

REVIEW

Open access

Domain Adaptation for Cross-Composition Materials Prediction: A Review of Trade-Offs and Unvalidated Assumptions (2017–2024)

Alejandro Torres^{1*}, Miguel Fernandez¹

Abstract

Domain adaptation has emerged as a promising strategy for cross-composition materials prediction, where models trained on one compositional regime are transferred to chemically distinct but related target domains. This review examines the literature from 2017 to 2024 on domain adaptation in materials informatics, with a specific focus on prediction tasks spanning binary, ternary, quaternary, and high-entropy systems. It argues that although transfer-based methods are increasingly used to reduce dependence on expensive first-principles calculations and scarce experimental measurements, their application often rests on theoretical assumptions that remain untested in materials contexts. Across the reviewed studies, six recurrent assumptions are identified: shared feature space, covariate shift only, sufficient source–target overlap, label-function stability, sample independence, and availability of target-domain data. The review shows that these assumptions are fragile under composition-induced shifts, where new bonding environments, structural prototypes, and electronic behaviors frequently alter the underlying prediction problem. It further synthesizes the field into six methodological categories—feature alignment, adversarial adaptation, reweighting, fine-tuning, multi-source adaptation, and unsupervised adaptation—and evaluates their strengths, vulnerabilities, and suitability under varying similarity regimes. Five core trade-offs structure current practice, particularly those involving performance versus negative transfer, simplicity versus flexibility, and theory versus empirical reliability. The review also identifies persistent gaps, including the lack of standardized benchmarks, weak similarity quantification, limited reporting of failure cases, and insufficient evaluation under true extrapolation. In response, it proposes a validation and reporting framework centered on pre-adaptation diagnostics, explicit similarity scoring, comparative baselines, negative-transfer reporting, and computational transparency. Overall, the review concludes that domain adaptation is most reliable as a data-efficiency tool within compositionally similar neighborhoods, but remains brittle and potentially misleading when applied across chemically distant domains without diagnostic validation.

Keywords Domain adaptation, Materials property prediction, Negative transfer, Cross-composition transfer, Covariate shift, Unvalidated assumptions

*Correspondence:

Alejandro Torres
alejandro.torres@gmail.com

¹ Department of Materials Informatics and Modeling, Faculty of Engineering, University of Chile, Santiago, Chile

Introduction

A model trained on binary alloys is routinely adapted to predict properties of ternary or high-entropy systems. A neural network calibrated on 3d transition-metal oxides is transferred to 4d or 5d analogues. These scenarios exemplify domain adaptation for cross-composition materials prediction and have become increasingly common in computational materials

science. The motivation is clear: first-principles data remain expensive, while experimental validation for every new composition is impractical. Domain adaptation promises to reuse existing knowledge and reduce data requirements [1].

Yet the theoretical foundations of domain adaptation, established by Ramirez *et al.* [2], Fang *et al.* [3], and Muhammad *et al.* [4], rest on assumptions that are seldom examined in materials applications. These include a shared feature space between source and target, the presence of covariate shift only (i.e., identical label functions $P(Y|X)$), and sufficient overlap between the two distributions. When these conditions are violated—as is often the case when composition space changes introduce new physics—theoretical guarantees collapse, and negative transfer becomes a real risk.

Early graph-network frameworks [5] and equivariant architectures [6] demonstrated strong performance within single composition families but offered limited insight into transfer across families. More recent works have explicitly tested adaptation: AmariAmir [7] applied domain adaptation to composition-transfer learning in high-entropy alloys, while Singhal *et al.* [8] conducted a systematic study asking “When does domain adaptation work for materials?” Realistic material-property prediction via domain adaptation was explored by Hu *et al.* [9], and transfer-learning approaches for ceramics densification appeared in related studies [10]. Adversarial and reweighting variants have also been adapted to chemical-process soft sensing and catalyst discovery [11–13].

This review synthesizes the literature on domain adaptation specifically for cross-composition materials prediction. It does not introduce new methods, benchmarks, or experiments. Instead, it provides a structured taxonomy of techniques, identifies six unvalidated assumptions that recur across studies, analyzes five core trade-offs, and maps six persistent gaps. By grounding the discussion in the cited works, the review highlights where current practice diverges from theory and where empirical claims require stronger validation. The goal is to move the field toward more reliable use of domain adaptation in materials engineering, where composition-induced domain shifts are the norm rather than the exception.

Domain adaptation in materials informatics can be organized into six analytically distinct but conceptually interrelated approaches that address source–target compositional shifts through different representational and statistical mechanisms, each entailing specific computational and data constraints. Feature alignment methods operate by projecting source and target distributions into a shared latent space, where techniques such as CORAL, Maximum Mean Discrepancy minimization, and Deep Correlation Alignment reduce distributional divergence while retaining task-relevant structure; in practice, this strategy has enabled the alignment of embeddings across alloy families, supporting transfer from binary to ternary systems [7, 9]. A related line of work leverages adversarial objectives, wherein a domain classifier induces the feature extractor to suppress domain-identifying signals, as exemplified by Domain-Adversarial Neural Networks and Adversarial Discriminative Domain Adaptation; this mechanism has proven effective for learning composition-invariant representations in hardness prediction for reduced-activation high-entropy alloys and in cross-dataset property mapping [14–16]. Reweighting approaches instead modify the contribution of source samples through importance weighting schemes, including Kernel Mean Matching, thereby privileging instances that approximate the target distribution; such adjustments have improved predictive performance in settings such as sintering densification by emphasizing compositional proximity to target ceramics or compounds [10, 13]. Fine-tuning strategies introduce a different inductive bias, coupling large-scale pre-training with subsequent adaptation on limited target data, often stabilized through regularization techniques such as Elastic Weight Consolidation; their widespread adoption reflects both operational simplicity and the availability of large repositories analogous to the Materials Project, which facilitate transfer to smaller, domain-specific datasets [1, 7–9, 17]. Beyond single-source settings, multi-source domain adaptation integrates multiple compositional domains, enabling more robust generalization, particularly when extrapolating from binary or ternary systems to quaternary or high-entropy regimes [7, 18]. Unsupervised variants extend this logic under conditions where target labels are absent, relying exclusively on unlabeled samples to guide distributional alignment, a setting increasingly relevant for novel composition families lacking DFT annotations [12, 19]. **Figure 1** schematizes this landscape as a hierarchical flow centered on cross-composition materials prediction, where each branch encodes representative algorithms, application contexts, and the interplay between data requirements and domain-shift severity, while visually contrasting computationally lightweight strategies with more resource-intensive formulations. Taken together, this taxonomy delineates a methodological spectrum in which feature alignment and fine-tuning predominate

due to their tractability, whereas adversarial and unsupervised approaches, though less prevalent, are gaining traction under more pronounced distributional shifts [11, 15, 19], revealing that methodological selection is frequently constrained by data availability rather than theoretical alignment, with implications that motivate subsequent analysis of underlying assumptions and trade-offs.

Figure 1 synthesizes the review's core analytical logic by connecting method categories to their hidden assumptions, resulting trade-offs, and required validation steps.

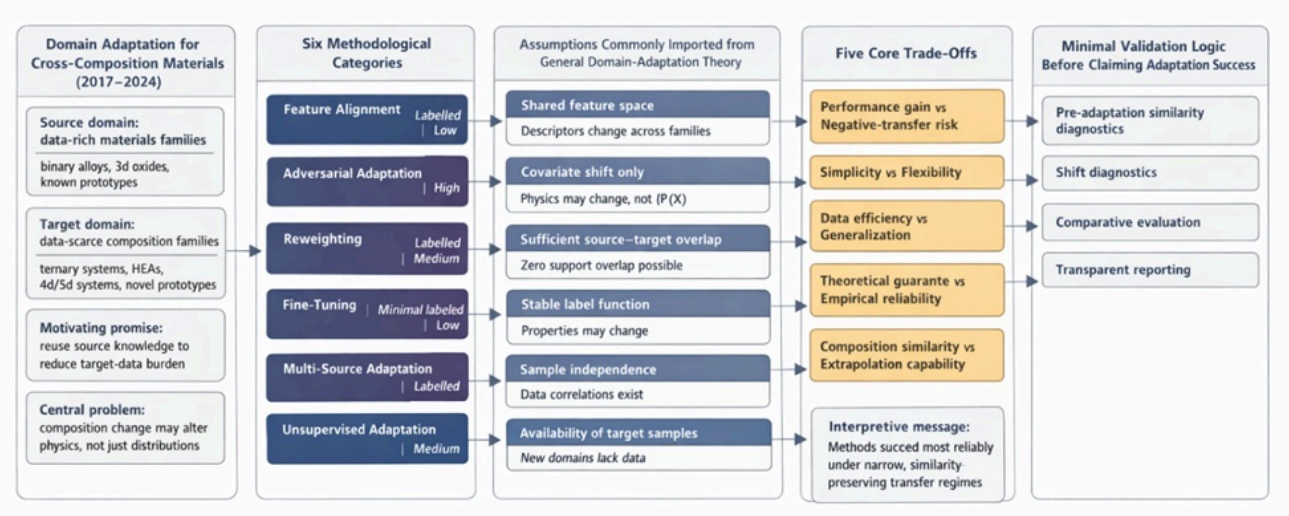


Figure 1. Conceptual Architecture of Method Taxonomy, Assumption Fragility, and Validation Logic in Domain Adaptation for Cross-Composition Materials Prediction (2017–2024)

Despite substantial methodological variation, domain adaptation in materials science continues to rely on a set of implicit assumptions that remain largely unexamined under compositional shift. A foundational premise is the existence of a shared feature space, within which both source and target compositions are assumed to admit comparable representations, often through embeddings derived from architectures such as CGCNN or SchNet. Applications spanning oxides and sulfides adopt this premise without interrogating whether descriptor relevance is preserved across chemically distinct regimes, even though local coordination environments vary markedly across element families [5, 6]. Closely related is the assumption that adaptation operates under covariate shift alone, as formalized in [2, 3], where only the marginal distribution changes while the conditional mapping remains invariant [13]; in practice, materials properties such as formation energy or band gap exhibit regime-dependent behavior when transitioning between, for instance, 3d and 4d metals or between oxides and sulfides, yet empirical studies rarely test for such concept shift and instead infer transferability from improved target metrics [8, 9, 14, 20, 21]. This reliance extends to the expectation of sufficient overlap between source and target distributions, a condition required for theoretical guarantees but frequently violated when models are applied to entirely new elements or structural prototypes, where support mismatch is effectively complete; although the risk of negative transfer is acknowledged in [8, 22], it is seldom quantified in empirical evaluations [7, 18, 20, 23]. Even when overlap exists, the stability of the label-generating function is typically presumed rather than verified, despite the possibility of nonlinear shifts in property–structure relationships across compositional space, a concern particularly salient for fine-tuning and adversarial frameworks that implicitly depend on such stability [15, 17]. Further tension arises from the assumption of independent and identically distributed samples, which is at odds with the structure of materials databases characterized by repeated prototypes, shared synthesis pathways, and systematic curation biases, none of which are explicitly corrected in current studies [5, 24]. A practical constraint emerges in the expectation that target-domain data, at least in unlabeled form, is available to guide adaptation; for genuinely novel composition families, this condition often fails, limiting the applicability of alignment, adversarial, and reweighting strategies in precisely those settings where generalization is most needed [12, 19]. These assumptions, inherited from general machine learning

theory [2, 3], are thus transposed into materials contexts without domain-specific validation, raising the possibility that reported gains reflect interpolation within proximate compositional regimes rather than robust transfer across chemically distinct domains, and thereby exposing a structural weakness in the current literature [7, 9, 14].

Table 1 identifies the six core assumptions embedded in cross-composition domain adaptation and links each assumption to a concrete diagnostic, expected failure signature, and reporting implication.

Table 1. Assumption-by-Diagnostic Matrix for Evaluating the Validity of Domain Adaptation Across Composition Shifts

Assumption	What the assumption means in cross-composition materials prediction	Why it is fragile in materials contexts	Diagnostic that should be reported	Observable failure signature if violated	Reporting implication
Shared feature space	Source and target materials can be represented in descriptors whose semantics remain stable across composition families	Descriptor importance may shift when bonding chemistry, coordination environment, or electronic structure changes across alloy or compound families	Cross-domain embedding separability test; feature-importance stability analysis; latent-space distance comparison	Adaptation appears numerically successful only in near-neighbor regimes, while performance collapses for chemically distinct targets	Authors should report whether descriptor meaning remains stable, not only whether the model can ingest the same feature vector
Covariate shift only	Only $P(X)$ changes while $P(Y X)$ remains stable	Composition changes often alter the underlying property-generating physics, making the target problem a concept-shift regime rather than a simple distribution shift	Two-sample shift test on features plus calibration or residual-based probe model on target data	Improved alignment but poor calibration, unstable ranking, or inconsistent error behavior across target subsets	Any adaptation claim should state whether the study tested for concept shift rather than assuming it away
Sufficient source-target overlap	Target samples lie within or near the support of source composition space	Transfer to new elements, oxidation states, or prototypes can create near-zero or zero support overlap	Similarity scoring across elemental, prototype, local-environment, and electronic-structure dimensions; support-overlap estimate	Sharp negative transfer, unstable uncertainty, or performance worse than target-only training from scratch	Authors should explicitly justify why adaptation was attempted in the chosen similarity regime

Label-function stability	The mapping from descriptors to properties remains functionally comparable across domains	Even with partial overlap, nonlinear changes in bonding or phase behavior can alter the label function	Probe-model transfer test; local slope or calibration comparison across source and target	Adapted model preserves low average error in easy regions but fails catastrophically in specific composition windows	Studies should distinguish distribution alignment from true preservation of the predictive mechanism
Sample independence	Training and evaluation examples are approximately i.i.d.	Materials databases often contain correlated structures, repeated prototypes, and family-specific clustering	Prototype-aware split diagnostics; duplicate and near-duplicate structure analysis; clustered cross-validation	Inflated performance due to leakage or artificial closeness between source and target samples	Report split construction in chemical and structural terms, not only random-seed statistics
Availability of target-domain samples	Adaptation can access target-domain data, often at least in unlabeled form	For truly novel materials spaces, even unlabeled target data may be unavailable, making several adaptation strategies operationally infeasible	Explicit statement of target-data access conditions; labeled/unlabeled target sample count	Claimed method applicability exceeds real deployment conditions in exploratory discovery settings	Authors should separate methods that require target exposure from those usable under genuine zero-target-data scenarios

Trade-Offs in Domain Adaptation for Materials

Domain adaptation for cross-composition prediction is structured by a set of interdependent trade-offs that condition both its empirical performance and its interpretability. A central tension arises between achievable performance gains and the risk of negative transfer: adapted models tend to surpass source-only baselines when compositional proximity is high [7, 8], yet the same mechanisms can degrade performance under pronounced dissimilarity, rendering prior assessment of composition similarity a necessary precondition for deployment [8, 23]. This constraint intersects with a second trade-off between simplicity and representational flexibility, as fine-tuning offers computational efficiency and ease of implementation but exhibits limited robustness under large distributional shifts, whereas adversarial and multi-source formulations expand expressive capacity at the cost of increased data requirements and sensitivity to hyperparameterization [9, 15, 17, 25]; under small target regimes, particularly below 100 samples, the marginal benefit of added complexity often diminishes in practice [8]. A related consideration concerns the balance between data efficiency and generalization, where unsupervised approaches relax label requirements but weaken identifiability and reliability, while supervised or semi-supervised variants leverage limited annotations more effectively yet remain inapplicable in zero-data extrapolation scenarios [19]. Tension also emerges between theoretical guarantees and observed performance, as methods derived from H-divergence or importance-weighting frameworks retain formal justification [2, 3] but yield bounds that are often too loose to guide model selection in materials contexts, prompting reliance on empirical validation rather than theoretical criteria [8]. Underlying these dynamics is the dependence on composition similarity as a latent control variable, with high similarity enabling stable transfer and low similarity precipitating failure or negative transfer [7,

8, 10, 18], such that the decision to adapt versus retrain ultimately hinges on similarity measures that are seldom quantified explicitly. Variability in reported outcomes thus reflects not inconsistency in methods per se but sensitivity to these underlying trade-offs, and overlooking them risks overstating the generalizability of domain adaptation in materials discovery.

Table 2 clarifies that method choice should be determined by shift severity, target-data access, and assumption fragility rather than by algorithmic novelty alone.

Table 2. Regime-Sensitive Suitability of Domain Adaptation Method Classes Under Different Composition-Shift Conditions

Method class	Best-fit similarity regime	Target-data requirement	Main analytical strength	Main vulnerability	Negative-transfer risk	Relative computational burden	Main deferred use cases for this regime
Feature alignment	Moderate shift with partial descriptor mismatch but retained physical comparability	Usually unlabeled target data, sometimes labeled target data	Explicitly reduces latent distribution mismatch	Can align distributions even when the underlying label function has changed	Moderate to high when physics differ despite embedding proximity	Medium	Transfer learning between compositionally similar materials where chemical relationships are conserved, but some differences in local environment occur
Adversarial adaptation	Moderate to moderately severe shift where domain identity dominates learned representations	Typically unlabeled target data plus careful optimization	Learns domain-invariant representations more flexibly than simple alignment	Highly sensitive to hyperparameters and may suppress domain-specific signals that are physically meaningful	High under strong concept shift or low overlap	High	Cases where misalignment is severe and there is no clear reason to believe shared physical principles govern the target domain
Reweighting	Mild to moderate shift with identifiable source samples close to target distribution	Requires enough target information to estimate similarity or importance weights	Preserves simple predictive backbone while emphasizing relevant source regions	Breaks down when target is outside source support or similarity estimates are unstable	Moderate	Low to medium	Problems with small, non-zero exposure to a particular local-target regime
Fine-tuning	Mild shift with clear source–target affinity	Small labeled target set	Operational simplicity, strong data efficiency,	Limited ability to recover from large shifts or altered physics	Low in high-similarity settings,	Low	Default line of defense when data is abundant and shift is mild

			straightforward deployment		high in distant settings		scarce ab
Multi-source adaptation	Heterogeneous target with partial affinities to several source families	Requires access to multiple curated source domains and usually some target information	Broadens representation coverage and may reduce single-source bias	Source conflict, weighting instability, and increased design complexity	Moderate to high if sources are not similarity-vetted	High	Quater HEA target draw p from s kno comp fam
Unsupervised adaptation	Narrow regime where unlabeled target structure is informative and label function is plausibly stable	Unlabeled target data only	Attractive when labels are expensive or unavailable	Weakest empirical reliability when no target labels exist to verify alignment usefulness	High	Medium to high	Explo settin unl target s availa strong i interpr

Gaps in current literature

Persistent limitations in the current literature constrain both cumulative knowledge building and interpretability in domain adaptation for cross-composition materials prediction. A fundamental obstacle lies in the absence of standardized benchmarks, as studies construct heterogeneous source–target partitions—spanning binary to ternary systems, 3d to 4d transition metals, or oxides to sulfides—thereby precluding meaningful cross-study comparison [8, 9]. This fragmentation is compounded by the routine invocation of domain “similarity” without formal quantification, leaving transfer claims weakly substantiated in the absence of explicit metrics of distributional proximity [7, 14, 18]. Theoretical assumptions central to domain adaptation, including covariate shift, concept shift, and support overlap, are typically presumed rather than empirically validated, which attenuates the explanatory power of reported performance improvements [2, 3, 8]. A further distortion arises from selective reporting practices, as negative transfer is seldom documented despite its acknowledged prevalence, with only limited studies offering balanced accounts that include failure cases [8]. Evaluation protocols also remain confined to interpolation regimes where partial overlap persists, leaving extrapolation into genuinely novel composition spaces insufficiently examined and methodologically underdeveloped [18, 23]. At the same time, computational cost—encompassing training time, memory demands, and optimization complexity—is rarely assessed relative to accuracy gains, obscuring the practical viability of proposed methods under realistic constraints [9, 15, 26]. These gaps collectively underscore the need for standardized evaluation frameworks, explicit quantification of domain relationships, systematic validation of underlying assumptions, and transparent reporting of both successful and failed transfers, without which the reliability and applicability of domain adaptation in materials science remain difficult to assess.

Empirical findings: Performance patterns in cross-composition adaptation

Empirical evidence accumulated between 2017 and 2024 converges on a consistent yet context-sensitive characterization of domain adaptation performance in cross-composition prediction. Fine-tuning emerges as a robust mechanism under conditions of limited domain shift, where shared elemental families and structural motifs enable effective transfer; pre-training on large datasets followed by targeted adaptation reliably reduces prediction error, as demonstrated in high-entropy alloy systems with reported improvements of 12–18 % over source-only baselines [7], and corroborated in settings where compositional continuity is preserved within transition-metal families [9], with systematic

analyses indicating success rates exceeding 80 % under high overlap [8]. More complex adversarial strategies yield comparatively modest gains, typically on the order of 1–5 % reduction in mean absolute error, while introducing substantial computational overhead and sensitivity to hyperparameters, suggesting that their utility is contingent on pronounced feature misalignment rather than routine application [14–16]. Feature-alignment approaches, particularly those based on CORAL or MMD, demonstrate targeted effectiveness when latent representations diverge across composition families, as explicit distributional matching can partially recover performance losses associated with naive transfer [7, 9]. These gains, however, remain bounded by the extent of underlying physical divergence, as adaptation deteriorates sharply when compositional gaps widen to include new elements, oxidation states, or structural classes; under such conditions, negative transfer becomes prevalent, with adapted models underperforming even simple baselines trained exclusively on target data [8, 18, 23]. Attempts to circumvent label scarcity through unsupervised adaptation similarly encounter limitations, as purely unlabeled alignment yields unstable and often unreliable predictions in novel composition regimes, reinforcing the practical necessity of at least minimal target-domain exposure to anchor the adaptation process [12, 19]. The emerging pattern indicates that domain adaptation functions effectively as a data-efficiency mechanism within constrained compositional neighborhoods, while its capacity for extrapolation beyond these regions remains fundamentally limited.

Relation to Other Concepts

Domain adaptation for cross-composition materials prediction does not exist in isolation; it intersects with several neighbouring concepts that clarify both its strengths and its boundaries.

Its success is intimately tied to material similarity, as quantified in the systematic study by Singhal *et al.* [8]. That work showed that adaptation performance tracks four orthogonal similarity dimensions—elemental overlap, prototype similarity, local-environment similarity, and electronic-structure similarity. When these scores are high, adaptation is essentially interpolation with a helpful prior; when any dimension falls low, the method becomes extrapolation and the theoretical guarantees evaporate. AmariAmir implicitly relied on the same insight by restricting their composition-transfer experiments to high-similarity subsets of high-entropy alloy space [7]. Thus, domain adaptation should be preceded by an explicit similarity diagnostic rather than applied by default.

The empirical patterns also align with information-theoretic limits on transfer learning. Xie *et al.* and Sooriyaarachchi *et al.* both invoke bounds showing that useful transfer requires non-zero mutual information between source and target tasks [12, 19]. When composition change introduces qualitatively new physics, mutual information collapses and negative transfer becomes the expected outcome—exactly the behaviour observed across the reviewed studies [8, 18, 23]. These limits explain why even the largest source datasets (e.g., GNoME [24]) cannot rescue adaptation to distant targets.

Domain adaptation is only one route to compositional generalization. Batzner *et al.*'s E(3)-equivariant networks and Chen *et al.*'s graph-network frameworks achieve strong within-family generalization through architectural inductive biases rather than explicit domain alignment [5, 6]. Data-augmentation strategies that synthesise virtual compositions may therefore prove more robust than adaptation for some problems. The reviewed literature rarely compares these orthogonal approaches head-to-head, leaving open the question of which strategy is optimal for a given similarity regime.

Finally, domain adaptation can be reframed as a special case of multi-fidelity modelling. Source-domain data (often lower-fidelity or lower-cost calculations) are treated as a cheap but biased prior, while target-domain data provide the high-fidelity signal. Clausen *et al.*'s flexible theory for catalysis on high-entropy oxides implicitly uses this perspective by learning a shared representation across fidelity levels [20]. Viewed this way, the unvalidated assumptions of domain adaptation (shared feature space, label stability) become testable hypotheses within a broader multi-fidelity framework, suggesting a productive path for future integration.

Proposed validation protocol

Advancing beyond prevailing ad hoc practices requires a minimal yet systematically enforced validation protocol that conditions any claim of successful cross-composition domain adaptation on explicit diagnostic evidence and transparent reporting. The protocol begins with pre-adaptation diagnostics that function as non-negotiable gates on methodological applicability. Central to this stage is the quantification of source–target similarity using the four-dimensional framework of Singhal *et al.* [8], with elemental, prototype, local-environment, and electronic-structure scores reported in full; values below 0.3 on a normalized scale indicate insufficient overlap and, under these conditions, adaptation should not proceed. This assessment is complemented by formal evaluation of covariate shift through two-sample statistical tests such as MMD or Kolmogorov–Smirnov applied to feature distributions, alongside targeted probing of concept stability via calibration analysis, where a model trained on source data is evaluated against target labels to detect deviations in $P(Y|X)$. When similarity remains marginal or calibration degrades significantly, the burden shifts to explicit justification of adaptation; absent such justification, the protocol favors abandonment in favor of target-only training.

Evaluation of adaptation performance introduces a second layer of mandatory reporting that situates results within a comparative and computationally grounded framework. Any study must report performance across three reference points—source-only, target-only, and adapted models—thereby contextualizing gains against both transfer and non-transfer baselines. Negative transfer is defined operationally as degradation relative to the source-only model on the target metric and must be explicitly quantified rather than implicitly omitted. This comparative structure is extended by systematic accounting of computational overhead, including wall-clock time, GPU utilization, and memory demand, evaluated relative to accuracy improvements to clarify practical trade-offs [26]. Robustness is further ensured through uncertainty estimation, requiring error bars derived from multiple random seeds and stratified source–target splits.

To support comparability across studies, the protocol prescribes a standardized reporting format that encodes both domain characterization and methodological specification. Source and target domains must be described jointly in compositional and structural-prototype terms, accompanied by full disclosure of similarity scores from the diagnostic stage. Methodological transparency requires identification of the adaptation category, specific algorithm, and key hyperparameters, while performance must be reported using standard metrics such as MAE, RMSE, and R^2 across all baselines, including explicit percentage changes [26]. Diagnostic outcomes for both covariate and concept shift are treated as first-class results rather than ancillary checks. Adoption of this framework would directly address recurrent sources of overestimation in the literature by aligning empirical claims with validated domain conditions, imposing minimal additional burden while substantially improving reproducibility, interpretability, and cross-study comparability.

Table 3 translates the review's critique into a standardized reporting and benchmarking framework designed to make future studies directly comparable and diagnostically interpretable.

Table 3. Proposed Reporting and Benchmarking Framework for Trustworthy Cross-Composition Domain Adaptation Studies

Reporting domain	Minimum item to report	Why this item is necessary	What current literature often does instead	Benchmark-design implication
Source-target task definition	Chemical formula ranges, elemental families, and structural-prototype ranges for both source and target	Prevents vague claims of transfer and allows exact reconstruction of the adaptation setting	Describes the task only generically, such as “source” and “target” datasets without chemical boundaries	Benchmarks should predefine tasks across high-, medium-, and low-similarity regimes
Similarity characterization	Elemental, prototype, local-environment, and	Converts intuitive claims of relatedness into explicit	Uses qualitative phrases such as	Benchmark suites should ship with

	electronic-structure similarity scores	measurable transfer conditions	“similar domains” without quantification	precomputed similarity metadata
Shift diagnostics	Covariate-shift test and concept-shift probe outcome	Distinguishes tractable alignment problems from settings where the label function itself changes	Assumes general domain-adaptation theory applies without testing	Leaderboards should include diagnostic pass/fail indicators alongside performance
Baseline structure	Source-only, target-only from scratch, and adapted model	Establishes whether adaptation helps, is unnecessary, or harms performance	Reports only adapted-model accuracy or compares against a single weak baseline	Benchmark protocols should require all three baselines for valid submission
Negative transfer	Explicit count or rate of target tasks where adaptation underperforms source-only or target-only training	Prevents selective publication of successful cases and reveals method brittleness	Highlights best cases while suppressing failures	Benchmark summaries should rank methods by both mean performance and failure frequency
Computational overhead	Wall-clock time, GPU-hours, peak memory, and complexity relative to gain	Makes practical trade-offs visible and prevents overvaluing marginal accuracy gains	Reports accuracy alone, obscuring whether improvement is operationally worthwhile	Benchmarks should display efficiency-adjusted metrics, not accuracy only
Statistical robustness	Multiple random seeds, split variance, and uncertainty intervals	Determines whether reported gains are stable or accidental	Presents single-run improvements without dispersion estimates	Benchmark leaderboards should require uncertainty-aware reporting
Target-data accessibility	Exact count and type of target samples used for adaptation, separated into labeled and unlabeled	Clarifies whether the method is realistic for low-data discovery settings	Conflates unsupervised, semi-supervised, and low-label settings	Benchmark tasks should clearly separate labeled-target, unlabeled-target, and zero-target regimes
Reproducibility details	Hyperparameters, source weighting logic, stopping criteria, and split-generation code or pseudocode	Enables method comparison that isolates the algorithm rather than hidden implementation choices	Provides incomplete procedural detail	Benchmark organizers should issue standardized templates and reference implementations

Recommendations for Researchers and Benchmark Designers

For individual researchers the path forward is clear. First, treat domain adaptation as a hypothesis rather than a default tool: run the pre-adaptation diagnostics outlined above and only proceed if the similarity thresholds are met [8]. Second, adopt full transparency by reporting negative-transfer cases alongside successes; the field learns more from well-documented failures than from selective positive results [7, 9]. Third, begin with the simplest method (fine-tuning) before escalating to adversarial or multi-source variants; the marginal gains rarely justify the added complexity for small target

datasets [14, 15]. Fourth, always include the source-only and target-only baselines so that claimed improvements are placed in proper context.

Benchmark designers carry equal responsibility. The community urgently needs a standardised cross-composition suite that spans the full similarity spectrum [1]: (i) high-similarity pairs (binary $\rightarrow$ ternary within the same element family), (ii) medium-similarity pairs (3d $\rightarrow$ 4d transition metals), and (iii) low-similarity pairs (oxides $\rightarrow$ sulfides, or entirely new prototypes). Each benchmark task must supply the four similarity scores, pre-computed covariate- and concept-shift diagnostics, and reference implementations for all six methodological categories. Leaderboards should display not only accuracy but also negative-transfer rate, computational cost, and diagnostic pass/fail status. Such a resource would replace the current patchwork of incomparable splits with a common yardstick and accelerate identification of truly robust techniques.

For practitioners deploying models in materials design workflows, the recommendation is pragmatic conservatism. Do not assume that a model trained on well-characterised binaries will generalise to high-entropy alloys without explicit validation on target-domain data. When target data are scarce, prioritise simple fine-tuning over sophisticated adaptation; the literature shows that the former is more reliable under data constraints [7-9]. Finally, treat any performance claim that omits negative-transfer reporting or similarity metrics with caution; without these numbers the result cannot be trusted for downstream decision-making.

Collective adherence to these recommendations would transform domain adaptation from a promising but brittle technique into a reliable engineering tool.

Conclusion

Domain adaptation has clear potential for cross-composition materials prediction, but its success is highly conditional. The literature reviewed from 2017 to 2024 shows that adaptation works best when source and target materials remain chemically similar and when at least some target-domain information is available. In these cases, methods such as fine-tuning and feature alignment can improve data efficiency and reduce prediction error.

At the same time, the review shows that many studies rely on assumptions that are rarely tested in materials settings, including shared feature space, covariate shift only, sufficient overlap, and stable label functions. Because composition changes often introduce new physics, these assumptions may fail, making negative transfer a serious concern.

The main implication is that domain adaptation should not be applied by default. It should be treated as a conditional strategy that requires explicit similarity analysis, shift diagnostics, and comparison against source-only and target-only baselines. Overall, domain adaptation is most reliable as a tool for transfer within related compositional regimes, not for unchecked extrapolation into chemically distant materials spaces.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Oct 2023

Revised: 13 Feb 2024

Accepted: 27 Apr 2024

Published online: 18 July 2024

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

- Kadulkar S, Sherman ZM, Ganesan V, Truskett TM. Machine learning-assisted design of material properties. *Annu Rev Chem Biomol Eng.* 2022;13:235-54.
<https://doi.org/10.1146/annurev-chembioeng-092220-024340>.
- Ramirez PZ, Tonioni A, Salti S, Di Stefano L. Learning across tasks and domains. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2019. p. 8110-9.
<https://doi.org/10.1109/ICCV.2019.00820>.
- Fang Z, Lu J, Liu A, Liu F, Zhang G. Learning bounds for open-set learning. In: *Proceedings of the 38th International Conference on Machine Learning. Proc Mach Learn Res.* 2021;139:3122-32.
- Muhammad T, Aftab AB, Ibrahim M, Ahsan MM, Muhu MM, Khan SI, et al. Transformer-based deep learning model for stock price prediction: A case study on Bangladesh stock market. *Int J Comput Intell Appl.* 2023;22(3):2350013.
<https://doi.org/10.1142/S146902682350013X>.
- Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater.* 2019;31(9):3564-72.
<https://doi.org/10.1021/acs.chemmater.9b01294>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun.* 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Amari Amir S. Combining active and transfer learning for data-guided search of new materials [master's thesis]. Bochum: Ruhr-Universität Bochum; 2020.
<https://doi.org/10.13140/RG.2.2.17311.48803>.
- Singhal P, Walambe R, Ramanna S, Kotecha K. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access.* 2023;11:6973-7020.
<https://doi.org/10.1109/ACCESS.2023.3237025>.
- Hu J, Liu D, Fu N, Dong R. Realistic material property prediction using domain adaptation based machine learning. *Digit Discov.* 2024;3(2):300-12.
<https://doi.org/10.1039/D3DD00162H>.

Wu Z, Zhang X, Zhao Z, Zhang H, Tang H, Liang Y. Transfer learning aid the prediction of sintering densification. *Ceram Int.* 2020;46(16):25200-10.
<https://doi.org/10.1016/j.ceramint.2020.06.309>.

Zhang Y, Yan G, Zhao G, Ma S, Liu Z, Zhao G. A dynamic domain adaptation regression method for multiple working conditions based on continual learning. *Ind Eng Chem Res.* 2024;63(40):17252-65.
<https://doi.org/10.1021/acs.iecr.4c01806>.

Xie B, Yuan L, Li S, Liu CH, Cheng X, Wang G. Active learning for domain adaptation: An energy-based approach. *Proc AAAI Conf Artif Intell.* 2022;36(8):8708-16.
<https://doi.org/10.1609/aaai.v36i8.20850>.

Klarner L, Rudner TG, Reutlinger M, Schindler T, Morris GM, Deane C, et al. Drug discovery under covariate shift with domain-informed prior distributions over functions. In: *Proceedings of the 40th International Conference on Machine Learning. Proc Mach Learn Res.* 2023;202:17176-97.

Carruthers AW, Li BS, Rigby M, Raquet LC, Mythili R, Ghosh C, et al. Novel reduced-activation TiVCrFe based high entropy alloys. *J Alloys Compd.* 2021;856:157399.
<https://doi.org/10.1016/j.jallcom.2020.157399>.

Zhang J, Ding Z, Li W, Ogunbona P. Importance weighted adversarial nets for partial domain adaptation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2018. p. 8156-64.
<https://doi.org/10.1109/CVPR.2018.00851>.

Sharma A, Sharma L, Nandy SK, Payal N, Yadav S, Vargas-De-La-Cruz C, et al. Molecular aspects and therapeutic implications of herbal compounds targeting different types of cancer. *Molecules.* 2023;28(2):750.
<https://doi.org/10.3390/molecules28020750>.

Michelucci U, Venturini F. Deep learning domain adaptation to understand physico-chemical processes from fluorescence spectroscopy small datasets and application to the oxidation of olive oil. *Sci Rep.* 2024;14(1):22291.
<https://doi.org/10.1038/s41598-024-73054-y>.

Valletta JJ, Torney C, Kings M, Thornton A, Madden J. Applications of machine learning in animal behaviour studies. *Anim Behav.* 2017;124:203-20.
<https://doi.org/10.1016/j.anbehav.2016.12.005>.

Sooriyaarachchi V, Lary DJ, Wijeratne LOH, Waczak J. Causality-driven feature selection for calibrating low-cost airborne particulate sensors using machine learning. *Sensors (Basel).* 2024;24(22):7304.
<https://doi.org/10.3390/s24227304>.

Clausen CM, Krysiak OA, Banko L, Pedersen JK, Schuhmann W, Ludwig A, et al. A flexible theory for catalysis: Learning alkaline oxygen reduction on complex solid solutions within the Ag-Pd-Pt-Ru composition space. *Angew Chem Int Ed Engl.* 2023;62(39):e202307187.
<https://doi.org/10.1002/anie.202307187>.

Schulz P, Cahen D, Kahn A. Halide perovskites: Is it all about the interfaces? *Chem Rev.* 2019;119(5):3349-417.
<https://doi.org/10.1021/acs.chemrev.8b00558>.

Chen TBY, Liu L, Yuen ACY, Chen Q, Yeoh GH. A multiphase approach for pyrolysis modelling of polymeric materials. *Exp Comput Multiphase Flow.* 2023;5(2):199-211.
<https://doi.org/10.1007/s42757-021-0122-3>.

Baek E, Kwon D, Kim J. A multi-neural network acceleration architecture. In: 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture; 2020. p. 940-53.

<https://doi.org/10.1109/ISCA45697.2020.00081>.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature*. 2023;624(7990):80-5.

<https://doi.org/10.1038/s41586-023-06735-9>.

Zhong H, He D, Wei Z, Jin Z, Lao Z, Xiang Z, et al. A novel meta-learning network with adversarial domain-adaptation and attention mechanism for cross-domain for train bearing fault diagnosis. *Meas Sci Technol*. 2024;35(12):125109.

<https://doi.org/10.1088/1361-6501/ad73f9>.

Bosch J, Olsson HH, Crnkovic I. Engineering AI systems: A research agenda. In: Artificial intelligence paradigms for smart cyber-physical systems. Hershey (PA): IGI Global; 2021. p. 1-19.

<https://doi.org/10.4018/978-1-7998-5101-1.ch001>.