

ORIGINAL RESEARCH

Open access

Systematic Collapse of Transfer Learning in Small-Data Materials Regimes: Why Fine-Tuning Fails

Chen Hao^{1*}, Liu Fang¹, Zhao Lin²

Abstract

Transfer learning has become a dominant paradigm for addressing the persistent small-data limitation in computational materials science, particularly through pre-training graph neural networks on large-scale repositories and subsequent fine-tuning on task-specific datasets. This strategy assumes that universal atomic-scale representations—such as local coordination environments, bonding characteristics, and elemental embeddings—transfer effectively across domains. However, accumulating empirical evidence reveals a systematic and reproducible failure: in small-data regimes, fine-tuned models frequently underperform both zero-shot pre-trained models and models trained from scratch. This work presents a comprehensive failure-mode analysis demonstrating that such collapse is not incidental but structurally inevitable under common conditions in materials applications. Four interdependent mechanisms are identified—catastrophic forgetting, negative transfer, representation distortion, and overfitting amplification—arising from irreducible source–target domain mismatch, feature entanglement in equivariant graph architectures, and the statistical insufficiency of target datasets typically containing fewer than 10^3 samples. Under these constraints, the corrective signal provided by target data is inadequate to overcome entrenched source biases or to reconfigure internal representations. Drawing exclusively on consolidated literature, we show that transfer learning failure emerges when domain similarity falls below critical thresholds and when data scarcity prevents stable optimization. The study further formalizes operational diagnostics—including baseline benchmarking, source-validation tracking, and learning-rate sensitivity analysis—and introduces mitigation strategies such as progressive unfreezing, elastic weight consolidation, replay-based regularization, and similarity-gated transfer decisions. By reframing transfer learning as a high-risk, condition-dependent strategy rather than a default solution, this work provides a principled foundation for robust model deployment in materials informatics. It challenges prevailing assumptions of universal transferability and advocates for systematic failure-mode reporting to ensure reliability in machine-learned interatomic potentials and property prediction workflows.

Keywords Small-data regimes, Catastrophic forgetting, Materials graph neural networks, Transfer learning failure, Fine-tuning collapse, Negative transfer

*Correspondence:

Chen Hao

chen.hao@outlook.com

¹ Department of Intelligent Materials Modeling, Faculty of Engineering, Zhejiang University, Hangzhou, China

² Department of Data-Driven Materials Engineering, Faculty of Science and Technology, Nanjing University, Nanjing, China

Introduction

Materials machine learning routinely confronts small-data regimes. A new high-entropy alloy composition space may be represented by only 80 DFT-relaxed structures; a specific grain-boundary defect type in a ceramic may have fewer than 150 labeled examples; a metastable polymorph discovered under extreme pressure may come with just 40 reliable property labels. In these settings, training a graph neural network or message-passing potential from scratch is

statistically fragile: the model lacks sufficient examples to learn chemically meaningful representations and generalizes poorly to unseen structures [1-3]. Transfer learning—pre-training a model on a large, diverse source corpus (often the Materials Project or a general-purpose interatomic-potential dataset) and then fine-tuning on the sparse target set—appears to be the rational response [4-7]. The source model has already internalized broad chemical and structural patterns; fine-tuning ostensibly adapts these patterns with minimal additional data.

Yet the opposite outcome is common. After fine-tuning, the model not only fails to surpass the scratch-trained baseline but frequently performs worse, sometimes catastrophically so. Performance on the target test set degrades, force predictions become unstable, and the model exhibits extreme sensitivity to random seeds and learning-rate schedules [8]. This systematic collapse is not a failure of implementation but a predictable consequence of the interaction between pre-trained representations, small target sample sizes, and materials-specific domain shifts.

Four failure modes are formalized—catastrophic forgetting, negative transfer, representation distortion, and overfitting amplification—each with detectable signatures and materials-specific manifestations. Detection principles and mitigation strategies are articulated so that practitioners can identify collapse before deployment. The paper thereby shifts the discourse from “transfer learning improves sample efficiency” to a more nuanced position: in small-data materials regimes, fine-tuning often collapses unless explicit safeguards are applied. This reframing is essential for the responsible advancement of data-driven materials engineering, where over-reliance on unexamined transfer pipelines risks propagating biased or unreliable predictions into downstream design workflows [9].

Why Transfer Learning Is Attractive (And Dangerous)

The attractiveness of transfer learning in materials science is immediate and intuitive. Pre-trained models have been exposed to millions of atomic configurations spanning diverse chemistries and crystal symmetries [4, 10, 11]. They encode chemically plausible inductive biases—bond-length distributions, coordination-number statistics, and elemental embedding spaces—that appear universal [5, 12, 13]. Fine-tuning then requires only a few hundred target structures, a fraction of the data needed for scratch training [6]. In principle, the practitioner inherits a feature extractor already optimized for interatomic interactions and merely adapts the final layers or the entire network with a modest learning rate. Recent successes in transferring neural network potentials across related chemical spaces reinforce this optimism [6].

The danger, however, lies in the unspoken assumptions. First, source and target domains are never perfectly aligned. Source datasets are overwhelmingly biased toward ordered, thermodynamically stable, cubic or high-symmetry crystals computed at 0 K [14, 15]. Target tasks frequently involve disordered solid solutions, amorphous phases, grain boundaries, or finite-temperature ensembles [4]. The feature distributions therefore shift dramatically. Second, small target datasets cannot supply a sufficiently strong corrective signal to realign the pre-trained representations. With fewer than 1000 samples, gradient updates remain noisy and insufficient to unlearn source biases while simultaneously learning target-specific physics.

Third, the very mechanisms that make pre-training powerful—deep feature hierarchies and equivariant message passing—become liabilities under fine-tuning. Parameters optimized for source tasks become entangled; altering them to fit the target risks distorting representations that were previously useful [16, 17]. Materials-specific challenges compound the problem. Graph networks trained on the Materials Project excel at predicting formation energies of stable inorganic compounds but struggle when transferred to metastable or defective structures whose local environments deviate sharply from the source distribution [2, 9]. The assumption that “more data is always better” fails precisely because the additional source data encode biases orthogonal or even antagonistic to the target.

Consequently, fine-tuning can produce models that are simultaneously worse than the zero-shot source model (because target adaptation has corrupted useful source knowledge) and worse than a scratch-trained model (because the inherited parameters introduce harmful inductive biases). This dual degradation is the hallmark of transfer collapse in small-data regimes and explains why practitioners repeatedly observe that “transfer did not help and sometimes hurt” [18]. The

danger is not merely suboptimal performance; it is the silent propagation of misleading predictions into materials discovery pipelines.

Why Small-Data Regimes Are Especially Vulnerable

Small-data regimes amplify every vulnerability inherent to transfer learning. With target datasets typically below 1000 structures—and often closer to 100—the fine-tuning process operates under extreme statistical constraints [1, 19, 20]. Five interlocking reasons render these regimes uniquely fragile.

First, fine-tuning updates exhibit high variance. Each gradient step is dominated by the limited target samples, producing noisy parameter updates that easily erase useful source knowledge before target patterns can be internalized [16, 21]. Second, the corrective signal is inherently weak. When source features are systematically misaligned with the target (e.g., pre-training on perfect crystals versus defective structures), the small target set lacks the statistical power to override those misalignments [9, 15]. The model therefore retains incorrect priors.

Third, validation for early stopping becomes unreliable. Small datasets force tiny validation splits, yielding high-variance performance estimates that cannot reliably signal the onset of forgetting or overfitting. Practitioners are left without trustworthy stopping criteria. Fourth, the source–target domain gap frequently exceeds the corrective capacity of the target data. A model pre-trained on 0 K DFT energies cannot be meaningfully adapted to 1000 K molecular-dynamics-derived forces when only 50 target configurations are available; the thermodynamic and kinetic differences overwhelm the sparse signal [14, 22].

Fifth, feature representations themselves become mismatched. Pre-trained layers have optimized internal activations for source-domain statistics. In the target domain these activations may be irrelevant or actively misleading, yet the small data volume prevents the network from learning new, disentangled features [18, 23]. Collectively, these factors create a regime in which transfer learning is not merely less effective but systematically counterproductive. The small target sample size prevents the model from either preserving source utility or acquiring target competence, leading to the observed collapse [1].

Figure 1 presents a strictly hierarchical failure architecture demonstrating how source–target misalignment propagates through fine-tuning dynamics to produce systematic transfer learning collapse.

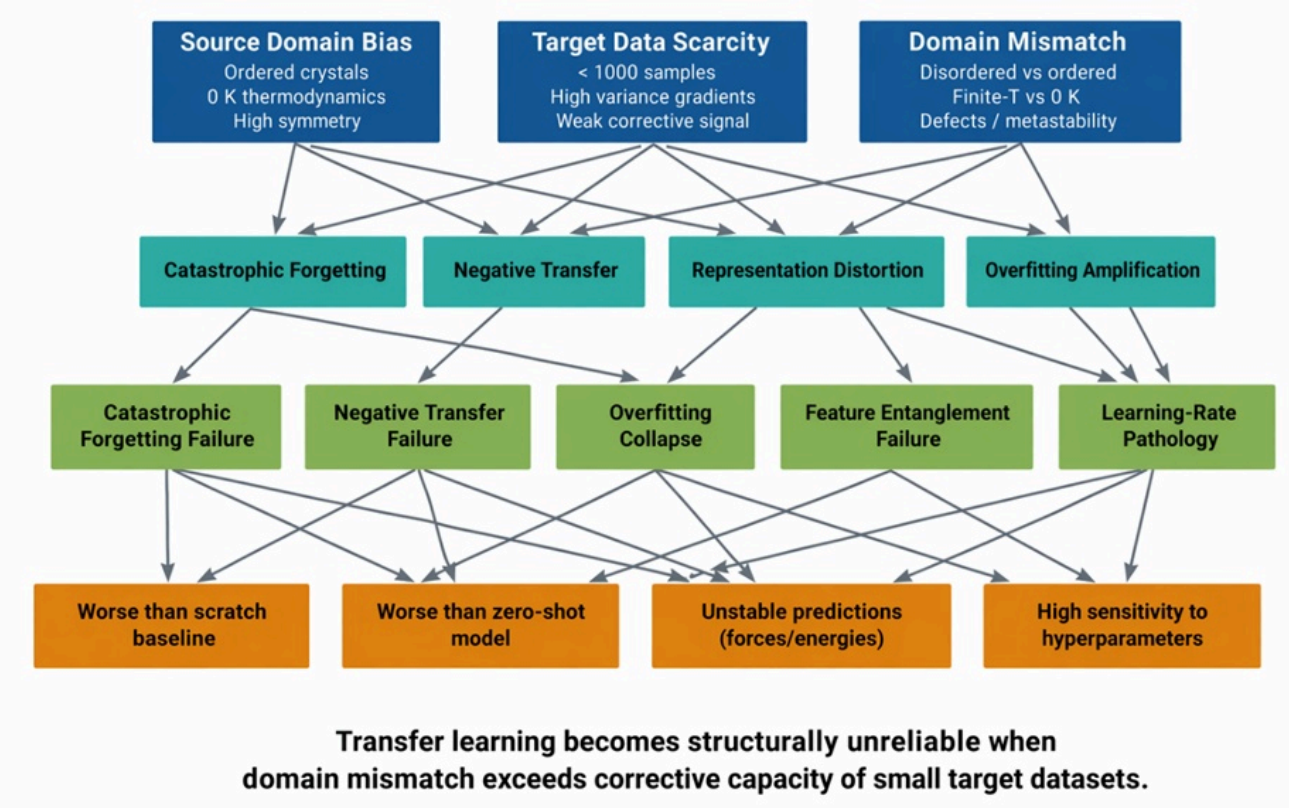


Figure 1. Hierarchical Failure Pathways of Transfer Learning Collapse in Small-Data Materials Regimes

Mechanisms of Transfer Collapse

The collapse of transfer learning reflects a tightly coupled set of mechanisms that manifest consistently across graph neural networks and machine-learned potentials [6, 16, 18, 24]. Catastrophic forgetting arises as fine-tuning disrupts parameters encoding source-domain structure without establishing comparably robust target representations, a process intensified by limited data availability such that prior chemical knowledge is lost more rapidly than new structure can be inferred, ultimately yielding performance that underperforms both frozen and randomly initialized baselines [16]. This degradation is further compounded when inherited inductive biases become misaligned with the target domain, giving rise to negative transfer in which pre-trained representations actively misguide prediction; biases favoring face-centered-cubic stability, for example, can obstruct accurate modeling of ionic oxides and suppress predictive fidelity below that of models trained from scratch [18]. A related consequence emerges at the level of internal representations, where fine-tuning induces distortions in feature space: embeddings originally optimized for the source domain are reshaped under data scarcity, yet insufficient signal prevents the emergence of genuinely task-specific structure, producing hybrid representations that generalize poorly across both domains [9]. This instability is amplified by the scale of contemporary pre-trained models, whose extensive parameterization enables rapid minimization of training error even in regimes with on the order of 100 samples, but in practice accelerates the memorization of noise rather than the extraction of physically meaningful patterns [1, 25]. Under these conditions, the interaction among forgetting, bias misalignment, representational degradation, and overfitting generates a self-reinforcing dynamic that systematically undermines fine-tuning in small-data materials contexts.

Table 1 consolidates the structural mapping between underlying mechanisms, observable failure modes, and their diagnostic signatures.

Table 1. Mechanism–Failure Mapping and Diagnostic Signatures in Transfer Learning Collapse

Mechanism	Core Process Description	Observable Failure Mode	Diagnostic Signature	Primary Risk Driver
Catastrophic Forgetting	Overwriting of source-optimized parameters during fine-tuning	Catastrophic forgetting failure	>30–50% drop in source validation performance	High learning rate + small target data
Negative Transfer	Misaligned source priors degrade target predictions	Negative transfer failure	Fine-tuned model < scratch baseline	Domain mismatch
Representation Distortion	Warping of internal feature space under limited supervision	Feature entanglement failure	Improvement in one property, degradation in another	Feature entanglement
Overfitting Amplification	Rapid memorization due to high parameter count	Overfitting collapse	Training–test error gap >2×	Data scarcity
Multi-Mechanism Interaction	Coupled effects across mechanisms	Learning-rate pathology	Extreme sensitivity across LR sweep (≥10× variation)	Hyperparameter instability

A Typology of Failure Modes

Five distinct yet overlapping failure modes characterize the collapse.

Catastrophic forgetting becomes evident when source validation accuracy collapses by more than half while target performance fails to surpass a scratch baseline, indicating that prior chemical regularities are erased without being productively reconfigured; this dynamic is exemplified when a model pre-trained on oxide chemistry and subsequently exposed to only a limited set of sulfide structures loses established oxide bonding rules yet does not internalize sulfide-specific patterns [16, 21]. A related degradation arises through negative transfer, where inherited inductive biases—such as cubic-metal assumptions—actively misguide predictions on structurally distinct targets, leading the fine-tuned model to underperform relative to one trained from scratch [18]. Beyond this misalignment, overfitting introduces a more insidious limitation: the model achieves near-perfect training accuracy while sustaining substantial test error, often exceeding a twofold gap, thereby undermining any capacity for extrapolation under sparse data conditions [1, 25]. This instability is further compounded when representational entanglement prevents selective adaptation, such that improvements in one predicted property, such as energy, coincide with deterioration in another, such as forces, reflecting an inability to disentangle correlated physical descriptors [6]. Under these conditions, optimization itself becomes unreliable, as learning dynamics exhibit acute sensitivity to parameterization, with modest adjustments in learning rate oscillating between stagnation and rapid forgetting without converging to a stable regime [16].

Detection Principles

Detection of transfer collapse in small-data materials regimes can be grounded in a set of operational diagnostics that require neither additional data nor architectural modification, instead leveraging only the existing source and target splits to surface failure prior to deployment. A central requirement is systematic baseline comparison, in which the fine-tuned model is evaluated alongside both the frozen source model under zero-shot conditions and a scratch-trained counterpart using identical target data; underperformance relative to either reference provides unambiguous evidence that transfer

has degraded rather than enhanced predictive capacity [1, 6, 15]. This evaluation gains interpretive depth when paired with continuous monitoring of source-domain validation behavior during fine-tuning, where declines exceeding 30 %—particularly when coupled with stagnant or deteriorating target metrics—indicate that previously encoded chemical regularities are being overwritten without compensatory learning of target-specific structure–property relationships [16]. Sensitivity of the optimization landscape further reveals itself through learning-rate perturbations, as performance instability spanning orders of magnitude, or the absence of any rate that simultaneously preserves source knowledge and improves target accuracy, signals a pathological training regime in which transfer cannot be reliably stabilized [16]. The emergence of overfitting can be diagnosed through the evolving divergence between training and test error on the target set, with rapidly widening gaps beyond a factor of two reflecting the mismatch between model capacity and the limited scale of available structures, often below 1000 instances [1, 25]. Complementing these internal signals, the degree of alignment between source and target domains can be quantified using established similarity measures, such as embedding-space proximity or compositional overlap, where values below 0.3 are associated with a heightened risk of negative transfer and suggest that continued fine-tuning is counterproductive relative to scratch training [4, 12, 18]. Together, these diagnostics reframe failure identification as an integrated, anticipatory process embedded within standard fine-tuning workflows, enabling early intervention before degraded models propagate into downstream materials design pipelines [9].

Mitigation Principles

Once collapse is identified, mitigation can proceed through a set of empirically grounded interventions that operate within the same data constraints while directly targeting the mechanisms underlying failure. A stabilizing entry point lies in progressive unfreezing, where early layers encoding general atomic features remain fixed as adaptation is initially confined to later layers and the output head, with deeper parameters only released once target performance ceases to fluctuate, thereby preserving transferable structure while enabling gradual specialization [3, 5, 6]. This controlled adaptation is reinforced by elastic weight consolidation, which introduces a Fisher-weighted quadratic penalty on parameter updates to selectively constrain changes to components most critical for source-domain competence, effectively counteracting the erosion of previously acquired knowledge [16, 21]. Complementing this constraint, reducing the learning rate by one to two orders of magnitude relative to scratch training moderates update magnitudes, limiting the risk of destabilizing shifts that would otherwise induce forgetting or exacerbate overfitting in low-data regimes [16, 18]. Retention of prior representations can be further sustained through intermittent replay of source-domain samples during fine-tuning, where even a small fraction within each mini-batch maintains alignment with earlier distributions and dampens both forgetting and bias misalignment [16]. A related refinement emerges in two-stage fine-tuning, wherein the predictor head is first optimized in isolation before the full network undergoes low-intensity adjustment, allowing the mapping to adapt ahead of deeper representational shifts [6, 13]. The propensity of high-capacity models to overfit under data scarcity can be attenuated through the targeted introduction of weight decay or dropout during fine-tuning, constraining parameter growth without disrupting learned structure [1, 25]. At a more strategic level, similarity-gated decision rules provide an ex ante filter, as low correspondence between source and target domains—particularly below 0.2 or under extremely limited sample sizes—signals that transfer is likely to be detrimental, favoring instead a carefully regularized scratch approach [4, 12, 18]. In combination, these interventions form an interdependent framework in which stabilization, constraint, and selective adaptation operate in concert, transforming transfer learning into a controlled and diagnostically informed process aligned with the data limitations of materials engineering [1, 19].

Table 2 introduces a decision-oriented framework that operationalizes when transfer learning should be applied, modified, or avoided.

Table 2. Decision Framework for Safe Deployment of Transfer Learning in Small-Data Materials Regimes

Condition	Threshold Indicator	Recommended Strategy	Rationale	Expected Outcome

High source–target similarity	Similarity > 0.5	Standard fine-tuning	Source features transferable	Performance gain likely
Moderate similarity + limited data	Similarity 0.2–0.5, 100–500 samples	Progressive unfreezing + low LR	Controlled adaptation required	Stable but modest improvement
Low similarity	Similarity < 0.2–0.3	Avoid transfer; train from scratch	High risk of negative transfer	More reliable generalization
Extremely small dataset	<100 samples	Scratch + strong regularization	Insufficient corrective signal	Reduced overfitting risk
Evidence of forgetting	Source validation drop >30%	Apply EWC or replay buffer	Preserve critical source knowledge	Stabilized training
High LR sensitivity	No stable performance across LR sweep	Reduce LR (1e-5–1e-4) + two-stage tuning	Prevent destructive updates	Improved convergence stability
Overfitting detected	Training–test gap >2×	Add dropout / weight decay	Limit memorization	Better generalization

Relation to Other Failure Modes

The collapse modes identified here intersect with—but remain distinct from—other documented failure mechanisms in materials machine learning. Catastrophic forgetting has been examined in the context of sequential learning across multiple alloy families [16], yet the present analysis focuses on single-shot source-to-target transfer rather than continual learning. The core mechanism (overwriting of source parameters) is shared, but the materials-specific trigger—abrupt domain shift from ordered crystals to disordered or metastable structures—adds a layer absent in purely sequential settings.

Representation distortion parallels the over-smoothing phenomenon observed in deep graph networks [2, 4, 9]. Over-smoothing collapses node representations toward a global mean as depth increases; transfer collapse similarly flattens useful feature hierarchies when fine-tuning distorts them on insufficient data. Both problems demand architectural or regularization interventions, yet transfer collapse is uniquely sensitive to dataset size rather than network depth alone.

Negative transfer, long recognized in general machine learning [26, 27], acquires distinct materials causes: the systematic bias of source datasets toward zero-kelvin ordered phases versus target needs for finite-temperature or defective configurations [14, 18]. This domain mismatch is more pronounced in chemistry and materials than in vision or language tasks, rendering generic negative-transfer mitigations insufficient without materials-aware similarity checks [18].

Overfitting amplification relates to capacity-control issues in high-dimensional potentials [1, 6, 10, 11, 23, 24, 28, 29] but is uniquely aggravated by the inheritance of millions of pre-optimized parameters. The failure modes therefore form a coherent family: each exploits the interaction of small target size with pre-trained inductive biases. Recognizing these relations allows researchers to borrow techniques across sub-fields while tailoring them to the unique statistical and chemical constraints of materials data.

Implications for Transfer Learning Practice

The systematic collapse of transfer learning carries immediate consequences for three stakeholder groups.

For materials ML researchers the central recommendation is epistemic humility: never assume transfer learning improves performance in small-data regimes. Every published claim must include scratch and zero-shot baselines, source-

validation curves, and learning-rate sensitivity analysis [1, 6, 15]. Reporting only “improvement over scratch” without these diagnostics risks propagating misleading conclusions. Default practice should shift toward progressive unfreezing plus EWC, with similarity gating applied before any fine-tuning begins [16, 18].

For benchmark designers the implication is equally clear. Future transfer-learning benchmarks must incorporate controlled domain shifts (ordered $\rightarrow$ disordered, 0 K $\rightarrow$ finite T, stoichiometric $\rightarrow$ defective) and report not merely average improvement but the rate of outright failure [4, 9, 13]. Success/failure rates, rather than mean performance gains, become the more informative metric when small-data regimes dominate real-world applications.

For practitioners confronting target datasets smaller than 100 structures the safest path is often to train from scratch using aggressive regularization and data-augmentation strategies native to the target domain [1, 19, 20]. Transfer should be attempted only when source–target similarity exceeds 0.5 and at least 200–300 reliable target samples are available. In all other cases, the risk of collapse outweighs the promised sample-efficiency gains.

Adopting these practices will raise the reliability of machine-learned interatomic potentials and property predictors, ensuring that data-driven materials engineering rests on reproducible rather than optimistic foundations.

Conclusion

This study demonstrates that transfer learning, while conceptually appealing, exhibits a systematic and predictable collapse in small-data materials regimes. Fine-tuning pre-trained graph neural networks and interatomic potentials frequently degrades performance, yielding models that underperform both frozen source models and scratch-trained baselines. This failure arises from the coupled interaction of catastrophic forgetting, negative transfer, representation distortion, and overfitting amplification—mechanisms fundamentally driven by domain mismatch, feature entanglement, and insufficient statistical signal in target datasets.

A key contribution of this work is the establishment of transfer failure as a *structural phenomenon* rather than an implementation artifact. In regimes characterized by fewer than $\sim 10^3$ samples, the optimization process cannot simultaneously preserve source knowledge and learn target-specific physics, leading to unstable and unreliable representations. This insight directly challenges the prevailing assumption that pre-training universally enhances sample efficiency in materials machine learning.

Importantly, the analysis demonstrates that transfer collapse is both detectable and, to a limited extent, controllable. Practical diagnostic protocols—such as rigorous baseline comparisons, monitoring of source-domain degradation, learning-rate sensitivity analysis, and domain-similarity quantification—enable early identification of failure. Complementary mitigation strategies, including progressive unfreezing, elastic weight consolidation, replay-based regularization, and conservative optimization schedules, provide a pathway toward stabilizing transfer when domain alignment is sufficiently high.

Nevertheless, the findings establish clear boundaries for safe deployment. When source–target similarity is low or target datasets are extremely limited, transfer learning is not merely ineffective but detrimental; under such conditions, carefully regularized training from scratch remains the more reliable approach. Accordingly, this work advocates a shift in practice from default adoption of transfer learning toward *conditional, diagnostically informed use*.

Looking forward, progress in materials machine learning will depend on embracing these limitations. Future research must prioritize failure-aware benchmarking, explicit reporting of negative transfer outcomes, and the development of domain-adaptive methodologies that account for the unique statistical and physical characteristics of materials data. Transfer learning should not be abandoned, but its application must be governed by principled criteria that reflect its inherent risks. Only through such rigor can data-driven materials science achieve both predictive accuracy and scientific reliability.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 31 Oct 2022

Revised: 25 Jan 2023

Accepted: 20 Apr 2023

Published online: 18 July 2023

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

- Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. *NPJ Comput Mater.* 2023;9(1):42.
<https://doi.org/10.1038/s41524-023-01000-z>.
- Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120(14):145301.
<https://doi.org/10.1103/PhysRevLett.120.145301>.
- Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet: A deep learning architecture for molecules and materials. *J Chem Phys.* 2018;148(24):241722.
<https://doi.org/10.1063/1.5019779>.
- 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>.

Smith JS, Nebgen BT, Zubatyuk R, Lubbers N, Devereux C, Barros K, et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun.* 2019;10(1):2903.
<https://doi.org/10.1038/s41467-019-10827-4>.

Schleder GR, Padilha AC, Acosta CM, Costa M, Fazzio A. From DFT to machine learning: Recent approaches to materials science—a review. *J Phys Mater.* 2019;2(3):032001.
<https://doi.org/10.1088/2515-7639/ab084b>.

Imbalzano G, Zhuang Y, Kapil V, Rossi K, Engel EA, Grasselli F, et al. Uncertainty estimation for molecular dynamics and sampling. *J Chem Phys.* 2021;154(7):074102.
<https://doi.org/10.1063/5.0036522>.

Chen C, Zuo Y, Ye W, Li X, Ong SP. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat Comput Sci.* 2021;1(1):46-53.
<https://doi.org/10.1038/s43588-020-00002-x>.

Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys Rev Lett.* 2018;120(14):143001.
<https://doi.org/10.1103/PhysRevLett.120.143001>.

Wang H, Zhang L, Han J, E W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput Phys Commun.* 2018;228:178-84.
<https://doi.org/10.1016/j.cpc.2018.03.016>.

Choudhary K, DeCost B, Tavazza F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys Rev Mater.* 2018;2(8):083801.
<https://doi.org/10.1103/PhysRevMaterials.2.083801>.

Schütt K, Kindermans PJ, Sauceda HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: *Advances in neural information processing systems* 30; 2017. p. 992-1002.

Bartók AP, De S, Poelking C, Bernstein N, Kermode JR, Csányi G, et al. Machine learning unifies the modeling of materials and molecules. *Sci Adv.* 2017;3(12):e1701816.
<https://doi.org/10.1126/sciadv.1701816>.

Zuo Y, Chen C, Li X, Deng Z, Chen Y, Behler J, et al. Performance and cost assessment of machine learning interatomic potentials. *J Phys Chem A.* 2020;124(4):731-45.
<https://doi.org/10.1021/acs.jpca.9b08723>.

Kirkpatrick J, Pascanu R, Rabinowitz N, Veness J, Desjardins G, Rusu AA, et al. Overcoming catastrophic forgetting in neural networks. *Proc Natl Acad Sci U S A.* 2017;114(13):3521-6.
<https://doi.org/10.1073/pnas.1611835114>.

Lubbers N, Smith JS, Barros K. Hierarchical modeling of molecular energies using a deep neural network. *J Chem Phys.* 2018;148(24):241715.
<https://doi.org/10.1063/1.5011181>.

Gupta V, Choudhary K, Tavazza F, Campbell C, Liao WK, Choudhary A, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat Commun.* 2021;12(1):6595.
<https://doi.org/10.1038/s41467-021-26921-5>.

Vanpoucke DEP, van Knippenberg OSJ, Hermans K, Bernaerts KV, Mehrkanoon S. Small data materials design with machine learning: When the average model knows best. *J Appl Phys.* 2020;128(5):054901.
<https://doi.org/10.1063/5.0012285>.

Jinnouchi R, Miwa K, Karsai F, Kresse G, Asahi R. On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *J Phys Chem Lett.* 2020;11(17):6946-55.
<https://doi.org/10.1021/acs.jpclett.0c01061>.

Liu H, Yang Y, Wang X. Overcoming catastrophic forgetting in graph neural networks. *Proc AAAI Conf Artif Intell.* 2021;35(10):8653-61.
<https://doi.org/10.1609/aaai.v35i10.17049>.

Chmiela S, Tkatchenko A, Sauceda HE, Poltavsky I, Schütt KT, Müller KR. Machine learning of accurate energy-conserving molecular force fields. *Sci Adv.* 2017;3(5):e1603015.
<https://doi.org/10.1126/sciadv.1603015>.

Behler J. First-principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew Chem Int Ed Engl.* 2017;56(42):12828-40.
<https://doi.org/10.1002/anie.201703114>.

Unke OT, Muwly M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J Chem Theory Comput.* 2019;15(6):3678-93.
<https://doi.org/10.1021/acs.jctc.9b00181>.

Wu YC, Feng JW. Development and application of artificial neural network. *Wirel Pers Commun.* 2018;102(2):1645-56.
<https://doi.org/10.1007/s11277-017-5224-x>.

Sze V, Chen YH, Yang TJ, Emer JS. Efficient processing of deep neural networks: A tutorial and survey. *Proc IEEE.* 2017;105(12):2295-329.
<https://doi.org/10.1109/JPROC.2017.2761740>.

Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre-training graph neural networks [Preprint]. *arXiv*; 2019. arXiv:1905.12265.

Bartók AP, Kermode JR, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. *Phys Rev X.* 2018;8(4):041048.
<https://doi.org/10.1103/PhysRevX.8.041048>.

Wood MA, Thompson AP. Extending the accuracy of the SNAP interatomic potential form. *J Chem Phys.* 2018;148(24):241721.
<https://doi.org/10.1063/1.5017641>.