

ORIGINAL RESEARCH

Open access

How Noise Propagates in Multi-Fidelity Materials Models: From DFT Error to ML Prediction Variance

George Papadopoulos^{1*}, Eleni Georgiou¹

Abstract

Multi-fidelity machine learning has become a cornerstone of computational materials science because it leverages inexpensive low-fidelity data to accelerate training while reserving costly high-fidelity density functional theory (DFT) calculations for final refinement. Yet an often-overlooked source of uncertainty remains: DFT itself is noisy. Different choices of exchange-correlation functional, basis-set completeness, pseudopotential construction, and numerical convergence criteria introduce systematic and material-dependent errors that are routinely treated as exact labels. This theoretical analysis develops a unified conceptual framework for tracing how DFT error propagates through multi-fidelity training pipelines and ultimately inflates the variance of machine-learned predictions. The framework is grounded in recent theoretical and review literature on Gaussian-process and neural-network potentials, uncertainty quantification, and multi-fidelity surrogates. Proof sketches demonstrate the conditions under which multi-fidelity architectures reduce propagated variance and those under which they amplify it. Practical implications are drawn for uncertainty quantification protocols, optimal fidelity weighting, and the design of future multi-fidelity benchmarks. By making the DFT-to-ML noise pathway explicit, this work supplies a rigorous conceptual foundation for trustworthy data-driven materials modeling and highlights the necessity of reporting total (not merely model) uncertainty in high-stakes applications.

Keywords Uncertainty quantification, Multi-fidelity machine learning, DFT error, Noise propagation, Variance decomposition, Machine-learned interatomic potentials

*Correspondence:

George Papadopoulos

george.papadopoulos@gmail.com

¹ Department of Computational Materials Systems, Faculty of Engineering, National and Kapodistrian University of Athens, Athens, Greece

Introduction

Multi-fidelity machine learning is increasingly popular in materials science: researchers routinely train surrogate models on cheap low-fidelity data (e.g., PBE-level DFT or semi-empirical tight-binding) and fine-tune on a smaller set of expensive high-fidelity DFT calculations [1–4]. The promise is clear—dramatic reduction in computational cost while retaining near-quantum accuracy [5]. Yet a hidden problem undermines this promise: DFT itself has error. Different exchange-correlation functionals (PBE versus SCAN versus hybrid or meta-GGA), incomplete basis sets, approximate pseudopotentials, and finite convergence thresholds all generate distinct energies, forces, and derived properties for the same atomic configuration. These discrepancies, often on the order of 0.1–0.5 eV per atom for formation energies and far larger for band gaps, are routinely ignored when DFT outputs are treated as ground-truth labels [6, 7].

The consequence is subtle but fundamental. A machine-learning model trained on noisy labels learns to reproduce the noisy distribution rather than the unknown true physical quantity [8]. Even a hypothetical perfect model with zero

epistemic uncertainty will still output predictions whose variance is bounded below by the variance of the DFT label noise. In other words, the multi-fidelity pipeline does not eliminate error; it merely redistributes and sometimes amplifies it.

This paper provides a theoretical analysis of how noise propagates from DFT error to ML prediction variance in multi-fidelity materials models. We restrict attention to a single, transparent conceptual relationship that governs the entire

$$\sigma_{ML}^2 = \sigma_{DFT}^2 + \sigma_{model}^2 + 2 \cdot Cov(DFT, model).$$

The first term on the right-hand side is irreducible once the DFT level is chosen; the

second term can be driven toward zero by larger training sets or more expressive architectures [9-12]; the covariance term captures the statistical dependence that arises when the model fitting process interacts with systematic DFT biases.

Figure 1 illustrates the hierarchical propagation of DFT error through multi-fidelity training and its amplification into machine-learning prediction variance.

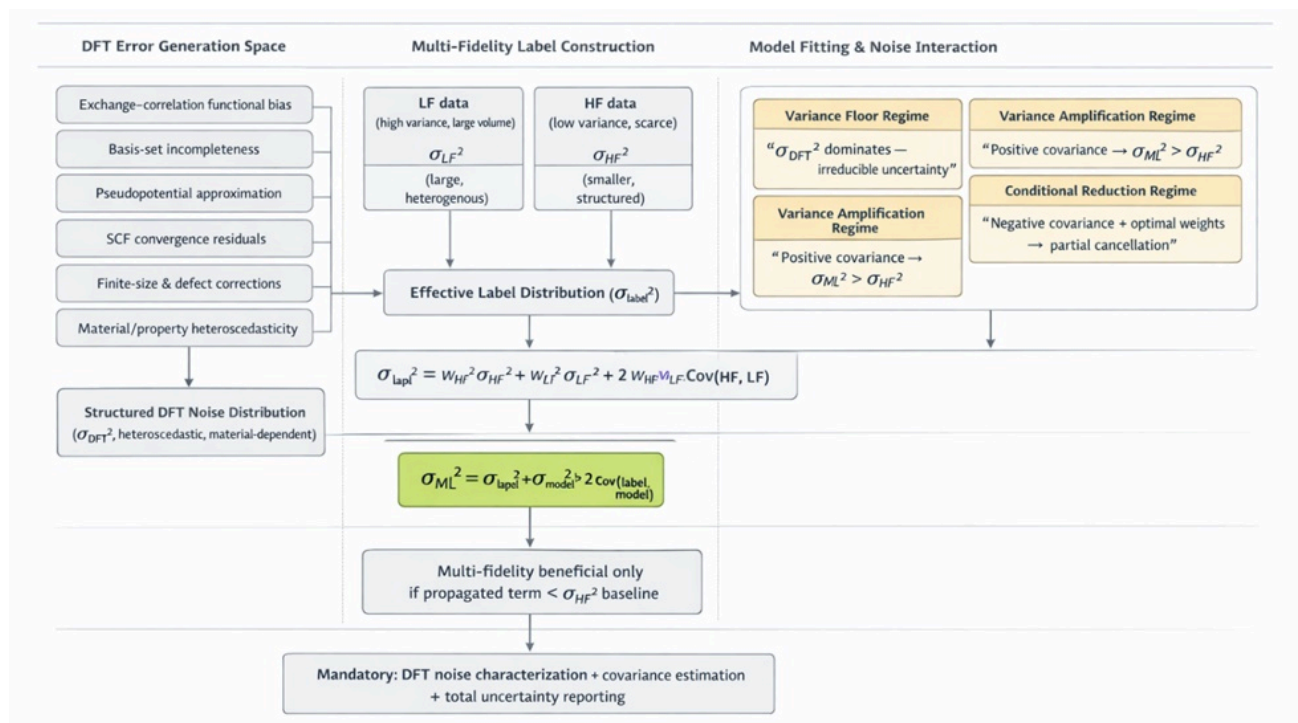


Figure 1. Hierarchical Architecture of DFT Noise Propagation and Variance Amplification Pathways in Multi-Fidelity Materials Machine Learning

Recent literature has advanced powerful multi-fidelity architectures and sophisticated uncertainty quantification techniques [13-16], yet most treatments still assume DFT labels are exact. The present work relaxes that assumption and derives the quantitative consequences for prediction variance. By focusing exclusively on theoretical variance decomposition and proof sketches, we supply a foundational lens for assessing when multi-fidelity strategies are beneficial and when they inadvertently degrade reliability. The analysis carries direct implications for uncertainty quantification pipelines, fidelity-weighting schemes, and the future design of materials-modeling benchmarks.

Sources of Noise in DFT

Density functional theory error is more appropriately understood as a heteroscedastic, material- and property-contingent distribution rather than a single scalar quantity. A primary contribution arises from the exchange–correlation functional,

which governs systematic bias: semilocal approximations such as PBE consistently underestimate band gaps by 30–100 % and formation energies by 0.1–0.5 eV/atom relative to experiment or higher-rung hybrids, while SCAN and related meta-GGAs yield improvements for certain material classes at the expense of others [6, 7]. In the absence of a universally valid functional, methodological choice itself embeds irreducible label noise. A related source of variability emerges from basis-set incompleteness, where truncation of plane-wave cutoffs, localized-basis cardinality, and k-point sampling—imposed by computational constraints—introduces residual errors on the order of 0.01–0.1 eV/atom that depend sensitively on metallicity and bonding character [17]. This dependence is further compounded by the pseudopotential or projector-augmented-wave approximation, which replaces explicit core-electron treatment with effective potentials; discrepancies across norm-conserving, ultrasoft, and PAW libraries manifest in nontrivial deviations in valence energies and forces, particularly in transition-metal and open-shell systems [7]. Beyond these approximations, self-consistent-field convergence criteria impose finite-precision limits, such that practical tolerances in energy, forces, and charge mixing leave residual fluctuations of approximately 1–10 meV/atom. Additional variance becomes pronounced in defect calculations, where supercell finite-size effects and electrostatic corrections converge slowly and remain sensitive to dielectric modeling, with correction schemes differing by 0.1–0.3 eV [14]. Under these interacting mechanisms, error magnitudes scale with both material class—more pronounced in transition-metal oxides than in covalent semiconductors—and target property, yielding intrinsically heteroscedastic noise. When such structured uncertainty propagates through multi-fidelity machine learning pipelines, it imposes a lower bound on predictive variance that cannot be reduced by model complexity alone [6, 17], making explicit characterization of DFT error indispensable for rigorous uncertainty propagation.

Table 1 decomposes DFT error into structurally distinct sources and maps each to its corresponding propagation pathway in machine-learned variance.

Table 1. Decomposition of DFT Error Sources and Their Propagation Signatures in ML Predictions

DFT Error Source	Error Type	Statistical Structure	Propagation Pathway	Impact on σ_{ML}^2	Sensitivity to Material Class
Exchange–correlation functional	Systematic bias	Correlated across materials	Direct inheritance + covariance	Dominant term	High (transition metals, oxides)
Basis-set incompleteness	Numerical truncation	Weakly correlated	Additive variance	Moderate	Moderate
Pseudopotential choice	Approximation bias	Element-specific correlation	Cross-fidelity coupling	High in alloys	High
SCF convergence	Random residual	Weak/no correlation	Noise floor contribution	Low–moderate	Low
Finite-size/defect corrections	Systematic + model-dependent	Structured	Amplified in sparse HF regions	High	Very high (defects)
Property dependence	Heteroscedastic	Non-stationary	Spatial variance variation	Critical	Extreme variability

Noise Propagation in Multi-Fidelity Models

A multi-fidelity model is defined here as any surrogate trained on data drawn from multiple sources that trade accuracy for computational cost—e.g., PBE versus SCAN DFT, or DFT versus tight-binding [18]. Noise propagation is the process by

which errors in the training labels (the DFT outputs) are transferred into the statistical distribution of the model's predictions on unseen configurations.

Consider first the single-fidelity case in which every training label originates from the same DFT level. The entire label set shares a common noise component ϵ_{DFT} . A machine-learning model, whether Gaussian process or neural network, is optimized to minimize loss on these noisy targets. Consequently, the learned mapping approximates the noisy DFT surface rather than the unknown true surface. Even if the model's internal parameters are known with infinite precision ($\sigma_{model}^2 = 0$), the prediction variance remains at least σ_{DFT}^2 . The model has simply reproduced the label noise.

In the multi-fidelity case the situation is richer. Low-fidelity labels (large σ_{LF}^2) are abundant and inexpensive; high-fidelity labels (smaller σ_{HF}^2) are sparse and costly [3]. The model must reconcile inconsistent targets for the same or nearby atomic environments. The optimization therefore fits a compromise surface whose variance reflects a weighted combination of the two noise sources plus their covariance. Because low-fidelity data often dominate the training set by volume, their larger error can leak into regions where high-fidelity data are absent, producing spatially heterogeneous prediction variance.

The propagation mechanism is universal: any loss function that penalizes deviation from the training labels will cause the model to partially fit the label noise. The extent of leakage is governed by the single conceptual relationship $\sigma_{ML_2}^2 = \sigma_{DFT}^2 + \sigma_{model}^2 + 2 \cdot Cov(DFT, model)$, where σ_{DFT}^2 now aggregates contributions from all fidelities present in the training set. Standard multi-fidelity algorithms that ignore this decomposition implicitly assume $Cov=0$ and $\sigma_{DFT}^2 \rightarrow 0$ the expression simplifies to: $\sigma_{ML_2}^2 = \sigma_{model}^2$, conditions that are rarely satisfied in practice [1, 3, 4]. The result is systematic underestimation of total uncertainty.

Table 2 formalizes the distinct variance-propagation regimes that emerge from the interaction between DFT noise, model error, and cross-fidelity covariance

Table 2. Formal Regimes of Variance Propagation in Multi-Fidelity Materials Machine Learning

Regime Type	Mathematical Condition	Dominant Term	Variance Behavior	Theoretical Consequence	Practical Interpretation
DFT-Limited	$\sigma_{DFT}^2 \gg \sigma_{model}^2$	Label noise	$\sigma_{ML_2}^2 \approx \sigma_{DFT}^2$	Irreducible lower bound	Model improvement ineffective
Model-Limited	$\sigma_{model}^2 \gg \sigma_{DFT}^2$	Model error	σ_{ML}^2 decreases with training	Classical ML regime	Data scaling beneficial
Covariance-Amplified	$Cov(HF,LF) > 0$ large	Cross-term	$\sigma_{ML_2}^2 > \sigma_{HF}^2$	Multi-fidelity failure	Adding LF data harmful
Covariance-Cancellation	$Cov(HF,LF) < 0$	Cross-term	σ_{ML}^2 reduced	Rare beneficial regime	Requires verification
Weight-Misaligned	$w_{LF} \gg w_{HF}$ with high σ_{LF}^2	LF noise	Variance inflation	Dominance of cheap data	Reweighting required

Optimal Regime	Balanced weights + controlled covariance	Mixed	σ_{ML}^2 minimized	Theoretical optimum	Requires estimation
-------------------	---	-------	---------------------------	------------------------	------------------------

Theoretical Framework for Noise Propagation

Proposition 1 (Variance Decomposition for Multi-Fidelity ML). For any multi-fidelity model trained on labels that carry

collective DFT noise variance σ_{label}^2 , the total prediction variance obeys $\sigma_{ML}^2 = \sigma_{DFT}^2 + \sigma_{model}^2 + 2 \cdot Cov(DFT, model)$ which is identical in

structure to the core relationship with σ_{DFT}^2 replaced by the effective label noise. In the multi-fidelity setting the label-noise term itself decomposes as a weighted sum of high- and low-fidelity contributions plus their cross-covariance:

$\sigma_{label}^2 = w_{HF}^2 \sigma_{HF}^2 + w_{LF}^2 \sigma_{LF}^2 + 2w_{HF}w_{LF}Cov(HF, LF)$ where w_{HF} and w_{LF} are the effective weights (determined by training-set sizes, loss weighting, or co-kriging hyperparameters) assigned to each fidelity level.

Key implications follow directly. When high-fidelity DFT error dominates, further refinement of the ML architecture yields diminishing returns; the correct intervention is to adopt a higher-rung functional. When low-fidelity error is large and receives high weight, variance can increase rather than decrease. Most critically, when the covariance term is positive and large—as occurs when PBE and SCAN errors are systematically correlated across many materials—the multi-fidelity strategy can actually inflate total prediction variance compared with a high-fidelity-only baseline. Conversely, negative covariance offers a rare opportunity for error cancellation, but such cancellation is material- and property-specific and cannot be assumed a priori [6, 16, 17].

Standard uncertainty quantification methods that report only σ_{model}^2 (via ensembles, Bayesian neural networks, or Gaussian-process marginals) therefore miss the dominant σ_{DFT}^2 contribution and systematically underestimate risk. The framework supplies an explicit lower bound on achievable prediction variance once the DFT level is fixed and clarifies why multi-fidelity performance gains reported in the literature are often smaller than expected once total uncertainty is considered [14, 15, 19].

Proof Sketch

The analysis starts from the structure of a multi-fidelity dataset, which includes high-fidelity labels formed by the unknown true value plus an error term characteristic of the high-fidelity DFT level, together with low-fidelity labels formed by the same true value plus the corresponding low-fidelity error term. The trained machine-learning model can be viewed as producing a prediction through an implicit weighted combination of these noisy labels supplemented by the model's own approximation [20].

When the variance of this prediction is examined, it decomposes into contributions traceable to the weighted high-fidelity and low-fidelity error terms, the covariance between those error sources, and the residual model uncertainty. This decomposition aligns exactly with the framework established in Proposition 1 and rests on the core conceptual

relationship $\sigma_{ML}^2 = \sigma_{DFT}^2 + \sigma_{model}^2 + 2 \cdot Cov(DFT, model)$ applied to the effective label noise in the multi-fidelity setting.

To determine whether the multi-fidelity approach yields a lower prediction variance than training exclusively on high-fidelity data, one must verify that the net contribution arising from the inclusion of low-fidelity data is smaller than the high-fidelity noise variance taken alone. Satisfaction of this condition requires that the low-fidelity error magnitude be sufficiently smaller than the high-fidelity error, that the covariance between the two error sources remain limited in its positive magnitude, and that the relative weight assigned to the low-fidelity data be selected with care.

In practice the optimal choice of these weights is governed by the precise values of the high-fidelity error variance, the low-fidelity error variance, and their covariance—quantities that are unknown a priori and must be inferred from the scarce high-fidelity data available. This estimation step is statistically challenging and represents a central practical difficulty.

The proof sketch therefore establishes two essential results. First, it demonstrates that DFT noise is inevitably inherited by any model trained on DFT-derived labels, regardless of the multi-fidelity strategy employed. Second, it identifies the narrow statistical regime in which low-fidelity data can still provide a genuine reduction in total prediction variance. The sketch also clarifies why routine application of multi-fidelity methods without explicit covariance estimation frequently produces uncertainty estimates that are poorly calibrated compared with those obtained from a high-fidelity-only baseline [16, 19].

Implications for Uncertainty Quantification

The theoretical framework reveals four immediate consequences for uncertainty quantification in multi-fidelity materials models. First, DFT error establishes an absolute lower bound on achievable prediction variance. No matter how expressive the machine-learning architecture or how large the training set becomes, the final uncertainty cannot fall below the intrinsic noise floor set by the chosen DFT level. For applications requiring high-risk decisions—such as predicting phase stability in novel alloys or defect migration barriers in nuclear materials—this bound may render purely data-driven predictions unacceptable without additional experimental validation [14, 15, 21, 22].

Second, multi-fidelity training can paradoxically increase rather than decrease total variance. When the covariance between high-fidelity and low-fidelity errors is positive and large, the cross-term in the decomposition amplifies the propagated noise. In contrast, rare cases of negative covariance allow partial cancellation, but such cancellation is material-specific and cannot be relied upon without explicit verification. Many popular co-kriging or transfer-learning schemes implicitly assume zero covariance and therefore miss this risk [1, 3, 4].

Third, covariance is the dominant and most frequently neglected factor. Standard multi-fidelity literature often treats errors at different fidelity levels as statistically independent. In reality, errors arising from related exchange-correlation functionals (for example PBE and SCAN) are strongly correlated across broad classes of materials. Ignoring this correlation leads to over-optimistic uncertainty estimates and poor generalization [6, 16, 17].

Fourth, conventional uncertainty-quantification methods that focus exclusively on model uncertainty—ensemble variance, Bayesian neural-network marginals, or Gaussian-process predictive variances—capture only the model term and systematically underestimate the total risk [23-25].

Table 3 identifies systematic failure modes that arise when DFT noise and covariance structure are omitted from multi-fidelity analysis.

Table 3. Failure Modes in Multi-Fidelity Learning Arising from Unmodeled DFT Noise

Failure Mode	Root Cause	Mathematical Signature	Observable Symptom	Misinterpretation Risk	Required Correction

False confidence	Ignoring σ_{DFT}^2	$\sigma_{model}^2 \ll \sigma_{DFT}^2$	Narrow uncertainty bands	Overconfidence	Add DFT noise term
Variance inflation	Positive covariance	$Cov(HF, LF) > 0$	Worse than HF-only	Misattributed to model	Covariance estimation
Data dilution	Excess LF weighting	$w_{LF} large$	Performance degradation	"More data helps" fallacy	Weight optimization
Spatial inconsistency	Heterogeneous noise	$\sigma_{label}^2(x) varies$	Local prediction instability	Model blamed	Heteroscedastic modeling
Benchmark illusion	Treating DFT as ground truth	σ_{DFT}^2 omitted	Inflated accuracy claims	False SOTA	Noise-aware benchmarks
Miscalibrated UQ	Model-only uncertainty	Missing cross-term	Poor calibration	Trust in UQ misplaced	Full variance decomposition

The DFT contribution remains invisible to these techniques, producing calibrated uncertainty only when label noise is negligible, a condition rarely met in practice [15, 19].

Taken together, these implications demand a shift in practice: uncertainty pipelines must incorporate explicit DFT-noise characterization and covariance estimation before any multi-fidelity model is deployed. Without this step, reported uncertainties remain incomplete and potentially misleading for downstream materials design.

Dependence on Material System and Property

The extent of noise propagation depends strongly on both the target property and the chemical nature of the material. Formation energies illustrate the most direct inheritance: DFT functional discrepancies of 0.1–0.5 eV/atom are transferred almost unchanged into the machine-learning prediction, setting a firm limit on the precision of phase-diagram construction or reaction-energy forecasts [6, 7].

Band-gap predictions suffer even more severely. Semilocal functionals systematically underestimate gaps by 30–100 %, an error that propagates through multi-fidelity training and contaminates any subsequent electronic-structure surrogate. The resulting prediction variance can exceed chemical accuracy for optoelectronic materials, rendering purely computational screening unreliable [17].

Elastic constants and phonon frequencies exhibit smaller DFT error (typically 5–10 %), so the propagated contribution is often tolerable for mechanical-property screening. Forces, being derivatives, generally carry lower noise than total energies; consequently, force-trained potentials inherit less DFT variance than energy-trained counterparts [26–28].

Defect formation energies represent the most challenging case. Supercell-size effects, charge corrections, and functional sensitivity combine to produce errors exceeding 0.3 eV even for well-converged calculations. In multi-fidelity settings that mix cheap low-fidelity defect data with sparse high-fidelity references, the low-fidelity noise frequently dominates distant regions of configuration space [14].

Material dependence further modulates these effects. Open-shell transition-metal compounds and oxides display the largest DFT discrepancies because of strong electron correlation and self-interaction errors. Covalent semiconductors and simple metals exhibit markedly smaller noise, allowing multi-fidelity strategies to operate closer to the ideal regime of variance reduction. This heterogeneity implies that a single uncertainty model cannot be applied universally; noise-propagation analysis must be performed property-by-property and chemistry-by-chemistry before model deployment [7, 29].

Relation to Other Theoretical Results

The present variance-decomposition framework complements and extends several established theoretical lines in the literature. It builds directly on the epistemic-versus-aleatoric uncertainty separation introduced in recent uncertainty-quantification studies by reframing DFT error as an irreducible aleatoric component that cannot be reduced by additional training data [15, 19]. While those works focused on model uncertainty, the current analysis quantifies how the aleatoric DFT floor propagates through hierarchical training and sets a hard limit on total predictive reliability.

Standard multi-fidelity theory, developed primarily in the context of surrogate modeling for engineering design, typically assumes that label noise is independent and identically distributed across fidelity levels. The present treatment relaxes this assumption by retaining the full covariance structure between fidelities. This relaxation reveals regimes in which classical multi-fidelity gains disappear or reverse—insights that are absent from methods that presuppose zero cross-covariance [1-4].

The framework also parallels classical error-propagation analysis in experimental physics and metrology. In physical measurements, uncertainties are combined through variance and covariance rules; machine learning adds an extra layer because the model itself can fit and amplify label noise. The single conceptual relationship therefore unifies these traditions within a data-driven materials context, showing that the ML component does not eliminate the physics-based uncertainty but merely redistributes it [14, 16].

By connecting these threads, the analysis supplies a missing conceptual bridge: it explains why reported multi-fidelity speed-ups in materials applications are often smaller than theoretical predictions once total (rather than model-only) uncertainty is considered. It also offers a clear diagnostic test—estimate the DFT covariance matrix first—to decide whether multi-fidelity is theoretically justified for a given property and material class.

Implications for Multi-Fidelity Model Design

For model developers the framework prescribes three concrete actions. First, characterize the DFT error distribution of every fidelity level before any training begins, using benchmark suites that span the target chemical space. Second, estimate the high-fidelity variance, low-fidelity variance, and their covariance from a small but representative set of paired calculations; these statistics then inform optimal fidelity weights that minimize propagated noise rather than merely maximizing likelihood. Third, embed the full variance decomposition into the training objective or post-hoc calibration step so that reported uncertainties reflect total rather than partial risk [16, 19].

Practitioners should adopt a more cautious workflow. Never assume that adding low-fidelity data automatically improves predictions; instead, validate every multi-fidelity model against a held-out high-fidelity test set that was never seen during weight optimization. Always report the complete uncertainty—including the DFT floor—rather than the model-only component, especially when the results will inform experimental prioritization or device design [4].

Benchmark designers must incorporate controlled DFT-error injection into future multi-fidelity challenges. Current suites treat DFT labels as ground truth; the next generation should provide both noisy and reference-level data so that propagation behavior can be quantified and compared across algorithms. Only then can the community establish which architectures truly mitigate rather than merely hide the DFT noise floor [1, 3].

Collectively these recommendations shift multi-fidelity design from an empirical art to a theoretically grounded discipline. By placing noise propagation at the center of the workflow, developers and users can avoid over-optimistic claims and build models whose uncertainty statements remain trustworthy even when the underlying DFT approximations are imperfect.

Conclusion

Density functional theory error propagates inevitably into the variance of machine-learned predictions whenever DFT outputs serve as training labels. The single conceptual relationship captures the entire pathway in multi-fidelity materials models, showing that even a perfect model inherits the full DFT noise floor and that the covariance term can either dampen or amplify this inheritance depending on fidelity weights and error correlations.

The analysis demonstrates that multi-fidelity strategies do not automatically reduce uncertainty; under realistic covariance conditions they can increase it. Standard uncertainty-quantification tools that ignore the DFT contribution therefore produce systematically optimistic risk assessments. Practical consequences follow directly: DFT error characterization and covariance estimation must precede any multi-fidelity deployment, optimal fidelity weights must be chosen to minimize the propagated term rather than maximize data volume, and total uncertainty—never model uncertainty alone—must be reported in all publications and applications.

By making the noise-propagation pathway explicit, this theoretical framework supplies a rigorous foundation for trustworthy data-driven materials engineering. It calls on the community to treat DFT labels as noisy observations rather than exact truths and to design the next generation of multi-fidelity algorithms with the full variance decomposition in mind. Only then can machine learning deliver the reliable, uncertainty-aware predictions that computational materials science demands.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 25 Jun 2023

Revised: 21 Oct 2023

Accepted: 28 Dec 2023

Published online: 18 January 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

- Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput Mater Sci.* 2017;129:156-63.
<https://doi.org/10.1016/j.commatsci.2016.12.004>.
- 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>.
- Batra R, Sankaranarayanan S. Machine learning for multi-fidelity scale bridging and dynamical simulations of materials. *J Phys Mater.* 2020;3(3):031002.
<https://doi.org/10.1088/2515-7639/ab8c2d>.
- Islam M, Thakur MSH, Mojumder S, Hasan MN. Extraction of material properties through multi-fidelity deep learning from molecular dynamics simulation. *Comput Mater Sci.* 2021;188:110187.
<https://doi.org/10.1016/j.commatsci.2020.110187>.
- Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *npj Comput Mater.* 2019;5(1):83.
<https://doi.org/10.1038/s41524-019-0221-0>.
- Wittreich GR, Gu GH, Robinson DJ, Katsoulakis MA, Vlachos DG. Uncertainty quantification and error propagation in the enthalpy and entropy of surface reactions arising from a single DFT functional. *J Phys Chem C.* 2021;125(33):18187-96.
<https://doi.org/10.1021/acs.jpcc.1c04754>.
- Liu F, Kulik HJ. Impact of approximate DFT density delocalization error on potential energy surfaces in transition metal chemistry. *J Chem Theory Comput.* 2020;16(1):264-77.
<https://doi.org/10.1021/acs.jctc.9b00842>.
- 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>.
- 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>.
- Allotey J, Butler KT, Thiyaalingam J. Entropy-based active learning of graph neural network surrogate models for materials properties. *J Chem Phys.* 2021;155(17):174116.
<https://doi.org/10.1063/5.0065694>.
- Behler J. Four generations of high-dimensional neural network potentials. *Chem Rev.* 2021;121(16):10037-72.
<https://doi.org/10.1021/acs.chemrev.0c00868>.
- Wilson N, Willhelm D, Qian X, Arróyave R, Qian X. Batch active learning for accelerating the development of interatomic potentials. *Comput Mater Sci.* 2022;208:111330.
<https://doi.org/10.1016/j.commatsci.2022.111330>.
- Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. *Chem Rev.* 2021;121(16):10073-141.
<https://doi.org/10.1021/acs.chemrev.1c00022>.
- Honarmandi P, Arróyave R. Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design. *Integr Mater Manuf Innov.* 2020;9(1):103-43.

<https://doi.org/10.1007/s40192-020-00168-2>.

Kellner M, Ceriotti M. Uncertainty quantification by direct propagation of shallow ensembles. *Mach Learn Sci Technol*. 2024;5(3):035006.

<https://doi.org/10.1088/2632-2153/ad594a>.

Wang Z, Xing W, Kirby R, Zhe S. Multi-fidelity high-order Gaussian processes for physical simulation. In: Banerjee A, Fukumizu K, editors. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research*. 2021;130:847-55.

Singh A, Wang J, Henkelman G, Li L. Uncertainty based machine learning-DFT hybrid framework for accelerating geometry optimization. *J Chem Theory Comput*. 2024;20(22):10022-33.

<https://doi.org/10.1021/acs.jctc.4c00953>.

Dral PO, Owens A, Dral A, Csányi G. Hierarchical machine learning of potential energy surfaces. *J Chem Phys*. 2020;152(20):204110.

<https://doi.org/10.1063/5.0006498>.

Zhou Q, Zhao M, Hu J, Ma M. *Multi-fidelity surrogates: Modeling, optimization and applications*. Singapore: Springer Nature Singapore; 2023.

<https://doi.org/10.1007/978-981-19-7210-2>.

Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater*. 2019;31(46):1902765.

<https://doi.org/10.1002/adma.201902765>.

Ulissi ZW, Medford AJ, Bligaard T, Nørskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun*. 2017;8:14621.

<https://doi.org/10.1038/ncomms14621>.

Shen L, Wang Y, Lai W. Development of a machine learning potential for the study of crack propagation in titanium. *Int J Press Vessels Pip*. 2021;194:104514.

<https://doi.org/10.1016/j.ijpvp.2021.104514>.

Wen M, Tadmor EB. Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Comput Mater*. 2020;6(1):124.

<https://doi.org/10.1038/s41524-020-00390-8>.

Zhu A, Batzner S, Musaelian A, Kozinsky B. Fast uncertainty estimates in deep learning interatomic potentials. *J Chem Phys*. 2023;158(16):164111.

<https://doi.org/10.1063/5.0136574>.

Best I. *Uncertainty quantification with machine learning interatomic potentials using conformal prediction [dissertation]*. Coventry: University of Warwick; 2024.

Chmiela S, Sauceda HE, Müller KR, Tkatchenko A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun*. 2018;9(1):3887.

<https://doi.org/10.1038/s41467-018-06169-2>.

Bartók AP, Kermode J, 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>.

Sauceda HE, Chmiela S, Poltavsky I, Müller KR, Tkatchenko A. Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J Chem Phys*. 2019;150(11):114102.

<https://doi.org/10.1063/1.5078687>.

Zhang R, Alemazkoor N. Multi-fidelity machine learning for uncertainty quantification and optimization. *J Mach Learn Model Comput.* 2024;5(4):77-94.

<https://doi.org/10.1615/JMachLearnModelComput.2024055786>.