MULTI-FIDELITY TRAINING RESHAPES UNCERTAINTY DECOMPOSITION IN ATOMISTIC MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-fidelity training, which combines abundant low-fidelity data with scarce high-fidelity labels, is increasingly common in atomistic modeling, but its effect on predictive uncertainty remains largely unexplored. We investigate this question using deep ensembles of mean–variance MACE models trained on ANI-1x, contrasting high-fidelity training from scratch with low-fidelity pretraining followed by fine-tuning. We characterize aleatoric, epistemic, and total uncertainty throughout training and under two distribution shifts: unseen molecular systems and higher-energy configurations. We find that pretraining reduces prediction error and improves uncertainty-based ranking across all test sets, but does not yield well-calibrated predictive variances; calibration deteriorates more severely under energy extrapolation. Training-dynamics analysis localizes most of this calibration degradation to high-fidelity optimization, during which the learned aleatoric variance becomes overconfident on held-out data. Moreover, multi-fidelity fine-tuning produces a more effective uncertainty-based acquisition function than high-fidelity training from scratch. These findings show that uncertainty ranking and calibration capture distinct properties and must be assessed separately in multi-fidelity atomistic models, especially under extrapolation.

1 INTRODUCTION

Reliable uncertainty estimation is essential in atomistic machine learning, where models are routinely deployed beyond their training distribution, for instance in active learning and molecular dynamics (Podryabinkin & Shapeev, 2017; Unke et al., 2021; Perego & Bonati, 2024). Recent work has investigated ensembles, calibration and the decomposition of predictive uncertainty into aleatoric and epistemic components for atomistic models (Dai et al., 2025; Grasselli et al., 2025; Tan et al., 2023; Valdenegro-Toro & Mori, 2022; Busk et al., 2023). Meanwhile, multi-fidelity (Musil et al., 2019; Tran et al., 2020; Kim et al., 2024; Gardner et al., 2025; Ko & Ong, 2025) and transfer learning (Falk et al., 2023; Rocken & Zavadlav, 2025; Novelli et al., 2025; Radova et al., 2025) have emerged as practical strategies for combining abundant low-fidelity (LF) data with scarce high-fidelity (HF) labels, typically obtained at different levels of electronic-structure theory (Smith et al., 2019; Chen et al., 2023; Zaverkin et al., 2023). Yet, how multi-fidelity training shapes predictive uncertainty remains largely unexplored.

In standard statistical settings, the aleatoric–epistemic decomposition separates irreducible data noise from uncertainty about the model (Bishop & Nasrabadi, 2006; Kendall & Gal, 2017). Its interpretation is less straightforward in atomistic modeling, where reference energies are deterministic and data coverage is highly structured. Moreover, uncertainty is commonly evaluated only at convergence and through a single score, conflating two distinct properties: whether uncertainty correctly *ranks* configurations by prediction error, as required for acquisition and filtering, and whether its magnitude matches the error, i.e. whether the error bars are *calibrated* (Kuleshov et al., 2018; Rasmussen et al., 2023). These properties need not evolve together across fidelity changes, distribution shifts, or training stages. The latter is particularly relevant in practice because deployment ultimately relies on a checkpoint selected during training.

Contributions. Using ANI-1x data (Smith et al., 2020), with DFT as low fidelity and CCSD(T)/CBS as high fidelity, we compare LF-only, HF-only and pretrain-then-fine-tune (LF→HF) training of mean–variance MACE ensembles. We evaluate two distribution shifts: unseen molecular systems and higher-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

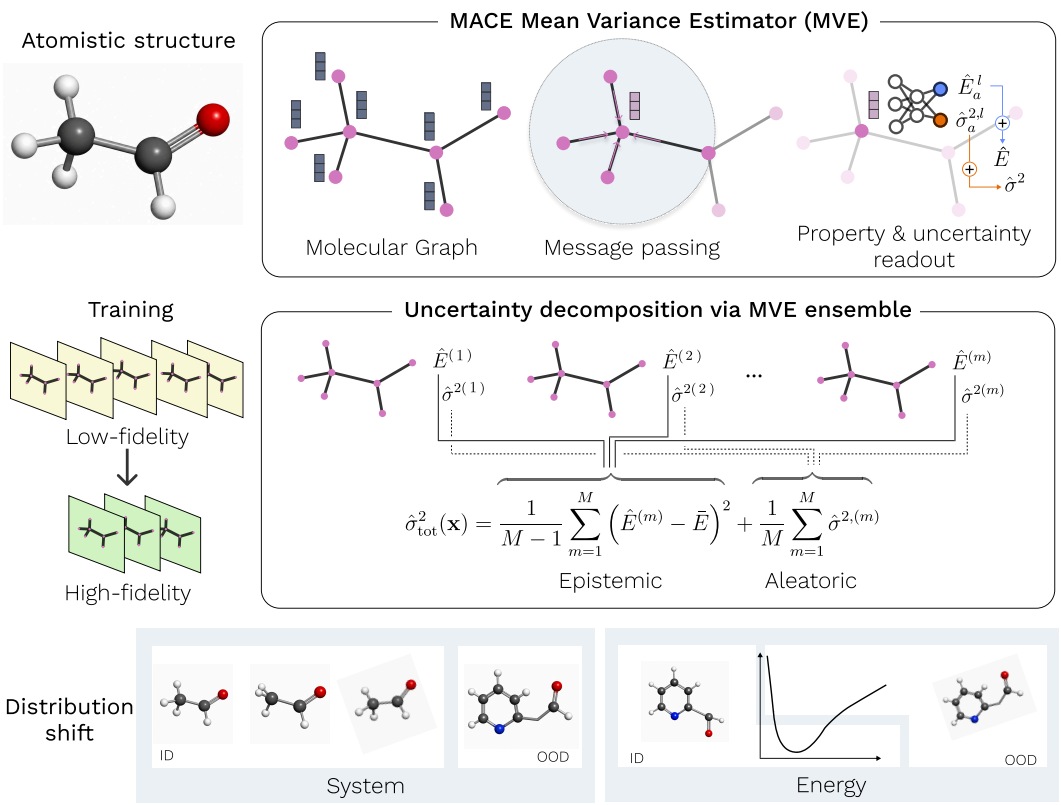

Figure 1: **Overview of the uncertainty estimation pipeline.** A mean–variance MACE model maps a structure to an energy and a variance. In an ensemble of such models, the average predicted variance gives the aleatoric uncertainty and the disagreement between members the epistemic uncertainty.

energy configurations of known systems. Across five independent data splits, we follow aleatoric (AU), epistemic (EU) and total uncertainty (TU) throughout training. Our study uses theoretically grounded mean–variance deep ensembles (Lakshminarayanan et al., 2017; Kendall & Gal, 2017) and, to our knowledge, provides the first systematic analysis of how multi-fidelity training affects uncertainty dynamics. We find that: (i) LF pretraining consistently improves predictive accuracy and the error *ranking* induced by AU, EU and TU across all test sets, but does not always improve *calibration*; (ii) unlike system shift, energy extrapolation decouples ranking from calibration: models rank high-energy errors more effectively while underestimating their magnitude more severely, and in-domain isotonic recalibration does not repair this; (iii) along training, most of the calibration is lost in the HF stage in both training protocols, and the loss emerges faster after pretraining: AU becomes markedly overconfident on unseen data, whereas EU mitigates the resulting degradation in TU; (iv) pretraining makes uncertainty a better guide for data acquisition: despite starting from higher accuracy, the pretrained model benefits more from uncertainty-guided selection than a model trained from scratch. We release splits, protocols, and evaluation code at https://anonymous.4open.science/r/multi-fidelity-training. Related work is reviewed in Appendix A.

2 Methodology

Figure 1 summarizes our pipeline: an ensemble of mean–variance MACE models, whose average predicted variance and disagreement define the aleatoric and epistemic components, respectively.

2.1 Error and predictive uncertainty

For a structure $\mathbf{x}$, defined by its atomic species and positions, with reference energy y, predictive uncertainty quantifies the reliability of a prediction $\hat{y}(\mathbf{x})$ before computing y. It is represented

by the predictive distribution $p(y|\mathbf{x}, \mathcal{D})$ conditioned on training data $\mathcal{D}$ (Murphy, 2012), and it is commonly decomposed into aleatoric and epistemic components (Gawlikowski et al., 2023). Epistemic uncertainty captures uncertainty in the predictor, arising from limited data coverage, model misspecification, or extrapolation, and is in principle reducible with additional data. Aleatoric uncertainty is conventionally associated with irreducible observation noise. However, since reference energies of a given electronic-structure method are deterministic (Helgaker et al., 2013), we instead interpret it operationally as input-dependent residual variability, or configuration-dependent difficulty. Whether the two terms actually behave as irreducible and reducible uncertainty is tested in Section 4.2.

With model parameters θ and target Y, the law of total variance (Bishop & Nasrabadi, 2006) gives

$$\mathrm{Var}(Y|\mathbf{x}, \mathcal{D}) = \underbrace{\mathbb{E}_{p(\theta|\mathcal{D})}[\mathrm{Var}(Y|\mathbf{x}, \theta)]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_{p(\theta|\mathcal{D})}[\mathbb{E}(Y|\mathbf{x}, \theta)]}_{\text{epistemic}}. \tag{1}$$

With deterministic labels, however, the learned variance captures any residual variation the model fails to explain, whether due to limited data, misspecification or optimization, while model disagreement provides only a proxy for epistemic uncertainty (Hüllermeier & Waegeman, 2021). We therefore use *aleatoric* and *epistemic* as operational labels for the two variance terms rather than as uniquely identified sources of uncertainty (Appendix B).

2.2 Predictive model and uncertainty parameterization

Mean–variance MACE model. We build on MACE, an E(3)-equivariant message-passing architecture with higher-body atomic-cluster-expansion features (Batatia et al., 2022; 2025). Given an atomistic configuration $\mathbf{x}$ with atom set $\mathcal{A}$, MACE represents the total energy as a sum of atomic contributions a across interaction layers ℓ: $\hat{E}(\mathbf{x}) = \sum_{a\in\mathcal{A}} \sum_{\ell=1}^{L} \hat{E}_a^{\ell}$. To obtain heteroscedastic (input-dependent) uncertainty estimates, we extend the final readout layer to output, for each atom a, both an energy contribution and a variance contribution, $(\hat{E}_a^L, \hat{\sigma}_a^{2,L})$, and obtain the structure-level predictive variance as their sum, $\hat{\sigma}^2(\mathbf{x}) = \sum_{a\in\mathcal{A}} \hat{\sigma}_a^{2,L}$. Modeling the reference energy as $E|\mathbf{x} \sim \mathcal{N}\big(\hat{E}(\mathbf{x}), \hat{\sigma}^2(\mathbf{x})\big)$, we train the model by minimizing the Gaussian negative log-likelihood (NLL)

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\frac{(E_i - \hat{E}(\mathbf{x}_i))^2}{2\hat{\sigma}^2(\mathbf{x}_i)} + \frac{1}{2}\log\hat{\sigma}^2(\mathbf{x}_i)\Big] + \text{const.} \tag{2}$$

The learned variance is the per-model estimate of aleatoric uncertainty in the ensemble decomposition below.

Ensemble-based uncertainty estimation. To capture epistemic uncertainty, we employ a deep ensemble (Lakshminarayanan et al., 2017) of M independently initialized mean–variance MACE models (Figure 1), closely related to the committee models used for active learning of interatomic potentials (Zhang et al., 2019; Schran et al., 2020; Yang et al., 2022). Each member m produces an energy $\hat{E}^{(m)}(\mathbf{x})$ and a variance $\hat{\sigma}^{2,(m)}(\mathbf{x})$, and the ensemble mean is $\bar{E}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M}\hat{E}^{(m)}(\mathbf{x})$. Treating the members as approximate samples from a posterior over parameters, we approximate the two terms of Equation 1 with averages over the M models, which gives the total uncertainty (TU)

$$\hat{\sigma}^2_{\mathrm{tot}}(\mathbf{x}) = \underbrace{\frac{1}{M}\sum_{m=1}^{M}\hat{\sigma}^{2,(m)}(\mathbf{x})}_{\text{aleatoric (AU)}} + \underbrace{\frac{1}{M}\sum_{m=1}^{M}\Big(\hat{E}^{(m)}(\mathbf{x}) - \bar{E}(\mathbf{x})\Big)^2}_{\text{epistemic (EU)}}. \tag{3}$$

We evaluate the ranking of AU, EU and TU separately, as three competing signals, while for calibration we focus on TU since it is the predictive variance (Section 3).

3 Experimental setup

Data and fidelities. All experiments use the ANI-1x data set (Smith et al., 2020), which provides molecular configurations labelled at several levels of theory. The low-fidelity (LF) target is the ωB97x/def2-TZVPP DFT energy and the high-fidelity (HF) target is the CCSD(T)/CBS coupled-cluster energy, which is available for far fewer configurations (the ANI-1ccx subset (Smith et al.,

2019)). Only energies are used. Each split has 50,000 DFT and 5,000 CC training configurations, 10,000 DFT and 1,000 CC validation configurations, and 5,000 CC configurations per test set; CC configurations are a subset of the DFT ones.

Distribution shifts. The *system-level shift* holds out 40% of the molecular systems entirely: System-ID tests on unseen configurations of the training systems, System-OOD on the held-out systems (Appendix D). The *energy-based shift* ranks the configurations of every system by their relative DFT energy, with within-system quantile $q_s(x) \in [0, 1]$. Configurations with $q_s \leq 0.5$ are used for training, validation and the Energy-ID test, configurations with $q_s \geq 0.55$ form the Energy-OOD test, and the band in between is discarded (Appendix E).

Training protocols. LF-only models are trained on DFT energies and **HF-only** models on CC energies, both from scratch for 300 epochs. **LF→HF** models are LF-only models fine-tuned on CC energies for 100 epochs with a tenfold smaller learning rate; every fine-tuned member starts from the LF-only member with the same split and seed. All protocols use the Gaussian NLL, the same architecture and $M = 10$ ensemble members (Appendix F). The checkpoint of each member is selected by its validation NLL; test sets are never used for selection. This selection criterion retains HF-only members after 30–111 of 300 epochs and LF→HF members after one to three fine-tuning epochs (Appendix F).

Replication and statistics. The pipeline, from data splitting to ensemble training, is repeated on five independent dataset splits, the unit of replication. Metrics are computed within a split from its ten-member ensemble and summarized by the five-split mean with a two-sided 95% Student-t interval (4 degrees of freedom); ENCE, RMSE and ratios use geometric means with log-space intervals.

Metrics. Both errors and uncertainties are normalized per atom, i.e. energies are divided by the number of atoms and variances by the square of the number of atoms. *Accuracy* is the RMSE of the ensemble mean. *Ranking* is measured by the area under the sparsification error curve (AUSE) (Lind et al., 2024), which compares discarding configurations by uncertainty with discarding them by their true error and rewards identifying the largest errors (lower is better); *calibration* by the expected normalized calibration error (ENCE) (Levi et al., 2022), $\frac{1}{B}\sum_b |\mathrm{MV}_b - \mathrm{MSE}_b|/\mathrm{MV}_b$ over $B = 15$ equal-count bins of predicted variance, with mean variance MV_b and mean squared error MSE_b (Appendix C). A bin where the variance overestimates the error contributes less than one, so ENCE above one can only come from overconfident bins. Note that, because TU is the predictive variance and AU and EU are its two components, the three cannot all match the magnitude of the error at once: calibration is expected of TU, while the ENCE of AU and EU only shows how much of the error each part accounts for. Our figures therefore highlight TU when showing calibration. All metrics exclude the handful of Energy-OOD configurations (13–16 of 5,000 per split) that are compressed below any interatomic distance seen in training (Appendix G).

Research questions. RQ1: how do AU, EU and TU rank and calibrate at convergence across fidelity regimes? **RQ2:** do system-level and energy-based shifts affect ranking and calibration differently? **RQ3:** how do the signals evolve during training, especially across the LF-to-HF transition? **RQ4:** how does pretraining affect uncertainty-guided data acquisition?

4 Results

We first analyze the converged models (Section 4.1). We then examine the full multi-fidelity training trajectory (Section 4.2) and compare its high-fidelity stage with HF-only training (Section 4.3). Finally, we present a small acquisition experiment (Section 4.4). All results are replicated over five data splits; figures show 95% intervals, and the complete set of metrics, reliability diagrams and training curves is in Appendices I and H.

4.1 Pretraining improves ranking but not always calibration

Figure 2 compares the HF-only and multi-fidelity training. As expected, pretraining reduces the prediction error (RMSE), most strongly in domain but also out of domain. Instead, the other metrics reveal a different behavior of the uncertainty components. On the one hand, pretraining consistently improves *ranking* (AUSE) (panel b): across all three signals and on all four test sets, pretrained ensembles order prediction errors more effectively than ensembles trained only on HF data. Notably,

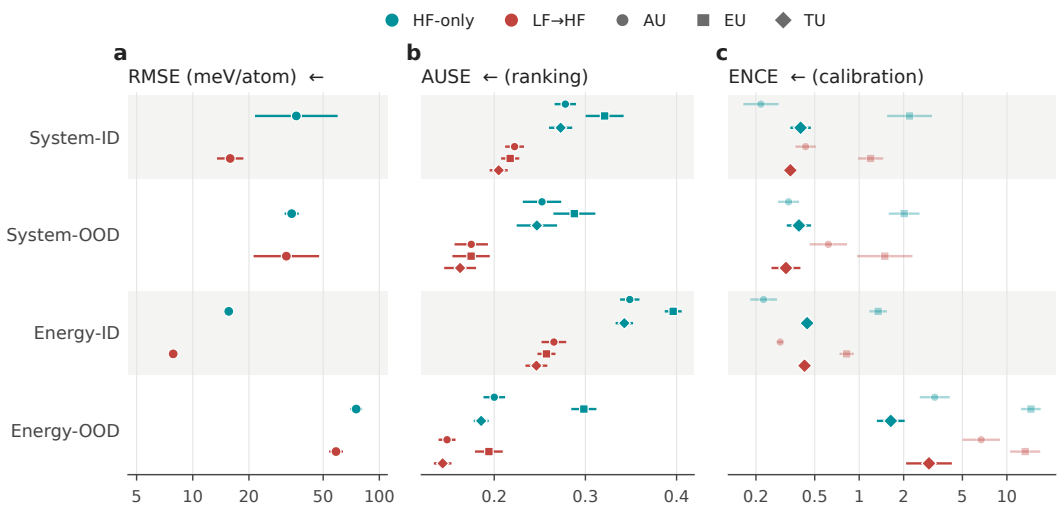

Figure 2: **Accuracy and uncertainty for the converged models**. **(a)** RMSE of the ensemble mean, **(b)** AUSE and **(c)** ENCE. Colour denotes the protocol (teal HF-only, red LF→HF) and marker the signal (circle AU, square EU, diamond TU); in (c) TU, the predictive variance, is highlighted and its components AU and EU are faded. Bars are 95% intervals over five splits, and RMSE and ENCE use log axes with geometric means. The wide HF-only System-ID RMSE interval comes from a single configuration in one split, which TU ranks as the most uncertain of 5,000.

multi-fidelity training reshapes the uncertainty decomposition rather than merely improving prediction accuracy. Indeed, there is a change in the ranking capabilities of the individual components: in HF-only, EU is the weakest ranking signal on every test set; with pretraining, EU ranks as well as or better than AU on three of the four test sets and approaches the ranking quality of total predictive uncertainty (see also Appendix I.2).

By contrast, pretraining has a mixed effect on *calibration* (ENCE; panel c). The relevant uncertainty (TU) is slightly better calibrated after pretraining in domain and on unseen systems, but substantially worse on Energy-OOD. Its components also diverge: EU improves across all test sets, while AU deteriorates. Thus, pretraining lowers error and improves ranking, but does not necessarily produce uncertainties with the correct magnitude.

Notably, we find that the relationship between ranking and calibration depends critically on the type of distribution shift. To highlight this, Figure 3 places each test set in the calibration–ranking plane and connects every in-domain result to its out-of-domain counterpart. Under system shift, moving from ID to unseen molecular systems improves ranking with little change in calibration. Under energy shift, however, where OOD means higher-energy configurations, ranking *improves* while calibration error increases several-fold: for TU, by approximately $4\times$ with HF-only training and $7\times$ with LF→HF. The Energy-OOD behavior in Figure 2c is therefore not specific to pretraining. Both protocols rank high-energy errors better than in-domain errors, yet underestimate them more severely. In other words, the models largely identify *which* high-energy configurations will incur larger errors, since those errors are larger and easier to order, but not *by how much*. Ranking and calibration can therefore move in opposite directions along some extrapolation axes, and isotonic recalibration fitted in domain does not repair this (Appendix I.3).

4.2 Uncertainty signals evolve differently during training

To better understand how uncertainty components behave in different training schemes, we examine their evolution over the course of training. Figure 4 follows the full LF→HF trajectory on the model's own training data (a, b), the ID test set (c, d) and the OOD test set (e, f). On the training set, ranking changes little during DFT pretraining and improves markedly during CC fine-tuning (panel a) while the calibration of TU remains stable along the training epochs (panel b). HF-only models behave the same way on their training data (Appendix H).

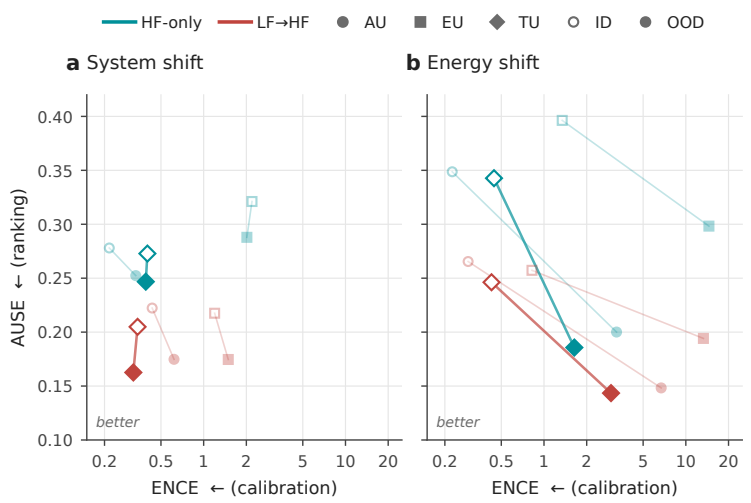

Figure 3: **Ranking and calibration behavior under different distribution shifts**. ENCE ($\log x$) against AUSE (y) for the converged models under **(a)** system shift and **(b)** energy shift; lower left is best. Colour: protocol; marker: signal, with TU highlighted and AU and EU faded. For every signal a thin line joins the ID (hollow) to the OOD test set (filled).

On the ID test set, ranking performance improves for all three uncertainty signals during pretraining (panel c). The change from the low- to the high-fidelity target produces a sharp discontinuity; during the subsequent fine-tuning stage, the ranking performance of EU and TU recovers, whereas that of AU changes little. Calibration follows the opposite trend (panel d): TU becomes better calibrated during pretraining but progressively less calibrated during fine-tuning. This deterioration is driven primarily by the aleatoric component, whose calibration error increases by approximately two orders of magnitude over the fine-tuning stage.

A similar ranking behavior is observed on the OOD test sets (panel e): ranking improves during pretraining under both distribution shifts and approaches a plateau during fine-tuning. The evolution of ENCE, however, depends on the type of shift (panel f). During pretraining, TU remains reasonably well calibrated on unseen molecular systems but progressively drifts on higher-energy configurations. Once fine-tuning begins, its calibration error increases under both shifts. However, the two uncertainty components behave differently: the calibration error associated with EU remains comparatively stable and ends near its pretraining level, whereas AU becomes strongly overconfident, with its predicted variance falling one to two orders of magnitude below the squared prediction error. The evolving balance between the two components is examined in Appendix H.1. Although the poor reliability of aleatoric uncertainty under distribution shift is well documented (Valdenegro-Toro & Mori, 2022), its deterioration on the ID test set indicates that the effect is also associated with high-fidelity training and is accelerated by pretraining (Section 4.3). A plausible explanation is that the variance head follows the shrinking residuals of the training configurations and transfers these small variance estimates to held-out configurations whose errors have not decreased, a known failure mode of mean–variance training (Nix & Weigend, 1994; Seitzer et al., 2022). Because the reference energies are deterministic, this behavior suggests that the aleatoric term acts less like irreducible noise and more like a reducible estimate of in-sample residual difficulty.

4.3 Pretraining speeds up both the gain in ranking and the loss of calibration

Having examined the full training dynamics of LF→HF, we now compare its high-fidelity stage with HF-only training (Figure 5), focusing on out-of-distribution performance. Under both distribution shifts and for all three uncertainty signals, the LF→HF models achieve a lower AUSE than the HF-only ones already from the first epoch. For both protocols, AUSE then decreases and levels off over the course of high-fidelity training.

Calibration follows a different trajectory (panels c and d). For HF-only training, ENCE initially decreases before increasing again, whereas for LF→HF it is lowest near the beginning of the high-

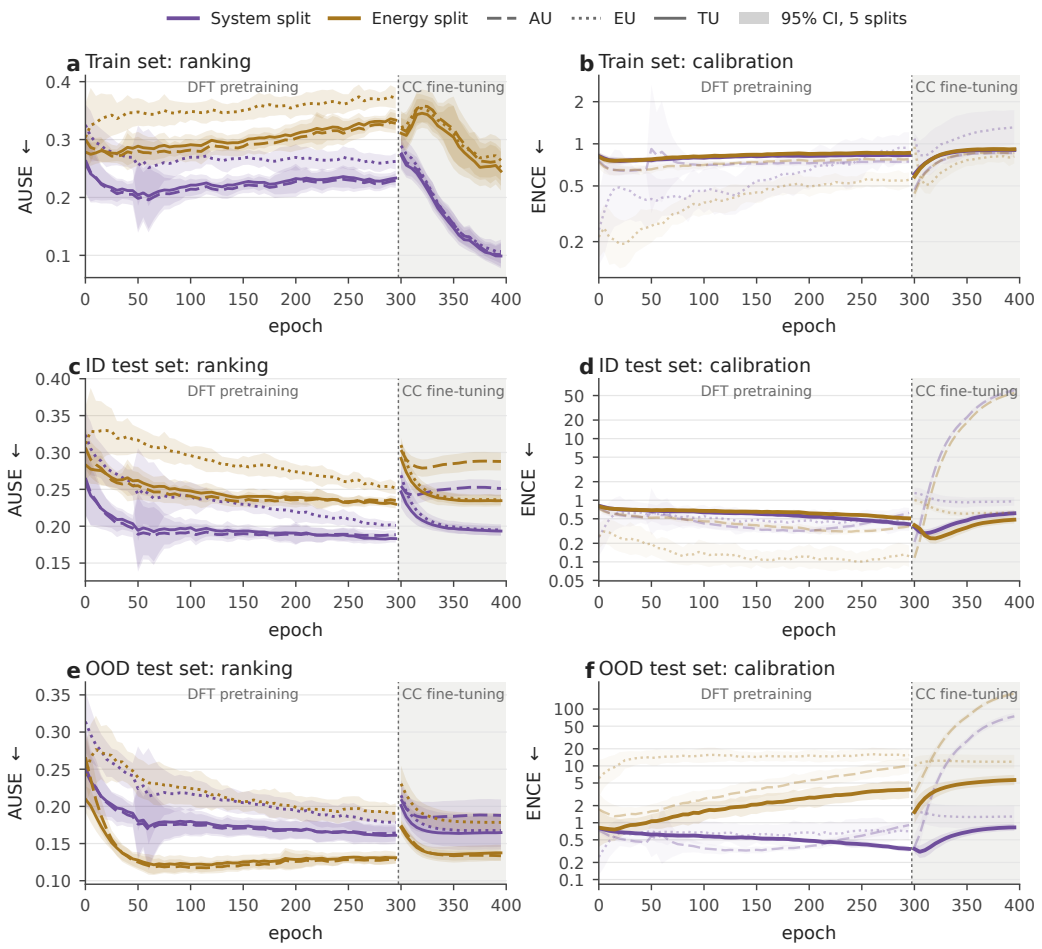

Figure 4: **Uncertainty evolution during training for the multi-fidelity case**. LF→HF through DFT pretraining (epochs 0–299, scored against DFT labels) and CC fine-tuning (300–399, shaded, CC labels) on its own training data (a, b), the ID (c, d) and the OOD test set (e, f). Note that at epoch 300 the target labels change and, except for Energy-OOD, the evaluated configurations also change; values across the transition are therefore not directly comparable. Colour: split (purple system, brown energy); line style: signal (dashed AU, dotted EU, solid TU), with TU highlighted in the calibration panels; bands: 95% intervals over five splits.

fidelity stage and then grows with fine-tuning. As observed above, this deterioration is driven primarily by the aleatoric component. Pretraining therefore does not appear to cause the degradation, but it accelerates it considerably: within 100 fine-tuning epochs, the AU ENCE reaches about 80 on System-OOD and 200 on Energy-OOD, compared with about 14 and 70, respectively, after 300 epochs of HF-only training. The epistemic component behaves differently: its calibration error ends close to or below its initial value in all four settings, limiting the deterioration of total uncertainty relative to that of the aleatoric component. Overall, these results show that uncertainty ranking and calibration follow distinct, partially decoupled trajectories during training: ranking can continue to improve even as calibration deteriorates. Checkpoint selection should therefore not rely on prediction error alone, but should jointly monitor accuracy, uncertainty ranking, and calibration throughout training.

4.4 Pretrained uncertainty estimates improve single-round acquisition

Finally, we compare single- and multi-fidelity training in a setting where uncertainty is used to acquire new data. We perform a single acquisition round on Energy-OOD, selecting 500 configurations either at random or according to the highest AU, EU, or TU scores. After adding the corresponding high-

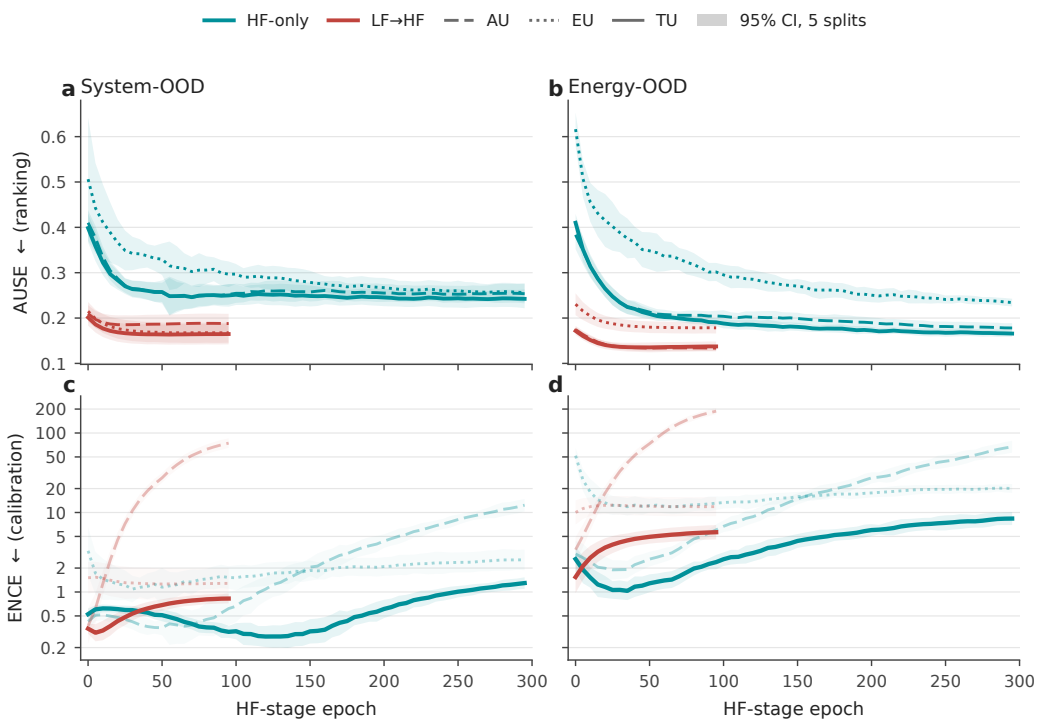

Figure 5: **Comparison of uncertainty evolution for the high-fidelity target.** AUSE (top) and ENCE (bottom, log scale) on System-OOD (a, c) and Energy-OOD (b, d) against the HF-stage epoch, for HF-only from scratch (teal, 300 epochs) and CC fine-tuning of LF→HF (red, 100 epochs). Line style: signal (dashed AU, dotted EU, solid TU); in the calibration panels TU is highlighted. Lines are exponential moving averages of the five-split mean (geometric for ENCE); bands are 95% intervals. Last-epoch values differ from Figure 2, whose models are best-validation checkpoints.

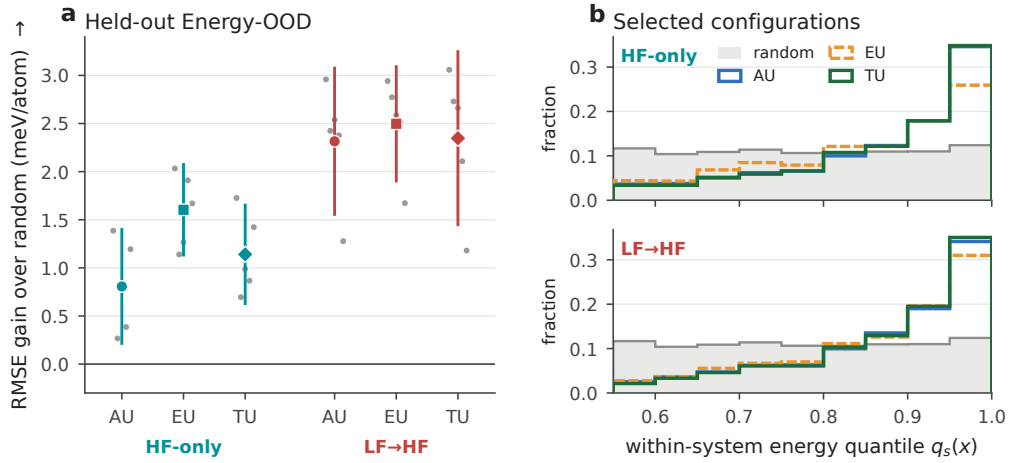

Figure 6: **Uncertainty-guided acquisition.** One round acquires 500 of 2,500 Energy-OOD pool configurations; models are evaluated on the disjoint held-out half. **(a)** RMSE reduction relative to random selection (positive favours uncertainty); markers are five-split means with 95% intervals, grey dots individual splits. **(b)** Within-system energy quantile $q_s(x)$ of the acquired configurations; grey is random selection, which reproduces the pool.

fidelity labels to the training set, each model continues training and is evaluated on a disjoint held-out set (Appendix J). All uncertainty-based acquisition strategies outperform random selection (Figure 6a).

However, all three uncertainty signals yield larger improvements for the pretrained ensembles than for those trained only on high-fidelity data: relative to random selection, their RMSE decreases by about 2.3–2.5 meV/atom, compared with 0.8–1.6 meV/atom for HF-only training. This is notable because the pretrained ensemble is already more accurate before acquisition, with an RMSE of about 59 meV/atom on the held-out set compared with 75 meV/atom for the HF-only ensemble, yet it benefits more from uncertainty-guided selection. The two protocols select configurations with similar energy-quantile distributions (Figure 6b), indicating that the larger improvement after pretraining cannot be explained simply by selecting systematically higher-energy configurations. Instead, it suggests that pretraining produces a more informative uncertainty-based acquisition function within comparable energy ranges.

5 Conclusion

We studied how low-fidelity pretraining followed by high-fidelity fine-tuning shapes the ranking, calibration, and decomposition of predictive uncertainty in atomistic models. Low-fidelity pretraining improves predictive accuracy and the error ranking for every uncertainty signal across all test sets, but does not consistently improve total uncertainty calibration. It also changes the uncertainty decomposition: without pretraining, the epistemic component is the weakest ranking signal; after pretraining it approaches total uncertainty on three of the four test sets (**RQ1**). Under both shifts, errors are ranked better out-of-domain than in-domain, whereas only energy extrapolation degrades calibration (**RQ2**). Along training, most calibration loss arises during the high-fidelity stage under both protocols, and develops faster after pretraining: aleatoric uncertainty remains conservative on the training data but becomes strongly overconfident on unseen configurations, while epistemic uncertainty mitigates the degradation of total uncertainty (**RQ3**). Finally, pretrained ensembles yield more effective uncertainty-based acquisition functions than HF-only models (**RQ4**).

These findings lead to clear practical implications. Ranking and calibration should be evaluated separately, and total uncertainty calibration should be monitored on held-out data throughout training because it may deteriorate even as ranking improves. Checkpoint selection should therefore account for uncertainty ranking and calibration in addition to prediction error. Aleatoric estimates obtained after training on scarce high-fidelity data require particular caution on unseen configurations, and in-domain recalibration should be validated explicitly before use under extrapolation (Appendix I.3). Although our analysis is limited to energy prediction, one uncertainty estimator, and a single acquisition setting (Appendix K), it establishes a broader conclusion: uncertainty in multi-fidelity atomistic models is not a static property, with its ranking, calibration, and decomposition changing across fidelity transitions, distribution shifts, and training stages.

AI use statement

In this work, we used generative AI tools (large language model coding assistants) to help write and refactor analysis and plotting code and to draft and edit parts of the manuscript text. We have not used generative AI tools to generate research ideas, experimental data or results. All AI-assisted code was reviewed by the authors, and every number reported in the paper was recomputed from the experiment outputs by scripts in the released repository. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

Ethics statement

This work uses only public data sets of molecular configurations labelled with electronic-structure methods; it involves no human subjects or personal data. Possible misuse and the broader impact of the work are discussed in Appendix L.

Reproducibility statement

All data are public: ANI-1x (Smith et al., 2020). Code, the exact data splits, configuration files, evaluation scripts and the scripts that produce every figure are available at the anonymous repository linked in Section 1. Section 3 summarizes the data, shifts, training protocols, checkpoint selection

and statistical protocol. Appendix B states the assumptions behind the estimator and the evaluation, Appendices D and E detail the two splitting procedures, Appendix F the model, the optimization, the random seeds, the selected checkpoints and the computational resources, Appendix G the training-support filter, Appendix C the metrics, Appendix I.3 the recalibration and Appendix J the acquisition protocol.

REFERENCES

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor NC Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The design space of e (3)-equivariant atom-centred interatomic potentials. *Nature Machine Intelligence*, 7(1):56–67, 2025.

Hubert Beck, Pavol Simko, Lars L Schaaf, Ondrej Marsalek, and Christoph Schran. Multi-head committees enable direct uncertainty prediction for atomistic foundation models. *The Journal of Chemical Physics*, 163(23), 2025.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Jonas Busk, Mikkel N Schmidt, Ole Winther, Tejs Vegge, and Peter Bjørn Jørgensen. Graph neural network interatomic potential ensembles with calibrated aleatoric and epistemic uncertainty on energy and forces. *Physical Chemistry Chemical Physics*, 25(37):25828–25837, 2023.

Jesús Carrete, Hadrián Montes-Campos, Ralf Wanzenböck, Esther Heid, and Georg KH Madsen. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *The Journal of Chemical Physics*, 158(20), 2023.

Michael S Chen, Joonho Lee, Hong-Zhou Ye, Timothy C Berkelbach, David R Reichman, and Thomas E Markland. Data-efficient machine learning potentials from transfer learning of periodic correlated electronic structure methods: Liquid water at afqmc, ccsd, and ccsd (t) accuracy. *Journal of Chemical Theory and Computation*, 19(14):4510–4519, 2023.

Jin Dai, Santosh Adhikari, and Mingjian Wen. Uncertainty quantification and propagation in atomistic machine learning. *Reviews in Chemical Engineering*, 41(4):333–357, 2025.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pp. 1184–1193. PMLR, 2018.

John Falk, Luigi Bonati, Pietro Novelli, Michele Parrinello, and Massimiliano Pontil. Transfer learning for atomistic simulations using gnns and kernel mean embeddings. *Advances in Neural Information Processing Systems*, 36:29783–29797, 2023.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

John LA Gardner, Hannes Schulz, Jean Helie, Lixin Sun, and Gregor NC Simm. Understanding multi-fidelity training of machine-learned force-fields. *arXiv preprint arXiv:2506.14963*, 2025.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial intelligence review*, 56(Suppl 1):1513–1589, 2023.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Federico Grasselli, Sanggyu Chong, Venkat Kapil, Silvia Bonfanti, and Kevin Rossi. Uncertainty in the era of machine learning for atomistic modeling. *Digital Discovery*, 4(10):2654–2675, 2025.

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep learning. *ACM Computing Surveys*, 58(7):1–35, 2026.

Esther Heid, Charles J McGill, Florence H Vermeire, and William H Green. Characterizing uncertainty in machine learning for chemistry. *Journal of Chemical Information and Modeling*, 63(13):4012–4029, 2023.

Trygve Helgaker, Poul Jorgensen, and Jeppe Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, 2013.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 5(3):035006, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Jaesun Kim, Jisu Kim, Jaehoon Kim, Jiho Lee, Yutack Park, Youngho Kang, and Seungwu Han. Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials. *Journal of the American Chemical Society*, 147(1):1042–1054, 2024.

Tsz Wai Ko and Shyue Ping Ong. Data-efficient construction of high-fidelity graph deep learning interatomic potentials. *npj Computational Materials*, 11(1):65, 2025.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.

Simon Kristoffersson Lind, Ziliang Xiong, Per-Erik Forssén, and Volker Krüger. Uncertainty quantification metrics for deep regression. *Pattern Recognition Letters*, 186:91–97, 2024.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Félix Musil, Michael J Willatt, Mikhail A Langovoy, and Michele Ceriotti. Fast and accurate uncertainty estimation in chemical machine learning. *Journal of chemical theory and computation*, 15(2):906–915, 2019.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Pietro Novelli, Giacomo Meanti, Pedro J Buigues, Lorenzo Rosasco, Michele Parrinello, Massimiliano Pontil, and Luigi Bonati. Fast and fourier features for transfer learning of interatomic potentials. *npj Computational Materials*, 11(1):293, 2025.

Simone Perego and Luigi Bonati. Data efficient machine learning potentials for modeling catalytic reactivity via active learning and enhanced sampling. *npj Computational Materials*, 10(1):291, 2024.

Evgeny V Podryabinkin and Alexander V Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.

Mariia Radova, Wojciech G Stark, Connor S Allen, Reinhard J Maurer, and Albert P Bartók. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. *npj Computational Materials*, 11(1):237, 2025.

Maria H Rasmussen, Chenru Duan, Heather J Kulik, and Jan H Jensen. Uncertain of uncertainties? a comparison of uncertainty quantification metrics for chemical data sets. *Journal of Cheminformatics*, 15(1):121, 2023.

Sebastien Rocken and Julija Zavadlav. Enhancing machine learning potentials through transfer learning across chemical elements. *Journal of Chemical Information and Modeling*, 65(14): 7406–7414, 2025.

Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153 (10), 2020.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Nicki Skafte, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of mean variance estimation neural networks. *Neurocomputing*, 597:127929, 2024.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):2903, 2019.

Justin S Smith, Roman Zubatyuk, Benjamin Nebgen, Nicholas Lubbers, Kipton Barros, Adrian E Roitberg, Olexandr Isayev, and Sergei Tretiak. The ani-1ccx and ani-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific data*, 7(1):134, 2020.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *International conference on artificial intelligence and statistics*, pp. 5593–5613. PMLR, 2023.

Aik Rui Tan, Shingo Urata, Samuel Goldman, Johannes CB Dietschreit, and Rafael Gómez-Bombarelli. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Computational Materials*, 9(1):225, 2023.

Anh Tran, Julien Tranchida, Tim Wildey, and Aidan P Thompson. Multi-fidelity machine-learning with uncertainty quantification and bayesian optimization for materials design: Application to ternary random alloys. *The Journal of Chemical Physics*, 153(7), 2020.

Oliver T Unke, Stefan Chmiela, Huziel E Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T Schutt, Alexandre Tkatchenko, and Klaus-Robert Muller. Machine learning force fields. *Chemical reviews*, 121(16):10142–10186, 2021.

Matias Valdenegro-Toro and Daniel Saromo Mori. A deeper look into aleatoric and epistemic uncertainty disentanglement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1508–1516. IEEE, 2022.

Manyi Yang, Luigi Bonati, Daniela Polino, and Michele Parrinello. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catalysis Today*, 387:143–149, 2022.

Viktor Zaverkin, David Holzmüller, Luca Bonfirraro, and Johannes Kästner. Transfer learning for chemically accurate interatomic neural network potentials. *Physical Chemistry Chemical Physics*, 25(7):5383–5396, 2023.

Linfeng Zhang, De-Ye Lin, Han Wang, Roberto Car, and Weinan E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials*, 3(2):023804, 2019.

A Related Work

Recent surveys and perspectives provide broad overviews of the uncertainty quantification (UQ) field, covering Bayesian methods, ensembles, calibration strategies, active learning, and the relationship between uncertainty, dataset composition, transferability, and model robustness (Dai et al., 2025; Grasselli et al., 2025; He et al., 2026). These works highlight that benchmarking and interpretation of UQ methods remain open challenges, especially in transfer and multi-fidelity settings.

A large body of work has focused on practical uncertainty estimators for neural network potentials. Ensemble-based methods remain among the most widely used and robust approaches. For instance, Tan et al. (2023) compare ensembles against single-model uncertainty methods such as mean-variance estimation, evidential regression, and Gaussian mixture models, finding that no method uniformly dominates across datasets and metrics, while ensembles remain strong for robustness and generalization. Other work has proposed more efficient ensemble variants, including multi-headed architectures, committee models, shallow or last-layer ensembles, and shared-representation ensembles, to reduce computational cost while retaining useful uncertainty estimates (Carrete et al., 2023; Beck et al., 2025; Kellner & Ceriotti, 2024).

Several works have specifically addressed calibration and uncertainty decomposition in atomistic models. Busk et al. (2023) propose a framework for calibrated aleatoric and epistemic uncertainty in graph neural network potentials, using deep ensembles to decompose total uncertainty and post-hoc recalibration to improve agreement between predicted uncertainty and observed error. More generally, Valdenegro-Toro & Mori (2022) study disentangled aleatoric and epistemic uncertainty across different UQ methods, showing that the interaction between these components can be non-trivial, that aleatoric uncertainty may become unreliable under distribution shift, and that ensembles provide the most consistent decomposition. In molecular property prediction, Heid et al. (2023) further analyze uncertainty sources by separating epistemic uncertainty into bias and variance contributions, showing that different sources of error dominate depending on noise level, dataset size, model architecture, and representation.

A related line of work studies multi-fidelity learning in atomistic modeling. Gardner et al. (2025) investigate multi-fidelity training strategies for atomistic models, including pretraining–fine-tuning and multi-headed training, showing that fine-tuning can improve accuracy but that learned representations may remain method-specific, while multi-headed models can learn more method-agnostic representations across fidelities. More broadly, multi-fidelity UQ has been studied in surrogate modeling and materials design, often using Gaussian process models to combine information across fidelity levels while quantifying predictive confidence (Tran et al., 2020).

Finally, the interpretation of uncertainty depends strongly on the evaluation metric. Rasmussen et al. (2023); Lind et al. (2024) analyze common regression UQ metrics, including AUSE, calibration error, Spearman correlation, and NLL, showing that they capture different properties and can lead to different conclusions.

Overall, existing work has made substantial progress on uncertainty estimation for atomistic models, uncertainty decomposition, calibration, efficient ensemble methods, and multi-fidelity training. However, these directions have largely been studied separately. In contrast, our work provides a controlled empirical analysis of aleatoric and epistemic uncertainty across multi-fidelity training regimes, system-level and energy-based distribution shifts, and training dynamics. This allows us to study not only whether uncertainty is accurate at convergence, but also how its components evolve and whether they can serve as diagnostic signals for model behavior under changing fidelity and data regimes.

B Technical background: error decomposition and mean–variance ensembles

This appendix collects standard results behind the estimator of Section 2.2 and the way we evaluate it. None of them is new; we state them in our notation to make explicit what AU, EU and TU can and cannot measure when reference energies are deterministic. All identities below are homogeneous of degree two in the energies, so they hold unchanged for the per-atom quantities used throughout the paper.

Bias–variance decomposition. Let $y(\mathbf{x})$ be the reference energy of configuration $\mathbf{x}$, deterministic for a given electronic-structure method, and let $f(\mathbf{x}; \mathcal{D}, \omega)$ be a model trained on data $\mathcal{D}$ with random seed ω (initialization and data order). With $\bar{f}(\mathbf{x}) = \mathbb{E}_{\mathcal{D},\omega}[f(\mathbf{x}; \mathcal{D}, \omega)]$, the expected squared error decomposes as (Geman et al., 1992)

$$\mathbb{E}_{\mathcal{D},\omega}\Big[\big(f(\mathbf{x}; \mathcal{D}, \omega) - y(\mathbf{x})\big)^2\Big] = \underbrace{\big(\bar{f}(\mathbf{x}) - y(\mathbf{x})\big)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D},\omega}\Big[\big(f(\mathbf{x}; \mathcal{D}, \omega) - \bar{f}(\mathbf{x})\big)^2\Big]}_{\text{variance}}. \tag{4}$$

With noisy labels, an irreducible noise term would add to the right-hand side; with deterministic labels it vanishes, so the whole expected error is reducible in principle, the variance by more data and the bias by a better model class. A deep ensemble whose members share the training set and differ only in their seed samples the variation over ω at fixed $\mathcal{D}$, which is substantial for neural networks (Fort et al., 2019). By the law of total variance, its expectation over data sets is at most the variance term of Equation 4, and it is blind to the bias, which all members share. EU is therefore an operational and generally partial estimate of reducible error.

Mean–variance estimation. Each member is trained with the Gaussian NLL of Section 2.2. For a fixed mean $\mu(\mathbf{x})$, the expected loss under a target distribution $p(y \mid \mathbf{x})$ is minimized by the conditional mean squared residual (Nix & Weigend, 1994),

$$\sigma^{2\star}(\mathbf{x}) = \mathbb{E}\big[(y - \mu(\mathbf{x}))^2 \,\big|\, \mathbf{x}\big]. \tag{5}$$

With deterministic labels, $p(y \mid \mathbf{x})$ is a point mass, and a variance head fitted to finitely many configurations approximates a local average of its member's squared training residuals, whatever their origin (limited data, misspecification or optimization), rather than observation noise. Two known properties of the estimator matter for our results (Skafte et al., 2019; Seitzer et al., 2022; Stirn et al., 2023; Sluijterman et al., 2024). First, the head is fitted to training residuals, which shrink as the network fits its training set, so on unseen configurations it can inherit the scale of the training error rather than that of the test error; this is the behavior of AU observed in Sections 4.2 and 4.3. Second, the gradient with respect to the mean is weighted by $1/\sigma^2$, so configurations that the head already deems easy dominate the fit of the mean (Seitzer et al., 2022). We use the plain Gaussian NLL, without the modifications proposed in these works.

Ensembles and the law of total variance. Treating the M members as samples from an approximate posterior, the predictive distribution is the mixture $\frac{1}{M}\sum_m \mathcal{N}\big(\hat{E}^{(m)}(\mathbf{x}), \hat{\sigma}^{2,(m)}(\mathbf{x})\big)$. Its mean is the ensemble mean $\bar{E}(\mathbf{x})$ and its variance is exactly AU + EU in Equation 3, the sample counterpart of Equation 1 (Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Depeweg et al., 2018). TU is therefore the variance of the predictive distribution around the prediction we evaluate, and the only one of the three signals that the observed error of $\bar{E}$ can calibrate. The split into AU and EU is operational (Hüllermeier & Waegeman, 2021): neither term has an observable target of its own against the error of the ensemble mean, so their ENCE in the main text describes how much of that error each term accounts for, not whether the term is calibrated.

The variance heads and the error of the ensemble mean. The members themselves do provide observable targets, since each head is trained to match its own member's squared error. Let $e_m = y - \hat{E}^{(m)}(\mathbf{x})$ be the error of member m and $e = y - \bar{E}(\mathbf{x})$ that of the ensemble mean. Because the member deviations $\hat{E}^{(m)} - \bar{E}$ average to zero, the ambiguity decomposition (Krogh & Vedelsby, 1994) holds exactly for every configuration,

$$e^2 = \frac{1}{M}\sum_{m=1}^{M} e_m^2 - \mathrm{EU}, \qquad \text{hence} \qquad \mathrm{TU} - e^2 = \Big(\mathrm{AU} - \frac{1}{M}\sum_{m=1}^{M} e_m^2\Big) + 2\,\mathrm{EU}, \tag{6}$$

where EU is the population variance of the member predictions, as in Equation 3. The bracket compares each head with the quantity it is trained to match, and, being linear, the identity also holds for the bin averages used by ENCE. Two consequences follow. If every head matched its member's squared error, TU would exceed the squared error of the ensemble mean by 2 EU, that is, it would be calibrated or conservative; this is consistent with TU lying on or slightly below the diagonal in domain (Appendix I.1). Conversely, TU can underestimate the error of the ensemble mean only where the heads underestimate their members' errors by more than 2 EU. The overconfidence of

TU on unseen data (Sections 4.2 and 4.3) is therefore a failure of the variance heads, a conclusion that needs no separate target for AU. Finally, the average member error exceeds e^2 by EU, which is typically smaller than AU (Appendix H.1); where AU falls far short of the error, as on unseen data after fine-tuning, comparing AU with e^2 rather than with the members' own errors therefore changes little.

C Details on uncertainty evaluation metrics

In this section, we provide the formal definitions of the metrics used in the main text. These metrics are designed to capture complementary aspects of predictive uncertainty, including ranking quality, calibration, and probabilistic correctness.

C.1 Area under the sparsification error (AUSE)

AUSE evaluates how well uncertainty ranks samples according to their prediction errors.

Let $S = \{(x_i, y_i)\}_{i=1}^N$ be a test set, and define the mean absolute error:

$$\mathrm{MAE}(S) = \frac{1}{|S|} \sum_{(x_i,y_i)\in S} |y_i - \hat{y}(x_i)|. \tag{7}$$

For a fraction $\alpha \in [0, 1]$, let:

- $S_V(\alpha)$ denote the subset obtained by removing the top-α fraction of samples with the largest prediction errors (oracle ranking),
- $S_U(\alpha)$ denote the subset obtained by removing the top-α fraction of samples with the largest predicted uncertainties.

The sparsification error compares the degradation in performance when removing samples based on uncertainty versus the oracle ranking. AUSE is defined as:

$$\mathrm{AUSE}(S) = \frac{1}{\mathrm{MAE}(S)} \int_0^1 \left(\mathrm{MAE}(S_U(\alpha)) - \mathrm{MAE}(S_V(\alpha))\right)\, d\alpha. \tag{8}$$

AUSE measures the *ranking quality* of uncertainty. Low AUSE indicates that uncertainty successfully identifies high-error samples. Compared to pure ordering metrics such as Spearman correlation, AUSE is sensitive to the *magnitude* of errors, penalizing misranking of large-error samples more strongly.

C.2 Spearman rank correlation

Spearman's rank correlation evaluates the monotonic relationship between uncertainty and prediction error:

$$\rho_S = \mathrm{Corr}_{\mathrm{rank}}\big(\sigma^2(x),\, e^2(x)\big), \tag{9}$$

where $e(x) = y - \hat{y}(x)$.

This metric measures whether uncertainty preserves the ordering of errors. It is invariant to monotonic transformations of the uncertainty scale (e.g., post-hoc calibration), and therefore isolates the *informativeness* of uncertainty independently of its magnitude.

C.3 Binned expected normalized calibration error (ENCE)

Calibration evaluates whether predicted uncertainty magnitudes match empirical errors.

We partition the test set into B bins based on predicted uncertainty. For each bin b, we compute:

$$\mathrm{RMSE}_b = \sqrt{\mathbb{E}[e^2|b]}, \qquad \mathrm{RMV}_b = \sqrt{\mathbb{E}[\sigma^2|b]}. \tag{10}$$

The Expected Normalized Calibration Error (ENCE) is defined as:

$$\mathrm{ENCE} = \frac{1}{B}\sum_{b=1}^{B}\frac{|\mathrm{RMV}_b^2 - \mathrm{RMSE}_b^2|}{\mathrm{RMV}_b^2}. \tag{11}$$

ENCE measures the *scale correctness* of uncertainty. Unlike AUSE and Spearman, it does not assess ranking quality, but instead quantifies whether predicted variances match observed squared errors.

We use $B = 15$ bins with an equal number of configurations each.

C.4 Negative log-likelihood (NLL)

NLL evaluates the probabilistic correctness of the predictive distribution:

$$\mathrm{NLL}(S) = -\frac{1}{N}\sum_{i=1}^{N}\log p(y_i|x_i). \tag{12}$$

Under the Gaussian predictive assumption:

$$\mathrm{NLL} = \frac{1}{2}\left(\frac{e^2}{\sigma^2} + \log\sigma^2\right) + C, \tag{13}$$

where C is a constant.

In our evaluation, e is the per-atom error of the ensemble mean and σ^2 the per-atom TU, i.e. the Gaussian that matches the first two moments of the ensemble mixture (Appendix B); the NLL, including the constant $\frac{1}{2}\log 2\pi$, is averaged over the configurations of a test set. NLL is a strictly proper scoring rule and jointly evaluates calibration and sharpness. However, it does not isolate ranking behavior and can be dominated by a small number of poorly calibrated predictions.

D System-level data splitting protocol

Dataset structure. The ANI-1x dataset consists of 3,114 molecular systems (distinct chemical formulas), each associated with a variable number of molecular configurations (geometries). For each configuration, multiple quantum-chemical annotations may be available, including density functional theory (DFT) energies and, for a subset, high-fidelity coupled-cluster energies (CCSD(T)/CBS).

D.1 System-OOD split

Let $\mathcal{S}$ denote the set of molecular systems. We first perform a random partition of $\mathcal{S}$ into two disjoint subsets:

$$\mathcal{S} = \mathcal{S}_{\text{seen}} \cup \mathcal{S}_{\text{OOD}}, \qquad \mathcal{S}_{\text{seen}} \cap \mathcal{S}_{\text{OOD}} = \varnothing, \tag{14}$$

where $\mathcal{S}_{\text{OOD}}$ contains a fixed proportion (40%) of systems and is reserved exclusively for *System-OOD testing*. All configurations belonging to systems in $\mathcal{S}_{\text{OOD}}$ are excluded from training and validation.

D.2 In-domain (ID) split within seen systems

The remaining systems $\mathcal{S}_{\text{seen}}$ define the in-domain (ID) regime. For each system $s \in \mathcal{S}_{\text{seen}}$, we randomly shuffle its configurations and partition them into three disjoint subsets according to fixed proportions:

$$\mathcal{X}_s = \mathcal{X}_s^{\text{train}} \cup \mathcal{X}_s^{\text{val}} \cup \mathcal{X}_s^{\text{test-ID}}, \tag{15}$$

where $\mathcal{X}_s^{\text{train}}$, $\mathcal{X}_s^{\text{val}}$, and $\mathcal{X}_s^{\text{test-ID}}$ denote the training, validation, and in-domain test configurations of system s, respectively.

Aggregating across systems yields:

$$\mathcal{X}_{\text{ID}}^{\text{train}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{train}}, \tag{16}$$

$$\mathcal{X}_{\text{ID}}^{\text{val}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{val}}, \tag{17}$$

$$\mathcal{X}_{\text{ID}}^{\text{test}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{test-ID}}. \tag{18}$$

This procedure ensures that: (i) no configuration leakage occurs across splits, and (ii) the in-domain test set contains unseen configurations from the same distribution of molecular systems as training.

D.3 Annotation-specific dataset construction

From each configuration split, we construct task-specific datasets:

- **DFT datasets:** configurations with available DFT annotations are selected from each split, optionally using stratified sampling to prevent over-representation of large systems.
- **CC datasets:** high-fidelity CC configurations are selected *only from the already selected DFT configurations*, ensuring
$$\mathcal{X}_{\text{CC}} \subseteq \mathcal{X}_{\text{DFT}}, \tag{19}$$
which guarantees consistency for multi-fidelity pretraining and fine-tuning.

D.4 Stratified sampling across molecular systems

After defining the system-level splits, we optionally subsample configurations within each split to reach fixed target sizes (e.g., $n_{\text{DFT}}^{\text{train}}$, $n_{\text{DFT}}^{\text{val}}$). To avoid over-representation of large systems (some of which contain orders of magnitude more configurations than others) we employ a *stratified sampling scheme at the system level*.

Available pools. For a given split $\sigma \in \{\text{train}, \text{val}, \text{test-ID}, \text{test-OOD}\}$ and theory level (DFT or CC), let

$$\mathcal{X}_{s,\sigma} \tag{20}$$

denote the set of configurations of system s that: (i) belong to split σ, and (ii) have a valid annotation for the selected theory level. Let

$$n_{s,\sigma} = |\mathcal{X}_{s,\sigma}| \tag{21}$$

be the number of available candidates for system s in split σ. The total number of available configurations in that split is

$$N_\sigma = \sum_s n_{s,\sigma}. \tag{22}$$

Quota allocation. Given a desired target size $N_\sigma^{\text{target}} \leq N_\sigma$, we allocate per-system quotas $q_{s,\sigma}$ proportionally to the number of available configurations:

$$\tilde{q}_{s,\sigma} = N_\sigma^{\text{target}} \frac{w_{s,\sigma}}{\sum_{s'} w_{s',\sigma}}, \tag{23}$$

where the weights are defined as

$$w_{s,\sigma} = \begin{cases} n_{s,\sigma}, & \text{if no cap is applied,} \\ \min(n_{s,\sigma}, \kappa), & \text{with optional cap } \kappa. \end{cases} \tag{24}$$

The optional cap κ limits the influence of very large systems while still allowing them to contribute more samples when necessary.

The final integer quotas $q_{s,\sigma}$ are obtained by:

1. taking the floor of $\tilde{q}_{s,\sigma}$, and

2. distributing the remaining samples to systems with the largest fractional remainders,

ensuring that

$$\sum_s q_{s,\sigma} = N_\sigma^{\text{target}}. \tag{25}$$

Sampling without replacement. For each system s, we then draw $q_{s,\sigma}$ configurations uniformly at random without replacement from $\mathcal{X}_{s,\sigma}$. This yields the final stratified subset

$$\widehat{\mathcal{X}}_\sigma = \bigcup_s \widehat{\mathcal{X}}_{s,\sigma}, \tag{26}$$

with $|\widehat{\mathcal{X}}_\sigma| = N_\sigma^{\text{target}}$.

Multi-fidelity constraint. For the high-fidelity CC dataset, candidate configurations are first restricted to those already selected in the corresponding DFT split, enforcing

$$\mathcal{X}_\sigma^{\text{CC}} \subseteq \widehat{\mathcal{X}}_\sigma^{\text{DFT}}. \tag{27}$$

Stratified quota allocation is then applied within this restricted pool, guaranteeing consistency between low- and high-fidelity datasets.

Properties. This procedure:

- preserves the system-level split,
- prevents dominance of large systems,
- enforces exact target sizes,
- maintains CC $\subseteq$ DFT,
- and is reproducible given the split seed (we use five seeds, 0–4).

D.5 Resulting evaluation regimes

This protocol produces four disjoint evaluation settings:

- **Train (ID)**: configurations from $\mathcal{S}_{\text{seen}}$ used for model optimization.
- **Validation (ID)**: configurations from $\mathcal{S}_{\text{seen}}$ used for checkpoint selection and for the post-hoc recalibration of Appendix I.3.
- **Test-ID**: unseen configurations from $\mathcal{S}_{\text{seen}}$.
- **Test-OOD (System-OOD)**: configurations from $\mathcal{S}_{\text{OOD}}$, representing novel molecular systems.

By construction, the System-OOD evaluation measures generalization to entirely unseen molecular systems, while the Test-ID split quantifies generalization within the same system distribution without configuration leakage. This split design enables a controlled comparison between in-domain performance and true structural distributional shift.

E Energy-based data splitting protocol

To evaluate generalization across energetic regimes within each molecular system, we construct an energy-based split that separates low-energy (in-domain) configurations from high-energy (out-of-domain) configurations.

Per-system relative energies. For each molecular system s, let $\mathcal{X}_s$ denote its full set of configurations. Given an energy annotation $E_s(x)$ (by default the DFT energy), we compute relative energies

$$\Delta E_s(x) = E_s(x) - \min_{x' \in \mathcal{X}_s} E_s(x'). \tag{28}$$

Relative energies are computed independently for each system to remove absolute offsets.

Empirical energy quantiles. Let $\{x_i\}_{i=1}^{N_s}$ be the configurations of system s with valid energies. We assign to each configuration an empirical quantile rank

$$q_s(x_i) = \frac{\mathrm{rank}(\Delta E_s(x_i))}{N_s - 1}, \tag{29}$$

where ranks are computed after sorting ΔE_s in ascending order. This yields $q_s(x) \in [0, 1]$ for all valid configurations.

Energy domains. Given two thresholds $q_{\text{low}} < q_{\text{high}}$ (we use 0.5 and 0.55), we define:

$$\mathcal{X}_s^{\text{low}} = \{x \in \mathcal{X}_s : q_s(x) \le q_{\text{low}}\}, \tag{30}$$

$$\mathcal{X}_s^{\text{high}} = \{x \in \mathcal{X}_s : q_s(x) \ge q_{\text{high}}\}. \tag{31}$$

Configurations with $q_{\text{low}} < q_s(x) < q_{\text{high}}$ form an intermediate band that is excluded from all splits to ensure a clear separation between in-domain and out-of-domain samples.

Low-energy ID splits. For each system, the low-energy set $\mathcal{X}_s^{\text{low}}$ is randomly shuffled and partitioned into:

$$\mathcal{X}_s^{\text{low}} = \mathcal{X}_s^{\text{train}} \cup \mathcal{X}_s^{\text{val}} \cup \mathcal{X}_s^{\text{test-ID}}, \tag{32}$$

according to fixed proportions $(p_{\text{train}}, p_{\text{val}}, p_{\text{test-ID}})$. Aggregating across systems yields:

$$\mathcal{X}_{\text{ID}}^{\text{train}} = \bigcup_s \mathcal{X}_s^{\text{train}}, \tag{33}$$

$$\mathcal{X}_{\text{ID}}^{\text{val}} = \bigcup_s \mathcal{X}_s^{\text{val}}, \tag{34}$$

$$\mathcal{X}_{\text{ID}}^{\text{test-ID}} = \bigcup_s \mathcal{X}_s^{\text{test-ID}}. \tag{35}$$

Energy-OOD set. All high-energy configurations are assigned to the OOD pool:

$$\mathcal{X}^{\text{test-OOD}} = \bigcup_s \mathcal{X}_s^{\text{high}}. \tag{36}$$

Theory-specific datasets. For each split $\sigma \in \{\text{train}, \text{val}, \text{test-ID}, \text{test-OOD}\}$:

- DFT configurations are selected from $\mathcal{X}_\sigma$ using stratified sampling across systems.
- CC configurations are selected exclusively from the corresponding DFT subset, enforcing

$$\mathcal{X}_\sigma^{\text{CC}} \subseteq \mathcal{X}_\sigma^{\text{DFT}}. \tag{37}$$

Properties. This protocol:

- defines energy shift independently within each molecular system,
- avoids configuration leakage across splits,
- enforces a strict low/high separation,
- preserves per-system diversity via stratified sampling,
- guarantees CC $\subseteq$ DFT,
- and is reproducible given the split seed (we use five seeds, 0–4).

F Training details and hyperparameters

Data. Low-fidelity labels are the ANI-1x ωB97x/def2-TZVPP energies (`wb97x_tz.energy`) and high-fidelity labels the CCSD(T)/CBS energies (`ccsd(t)_cbs.energy`). Forces are not used at either fidelity. Table 1 lists the number of configurations per split; each CC set is a subset of the corresponding DFT set. The system split holds out 40% of the molecular systems ($p_{\text{seen}} = 0.6$) and divides the configurations of seen systems into train, validation and test-ID with proportions $0.6/0.2/0.2$. The energy split uses $q_{\text{low}} = 0.5$ and $q_{\text{high}} = 0.55$ with the same proportions. Stratified sampling uses a per-system soft cap $\kappa = 64$. The whole construction is repeated with five split seeds (0–4).

Table 1: Number of configurations per split for the DFT and CC data sets, identical for both shifts and all five split seeds.

Split	DFT	CC
Train	50,000	5,000
Validation	10,000	1,000
Test-ID	50,000	5,000
Test-OOD	50,000	5,000

Model. All models are MACE networks with 128 channels, $L_{\max} = 1$ equivariant features, a 6 Å cutoff, two interaction layers of correlation order three and spherical harmonics up to $\ell = 3$, with the mean–variance readout of Section 2.2 and atomic reference energies fitted to the training set by least squares. The readout outputs, for every atom, an energy and an unconstrained scalar that a softplus maps to a positive variance, plus 10^{-12} for numerical stability; the per-atom variances are summed over the configuration and multiplied by the square of the model's energy scale factor, while the reference energies carry no variance. Each ensemble has $M = 10$ members that differ only in their random seed (0–9).

Optimization. Models are trained on the Gaussian NLL of the energy, with the term of each configuration divided by its number of atoms, using Adam (AMSGrad), batch size 64, weight decay $5 \cdot 10^{-7}$ and a reduce-on-plateau learning-rate schedule (factor 0.8, patience 50 epochs), without stochastic weight averaging or an exponential moving average of the weights. LF-only and HF-only models are trained from scratch for 300 epochs with learning rate 10^{-2}. LF→HF models are initialized from the weights of the LF-only member with the same split and seed after its last pretraining epoch, including the mean–variance readout, and fine-tuned on CC energies for 100 epochs with learning rate 10^{-3}; the atomic reference energies are refitted on the CC training set, and the optimizer and learning-rate schedule start afresh. For every member, the checkpoint with the lowest validation NLL is kept; test sets are never used for model selection. Across the 50 members of each experiment, the selected epoch has a median of 60 (range 30–111) for HF-only under system shift and 67 (34–111) under energy shift, and a median of 2 (range 1–3) fine-tuning epochs for LF→HF under both shifts.

F.1 Computational resources

All experiments were conducted on a high-performance computing node equipped with two AMD EPYC 7742 CPUs (128 cores total, 256 threads) with 1TB of total RAM, and 8 NVIDIA A100 GPUs (40GB memory each).

Model training and evaluation were performed using GPU acceleration. Training the 450 ensemble members reported in the paper took about 1,900 GPU hours in total, dominated by the 100 low-fidelity pretraining runs, each of which takes about 15–16 hours on one A100 GPU and occupies about 20GB of GPU memory; high-fidelity training from scratch takes 1–1.5 hours and fine-tuning about half an hour. Per-epoch evaluation and the acquisition experiment add to this budget. The use of deep ensembles, involving multiple independently initialized models, constitutes the primary computational cost of the study.

Data preprocessing and auxiliary computations were performed on CPU resources.

The full research project required more computing than the experiments reported in the paper due to preliminary or failed experiments.

G Training-support filter

Energy-OOD contains a small number of configurations compressed far below the repulsive wall covered by training. For each split we compute the smallest interatomic distance occurring anywhere in that split's DFT and CC training and validation sets, and we exclude every test configuration whose smallest interatomic distance falls below it. The rule reads geometry only: it never inspects

predictions, errors or uncertainties, and it removes the same configurations for every protocol, signal and metric within a split. It removes 13–16 of the 5,000 Energy-OOD configurations per split, at most two Energy-ID configurations, and none on the system split, whose training data already cover compressed geometries.

The filter matters because all our metrics average squared errors or variances. On the excluded configurations the models extrapolate very poorly, and a single one of them can dominate the RMSE, the ENCE of a bin or the AUSE of a whole test set. We therefore treat them as outside the domain of validity of the models rather than as evidence about the quality of uncertainty. The same rule is applied to the acquisition experiment (Appendix J).

H TRAINING DYNAMICS

This section complements Sections 4.2 and 4.3 with the full per-epoch trajectories of every metric, on the training, ID and OOD sets side by side: HF-only training (Figures 7 and 9) and the whole LF→HF trajectory (Figures 8 and 10), under system and energy shift. The Energy-OOD columns use the training-support filter; the training and ID curves are computed on all configurations, of which the filter would remove at most two per split. Figure 11 gives the training and validation curves of every experiment.

Four regularities recur. First, ranking on held-out data settles within a few dozen epochs, while calibration keeps changing: during HF-only training the TU calibration error on the ID and OOD sets passes through a minimum and then rises for the rest of the run, driven by the aleatoric component, and during fine-tuning the aleatoric calibration error rises from the first epoch. Second, on the training set the calibration error of every signal stays close to, and below, one throughout, without the rise seen on held-out data. Third, on the training set RMSE and NLL keep decreasing, whereas on the ID and OOD sets they flatten early; the gap opens most clearly during CC fine-tuning, where the training error keeps falling while the held-out error barely moves. This widening gap between training and held-out error is plausibly what the variance head, fitted to the training residuals, fails to capture. Fourth, the validation loss (Figure 11) keeps improving over 300 epochs for LF-only training, reaches a minimum after a few dozen epochs for HF-only training and then slowly worsens, and stops improving within the first few epochs of fine-tuning while the training loss keeps falling, which is why validation-based selection keeps such early fine-tuning checkpoints (Appendix F). System-split pretraining also shows a transient instability around epochs 50–60 in two of the five splits, visible in every metric and in the EU/AU ratio of Appendix H.1.

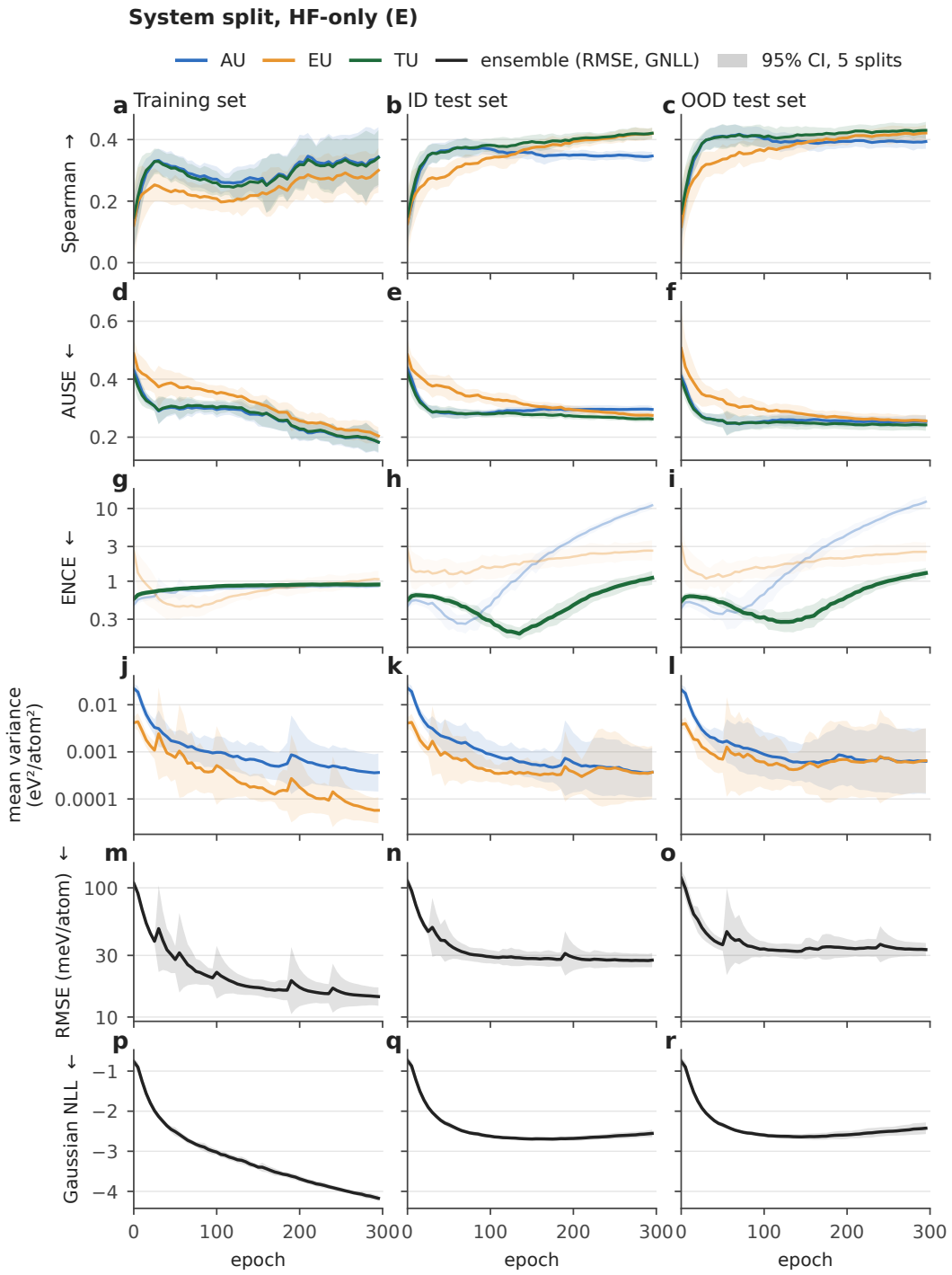

Figure 7: **Training dynamics of HF-only ensembles under system shift.** Columns: training set, ID and OOD test set. Rows: Spearman (↑), AUSE (↓), ENCE (↓, TU highlighted, AU and EU faded), mean per-atom AU and EU, RMSE of the ensemble mean and Gaussian NLL of the predictive distribution, against epoch. Lines are exponential moving averages of the five-split means (geometric on log axes); bands are 95% intervals.

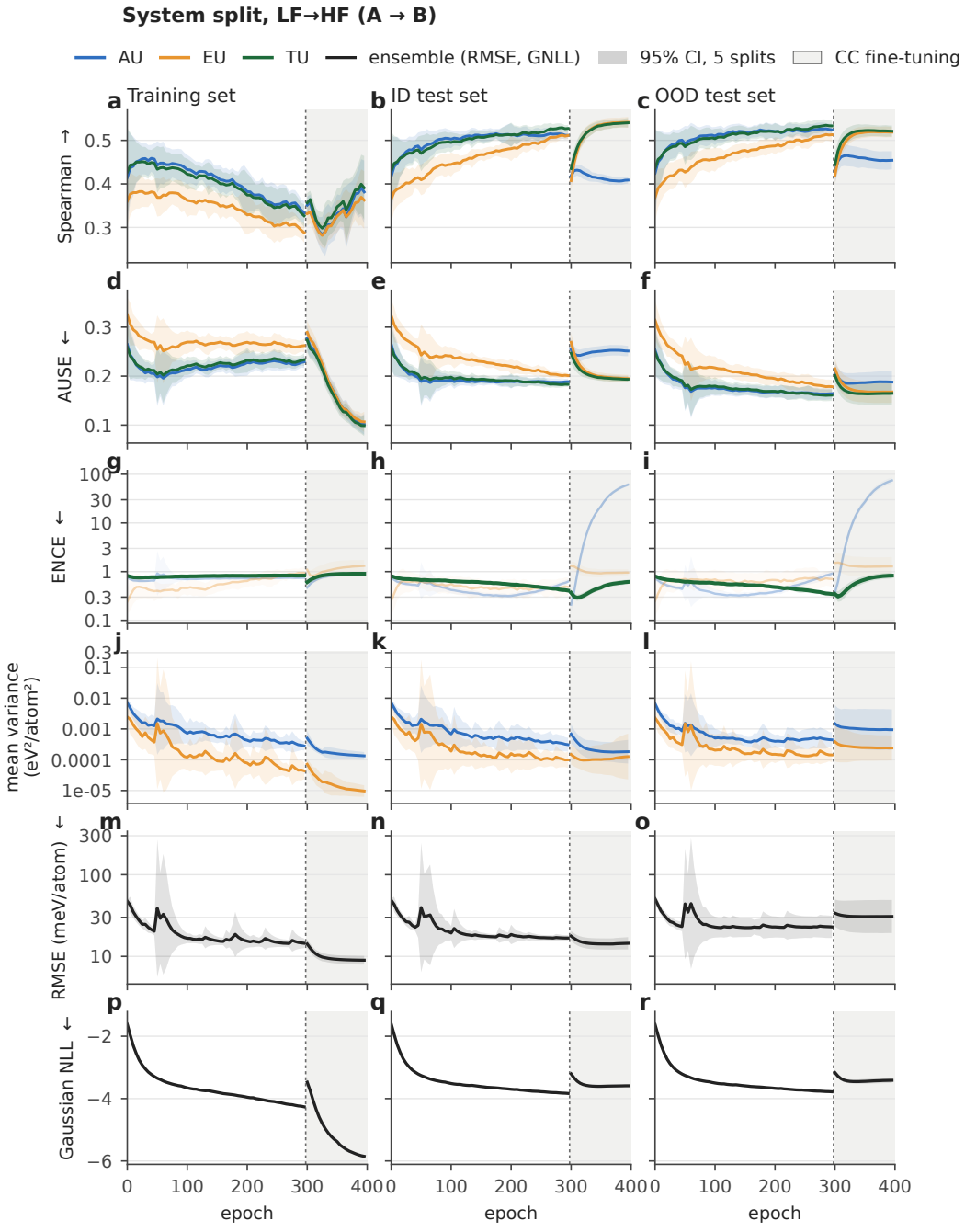

Figure 8: **Training dynamics of LF→HF ensembles under system shift.** As Figure 7. Epochs 0–299 are DFT pretraining (the LF-only ensemble, DFT labels), epochs 300–399 (shaded) CC fine-tuning; RMSE and NLL jump at epoch 300 partly because the target changes.

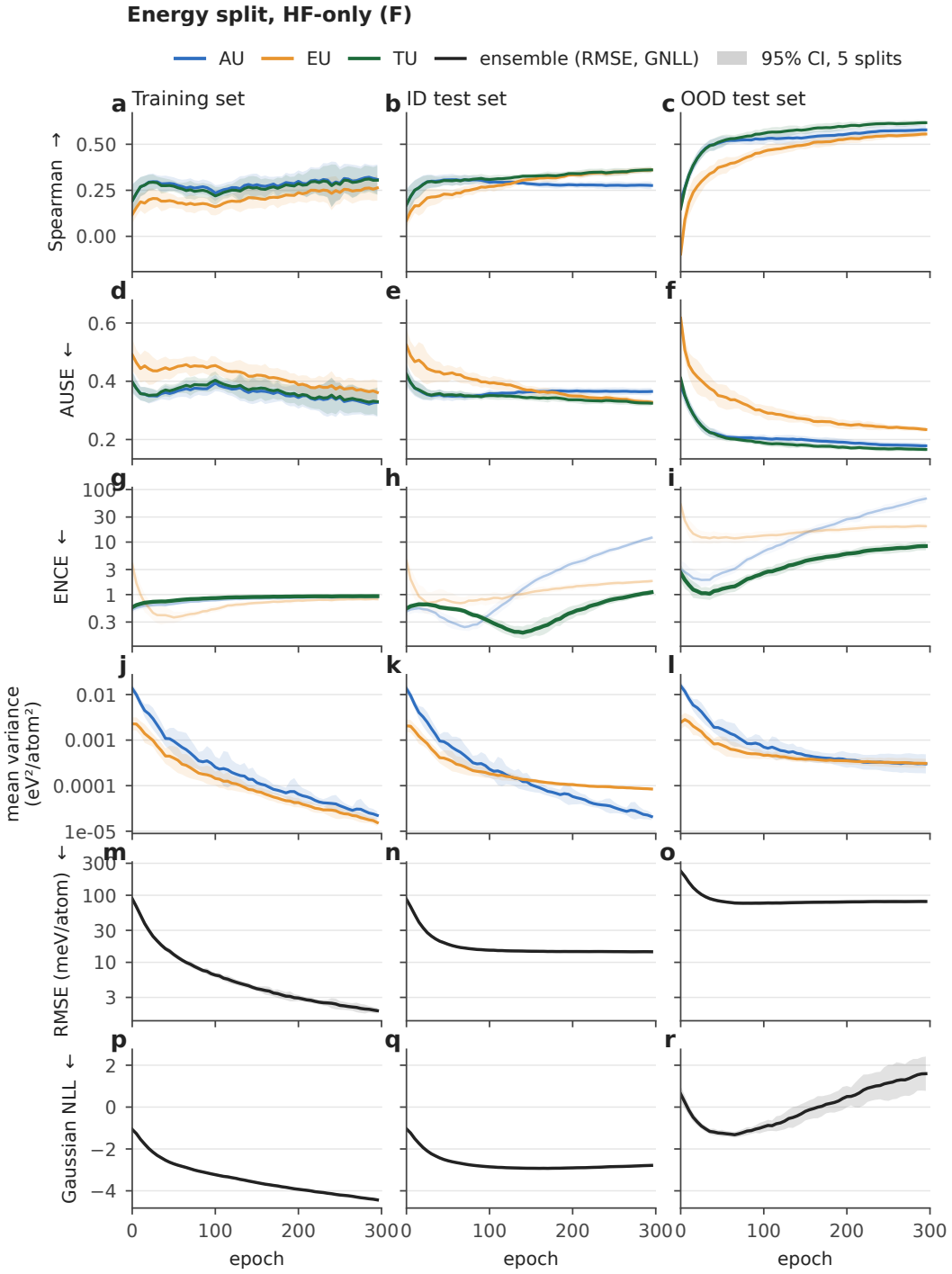

Figure 9: **Training dynamics of HF-only ensembles under energy shift.** As Figure 7; the Energy-OOD column is support-filtered.

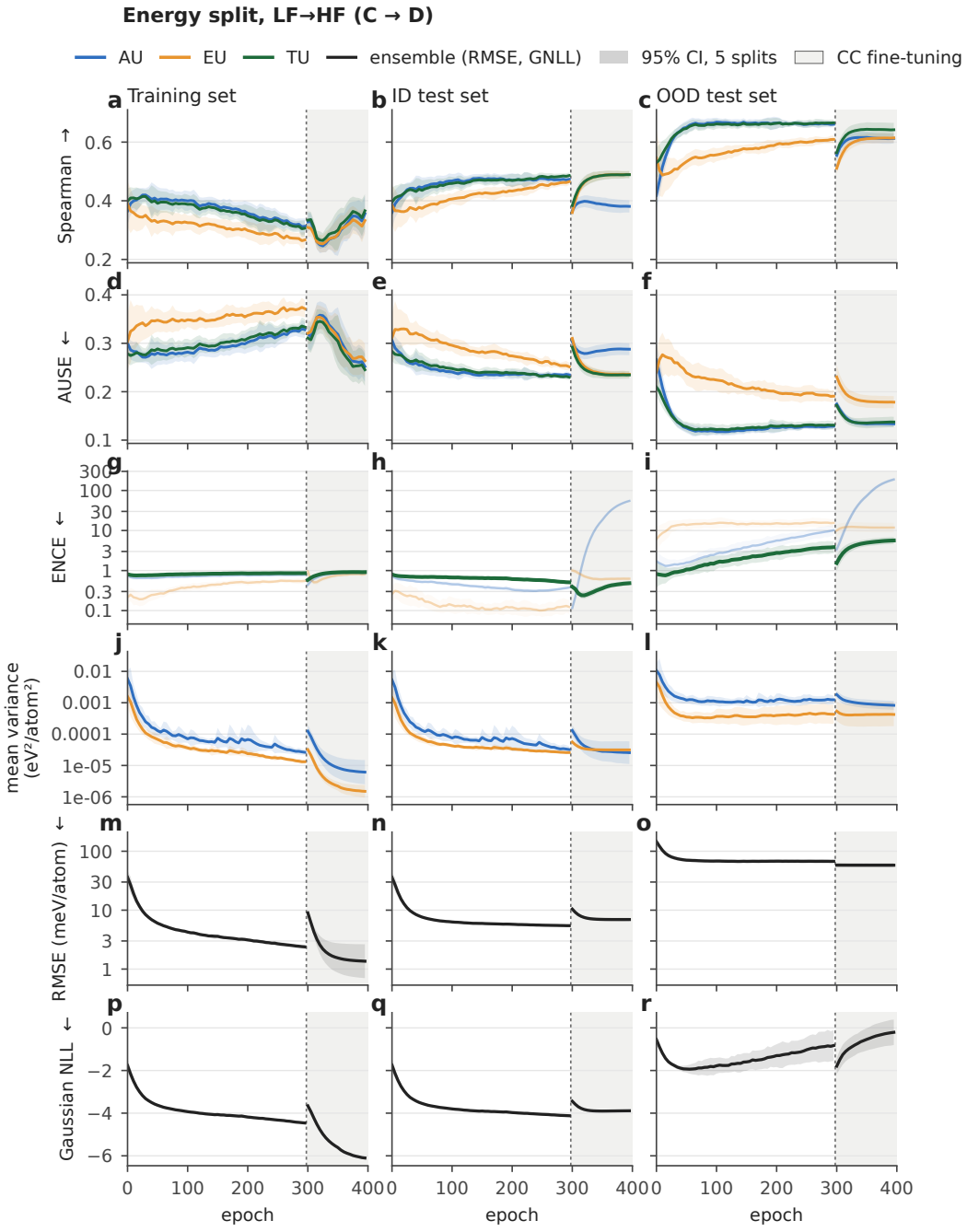

Figure 10: **Training dynamics of LF→HF ensembles under energy shift.** As Figure 8; the Energy-OOD column is support-filtered and, during pretraining, scored against DFT labels.

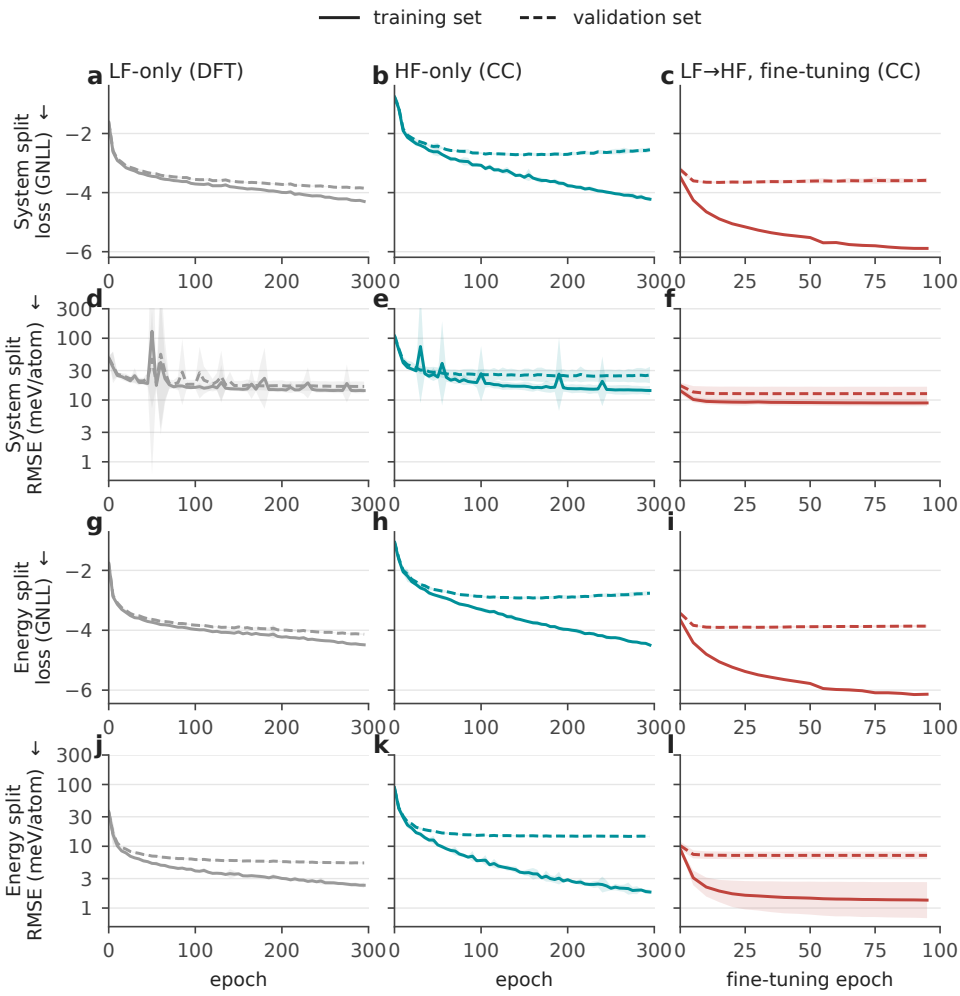

Figure 11: **Training and validation curves.** Training loss (Gaussian NLL, solid) and validation loss (dashed), and the corresponding RMSE, for LF-only (DFT), HF-only (CC) and the fine-tuning stage of LF→HF (CC), under system (top two rows) and energy shift (bottom two rows). Loss: five-split arithmetic means; RMSE: geometric means; bands are 95% intervals. For HF-only and LF→HF the validation loss flattens or rises while the training loss keeps falling, which is why validation-based selection keeps early fine-tuning checkpoints.

H.1 Balance between epistemic and aleatoric uncertainty

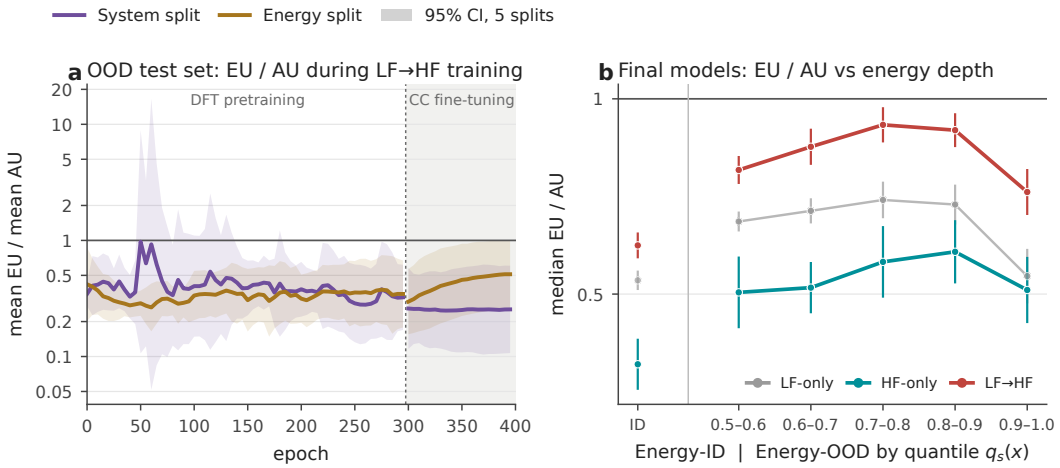

Figure 12: **EU is smaller than AU at every energy depth and for nearly all of training. (a)** Mean EU over mean AU on the OOD test sets during LF→HF training, DFT pretraining (epochs 0–299) then CC fine-tuning (300–399, shaded); bands are 95% intervals. **(b)** Converged energy-split models: median per-configuration EU/AU on Energy-ID and in within-system energy-quantile bins of Energy-OOD, for LF-only (grey, DFT labels), HF-only and LF→HF.

Figure 12a tracks the ratio of mean EU to mean AU on the OOD test sets. Apart from two early pretraining epochs on the system split, where the five-split mean exceeds one with an interval spanning several orders of magnitude, it stays below one. It does not grow during pretraining (energy split $0.42 \rightarrow 0.35$, system split $0.35 \rightarrow 0.33$), and changes only during fine-tuning, rising to 0.51 on the energy split and settling near 0.26 on the system split. The wide system-split band during pretraining reflects split-to-split variation in a ratio of two small means. Figure 12b uses the median per-configuration ratio, which is robust to the few configurations with extreme EU. The median EU/AU stays below one in every bin and for every protocol. It rises from Energy-ID to the middle of the tail (LF→HF $0.59 \rightarrow 0.91$, HF-only $0.39 \rightarrow 0.58$, LF-only $0.53 \rightarrow 0.70$) and falls again in the highest-energy bin (0.72, 0.51 and 0.53). LF→HF has the largest EU share in every bin, above both single-fidelity baselines.

I Supplementary results for the converged models

This section complements Section 4.1. Statistics follow Section 3: five-split means with two-sided 95% Student-t intervals, geometric means with log-space intervals for ENCE and RMSE, and the training-support filter applied throughout.

I.1 Reliability diagrams

Figures 13 and 14 show the reliability diagrams of every converged ensemble, including the LF-only ensembles evaluated against DFT labels, from which the calibration results of Section 4.1 are computed. On the system split the bands are wide in the few highest-variance bins, which hold a handful of large-error configurations per split; on the energy split they are narrow throughout.

Under system shift (Figure 13), the ID and OOD diagrams are nearly indistinguishable for every protocol: TU follows the diagonal or lies slightly below it. Of its components, AU alone behaves similarly, whereas EU alone lies above the diagonal, as expected of one part of the predictive variance. Under energy shift (Figure 14), TU stays close to the diagonal in domain, while on Energy-OOD it lies far above it for every protocol, and so do both of its components. This includes the LF-only ensembles, which suggests that overconfidence under energy extrapolation is not specific to high-fidelity training; high-fidelity training adds the progressive degradation of the aleatoric term (Section 4.3).

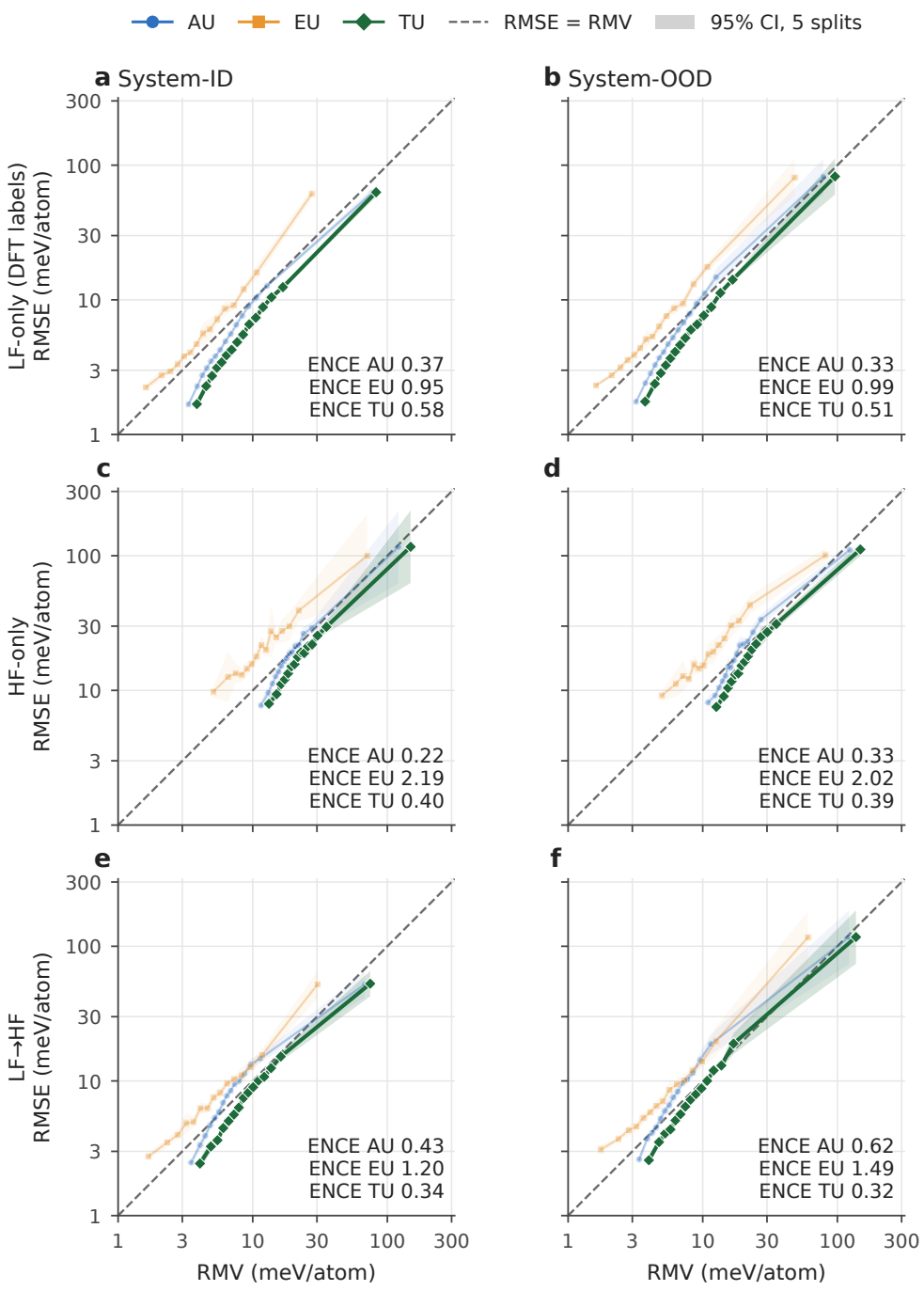

Figure 13: **Reliability of the converged ensembles under system shift, before recalibration.** Rows: LF-only (DFT labels), HF-only and LF→HF (CC labels); columns: ID and OOD test set. In each split, the support-filtered test configurations are sorted by each signal and cut into 15 equal-count bins; each point is the five-split geometric mean of the root mean predicted variance (RMV, x) and of the RMSE (y) of one bin, per atom, on common log-log axes. Bands are 95% log-space intervals of the RMSE over five splits (RMV intervals not drawn); the dashed diagonal marks perfect calibration, and points above it are overconfident. The corner text gives the five-split geometric-mean ENCE of each signal.

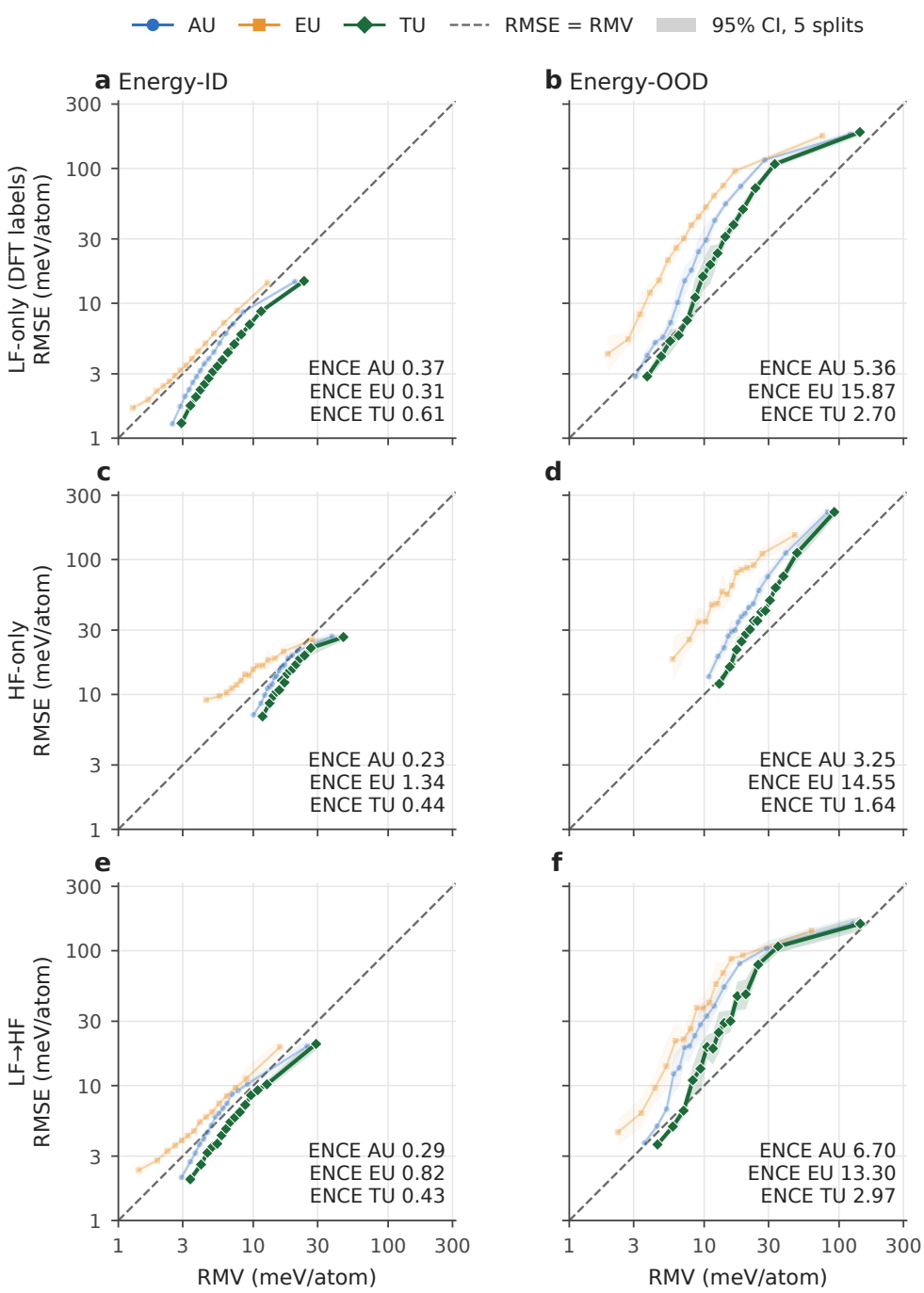

Figure 14: **Reliability of the converged ensembles under energy shift, before recalibration.** As Figure 13.

I.2 Rank correlation of the converged models

Table 2 reports the Spearman rank correlation between predicted uncertainty and squared error for the converged models of Figure 2. It agrees with AUSE: LF→HF has a higher correlation than HF-only for every signal on every test set, EU is the weakest signal without pretraining, and after pretraining EU exceeds AU on System-ID, System-OOD and Energy-ID. Both protocols rank Energy-OOD errors better than Energy-ID errors, as in Figure 3.

Table 2: Spearman rank correlation (↑) between uncertainty and squared error of the converged models, five-split mean ± 95% half-width.

	HF-only			LF→HF		
Test set	AU	EU	TU	AU	EU	TU
System-ID	$.365 \pm .013$	$.338 \pm .010$	$.377 \pm .012$	$.463 \pm .009$	$.492 \pm .013$	$.500 \pm .009$
System-OOD	$.402 \pm .045$	$.370 \pm .040$	$.411 \pm .044$	$.490 \pm .025$	$.501 \pm .013$	$.518 \pm .019$
Energy-ID	$.310 \pm .020$	$.255 \pm .020$	$.315 \pm .022$	$.420 \pm .024$	$.449 \pm .015$	$.457 \pm .019$
Energy-OOD	$.533 \pm .022$	$.445 \pm .015$	$.556 \pm .018$	$.626 \pm .018$	$.589 \pm .019$	$.640 \pm .019$

I.3 Post-hoc isotonic recalibration

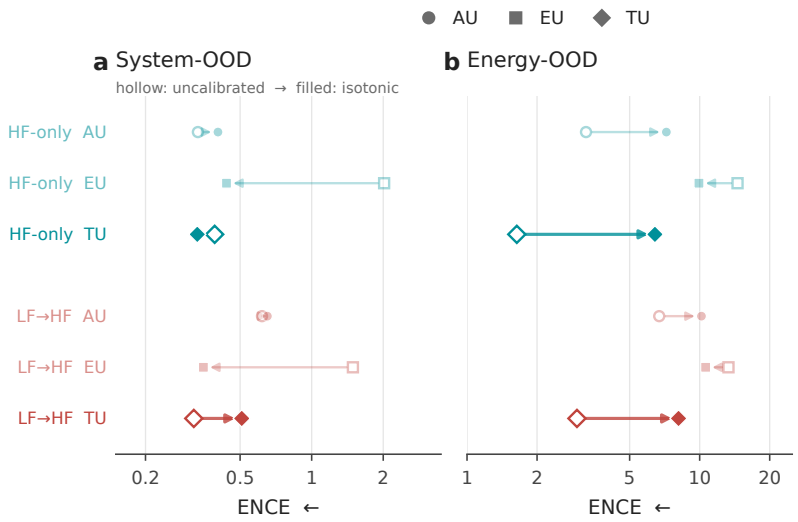

Figure 15: **Isotonic recalibration fitted in domain transfers poorly to energy extrapolation.** ENCE of the converged models on **(a)** System-OOD and **(b)** Energy-OOD before (hollow) and after (filled) isotonic recalibration fitted on the in-domain validation set, per protocol and signal.

For each signal we fit an isotonic regression from predicted variance to squared error on the ensemble's validation predictions in each split (Kuleshov et al., 2018; Busk et al., 2023), and apply it to the test sets. The map is monotone and is held constant beyond the range of validation variances. Recalibration changes only the predicted variances: the ensemble mean, and hence the RMSE, are those of Figure 2, and AUSE and Spearman change only through the ties that the piecewise-constant map introduces. In domain it works as intended, reducing TU ENCE from 0.444 to 0.196 (HF-only) and from 0.427 to 0.150 (LF→HF) on Energy-ID, and from 0.401 to 0.250 and from 0.342 to 0.174 on System-ID. Figure 15 shows the OOD test sets. Under system shift, recalibration repairs the badly scaled epistemic signal ($2.02 \to 0.44$ for HF-only, $1.49 \to 0.35$ for LF→HF), with mixed effects on TU ($0.391 \to 0.330$ for HF-only, $0.319 \to 0.508$ for LF→HF). Under energy extrapolation it inflates the calibration error of AU ($3.25 \to 7.19$ and $6.70 \to 10.2$) and TU ($1.64 \to 6.42$ and $2.97 \to 8.11$) for both protocols, and improves EU only slightly ($14.5 \to 9.93$ and $13.3 \to 10.6$). Because the map preserves the ordering up to ties, ranking is essentially unchanged (Energy-OOD TU Spearman $0.556 \to 0.547$ for HF-only and $0.640 \to 0.632$ for LF→HF). The failure appears to be one of scale: high-energy errors exceed anything seen in the validation set, and a map that is constant beyond the validation range cannot extend the variance scale to them.

Figures 16 and 17 show the reliability diagrams after recalibration. In domain, all three signals of every protocol collapse onto the diagonal, and on System-OOD they stay close to it. On Energy-OOD the three signals become nearly indistinguishable but remain far above the diagonal, because the recalibrated variances cannot exceed the largest value seen on the validation set.

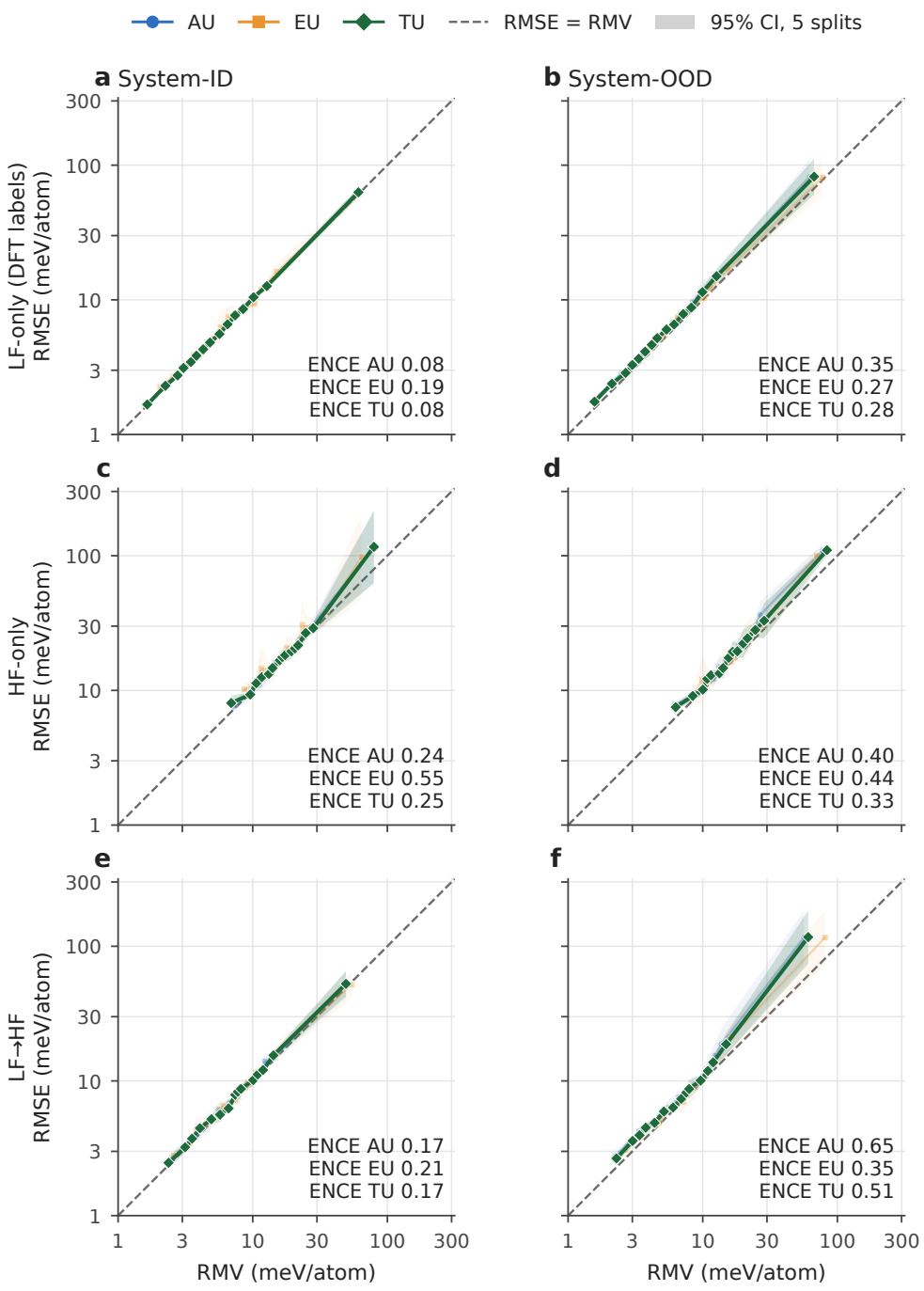

Figure 16: **Reliability of the converged ensembles under system shift, after isotonic recalibration.** Rows: LF-only (DFT labels), HF-only and LF→HF (CC labels); columns: ID and OOD test set. In each split, the support-filtered test configurations are sorted by each signal and cut into 15 equal-count bins; each point is the five-split geometric mean of the root mean predicted variance (RMV, x) and of the RMSE (y) of one bin, per atom, on common log-log axes. Bands are 95% log-space intervals of the RMSE over five splits (RMV intervals not drawn); the dashed diagonal marks perfect calibration, and points above it are overconfident. The corner text gives the five-split geometric-mean ENCE of each signal. The isotonic map of each signal is fitted on the in-domain validation set of each split.

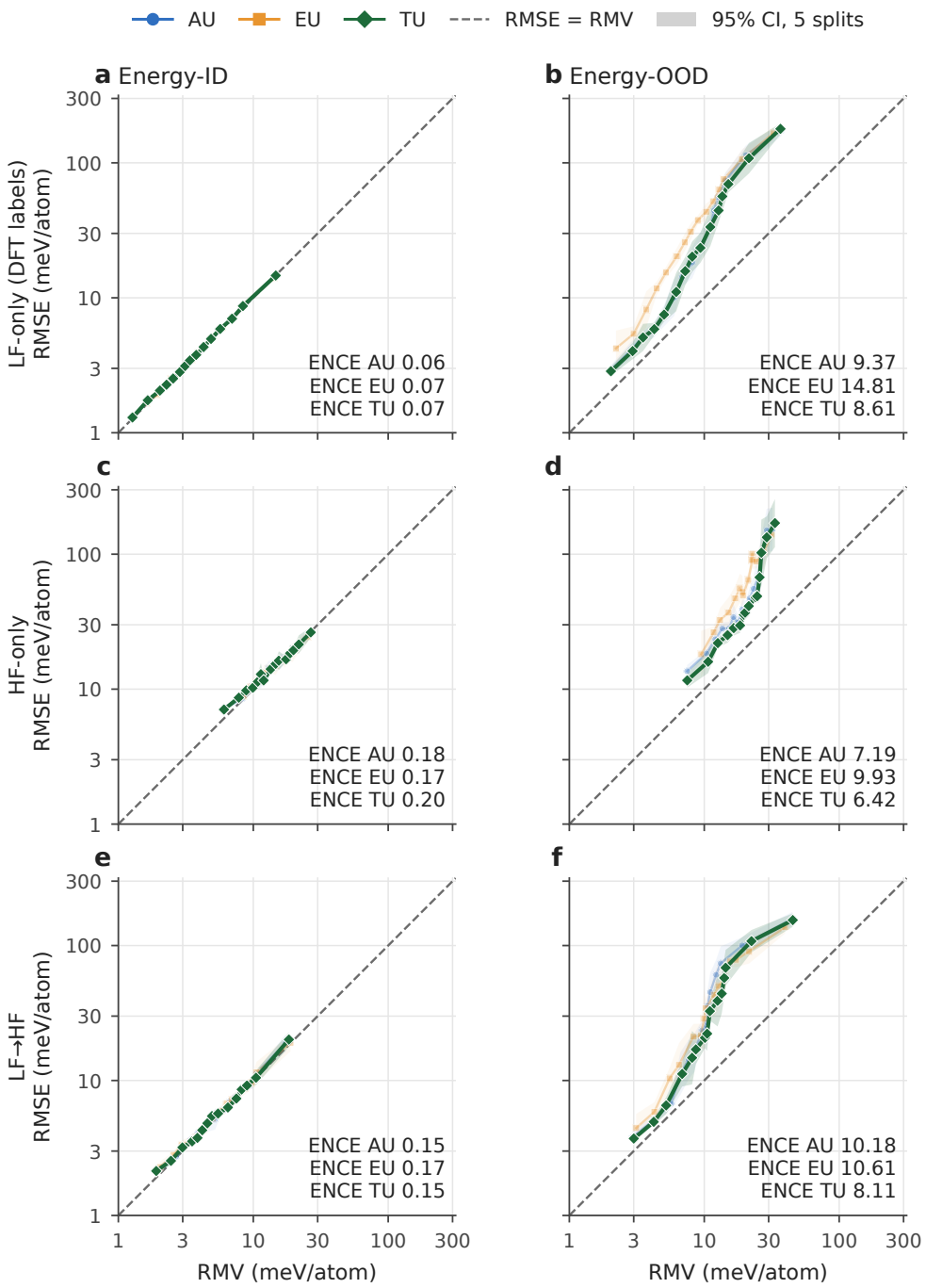

Figure 17: **Reliability of the converged ensembles under energy shift, after isotonic recalibration.** As Figure 16.

J ACQUISITION EXPERIMENT

Protocol. For each split, the 5,000-configuration Energy-OOD CC test set is divided, stratified by molecular system, into an acquisition pool and a held-out test set of 2,500 configurations each; the held-out half is not a higher-energy half. The pool labels are hidden. Each strategy selects 500 pool configurations: uniformly at random (one batch per split, shared by both protocols), or the 500 with the largest AU, EU or TU under the converged HF-only or LF→HF ensemble. Every member of the

corresponding ensemble is then warm-started from its converged checkpoint, with a fresh optimizer and learning-rate schedule, and trained for 50 additional epochs with learning rate 10^{-3} on its CC training set plus the 500 acquired configurations. Checkpoints are selected on the original validation set. Gains relative to random selection are computed within each split before averaging over splits. The training-support filter (Appendix G) removes 6–10 held-out configurations per split. In addition, one held-out configuration of one split, on which the LF→HF ensemble extrapolated catastrophically, was removed after inspection; that split therefore has 2,499 held-out configurations before filtering.

Results. Figure 6a shows the RMSE reduction on the held-out configurations relative to random selection. Uncertainty-guided acquisition beats random selection for all six combinations of protocol and signal. The benefit is larger after pretraining, roughly 2.3–2.5 meV/atom against 0.8–1.6 meV/atom without it, and the difference between the two protocols is resolved for every signal. Without pretraining, EU is the best signal, ahead of AU by 0.80 ± 0.55 and of TU by 0.46 ± 0.41 meV/atom (paired over splits); after pretraining the three signals are indistinguishable. Figure 6b shows why uncertainty helps: every signal concentrates its batch on the high-energy end of each molecule's range, where the model is weakest.

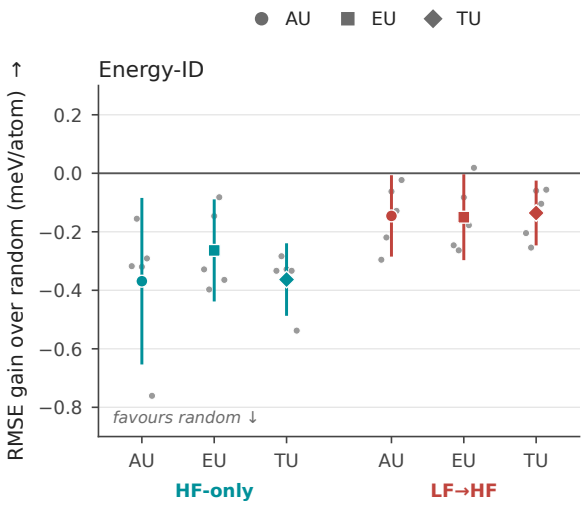

Figure 18: **Uncertainty-guided acquisition from the Energy-OOD pool costs little in domain.** Energy-ID RMSE reduction relative to random selection of the same budget, for the acquisition experiment of Section 4.4; negative values favour random selection. Markers are five-split means with 95% intervals, grey dots the individual splits.

In-domain cost. Figure 18 evaluates the acquisition experiment of Section 4.4 on Energy-ID. All acquired configurations come from the Energy-OOD pool, and continued training improves Energy-ID RMSE for every strategy (random selection: $15.60 \to 15.06$ meV/atom for HF-only and $7.86 \to 7.42$ meV/atom for LF→HF). Uncertainty-guided batches improve it slightly less than random ones, consistent with their concentration on the highest-energy configurations (Figure 6b), which are furthest from the in-domain region. The cost relative to random selection is 0.37 [0.08, 0.65] (AU), 0.26 [0.09, 0.44] (EU) and 0.36 [0.24, 0.49] meV/atom (TU) for HF-only, and 0.15 [0.01, 0.28], 0.15 [0.00, 0.30] and 0.14 [0.02, 0.25] meV/atom for LF→HF. After pretraining the cost is smaller, by 0.23 [0.02, 0.44] meV/atom for TU, while for AU and EU the difference between protocols is not resolved with five splits.

K Limitations and Future Work

We focus only on energy prediction. While energies are central to atomistic modeling and provide a clean setting for uncertainty analysis, practical molecular simulations also require accurate and reliable forces. Extending the analysis to force uncertainty is an important direction for future work,

especially because force errors are local, vector-valued, and directly affect molecular dynamics stability.

We use mean–variance deep ensembles as our uncertainty estimator. This choice provides a strong, theoretically founded, and widely used baseline for decomposing predictive uncertainty into aleatoric and epistemic components, but it is computationally expensive because it requires training and evaluating multiple models. Future work should investigate whether cheaper alternatives, such as shallow ensembles, last-layer ensembles, multi-head models, evidential methods, or approximate Bayesian approaches, reproduce the same uncertainty dynamics observed here. We also do not ablate the parts of the estimator, for example by training ensembles with a squared loss, using a single mean–variance model, or resetting the variance head at the start of fine-tuning.

Our analysis relies on operational definitions. With deterministic labels, there is no ground truth for the aleatoric and epistemic terms separately, so we judge calibration on TU and use the ENCE of AU and EU only as a diagnostic (Appendix B). The comparison between HF-only and LF→HF is one of data access: LF→HF also sees the DFT labels, including the geometries of the CC training set, and we do not separate the effect of this additional information from that of the change of fidelity. Likewise, the two shifts use different training sets, so we compare each OOD test set with the ID test set of its own split rather than the two shifts with each other in absolute terms.

Our main experiments are based on ANI-1x and on controlled system-level and energy-based splits. This design is useful for isolating the effect of data fidelity and distributional shift, but it does not capture the full complexity of realistic simulation workflows, where chemical composition, molecular size, temperature, conformation, and energy can shift simultaneously. Evaluating the same questions on larger and more diverse datasets, including materials and biomolecular systems, would test the generality of our conclusions.

L Broader impact

This work is primarily methodological and focuses on understanding predictive uncertainty in atomistic machine learning models. As such, it does not directly introduce new capabilities for generating data or making decisions affecting individuals or groups, and its societal impact is indirect.

On the positive side, improving the understanding of uncertainty in atomistic models can contribute to more reliable simulations in chemistry and materials science. This has potential downstream benefits in areas such as drug discovery, energy storage, and materials design, where more accurate and trustworthy models can accelerate scientific progress while reducing computational cost. In particular, better uncertainty estimation can enable more efficient data acquisition strategies, reducing the need for expensive quantum-mechanical calculations and lowering the environmental footprint of large-scale simulations.

Potential negative impacts are limited but not entirely absent. More efficient and reliable atomistic modeling tools could be used to accelerate the design of harmful chemical compounds or materials if applied inappropriately. Additionally, incorrect or misinterpreted uncertainty estimates could lead to overconfidence in model predictions, potentially resulting in flawed scientific conclusions or unsafe downstream applications. However, our work explicitly highlights the limitations and regime-dependent behavior of uncertainty estimates, which may help mitigate such risks by discouraging overreliance on poorly calibrated models.

Overall, we believe the primary impact of this work is to improve the transparency and reliability of uncertainty estimation in atomistic machine learning, contributing to safer and more informed use of these models in scientific and engineering applications.