

ORIGINAL RESEARCH

Open access

Separating What We Know from What We Guess: A Modular Framework for Epistemic vs. Aleatoric Uncertainty in ML Potentials

Ahmed El-Kholy^{1*}, Nour Abdelrahman¹, Karim Hassan²

Abstract

Uncertainty quantification in machine learning (ML) interatomic potentials remains fundamentally limited by the conflation of epistemic uncertainty, arising from incomplete sampling of configuration space, and aleatoric uncertainty, embedded in reference data generated by density-functional theory. Existing approaches provide internally consistent uncertainty estimates but collapse these distinct sources into a single scalar, obscuring the mechanisms governing model reliability and limiting principled decision-making. This work introduces a modular, architecture-agnostic framework that enforces explicit separation of epistemic and aleatoric contributions at the level of model design rather than post hoc analysis. The framework defines five interoperable components—a shared feature extractor, dedicated epistemic and aleatoric modules, an aggregation mechanism, and a calibration stage—whose interactions preserve disentanglement throughout training, inference, and downstream application. The resulting formulation transforms uncertainty into an operational diagnostic. Epistemic uncertainty identifies regions where additional data acquisition is informative, whereas aleatoric uncertainty defines the intrinsic accuracy ceiling imposed by the reference method. This separation restructures active learning by directing sampling toward reducible error, enables meaningful comparison between models through their uncertainty composition, and grounds performance evaluation relative to an explicit noise floor. The framework further introduces operational criteria that provide falsifiable tests of successful separation, ensuring that reported uncertainties remain interpretable and consistent across architectures. By decoupling learnable structure from irreducible variability, the proposed approach establishes a principled foundation for uncertainty-aware ML potentials, supporting more efficient data allocation, more reliable atomistic simulations, and more rigorous standards for model development in computational materials science.

Keywords Epistemic uncertainty, Active learning, Aleatoric uncertainty, Machine learning potentials, Uncertainty decomposition, Modular framework

*Correspondence:

Ahmed El-Kholy
ahmed.elkholy@outlook.com

¹ Department of Smart Materials Systems Engineering, Faculty of Engineering, Alexandria University, Alexandria, Egypt

² Department of Computational Materials Analytics, Faculty of Technology, Ain Shams University, Cairo, Egypt

Introduction

When an ML potential predicts the energy of a crystal configuration as -5.32 eV/atom with an uncertainty of ± 0.05 eV, the practitioner immediately confronts a practical question: does this uncertainty stem from the model's limited exposure to similar atomic environments, or from irreducible noise already present in the training data generated by density-functional theory (DFT)? Current practice offers no answer. Most uncertainty-quantification pipelines for ML potentials—whether deep ensembles, dropout-based Bayesian approximations, or Gaussian-process surrogates—collapse both sources into

a single reported variance [1-5]. The resulting scalar is internally consistent yet externally opaque, rendering it impossible to decide whether the next computational investment should target more training structures, a higher-fidelity reference method, or simply accept an intrinsic floor.

This ambiguity is particularly costly in materials science, where ML potentials are deployed for discovery campaigns spanning millions of configurations, phase-diagram explorations, or defect energetics [6-12]. A misallocated data-acquisition budget can delay promising materials by months, while an unrecognized aleatoric ceiling can lead researchers to chase phantom improvements beyond the accuracy of the underlying electronic-structure method. The problem is not new. Kendall and Gal demonstrated in computer-vision settings that epistemic and aleatoric uncertainties require distinct handling [13], and Bengio *et al.* showed how dropout can approximate posterior uncertainty [14]. Yet these insights have not been systematically translated to the structured, permutation-invariant, many-body domain of interatomic potentials.

The present work closes that gap by proposing a modular conceptual framework that enforces separation at the architectural level. The framework is deliberately abstract: it prescribes no specific neural-network architecture, no particular ensemble size, and no fixed loss function. Instead, it defines five interoperable components whose interactions guarantee that epistemic and aleatoric contributions remain distinguishable at every stage—training, inference, and downstream decision-making. The only mathematical commitment is the additive decomposition of total predictive variance shown above, a relation that follows directly from the law of total variance once the two sources are conditionally independent given the input representation.

In short, this framework reframes uncertainty quantification for ML potentials from an after-the-fact statistical exercise into a first-principles engineering discipline. It supplies the conceptual scaffolding needed to move beyond “how uncertain are we?” toward the sharper question “what exactly are we uncertain about—and what should we do about it?”

Defining Epistemic and Aleatoric Uncertainty

Epistemic uncertainty is model uncertainty. It quantifies the degree to which the learned potential deviates from the true underlying interatomic energy surface because the training data are finite and non-exhaustive. In the language of Bayesian statistics, it corresponds to posterior variance over model parameters once the data are observed [13, 14]. Because epistemic uncertainty arises solely from lack of information, it is, in principle, reducible: acquiring additional high-quality reference calculations in under-sampled regions of configuration space will shrink it [15, 16]. For ML potentials, epistemic uncertainty is highest for novel compositions, extreme pressures, or rare defect geometries that lie far from the training distribution [17].

Aleatoric uncertainty, by contrast, is data uncertainty. It captures irreducible stochasticity or systematic bias already embedded in the reference calculations themselves—numerical noise from k-point sampling, basis-set incompleteness, pseudopotential approximations, or the intrinsic exchange-correlation error of the chosen DFT functional [18]. Once a particular reference method is fixed, this floor cannot be lowered by collecting more data generated with that same method. Two independently trained ML potentials that share identical training data will therefore share the same aleatoric component, even if their epistemic uncertainties differ. This shared floor is what Hüllermeier and Waegeman term “data noise” in the general machine-learning taxonomy [18].

The key insight for materials modeling is that aleatoric uncertainty is heteroscedastic: it varies with local atomic environment. A perfect crystal may exhibit low aleatoric noise because DFT is highly reproducible there, whereas a disordered alloy or a transition-state configuration may carry substantially larger reference uncertainty due to self-interaction errors or strong electron correlation. Consequently, any framework that assumes homoscedastic noise—as many early Gaussian-process models did [6]—will systematically misattribute environment-dependent reference error to the model.

The two uncertainties are statistically independent once the input representation is fixed. The total predictive distribution for an energy (or force) can therefore be decomposed without loss of information into an epistemic spread around a mean prediction and an aleatoric spread around each realization of that mean. This decomposition is not merely additive in variance; it is conceptually additive in interpretability. An epistemic uncertainty of 0.08 eV/atom signals “collect more data in this region,” whereas an aleatoric uncertainty of 0.08 eV/atom signals “the reference method itself is the limiting factor.”

Importantly, the distinction survives across different MLP families. Whether one uses an ensemble of message-passing networks [2, 5], Monte-Carlo dropout on a Deep Potential model [4, 7, 19], or a Bayesian neural network with variational inference [20], the epistemic module always probes variability across hypotheses consistent with the data, while the aleatoric module isolates the residual variance that no amount of re-training can explain [21]. The modular framework formalizes this separation at the level of architecture rather than post-hoc analysis, ensuring that the two components remain disentangled even when modules are swapped or upgraded independently.

A precise conceptual distinction between epistemic and aleatoric uncertainty is formalized in Table 1, highlighting their fundamentally different origins, behaviors, and operational implications.

Table 1. Conceptual Distinction between Epistemic and Aleatoric Uncertainty in Machine Learning Potentials

Dimension	Epistemic Uncertainty	Aleatoric Uncertainty
Definition	Model uncertainty due to limited or incomplete training data	Data uncertainty arising from noise or bias in reference calculations
Origin	Finite sampling of configuration space	Intrinsic limitations of DFT or experimental data
Reducibility	Reducible with additional high-quality data	Irreducible under fixed reference method
Dependence on Model	Strongly model-dependent	Largely model-independent (shared across models trained on same data)
Behavior with Data Scaling	Decreases as dataset grows	Remains constant
Spatial Characteristics	High in extrapolative regions	Environment-dependent (heteroscedastic)
Diagnostic Interpretation	“Collect more data”	“Improve reference method”
Role in Active Learning	Primary acquisition signal	Should be excluded from acquisition
Cross-Model Comparability	Varies across architectures	Stable across architectures
Practical Implication	Guides data acquisition strategy	Defines accuracy ceiling

Why Separation Matters for ML Potentials

Separation of epistemic and aleatoric uncertainty is not merely conceptual; it directly conditions decision-making across materials modeling pipelines. In active learning, reliance on total uncertainty misallocates computational effort toward configurations where the reference method itself is unreliable, whereas an epistemic-focused acquisition strategy preferentially targets regions that reduce model ignorance while respecting the irreducible noise floor [15-17]. As

epistemic uncertainty approaches a user-defined threshold, marginal gains from additional data diminish, yielding a principled stopping criterion absent from total-uncertainty formulations [3]. A related implication emerges in model comparison: equivalent total uncertainty across competing ML potentials can conceal fundamentally different error structures, where dominance of epistemic uncertainty signals insufficient or unrepresentative training data, while aleatoric dominance indicates limitations of the underlying electronic-structure reference. Without explicit separation, neither diagnosis nor targeted improvement is feasible [2, 22].

This distinction extends into downstream simulations, where uncertainty propagation governs the interpretation of molecular-dynamics trajectories used in phase-diagram construction or defect migration analyses. Epistemic-dominated uncertainty reflects extrapolative regimes that undermine predictive reliability, whereas aleatoric-dominated behavior indicates intrinsic variability in the physical observable that cannot be mitigated through model refinement. Such differentiation becomes critical in thermodynamic integration and free-energy workflows, where uncertainty structure directly shapes confidence in derived quantities [6, 10, 23, 24]. Beyond immediate applications, benchmark design is similarly reframed: claims of predictive accuracy must be evaluated against the aleatoric floor imposed by the reference method, rendering sub-threshold improvements statistically uninformative and redirecting attention toward minimizing epistemic contributions relative to this baseline [18, 25]. Under these conditions, uncertainty separation transforms from passive metric to operational variable, guiding allocation of computational resources between data generation, methodological refinement, and acceptance of intrinsic limits, thereby enabling a level of control inaccessible to single-number uncertainty paradigms.

Existing Methods and Their Limitations

Contemporary uncertainty quantification for ML potentials falls into three broad families, each of which ultimately collapses epistemic and aleatoric contributions.

Deep ensembles remain the workhorse [1-3, 5]. Variance across independently trained networks is interpreted as predictive uncertainty. Because every network is trained on the same noisy DFT data, the ensemble variance inevitably mixes model disagreement (epistemic) with the shared residual noise (aleatoric). No post-processing step can disentangle them once the models have converged to the same noisy target.

Bayesian neural networks and their practical approximations (variational inference, Monte-Carlo dropout) suffer the same fate [4, 14, 20]. The posterior variance over weights captures epistemic uncertainty, yet the likelihood function itself must assume a noise model. When that noise is learned jointly with the weights—as in Kendall and Gal's heteroscedastic regression [13]—the two terms become entangled at prediction time. The resulting output variance again mixes both sources, and the mixing is architecture-dependent, rendering cross-model comparisons unreliable.

Gaussian-process regression and its sparse variants have long been used for materials potentials [6, 11, 26]. The predictive variance naturally decomposes into a model term plus an explicit noise variance; however, the noise hyperparameter is almost universally assumed constant across configuration space. This homoscedastic assumption fails for real materials, where DFT error varies sharply with bonding type, coordination number, or magnetic state. Attempts to learn a position-dependent noise term still leave the final prediction with a single variance that practitioners cannot attribute to model versus data [6, 27].

Even recent proposals that explicitly mention “epistemic” and “aleatoric” in the context of graph neural networks [2, 21] or fast uncertainty estimates [22] ultimately report a combined uncertainty for downstream use. The literature therefore contains valuable building blocks—ensemble variance, learned noise heads, calibration techniques—but lacks an architectural commitment to separation. The result is that uncertainty remains a black-box scalar rather than a pair of interpretable, actionable diagnostics. The modular framework addresses this gap by isolating each source in its own module, enforcing the additive relation at every stage, and thereby overcoming the fundamental limitation shared by all prior approaches.

Modular Framework: Components

The functional responsibilities and independence properties of the five architectural modules are systematically consolidated in **Table 2**.

Table 2. Functional Roles of the Five-Module Architecture in Enabling Actionable Uncertainty Decomposition

Module	Primary Function	Input	Output	Independence Property	Example Implementations
Feature Extractor	Encodes atomic configurations into invariant representation	Atomic structure	Representation vector	Shared across all modules	GNN, MPNN, Equivariant Transformer
Epistemic Module	Quantifies model uncertainty	Representation vector	$\sigma_{epistemic}^2$	Independent of data noise modeling	Deep ensembles, MC dropout, Bayesian NN
Aleatoric Module	Learns irreducible data noise	Representation vector	$\sigma_{aleatoric}^2$	Independent of model hypothesis space	Heteroscedastic noise head, residual model
Aggregator	Enforces variance decomposition	$\sigma_{epistemic}^2, \sigma_{aleatoric}^2$	σ_{total}^2	Deterministic combination	Additive variance rule
Calibration Module	Ensures statistical reliability	All variance outputs	Calibrated uncertainties	Post-hoc independent adjustment	Conformal prediction, temperature scaling

The proposed framework comprises exactly five modules whose interfaces are deliberately minimal and whose interactions are governed by a single variance-additivity rule. Each module can be realized by multiple concrete implementations, allowing plug-and-play extensibility. The architecture is organized around a shared representation that anchors all subsequent uncertainty quantification. A feature extractor maps raw atomic configurations—positions, species, and cell vectors—into a fixed-dimensional, permutation-invariant embedding, ensuring that both epistemic and aleatoric pathways operate over an identical informational substrate. This shared basis enables consistent attribution of uncertainty sources and is compatible with modern equivariant or graph-based backbones [7, 19, 28, 29]. Within this representation space, epistemic uncertainty is expressed as model-dependent variability, typically realized through ensemble variance, Monte Carlo dropout at inference [14], or variational Bayesian formulations, yielding $\sigma_{epistemic}^2$ that contracts as data coverage improves. By contrast, aleatoric uncertainty is learned as an intrinsic property of the data-generating process, either via a heteroscedastic noise head conditioned on the same representation [13] or through residual models trained against higher-level reference calculations; its magnitude, $\sigma_{aleatoric}^2$, remains effectively invariant once the reference method is fixed.

These components are coupled through a minimal aggregation constraint,

$$\begin{aligned} \sigma_{total}^2 &= \sigma_{epistemic}^2(1) \\ &+ \sigma_{aleatoric}^2 \end{aligned}$$

applied element-wise to preserve physical units and tensor structure across atomic or force-level predictions. Such decomposition enables post hoc calibration to operate with greater precision: lightweight schemes such as temperature

scaling, conformal prediction, or isotonic regression adjust uncertainty estimates to achieve statistical reliability on validation data, with the added flexibility that each component can be calibrated independently once disentangled.

Addition occurs element-wise for per-atom or per-force predictions, preserving physical units and tensorial character. Post-aggregation, a lightweight calibrator (temperature scaling, conformal prediction intervals, or isotonic regression) adjusts both variance components to achieve statistical reliability on a validation set. Because calibration is applied after separation, each component can be calibrated independently if desired.

The hierarchical architecture enforcing the strict separation and recombination of epistemic and aleatoric uncertainty is illustrated in **Figure 1**.

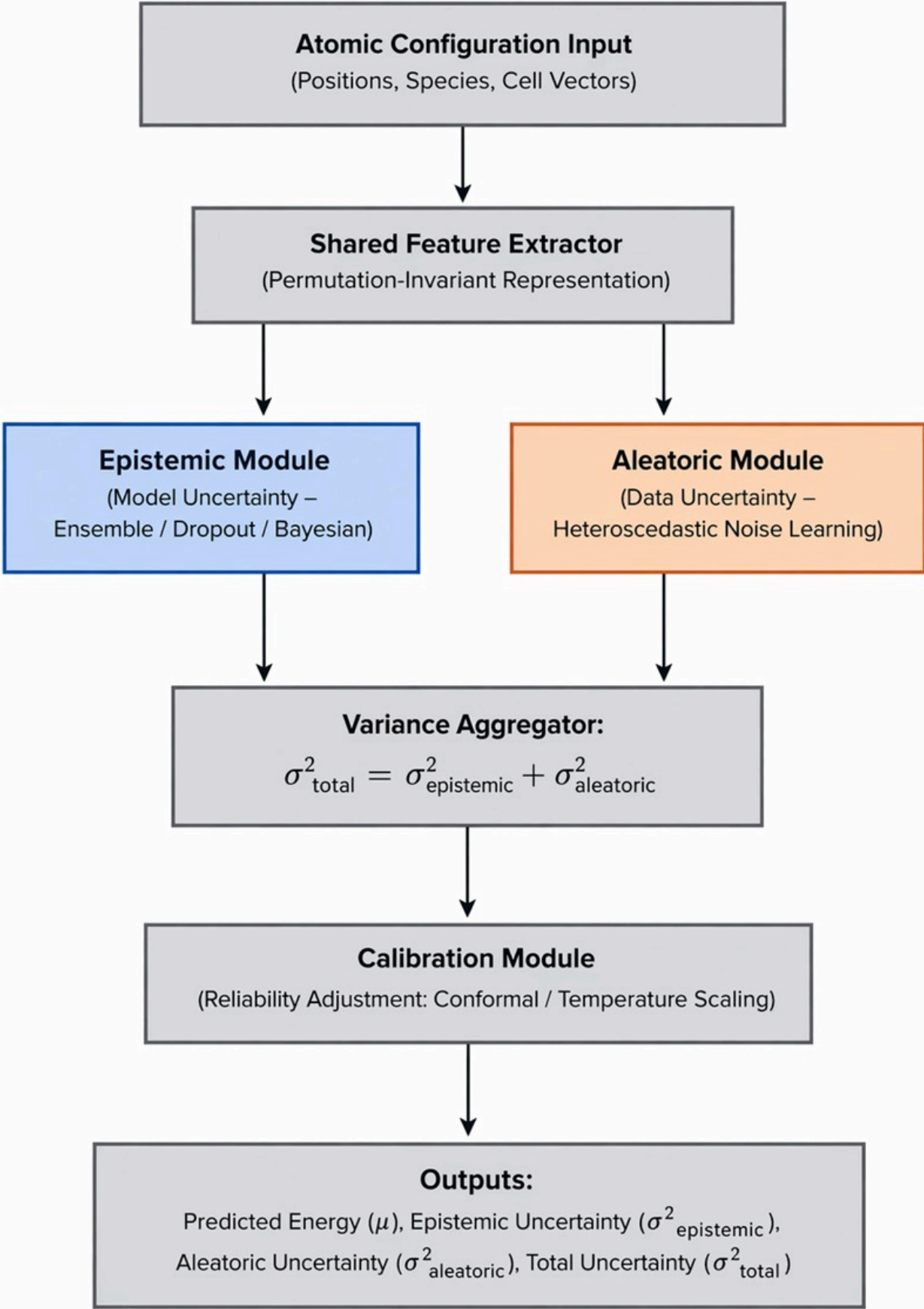


Figure 1. Modular Hierarchical Architecture for the Separation of Epistemic and Aleatoric Uncertainty in Machine Learning Potentials

By construction, the five-component design satisfies three desiderata that no prior method simultaneously meets: (i) clean separation at inference time, (ii) compatibility with any underlying MLP backbone, and (iii) direct actionability for downstream tasks. The remaining sections of the manuscript elaborate how these modules interact during training and inference, define operational tests that confirm successful separation, and outline the implications for active learning, model reporting, and the future development of uncertainty-aware ML potentials.

How the Modules Interact

The modular framework is defined not only by its five independent components but by the precise protocols that govern their interactions during both training and inference. These protocols ensure that epistemic and aleatoric uncertainties remain disentangled at every computational step while preserving complete interchangeability of individual modules.

In the training phase, the shared Feature Extractor first processes every atomic configuration into a representation vector. This vector is passed simultaneously to the Epistemic Module and the Aleatoric Module, which are optimized under separate loss terms. The Epistemic Module minimizes disagreement across its internal hypotheses (ensemble members, dropout masks, or variational samples) while fitting the mean energy surface. The Aleatoric Module, by contrast, is trained exclusively on the residuals between the mean prediction and the reference data; its objective is to reproduce the observed noise structure without influencing the primary energy prediction. Because the two modules share only the representation and the mean target, their parameter updates remain orthogonal. The Aggregator then receives the two variance outputs and applies the additive rule.

$$\begin{aligned}\sigma^{2total} \\ &= \sigma^{2epistemic} (2) \\ &+ \sigma^{2aleatoric}\end{aligned}$$

before the Calibration Module performs a joint or separate recalibration pass on a held-out validation split. Crucially, the training schedule itself is modular: one may freeze the Aleatoric Module after an initial warm-up and continue refining only the Epistemic Module, or vice versa, without retraining the entire network.

At inference time the interaction simplifies further. A new configuration enters the Feature Extractor once. The Epistemic Module produces its variance estimate in a single forward pass (or M passes for ensembles), while the Aleatoric Module produces its environment-dependent noise estimate in parallel. The Aggregator combines them instantaneously, and the Calibration Module applies pre-learned scaling factors. The final output is therefore a quadruple—predicted energy, epistemic uncertainty, aleatoric uncertainty, and total uncertainty—delivered with zero additional overhead beyond the cost of the individual modules.

This training-versus-inference separation is what enables true modularity. A practitioner may replace the Epistemic Module with a different ensemble size or a dropout-based approximation [14] while retaining the identical Aleatoric Module and Aggregator. The interface contracts are minimal: each module accepts the same representation vector and returns a scalar (or tensor) variance. No module ever sees the internal parameters of another. Consequently, the framework supports incremental upgrades—improving epistemic estimation without touching aleatoric modeling, or upgrading the noise head when a new reference method becomes available—without invalidating prior calibration or downstream workflows. The result is an architecture in which epistemic and aleatoric uncertainties are not merely reported separately; they are computed, stored, and propagated as first-class, independently addressable quantities throughout the entire model lifecycle [1, 13, 18].

Operational Criteria for Separation

To guarantee that a given implementation truly separates epistemic from aleatoric uncertainty, the framework supplies four falsifiable operational criteria. These criteria are model-agnostic and can be applied to any MLP architecture once the five modules are in place.

Validation of the decomposition hinges on its behavior under controlled perturbations of data, reference fidelity, and acquisition strategy. Under progressive data scaling, epistemic variance should contract monotonically toward zero as coverage of configuration space improves, while the aleatoric term remains statistically stationary once the reference method is fixed; any systematic drift in the latter signals leakage between components and undermines interpretability [5, 24]. A related diagnostic arises when the same architecture is retrained across distinct DFT functionals or basis sets on identical configurations: epistemic uncertainty should remain invariant, reflecting unchanged structural ignorance, whereas the aleatoric contribution must shift in accordance with the altered reference error, thereby confirming that noise attribution is decoupled from model parameters [18].

This separation imposes measurable consequences for acquisition dynamics. When active learning is driven exclusively by epistemic output, reductions in total prediction error should outpace strategies based on random or aggregate uncertainty, while the aleatoric component remains flat across iterations, indicating that irreducible noise is not being misinterpreted as informative signal. In the asymptotic regime of extensive training data, total uncertainty in well-sampled regions should converge to the aleatoric floor; failure of the epistemic term to vanish alongside a nonzero plateau in total variance indicates incomplete decomposition and residual entanglement. Such convergence behavior provides a concrete empirical marker that the model has saturated its learnable capacity, shifting the locus of improvement from data acquisition to refinement of the reference methodology.

When all four criteria are satisfied simultaneously, the practitioner can be confident that the reported epistemic and aleatoric uncertainties are cleanly separated, interpretable, and actionable. These criteria therefore serve as both validation protocol and diagnostic tool, closing the loop between conceptual design and practical deployment [3, 13, 17].

Relation to Active Learning and Decision-Making

The separation of epistemic and aleatoric uncertainty directly redefines how active learning loops and high-stakes materials-discovery decisions are structured. In conventional pipelines a single uncertainty scalar drives query selection, often leading to inefficient sampling of regions where the reference method itself is noisy. Within the modular framework the Epistemic Module alone supplies the acquisition function, directing computational effort exclusively toward reducing model ignorance. The Aleatoric Module, meanwhile, provides an explicit stopping criterion: once epistemic uncertainty falls below a user-chosen threshold relative to the local aleatoric floor, further data acquisition is demonstrably futile.

For materials discovery campaigns the joint information yields a compact decision matrix. High epistemic plus low aleatoric signals “collect more reference data in this region.” Low epistemic plus high aleatoric signals “the reference method must be upgraded before this configuration can be trusted.” High values of both indicate a region that is both unexplored and intrinsically noisy, warranting immediate reference-method refinement followed by targeted data generation. Low values of both confirm that the potential is ready for production use in that domain.

This decision logic extends beyond active learning to any downstream task that consumes uncertainty. When propagating errors into thermodynamic integration or phase-diagram construction, the epistemic component can be reduced by ensemble averaging or additional training, while the aleatoric component must be carried forward as an irreducible contribution to the final property variance. The framework therefore converts uncertainty from a passive confidence interval into a precise map of leverage points: where to invest data, where to invest methodological improvement, and where to accept intrinsic limits [1, 15–17, 23].

Implications for Model Development

Adoption of the modular framework reconfigures how ML potentials are specified, validated, and disseminated. Architectures are required to expose epistemic and aleatoric outputs as intrinsic interface elements rather than retrospective augmentations, thereby repositioning uncertainty separation as a design constraint. This shift propagates to reporting practices, where model cards and benchmark studies must present both components independently on standardized test sets, enabling comparison of epistemic efficiency across architectures. It further necessitates validation pipelines in which the four operational criteria function as non-negotiable checkpoints; failure to satisfy any condition precludes claims of separation and instead directs attention to architectural refinement, often through modification of loss formulations or the coupling between epistemic and aleatoric pathways.

Beyond individual models, the framework promotes a modular ecosystem in which components are treated as interoperable units. Pre-trained aleatoric modules calibrated to specific DFT functionals can be disseminated and reused, allowing independent development of epistemic estimators without recalibration of noise statistics, while advances in epistemic modeling can be integrated into established pipelines without perturbing aleatoric baselines. Such composability accelerates methodological iteration while preserving conceptual rigor [2, 6, 18, 22]. Under these conditions, uncertainty quantification transitions from a peripheral diagnostic to a governing principle of model evaluation, where performance is judged not only by predictive accuracy but by the fidelity with which models distinguish between reducible ignorance and intrinsic variability.

Conclusion

The central problem addressed by this conceptual framework is simple yet pervasive: contemporary ML potentials report a single uncertainty number that conflates model ignorance with irreducible data noise, thereby obscuring the very information practitioners need to make informed decisions. By enforcing a clean additive decomposition

through five interoperable modules, the framework supplies a modular, architecture-agnostic solution that applies to any interatomic potential family.

The five components—shared Feature Extractor, Epistemic Module, Aleatoric Module, Aggregator, and Calibration Module—interact via well-defined protocols that keep the two uncertainty sources distinguishable from training through inference. Four operational criteria provide rigorous, falsifiable tests that confirm successful separation in practice. When combined with targeted active-learning strategies and risk-aware decision matrices, the framework transforms uncertainty from a passive error bar into an actionable diagnostic that tells the materials scientist exactly where to allocate the next computational resource.

The broader implication is a shift in reporting standards: future ML potential publications should routinely disclose epistemic and aleatoric uncertainties separately, validated against the criteria set forth here. Only then can the community move beyond aggregated confidence scores toward a mature, uncertainty-aware discipline in computational materials engineering. The modular framework presented offers the conceptual scaffold required for that transition.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Jul 2022

Revised: 14 Sep 2022

Accepted: 22 Dec 2022

Published online: 18 January 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

- Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv Neural Inf Process Syst.* 2017;30:6402-13.
<https://doi.org/10.5555/3295222.3295387>.
- Busk J, Schmidt MN, Winther O, Vegge T, Jørgensen PB. Graph neural network interatomic potential ensembles with calibrated aleatoric and epistemic uncertainty on energy and forces. *Phys Chem Chem Phys.* 2023;25(37):25828-37.
<https://doi.org/10.1039/d3cp02143b>.
- Carrete J, Montes-Campos H, Wanzenböck R, Heid E, Madsen GKH. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *J Chem Phys.* 2023;158(20):204801.
<https://doi.org/10.1063/5.0146905>.
- Thaler S, Doehner G, Zavadlav J. Scalable Bayesian uncertainty quantification for neural network potentials: Promise and pitfalls. *J Chem Theory Comput.* 2023;19(14):4520-32.
<https://doi.org/10.1021/acs.jctc.2c01267>.
- Liu JZ, Paisley J, Kioumourtzoglou MA, Coull BA. Accurate uncertainty estimation and decomposition in ensemble learning. *Adv Neural Inf Process Syst.* 2019;32:8952-63.
<https://doi.org/10.5555/3454287.3455090>.
- 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>.
- 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>.

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>.

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

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, 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>.

Chen H, Stoddart JF. From molecular to supramolecular electronics. *Nat Rev Mater*. 2021;6(9):804-28.
<https://doi.org/10.1038/s41578-021-00302-2>.

Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Adv Neural Inf Process Syst*. 2017;30:5574-84.
<https://doi.org/10.5555/3295222.3295309>.

Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge (MA): MIT Press; 2016.

Jinnouchi R, Karsai F, Kresse G. On-the-fly machine learning force field generation: Application to melting points. *Phys Rev B*. 2019;100(1):014105.
<https://doi.org/10.1103/PhysRevB.100.014105>.

Xie Y, Vandermause J, Sun L, Cepellotti A, Kozinsky B. Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene. *NPJ Comput Mater*. 2021;7(1):40.
<https://doi.org/10.1038/s41524-021-00510-y>.

Kulichenko M, Barros K, Lubbers N, Li YW, Messerly R, Tretiak S, et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat Comput Sci*. 2023;3(3):230-9.
<https://doi.org/10.1038/s43588-023-00406-5>.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506.
<https://doi.org/10.1007/s10994-021-05946-3>.

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>.

Olivier A, Shields MD, Graham-Brady L. Bayesian neural networks for uncertainty quantification in data-driven materials modeling. *Comput Methods Appl Mech Eng*. 2021;386:114079.
<https://doi.org/10.1016/j.cma.2021.114079>.

Vazquez-Salazar LI, Boittier ED, Meuwly M. Uncertainty quantification for predictions of atomistic neural networks. *Chem Sci*. 2022;13(44):13068-84.
<https://doi.org/10.1039/d2sc04056e>.

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>.

Kahle L, Zipoli F. Quality of uncertainty estimates from neural network potential ensembles. *Phys Rev E*. 2022;105(1):015311.
<https://doi.org/10.1103/PhysRevE.105.015311>.

Vassaux M, Wan S, Edeling W, Coveney PV. Ensembles are required to handle aleatoric and parametric uncertainty in molecular dynamics simulation. *J Chem Theory Comput*. 2021;17(8):5187-97.
<https://doi.org/10.1021/acs.jctc.1c00526>.

Heid E, McGill CJ, Vermeire FH, Green WH. Characterizing uncertainty in machine learning for chemistry. *J Chem Inf Model*. 2023;63(13):4012-29.
<https://doi.org/10.1021/acs.jcim.3c00373>.

Klawohn S, Darby JP, Kermode JR, Csányi G, Caro MA, Bartók AP. Gaussian approximation potentials: Theory, software implementation and application examples. *J Chem Phys*. 2023;159(17):174108.
<https://doi.org/10.1063/5.0160898>.

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>.

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>.

Zhang L, Han J, Wang H, Saidi WA, Car R, E W. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. *Adv Neural Inf Process Syst*. 2018;31:4436-46.
<https://doi.org/10.5555/3327345.3327356>.