
Multi-Fidelity Transfer Learning Reshapes Uncertainty Quantification in Atomistic Models

Matteo Tolloso

Department of Computer Science
University of Pisa
matteo.tolloso@phd.unipi.it

Davide Bacciu

Department of Computer Science
University of Pisa
davide.bacciu@unipi.it

Luigi Bonati

Italian Institute of Technology
luigi.bonati@iit.it

Abstract

Multi-fidelity training is an increasingly important form of transfer learning in atomistic machine learning, where models pretrained on abundant low-fidelity calculations are adapted using scarce high-fidelity labels. Although uncertainty quantification is critical for deploying these models beyond their training distribution, it remains unclear how predictive uncertainty and its decomposition change during transfer across fidelity levels. Here we extend the MACE architecture with mean-variance outputs and use deep ensembles to study aleatoric and epistemic uncertainty across single-fidelity training and low-to-high fidelity fine-tuning under system-level distribution shift, over five independent data splits. Our results show that the transfer through high-fidelity fine-tuning not only improves in-domain accuracy but also produces a more reliable and stable ranking of errors. Furthermore, it changes the role of uncertainty decomposition and particularly which components are informative for generalization. Aleatoric uncertainty loses calibration and ranking power on unseen data during high-fidelity training, much faster after pretraining, whereas epistemic uncertainty becomes a more reliable indicator of errors in unseen systems. These findings show that the behavior and quality of uncertainty estimates in atomistic models are not fixed properties of the model architecture, but depend on the training pathway.

1 Introduction

Uncertainty estimation plays a central role in atomistic machine learning, where models are routinely deployed beyond their training distribution, for instance in active learning, structure exploration, and molecular dynamics [1–3]. Recent work has studied uncertainty quantification in atomistic models through ensemble methods, calibration strategies, and decompositions into aleatoric and epistemic components [4–8]. In parallel, multi-fidelity learning [9–13] and transfer learning [14–17] have emerged as practical strategies for balancing computational cost and accuracy in atomistic modeling. In multi-fidelity settings, large low-fidelity (LF) datasets are combined with smaller high-fidelity (HF) ones, often corresponding to different levels of electronic-structure theory [18–20]. However, the interaction between multi-fidelity learning and uncertainty estimation remains largely unexplored.

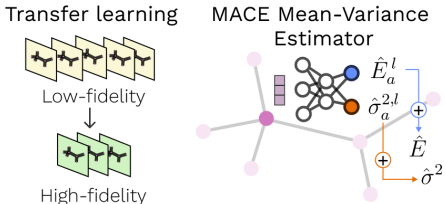
A key missing piece is understanding how predictive uncertainty, particularly its decomposition into aleatoric and epistemic components, behaves under multi-fidelity training and distributional shifts. In standard statistical settings, this decomposition provides information on noise in the data

versus uncertainty in the model [21, 22]. However, in atomistic modeling, where targets are often deterministic and data coverage is highly structured, the interpretation of these components is less clear. Moreover, existing work typically evaluates uncertainty only at convergence, overlooking how uncertainty evolves during training and how it is affected by changes in data fidelity. As a result, it remains unclear whether aleatoric and epistemic uncertainty can be used as *diagnostic signals* to understand model behavior and guide improvements in multi-fidelity settings.

Contributions. We study uncertainty in multi-fidelity atomistic learning as a transfer problem, using a controlled benchmark based on the ANI-1x dataset [23] and its coupled-cluster subset [18]. We compare low-fidelity training, high-fidelity training, and low-to-high-fidelity fine-tuning, and evaluate how this transfer affects aleatoric and epistemic uncertainty under structured shifts to unseen molecular systems. Our results show that fine-tuning improves accuracy and error ranking over training directly on scarce high-fidelity data, while also reshaping the uncertainty decomposition: epistemic uncertainty becomes an informative signal on unseen systems, whereas aleatoric uncertainty remains reliable mainly in-domain and loses its calibration on unseen data during high-fidelity training. Our contributions are: (i) a controlled analysis of uncertainty decomposition across multi-fidelity transfer, distribution shift, and training dynamics; (ii) evidence that fidelity transfer regularizes the ranking of errors but not the calibration of aleatoric uncertainty, and increases reliance on epistemic uncertainty under shift; and (iii) standardized splits, protocols, and metrics for reproducible evaluation at <https://anonymous.4open.science/r/multi-fidelity-training>. Related work is discussed in Appendix B.

2 Methodology

Starting from an atomistic structure $\mathbf{x}$, we use a mean-variance extension of the MACE graph neural network architecture [24] to predict both the total energy and an input-dependent variance. The model represents the energy as a sum of atomic contributions and augments the final readout to also produce atomic variance contributions, which are summed to obtain a structure-level predictive variance. Assuming a Gaussian predictive distribution, each model is trained by minimizing the negative log-likelihood (NLL), so that the learned variance captures heteroscedastic, configuration-dependent residual uncertainty. To separate uncertainty sources, we train an ensemble of $M = 10$ independently initialized mean-variance MACE models, following the mean-variance ensemble framework of Lakshminarayanan et al. [25]. The ensemble mean is used as the final energy prediction, the average predicted variance across members defines *aleatoric uncertainty* (AU), and the variance of the member energy predictions defines *epistemic uncertainty* (EU), following the law of total variance [21]. Their sum, the *total uncertainty* (TU), is the variance of the ensemble’s predictive distribution. As reference energies are deterministic, *aleatoric* and *epistemic* are operational labels for the two variance terms rather than uniquely identified sources of uncertainty [26] (Appendix A).



We evaluate this framework on the ANI-1x dataset [23], which provides molecular structures evaluated at multiple levels of theory, using density functional theory (DFT) energies as the low-fidelity target and coupled-cluster (CC) energies as the high-fidelity target. We consider three training regimes: training only on low-fidelity data (LF-only), training directly on high-fidelity data (HF-only), and low-to-high-fidelity transfer through fine-tuning (LF→HF). In particular, we evaluate models on an in-distribution test set of unseen configurations of the training systems (System-ID) and on an out-of-distribution test set of unseen molecular systems (System-OOD), and analyze both the final models, selected by their validation loss, and the evolution of uncertainty across training epochs. The whole pipeline is repeated on five independent data splits, and all results are reported as five-split means with 95% confidence intervals. Details on dataset splits, model hyperparameters, computational resources, and evaluation metrics are provided in Appendices D, E and C.

3 Results

Throughout this section, we analyze uncertainty along two complementary dimensions: its ability to *rank* high-error samples, measured by AUSE and Spearman correlation, and its *calibration*, measured

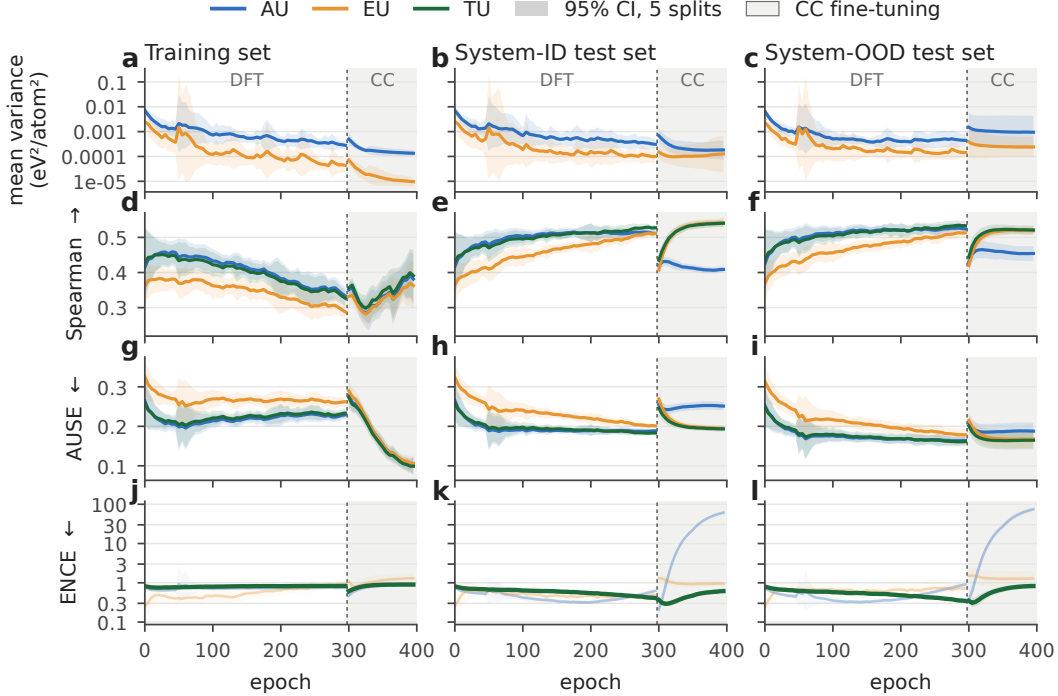


Figure 1: **Training dynamics of uncertainty for LF→HF training under system-level shift.** Columns: training set, System-ID and System-ODD. Rows: mean per-atom AU and EU (a–c), Spearman (d–f, $\uparrow$), AUSE (g–i, $\downarrow$) and ENCE (j–l, $\downarrow$; TU highlighted). Epochs 0–299: DFT pretraining (the LF-only ensemble, DFT labels); 300–399 (shaded): CC fine-tuning (CC labels and configurations), so values across epoch 300 are not directly comparable. Lines: five-split means (geometric on log axes), smoothed by an exponential moving average; bands: 95% intervals; every axis covers all bands.

by ENCE. This distinction is critical, as these two properties do not necessarily evolve together [27]. Since total uncertainty is the variance of the predictive distribution, it is the signal expected to be calibrated, while the ENCE of its aleatoric and epistemic parts indicates how much of the error each of them accounts for.

We first study how uncertainty components evolve during training. We analyze two complementary questions: first, how the roles of AU and EU change during the LF→HF transition; second, whether the high-fidelity stage of multi-fidelity training produces different uncertainty dynamics than training directly on high-fidelity data.

Figure 1 shows the uncertainty dynamics for LF→HF training under system-level shift.¹ During the low-fidelity pretraining phase, both AU and EU decrease in magnitude (Fig. 1a–c). On the train set, AU remains the dominant uncertainty component and provides better ranking performance than EU (Fig. 1d, g), although the ranking of both components does not improve with training epochs. On unseen data (both ID and OOD), instead, all signals rank errors increasingly well during pretraining, and total uncertainty becomes better calibrated (Fig. 1e, f, h, i, k, l). The transition to high-fidelity fine-tuning changes this behavior. On unseen data, AU becomes significantly less reliable: its ranking falls behind that of EU and TU, as shown by lower Spearman correlation (Fig. 1e, f) and higher AUSE (Fig. 1h, i), and its calibration deteriorates sharply, with an ENCE growing by about two orders of magnitude (Fig. 1k, l). In contrast, after fine-tuning, EU becomes more informative than AU on unseen data and its calibration remains stable, while total uncertainty provides the most stable signal and maintains accurate ranking even under OOD evaluation. On the training set, AU instead remains conservative (Fig. 1j). This suggests that the transition from LF pretraining to HF

¹In the reviewed version, the System-ID and System-ODD columns of this figure plotted AUSE, ENCE and Spearman in the rows labelled Spearman, AUSE and ENCE, which caused the apparent disagreement with Figure 2. All figures now use five data splits.

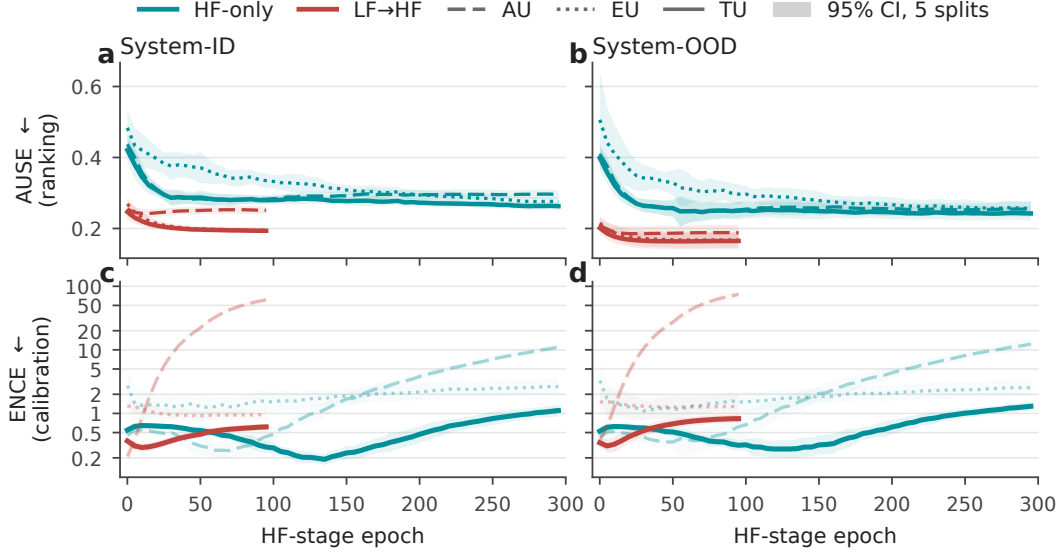


Figure 2: **Training dynamics of uncertainty for HF-only training vs. high-fidelity fine-tuning.** AUSE (a, b) and ENCE (c, d; TU highlighted) on System-ID and System-OOD, for HF-only training from scratch (teal) and the CC fine-tuning of LF→HF (red; the runs of the shaded region of Figure 1); line style: signal. Lines and bands as in Figure 1.

fine-tuning changes the relative role of AU and EU: AU remains effective on fitted training data, whereas EU becomes more relevant for generalization. Overall, high-fidelity fine-tuning redistributes uncertainty informativeness from the aleatoric component toward the epistemic component, while total uncertainty remains the most robust diagnostic signal.

Figure 2 directly compares HF-only training with the high-fidelity fine-tuning stage of LF→HF training, i.e. the same runs as the shaded region of Figure 1, showing how error ranking (AUSE) and calibration (ENCE) vary across epochs on System-ID and System-OOD. The benefit of multi-fidelity learning appears in ranking from the very first fine-tuning epoch: LF→HF achieves a lower AUSE than HF-only for all three signals, on System-OOD already lower than HF-only after its full training, and its ranking remains stable during fine-tuning (Fig. 2a, b). Calibration tells a different story (Fig. 2c, d). In both protocols, the calibration of AU on unseen data deteriorates as high-fidelity training proceeds, but this happens much faster after pretraining: within a few fine-tuning epochs for LF→HF, against more than a hundred epochs for HF-only. EU remains comparatively stable in both cases, which limits the degradation of TU. Thus, low-fidelity pretraining regularizes the ranking of errors, but not the calibration of the aleatoric component, whose loss it accelerates.

Limits of aleatoric variance transfer across fidelities. A variance head fitted to DFT residuals carries no direct information about the residuals of another level of theory, so some loss of aleatoric calibration at the fidelity switch is expected. The overconfidence we observe, however, is not caused by the switch alone: it also develops when training on CC energies from scratch, and in both protocols AU stays conservative on the training data. A plausible explanation is that the variance head follows the training residuals, which keep shrinking during high-fidelity training while held-out errors level off (Appendix F), and transfers these small variances to unseen configurations, a known failure mode of mean–variance training [28, 29]. With deterministic reference energies, AU thus behaves less like irreducible noise and more like an estimate of in-sample residual difficulty, which neither transfers across fidelities nor generalizes to unseen configurations.

We now analyze the final models by evaluating the behavior of aleatoric and epistemic uncertainty at convergence. These are the checkpoints with the lowest validation loss, which for LF→HF are reached after only a few fine-tuning epochs, before most of the loss of aleatoric calibration seen in Figures 1 and 2. Figure 3 summarizes their accuracy, ranking and calibration; reliability diagrams are reported in Appendix G. The ranking quality of the uncertainty estimates is stable between System-ID and System-OOD, and slightly better on unseen systems. Among the two CC-trained protocols,

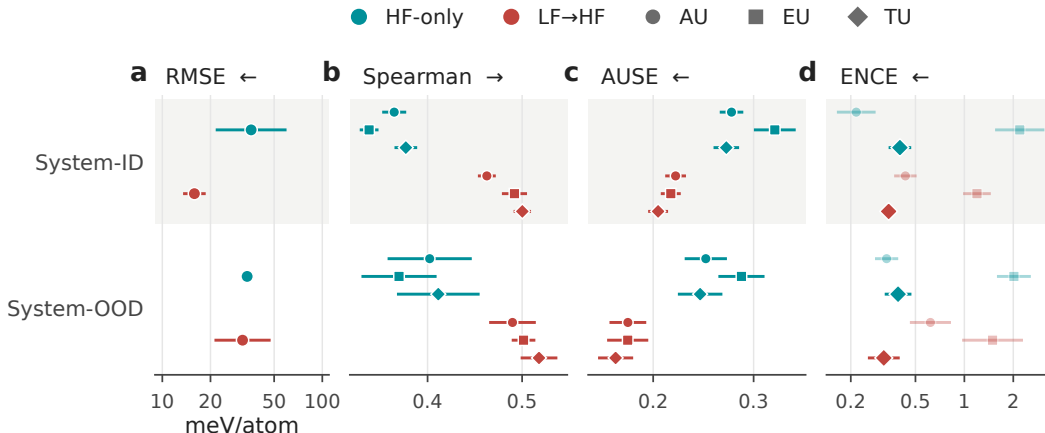


Figure 3: **Accuracy and uncertainty of the converged models under system-level shift.** (a) RMSE of the ensemble mean, (b) Spearman and (c) AUSE (ranking), (d) ENCE (calibration; TU highlighted) of the best-validation checkpoints. Bars: 95% intervals over five splits; log axes use geometric means. The wide HF-only System-ID RMSE interval stems from one configuration in one split, which TU ranks as the most uncertain of 5,000.

LF→HF clearly outperforms HF-only in ranking quality, with higher Spearman correlation and lower AUSE for every signal in both regimes; it also more than halves the in-domain error, whereas on unseen systems the errors of the two protocols are similar. Multi-fidelity training also reshapes the decomposition: without pretraining, EU is the weakest ranking signal, whereas after pretraining it matches or exceeds AU and approaches TU, the best signal in every case. Total uncertainty is well calibrated for both protocols, with EU accounting for a larger share of the error after pretraining. This indicates that low-fidelity pretraining provides a useful initialization for high-fidelity uncertainty estimation, whereas training directly on CC data leads to substantially weaker error ranking.

4 Limitations

Our analysis is limited to energies; extending it to forces, which govern the stability of molecular dynamics, is the most important next step. We use one dataset (ANI-1x) and one shift (unseen molecular systems); other chemistries, materials and shifts are needed to establish generality. We study one estimator, mean-variance deep ensembles: the aleatoric overconfidence is tied to fitting a variance head with the Gaussian NLL, so other ensembles, evidential or Bayesian methods, or modified objectives [29, 30] may behave differently, and we do not ablate choices such as resetting the variance head before fine-tuning. Finally, LF→HF also sees the DFT labels, so its comparison with HF-only concerns data access as well as fidelity transfer.

5 Conclusion

We examined aleatoric (AU) and epistemic (EU) uncertainty in deep ensembles for multi-fidelity atomistic models across data fidelity, distribution shift, and training trajectory. Our results show that uncertainty decomposition is not fixed: the relative usefulness of AU and EU depends on the training pathway, the fidelity transition, and the stage of training. Low-fidelity pretraining followed by high-fidelity fine-tuning consistently improved error ranking, and in-domain accuracy, over direct high-fidelity training, but it did not regularize calibration in the same way: on unseen data, AU lost its calibration during high-fidelity training, much faster after pretraining, while remaining reliable on the fitted training data. After fine-tuning, EU became the more informative signal on unseen data. Across regimes and throughout training, total uncertainty provided the most robust signal; ranking and calibration should therefore be monitored separately on held-out data.

References

- [1] Evgeny V Podryabinkin and Alexander V Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.
- [2] 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.
- [3] 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.
- [4] Jin Dai, Santosh Adhikari, and Mingjian Wen. Uncertainty quantification and propagation in atomistic machine learning. *Reviews in Chemical Engineering*, 41(4):333–357, 2025.
- [5] 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.
- [6] 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.
- [7] 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)*, pages 1508–1516. IEEE, 2022.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] 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.

- [18] 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.
- [19] 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, ccsc, and ccsc (t) accuracy. *Journal of Chemical Theory and Computation*, 19(14):4510–4519, 2023.
- [20] Viktor Zaverkin, David Holzmüller, Luca Bonferraro, and Johannes Kästner. Transfer learning for chemically accurate interatomic neural network potentials. *Physical Chemistry Chemical Physics*, 25(7):5383–5396, 2023.
- [21] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.
- [26] 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.
- [27] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR, 2018.
- [28] 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, pages 55–60. IEEE, 1994.
- [29] 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.
- [30] 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*, pages 5593–5613. PMLR, 2023.
- [31] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [32] Jakob Gawlikowski, Cedrique Rovile Njiteutcheu 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.
- [33] Trygve Helgaker, Poul Jorgensen, and Jeppe Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, 2013.
- [34] 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.

- [35] 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.
- [36] 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.
- [37] 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.
- [38] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.
- [39] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.
- [40] Nicki Skafté, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *Advances in Neural Information Processing Systems*, 32, 2019.
- [41] Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of mean variance estimation neural networks. *Neurocomputing*, 597:127929, 2024.
- [42] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udfluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.
- [43] Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.
- [44] 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.
- [45] 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.
- [46] 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.
- [47] Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 5(3):035006, 2024.
- [48] 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.
- [49] 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.
- [50] Simon Kristoffersson Lind, Ziliang Xiong, Per-Erik Forssén, and Volker Krüger. Uncertainty quantification metrics for deep regression. *Pattern Recognition Letters*, 186:91–97, 2024.
- [51] Dan Levi, Liran Gispan, Niv Giladi, and Ethan Fetaya. Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors*, 22(15):5540, 2022.

A Details on methodology

Starting from an atomistic structure $\mathbf{x}$, we use a mean–variance MACE model to predict both a total energy and an input-dependent variance. To separate uncertainty sources, we train an ensemble of independently initialized mean–variance models: the average predicted variance defines the aleatoric component, while the disagreement between ensemble energy predictions defines the epistemic component. We then study how these uncertainty signals behave under controlled changes in training fidelity and under a system-level distributional shift across molecular systems. The following sections formalize this decomposition (Section A.1), describe the predictive model (Section A.2) and state the assumptions behind the estimator (Section A.3).

A.1 Error and predictive uncertainty

For an atomistic structure (such as a molecular configuration) $\mathbf{x}$, defined by atomic positions and chemical species, and a reference property y , such as the total energy from a quantum-chemical calculation, the prediction error is the discrepancy between the model prediction $\hat{y}(\mathbf{x})$ and the reference value. This error can only be evaluated once the reference calculation is available. Predictive uncertainty, instead, aims to quantify the reliability of the prediction before observing y , and is naturally associated with the predictive distribution $p(y|\mathbf{x}, \mathcal{D})$, where $\mathcal{D}$ denotes the training dataset [31].

Predictive uncertainty is commonly decomposed into aleatoric and epistemic components [32]. Epistemic uncertainty reflects uncertainty in the predictor itself, arising from limited data coverage, model misspecification, or extrapolation beyond the training distribution, and is in principle reducible with additional data. Aleatoric uncertainty is classically associated with irreducible observation noise. In this case, however, reference energies computed with a given electronic-structure method and numerical protocol are typically deterministic for a given configuration [33]. We can therefore interpret aleatoric uncertainty in a generalized, operational sense: it captures input-dependent residual variability, including heteroscedasticity and configuration-dependent difficulty.

Formally, if θ denotes the model parameters and Y the random variable associated with the reference property of structure $\mathbf{x}$, the predictive variance can be decomposed through the law of total variance [21] as

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

The first term is the expected conditional variance of the target under a fixed model, and therefore captures variability intrinsic to the prediction task itself; for this reason, it is associated with aleatoric uncertainty. The second term measures how much the prediction changes across plausible models consistent with the training data, and therefore reflects uncertainty about the predictor; this is the epistemic component. 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 [26]. We therefore use *aleatoric* and *epistemic* as operational labels for the two variance terms (Section A.3).

A.2 Predictive model and uncertainty parameterization

Mean–variance MACE model. We build on MACE, an E(3)-equivariant message-passing architecture that combines message passing with higher-body atomic-cluster-expansion features for accurate and efficient modeling of the potential energy surface [24, 34]. 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 a Gaussian random variable E conditioned on the structure, $E|\mathbf{x} \sim \mathcal{N}(\hat{E}(\mathbf{x}), \hat{\sigma}^2(\mathbf{x}))$, we train the model by minimizing the Gaussian negative log-likelihood

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

The variance learned through this objective forms 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 [25] of M independently initialized mean–variance MACE models. This strategy is closely related to committee models widely used for active learning in interatomic potentials [35–37]. 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 independently trained 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)

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

The first term (average heteroscedastic variance predicted by individual models) is interpreted as aleatoric uncertainty, the second (disagreement among independently trained ensemble members) as epistemic uncertainty [25]. This decomposition defines the three uncertainty signals analyzed throughout the paper, which we evaluate separately in terms of error ranking and calibration in order to test how their role changes across fidelity regimes and distribution shift.

A.3 Assumptions behind the estimator

This section collects standard results behind the estimator of Section A.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 [38]

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

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 [39]. 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 A.2. For a fixed mean $\mu(\mathbf{x})$, the expected loss under a target distribution $p(y | \mathbf{x})$ is minimized by the conditional mean squared residual [28],

$$\sigma^{2*}(\mathbf{x}) = \mathbb{E}[(y - \mu(\mathbf{x}))^2 | \mathbf{x}]. \quad (5)$$

With deterministic labels, $p(y | \mathbf{x})$ 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 [40, 29, 30, 41]. 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 Section 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 [29]. 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 $\frac{1}{M} \sum_m \mathcal{N}(\hat{E}^{(m)}(\mathbf{x}), \hat{\sigma}^{2,(m)}(\mathbf{x}))$. 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 [25, 22, 42]. 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 [26]: 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 [43] holds exactly for every configuration,

$$e^2 = \frac{1}{M} \sum_{m=1}^M e_m^2 - \text{EU}, \quad \text{hence} \quad \text{TU} - e^2 = \left(\text{AU} - \frac{1}{M} \sum_{m=1}^M e_m^2 \right) + 2 \text{EU}, \quad (6)$$

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 of the reliability diagrams (Appendix G). 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 during fine-tuning (Section 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 (the median per-configuration EU/AU of the converged models is between 0.40 and 0.69 on both test sets); 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.

B 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 [4, 5, 44]. 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. [6] 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 [45–47].

Several works have specifically addressed calibration and uncertainty decomposition in atomistic models. Busk et al. [8] 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 and Mori [7] 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. [48] 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. [12] 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 [10].

Finally, the interpretation of uncertainty depends strongly on the evaluation metric. Rasmussen et al. [49], Lind et al. [50] 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 distribution shift, 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.

C Details on uncertainty evaluation metrics

A central aspect of our analysis is to disentangle different properties of predictive uncertainty. In particular, we distinguish between *ranking quality*, whether uncertainty correctly identifies high-error samples, and *calibration*, whether uncertainty magnitudes match the scale of prediction errors. To this end, we employ the following complementary metrics. Both errors and uncertainties are normalized per atom: energies are divided by the number of atoms and variances by the square of the number of atoms.

RMSE (Root Mean Squared Error) of the ensemble mean is the standard metric for evaluating the accuracy of atomistic models.

Predicted uncertainty is the model-predicted variance (AU, EU or TU), which serves as an estimate of the expected squared error.

Spearman rank correlation measures the monotonic relationship between predicted uncertainty and prediction error. It evaluates whether uncertainty preserves the ordering of samples by difficulty, independently of its scale.

AUSE (Area Under the Sparsification Error) [50] measures how effectively uncertainty ranks samples according to their prediction error. Low AUSE indicates that uncertainty is useful for identifying unreliable predictions, with greater emphasis on correctly ranking high-error samples.

Binned Expected Normalized Calibration Error (ENCE) [51] evaluates whether predicted uncertainty magnitudes are consistent with empirical errors, and therefore measures the *calibration* of uncertainty, independently of its ranking quality.

C.1 Area under the sparsification error (AUSE)

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

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

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), and $S_U(\alpha)$ 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, and AUSE is defined as

$$\text{AUSE}(S) = \frac{1}{\text{MAE}(S)} \int_0^1 (\text{MAE}(S_U(\alpha)) - \text{MAE}(S_V(\alpha))) d\alpha. \quad (8)$$

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,

$$\rho_S = \text{Corr}_{\text{rank}}(\sigma^2(x), e^2(x)), \quad (9)$$

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

We partition the test set into $B = 15$ bins with an equal number of configurations each, sorted by predicted uncertainty. For each bin b , we compute

$$\text{RMSE}_b = \sqrt{\mathbb{E}[e^2|b]}, \quad \text{RMV}_b = \sqrt{\mathbb{E}[\sigma^2|b]}, \quad (10)$$

and define

$$\text{ENCE} = \frac{1}{B} \sum_{b=1}^B \frac{|\text{RMV}_b^2 - \text{RMSE}_b^2|}{\text{RMV}_b^2}. \quad (11)$$

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. A bin where the variance overestimates the error contributes less than one, so an ENCE above one can only come from overconfident bins.

C.4 Negative log-likelihood (NLL)

NLL evaluates the probabilistic correctness of the predictive distribution,

$$\text{NLL}(S) = -\frac{1}{N} \sum_{i=1}^N \log p(y_i|x_i), \quad (12)$$

which under the Gaussian predictive assumption reads $\frac{1}{2} (e^2/\sigma^2 + \log \sigma^2) + C$, with C 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 (Section A.3); 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.

C.5 Training-support check

Metrics that average squared errors or variances can be dominated by a single configuration on which the models extrapolate far outside their training data. For the converged models, we therefore exclude every test configuration whose smallest interatomic distance falls below the smallest one present in that split’s DFT and CC training and validation sets. The rule reads geometry only, never predictions or uncertainties, and applies to every protocol and signal alike. On the system split it removes a single System-ID configuration (in one of the five splits) and no System-OOD configuration. The per-epoch curves of Figures 1 and 2 and of Appendix F use all configurations.

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:

$$\mathcal{S} = \mathcal{S}_{\text{seen}} \cup \mathcal{S}_{\text{OOD}}, \quad \mathcal{S}_{\text{seen}} \cap \mathcal{S}_{\text{OOD}} = \emptyset, \quad (13)$$

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:

$$\mathcal{X}_s = \mathcal{X}_s^{\text{train}} \cup \mathcal{X}_s^{\text{val}} \cup \mathcal{X}_s^{\text{test-ID}}, \quad (14)$$

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:

$$\mathcal{X}_{\text{ID}}^{\text{train}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{train}}, \quad (15)$$

$$\mathcal{X}_{\text{ID}}^{\text{val}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{val}}, \quad (16)$$

$$\mathcal{X}_{\text{ID}}^{\text{test}} = \bigcup_{s \in \mathcal{S}_{\text{seen}}} \mathcal{X}_s^{\text{test-ID}}. \quad (17)$$

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

$$\mathcal{X}_{\text{CC}} \subseteq \mathcal{X}_{\text{DFT}}, \quad (18)$$

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 $\sigma \in \{\text{train, val, test-ID, test-OOD}\}$ and theory level (DFT or CC), let

$$\mathcal{X}_{s,\sigma} \quad (19)$$

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

$$n_{s,\sigma} = |\mathcal{X}_{s,\sigma}| \quad (20)$$

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

$$N_\sigma = \sum_s n_{s,\sigma}. \quad (21)$$

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:

$$\tilde{q}_{s,\sigma} = N_\sigma^{\text{target}} \frac{w_{s,\sigma}}{\sum_{s'} w_{s',\sigma}}, \quad (22)$$

where the weights are defined as

$$w_{s,\sigma} = \begin{cases} n_{s,\sigma}, & \text{if no cap is applied,} \\ \min(n_{s,\sigma}, \kappa), & \text{with optional cap } \kappa. \end{cases} \quad (23)$$

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

$$\sum_s q_{s,\sigma} = N_\sigma^{\text{target}}. \quad (24)$$

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

$$\hat{\mathcal{X}}_\sigma = \bigcup_s \hat{\mathcal{X}}_{s,\sigma}, \quad (25)$$

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

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

$$\mathcal{X}_\sigma^{\text{CC}} \subseteq \hat{\mathcal{X}}_\sigma^{\text{DFT}}. \quad (26)$$

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 $\text{CC} \subseteq \text{DFT}$,
- 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 G.2.
- **Test-ID (System-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 System-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 Training details and hyperparameters

Data. Low-fidelity labels are the ANI-1x ω B97x/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. 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 datasets, identical for 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, $L_{\text{max}} = 1$ 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 A.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 checkpoint lies at HF-stage epoch 60 in the median (range 30–111) for HF-only and 2 (range 1–3) for LF→HF, on the epoch axis of Figure 2, where epoch 0 is the first checkpoint of the high-fidelity stage.

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). Training the 150 ensemble members reported in this paper took 825.5 GPU hours for LF-only pretraining (about 15 hours per member on one A100 GPU, occupying about 20GB of GPU memory), 75.9 hours for HF-only training and 25.8 hours for LF→HF fine-tuning, about 930 GPU hours in total. Per-epoch evaluation adds to this budget; 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.

F Training dynamics

This section complements Figures 1 and 2 with the full per-epoch trajectories of every metric, on the training, ID and OOD sets side by side: HF-only training (Figure 4) and the whole LF→HF trajectory (Figure 5), which also shows the LF-only ensemble during the pretraining epochs. Figure 6 gives the training and validation curves of every training regime.

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 epochs. Second, on the training set the calibration error of AU and TU stays 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 decreases much less. 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 6) 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 E). Pretraining also shows a transient instability around epochs 50–60, visible as wide bands in every metric.

G Supplementary results for the converged models

This section complements Figure 3. Statistics follow Section 2: five-split means with two-sided 95% Student-*t* intervals, and geometric means with log-space intervals for ENCE and RMSE.

G.1 Reliability diagrams

Figure 7 shows the reliability diagrams of every converged ensemble, including the LF-only ensembles evaluated against DFT labels, from which the calibration results of Figure 3 are computed. The bands are wide in the few highest-variance bins, which hold a handful of large-error configurations per split. 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.

G.2 Post-hoc isotonic recalibration

For each signal we fit an isotonic regression from predicted variance to squared error on the ensemble’s validation predictions in each split [27, 8], 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 3, and AUSE and Spearman change only through the ties that the piecewise-constant map introduces. In domain it works as intended, reducing the TU ENCE on System-ID from 0.401 to 0.250 (HF-only) and from 0.342 to 0.174 (LF→HF). On System-OOD, recalibration repairs the badly scaled epistemic signal ($2.02 \rightarrow 0.44$ for HF-only, $1.49 \rightarrow 0.35$ for LF→HF), with mixed effects on TU ($0.391 \rightarrow 0.330$ for HF-only, $0.319 \rightarrow 0.508$ for LF→HF), while ranking is essentially unchanged (System-OOD TU Spearman $0.411 \rightarrow 0.410$ and $0.518 \rightarrow 0.517$). Figure 8 shows 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, so the map learned on in-domain data transfers reasonably well to unseen molecular systems. Post-hoc recalibration, however, assumes that the relation between uncertainty and error is the same on the validation and test data; it should therefore be validated explicitly before it is relied upon under stronger extrapolation.

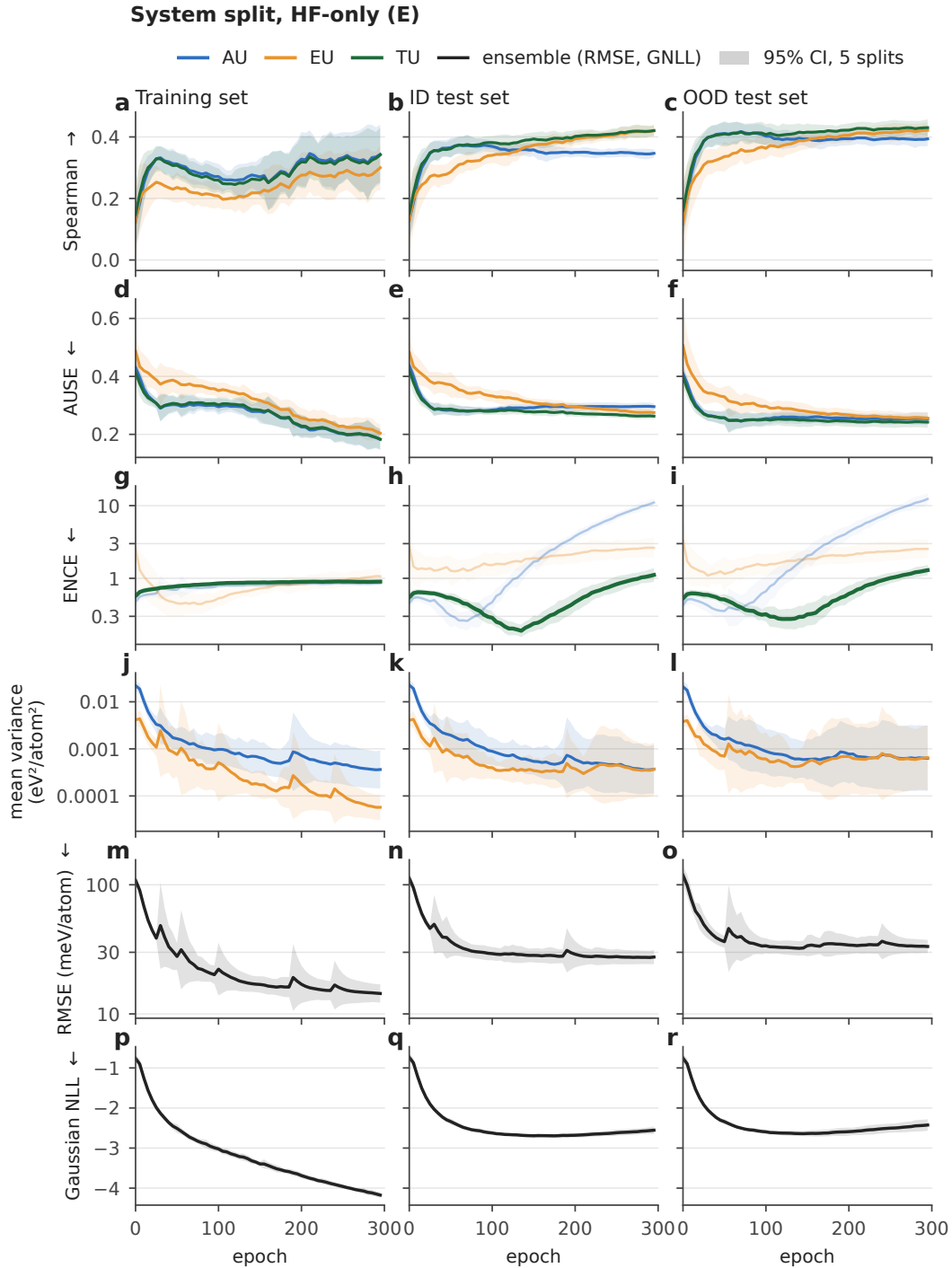


Figure 4: **Training dynamics of HF-only ensembles under system shift.** Columns: training set, ID and OOD test set. Rows: Spearman (↑), AUSE (↓), ENCE (↓, TU highlighted, AU and EU faded), mean per-atom AU and EU, RMSE of the ensemble mean and Gaussian NLL of the predictive distribution, against epoch. Lines are exponential moving averages of the five-split means (geometric on log axes); bands are 95% intervals.

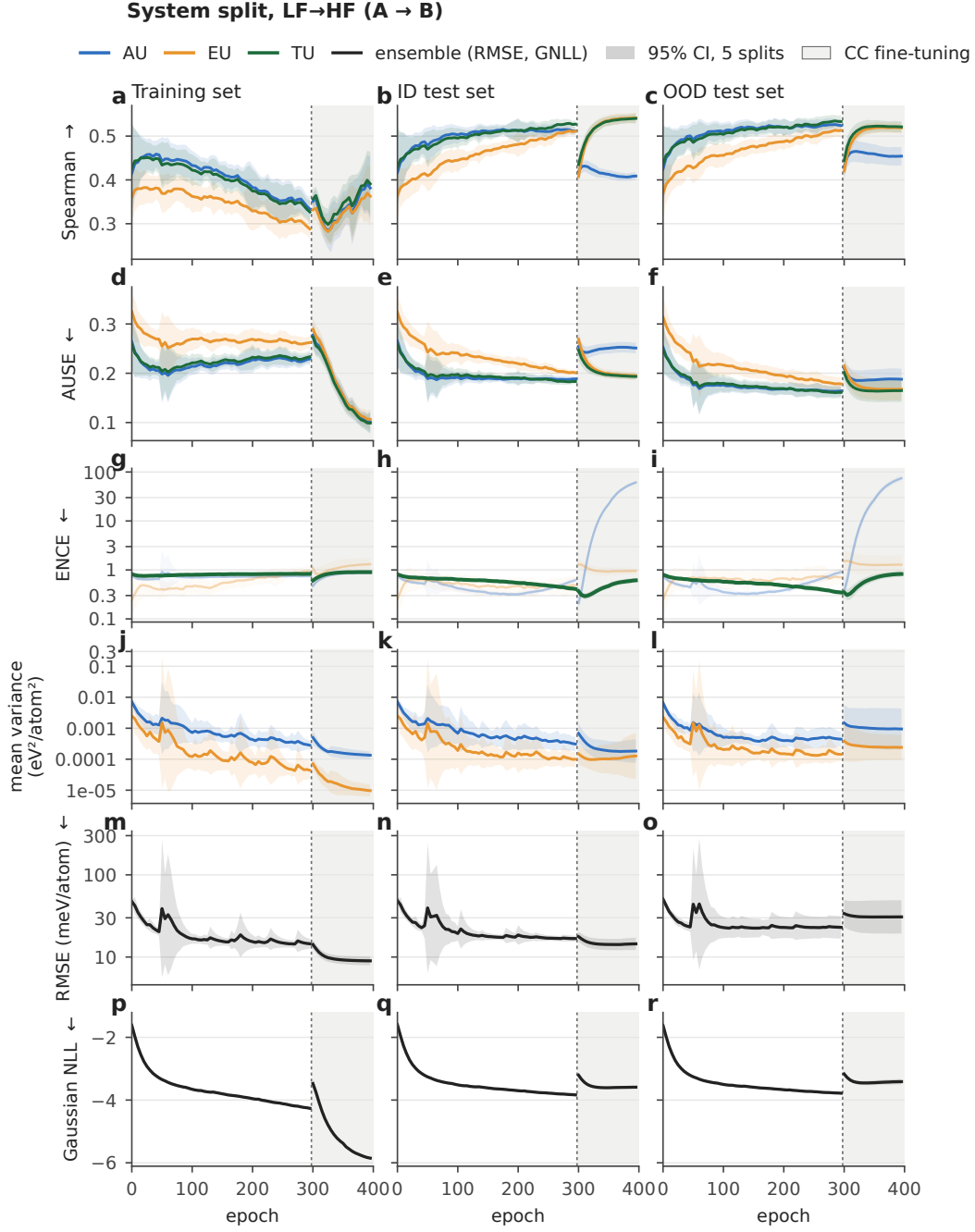


Figure 5: **Training dynamics of LF→HF ensembles under system shift.** As Figure 4. Epochs 0–299 are DFT pretraining (the LF-only ensemble, DFT labels), epochs 300–399 (shaded) CC fine-tuning; RMSE and NLL jump at epoch 300 partly because the target changes. The first four rows show the data of Figure 1 in a different order.

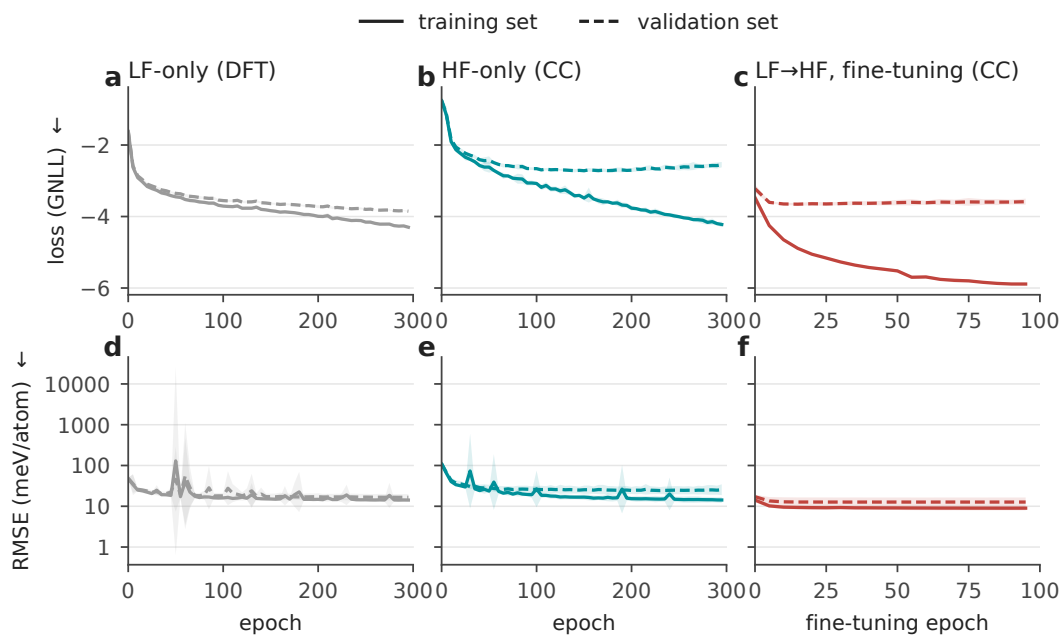


Figure 6: **Training and validation curves.** Training loss (Gaussian NLL, solid) and validation loss (dashed), and the corresponding RMSE of the ensemble mean, for LF-only (DFT), HF-only (CC) and the fine-tuning stage of LF→HF (CC). Loss: five-split arithmetic means; RMSE: geometric means; bands are 95% intervals, wide around epochs 50–60 of LF-only training. 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.

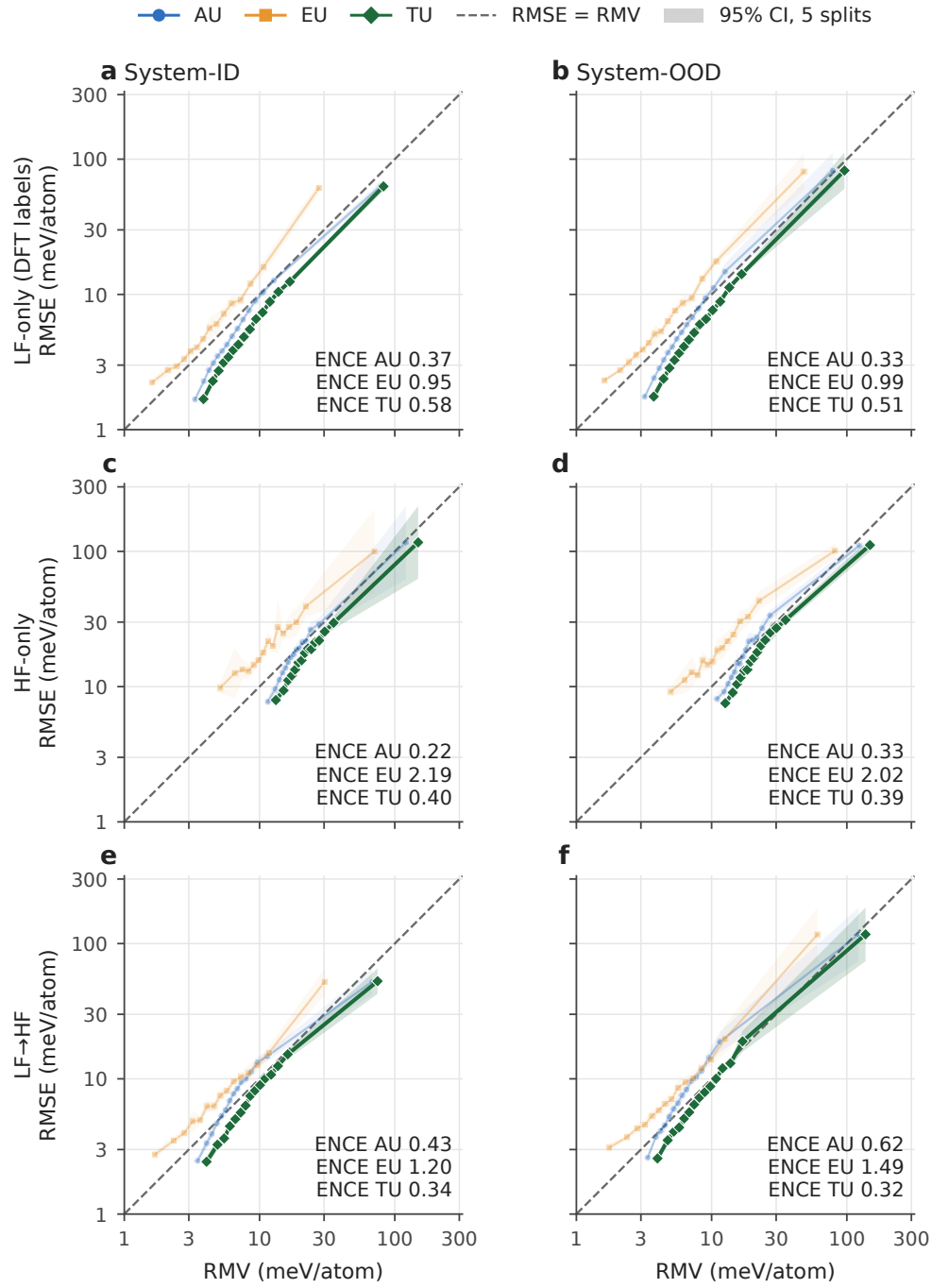


Figure 7: **Reliability of the converged ensembles under system shift, before recalibration.** Rows: LF-only (DFT labels), HF-only and LF→HF (CC labels); columns: ID and OOD test set. In each split, the test configurations are sorted by each signal and cut into 15 equal-count bins; each point is the 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.

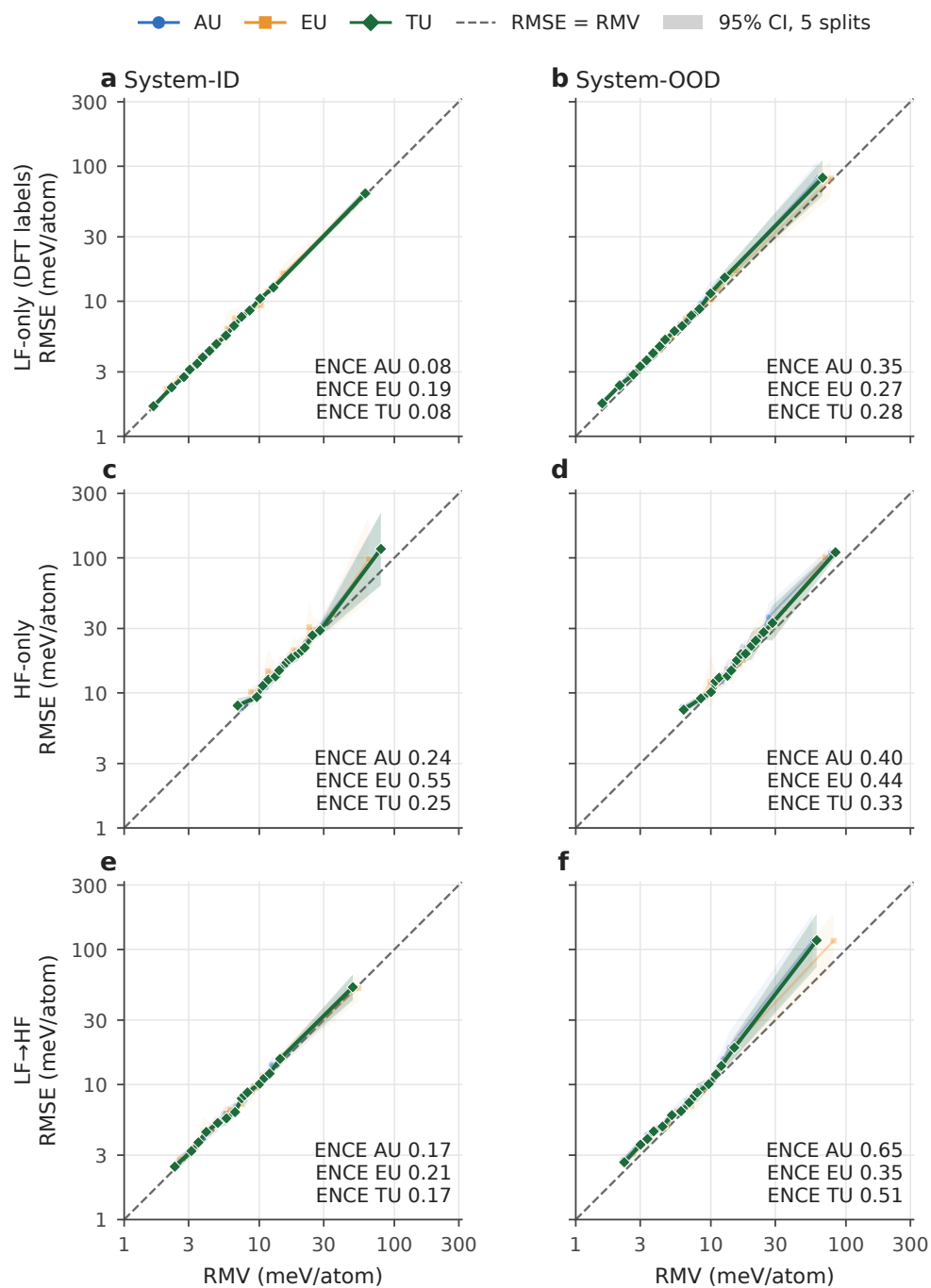


Figure 8: **Reliability of the converged ensembles under system shift, after isotonic recalibration.** As Figure 7, after the isotonic map of each signal, fitted on the in-domain validation set of each split.