

REVIEW

Open access

Uncertainty Quantification for ML Interatomic Potentials: A Review of Methods, Hidden Assumptions, and Unresolved Questions

Daniel Fischer^{1*}, Laura Meier¹, Thomas Braun², Stefan Koch², Felix Roth¹

Abstract

Uncertainty quantification (UQ) has become indispensable for the trustworthy deployment of machine learning interatomic potentials (MLIPs) in materials science and molecular modeling, where predictions of energies, forces, and derived properties directly inform high-stakes decisions in materials discovery, long-time-scale molecular dynamics, and autonomous design workflows. Without reliable uncertainty estimates, MLIPs risk propagating errors that compromise simulation stability, mislead experimental prioritization, or produce unphysical results in extrapolation regimes critical to novel alloy or molecular discovery. This review synthesizes the literature on UQ methods specifically developed for or applied to MLIPs, drawing exclusively from the compiled reference set to provide a focused, critical overview of progress during this formative period. The scope is deliberately restricted to UQ techniques for interatomic potentials themselves (including GAP, DeepMD, ANI-series, SchNet-derived, and E(3)-equivariant models such as NequIP), excluding standalone ML property prediction unless the method directly supports force-field uncertainty. A systematic taxonomy organizes existing approaches into five methodological families—Bayesian and probabilistic methods, ensemble methods, Gaussian process and kernel methods, conformal prediction and frequentist methods, and heuristic and ad hoc methods—highlighting their distinct mathematical foundations and practical implementations in MLIP contexts. Hidden assumptions pervading these families are identified and dissected, including independence of atomic errors, Gaussianity of predictive distributions, homoscedasticity across chemical space, kernel-imposed smoothness in Gaussian processes, approximation quality in variational or Monte-Carlo inference, and exchangeability in conformal frameworks. These assumptions frequently remain unstated yet profoundly influence calibration and reliability when MLIPs are deployed in production simulations. Unresolved questions are articulated with precision: how to treat correlated uncertainties along molecular-dynamics trajectories, the absence of a true ground-truth uncertainty given DFT approximations, the prohibitive computational overhead of scalable UQ, evaluation under distribution shift, vectorial uncertainty for forces rather than scalar energies, detection of physical inconsistencies, and hierarchical fusion of model, data, and ab-initio uncertainties. Future outlook points toward integrated UQ-driven active learning, force-aware uncertainty representations, and hybrid methods that balance calibration, sharpness, and efficiency for next-generation autonomous materials engineering.

Keywords Uncertainty quantification, Ensemble methods, Machine learning interatomic potentials, Gaussian processes, Bayesian neural networks, Conformal prediction

*Correspondence:

Daniel Fischer

daniel.fischer@outlook.com

¹ Department of Computational Materials Analytics, Faculty of Engineering, University of Freiburg, Freiburg, Germany

² Department of Materials Data Science, Faculty of Technology, Karlsruhe Institute of Technology, Karlsruhe, Germany

Introduction

Machine learning interatomic potentials have transformed computational materials science by enabling quantum-mechanical accuracy at force-field speeds, powering large-scale molecular dynamics, high-throughput screening, and inverse design of novel compounds [1]. Yet, as these models are increasingly embedded in decision-making pipelines, the absence of reliable uncertainty estimates severely limits their utility [2, 3]. Predictions without accompanying confidence measures cannot support risk-aware materials discovery, where erroneous extrapolation to unseen compositions or structures may waste experimental resources or lead to unsafe conclusions [2, 4, 5]. Early recognition of this limitation appears in foundational works that unified materials and molecular modeling while implicitly acknowledging the need for uncertainty-aware frameworks [6, 7].

Deringer et al. provided a comprehensive overview of Gaussian process regression for materials and molecules, underscoring that predictive variance is not merely a byproduct but a core enabler of trustworthy modeling [6]. Similarly, Bartók et al. demonstrated how machine learning can unify materials and molecules, yet their 2017 framework already hinted at the necessity of uncertainty propagation for practical deployment [7]. Subsequent developments in deep potential molecular dynamics and E(3)-equivariant graph neural networks further accelerated adoption [8, 9], yet these scalable architectures inherited the black-box nature of deep learning, amplifying the demand for explicit UQ [2, 10].

This review surveys UQ methods applied to MLIPs strictly within the 2017–2022 window, organizing them into a taxonomy that spans Bayesian, ensemble, Gaussian-process, conformal, and heuristic families [2, 11–15]. The analysis identifies patterns across disparate literatures, critically evaluates hidden assumptions that undermine reliability, and articulates unresolved challenges that must be addressed before MLIPs can be considered production-ready for autonomous discovery [2, 13, 16]. Scope is tightly delimited to uncertainty in interatomic potentials themselves—energies and forces—rather than downstream properties such as band gaps or formation energies, unless the UQ technique directly augments force-field construction [2, 10]. Methods covered include those applied to GAP, DeepMD, ANI-1, SchNet derivatives, and NequIP-style equivariant networks [1, 8, 9, 17, 18].

The motivation is both scientific and practical. In active-learning loops, uncertainty guides data acquisition; in molecular dynamics, it flags when trajectories drift into unreliable regimes; in high-entropy alloy design, it quantifies confidence in phase stability predictions [19–21]. Peterson et al. explicitly addressed uncertainty in atomistic machine learning [2], while Imbalzano et al. examined uncertainty estimation for molecular dynamics and sampling, setting the stage for the systematic treatment presented here [16]. Despite these advances, the literature remains fragmented: Bayesian approaches emphasize posterior inference but struggle with scale [13, 14], ensembles offer pragmatic calibration at high computational cost [5, 15], Gaussian processes deliver closed-form variance yet suffer from kernel sensitivity [6, 22], and conformal methods promise distribution-free guarantees that have seen limited adoption in force-field contexts [11].

This review therefore performs three essential functions: (i) a taxonomy that synthesizes methodological families with their mathematical bases, applications, and limitations; (ii) a critical dissection of hidden assumptions—independence, Gaussianity, homoscedasticity, kernel choice, approximation quality, and exchangeability—that are routinely overlooked [2, 13, 16]; and (iii) an articulation of open questions surrounding correlated errors, ground-truth definition, computational efficiency, extrapolation, force-vector uncertainty, and multi-source uncertainty fusion [2, 4, 5, 10].

Taxonomy of UQ Methods for MLIPS

A systematic taxonomy is essential to impose order on the heterogeneous UQ landscape for ML interatomic potentials. Five primary families emerge from the 2017–2022 literature, each characterized by distinct mechanisms for representing and propagating uncertainty.

Figure 1 organizes the review as a hierarchical analytical framework that links the five major UQ families for MLIPs to their underlying assumptions, recurring evaluation weaknesses, and unresolved research challenges.

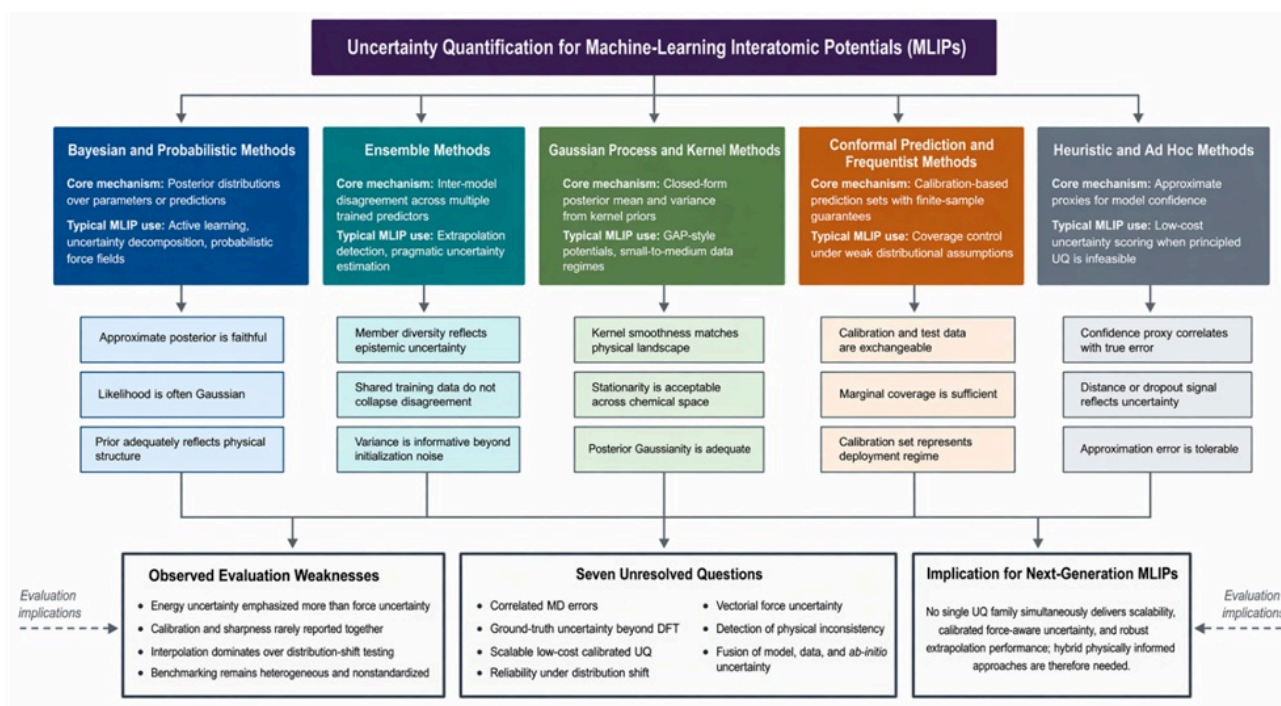


Figure 1. Hierarchical framework of uncertainty quantification methods for machine learning interatomic potentials

Bayesian and probabilistic formulations anchor one end of the methodological spectrum by treating model parameters as latent random variables, allowing prior beliefs to be updated into posterior distributions that propagate uncertainty directly into predictions. This probabilistic grounding provides a principled route to uncertainty quantification, yet its practical realization often encounters computational and approximation challenges in high-dimensional interatomic models. A related perspective emerges from ensemble-based strategies, where multiple models—differentiated through initialization, data partitioning, or stochastic training dynamics—serve as an empirical proxy for epistemic uncertainty, with variance across predictions interpreted as a signal of model disagreement. While less formally grounded than full Bayesian inference, this approach has gained traction due to its scalability and compatibility with deep learning architectures.

Beyond these parametric viewpoints, Gaussian process and kernel-based methods offer a nonparametric alternative in which uncertainty is encoded directly through a prior over functions. By defining a Gaussian prior over the potential energy surface, these approaches yield closed-form posterior mean and variance estimates, thereby providing analytically tractable uncertainty measures under specified kernel assumptions. In contrast, conformal prediction and related frequentist techniques shift attention away from model structure toward calibration guarantees, using nonconformity scores and held-out calibration sets to produce prediction intervals with finite-sample validity under exchangeability assumptions. This distribution-free perspective introduces a distinct notion of reliability, one that is decoupled from parametric assumptions but contingent on the stability of data-generating conditions.

In practice, however, neither fully Bayesian nor strictly calibrated approaches are always computationally feasible, particularly in large-scale atomistic simulations. This constraint has motivated the adoption of heuristic and approximation-driven methods, including Monte Carlo dropout and distance-based proxies tied to the training distribution. Such techniques trade theoretical rigor for tractability, offering lightweight uncertainty indicators that can be integrated into existing workflows, albeit with less transparent interpretability.

These methodological strands can be organized within a unified taxonomy, conceptually represented as a rooted tree in which “UQ for MLIPs” occupies the central node. From this origin, distinct branches capture the major families of

approaches, each associated with characteristic mathematical structures and representative contributions [6, 11-13, 15]. The internal structure of the tree reflects further specialization, linking methodological variants—such as variational inference or deep ensembles—to specific model classes including GAP, DeepMD, and NequIP [8, 9, 18]. Complementing this hierarchical view, a matrix-style representation can illuminate how these families perform across key operational criteria, including scalability, calibration guarantees, support for force-vector uncertainty, and robustness under extrapolation. When viewed through this comparative lens, a temporal pattern becomes apparent: Gaussian-process methods dominate earlier developments, whereas ensemble techniques have become increasingly prominent with the rise of deep neural potentials [5, 6, 15].

This taxonomy is not merely descriptive but serves as an analytical framework for the sections that follow, where each methodological family is examined in terms of its mathematical foundations, practical deployment, and underlying assumptions [2, 13, 16]. Taken together, these analyses point to a structural limitation that cuts across the literature. No single approach simultaneously satisfies the competing demands of modern MLIP applications, particularly when scalability, physically meaningful uncertainty over forces, and robustness to distributional shift are considered in combination. This tension suggests that progress is unlikely to arise from further refinement of isolated methods alone, but rather from the development of hybrid strategies that integrate complementary strengths across paradigms [2, 5, 13, 14].

Table 1 clarifies that the major UQ families differ not only in mathematical machinery, but also in the specific assumptions that determine when their uncertainty estimates remain scientifically trustworthy.

Table 1. Comparative analytical structure of uncertainty quantification families for machine-learning interatomic potentials

UQ family	Statistical basis	Primary uncertainty target	Main strength in MLIP setting	Hidden assumption that most affects trustworthiness	Main failure mode in practice	Best-fit MLIP context
Bayesian and probabilistic methods	Posterior inference over parameters and/or predictions	Epistemic, sometimes aleatoric	Principled decomposition of uncertainty; natural fit for active learning	Approximate posterior is a faithful proxy for true posterior; likelihood is often assumed Gaussian	Overconfident inference under posterior approximation error or prior misspecification	Data-efficient workflows, rare-event exploration, uncertainty-guided sampling
Ensemble methods	Empirical variance across independently trained models	Mostly epistemic	Strong pragmatic performance; easy parallelization; often better empirical calibration than single models	Ensemble diversity genuinely reflects knowledge uncertainty rather than shared bias or initialization noise	High computational cost; disagreement may understate uncertainty when members are too similar	Deep neural MLIPs, extrapolation monitoring, production screening
Gaussian process and kernel methods	Nonparametric Bayesian regression with kernel-defined covariance	Predictive mean and variance, often highly interpretable	Closed-form uncertainty estimates; strong small-data behavior;	Kernel smoothness, stationarity, and Gaussian predictive form	Sensitivity to descriptor/kernel choice; scalability limitations without sparsification	GAP-style potentials, small-to-medium datasets,

			natural derivative handling	are physically adequate		exact variance prioritization
Conformal prediction and frequentist methods	Calibration-based coverage control using nonconformity scores	Predictive intervals/sets with marginal coverage guarantees	Distribution-light validity; attractive post hoc calibration framework	Calibration and deployment data are exchangeable; marginal coverage is scientifically sufficient	Coverage weakens under sequential sampling, MD dependence, or distribution shift	Post hoc calibration, risk control, uncertainty reporting layers over existing MLIPs
Heuristic and ad hoc methods	Proxy-based confidence indicators such as dropout or distance-to-training-set	Mixed or unspecified	Low implementation cost; scalable; easy integration into existing pipelines	Proxy signal is monotonic with true prediction error	Apparent confidence can be poorly calibrated and method-specific	Fast screening, preliminary safety checks, low-cost active-learning triggers

Bayesian and Probabilistic Methods

Bayesian methods treat the parameters θ of an MLIP as random variables endowed with a prior $p(\theta)$. Given training data $D=\{(x_i, y_i)\}_{i=1}^N$, the posterior $p(\theta|D)$ is obtained via Bayes' rule, and predictions are formed from the posterior predictive

$$\text{distribution: } p(y^* | x^*, D) = \int p(y^* | x^*, \theta) p(\theta | D) d\theta.$$

This formulation naturally separates epistemic uncertainty (posterior variability) from aleatoric uncertainty (likelihood noise). In practice, exact inference is intractable for neural-network potentials, so variational inference or MCMC approximations are employed. Xie *et al.* demonstrated Bayesian force fields constructed via active learning for inter-dimensional transformations, showing how posterior samples can guide exploration of rare events. Kőfing *et al.* applied Bayesian inference to optimize empirical force fields, illustrating the value of physically informed priors [23]. Venturi *et al.* quantified uncertainties on ab-initio potential energy surfaces using Bayesian machine learning, highlighting the propagation of parametric uncertainty into thermodynamic observables.

Hidden assumptions include the choice of prior (frequently Gaussian or uninformative, rarely encoding known physical symmetries) and the quality of the approximate posterior (variational gap or MCMC convergence diagnostics are seldom reported). The likelihood is typically assumed Gaussian, yet force errors in MLIPs often exhibit heavy tails. Scalability remains a bottleneck: full Bayesian neural networks for large systems are rare, and even sparse approximations incur overhead. Nevertheless, these methods provide principled uncertainty decomposition and have been successfully coupled with active learning to reduce data requirements in GAP-like frameworks. Limitations are most acute in high-dimensional composition spaces where prior misspecification leads to overconfident posteriors.

Ensemble Methods

Ensemble methods train M independent models $\{f_m\}_{m=1}^M$ and estimate predictive uncertainty from the empirical

$$\begin{aligned} \mu(x^*) &= \frac{1}{M} \sum_{m=1}^M f_m(x^*) \\ \text{variance: } \sigma^2(x^*) &= \frac{1}{M} \sum_{m=1}^M [f_m(x^*) - \mu(x^*)]^2 \end{aligned}$$

Lakshminarayanan *et al.* established that deep ensembles yield well-calibrated uncertainty with minimal hyperparameter tuning. Vassaux *et al.* showed that ensembles are required to disentangle aleatoric and parametric uncertainty in molecular dynamics, demonstrating improved trajectory stability when ensemble variance is monitored. Applications to MLIPs include bootstrap ensembles for GAP variants and random-initialization ensembles for SchNet-derived potentials, where inter-model disagreement flags extrapolation.

The core assumption is that ensemble diversity captures epistemic uncertainty rather than mere initialization noise; this holds only when training data subsets or architectures differ meaningfully. Independence among members is frequently violated when all models share the same training set. Advantages include straightforward parallelization and empirical calibration superior to single-model baselines. Limitations center on the M -fold training cost, which becomes prohibitive for production-scale MD, and the lack of theoretical coverage guarantees. In the reviewed literature, ensembles are predominantly evaluated on energy predictions; force-vector covariances are rarely reported.

Gaussian Process and Kernel Methods

Gaussian processes place a prior $f \sim GP(m, k)$ over the energy surface. Given observations, the posterior predictive

$$\begin{aligned} \text{variance is: } \sigma^2(x^*) &= k(x^*, x^*) - k^{*\top} (K + \sigma^2 I)^{-1} k^* \end{aligned}$$

where K is the kernel matrix and k^* the cross-covariance vector.

Deringer *et al.* reviewed Gaussian process regression for materials and molecules, emphasizing its exact uncertainty estimates. Bartók *et al.* developed machine-learning general-purpose interatomic potentials for silicon, while their earlier work on improved UQ for GAP explicitly addressed predictive variance in force predictions. Sparse GPs and derivative-informed kernels (incorporating forces into training) have been used to mitigate $O(N^3)$ scaling.

Kernel choice—squared-exponential versus Matérn—implicitly encodes smoothness and extrapolation behavior; stationarity is routinely assumed despite the non-stationary nature of potential energy surfaces across chemical space. The Gaussianity of the posterior is convenient but may mismatch true error distributions. Limitations include computational cost without sparsification and the under-discussed sensitivity of uncertainty to kernel hyperparameters. These methods remain the gold standard for small-to-medium systems where exact variance is paramount.

Conformal Prediction and Frequentist Methods

Conformal prediction constructs prediction sets with finite-sample coverage guarantees under exchangeability. A nonconformity score $s(y, \hat{y})$ is computed on a calibration set; the $(1-\alpha)$ -quantile of scores defines the threshold for test-set intervals. Hu *et al.* introduced robust and scalable conformal prediction for machine-learned interatomic potentials, demonstrating marginal coverage for molecular properties while highlighting the need for careful calibration-set design.

The exchangeability assumption is weaker than i.i.d. but still strained by temporally or spatially correlated MD data. Coverage is marginal rather than conditional, and intervals can become uninformative under severe distribution shift. Applications to MLIPs remain sparse; jackknife-style variants have been explored for force uncertainty but lack widespread adoption. Advantages include distribution-free guarantees and computational efficiency at inference. Limitations include dependence on a held-out calibration set and the absence of natural vectorial extensions for forces.

Hidden Assumptions in Current UQ Methods

Across the major families of uncertainty quantification (UQ) methods used in machine-learning interatomic potentials (MLIPs), a relatively small set of hidden assumptions repeatedly structures model design, calibration practice, and empirical interpretation. These assumptions are often reasonable as first-order approximations, yet they become consequential when MLIPs are deployed in precisely the settings where uncertainty matters most: long molecular dynamics (MD) trajectories, rare-event exploration, extrapolative chemistry, multicomponent materials, and active-learning workflows. The problem is not merely that assumptions are made—every statistical method rests on assumptions—but that in the MLIP literature they are often left implicit, weakly diagnosed, and rarely stress-tested under realistic violations. As a result, uncertainty estimates that appear numerically well behaved in static benchmark settings may become unreliable in downstream scientific use.

Table 2 consolidates the manuscript’s central critical claim by linking each hidden assumption to its downstream empirical consequence and to the validation test required for a scientifically credible UQ pipeline.

Table 2. Hidden assumptions, empirical consequences, and priority validation tests for uncertainty quantification in MLIPs

Hidden assumption	Where it appears most strongly	Why it is often left implicit	Consequence if violated	What should be tested empirically	Priority implication for future work
Independence of errors	Standard likelihoods, resampling logic, frame-wise calibration summaries	Benchmarking usually treats configurations as isolated test points	Trajectory-level risk is understated; correlated force errors accumulate over MD time	Long-horizon trajectory calibration, temporally blocked test splits, path-dependent error growth	Develop trajectory-aware UQ and correlation-sensitive evaluation
Gaussianity of predictive errors	GPs, Gaussian NLL objectives, variance-only summaries of ensembles	Mean–variance summaries are convenient and easy to compare	Heavy tails, skew, or multimodality are hidden; intervals become misleading	Tail-risk diagnostics, quantile calibration, non-Gaussian residual analysis	Move beyond variance-only UQ toward distribution-shape-aware methods
Homoscedasticity across chemical space	Global variance terms, pooled validation summaries	Simpler training and reporting pipelines	Rare regimes are smoothed into average uncertainty levels	Subgroup calibration across composition, coordination, defects, temperature, and strain	Build explicitly heteroscedastic, regime-aware UQ for MLIPs
Kernel-imposed smoothness and stationarity	GP and kernel-based methods	Kernel defaults are mathematically elegant and historically standard	Abrupt regime changes or heterogeneous PES structure are underrepresented	Sensitivity to kernel choice, nonstationary splits, regime-specific calibration	Use physically informed and nonstationary kernels where appropriate

Approximation quality of Bayesian surrogates	Variational inference, dropout, Laplace methods, posterior surrogates	Scalability pressures encourage approximation without posterior diagnostics	Epistemic uncertainty may be sharply underestimated in poorly constrained regions	Posterior-quality diagnostics, approximation ablations, agreement with independent uncertainty baselines	Report approximation uncertainty explicitly rather than implicitly trusting surrogates
Exchangeability in conformal calibration	Conformal prediction and post hoc calibration workflows	Formal guarantees appear attractive and transferable	Coverage claims degrade in active learning, sequential MD, and curated rare-event datasets	Sequential calibration tests, covariate-shift stress tests, time-ordered calibration splits	Extend conformal methods to realistic sequential materials pipelines
Ground-truth availability	All calibration frameworks using DFT labels as reference truth	DFT is operationally treated as exact target data	UQ may be calibrated to surrogate error while ignoring electronic-structure uncertainty	Cross-functional comparison, multi-fidelity validation, uncertainty decomposition beyond ML error	Integrate model, data, and ab-initio uncertainty into hierarchical UQ frameworks

Independence

A large share of current UQ procedures either explicitly or implicitly treats prediction errors as independent across observations [2, 16]. This assumption is deeply embedded in standard likelihood functions, bootstrap-style resampling logic, many calibration summaries, and the interpretation of pointwise confidence intervals. In atomistic simulation, however, errors are rarely independent in any practically meaningful sense. MD trajectories generate temporally correlated configurations; neighboring atoms induce strong spatial dependence in local environments; and force errors on nearby degrees of freedom often co-move because they are produced by the same latent representational failure. Independence-based reasoning can therefore make uncertainty accumulation appear smaller than it really is. A model may look acceptably calibrated on randomly sampled frames while being dangerously overconfident along a continuous trajectory, where small but correlated force errors compound over many steps into biased diffusion constants, incorrect transition rates, or qualitatively wrong phase behavior [2, 4]. This issue is especially severe in long-horizon simulations, where the operational object of interest is not a single prediction but a path-dependent dynamical process. Treating each frame as an isolated test point obscures this cumulative structure and can materially understate epistemic risk.

Gaussianity

Many UQ methods assume, or are interpreted through the lens of, approximately Gaussian predictive distributions [6, 13]. Gaussian processes encode Gaussian predictive marginals by construction; deep ensembles are often summarized only by mean and variance; heteroscedastic regressors are commonly trained under Gaussian negative log-likelihood objectives; and calibration is frequently evaluated with metrics whose meaning is clearest under symmetric, light-tailed predictive error. Yet force and energy errors in MLIPs are often neither symmetric nor light-tailed, especially in the presence of sparse training coverage, structural motifs absent from the training set, bond rearrangements, charged environments, or competing metastable states. Under extrapolation, error distributions can become skewed, heavy-tailed, or multimodal [2, 13]. In such settings, variance alone is an inadequate summary of uncertainty, and Gaussian assumptions can yield intervals that are too narrow in the tails or centered at unrepresentative means

Homoscedasticity

A further hidden assumption is that predictive noise or uncertainty has constant scale across the input space [2, 10]. Even when authors conceptually distinguish aleatoric and epistemic uncertainty, implementation choices often collapse to effectively homoscedastic behavior, especially when a global variance parameter is used or when uncertainty is post hoc derived from residual dispersion on pooled validation data. This is difficult to justify for atomistic systems. Error magnitudes differ substantially across local coordination motifs, chemical compositions, strain regimes, temperatures, pressures, defect environments, and regions of configuration space near versus far from equilibrium. For example, highly coordinated bulk atoms may be predicted with relative confidence, while undercoordinated surface atoms, transition states, dopant environments, or rare collision geometries may exhibit much larger and qualitatively different uncertainty. A constant-variance view therefore smooths over the very structure the model should detect.

Kernel/Smoothness in GPs

Gaussian-process-based UQ brings additional assumptions through the kernel [6, 22]. Popular default kernels often encode stationarity, isotropy, and high—sometimes effectively infinite—degrees of differentiability. These priors are mathematically convenient and often beneficial in small-data regression, but they may be physically mismatched to materials problems. Interatomic energy landscapes can be locally smooth and globally heterogeneous; they may contain abrupt changes in relevant curvature across bonding regimes, compositional neighborhoods, or near structural instabilities. Stationary kernels treat covariance as a function of relative distance in representation space rather than of physically meaningful regime changes, while overly smooth kernels can suppress sharp but real variations in forces or fail to reflect piecewise structure induced by different bonding patterns. In addition, kernel behavior depends on the descriptor representation supplied to the GP, so misspecification can enter both through the kernel and through the geometry of the feature space.

Approximation quality

Several scalable Bayesian-inspired approaches rely on approximations—variational inference, Monte Carlo dropout, Laplace approximations, low-rank posterior surrogates, or simplified ensemble interpretations—and then proceed as though the resulting predictive distribution is a faithful proxy for the true posterior uncertainty [13, 15]. This step is often accepted pragmatically, but the approximation gap is rarely measured. In high-dimensional neural potentials, posterior geometry is complex, multimodal, and strongly affected by symmetries, optimizer bias, finite-width effects, and non-identifiability. Approximate methods may capture only a narrow slice of this structure, sometimes underestimating epistemic uncertainty in exactly those regions where the model is least constrained [2, 13]. MC dropout, for instance, may provide a useful stochastic heuristic without constituting a reliable posterior approximation across architectures and training regimes. Variational methods may prefer overly concentrated solutions because of tractable objective design. Ensemble diversity may reflect optimization randomness more than principled posterior spread.

Exchangeability in conformal prediction

Conformal prediction has attracted attention because it offers finite-sample coverage guarantees under weak assumptions, most notably exchangeability between calibration and test data [11]. In MLIPs, however, that requirement is easy to violate. Active-learning loops deliberately alter the training and calibration distribution over time by querying informative or high-uncertainty structures. MD-generated samples are temporally dependent. Data curation often reweights rare events, unstable structures, or chemically important motifs. As a result, calibration residuals collected at one stage of model development may not be representative of the configurations encountered later in deployment. The formal coverage guarantee of conformal prediction is therefore less portable than it may appear [2, 11]. The danger is subtle: a method can retain the language of distribution-free validity while operating under conditions where the guarantee no longer strictly applies. In sequential discovery settings, exchangeability violation is not an edge case but a structural feature of the pipeline.

Ground-truth availability

Most calibration frameworks presume access to exact target values against which residuals can be measured and uncertainty can be tuned [2]. In MLIP development, those targets are often density functional theory (DFT) energies and forces, treated operationally as ground truth. Yet DFT itself contains functional-dependent systematic errors, basis-set limitations, numerical tolerances, and approximations that vary across material classes and chemical processes. For some applications, the dominant uncertainty may therefore arise not from the ML surrogate alone but from the reference electronic-structure method on which the surrogate is trained [13]. Calibrating an MLIP only to DFT residuals can produce internally consistent uncertainty estimates relative to the surrogate task while obscuring broader scientific uncertainty relative to experiment or higher-level theory. This issue becomes especially important when UQ is used for decision-making—screening candidate materials, estimating barriers, prioritizing active-learning queries, or terminating simulations upon threshold exceedance.

Taken together, these seven assumptions illuminate why apparently sophisticated UQ pipelines can still fail in practice. Their joint effect is to narrow the space of acknowledged uncertainty: correlations are ignored, non-Gaussian tails are compressed, heterogeneity is averaged away, prior smoothness is overimposed, approximation error is hidden, sequential sampling breaks formal guarantees, and label uncertainty is externalized. The result is not random noise around otherwise valid uncertainty estimates, but the possibility of systematic miscalibration. In this sense, current limitations are structural rather than incidental.

Empirical Evaluation Practices

Empirical evaluation practices for UQ in the MLIP literature remain heterogeneous, fragmented, and often underpowered relative to the claims being made [2, 4, 16]. Although many papers now report some measure of uncertainty quality, there is still no shared evaluation protocol comparable to what mean absolute error (MAE), root mean square error (RMSE), and benchmark test sets provide for point prediction. As a result, methods that differ substantially in what they quantify—and in the assumptions under which they are valid—are often compared using incomplete or noncomparable evidence [2, 13]. The field has made progress in recognizing calibration as distinct from accuracy, but evaluation practice has not yet caught up with that conceptual distinction [11, 24].

Other metrics appear only sporadically. Negative log-likelihood (NLL) is sometimes used because it rewards both accuracy and calibrated sharpness, but its interpretation depends strongly on the assumed predictive distribution; under Gaussian likelihoods, a low NLL may still be misleading if the true error distribution is skewed or heavy-tailed [13]. Sharpness—often operationalized as average predictive variance, entropy, or interval width—is even less consistently reported, despite being essential for distinguishing informative uncertainty from trivial uncertainty inflation [12, 15]. A model can achieve nominal coverage simply by issuing wide intervals; without a sharpness metric, such performance may be incorrectly viewed as strong UQ. The central evaluative principle is that calibration and sharpness must be examined jointly [13, 24]. Well-calibrated but uninformative uncertainty is not enough for scientific deployment, especially when uncertainty is meant to guide expensive data acquisition or simulation control [2, 4].

A second weakness in current practice is the dominance of energy-only evaluation. Many studies assess uncertainty on total energies while giving little or no attention to force uncertainty [2, 10]. This is a serious limitation because forces, not energies, drive MD trajectories, geometry optimization, phonon calculations, and many downstream dynamical observables [1, 3]. A model whose energy uncertainty appears reasonable may still provide poorly calibrated force uncertainties, particularly at the vector level where directional errors matter [10, 16]. Scalar summaries of force magnitude alone are insufficient; force uncertainty should ideally be assessed component-wise, norm-wise, and in relation to directional consistency. Yet such analyses remain rare [2, 16]. This creates a mismatch between evaluation and use: the community often tests UQ on the easiest prediction target and then implicitly generalizes the conclusion to simulation behavior governed by a harder one.

A third major inconsistency concerns distribution shift. Calibration is seldom tested explicitly under interpolation versus extrapolation regimes, even though the latter is where epistemic uncertainty should, in principle, be most valuable [2, 13].

Many benchmark splits remain random or near-random, producing train and test sets that are similar in local geometry and composition. Such splits are useful for measuring supervised generalization but weak for evaluating whether uncertainty rises appropriately when the model leaves the support of the training distribution [2, 4]. More demanding tests would separate equilibrium from nonequilibrium structures, low- from high-temperature samples, pristine from defect-containing systems, in-domain from compositionally novel materials, and relaxed from strained or reactive configurations. They would also include controlled extrapolation axes, allowing researchers to ask not merely whether errors increase, but whether predicted uncertainty increases in proportion to error growth. Hu *et al.* and Wen *et al.* provide rare examples of more rigorous calibration analysis in this direction [11, 12], and their work points toward what a stronger evaluation norm could look like.

The lack of standardized benchmarks further compounds these problems. Different papers use different datasets, different descriptors of uncertainty, different train/test partitions, and different target definitions [2, 13]. Some evaluate atomwise properties, others structure-level aggregates; some use validation splits for both tuning and assessment; others pool results across chemically distinct systems without checking whether calibration is stable across subgroups. This heterogeneity makes cross-paper comparison difficult. A method may appear better simply because it is evaluated on an easier system, a milder shift, a narrower uncertainty target, or a more forgiving metric. In the absence of benchmark standardization, the literature risks rewarding presentation rather than reliability.

Empirical practice is also weakened by limited attention to subgroup and regime-specific calibration. Reporting a single global ECE or one overall reliability diagram can conceal substantial heterogeneity across chemical environments [2, 16]. Calibration may differ between bulk and surface atoms, between majority and minority species in alloys, between low- and high-temperature states, or between near-equilibrium and transition-state-like geometries. For scientific applications, these subgroup failures are often more important than the global average because they concentrate in the rare regimes where conclusions are most sensitive. A more informative evaluation would stratify calibration by physically meaningful slices of the input space and report where uncertainty succeeds or fails.

There is, in addition, insufficient practice of uncertainty-aware ablation. Many UQ papers compare final methods without isolating the contribution of architecture, representation, dataset composition, training objective, posterior approximation, and calibration procedure [2, 13]. This makes it hard to know whether gains come from better uncertainty modeling or simply from better point prediction. For example, an ensemble may appear better calibrated partly because its members are more accurate, not because ensemble variance itself is a superior epistemic proxy [5, 15]. Careful ablations would help disentangle these effects and clarify whether improvements are methodological or incidental.

Given these limitations, several methodological recommendations emerge from the corpus. At a minimum, papers should report reliability diagrams alongside standard error metrics such as MAE, rather than presenting predictive accuracy alone. They should evaluate calibration separately on interpolation and extrapolation splits, with the latter designed to reflect realistic scientific deployment rather than random holdout convenience [11, 13]. They should include force-vector uncertainty analysis rather than relying primarily on scalar energy prediction [2, 10]. They should report both calibration and sharpness, so that nominal validity is not confused with inflated intervals [12, 24]. They should test performance on long MD trajectories and under explicit distribution shift [2, 4]. And where possible, they should assess calibration by physically meaningful subgroups, making it clear whether uncertainty remains trustworthy in rare but important regimes.

Overall, current empirical practice has reached an important transitional stage. The literature no longer treats UQ as a purely optional add-on; many authors now recognize that uncertainty must itself be evaluated [2, 16]. Yet evaluation remains too selective, too static, and too weakly standardized to support strong claims of reliability. Hu *et al.* and Wen *et al.* stand out because they move beyond token uncertainty reporting toward more rigorous calibration analysis [11, 12]. Their examples suggest a path forward, but they remain exceptions rather than the norm. For UQ in MLIPs to become scientifically credible as a basis for simulation control, active learning, and discovery, empirical evaluation must shift from opportunistic reporting to disciplined validation. The key question is no longer whether a model can output an uncertainty score, but whether that score has been tested in the regimes where users would actually trust it.

Unresolved Questions and Open Challenges

Seven open questions persist.

Question 1: How should UQ handle correlated errors in MD trajectories? Standard independence assumptions fail.

Question 2: What constitutes “ground truth” uncertainty when DFT is itself approximate?

Question 3: Can computationally cheap UQ scale to million-atom simulations without sacrificing calibration?

Question 4: How should UQ be evaluated and guaranteed under distribution shift?

Question 5: How to represent and propagate uncertainty for force vectors rather than scalar energies?

Question 6: Can UQ detect physical violations (negative phonons, Born instability) beyond mere prediction error?

Question 7: How to hierarchically combine model, data, and ab-initio uncertainties?

Novikov *et al.*, Xie *et al.*, and Vassaux *et al.* touch on subsets of these issues but offer no comprehensive solutions. Addressing them is prerequisite for trustworthy autonomous discovery.

Conclusion

This review has synthesized literature on UQ for MLIPs through a five-family taxonomy, detailed mathematical formulations, applications to GAP/DeepMD/NequIP architectures, and critical examination of seven hidden assumptions. Empirical practices remain inadequate, and seven unresolved questions highlight persistent gaps in correlation handling, ground-truth definition, efficiency, extrapolation, force-vector treatment, physical consistency, and multi-source fusion.

The outlook is cautiously optimistic. Integration of UQ into active-learning loops, development of force-aware and distribution-shift-robust methods, and creation of community benchmarks will accelerate the transition from proof-of-concept MLIPs to production-grade tools for materials engineering. Future research must prioritize rigorous calibration under realistic conditions, hybrid Bayesian–conformal frameworks, and physically informed priors to realize the full promise of uncertainty-aware computational discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Sep 2021

Revised: 16 Jan 2022

Accepted: 28 May 2022

Published online: 18 July 2022

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

- Mueller T, Hernandez A, Wang C. Machine learning for interatomic potential models. *J Chem Phys.* 2020;152(5):050902.
<https://doi.org/10.1063/1.5126336>.
- Peterson AA, Christensen R, Khorshidi A. Addressing uncertainty in atomistic machine learning. *Phys Chem Chem Phys.* 2017;19(18):10978-85.
<https://doi.org/10.1039/C7CP00375G>.
- Chan H, Narayanan B, Cherukara MJ, Sen FG, Sasikumar K, Gray SK, et al. Machine learning classical interatomic potentials for molecular dynamics from first-principles training data. *J Phys Chem C.* 2019;123(12):6941-57.
<https://doi.org/10.1021/acs.jpcc.8b09917>.
- Wan S, Sinclair RC, Coveney PV. Uncertainty quantification in classical molecular dynamics. *Philos Trans A Math Phys Eng Sci.* 2021;379(2197):20200082.
<https://doi.org/10.1098/rsta.2020.0082>.
- 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>.
- 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>.
- 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>.
- Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun.* 2022;13(1):2453.
<https://doi.org/10.1038/s41467-022-29939-5>.
- Li Y, Xiao W, Wang P. Uncertainty quantification of artificial neural network based machine learning potentials. In: ASME 2018 international mechanical engineering congress and exposition; 2018; Pittsburgh, PA. New York: American Society of Mechanical Engineers; 2018. p. V012T11A030.
<https://doi.org/10.1115/IMECE2018-88071>.

Hu Y, Musielewicz J, Ulissi ZW, Medford AJ. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Mach Learn Sci Technol.* 2022;3(4):045028.
<https://doi.org/10.1088/2632-2153/aca7b1>.

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>.

Venturi S, Jaffe RL, Panesi M. Bayesian machine learning approach to the quantification of uncertainties on ab initio potential energy surfaces. *J Phys Chem A.* 2020;124(25):5129-46.
<https://doi.org/10.1021/acs.jpca.0c02395>.

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>.

Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: *Advances in neural information processing systems 30*. Red Hook (NY): Curran Associates; 2017. p. 6402-13.

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

Smith JS, Isayev O, Roitberg AE. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci.* 2017;8(4):3192-203.
<https://doi.org/10.1039/C6SC05720A>.

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>.

Rao Z, Tung PY, Xie R, Wei Y, Zhang H, Ferrari A, et al. Machine learning-enabled high-entropy alloy discovery. *Science.* 2022;378(6615):78-85.
<https://doi.org/10.1126/science.abo4940>.

Novikov IS, Gubaev K, Podryabinkin EV, Shapeev AV. The MLIP package: Moment tensor potentials with MPI and active learning. *Mach Learn Sci Technol.* 2021;2(2):025002.
<https://doi.org/10.1088/2632-2153/abc9fe>.

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>.

Bartók AP, Kermode JR. Improved uncertainty quantification for Gaussian process regression based interatomic potentials. *arXiv [Preprint].* 2022.
<https://doi.org/10.48550/arXiv.2206.08744>.

Köfinger J, Hummer G. Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur Phys J B.* 2021;94(12):245.
<https://doi.org/10.1140/epjb/s10051-021-00234-4>.

Nigam A, Pollice R, Hurley MFD, Hickman RJ, Aldeghi M, Yoshikawa N, et al. Assigning confidence to molecular property prediction. *Expert Opin Drug Discov.* 2021;16(9):1009-23.
<https://doi.org/10.1080/17460441.2021.1925247>.