PAPER • OPEN ACCESS

Model accuracy and data heterogeneity shape uncertainty quantification in machine learning interatomic potentials

IOP Publishing Mach. Learn.: Sci. Technol. 7 (2026) 025002 https://doi.org/10.1088/2632-2153/ae3d80

MACHINE LEARNING
Science and Technology

PAPER

Model accuracy and data heterogeneity shape uncertainty quantification in machine learning interatomic potentials

Fei Shuang[1,*], **Zixiong Wei**[1], **Kai Liu**[1], **Wei Gao**[2,3] **and Poulumi Dey**[1,*]

[1] Department of Materials Science and Engineering, Faculty of Mechanical Engineering, Delft University of Technology, Mekelweg 2, Delft 2628CD, The Netherlands
[2] J Mike Walker'66 Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, United States of America
[3] Department of Materials Science & Engineering, Texas A&M University, College Station, TX 77843, United States of America
[*] Authors to whom any correspondence should be addressed.

E-mail: shuangfei1991@gmail.com and P.Dey@tudelft.nl

Keywords: machine learning interatomic potentials, uncertainty quantification, D-optimality, ensemble learning, atomic cluster expansion

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
12 September 2025

REVISED
5 January 2026

ACCEPTED FOR PUBLICATION
26 January 2026

PUBLISHED
13 February 2026

Abstract

Machine learning interatomic potentials (MLIPs) enable accurate atomistic modeling, but reliable uncertainty quantification (UQ) remains elusive. In this study, we investigate two UQ strategies, ensemble learning and D-optimality, within the atomic cluster expansion framework. It is revealed that higher model accuracy strengthens the correlation between predicted uncertainties and actual errors and improves novelty detection, with D-optimality yielding more conservative estimates. Both methods deliver well calibrated uncertainties on homogeneous training sets, yet they underpredict errors and exhibit reduced novelty sensitivity on heterogeneous datasets. To address this limitation, we introduce clustering enhanced local D-optimality, which partitions configuration space into clusters during training and applies D-optimality within each cluster. This approach substantially improves the detection of novel atomic environments in heterogeneous datasets. Our findings clarify the roles of model fidelity and data heterogeneity in UQ performance and provide a practical route to robust active learning and adaptive sampling strategies for MLIP development.

1. Introduction

Machine learning interatomic potentials (MLIPs) have reshaped computational materials science by bridging the accuracy of quantum-mechanical methods with the scale of classical molecular dynamics (MD) [1, 2]. By learning the mapping from local atomic environments (LAEs) to potential energy surfaces using first-principles data, MLIPs routinely approach near-quantum fidelity at a fraction of the computational cost [3, 4]. This advance has enabled simulations of complex, previously inaccessible phenomena, from phase transformations and defect kinetics to catalyst discovery and non-equilibrium transport, at time and length scales far beyond *ab initio* MD (AIMD) [5–7].

Unlike traditional, physically motivated functional forms such as the embedded-atom model, MLIPs are constrained by their training distributions. When presented with out-of-distribution (OOD) atomic environments, they may yield unreliable or unphysical predictions, limiting transferability in practical workflows. This challenge has motivated a rich set of uncertainty quantification (UQ) strategies to assess reliability of energies and forces. Among these, D-optimality and ensemble-based methods have been particularly influential owing to their practical implementation across multiple frameworks. The D-optimality criterion, implemented in moment tensor potentials (MTPs) [8–10], the atomic cluster expansion (ACE) [11–13], and neuroevolution potentials (NEPs) [14], identifies informative configurations via their contribution to feature-space volume (e.g. extrapolation grade). In parallel, ensemble approaches estimate epistemic uncertainty by measuring the spread of predictions across models trained with different initializations, data bootstraps, or hyperparameters.

Beyond their role in diagnosing reliability, UQ methods have become central to data generation via active learning. In UQ-guided loops, candidate configurations discovered during exploration are selectively labeled and appended to the training set, yielding automated, recursive improvement in both accuracy and robustness. This paradigm has matured into a standard practice for MLIP development: it reduces the size (and cost) of reference datasets while enhancing stability under demanding conditions. In applications, D-optimality-based selection within MTPs is a mainstay for metals and alloys [15–19], whereas ensemble-force criteria are particularly effective in complex, heterogeneous systems such as silicon–oxygen networks [7]. Recent hyperactive learning strategies further accelerate sampling by biasing dynamics toward uncertain regions, efficiently generating information-rich configurations for linear ACE potentials [20]. Collectively, these developments underscore the pivotal role of UQ in both the application and advancement of MLIPs [21, 22].

Despite this progress, key questions remain regarding calibration and transferability of UQ metrics. Notably, Lysogorskiy *et al* reported within the ACE framework that D-optimality and ensemble indicators offer broadly comparable reliability [23]. Two issues are particularly pressing. First, how does the baseline predictive accuracy of a fitted MLIP influence the fidelity of its uncertainty estimates? Second, how does increasing dataset heterogeneity (e.g. mixing simple elastic deformations with defect-rich clusters, surface reconstructions, liquid-like motifs, and high-strain-rate configurations) affect the calibration and sensitivity of UQ measures? These questions are especially relevant for on-the-fly active learning, wherein the training set evolves to include progressively more diverse atomic environments, potentially improving coverage while challenging model generalization.

In this work, we systematically evaluate ensemble-based and D-optimality UQ within the ACE framework. We quantify how model accuracy and dataset heterogeneity together govern (i) the alignment between predicted uncertainties and realized errors and (ii) each method's capability to flag novel configurations and LAEs. Building on these insights, we introduce a *clustering-enhanced local D-optimality* criterion: configuration space is partitioned into clusters of similar atomic motifs, and extrapolation grades are computed within each cluster rather than globally. This strategy improves calibration, tracks true errors more faithfully, and more reliably detects OOD LAEs in large-scale deformation simulations. The resulting protocol maintains the computational efficiency of ACE models while providing uncertainty estimates that are both sensitive and robust across heterogeneous datasets.

2. Methods

2.1. MLIPs

In this study, we use the ACE framework for UQ analysis for three main reasons. First, ACE [11–13, 23, 24] provides a general, mathematically complete formalism [25] that can be extended to other descriptors such as spectral neighbor analysis potential [26] and MTP [8–10]. Second, it strikes an optimal balance between accuracy and computational efficiency [12]. Third, ACE's built-in support for extrapolation-grade evaluation in both ASE [27] and LAMMPS [28] makes it straightforward to apply from small clusters up to million-atom configurations. To keep our analysis focused, we consider only linear ACE models. We explore six models of increasing complexity, ranging from 15 to 945 functions, covering a corresponding span of training accuracies. Throughout fitting, we fix the force-weighting parameter κ at 0.01 and cap the number of training steps at 2000. The *pacemaker* package manages ACE training [12, 13].

2.2. UQ

2.2.1. Ensemble learning

Following the [23], we compute the maximum deviations of configurational energies and atomic forces, which serve as quantitative measures of uncertainty for each atom ($U_{F,\mathrm{atom}}$) and the whole configuration ($U_{E,\mathrm{cfg}}$ and $U_{F,\mathrm{cfg}}$), respectively, formulated as:

$$U_{E,\mathrm{cfg}} = \max_k |E_j^k - \langle E_j \rangle| , \tag{1}$$

$$U_{F,\mathrm{atom}} = \max_k |\mathbf{F}_i^k - \langle \mathbf{F}_i \rangle| , \tag{2}$$

$$U_{F,\mathrm{cfg}} = \max_{i \in j} \left(\max_k |\mathbf{F}_i^k - \langle \mathbf{F}_i \rangle| \right) , \tag{3}$$

where $k = 1, \ldots, K$ are the indices of the ACE models in the ensemble, E_j^k is the energy predicted by model k for the configuration j, and $\langle E_j \rangle$ is the ensemble average of the energy for the corresponding

configuration. The force on atom i in ensemble model k is given by $\mathbf{F}_i^k$, while $\langle \mathbf{F}_i \rangle$ is the ensemble force average. Then, we compare the uncertainties $U_{F,\mathrm{atom}}$, $U_{F,\mathrm{cfg}}$ and $U_{E,\mathrm{cfg}}$, to their respective ground-truth errors, defined as:

$$e_{E,\mathrm{cfg}} = |E_j^{\mathrm{DFT}} - \langle E_j \rangle|\,, \tag{4}$$

$$e_{F,\mathrm{atom}} = |\mathbf{F}_i^{\mathrm{DFT}} - \langle \mathbf{F}_i \rangle|\,, \tag{5}$$

$$e_{F,\mathrm{cfg}} = \max_{i \in j} |\mathbf{F}_i^{\mathrm{DFT}} - \langle \mathbf{F}_i \rangle|. \tag{6}$$

In an active learning loop, new configurations or LAEs are selected when their uncertainties in predicted energy ($U_{E,\mathrm{cfg}}$) or force ($U_{F,\mathrm{cfg}}$ or $U_{F,\mathrm{atom}}$) exceed specified thresholds ε_E or ε_F. Previous studies have typically relied on a force-based criterion, but the choice of ε_F varies widely: hyperactive learning with linear ACE models often uses 0.2–0.4 eV Å^{-1} [7], whereas active learning for MTPs in silicon-oxygen systems employs 1–2 eV Å^{-1} [20]. Lysogorskiy *et al* proposes a consistent threshold for both energy and force:

$$\varepsilon = Q_3 + 1.5 \times \mathrm{IQR}\,, \tag{7}$$

where Q_3 is the third quartile of the training-error distribution of configurational energies or atomic forces and IQR = Q_3–Q_1 is its interquartile range [23]. In this work, we adopt equation (7) because it automatically adapts to each ACE model's specific training-error characteristics.

Using equation (1), we establish the configuration-based energy criterion (CBE), noting that an atom-level energy criterion is physically meaningless since energy cannot be properly partitioned at the atomic scale [23]. Similarly, we derive two force-based uncertainty metrics from equations (2) and (3): the atom-based force criterion (ABF) and the configuration-based force criterion (CBF). For small systems, we employ both CBE and CBF to detect novel configurations, while ABF serves as the primary metric for identifying new LAEs in large-scale simulations. A configuration is identified as novel if its energy uncertainty exceeds the threshold ($U_{E,\mathrm{cfg}} > \varepsilon_E$) or its force uncertainty surpasses the critical value ($U_{F,\mathrm{cfg}} > \varepsilon_{F,\mathrm{cfg}}$), while an atom is flagged as new when its local force uncertainty exceeds the threshold ($U_{F,\mathrm{atom}} > \varepsilon_{F,\mathrm{atom}}$).

2.2.2. D-optimality

The *pacemaker* package is used to construct the D-optimal active set and to evaluate extrapolation grades γ for our linear-in-parameters ACE models [12, 13, 23]. Following [23], we consider, for each chemical species μ, a reference dataset with N_μ atomic environments. For every environment i we form a vector of n_v ACE basis functions

$$\mathbf{B}_i = (B_{i1}, B_{i2}, \ldots, B_{in_v})\,, \tag{8}$$

and collect all such rows into the matrix

$$\hat{\mathbf{B}}_\mu = \begin{bmatrix} B_{11} & \cdots & B_{1n_v} \\ \vdots & \ddots & \vdots \\ B_{N_\mu 1} & \cdots & B_{N_\mu n_v} \end{bmatrix}, \tag{9}$$

where typically $N_\mu \gg n_v$, i.e. many more environments than basis functions are available. Using the MaxVol algorithm [23, 29], *pacemaker* selects n_v rows whose $n_v \times n_v$ submatrix has (approximately) maximal determinant; these rows define the active-set matrix $\hat{\mathbf{A}}_\mu$.

Any environment from the reference set can then be written as a linear combination of the active environments,

$$\mathbf{B}_i = \sum_{k=1}^{n_v} \gamma_k^{(i)} \mathbf{B}_k^A, \tag{10}$$

which can be expressed compactly as

$$\boldsymbol{\gamma}^{(i)} = \mathbf{B}_i \hat{\mathbf{A}}_\mu^{-1}. \tag{11}$$

The atomic D-optimality extrapolation grade is defined as the largest absolute coefficient [23]

$$\gamma_i = \max_k \left|\gamma_k^{(i)}\right| = \max \left|\mathbf{B}_i \hat{\mathbf{A}}_\mu^{-1}\right|. \tag{12}$$

By construction, environments that belong to the active set satisfy $\gamma_{\text{atom}} \leqslant 1$, whereas $\gamma_{\text{atom}} > 1$ indicates extrapolation with respect to the current active set. Moreover, multiplying any basis function by a constant for all environments rescales both $\mathbf{B}_i$ and $\hat{\mathbf{A}}_\mu$ in the same way, so that γ_{atom} is scale-invariant, as discussed in [23].

In our implementation, *pacemaker* additionally performs a simple force-based outlier filtering when constructing $\hat{\mathbf{B}}_\mu$: atoms whose force magnitudes exceed

$$\varepsilon = Q_3 + 1.5 \times \text{IQR}, \tag{13}$$

with Q_3 the third quartile and IQR the interquartile range of the force distribution, are discarded from the active-set search. By contrast, the standard MTP workflow retains every atom when computing γ [9, 10]. This preliminary filtering reduces the impact of a few extreme configurations on the D-optimality matrix and makes the extrapolation grades more stable (see section 3.4).

Within standard active-learning frameworks, two complementary extrapolation grades are typically used: the atomic grade γ_{atom} defined above and a configuration-level grade

$$\gamma_{\text{cfg}} = \max_{i \in \text{cfg}} \gamma_{\text{atom}}(i), \tag{14}$$

which aggregates the most extrapolative atom in each configuration. Configurations or atoms with $\gamma > 1$ are treated as extrapolative, and the MaxVol algorithm is then used to select determinant-optimizing environments from these candidates for subsequent DFT calculations and active-set updates [23]. Although our present work focuses specifically on UQ methodology, this canonical active learning procedure provides important context for evaluating the performance of detection metrics and their implications for active learning efficiency.

2.3. Simulation and visualization

We utilize the Vienna *ab initio* simulation package (VASP) to perform first-principles calculations of all new configurations [30]. A gradient-corrected functional in the Perdew–Burke–Ernzerhof form is used to describe the exchange and correlation interactions [31]. Electron-ion interactions are treated within the projector-augmented-wave (PAW) method, using the standard PAW pseudopotentials provided by VASP [32]. The energy convergence criterion is set to 10^{-6} eV for electronic self-consistency calculations. The plane-wave cutoff energy is chosen to be 520 eV. The KPOINTS are generated by VASPKIT [33], based on the Monkhorst–Pack scheme [34], with a consistent density of $2\pi \times 0.03$ Å^{-1}. Additionally, LAMMPS is used for force calculations and atomic extrapolation grade (γ_{atom}) for million-atom configurations [28]. OVITO is employed for the visualization of the atomic structures [35].

3. Results

3.1. Dataset preparation and analysis

We employ the body-centered cubic tungsten (BCC W) dataset from our recent work [36] for UQ. Figure 1 displays all the configurations of the six subsets (A to F) using the first two principal components of the MACE descriptor [37], with each subset annotated by its representative configuration. Details of these subsets are summarized below:

A Unit cells undergoing elastic deformation (two atoms per cell).
B AIMD snapshots and simple defects, including vacancies, dislocations, grain boundaries (GBs), and surfaces.
C Atomic clusters extracted from complex defects in large-scale MD simulations, with periodic boundary conditions reconstructed using an empirical interatomic potential-guided grand-canonical Monte Carlo (EIP-GCMC) method. Methodological details are provided in [36].
D Spherical BCC clusters embedded in vacuum within a periodic box, introducing a large fraction of free surfaces.
E Atomic clusters cut from complex defects using the MLIP-3 package [10].
F A comprehensive validation set from our previous study [36], spanning diverse defect and deformation scenarios, including GBs with random perturbations, GBs under severe compression, two- and three-dimensional random GBs, and crack tip originally from [38].

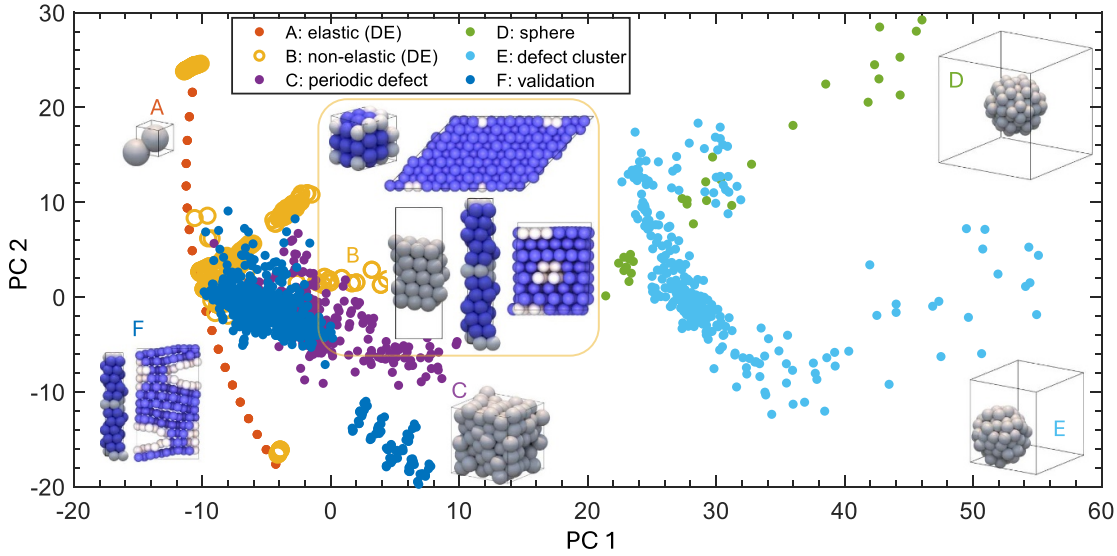

Figure 1. Six subsets of W configurations used in this study. Each point represents a configuration projected onto the first two principal components of its MACE descriptor. (A) Elastic deformations; (B) supercell configurations from AIMD simulations and defective structures (grain boundaries, vacancies, and dislocations); (C) defect-related structures with reconstructed periodic boundary conditions; (D) BCC spherical clusters with free surfaces; (E) atomic clusters from defective regions; (F) test set configurations. Representative atomic structures for each subset are shown in the insets. The union of datasets A and B forms the domain expertise (DE) set, with B termed the non-elastic subset. Datasets C, E, and F are from [36].

Subsets A and B together form the typical foundation for initial MLIP training through domain expertise (DE). This progression, which starts from simple elastic strains in A, moves through increasingly complex defect structures and surfaces in B to E, and culminates in the broad validation collection in F, enables systematic assessment of MLIP performance and UQ behavior across increasing configurational complexity.

In the following sections, we perform UQ analysis using two dataset combinations. The first employs A + B for training and C + E + F for testing, representing a typical scenario where MLIPs predict atomic environments for unseen defects from standard DE datasets. The second, more challenging combination uses A + D for training and B + C + E + F for testing, where elastic deformations (A) and free surfaces (D) create highly heterogeneous features. In this case, all test configurations become OOD relative to the training set. Our results demonstrate that while both ensemble learning and D-optimality provide satisfactory UQ performance in the first scenario, they struggle with the increased complexity of the second case.

3.2. Ensemble learning method

In this section, we employ the maximum deviation of ACE predictions to quantify uncertainty via the ensemble learning method, following [23] and as detailed in 2. At the configurational level, we consider the CBE and CBF criteria, quantified by $U_{E,\mathrm{cfg}}$ and $U_{F,\mathrm{cfg}}$, respectively. At the atomic level, we adopt the ABF criterion, denoted by $U_{F,\mathrm{atom}}$. CBE and CBF facilitate active learning or sampling of entire configurations, whereas ABF is tailored to select LAEs in large scale simulations. We then compute the corresponding errors $e_{E,\mathrm{cfg}}$, $e_{F,\mathrm{cfg}}$, and $e_{F,\mathrm{atom}}$ (defined in 2) and examine the correlation between each uncertainty metric and its error. We consider A + B as the training set with C + E + F for testing. A six member ensemble is employed to quantify predictive uncertainty.

To illustrate the impact of model accuracy on UQ, we first present two ACE models at opposite ends of the basis-set complexity: the compact Func 15, which uses just 15 basis functions, and the expansive Func 945, which employs 945 basis functions. We evaluate three UQ metrics: CBE (figures 2(a) and (d)), CBF (figures 2(b) and (e)), and ABF (figures 2(c) and (f)). The results in figure 2 reveal three key observations. First, CBE demonstrates weak error correlations for both models, with Func-945 showing only slight improvement. Second, both force-based metrics (CBF and ABF) achieve substantially stronger correlations, where CBF's superior performance stems from its integration of structural information across all atoms in a configuration. Third, while increased model complexity significantly reduces training-set errors and uncertainties, test-set performance remains relatively unaffected, as indicated by the dashed lines and arrows in figure 2. This persistent gap reflects the test data's OOD nature and the growing separation between training and test distributions as models become more accurate.

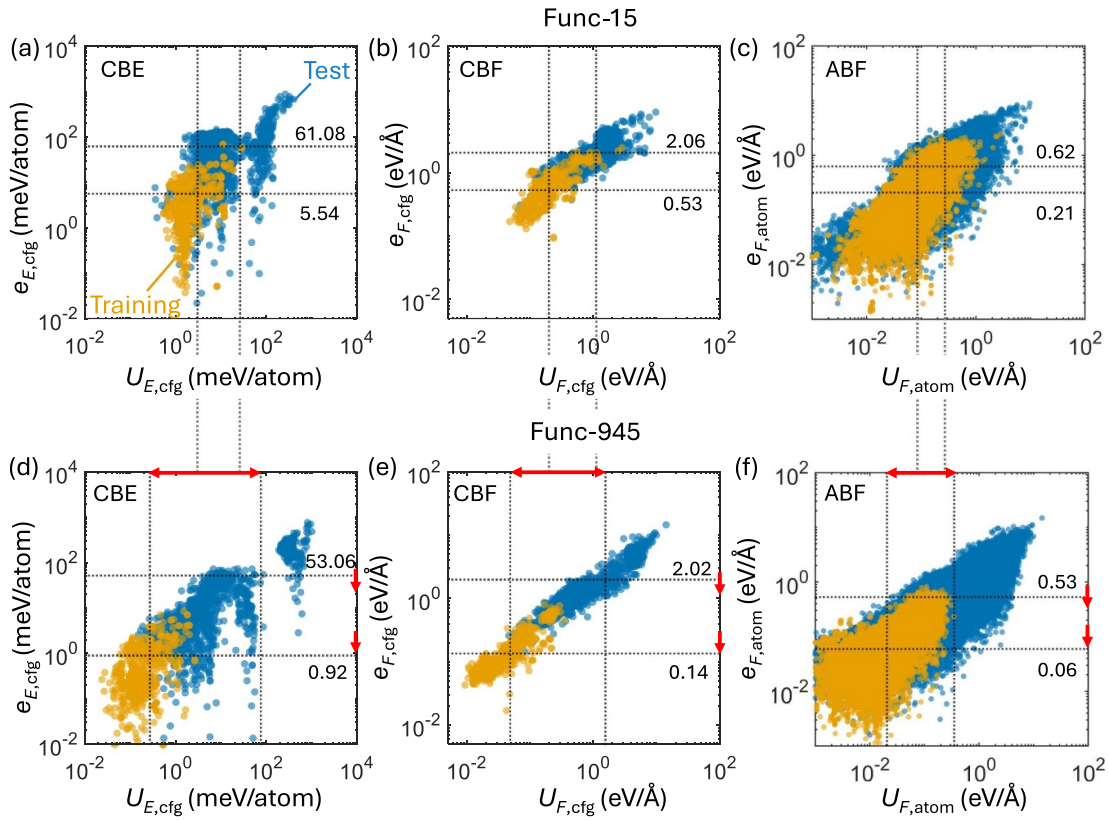

Figure 2. Correlation between model error and uncertainty evaluated using ensemble learning with ACE models. (a)–(c) Func-15 models (15 functions) with $F_{\mathrm{RMSE}} = 171.62$ meV Å^{-1} and $E_{\mathrm{RMSE}} = 9.05$ meV/atom. (d)–(f) Func-945 models (945 functions) with $F_{\mathrm{RMSE}} = 51.37$ meV Å^{-1} and $E_{\mathrm{RMSE}} = 1.46$ meV/atom. Columns represent different uncertainty criteria: (a), (d) configuration-based energy (CBE), (b), (e) configuration-based force (CBF), and (c), (f) atom-based force (ABF). Dashed lines show mean error and uncertainty values. Arrows indicate systematic performance shifts between Func-15 and Func-945 ensembles.

To systematically evaluate the impact of model accuracy, we compute Spearman's rank correlation coefficient (ρ), a nonparametric measure of how closely the ordering of predicted uncertainties matches the ordering of observed errors, across models with progressively lower force root mean square error (F_{RMSE}). Figure 3(a) demonstrates that for the CBE criterion, correlation strength increases monotonically with increase in model accuracy for both training and test datasets, showing particularly dramatic increase in test data. The CBF criterion (figure 3(b), solid line) shows analogous accuracy dependence while achieving substantially stronger correlations than CBE. Notably, the ABF criterion (dashed line) reveals divergent behavior: test data correlations increase steadily with accuracy, training set correlations remain consistently low ($\rho < 0.7$) and show no systematic relationship with model accuracy. Three fundamental insights emerge from this analysis. First, force-based criteria (CBF and ABF) universally surpass the energy-based CBE in correlation strength. Second, CBF consistently outperforms ABF. Third, and most significantly, test data correlations not only benefit more from improved model accuracy than training data, but also maintain superior absolute correlation strength across all accuracy levels. These findings collectively establish that robust UQ requires both careful metric selection and ongoing model refinement, with force-based configuration-level analysis delivering optimal performance for practical applications involving defection of novel configurations or LAEs.

The primary goal of UQ is to detect unseen configurations and LAEs. We derive UQ thresholds for CBE ($\varepsilon_{E,\mathrm{cfg}}$), CBF ($\varepsilon_{F,\mathrm{cfg}}$), and ABF ($\varepsilon_{F,\mathrm{atom}}$) (see section 2) to flag OOD configurations and LAEs. Applying these thresholds to the combined C, E, and F test sets (figure 1), we identify OOD configurations using CBE (figure 3(c)) and CBF (figure 3(d)), and detect OOD LAEs using ABF (figure 3(e)) for both the Func-15 and Func-945 models. The Func-15 model selects very few new configurations or LAEs, classifying most test cases as ID despite high errors. In contrast, the more accurate Func-945 model flags a substantial fraction of new configurations and LAEs, due to the clearer separation between training and test data (figure 2). Figure 3(f) illustrates how the selection rate of each criterion scales with model accuracy, defined as the fraction of flagged configurations (relative to total test configurations) or LAEs (relative to total test-set atoms). Higher model accuracy consistently yields more flagged items.

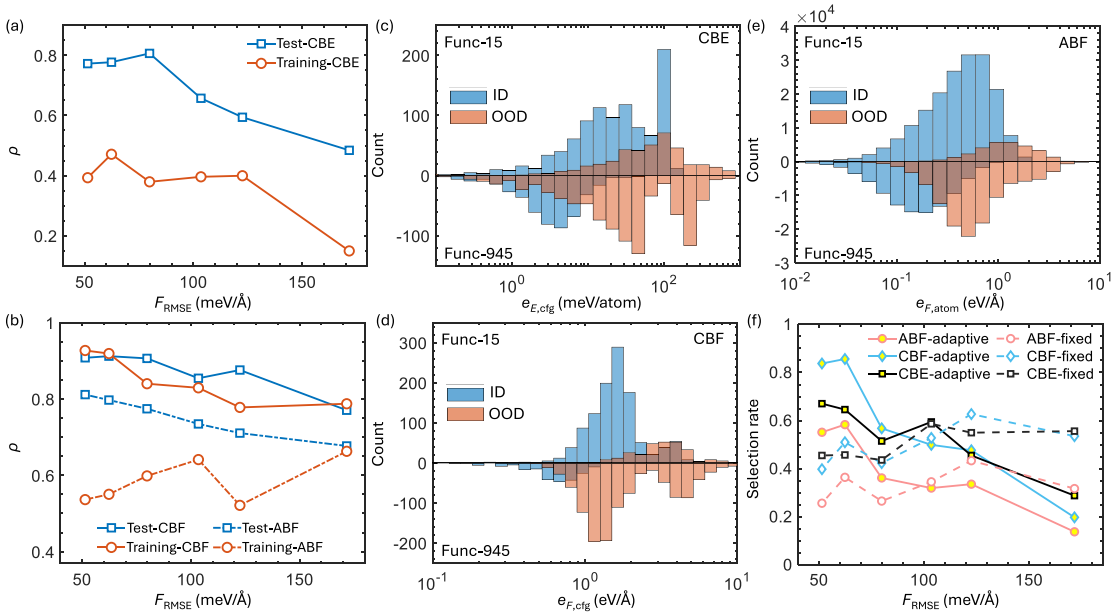

Figure 3. Uncertainty quantification performance in ensemble learning: role of model accuracy. (a), (b) Spearman's ρ (uncertainty-error correlation) vs force RMSE (F_{RMSE}) for training/test data across three criteria: configuration-based energy (CBE), force (CBF), and atomic force (ABF). (c)–(e) Comparison between Func-15 and Func-945 models for in-distribution (ID) and out-of-distribution (OOD) detection across CBE, CBF, and ABF criteria for the test data. (f) Selection rate vs F_{RMSE} using adaptive/fixed thresholds for all criteria.

Notably, at comparable accuracy levels, CBF outperforms CBE in detecting novel configurations, a trend particularly evident for the highest-fidelity ACE models.

A key remaining question concerns the relative performance of adaptive versus fixed thresholds for OOD detection. We assess this by applying the mean thresholds of our three criteria (CBE, CBF, and ABF) across different F_{RMSE} levels (figure S1) as fixed thresholds to evaluate selection rates. As shown by the dashed lines in figure 3(f), fixed thresholds exhibit selection rates with minimal dependence on F_{RMSE}. While both approaches demonstrate similar selection rates at $F_{RMSE} = 100\,\text{meV Å}^{-1}$, fixed thresholds identify more configurations/LAEs below this value and fewer above it. However, while fixed thresholds may select more configurations/LAEs at low F_{RMSE}, this does not necessarily indicate better OOD detection accuracy. These findings collectively demonstrate the superior reliability of adaptive thresholds for OOD detection.

We also evaluate how ensemble size affects the detection of novel configurations and LAEs. Using our most-accurate ACE model (Func-945) with ensemble sizes ranging from 3 to 30 models, figure S2(a) shows that force-based metrics (CBF and ABF) exhibit strong ensemble-size dependence, while CBE remains relatively stable. All three criteria achieve consistent selection rates only when the ensemble contains $\geqslant$10 models, which is twice the conventional five-model standard [7]. To understand this dependence, we compute the Spearman correlation ρ between prediction error and uncertainty for both training (A + B) and test (C + E + F) sets (figures S2(b), (c)). The fluctuating ρ values reveal no systematic trend with ensemble size, indicating Spearman's ρ alone cannot explain the detection trends. Analysis of prediction errors (figures S2(d)–(i)) shows larger ensembles simultaneously increase test-set errors while decreasing training-set errors. This growing train-test divergence enhances novel configuration/LAE detection, an effect distinct from model accuracy effects in figure 2. Moreover, larger ensembles provide two key advantages: (1) increased mean test-set uncertainty (figure S2(j)), and (2) reduced novelty-detection thresholds ε (figures S2(k)–(m)), except for CBE (figure S2(m)). These lower thresholds enable more OOD flagging, fully explaining the rising selection rates in figure S2(a).

3.3. D-optimality criterion and MaxVol algorithm

In our analysis of the D optimality criterion, we use the extrapolation grade γ computed via the MaxVol algorithm for UQ (see section 2). Analogous to the ensemble approach, we derive γ_{cfg} and γ_{atom} to assess the uncertainty of entire configurations and individual atoms, respectively.

We first consider A + B as the training set with C + E + F for testing. Figure 4 presents extrapolation grades at both configuration and atom level (γ_{cfg} and γ_{atom}), plotted against energy and force errors. The threshold $\gamma = 1$ (dashed line in the figure) separates ID ($\gamma < 1$) from OOD ($\gamma > 1$) regimes across

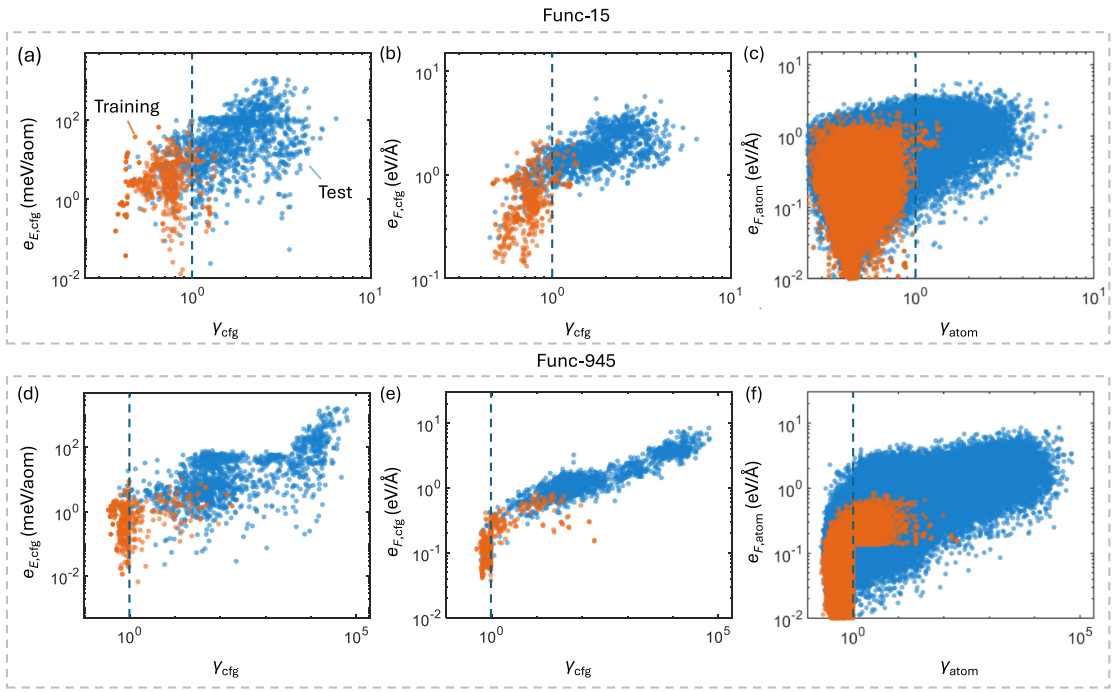

Figure 4. Prediction errors vs extrapolation grade (γ) in ACE models. Results are shown for two models: (a)–(c) Func-15 and (d)–(f) Func-945. (a), (d) Configurational energy errors. (b), (e) Configurational force errors. (c), (f) Atomic-level force errors. Vertical dashed lines mark the extrapolation threshold ($\gamma = 1$).

both models. Our D-optimality analysis reveals distinct patterns in UQ when comparing the Func-15 and Func-945 models. The more accurate Func-945 model (panels d–f) shows significantly stronger error-grade correlations than Func-15 (panels a–c), consistent with ensemble method results in figure 2. The range of γ values also differs by orders of magnitude: Func-15 yields grades around 10^2, whereas Func-945 reaches values near 10^6, highlighting how higher model accuracy improves discrimination among configurations and LAEs. For both models, configurational energy errors (figures 4(a) and (d)) and force errors (figures 4(b) and (e)) remain random below $\gamma_{\mathrm{cfg}} = 1$ but increase markedly once γ_{cfg} exceeds 1. Overall, these results confirm that D optimality effectively identifies OOD configurations and that γ_{cfg} correlates more strongly with configuration force errors than with energy errors, consistent with the ensemble learning trends shown in figure 3. At the atomic level (figures 4(c) and (f)), γ_{atom} identifies more OOD LAEs in the Func-945 case, yet the per-atom force errors show only a weak dependence on γ_{atom}. Notably, many atoms with $\gamma_{\mathrm{atom}} > 1$ exhibit very low errors, indicating potential extrapolation capability of the MLIP. These results collectively establish D-optimality as a robust method for configuration-level UQ, while revealing inherent limitations in atomic-level analysis.

We then compare OOD detection performance between ensemble learning and D-optimality approaches in figure 5. The solid lines in figure 5(a) demonstrate that D-optimality achieves consistently high configuration-level detection (>90%) across all model accuracies, while LAE detection improves from ∼5% to ∼70% with increasing accuracy. Compared to both 6-member and 30-member ensemble results, D-optimality shows superior configuration-level detection and comparable atomic-level performance, despite requiring only a single ACE model. This reveals D-optimality's dual advantages of more conservative detection and greater computational efficiency relative to ensemble methods.

The detailed comparison between ensemble learning and D-optimality is shown in figures 5(b)–(d), contrasting their ability to identify ID and OOD configurations/LAEs in the combined C + E + F test set using Func-945 potentials. D-optimality demonstrates superior detection performance, flagging over 99% of test configurations as OOD (upper panels in figures 5(b) and (c)). In contrast, ensemble methods miss significant fractions of high-error cases: the energy-based ensemble overlooks ∼33% and the force-based ensemble ∼16%, incorrectly labeling them as ID (lower panels). At the atomic level (figure 5(d)), D-optimality identifies 64% of atoms as OOD LAEs versus 55% for ensembles, demonstrating more comprehensive local environment sampling. However, both approaches exhibit characteristic limitations: they incorrectly classify high-error atomic sites (up to 1 eV Å^{-1}) as ID (demonstrating overconfidence) while flagging low-error sites (0.05 eV Å^{-1}) as OOD (showing underconfidence), as

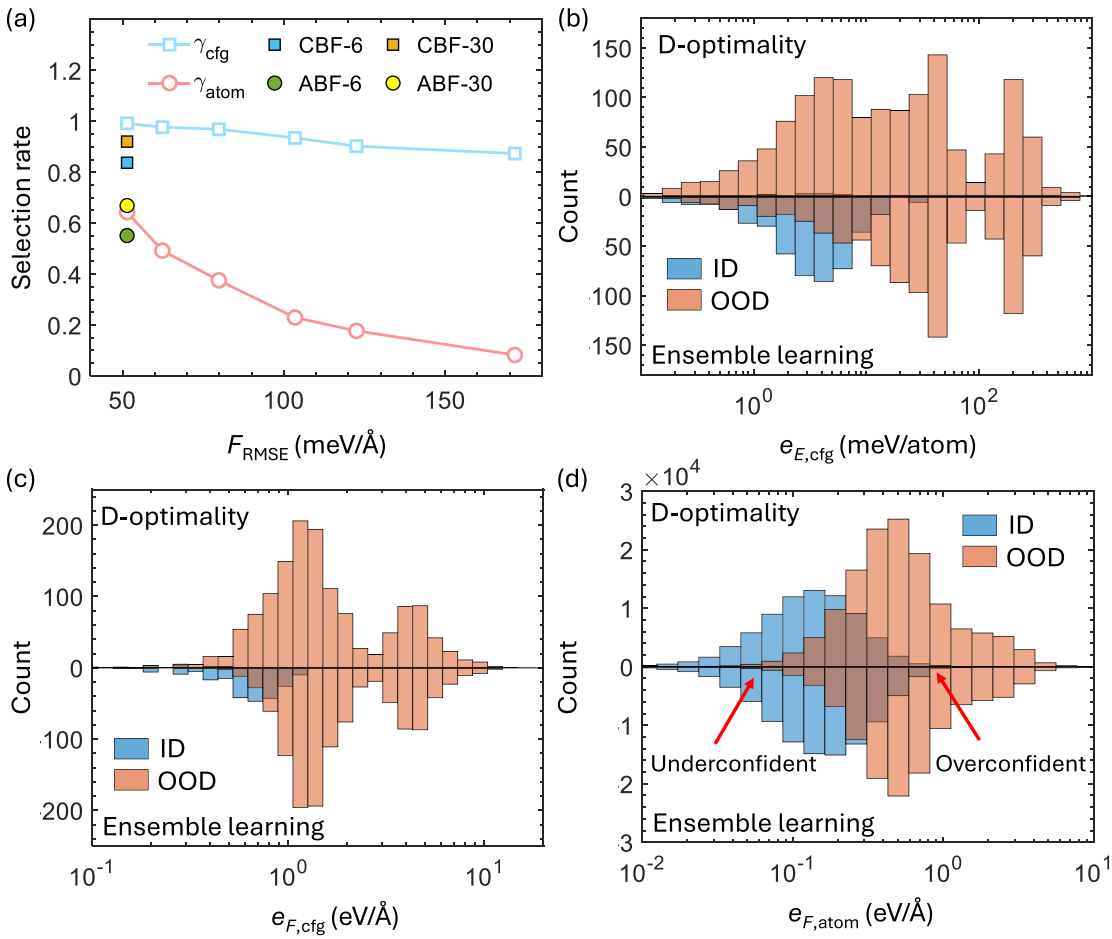

Figure 5. Comparison of ensemble learning and D-optimality methods. (a) OOD detection rate versus F_{RMSE}. (b)–(d) Classification performance for in-distribution (ID) and out-of-distribution (OOD) configurations/atoms. In panel (d), left arrow indicate underconfident predictions (low-error LAEs misidentified as OOD), right arrow show overconfident cases (high-error LAEs misclassified as ID).

highlighted by the arrows. This reflects the fundamental challenge of atomic-level active learning compared to whole-configuration sampling. Neither method achieves perfect discrimination—both systematically miss critical high-error sites while oversampling well-predicted regions, leading to inefficient computational resource allocation that undermines overall sampling efficiency.

3.4. Influence of data heterogeneity

To probe the limitations of ensemble learning and D optimality on structurally heterogeneous data, we devise a stringent scenario. The training set consists of 30 elastic deformation configurations (dataset A) and 30 nanospheres (dataset D), while datasets B, C, E, and F serve as the test set. This arrangement echoes the neighborhood mode of MLIP 3's active learning framework [10], in which vacuum embedded clusters are constructed so that novel LAEs occupy the cluster center. By applying both UQ methods in this context, we uncover their respective blind spots and derive practical lessons for optimizing active learning protocols to heterogeneous training sets.

All ensemble learning uncertainty calculations employ Func-945 models. Figure 6 reveals a fundamental paradox in ensemble-based UQ: despite strong force error-uncertainty correlations at both configurational (figure 6(a)) and atomic (figure 6(b)) levels, the method fails catastrophically for novelty detection. The detected OOD fractions (only 0.076% of configurations and 0.265% of atoms, corresponding to data points beyond the dashed uncertainty thresholds) represent complete failure, since the entire test set should be identified as OOD by design. This conclusion is unequivocal given that the training set contained just two structural motifs (elastically deformed bulk structures and BCC nanospheres), while the test set consists entirely of different defect-bearing configurations.

This critical failure originates from the training data's intrinsic heterogeneity. Figure 6(c) reveals that the training-set force errors exhibit bimodal distribution: one mode corresponds to easily predicted

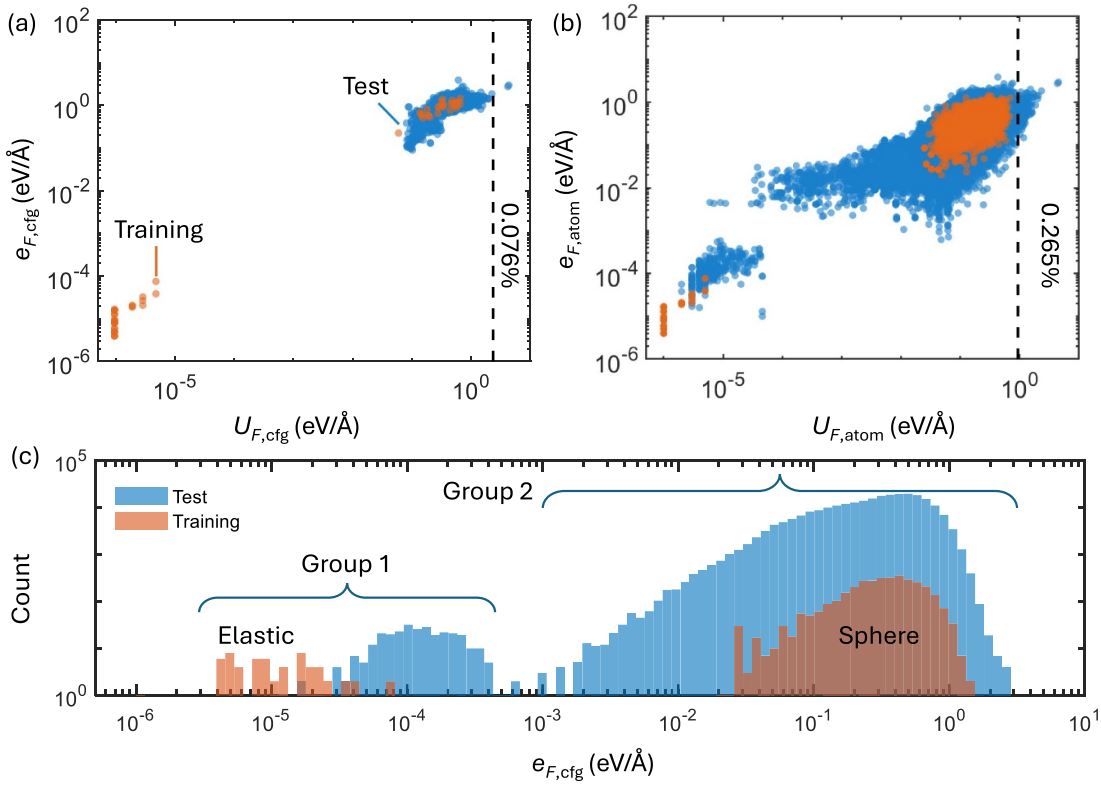

Figure 6. Limitations of ensemble learning for datasets containing subsets with heterogeneous complexity. (a) Force error vs uncertainty at the configurational level. The vertical dashed line marks the critical uncertainty threshold for identifying OOD configurations. Only 0.076% of test configurations are flagged as OOD. (b) Force error vs uncertainty at the atomic level, where just 0.265% of test atoms are classified as OOD. (c) Distribution of atomic force errors in training and test datasets, revealing two distinct subgroups with differing complexity in both datasets (elastic and sphere for the training set, and group 1 and 2 for the test set).

elastic-deformation configurations, while the other reflects the inherently more complex nanosphere surface environments. A single global uncertainty threshold, forced to accommodate both regimes, becomes dominated by the high-error nanosphere population and consequently sets an excessively high threshold for the elastic-deformation cases. The test set replicates this bimodal structure, with clusters centered near 10^{-4} eV Å^{-1} (Group 1) and 10^{-1} eV Å^{-1} (Group 2). As a result, the unified cutoff even fails to identify high-error Group 2 sites as OOD. This prevalence of false negatives in the high-error regime not only compromises UQ's reliability for active learning and adaptive sampling but also exposes the fundamental limitation of single-threshold methods when applied to multimodal error distributions.

Using Func 945 models, we compute D optimality extrapolation grades by training on 30 configurations each from datasets A and D and testing on the combined B + C + E + F set. Figure 7 compares force errors against these grades at both the configurational and atomic scales. At the configuration level in figures 7(a) and (D) optimality flags 75.4% of test structures as OOD, improving on the ensemble method (figure 6) yet still inadequate given that every test configuration is, by design, OOD. At the atomic level in figure 7(b), only 10% of local environments are detected as OOD. Compared with the ensemble results in figure 6, extrapolation grades show much better selection rates but weaker correlation with force errors. Moreover, the γ values span just 0.1–10, a dramatically narrower range than the 10^6 observed for the homogeneous A + B training set as shown in figure 4. These observations demonstrate that structural heterogeneity constrains both the magnitude and the predictive reliability of D optimality grades.

To further elucidate the limitations of D-optimality and MaxVol algorithm, we present a simplified two-dimensional example in figure 7(c) demonstrating the MaxVol algorithm's active set selection and extrapolation grade calculation, where three distinct non-overlapping subsets (a, b, c) with respective active sets ($\mathbf{v_1}$,$\mathbf{v_2}$), ($\mathbf{v_3}$, $\mathbf{v_4}$), and ($\mathbf{v_5}$, $\mathbf{v_6}$) combine to form a new active set ($\mathbf{v_1}$, $\mathbf{v_5}$). This analysis reveals critical inconsistencies in extrapolation grade determination: while point A appears ID ($\gamma_{15} = 0.76$) and point B OOD ($\gamma_{15} = 1.17$) in the combined dataset, examination of individual subsets shows the opposite behavior: point A consistently demonstrates OOD character ($\gamma_{12} = 6.36$, $\gamma_{34} = 2.41$, $\gamma_{56} = 2.34$) while point B is clearly ID ($\gamma_{34} = 0.88$) as it belongs to subset b. A comprehensive regional scan (figure 7(d))

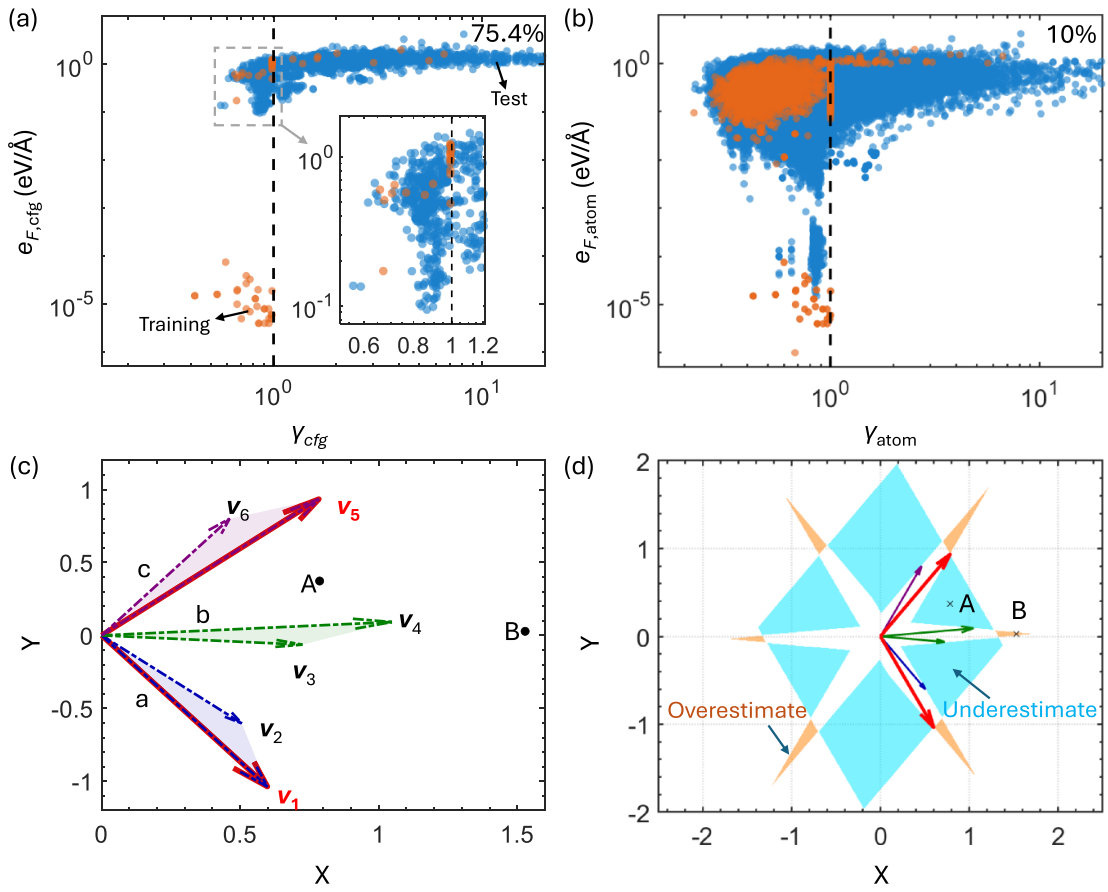

Figure 7. Limitations of D-optimality for datasets with subsets of heterogeneous complexity. (a) Force error vs extrapolation grade (γ_{cfg}) at the configurational level. The vertical dashed line marks the critical grade ($\gamma_{cfg} > 1$) for flagging OOD configurations. The inset zooms in on the data within the dashed rectangle. 75.4% of test configurations are labeled as OOD. (b) Force error vs extrapolation grade (γ_{atom}) at the atomic level, with only 10% of test atoms identified as OOD. (c) Three groups of vectors, each containing two vectors as an active set. Vectors $\mathbf{v}_1$ and $\mathbf{v}_5$ form the active set for the combined data. Test points A and B are highlighted. (d) Overestimated (false OOD) and underestimated (missed OOD) regions when calculating the extrapolation grade (γ) using the combined dataset.

further demonstrates that grade calculations based on the combined dataset overwhelmingly tend toward underestimation, with only rare cases of overestimation, as exemplified by points A and B respectively. These results highlight a core weakness of MaxVol: it targets only the extreme vertices of training dataset and ignores interior points. Novel data that lie within this hull receive low γ values, remain unselected, and leave large regions of configuration space unsampled, ultimately constraining the reach of D-optimality based active learning in MLIP development.

3.5. Improved D-optimality approach

To overcome the D-optimality limitations revealed in figure 7, we propose a clustering-enhanced local D-optimality approach that significantly improves UQ for structurally diverse datasets, as shown in figure 8. The key insight stems from recognizing that conventional single-grade calculations (γ_{cfg} or γ_{atom}) systematically underestimate novelty in heterogeneous dataset (figure 7), prompting our modified algorithm to instead compute subset-specific grades ($\gamma_{cfg,i}$ or $\gamma_{atom,i}$) and select their minimum as the final metric, a strategy that simultaneously prevents both underestimation by combined datasets and overestimation from individual subsets (as shown in figures 8(a) and (b)). This approach proves particularly useful for identifying transitional configurations between distinct structural regimes, as demonstrated by the point A in figures 7(c) and (d): where traditional methods would erroneously classify this boundary-spanning environment as ID, our minimum-grade criterion correctly flags it as OOD, thereby capturing crucial yet easily overlooked atomic environments that are essential for developing truly comprehensive MLIP.

To validate our clustering enhanced D optimality approach, we apply it to the W dataset sourced from [39]. This dataset comprises a diverse set of pre labeled subgroups, including distorted BCC unit

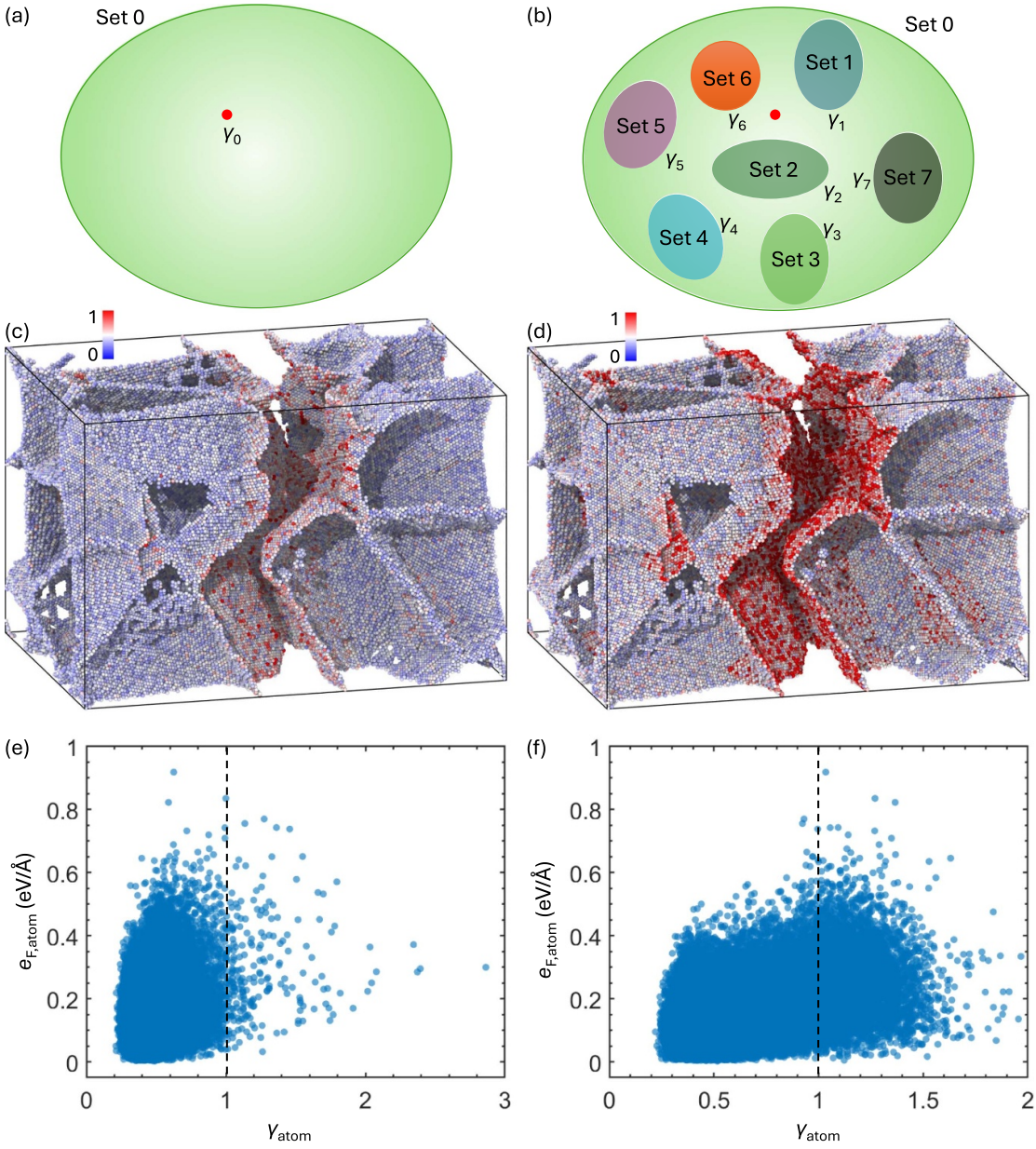

Figure 8. A new D-optimality approach for uncertainty quantification. Schematic illustrations compare (a) the original D-optimality with (b) our clustering-enhanced local D-optimality method. Atomistic configurations of a fractured tungsten (W) polycrystal are shown, with atomic colors indicating the extrapolation grade (γ_{atom}) computed using (c) the original D-optimality and (d) its improved variant. Scatter plots demonstrate the correlation between atomic force errors and extrapolation grades for (e) the original and (f) the refined approach.

cells, FCC and HCP crystals, high temperature BCC phases, vacancies, self interstitials, surface configurations, liquids, and others. Rather than using the original DFT energies and forces, we employ predictions from the universal NEP89 potential [40] to label all structures, thereby enabling the calculation of true errors for large scale configurations. For each pre labeled subgroup, we train a dedicated ACE model and assemble its active set. We then compute the extrapolation grade γ for every atom with respect to each active set and assign each atom the minimum γ value across all subgroup models as its final extrapolation grade. We test this procedure on a fractured polycrystal model from our recent work [36]. As shown in figure 8, the original D optimality method (figure 8(c)) flags only a few fracture surface atoms as OOD, despite leaving many high error ID atoms undetected (figure 8(e)). By contrast, our clustering enhanced version (figure 8(d)) correctly identifies a much larger set of fracture surface atoms as OOD, all with $\gamma > 1$. Crucially, figure 8(f) confirms that these newly detected atoms consistently exhibit higher force errors, demonstrating the superior reliability of our method for UQ.

4. Discussion

Our study reveals consistent principles and key distinctions between both UQ methods. For ensemble learning, we establish three critical findings. First, force-based criteria (CBF/ABF) show superior error-uncertainty correlations compared to energy-based metrics (CBE), with configuration-level analysis proving more reliable than atomic-level assessment. Second, model accuracy plays a crucial role in effective novelty detection. Third, robust detection requires larger ensembles of at least 10 models for stable performance. These principles also apply to D-optimality approaches, where configuration-level metrics similarly outperform atomic-level analysis in error correlation. However, a key difference emerges regarding accuracy dependence: atomic-level D-optimality detection shows strong sensitivity to model accuracy, while configuration-level performance remains largely accuracy-independent. Both methods exhibit qualitatively similar novelty identification behavior, with D-optimality offering a more conservative and computationally efficient alternative to ensemble learning. While increasing MLIP count can improve ensemble detection capability, this comes at substantial computational cost during both training and inference. We therefore recommend D-optimality as the preferred acquisition criterion. When unavailable (e.g. for universal MLIPs), ensemble methods must incorporate force-based analyses, high-fidelity models, and sufficiently large ensemble sizes (minimum 10 models) to ensure adequate performance.

Critically, our analysis reveals fundamental limitations in both ensemble and D-optimality UQ methods when handling heterogeneous training data. These approaches systematically fail to properly quantify uncertainty across multimodal distributions, leading to unreliable novelty detection. This failure stems from their inability to simultaneously accommodate diverse atomic environments. Yet this heterogeneity is unavoidable in practice. Proper MLIP training sets must encompass the complete spectrum of atomic environments found in real materials, including surfaces, interfaces, point defects, and bulk polymorphs across multiple space groups [41]. They must also incorporate extreme configurations like isolated atoms, dimers at varying separations, and collision geometries relevant to radiation-damage cascades [39]. The RANDSPG algorithm's material-agnostic approach, enumerating all 230 space groups with random primitive cells of 3–10 atoms [42], further demonstrates this inherent diversity. For high-entropy alloys, the challenge compounds as structural and chemical diversity interact in ways not yet fully understood. This unavoidable heterogeneity creates a fundamental tension: while current UQ methods work well for near-homogeneous data, they break down for the complex, multimodal distributions required for robust MLIP development. Our results expose this critical gap in the workflow of MLIP development, where inadequate UQ leads to persistent undersampling of precisely those atomic environments that are most informative yet most challenging to model.

Our findings have significant implications for on-the-fly active learning of LAEs in large-scale simulations, where atom-based UQ is required. In the standard MLIP-3 and *pacemaker* workflows, a spherical cluster around each candidate 'core' atom is extracted, enclosed in vacuum layers, and appended to the training set. However, this practice inadvertently incorporates surface atoms that are irrelevant to bulk-focused simulations. Because these extreme surface configurations substantially enlarge the envelope of active set in the MaxVol algorithm as illustrated in figures 7(a) and (b), the extrapolation grade underestimates the novelty of true bulk environments in the following active learning; genuinely new local structures are misclassified as ID simply because they are less exotic than the spurious surface atoms. Consequently, the original extrapolation-grade criterion renders on-the-fly active learning in MLIP-3 and *pacemaker* ineffective for generating truly local, bulk-specific MLIPs. A simple remedy is to construct the active set using only the core atoms, thereby excluding those with artificially truncated coordination. Alternatively, one can fill the vacuum region via EIP-GCMC and retain only the lowest-energy configurations [36]. Both strategies preserve structural relevance to the target simulation, prevent dilution of the uncertainty metric by spurious surfaces, and restore the extrapolation grade's sensitivity to genuinely novel local structures.

Yet the most reliable ensemble learning and D-optimality based UQ must be performed locally, gradually and independently for each candidate environment during on-the-fly active learning. Hodapp *et al* recently exemplified this approach by embedding an isolated screw dislocation in BCC metals or partial dislocation in FCC metals into a fully periodic supercell while excluding all atomic environments outside the defect core [43, 44]. By calculating the extrapolation grade solely within this narrowly defined region, their acquisition algorithm accurately identifies truly novel dislocation configurations and discards spurious outliers. The resulting MTP achieves remarkably low fitting errors and accurately reproduces the Peierls barrier, demonstrating that a defect-centered, locality-preserving sampling strategy is essential for reliable active learning. If the initial training set is heterogeneous, important environments will remain undersampled. A practical solution is to partition active learning by structural motif,

handling bulk phases, interfaces, and dislocations in separate acquisition loops in order to maintain extrapolation-grade accuracy and ensure comprehensive coverage of every relevant atomic environment.

The clustering-enhanced local D-optimality scheme proposed in this study reduces the impact of structural heterogeneity by evaluating uncertainty within clusters of geometrically similar environments. This partitioned analysis maintains the accuracy of MLIPs and supports their transferability across defect-rich configurational landscapes. The approach is particularly helpful when expanding an existing database that already contains several defect classes. Unsupervised algorithms such as *k*-means or BIRCH [45] can be used to divide the dataset into structurally coherent clusters before active learning or DIRECT sampling is applied. A comparable cluster-wise strategy could also be adopted for ensemble-based acquisition by assigning separate uncertainty thresholds to each subset; however, training many independent models would raise the computational cost substantially.

When using the original D-optimality method, it is important to note that the MaxVol algorithm focuses only on the most exotic atomic environments and therefore considers only the outer boundary of the dataset when constructing the active set. The major advantage is speed in the evaluation of the extrapolation grades, even for very large structures containing million atoms, but the drawback is reduced accuracy. In practice, the extrapolation grades calculated on heterogeneous datasets are often underestimated. Even with clustering-enhanced local D-optimality, capturing the fine details of every motif is difficult unless the acquisition step is carefully designed like Hodapp *et al* [43]. As a result, D-optimality-based active learning that starts from a global dataset tends to add very exotic structures or LAEs, which primarily guarantees the numerical stability of MD but may fall short of reproducing specific properties, such as dislocation migration or grain-boundary phase transitions, with DFT-level fidelity. A complementary route is provided by the QUESTS framework of Schwalbe-Koda *et al* [46], which measures the information-entropy increment that a candidate environment would contribute to a kernel-density estimate of the training distribution. Because this metric is model-free and depends only on geometric descriptors, it remains sensitive to rare motifs even in strongly heterogeneous datasets and can flag genuinely novel environments before any potential is fitted. Future studies could explore incorporating more expressive descriptors, such as SOAP [47] or the message passing MACE representation [37, 48, 49], to better handle multi-element systems.

5. Conclusions

In summary, we have advanced the theoretical foundations of UQ for MLIPs within the ACE framework and delivered practical improvements that reinforce active learning workflows. By integrating high fidelity base models with both configuration and atom resolved diagnostics, we enhance ensemble learning and D optimality's capacity to detect truly novel atomic environments. We further expose the key failure modes that arise in heterogeneous configuration spaces and introduce a clustering enhanced local D optimality criterion that restores reliable uncertainty estimates across diverse datasets. These developments are essential for robust adaptive sampling and active learning, underpinning the efficient and confident development of MLIPs.

Data availability statement

All code and DFT datasets are publicly available via GitHub (https://github.com/ufsf/UQ-ACE) [50] and Zenodo (https://zenodo.org/records/17899237). All simulations are executed using open-source software LAMMPS [28]. The machine learning force field was trained and validated by the *pacemaker* package [12, 13].

Supplementary data 1 available at https://doi.org/10.1088/2632-2153/ae3d80/data1.

Acknowledgments

This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO; the Netherlands Organization for Scientific Research), Domain Science, for access to supercomputing facilities. We also acknowledge the use of the DelftBlue supercomputer provided by the Delft High Performance Computing Center (DHPC; www.tudelft.nl/dhpc).

Author contributions

F. S.: Writing—original draft, Writing—review & editing, Validation, Methodology, Data curation, Conceptualization. Z. W: Writing—review & editing, Data curation and Analysis. K. L.: Writing—review

& editing, Data curation and Analysis. W. G.: Writing—review & editing. P. D.: Writing—review & editing, Supervision, Methodology, Funding acquisition, Conceptualization.

ORCID iDs

Fei Shuang 0000-0001-6333-9237
Zixiong Wei 0009-0000-3900-3995

References

[1] Friederich P, Häse F, Proppe J and Aspuru-Guzik A 2021 Machine-learned potentials for next-generation matter simulations *Nat. Mater.* **20** 750–61
[2] Ceriotti M 2022 Beyond potentials: integrated machine learning models for materials *MRS Bull.* **47** 1045–53
[3] Zuo Y *et al* 2020 Performance and cost assessment of machine learning interatomic potentials *J. Phys. Chem. A* **124** 731–45
[4] Jacobs R *et al* 2025 A practical guide to machine learning interatomic potentials - status and future *Curr. Opin. Solid State Mater. Sci.* **35** 101214
[5] Qamar M, Mrovec M, Lysogorskiy Y, Bochkarev A and Drautz R 2023 Atomic cluster expansion for quantum-accurate large-scale simulations of carbon *J. Chem. Theory Comput.* **19** 5151–67
[6] Liang Y, Mrovec M, Lysogorskiy Y, Vega-Paredes M, Scheu C and Drautz R 2023 Atomic cluster expansion for Pt-Rh catalysts: from *ab initio* to the simulation of nanoclusters in few steps *J. Mater. Res.* **38** 5125–35
[7] Erhard L C, Rohrer J, Albe K and Deringer V L 2024 Modelling atomic and nanoscale structure in the silicon-oxygen system through active machine learning *Nat. Commun.* **15** 1927
[8] Shapeev A V 2016 Moment tensor potentials: a class of systematically improvable interatomic potentials *Multiscale Model. Simul.* **14** 1153–73
[9] Novikov I S, Gubaev K, Podryabinkin E V and Shapeev A V 2021 The MLIP package: moment tensor potentials with MPI and active learning *Mach. Learn.: Sci. Technol.* **2** 025002
[10] Podryabinkin E, Garifullin K, Shapeev A and Novikov I 2023 MLIP-3: active learning on atomic environments with moment tensor potentials *J. Chem. Phys.* **159** 084112
[11] Drautz R 2019 Atomic cluster expansion for accurate and transferable interatomic potentials *Phys. Rev. B* **99** 014104
[12] Lysogorskiy Y *et al* 2021 Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon *npj Comput. Mater.* **7** 97
[13] Bochkarev A, Lysogorskiy Y, Menon S, Qamar M, Mrovec M and Drautz R 2022 Efficient parametrization of the atomic cluster expansion *Phys. Rev. Mater.* **6** 013804
[14] Xu K *et al* 2025 GPUMD 4.0: a high performance molecular dynamics package for versatile materials simulations with machine learned potentials *Mater. Genome Eng. Adv.* **3** e70028
[15] Carral A D, Xu X, Gravelle S, YazdanYar A, Schmauder S and Fyta M 2023 Stability of binary precipitates in Cu-Ni-Si-Cr alloys investigated through active learning *Mater. Chem. Phys.* **306** 128053
[16] Xu X, Zhang X, Bitzek E, Schmauder S and Grabowski B 2024 Origin of the yield stress anomaly in $L1_2$ intermetallics unveiled with physically informed machine-learning potentials *Acta Mater.* **281** 120423
[17] Rybin N, Novikov I S and Shapeev A 2025 Accelerating structure prediction of molecular crystals using actively trained moment tensor potential *Phys. Chem. Chem. Phys.* **27** 5141–8
[18] Klimanova O, Rybin N and Shapeev A 2025 Accelerating the global search of adsorbate molecule positions using machine-learning interatomic potentials with active learning *Phys. Chem. Chem. Phys.* **27** 9201–10
[19] Kotykhov A S, Hodapp M, Tantardini C, Kravtsov K, Kruglov I, Shapeev A V and Novikov I S 2025 Actively trained magnetic moment tensor potentials for mechanical, dynamical and thermal properties of paramagnetic CrN *Phys. Rev. B* **111** 094438
[20] van der Oord C, Sachs M, Kovács D P, Ortner C and Csányi G 2023 Hyperactive learning for data-driven interatomic potentials *npj Comput. Mater.* **9** 168
[21] Perez D, Subramanyam A P A, Maliyov I and Swinburne T D 2025 Uncertainty quantification for misspecified machine learned interatomic potentials *npj Comput. Mater.* **11** 263
[22] Bilbrey J A, Firoz J S, Lee M-S and Choudhury S 2025 Uncertainty quantification for neural network potential foundation models *npj Comput. Mater.* **11** 109
[23] Lysogorskiy Y, Bochkarev A, Mrovec M and Drautz R 2023 Active learning strategies for atomic cluster expansion models *Phys. Rev. Mater.* **7** 043801
[24] Drautz R 2020 Atomic cluster expansion of scalar, vectorial and tensorial properties including magnetism and charge transfer *Phys. Rev. B* **102** 024104
[25] Dusson G, Bachmayr M, Csányi G, Drautz R and Etter S 2022 van der Oord C, Ortner C, Atomic cluster expansion: completeness, efficiency and stability *J. Comput. Phys.* **454** 110946
[26] Thompson A, Swiler L, Trott C, Foiles S and Tucker G 2015 Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials *J. Comput. Phys.* **285** 316–30
[27] Larsen A H *et al* 2017 The atomic simulation environment-a python library for working with atoms *J. Phys.: Condens. Matter* **29** 273002
[28] Thompson A P *et al* 2022 LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso and continuum scales *Comput. Phys. Commun.* **271** 108171
[29] Podryabinkin E V and Shapeev A V 2017 Active learning of linearly parametrized interatomic potentials *Comput. Mater. Sci.* **140** 171–80
[30] Kresse G and Furthmüller J 1996 Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set *Phys. Rev. B* **54** 11169–86
[31] Perdew J P, Burke K and Ernzerhof M 1996 Generalized gradient approximation made simple *Phys. Rev. Lett.* **77** 3865–8
[32] Blöchl P E 1994 Projector augmented-wave method *Phys. Rev. B* **50** 17953–79
[33] Wang V, Xu N, Liu J-C, Tang G and Geng W-T 2021 VASPKIT: a user-friendly interface facilitating high-throughput computing and analysis using VASP code *Comput. Phys. Commun.* **267** 108033

[34] Monkhorst H J and Pack J D 1976 Special points for brillouin-zone integrations *Phys. Rev. B* **13** 5188–92
[35] Stukowski A 2010 Visualization and analysis of atomistic simulation data with OVITO-the open visualization tool *Modelling Simul. Mater. Sci. Eng.* **18** 015012
[36] Shuang F, Liu K, Ji Y, Gao W, Laurenti L and Dey P 2025 Modeling extensive defects in metals through classical potential-guided sampling and automated configuration reconstruction *npj Comput. Mater.* **11** 118
[37] Batatia I, Kovacs D P, Simm G N C, Ortner C and Csanyi G 2022 MACE: higher order equivariant message passing neural networks for fast and accurate force fields *Advances in Neural Information Processing Systems* ed A H Oh, A Agarwal, D Belgrave and K Cho (available at: https://openreview.net/forum?id=YPpSngE-ZU)
[38] Zhang L, Csányi G, van der Giessen E and Maresca F 2023 Atomistic fracture in bcc iron revealed by active learning of Gaussian approximation potential *npj Comput. Mater.* **9** 217
[39] Byggmästar J, Nordlund K and Djurabekova F 2020 Gaussian approximation potentials for body-centered-cubic transition metals *Phys. Rev. Mater.* **4** 093802
[40] Liang T *et al* 2025 NEP89: universal neuroevolution potential for inorganic and organic materials across 89 elements (arXiv:2504.21286)
[41] Poul M, Huber L and Neugebauer J 2025 Automated generation of structure datasets for machine learning potentials and alloys *npj Comput. Mater.* **11** 174
[42] Poul M, Huber L, Bitzek E and Neugebauer J 2023 Systematic atomic structure datasets for machine learning potentials: application to defects in magnesium *Phys. Rev. B* **107** 104103
[43] Hodapp M and Shapeev A 2020 In operando active learning of interatomic interaction during large-scale simulations *Mach. Learn.: Sci. Technol.* **1** 045005
[44] Mismetti L and Hodapp M 2024 Automated atomistic simulations of dissociated dislocations with *ab initio* accuracy *Phys. Rev. B* **109** 094120
[45] Qi J, Ko T W, Wood B C, Pham T A and Ong S P 2024 Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling *npj Comput. Mater.* **10** 43
[46] Schwalbe-Koda D, Hamel S, Sadigh B, Zhou F and Lordi V 2025 Model-free estimation of completeness, uncertainties and outliers in atomistic machine learning using information theory *Nat. Commun.* **16** 4014
[47] Bartók A P, Kondor R and Csányi G 2013 On representing chemical environments *Phys. Rev. B* **87** 184115
[48] Batatia I *et al* 2025 A foundation model for atomistic materials chemistry *J. Chem. Phys.* **163** 184110
[49] Batatia I, Batzner S, Kovács D P, Musaelian A, Simm G N C, Drautz R, Ortner C, Kozinsky B and Csányi G 2025 The design space of E(3)-equivariant atom-centred interatomic potentials *Nat. Mach. Intell.* **7** 56–67
[50] Fei S 2025 Data for: Model accuracy and data heterogeneity shape uncertainty quantification in machine learning interatomic potentials *GitHub* (available at: https://github.com/ufsf/UQ-ACE)