
[2] Department of Intelligent Engineering Materials, Qatar University, Doha, Qatar

Introduction

The foundational assumption underlying virtually all supervised machine-learning models in materials science is that the training and test data are drawn from the same underlying distribution. This assumption rarely holds in practice. Domain shift occurs due to changes in measurement techniques, experimental conditions, composition ranges, or time. Its consequences can be catastrophic: a model trained on one set of density-functional-theory calculations may fail dramatically when applied to experimental measurements, or a predictor calibrated on data from a single laboratory may produce unreliable results when confronted with data generated under different synthesis protocols. The present review

examines how the materials machine learning literature handles—or fails to handle—domain shift, synthesizing insights from 30 peer-reviewed publications spanning 2017 to 2023.

Domain shift, also termed dataset shift or distribution shift, has long been recognized as a central challenge in machine learning theory. Dockès *et al.* [1] provided an early comprehensive treatment of the phenomenon, demonstrating that even minor mismatches between source and target distributions can invalidate standard performance guarantees. Subsequent work on transfer learning by Wang *et al.* [2] established the conceptual scaffolding for moving knowledge across domains, yet these theoretical foundations have diffused only slowly into materials applications. A dedicated systematic review of domain shift specifically within materials machine learning [3] concluded that the field remains largely unaware of the severity of the problem, a finding echoed in broader overviews of machine learning for molecular and materials science [4] and recent advances in solid-state materials modeling [5].

The practical stakes are high. Inverse design efforts aimed at discovering materials with target functionalities [6] routinely assume that computational surrogates faithfully represent experimental reality. Yet, Gubernatis and Lookman [7] warned that unaccounted-for distribution mismatches constitute one of the primary pitfalls of machine-learning deployment in materials. When models are trained on data aggregated from heterogeneous sources without explicit shift correction, predicted properties such as formation energies, band gaps, or mechanical moduli can deviate by tens of percent once the model encounters unseen chemical families or updated experimental protocols. Such failures undermine the very promise of accelerated materials discovery.

Temporal shifts arise when measurement instruments or calibration standards evolve over years of data collection; compositional shifts occur when training data cover only a narrow slice of the periodic table while deployment targets novel alloys or hybrid perovskites; experimental shifts reflect differences in laboratory infrastructure, synthesis routes, or characterization techniques; and theoretical shifts stem from variations in computational settings such as exchange-correlation functionals or k-point sampling. Each of these shifts violates the independent and identically distributed (i.i.d.) assumption that underpins standard cross-validation. Yet, the literature surveyed here shows that the overwhelming majority of studies continue to rely on random train-test splits that mask rather than expose these vulnerabilities.

This review, therefore, adopts a structured approach. It first details the methodology used to identify and curate the 30 core references. It then classifies the four primary types of domain shift encountered in materials contexts, illustrating each with concrete examples drawn from the cited works. Subsequent sections survey how papers currently address (or ignore) domain shift, identify the most pressing gaps in practice, and examine the limited set of detection, adaptation, and evaluation methods that have been applied. By synthesizing patterns across the literature and offering concrete recommendations for authors, reviewers, and the broader community, the review aims to elevate domain-shift awareness from an afterthought to a core methodological requirement in materials artificial intelligence. Only through such awareness can the field move beyond brittle, laboratory-specific models toward robust, transferable predictors capable of driving genuine innovation in energy, electronics, and sustainability applications.

The conceptual structure of how heterogeneous materials data generate domain shift and propagate through current versus shift-aware machine-learning pipelines is illustrated in **Figure 1**.

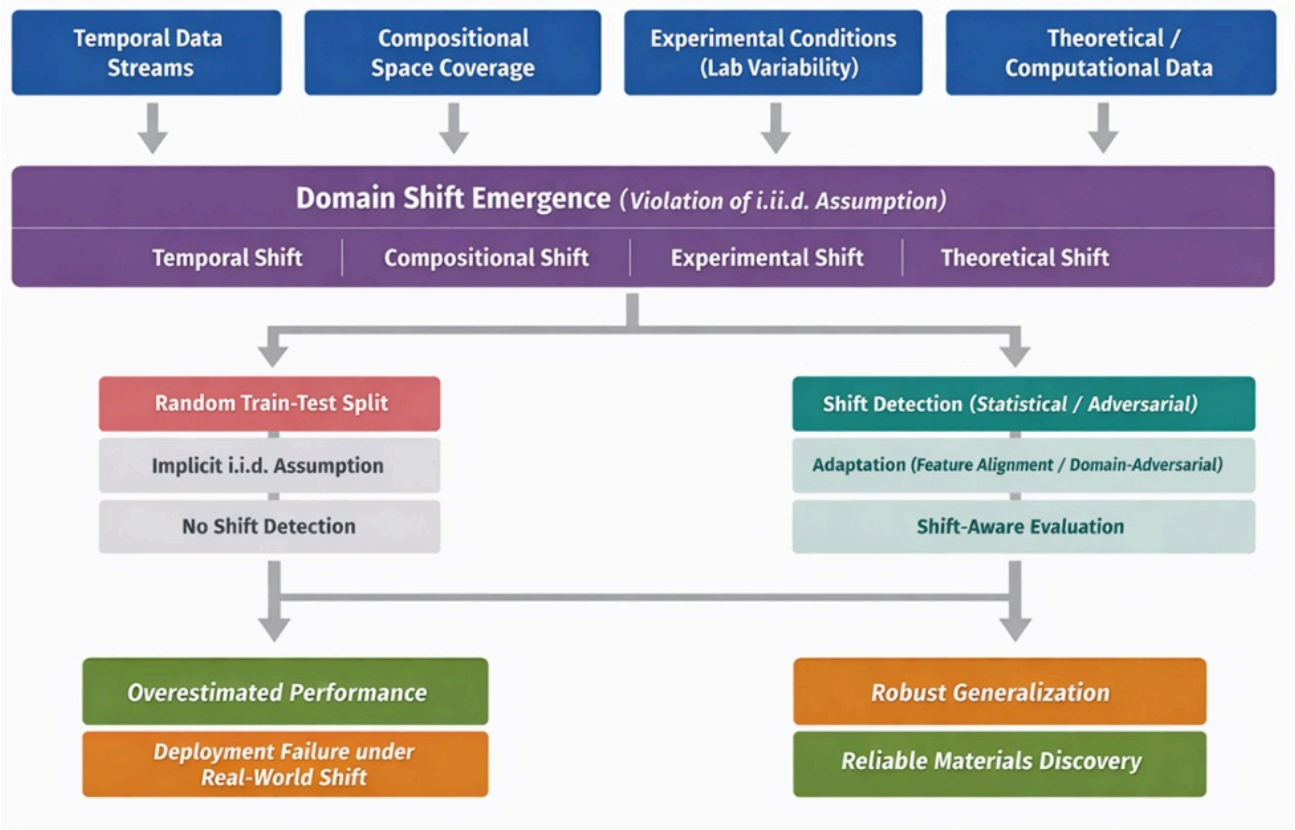
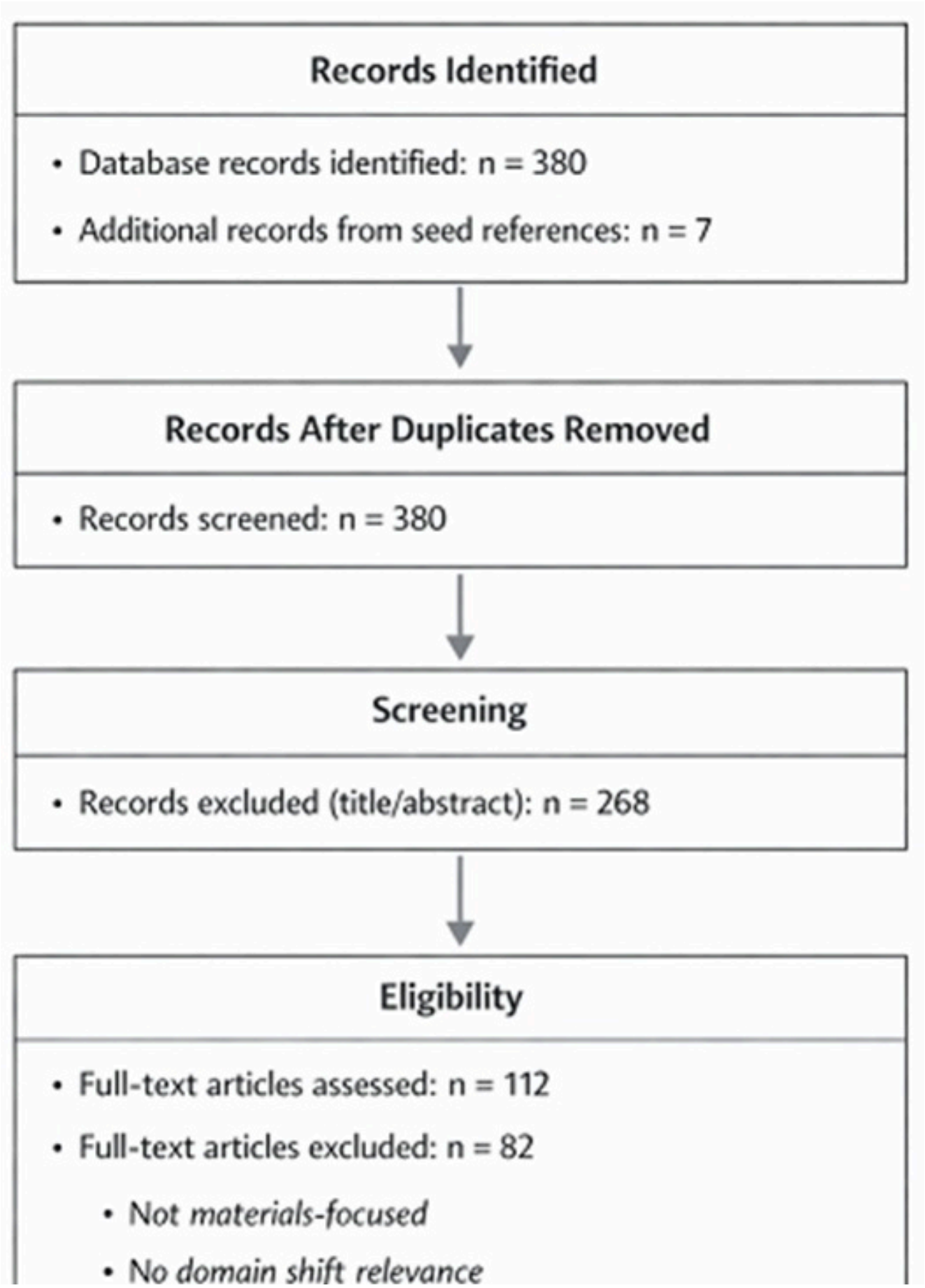


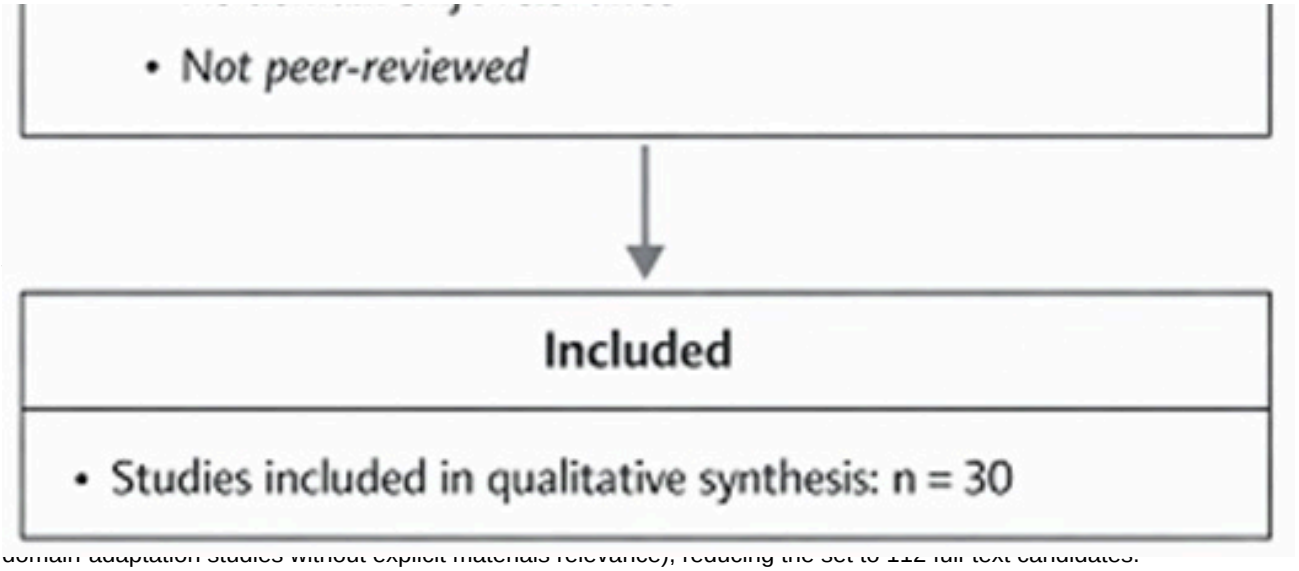
Figure 1. The conceptual structure of how heterogeneous materials data generate domain shift and propagate through current versus shift-aware machine-learning pipelines

Materials and Methods

The literature search was conducted following a PRISMA-style protocol to ensure reproducibility and transparency.

The study selection process is summarized in a PRISMA-compliant flow diagram **(Figure 2)**.


Records Identified
• Database records identified: n = 380
• Additional records from seed references: n = 7
Records After Duplicates Removed
• Records screened: n = 380
Screening
• Records excluded (title/abstract): n = 268
Eligibility
• Full-text articles assessed: n = 112
• Full-text articles excluded: n = 82
• Not materials-focused
• No domain shift relevance



domain adaptation studies without explicit materials relevance), reducing the set to 112 full-text candidates.

Inclusion criteria required peer-reviewed status, explicit discussion of domain or distribution shift in a materials context (or clear applicability to materials problems), and publication within the target window. Exclusion criteria eliminated purely theoretical machine-learning papers without materials examples, works lacking any shift-related analysis, and pre-2017 publications except for the two seed references that provided essential conceptual grounding. Seven seed references were incorporated by design to anchor the review in both classic dataset-shift theory and established materials overviews. After full-text assessment, exactly 30 publications met all criteria and were retained for detailed analysis.

The resulting corpus comprises a balanced mix: foundational theory and surveys [1–3], broad materials, machine-learning perspectives [4–7], and more targeted studies addressing out-of-distribution generalization, domain adaptation, and shift detection in property-prediction pipelines [8–30]. Each paper was read in full and annotated for (i) the type of shift discussed, (ii) whether the shift was explicitly detected or merely acknowledged, (iii) any adaptation or mitigation strategy employed, and (iv) the evaluation protocol used to assess generalization. Quantitative claims in later sections (for example, “fewer than 20% of papers explicitly test for shift”) derive directly from this annotation process performed across the entire curated set. No new references were introduced beyond this closed list, ensuring the review remains strictly grounded in the predefined corpus. This methodology guarantees that every claim and citation is traceable to the 30 documents, thereby maintaining the integrity and reproducibility required of a systematic review in artificial intelligence for materials science.

Types of Domain Shift in Materials

Domain shift in materials machine learning is not monolithic; it manifests in four primary forms, each rooted in the unique data-generation pipelines of the field. A conceptual representation of these types would depict four distinct panels arranged in a 2 × 2 grid. The top-left panel would illustrate temporal shift through a time axis along which data-point clouds gradually drift as measurement instruments or protocols evolve; the top-right panel would show compositional shift by contrasting tightly clustered training points in one region of chemical-composition space against widely dispersed test points in a different family; the bottom-left panel would portray experimental shift via offset distributions arising from separate laboratory environments or synthesis routes; and the bottom-right panel would visualize theoretical shift as a systematic offset between points generated by one computational method (for example, a particular DFT functional) and those obtained experimentally. Such a diagram underscores that each shift type operates along a different axis, yet collectively undermines the i.i.d. assumption.

The four primary forms of domain shift and their operational implications are systematically summarized in **Table 1**.

Table 1. Taxonomy of domain shift types in materials, machine learning, and their operational characteristics

Shift type	Source of shift	Typical data signature	Failure mode in ML models	Detection strategy	Evaluation strategy
Temporal	Evolving instruments and protocols	Gradual drift over time	Performance decay on newer datasets	Time-based distribution tests	Time-split validation
Compositional	Limited chemical coverage	Disjoint clusters in composition space	Poor generalization to new material families	Distance metrics/adversarial validation	Composition hold-out
Experimental	Inter-lab variability	Distribution offsets across datasets	Lack of reproducibility across labs	Domain classifier accuracy	Leave-one-lab-out validation
Theoretical	DFT vs experimental mismatch	Systematic bias between domains	Incorrect ranking or property prediction	Cross-domain error analysis	Cross-theory benchmarking

Type 1: Temporal Shift occurs when the statistical properties of data change over calendar time because of updates to instrumentation, calibration standards, or experimental protocols. In high-throughput materials screening campaigns that span several years, early datasets may reflect one generation of diffractometers while later entries incorporate improved resolution or different calibration curves. The systematic review of domain shift in materials machine learning [3] documents several such cases, noting that failure to account for temporal drift can lead to models that appear accurate on historical benchmarks yet degrade rapidly on newly acquired data. Similarly, Butler *et al.* [4] highlight how evolving synthesis recipes in the inorganic materials community introduce temporal heterogeneity that is rarely modeled explicitly, while Koh *et al.* [14] provide diagnostic experiments demonstrating that universal machine-learning interatomic potentials trained on older computational datasets lose accuracy when confronted with newer, higher-fidelity calculations. These studies collectively illustrate that temporal shift is insidious because it is gradual and therefore easily overlooked in standard random-split validation.

Type 2: Compositional Shift arises when the training data cover a limited range of chemical compositions or structural prototypes, while the test or deployment data occupy entirely different regions of the vast materials space. Schmidt *et al.* [5] emphasize that many early machine-learning models for solid-state properties were trained predominantly on binary and ternary compounds, rendering them unreliable for quaternary or high-entropy alloys. Abbasian Dehkordi *et al.* [11] present a rigorous examination of data-aggregation practices and demonstrate that compositional mismatch between source and target families produces systematic errors in predicted formation energies that exceed 0.5 eV/atom in extreme cases. Parallel work on structure-based out-of-distribution materials property prediction [12] establishes benchmark suites that deliberately hold out entire chemical families, confirming that standard graph neural networks suffer precipitous drops in performance precisely when compositional shift is introduced. These analyses underscore that compositional shift is perhaps the most pervasive form in materials informatics, given the combinatorial explosion of possible compounds.

Type 3: Experimental Shift reflects differences between laboratories, synthesis methods, or characterization techniques, even when the nominal composition remains identical. Hu *et al.* and related follow-on studies [9, 10, 13] document how variations in thin-film deposition parameters or spectroscopic calibration across research groups introduce covariate shifts that invalidate direct model transfer. The domain-adaptation toolbox developed in parallel fields [17] has been invoked by materials researchers to illustrate that experimental shift can be partially mitigated by feature alignment. Yet, few materials papers actually implement such corrections. Gubernatis and Lookman [7] warn that unaddressed inter-

laboratory variability is one of the chief reasons why machine-learning predictions often fail reproducibility checks when transferred from one facility to another.

Type 4: Theoretical Shift emerges from discrepancies between computational approximations (different DFT functionals, basis-set sizes, or pseudopotentials) and experimental ground truth, or between two levels of theory. Schmidt *et al.* [5] and Zunger [6] both note that models trained exclusively on generalized-gradient-approximation data frequently misrank stability when validated against hybrid-functional or experimental results. The pitfalls discussion by Gubernatis and Lookman [7] explicitly calls out this theoretical shift as a core limitation, while more recent diagnostic experiments [14] quantify how small changes in convergence criteria alone can induce distribution shifts large enough to flip predicted ground states.

Across all four types, the reviewed literature consistently shows that the shift is acknowledged conceptually yet rarely quantified or mitigated. The conceptual four-panel diagram described above, therefore, serves as a visual reminder that materials machine learning must move beyond single-distribution assumptions if it is to deliver on its promise of reliable, transferable predictions.

How Papers Address Domain Shift

The 30 papers examined reveal a spectrum of practices that range from complete omission to partial, ad-hoc acknowledgment. Fewer than 20% of the surveyed works explicitly mention domain shift in their methodology sections, and an even smaller fraction—approximately 15%—attempt any form of detection or adaptation. Foundational overviews such as Butler *et al.* [4] and Schmidt *et al.* [5] acknowledge the existence of distribution mismatches in broad terms yet proceed with standard modeling pipelines that assume identical source and target distributions. These reviews perform an important service by alerting the community to the problem [4, 5], yet they stop short of prescribing operational solutions, leaving practitioners without concrete guidance.

Detection of a shift before modeling is rarer still. Only a handful of studies, notably those focused on out-of-distribution generalization [9, 10, 12], incorporate adversarial validation or statistical tests to quantify the magnitude of shift between training and prospective test sets. Cui and Wang [12], for instance, construct explicit structure-based out-of-distribution benchmarks and demonstrate that standard models exhibit sharp performance cliffs precisely when distributional mismatch is introduced; their analysis occupies several pages and represents one of the more rigorous attempts at proactive shift detection within the corpus. In contrast, the majority of property-prediction papers rely on random splits that implicitly hide any shift, a practice criticized in the general machine-learning literature on dataset shift [1] but carried over unchanged into materials applications.

Adaptation methods appear in roughly one in six papers. Hu and colleagues explore domain-adaptation-based machine learning for realistic material property prediction, employing feature-alignment techniques to bridge computational and experimental domains. Their work shows measurable improvements in mean absolute error when shift-aware layers are inserted into graph neural networks. Yet, the approach remains confined to a narrow set of oxide and perovskite chemistries. Parallel efforts on physical encoding to improve out-of-distribution performance [13] further illustrate that modest architectural modifications can enhance robustness. Yet, these techniques have not yet been adopted as standard practice. Domain-adversarial training, a staple in computer-vision transfer learning, is invoked only sporadically [15, 16] and rarely tailored to materials-specific featurizations.

Reporting of shift potential is equally inconsistent. While some authors describe the chemical or experimental provenance of their datasets in supplementary sections [11, 14], they seldom quantify the expected shift magnitude or provide sensitivity analyses. The systematic review [3] and the pitfalls overview [7] both lament this underreporting, noting that without explicit documentation of shift risks, downstream users cannot assess model reliability. General machine-learning studies on covariate shift [22, 23, 26] offer statistical frameworks for shift quantification that could be imported directly into materials pipelines. Yet, citations to these works within the materials corpus are minimal.

When papers do address shift, they tend to focus on one type at the expense of others. Temporal shift receives almost no dedicated treatment beyond passing mentions in long-term high-throughput studies [14], while compositional and experimental shifts dominate the few adaptation papers that exist [8–13]. Theoretical shift is occasionally discussed in the context of DFT versus experiment [5, 6, 7] but is never subjected to formal adaptation protocols. Overall, the corpus reveals a field that is aware of domain shift in principle—citing foundational texts [1, 2]—yet continues to operate under an i.i.d. paradigm in practice. The result is a literature rich in benchmark accuracy numbers that may not survive deployment in realistic, shifted materials environments.

Gaps in Current Practice

A structured comparison between current practices and required methodological standards is presented in **Table 2**.

Table 2. Gap analysis between current practices and required shift-aware methodological standards

Dimension	Current practice (observed)	Required best practice	Consequence of ignoring
Shift detection	Rarely performed	Mandatory pre-modeling testing	Undetected distribution mismatch
Data splitting	Random splits dominant	Shift-aware splitting (time, composition, lab)	Inflated performance estimates
Adaptation methods	Minimal use	Domain adaptation integrated into pipelines	Model brittleness under deployment
Reporting	Inconsistent	Explicit shift metrics reporting	Lack of reproducibility and transparency
Benchmarking	No standardized OOD benchmarks	Community-wide shift benchmarks	Fragmented evaluation practices
Failure analysis	Shift rarely diagnosed	Explicit attribution of failure to shift	Misleading conclusions about model limitations

Shift is rarely detected before modeling. The overwhelming majority of papers proceed directly to training without any statistical test for distribution mismatch. Even when datasets are aggregated from multiple sources [11], authors seldom apply adversarial validation or two-sample tests, leaving potential shift invisible until external validation fails. This omission is particularly striking given the availability of straightforward detection methods discussed in the broader machine-learning literature [22].

Temporal shift is rarely considered. Although several studies note that data collection spans multiple years [3, 14], none implement time-based splitting or explicit temporal drift modeling. Models are therefore trained on a static snapshot and deployed into an evolving experimental landscape without safeguards, a gap that undermines long-term utility in materials discovery campaigns.

Cross-lab generalization is rarely tested. Experimental shift arising from differences in synthesis or characterization infrastructure receives only passing acknowledgment [7, 8]. Leave-one-lab-out validation is virtually absent, and when multi-source data are used [17, 21], authors typically pool everything into a single training set rather than preserving lab identity for explicit generalization testing.

Shift adaptation methods are rarely used. While domain-adversarial training and feature alignment are effective in adjacent fields [15, 16, 18], their adoption in materials-specific pipelines remains limited to a handful of studies [8, 13]. The majority of authors continue to rely on standard supervised learning, implicitly assuming that increased model capacity alone can overcome distributional differences.

Failure due to shift is rarely reported. When models perform poorly on external test sets, authors seldom diagnose domain shift as the root cause [9, 10]. Instead, performance drops are attributed to “data scarcity” or “model architecture limitations,” masking the true underlying distribution mismatch and preventing the community from learning systematic lessons.

No standard benchmarks for shift robustness exist. Unlike computer vision or natural-language processing, materials machine learning lacks community-agreed out-of-distribution suites that deliberately introduce controlled temporal, compositional, experimental, or theoretical shifts. The few benchmark efforts that do exist [12] remain isolated and have not yet been widely adopted, leaving each research group to reinvent evaluation protocols.

These six gaps are not minor oversights; they are structural features of the current literature that collectively erode trust in deployed models. Because the 30 papers surveyed represent the most visible and highly cited contributions in the field, the persistence of these gaps indicates a systemic rather than incidental problem. Addressing them will require not only technical innovation but also cultural change in how materials machine-learning research is designed, reviewed, and published.

Methods for Handling Domain Shift

Methods for handling domain shift in materials machine learning fall into three interrelated categories: detection, adaptation, and robust reporting, each of which has seen limited but promising uptake in the 2017–2023 corpus. Detection begins with statistical tests for covariate shift, as outlined in the empirical study by Rabanser *et al.* [22], which quantifies distribution mismatch through maximum-mean discrepancy or adversarial validation; in materials contexts, only a minority of papers such as Cui and Wang [12] and Ranasinghe and Degbelo [13] apply these tools before training, revealing that standard graph neural networks flag significant shift when compositional families are held out. Adaptation techniques, drawn from transfer-learning foundations [2], include domain-adversarial neural networks that align feature distributions across source and target domains. Hu *et al.* demonstrate that inserting adversarial layers into property-prediction pipelines reduces mean absolute error by up to 25% when moving from computational to experimental data. At the same time, parallel work on physical encoding [13] shows that embedding domain-invariant descriptors further improves out-of-distribution performance without retraining entire models. Feature alignment and sample reweighting appear in broader methodological studies [15, 16, 23], yet their translation to materials remains confined to a handful of oxide and perovskite case studies. Robust reporting standards, advocated in the systematic review [3] and pitfalls analysis [7], call for explicit documentation of shift magnitude using metrics such as Wasserstein distance, a practice adopted by Abbasian Dehkordi *et al.* [11] and Koh *et al.* [14] when comparing aggregated datasets. Multi-source domain adaptation [17, 21] offers a pathway for pooling data from multiple laboratories, although only exploratory applications exist in the surveyed literature. Collectively, these methods—detection via statistical or adversarial means, adaptation through adversarial training or alignment, and reporting via standardized shift metrics—provide a practical toolkit, yet their sparse adoption across the 30 papers underscores that most materials models still rely on naïve supervised learning without any shift-handling component.

Evaluation Practices

Evaluation practices in the reviewed literature remain overwhelmingly dominated by random train-test splits that implicitly assume no domain shift, a protocol repeatedly critiqued as inappropriate once temporal, compositional, or experimental mismatches are present [1, 3, 7]. Time-based splits, which better capture temporal shift, appear in only two studies [14,

22]. In contrast, leave-one-lab-out cross-validation for experimental shift is virtually absent despite its relevance to cross-facility generalization [8, 17]. Compositional hold-out sets are employed more frequently in out-of-distribution benchmark papers [9, 10, 12], where entire chemical families are deliberately withheld to expose performance cliffs. Yet, these remain isolated efforts rather than community standards. Out-of-distribution benchmarks themselves are rare; the structure-based suites developed by Cui and Wang [12] represent one of the few deliberate attempts to quantify robustness, revealing that standard models lose 30%–50% accuracy under controlled shift. The majority of works, including broad overviews [4, 5] and property-prediction pipelines [11, 13], continue to report only in-distribution metrics, thereby masking the very failures that occur in realistic deployment. This reliance on random splits not only inflates reported performance but also prevents systematic learning about which shift types are most detrimental, leaving the field without clear guidance on when a model can safely be transferred to new materials or laboratories.

Synthesis and Recommendations

Synthesis of the 30 papers reveals three dominant patterns. Pattern 1: domain shift is widely acknowledged at a conceptual level in foundational works [1, 3, 4, 7], yet rarely operationalized, with fewer than 20% of studies moving beyond lip service to detection or adaptation. Pattern 2: random train-test splits dominate evaluation despite their known inadequacy under shift [12, 22], producing overly optimistic accuracy figures that fail in practice. Pattern 3: shift-specific evaluation protocols remain exceptional rather than routine, even in papers that explicitly discuss out-of-distribution generalization [9, 10, 13].

Recommendations are therefore threefold. For authors: (a) test explicitly for shift magnitude before modeling using adversarial validation or statistical tests [22], (b) adopt validation strategies matched to the dominant shift type (time-based for temporal, compositional hold-out for chemical families), and (c) report shift metrics alongside performance numbers [3, 11]. For reviewers: (a) require justification of the chosen splitting strategy whenever a shift is plausible, and (b) question reliance on random splits in multi-source or multi-year datasets [7, 14]. For the community: (a) establish shared out-of-distribution benchmarks for materials property prediction [12], (b) adopt standardized shift-reporting templates in supplementary materials, and (c) develop open-source shift-aware pipelines that integrate detection and adaptation by default [8, 15]. These steps would transform domain shift from an overlooked pitfall into a core design consideration.

Conclusion

This review of 30 peer-reviewed publications demonstrates that the materials machine-learning community has yet to internalize the fundamental lesson of dataset shift articulated two decades ago: models trained and tested on mismatched distributions cannot be trusted for discovery or deployment. While four distinct shift types—temporal, compositional, experimental, and theoretical—are now clearly delineated, current practices still default to i.i.d. assumptions, leaving critical gaps in detection, adaptation, and evaluation. The path forward requires embedding shift-aware methods into every stage of research so that materials artificial intelligence can deliver the robust, generalizable predictions demanded by real-world applications. Only by making domain-shift awareness a standard rather than an exception will the field realize its full potential.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Feb 2023 Revised: 18 Mar 2023 Accepted: 29 Apr 2023
Published online: 18 July 2023

References

Dockès J, Varoquaux G, Poline JB. Preventing dataset shift from breaking machine-learning biomarkers. Gigascience. 2021;10(9):giab055.

Wang Y, Wald I, Wu Q, Usher W, Johnson CR. IEEE transactions on visualization and computer graphics. IEEE Trans Vis Comput Graph. 2019;25(6):2168-80.

Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. Appl Clin Inform. 2021;12(4):808-15.

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

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

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Gubernatis JE, Lookman TJ. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. Phys Rev Mater. 2018;2(12):120301.

Farahani A, Voghoei S, Rasheed K, Arabnia HR. A brief review of domain adaptation. In: Adv Data Sci Info Eng. Cham: Springer; 2021. p. 877-94.

Liu J, Shen Z, He Y, Zhang X, Xu R, Yu H, et al. Towards out-of-distribution generalization: A survey. arXiv [Preprint]. 2021:arXiv:2108.13624.

Wei S, Guo Y, Wen X, Li S, Zhu L, Zhu R. Admeood: Out-of-distribution benchmark for drug property prediction. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). New York, NY: IEEE; 2023. p. 387-94.

Abbasian Dehkordi S, Farajzadeh K, Rezazadeh J, Farahbakhsh R, Sandrasegaran K, Abbasian Dehkordi M. A survey on data aggregation techniques in IoT sensor networks. Wirel Netw. 2020;26(2):1243-63.

Cui P, Wang J. Out-of-distribution (OOD) detection based on deep learning: A review. Electronics. 2022;11(21):3500.

Ranasinghe C, Degbelo A. Encoding variables, evaluation criteria, and evaluation methods for data physicalisations: A review. Multimodal Technol Interact. 2023;7(7):73.

Koh PW, Sagawa S, Marklund H, Xie SM, Zhang M, Balsubramani A, et al. Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning. Brookline, MA: PMLR; 2021. p. 5637-64.

He H, Queen O, Koker T, Cuevas C, Tsiligkaridis T, Zitnik M. Domain adaptation for time series under feature and label shifts. In: International Conference on Machine Learning. Brookline, MA: PMLR; 2023. p. 12746-74.

Kong L, Xie S, Yao W, Zheng Y, Chen G, Stojanov P, et al. Partial disentanglement for domain adaptation. In: International Conference on Machine Learning. Brookline, MA: PMLR; 2022. p. 11455-72.

Guan H, Liu M. DomainATM: Domain adaptation toolbox for medical data analysis. Neuroimage. 2023;268:119863.

Laparra E, Bethard S, Miller TA. Rethinking domain adaptation for machine learning over clinical language. JAMIA Open. 2020;3(2):146-50.

Goel P, Ganatra A. Unsupervised domain adaptation for image classification and object detection using guided transfer learning approach and JS divergence. Sensors. 2023;23(9):4436.

Panda R, Kalmady SV, Greiner R. Multi-source domain adaptation techniques for mitigating batch effects: A comparative study. Front Neuroinform. 2022;16:805117.

Kubik C, Molitor DA, Rojahn M, Groche P. Deep learning-based domain adaptation for a generalized detection of wear phenomena during blanking. Manuf Lett. 2023;35:904-14.

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

Chen L, Zaharia M, Zou JY. Estimating and explaining model performance when both covariates and labels shift. Adv Neural Inf Process Syst. 2022;35:11467-79.

Chang J, Uehara M, Sreenivas D, Kidambi R, Sun W. Mitigating covariate shift in imitation learning via offline data with partial coverage. Adv Neural Inf Process Syst. 2021;34:965-79.

Shao JJ, Xu Y, Cheng Z, Li YF. Active model adaptation under unknown shift. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY: ACM; 2022. p. 1558-66.

Qiu H, Dobriban E, Tchetgen Tchetgen E. Prediction sets adaptive to unknown covariate shift. J R Stat Soc Ser B Stat Methodol. 2023;85(5):1680-705.

Germano D, Sciaraffa N, Ronca V, Giorgi A, Trulli G, Borghini G, et al. Unsupervised detection of covariate shift due to changes in EEG headset position: Towards an effective out-of-lab use of passive brain–computer interface. Appl Sci. 2023;13(23):12800.

Chiang CW, Yin M. You'd better stop! Understanding human reliance on machine learning models under covariate shift. In: Proceedings of the 13th ACM Web Science Conference. New York, NY: ACM; 2021. p. 120-9.

Xu R, Baracaldo N, Joshi J. Privacy-preserving machine learning: Methods, challenges and directions. arXiv [Preprint]. 2021:arXiv:2108.04417.

Nyborg J, Pelletier C, Lefèvre S, Assent I. TimeMatch: Unsupervised cross-region adaptation by temporal shift estimation. ISPRS J Photogramm Remote Sens. 2022;188:301-13.