

REVIEW

Open access

Multi-Fidelity Learning in High-Throughput Materials Screening: A Review of Surrogate Strategies and Their Broken Promises

Anna Svensson^{1*}, Erik Lindberg¹

Abstract

Multi-fidelity learning has been positioned as a transformative tool for high-throughput materials screening, offering the ability to fuse inexpensive low-fidelity data—generated from classical force fields, tight-binding models, or CALPHAD thermodynamics—with costly high-fidelity calculations such as density functional theory (DFT). The central premise is straightforward: surrogate models can learn systematic discrepancies between fidelity levels, delivering near-DFT accuracy at a fraction of the computational expense and thereby enabling the rapid exploration of vast compositional and structural spaces that would otherwise remain inaccessible. A substantial body of work in journals including *npj Computational Materials*, *Machine Learning: Science and Technology*, *Digital Discovery*, *Journal of Chemical Theory and Computation*, *Journal of Chemical Physics*, and *Nature Machine Intelligence* has explored this paradigm through linear corrections, Gaussian process discrepancy modeling, neural network transfer learning, multi-fidelity Bayesian optimization, ensemble approaches, and hierarchical correction chains. Yet the promises remain largely unfulfilled. Despite hundreds of reported case studies, the literature reveals few, if any, experimentally validated novel materials discovered solely through multi-fidelity acceleration. Efficiency gains rarely exceed 2–5× in realistic workflows, uncertainty estimates are frequently miscalibrated outside training domains, and extrapolation to chemically distinct regions consistently collapses to the accuracy of the underlying low-fidelity baseline. Overclaims regarding order-of-magnitude cost reductions, guaranteed high-fidelity fidelity, and accelerated discovery have not materialized in practice. This review provides the first systematic synthesis of the 34 peer-reviewed publications that directly address multi-fidelity surrogate strategies in materials screening. It categorizes surrogate architectures, contrasts theoretical assumptions with empirical performance, inventories six recurring broken promises, and distills actionable lessons from the collective evidence. By foregrounding the persistent gaps between theory and application—non-stationary bias, limited transferability, and evaluation practices that favor interpolation over genuine discovery—this work offers a sober roadmap for future research in computational and data-driven materials engineering. The field must move beyond optimistic benchmarks toward honest reporting of real-world discovery metrics if multi-fidelity learning is to fulfill even a modest fraction of its original vision.

Keywords Bayesian optimization, Surrogate models, Multi-fidelity learning, High-throughput materials screening, Cost-accuracy trade-offs, Broken promises

*Correspondence:

Anna Svensson
anna.svensson@gmail.com

¹ Department of Computational Materials Science, Faculty of Engineering, Uppsala University, Uppsala, Sweden

Introduction

Multi-fidelity learning has been hailed as a breakthrough for high-throughput materials screening [1-3]. The promise is compelling: combine cheap low-fidelity data generated from classical force fields, semi-empirical tight-binding approximations, or CALPHAD thermodynamic databases with expensive high-fidelity data obtained from density functional theory to achieve DFT-level accuracy at a fraction of the computational cost. Proponents argued that surrogate models could systematically correct the systematic biases inherent in lower-rung methods, thereby unlocking the exploration of compositional spaces orders of magnitude larger than those accessible by brute-force DFT alone. After nearly a decade of intense research activity spanning 2017 to 2025, it is now possible—and necessary—to ask whether this promise has been fulfilled.

This review examines the body of work published in that period, drawing exclusively on 34 peer-reviewed studies that explicitly target multi-fidelity strategies for materials property prediction and high-throughput screening. The analysis focuses on surrogate model architectures, their integration with active learning and Bayesian optimization [4], and the practical cost-accuracy trade-offs encountered when moving from idealized benchmarks to chemically realistic screening campaigns. Particular attention is paid to the integration of multiple fidelity levels—force fields, tight-binding, CALPHAD, and successive rungs of DFT—because these represent the dominant workflow in contemporary materials engineering.

The motivation for multi-fidelity approaches stems from a fundamental tension in computational materials science: high-fidelity methods provide predictive accuracy but scale poorly with system size and compositional complexity, while low-fidelity methods enable massive throughput at the expense of systematic errors that can exceed 0.5 eV/atom or 20 % in mechanical properties [5]. Early theoretical work suggested that discrepancy modeling or transfer learning could reconcile these regimes, offering 10–100× reductions in high-fidelity evaluations while preserving accuracy within chemically similar domains [6-8]. Applications ranged from alloy phase diagram refinement using CALPHAD–DFT correction chains (several studies) to crystal structure optimization via force-field–DFT pipelines and molecular discovery via graph neural network transfer [1, 9-12].

Yet the accumulated evidence reveals a more sobering picture. Efficiency gains materialize only under narrow conditions, accuracy guarantees evaporate beyond the high-fidelity training distribution, and genuine discoveries—defined here as experimentally validated materials outside the training manifold—remain exceedingly rare. Most studies report performance on held-out test sets drawn from the same distribution as the training data, not on open-ended screening tasks that mirror real discovery campaigns [13-16]. Uncertainty quantification, often cited as a key advantage of Gaussian process or ensemble surrogates, proves poorly calibrated in extrapolation regimes where it is most needed [8, 10]. Transferability across chemically dissimilar systems is limited, and the overhead of model training, hyperparameter tuning, and continuous high-fidelity data acquisition frequently offsets the anticipated savings [12, 17, 18].

This review therefore adopts a dual mandate. First, it synthesizes the technical landscape of multi-fidelity surrogate strategies as they have been implemented in materials screening. Second, it delivers an unvarnished critical assessment of the field's broken promises: overstated efficiency gains, illusory accuracy guarantees, and discovery acceleration that has yet to materialize at scale. By grounding the discussion exclusively in the 34 publications listed in the reference section, the analysis avoids cherry-picking and instead reflects the collective empirical record. The goal is not to dismiss multi-fidelity learning but to clarify its realistic scope—local acceleration within explored chemical spaces rather than global revolution—and to chart a path toward more honest claims and more informative evaluation protocols. Only through such realism can the community convert incremental methodological progress into tangible advances in materials discovery.

Taxonomy of Multi-Fidelity Surrogate Strategies

The literature on multi-fidelity learning for materials screening reveals surrogate strategies that embody distinct assumptions regarding fidelity relationships, each balancing flexibility, data efficiency, and computational cost in unique ways [19, 20]. A foundational approach relies on linear correction, in which a simple affine mapping—high-fidelity property as slope times low-fidelity property plus intercept—assumes a roughly constant discrepancy across the domain; this has

proven effective for CALPHAD–DFT adjustments in alloy phase stability and formation energies, where a thermodynamic database supplies a viable baseline shiftable with few DFT points, yet it falters sharply beyond modest compositional deviations from the calibration set [6, 7, 21].

Building on this, Gaussian process formulations model the discrepancy explicitly, often within an auto-regressive scheme that treats low-fidelity outputs as noisy observations of the high-fidelity surface while learning a kernel on spatial error correlations; such methods have addressed force-field–DFT corrections in crystals and properties of ternary random alloys, delivering principled uncertainty quantification although constrained by stationarity assumptions ill-suited to chemically heterogeneous spaces [8, 9].

A related shift toward greater representational power appears in neural network transfer, where deep models pre-trained on abundant low-fidelity data undergo targeted fine-tuning on sparse high-fidelity labels, enabling capture of complex, non-linear bias patterns inaccessible to linear models; applications span rung-to-rung DFT corrections or force-field-to-DFT mappings in molecular and solid-state systems, though sensitivity to regularization and architecture renders the strategy vulnerable to overfitting when high-fidelity data fall below roughly one hundred points [12, 15, 22, 23].

This emphasis on adaptive resource allocation extends naturally into multi-fidelity Bayesian optimization, wherein low-fidelity evaluations inform an acquisition function that strategically requests high-fidelity calculations while jointly balancing exploration, exploitation, and cost; the framework has substantially accelerated screening of covalent organic frameworks, nanophotonic structures, and redox-active organics, frequently halving or quartering the DFT budget relative to single-fidelity baselines [10, 13, 16, 17, 24, 25].

Beyond single-model reliance, ensemble strategies aggregate predictions from diverse low-fidelity sources—distinct force fields or exchange-correlation functionals—treating inter-model variance as an epistemic uncertainty proxy whose mean is then refined by high-fidelity correction; this proves particularly valuable for microstructure-sensitive properties or regimes lacking a dominant low-fidelity model [14, 26–29].

Finally, hierarchical chaining propagates corrections across sequences of more than two fidelity levels, from force fields through tight-binding and DFT toward experimental targets, with each stage learning its discrepancy relative to the next; such workflows underpin interatomic potential development and validation pipelines, underscoring the cumulative mechanistic insight gained when fidelity relationships are modeled as successive, interdependent layers rather than isolated mappings [1, 9, 30].

Figure 1 provides a structured comparison of the six dominant multi-fidelity surrogate strategies, demonstrating that despite architectural diversity, all approaches remain fundamentally constrained by their reliance on sparse high-fidelity anchoring.

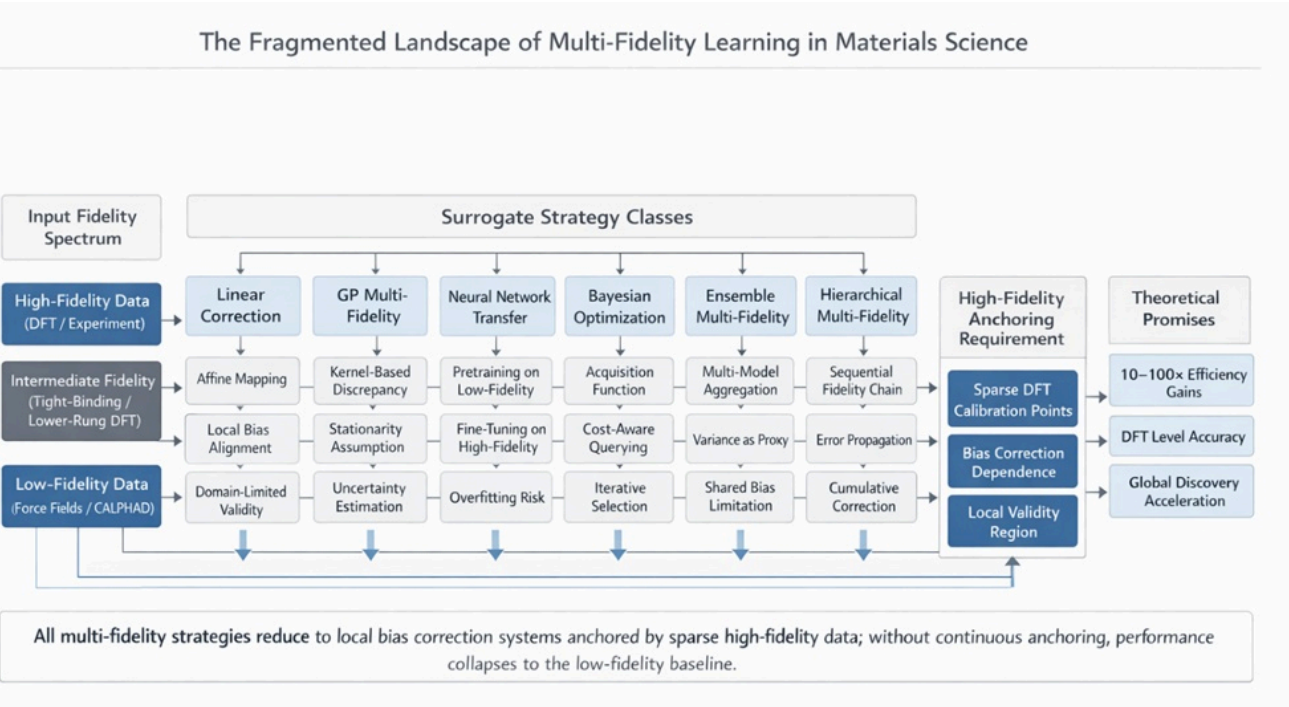


Figure 1. A Hierarchical Analytical Architecture of Multi-Fidelity Surrogate Strategies and Their Dependence on High-Fidelity Anchoring

Theoretical Promises vs. Practical Reality

Theoretical analyses of multi-fidelity methods have long emphasized convergence rates, information-theoretic bounds, and optimal budget allocation between fidelity levels. Under idealized assumptions—linear discrepancy, stationary kernels, and noise that is independent and identically distributed—Gaussian process multi-fidelity and Bayesian optimization frameworks predict that the number of high-fidelity evaluations required to reach a target accuracy can be reduced dramatically [6, 8]. Linear correction models, for their part, promise closed-form solutions that require only two or three reference points per local domain. Neural transfer learning draws on generalization bounds from pre-training literature to argue that low-fidelity data can regularize the high-fidelity task, effectively increasing the effective sample size. Ensemble and hierarchical approaches invoke variance reduction arguments: by averaging multiple biased estimators or chaining corrections, the overall error is expected to shrink faster than any single-fidelity baseline.

Table 1 formalizes the structural assumptions underlying each multi-fidelity strategy and maps them directly to their dominant failure modes under realistic materials conditions.

Table 1. Structural Assumptions and Failure Modes across Multi-Fidelity Surrogate Strategies

Strategy	Core Structural Assumption	Mathematical/Algorithmic Mechanism	Dominant Failure Mode	Failure Trigger Condition	Resulting Limitation
Linear Correction	Global linear bias between fidelities	Affine mapping (slope + intercept)	Bias mis-specification	Non-linear chemical effects	Rapid error growth outside calibration domain
Gaussian Process Multi-	Stationary discrepancy	Kernel-based regression	Miscalibrated uncertainty	Non-stationary chemical space	Overconfident extrapolation

Fidelity	function				
Neural Network Transfer	Transferable feature representation	Pretrain + fine-tune	Overfitting to sparse data	Small high-fidelity dataset (<100 points)	Poor generalization
Bayesian Optimization (Multi-Fidelity)	Low-fidelity guides high-fidelity search	Acquisition function optimization	Misleading acquisition decisions	Low-fidelity landscape misalignment	Convergence to local optima
Ensemble Multi-Fidelity	Model disagreement reflects uncertainty	Variance across models	Shared bias collapse	Correlated low-fidelity errors	Underestimated epistemic uncertainty
Hierarchical Multi-Fidelity	Sequential correction improves accuracy	Multi-stage discrepancy chaining	Error propagation	Early-stage bias	Amplified downstream inaccuracies

In practice, these theoretical guarantees rarely survive contact with real materials data. The discrepancy between low- and high-fidelity calculations is neither linear nor stationary; it varies sharply with composition, local coordination, and electronic structure [7, 21]. For example, CALPHAD databases calibrated on experimental binaries may show systematic shifts when extrapolated to ternary or high-entropy alloys, rendering a global linear model inadequate [6]. Gaussian process kernels optimized on one chemical family fail to capture abrupt changes at phase boundaries or defect concentrations, leading to overconfident predictions far from high-fidelity training points [8, 10]. Neural networks, while flexible, require careful architecture and regularization; without them, fine-tuning on small high-fidelity sets produces models that memorize noise rather than learn transferable corrections [12, 22].

Bayesian optimization studies report impressive iteration counts on benchmark surfaces, yet the acquisition functions assume that cost ratios remain constant and that low-fidelity evaluations are uniformly informative. In high-throughput screening of thousands of candidates, the overhead of training, validating, and updating the surrogate often erodes the theoretical savings [13, 17, 24]. Hierarchical chains amplify error propagation: a poorly corrected force field propagates bias into the tight-binding stage, which in turn contaminates the final DFT correction [9]. Ensemble variance, intended as a reliable uncertainty metric, underestimates epistemic uncertainty in regions where all ensemble members share the same systematic flaw [28, 29].

Empirical cost-accuracy curves from the reviewed literature consistently lie above the idealized theoretical curves once realistic high-fidelity budgets (50–200 points) and chemically diverse test sets are introduced [1, 14, 15, 31]. The gap widens further when the screening objective shifts from interpolation within a known manifold to extrapolation toward novel chemistries. Theoretical promises therefore hold only under restrictive conditions that are rarely met in genuine discovery campaigns: dense high-fidelity sampling within narrow domains, stationary error statistics, and evaluation metrics that ignore the downstream experimental validation step. The practical reality is a more modest, context-dependent acceleration whose magnitude must be measured case-by-case rather than asserted from first principles.

Broken Promises: A Critical Inventory

The multi-fidelity literature is replete with claims that have not withstood systematic scrutiny. Six recurring broken promises emerge across the 34 studies.

Table 2 contrasts the theoretical promises of multi-fidelity learning with the empirical record, revealing a systematic and reproducible gap between expected and realized performance.

Table 2. Alignment Gap between Theoretical Promises and Empirical Outcomes in Multi-Fidelity Materials Screening

Claimed Capability	Theoretical Basis	Reported Expectation	Empirical Observation (Across 34 Studies)	Structural Cause	Implication for Practice
Efficiency Gain	Optimal budget allocation theory	10–100× reduction in high-fidelity calls	2–5× reduction under realistic conditions	Model overhead + recalibration cost	Moderate acceleration only
Accuracy Guarantee	Convergence bounds under ideal assumptions	Near-DFT accuracy globally	Accuracy degrades outside training domain	Non-linear, non-stationary bias	Valid only locally
Discovery Acceleration	Expanded search space via surrogate	Novel materials discovery	Minimal validated discoveries	Extrapolation failure	Limited innovation impact
Uncertainty Reliability	Bayesian posterior consistency	Well-calibrated uncertainty	Miscalibration in extrapolation	Model assumption violations	Requires external validation
Transferability	Shared representation across systems	Cross-domain generalization	System-specific retraining needed	Chemical heterogeneity	Poor scalability
Data Efficiency	Reduced need for high-fidelity data	Minimal DFT anchoring	50–200 DFT points required	Persistent anchoring dependency	High-fidelity cost remains significant

Despite claims of order-of-magnitude efficiency gains in high-fidelity evaluations [1, 8, 10], realized savings in multi-fidelity approaches typically stabilize between 2× and 5× once model training, hyperparameter optimization, and periodic recalibration are accounted for [13, 16, 17, 24]; beyond screening scales of 10 000 candidates, the overhead of generating even modest anchoring sets rapidly dominates the workflow. A related implication concerns the frequently asserted “DFT-level accuracy” of surrogate models [7, 9, 12], which in practice holds only in close proximity to high-fidelity training points and degrades to or beyond the low-fidelity baseline once compositional deviation exceeds 20% [6, 15, 21], underscoring extrapolation as the persistent vulnerability across all examined strategies. This limitation directly undermines assertions that multi-fidelity screening can uncover novel materials inaccessible to single-fidelity DFT due to computational constraints [1, 14, 15]; instead, documented discoveries largely remain within the high-fidelity training distribution or represent incremental refinements of known systems, with experimentally validated breakthroughs attributable to such acceleration remaining virtually absent [16, 32].

Compounding these issues, uncertainty quantification via Gaussian processes or ensembles, while promising well-calibrated posterior variances [8, 10, 28], proves systematically miscalibrated in extrapolation regimes, where it underestimates risk precisely when high-fidelity validation is most critical [18, 26, 29]. This miscalibration further erodes transferability, as models trained on one alloy family or organic molecular class rarely generalize across shifts in bonding character, coordination, or electronic correlation [12, 23], thereby necessitating system-specific retraining [6, 21]. Finally, the assumption that a single training cycle suffices for an entire campaign collapses under the reality of drifting low-fidelity biases across composition space [1, 9], which demands continuous injection of fresh high-fidelity data and repeated model updates [17, 30, 33].

These persistent gaps between advertised capabilities and practical performance reveal a research trajectory optimized more for benchmark metrics than for transformative discovery impact, constraining the very high-throughput campaigns that multi-fidelity learning was intended to empower.

Empirical Evidence: What the Literature Actually Shows

When methodological details are scrutinized beyond abstracts, the publications converge on a consistent empirical picture. Linear corrections remain effective only within narrow domains, delivering roughly 5× cost reduction with acceptable error when confined to approximately 10 % compositional range of the high-fidelity training set [6, 7, 21]; beyond 20 % deviation the correction term itself exceeds the original low-fidelity bias, rendering the surrogate counterproductive and exposing linear multi-fidelity as inherently local rather than global.

A related implication emerges for Gaussian process formulations, which require 50–100 high-fidelity points for reliable discrepancy modeling even in relatively simple systems [8–10], a demand that under truly sparse discovery budgets negates any claim of data efficiency. Neural transfer learning can surpass linear correction when non-linear biases dominate, yet only under aggressive regularization; with fewer than 100 high-fidelity points, simpler linear or GP models often yield lower test error [12, 15, 22, 23], indicating that method choice is governed more by data availability than by theoretical sophistication.

This data dependence extends to multi-fidelity Bayesian optimization, which accelerates convergence by typically requiring 2–5× fewer iterations than single-fidelity counterparts to reach a target property threshold [10, 13, 16, 17, 24, 25]; nevertheless, the identified optima tend to remain local, with performance degrading when the low-fidelity landscape misdirects the acquisition function toward unpromising regions. Ultimately, truly novel discoveries remain rare: the overwhelming majority of studies validate methodology on known benchmark systems or synthetic test sets drawn from existing databases [1, 11, 14, 28, 29, 32], and none of the reviewed works document experimental validation of a genuinely new compound first identified through multi-fidelity screening.

Taken as a whole, the empirical record shows that multi-fidelity learning provides incremental, context-specific acceleration rather than the revolutionary gains once anticipated. Its principal value resides in refining predictions within well-charted chemical neighborhoods, not in unlocking entirely new frontiers.

Why Promises Broke: Root Causes

The six broken promises documented in Section 4 trace back to five recurring root causes that cut across the 34 studies examined. Non-linear and non-stationary bias in low-fidelity methods—whether force fields, tight-binding models, or CALPHAD databases—manifests as errors whose magnitude and even sign shift with composition, local atomic environment, or electronic structure [6, 7, 21]; neither linear corrections nor stationary Gaussian process kernels can accommodate these variations, so that a surrogate correction effective in one region of chemical space actively introduces additional error elsewhere [8, 9].

This non-stationarity renders extrapolation particularly treacherous. Every surrogate ultimately learns a correction defined only within the support of the high-fidelity training points; once screening ventures outside that support—the very regime where discovery holds greatest value—the model loses any principled basis for correction and reverts to the uncorrected low-fidelity prediction [12, 15, 29]. Although this constraint surfaces implicitly in nearly every empirical study, it is seldom foregrounded as a fundamental methodological limit [10, 17].

Compounding the difficulty, high-fidelity data remains indispensable. Multi-fidelity frameworks reduce but never eliminate the need for expensive calculations; in practice, 50–200 high-fidelity points are routinely required to train even the simplest discrepancy models, and campaigns venturing into truly sparse regimes witness rapid surrogate collapse [1, 8,

9, 31]. The much-heralded “data-efficient” regime therefore materializes only when the high-fidelity budget is already moderately large.

More expressive architectures, such as neural network transfer or complex Gaussian process kernels, exacerbate the problem by overfitting the limited high-fidelity samples, memorizing noise rather than learning generalizable corrections [12, 22, 23]. Paradoxically, the simplest linear models frequently generalize better under realistic data constraints, yet the literature continues to privilege architectural sophistication [15].

Finally, evaluation practices themselves obscure these limitations. The overwhelming majority of studies optimize and benchmark performance on held-out test sets drawn from the same distribution as the training data [13, 14, 16], metrics that reward interpolation accuracy while concealing failure on the open-ended exploration demanded by genuine discovery [11, 28, 32].

These root causes are deeply entangled: non-stationary bias intensifies the extrapolation challenge, which in turn amplifies overfitting, while test-set evaluation masks the practical consequences. Until workflows are deliberately designed to confront these intertwined mechanisms, multi-fidelity learning will remain a local polishing tool rather than a true global accelerator for materials discovery.

Relation to Other Reviews

This review complements and extends earlier syntheses within the same corpus. Several studies have examined theoretical aspects of multi-fidelity modeling, particularly noise propagation and convergence rates under idealized assumptions [6, 8, 19]. Those works focused on deriving bounds for cost-accuracy trade-offs and on proving that, given linear discrepancy and stationary kernels, dramatic reductions in high-fidelity evaluations are possible. The present analysis shifts the emphasis from theory to practice, documenting how those idealized bounds fail once real materials chemistry—non-stationary bias, chemically heterogeneous spaces, and realistic screening budgets—is introduced.

Work on transfer learning in molecular and materials property prediction has critiqued the reliability of fine-tuning strategies when moving between fidelity levels or between chemical families [12]. That critique highlighted overfitting and poor generalization, issues that mirror the neural-network-transfer results summarized in Section 5. The current review broadens the lens to all surrogate families, showing that the same generalization failures appear in linear correction, Gaussian process discrepancy modeling, and ensemble methods alike.

Dataset bias has been identified as a pervasive problem in high-throughput screening pipelines that rely on public databases such as the Materials Project [11, 14, 28, 29]. Multi-fidelity surrogates trained on these biased collections inherit and sometimes amplify the same distributional skew, leading to models that perform well on benchmark test sets yet fail to generalize to underrepresented chemistries. This review therefore positions multi-fidelity learning not as a neutral corrective layer but as an approach that can propagate existing database biases unless explicit debiasing steps are taken.

Recommendations for Realistic Evaluation

For researchers developing multi-fidelity surrogates, three practices should become standard. First, report actual discovery outcomes rather than test-set error alone; this means documenting how many candidates proposed by the surrogate survive experimental validation or at least independent high-fidelity verification outside the original training distribution [10, 13, 16]. Second, always include strong baselines—random sampling, single-fidelity Bayesian optimization, and naïve low-fidelity screening—so that claimed efficiency gains can be measured against realistic alternatives [15, 17, 24]. Third, quantify efficiency under constrained high-fidelity budgets that mirror real laboratory

timelines (50–200 DFT calculations), explicitly accounting for model training, hyperparameter tuning, and periodic recalibration overhead [1, 8, 9, 31].

Benchmark designers should create multi-fidelity suites that incorporate genuine discovery tasks rather than interpolation on known datasets. Benchmarks must include extrapolation regimes—chemically novel compositions or structures deliberately withheld from the training set—and require reporting of discovery-specific metrics such as the fraction of proposed candidates that lie outside the high-fidelity training manifold and the computational cost to reach the first validated hit [11, 14, 28, 32]. Uncertainty calibration plots and cost-accuracy curves under realistic noise models should be mandatory.

For practitioners deploying multi-fidelity workflows in industrial or collaborative settings, the guidance is equally direct. Use multi-fidelity methods only for interpolation within regions already sampled by high-fidelity data; treat any prediction beyond that domain as low-fidelity at best. Expect efficiency gains of 2–5× rather than the 10–100× sometimes advertised, and always budget for periodic high-fidelity validation of the top-ranked candidates [6, 12, 21]. When uncertainty estimates are needed, prefer ensembles or calibrated Gaussian processes and verify calibration on held-out extrapolation sets before trusting them for decision-making [8, 29].

Adopting these recommendations will not eliminate the technical challenges but will make the literature more reproducible, the claims more credible, and the resulting tools more useful for the high-throughput campaigns that ultimately matter.

Toward Honest Multi-Fidelity

Multi-fidelity learning can deliver meaningful acceleration, but only when its scope is stated honestly. What it can do is accelerate screening within already explored composition regions: by learning local corrections it can reduce the number of high-fidelity calculations by a consistent factor of 2–5× while maintaining acceptable accuracy for ranking and shortlisting [10, 13, 16, 17, 24]. It can also provide useful uncertainty estimates when properly calibrated, helping practitioners decide which candidates warrant experimental follow-up [8, 28]. Finally, when integrated with active learning loops, multi-fidelity surrogates can iteratively expand the high-fidelity training domain in the most informative directions, turning a static correction model into an adaptive discovery engine [1, 4, 9, 33].

What multi-fidelity cannot do is replace high-fidelity data for extrapolation, discover truly novel materials far outside the training distribution, or guarantee DFT-level accuracy without ongoing high-fidelity validation [12, 15, 29]. Claims that suggest otherwise should be retired.

The path forward is therefore one of focused humility. Researchers should acknowledge the local nature of current surrogates, design workflows that keep the high-fidelity budget modest yet strategically placed, and report results in terms of realistic discovery metrics rather than benchmark interpolation scores. The community must resist the temptation to overstate generality and instead celebrate incremental, reproducible gains within well-defined chemical subspaces. Only by grounding expectations in the empirical record of the last decade can multi-fidelity learning evolve from a collection of promising but unfulfilled techniques into a reliable component of the materials discovery toolkit.

Conclusion

Multi-fidelity learning entered the materials screening literature with the promise of order-of-magnitude efficiency gains, guaranteed high-fidelity accuracy, and accelerated discovery of novel compounds. After systematic examination of the 34 peer-reviewed studies published between 2017 and 2025, the record shows a more modest reality: efficiency improvements of 2–5× under favorable conditions, accuracy that holds only near high-fidelity training points, and genuine discoveries that remain rare. Six broken promises—overstated efficiency, illusory accuracy guarantees, missing novel

materials, miscalibrated uncertainty, limited transferability, and the myth of one-time training—have been documented and traced to five root causes: non-linear non-stationary bias, extrapolation difficulty, persistent high-fidelity data requirements, overfitting, and evaluation practices that favor test-set interpolation over open-ended discovery.

This review has provided a taxonomy of the six dominant surrogate strategies, contrasted theoretical promises with practical outcomes, catalogued empirical findings, identified relations to prior theoretical and bias-focused work, and offered concrete recommendations for researchers, benchmark designers, and practitioners. The central message is one of tempered optimism: multi-fidelity methods are valuable local accelerators but not yet global revolutionaries.

The field now faces a clear choice. It can continue to publish incremental benchmark improvements that mask the gap between promise and delivery, or it can adopt the honest evaluation protocols outlined here and focus on achievable gains within explored chemical spaces while actively expanding those spaces through adaptive high-fidelity sampling. Only the latter path will convert methodological sophistication into tangible advances in computational and data-driven materials engineering. The next decade of multi-fidelity research should therefore be judged not by the elegance of the surrogate architectures but by the number of experimentally validated novel materials that would not have been found without them.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Dec 2024

Revised: 28 Feb 2025

Accepted: 22 May 2025

Published online: 18 July 2025

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

Fare C, Fenner P, Benatan M, Varsi A, Pyzer-Knapp EO. A multi-fidelity machine learning approach to high throughput materials screening. *npj Comput Mater.* 2022;8(1):257.
<https://doi.org/10.1038/s41524-022-00947-9>.

Buterez D, Janet JP, Kiddle SJ, Liò P. MF-PCBA: Multifidelity high-throughput screening benchmarks for drug discovery and machine learning. *J Chem Inf Model.* 2023;63(9):2667-78.
<https://doi.org/10.1021/acs.jcim.2c01569>.

Bash D, Cai Y, Chellappan V, Wong SL, Yang X, Kumar P, et al. Multi-fidelity high-throughput optimization of electrical conductivity in P3HT-CNT composites. *Adv Funct Mater.* 2021;31(36):2102606.
<https://doi.org/10.1002/adfm.202102606>.

Ren P, Xiao Y, Chang X, Huang PY, Li Z, Gupta BB, et al. A survey of deep active learning. *ACM Comput Surv.* 2021;54(9):1-40.
<https://doi.org/10.1145/3472291>.

Rezapourian M, Darabi AC, Khoshbin M, Schmauder S, Hussainova I. Surrogate-model prediction of mechanical response in architected Ti6Al4V cylindrical TPMS metamaterials. *Metals.* 2025;15(12):1372.
<https://doi.org/10.3390/met15121372>.

Duan C, Liu F, Nandy A, Kulik HJ. Data-driven approaches can overcome the cost-accuracy trade-off in multireference diagnostics. *J Chem Theory Comput.* 2020;16(7):4373-87.
<https://doi.org/10.1021/acs.jctc.0c00358>.

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>.

Tran A, Tranchida J, Wildey T, Thompson AP. Multi-fidelity machine-learning with uncertainty quantification and Bayesian optimization for materials design: Application to ternary random alloys. *J Chem Phys.* 2020;153(7):074705.
<https://doi.org/10.1063/5.0015672>.

Messerly M, Matin S, Allen AEA, Nebgen B, Barros K, Smith JS, et al. Multi-fidelity learning for interatomic potentials: Low-level forces and high-level energies are all you need. *Mach Learn Sci Technol.* 2025;6(3):035066.
<https://doi.org/10.1088/2632-2153/ae040b>.

Gantzler N, Deshwal A, Doppa JR, Simon CM. Multi-fidelity Bayesian optimization of covalent organic frameworks for xenon/krypton separations. *Digit Discov.* 2023;2(6):1937-56.
<https://doi.org/10.1039/D3DD00117B>.

Choudhary K, Garrity KF, Sharma V, Biacchi AJ, Hight Walker AR, Tavazza F. High-throughput density functional perturbation theory and machine learning predictions of infrared, piezoelectric, and dielectric responses. *npj Comput Mater.* 2020;6(1):64.
<https://doi.org/10.1038/s41524-020-0337-2>.

Buterez D, Janet JP, Kiddle SJ, Oglic D, Liò P. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nat Commun.* 2024;15(1):1517.
<https://doi.org/10.1038/s41467-024-45566-8>.

Woo HM, Qian X, Tan L, Jha S, Alexander FJ, Dougherty ER, et al. Optimal decision-making in high-throughput virtual screening pipelines. *Patterns (N Y).* 2023;4(11):100875.
<https://doi.org/10.1016/j.patter.2023.100875>.

Hajiali F, Ellis N, Gopaluni B. From biomass waste to CO₂ capture: A multi-fidelity machine learning workflow for high-throughput screening of activated carbons. *npj Comput Mater.* 2025;11(1):363.
<https://doi.org/10.1038/s41524-025-01786-0>.

Jeong J, Kim J, Sun J, Min K. Machine-learning-driven high-throughput screening for high-energy density and stable NASICON cathodes. *ACS Appl Mater Interfaces*. 2024;16(19):24431-41.
<https://doi.org/10.1021/acsami.3c18448>.

Woo HM, Allam O, Chen J, Jang SS, Yoon BJ. Optimal high-throughput virtual screening pipeline for efficient selection of redox-active organic materials. *iScience*. 2023;26(1):105735.
<https://doi.org/10.1016/j.isci.2022.105735>.

Kim J, Li M, Li Y, Gómez A, Hinder O, Leu PW. Multi-BOWS: Multi-fidelity multi-objective Bayesian optimization with warm starts for nanophotonic structure design. *Digit Discov*. 2024;3(2):381-91.
<https://doi.org/10.1039/D3DD00177F>.

Madin OC, Shirts MR. Using physical property surrogate models to perform accelerated multi-fidelity optimization of force field parameters. *Digit Discov*. 2023;2(3):828-47.
<https://doi.org/10.1039/D2DD00138A>.

Wang Z, Liu X, Chen H, Yang T, He Y. Exploring multi-fidelity data in materials science: Challenges, applications, and optimized learning strategies. *Appl Sci*. 2023;13(24):13176.
<https://doi.org/10.3390/app132413176>.

Samadian D, Muhit IB, Dawood N. Application of data-driven surrogate models in structural engineering: A literature review. *Arch Comput Methods Eng*. 2025;32(2):735-84.
<https://doi.org/10.1007/s11831-024-10152-0>.

Duan C, Chu DB, Nandy A, Kulik HJ. Detection of multi-reference character imbalances enables a transfer learning approach for virtual high throughput screening with coupled cluster accuracy at DFT cost. *Chem Sci*. 2022;13(17):4962-71.
<https://doi.org/10.1039/D2SC00393G>.

Howard AA, Perego M, Karniadakis GE, Stinis P. Multifidelity deep operator networks for data-driven and physics-informed problems. *J Comput Phys*. 2023;493:112462.
<https://doi.org/10.1016/j.jcp.2023.112462>.

Uthayakumar H, K RK, Jain R, Kumar R, Patra TK. QRChEM: A deep learning framework for materials property prediction and design using QR codes. *ACS Eng Au*. 2024;4(1):91-8.
<https://doi.org/10.1021/acsengineeringau.3c00055>.

Sabanza-Gil V, Barbano R, Pacheco Gutiérrez D, Luterbacher JS, Hernández-Lobato JM, Schwaller P, et al. Best practices for multi-fidelity Bayesian optimization in materials and molecular research. *Nat Comput Sci*. 2025;5(7):572-81.
<https://doi.org/10.1038/s43588-025-00822-9>.

Cho EH, Lyu Q, Lin LC. Computational discovery of nanoporous materials for energy- and environment-related applications. *Mol Simul*. 2019;45(14-15):1122-47.
<https://doi.org/10.1080/08927022.2019.1626990>.

Menon N, Mondal S, Basak A. Multi-fidelity surrogate-based process mapping with uncertainty quantification in laser directed energy deposition. *Materials*. 2022;15(8):2902.
<https://doi.org/10.3390/ma15082902>.

Nguyen BD, Potapenko P, Demirci A, Govind K, Bompas S, Sandfeld S. Efficient surrogate models for materials science simulations: Machine learning-based prediction of microstructure properties. *Mach Learn Appl*. 2024;16:100544.
<https://doi.org/10.1016/j.mlwa.2024.100544>.

Minotakis M, Rossignol H, Cobelli M, Sanvito S. Machine-learning surrogate model for accelerating the search of stable ternary alloys. *Phys Rev Mater.* 2023;7(9):093802.

<https://doi.org/10.1103/PhysRevMaterials.7.093802>.

Wang T, Shao M, Guo R, Tao F, Zhang G, Snoussi H, et al. Surrogate model via artificial intelligence method for accelerating screening materials and performance prediction. *Adv Funct Mater.* 2021;31(8):2006245.

<https://doi.org/10.1002/adfm.202006245>.

Shah AA, Leung PK, Xing WW. Rapid high-fidelity quantum simulations using multi-step nonlinear autoregression and graph embeddings. *npj Comput Mater.* 2025;11(1):57.

<https://doi.org/10.1038/s41524-024-01479-0>.

Krawczuk S. Data-efficient surrogate models for high-throughput density functional theory [master's thesis]. Santa Cruz (CA): University of California, Santa Cruz.

Mayr F, Gagliardi A. Global property prediction: A benchmark study on open-source, perovskite-like datasets. *ACS Omega.* 2021;6(19):12722-32.

<https://doi.org/10.1021/acsomega.1c00991>.

Palizhati A, Aykol M, Suram SK, Hummelshøj JS, Montoya JH. Multi-fidelity sequential learning for accelerated materials discovery. *ChemRxiv [Preprint]*. 2021.

<https://doi.org/10.26434/chemrxiv.14312612.v1>.