Predicting Precision: Equivariant Delta-learning, Recommender Systems and Uncertainty Quantification for High-Throughput DFT

Daniel T. Speckhard [1 2] **Christian Carbogno** [2] **Matthias Scheffler** [3] **Claudia Draxl** [1]

Abstract

Density Functional Theory (DFT) is a cornerstone of materials discovery, but running highly precise calculations requires expensive convergence testing for parameters like basis-set size and k-point density. To bypass this, we use Δ-learning to predict the difference between fast, low-fidelity calculations and their converged, high-fidelity counterparts. Current machine learning models employed for similar tasks, however, are deterministic and often fail silently on out-of-distribution crystal structures. We introduce a framework that integrates Uncertainty Quantification directly into an $E(3)$-equivariant graph neural network to predict these Δ-corrections. By conditioning the network on the low-fidelity DFT settings using Feature-wise Linear Modulation (FiLM), our model adapts to the specific physical approximations made. We evaluate our approach on a newly generated dataset of over 60,000 multi-fidelity relaxations for 4,220 binary semiconductors. The framework not only acts as a highly accurate error-corrector, outperforming heavily tuned Random Forests, but also functions as a zero-cost recommender system that predicts calculation errors and recommends what settings to use before the calculations are run. Finally, we show that Deep Ensembles and latent-space distance metrics provide explicit out-of-distribution detection, allowing the model to flag its own catastrophic errors and ensure out-of-distribution reliability for high-throughput screening.

[1]Department of Physics and CSMB, Humboldt-Universität zu Berlin, Berlin, Germany [2]Fritz Haber Institute of the Max Planck Society, Berlin, Germany [3]The NOMAD Laboratory at BIFOLD, Technical University Berlin, Berlin, Germany. Correspondence to: Daniel Speckhard <dts@physik.hu-berlin.de>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Density Functional Theory (DFT) calculations are responsible for a large portion of the world's supercomputing time (Marks, 2021). Researchers rely on DFT to discover new materials for batteries, solar cells, and carbon capture (Jain et al., 2016; Tabor et al., 2018). However, getting highly precise results requires running calculations with dense k-point meshes and large basis sets, which is computationally expensive. Note, we make a distinction between an *accurate* DFT calculation where the exchange-correlation (xc) functional follows Jacob's ladder (Giustino, 2014) and a precise calculation, that for a given xc-functional has all of the other settings set to their most computationally expensive values (e.g., potential cutoff radius) (Lejaeghere et al., 2016). To save time and compute budget, researchers often run high-throughput workflows using cheaper, *low-fidelity* settings, but this inherently sacrifices precision.

To bridge this gap, the machine learning community has increasingly turned to Δ-learning (Ramakrishnan et al., 2015), and they have largely focused on accuracy. Instead of training a neural network to predict the absolute energy or volume of a crystal from scratch, the model is trained to predict the difference (the Δ) between the cheaper, less accurate calculation and the more expensive more accurate one. While this approach has shown great promise, current models have two major flaws. First, they typically treat all *cheap* calculations as if they were the same, without knowing exactly which computational shortcuts were taken to generate the baseline data. Second, they are deterministic point estimators. Machine learning models are inherently statistical and should, in principle, provide predictive distributions rather than point estimates. In materials science, where datasets are typically highly biased and not i.i.d., epistemic uncertainty is critical. When these models are fed a new, exotic crystal structure that falls into an extrapolative regime, they confidently guess a correction value without providing any measure of uncertainty.

We formulate the precision Δ-learning problem in two ways. First, our model functions as a zero-cost recommender system to suggest settings to use for a DFT relaxation as shown in Figure 1. A large amount of compute is currently spent on multi-dimensional convergence sweeps to ensure suf-

ficient calculation precision, or on using overconverged settings throughout a high-throughput workflow when no such sweeps are run. By operating our model before any calculation is run, a user can input a crystal structure and query the network to find the cheapest requested settings that still guarantee a desired precision threshold within a strict, high-confidence interval.

Second, the model acts as a post-calculation data harmonization tool for heterogeneous databases. Repositories like NOMAD (Draxl & Scheffler, 2019) host over 100 million DFT calculations, but because many of the data were run with varying settings for varying purposes (e.g., molecular dynamics or sub-meV elemental bulk calculations), the machine learning community struggles to repurpose these data into a single dataset by placing the data on equal footing. By taking an existing low-fidelity calculation and its settings as inputs, our model provides a Δ-correction to the highest-precision baseline alongside an explicit uncertainty estimate. This allows researchers to confidently stitch together big, harmonized datasets, discarding only the calculations the model flags as out-of-distribution.

In this paper, we propose a solution: an uncertainty-quantification-aware, $E(3)$-equivariant graph neural network designed specifically for multi-fidelity DFT data that acts as a recommender system and post-calculation error correction model. We feed the specific low-fidelity DFT settings directly into the network's intermediate layers so it knows exactly what baseline it is correcting. We also equip the model with deep UQ so it can reject unreliable predictions when it encounters entirely new materials and/or settings.

Our main contributions are:

- We use Feature-wise Linear Modulation (FiLM) to inject the low-fidelity DFT settings into the equivariant message-passing layers, so the network knows which approximations were taken in the baseline calculation.
- We compare uncertainty quantification approaches and find that a k-nearest-neighbors distance in the model's latent space provides an alternative to Deep Ensembles, capturing substantial UQ signal at a fraction of the computational cost.
- We show our architecture also successfully operates as a *zero-cost* recommender system, outperforming tabular baseline models by predicting the computational error before a DFT calculation is run.
- We open-source a newly generated dataset of over 60,000 multi-fidelity structural relaxations for 4,220 binary semiconductors to accelerate studies into DFT precision.

2. Related Work

Δ-learning in DFT: In computational chemistry and solid-state physics, researchers often have to choose between fast, low-accuracy or low-precision computational methods and slow, high-accuracy or high-precision ones. To date the community has largely focused on the problem of accuracy, and increasingly turned to a machine learning technique known as Δ-learning. Instead of training a model to predict a molecular property from scratch, Δ-learning trains the model to simply predict the difference, or the Δ, between a cheap baseline calculation and an expensive, high-fidelity target. This approach has been highly successful across a variety of tasks. For example, it has been used to predict Coupled Cluster energy corrections (Ruth et al., 2022), to boost the accuracy of subsystem density functional theory (DFT) to match full Kohn-Sham calculations (Pauletti et al., 2021), to predict complex properties like dielectric and Raman spectra (Grumet et al., 2024) and more recently to predict the precision of a calculation based on its basis-set size (Speckhard et al., 2025).

A major drawback of standard Δ-learning models is that they are almost always deterministic. This means they output a single number without any measure of confidence. When these models are fed a material or molecule that looks completely different from the training data (i.e., out-of-distribution data), they tend to fail silently. They will confidently output an incorrect Δ-correction without giving the end-user any warning that the model is essentially guessing.

Uncertainty quantification and graph models: To know if we can actually trust a theoretical calculation, we need UQ. In traditional physics and chemistry simulations, UQ is well-recognized as a critical step. For instance, researchers rely on it to track errors in nuclear density functional theory (Schunck et al., 2015) and better understand convergence parameters in plane-wave DFT calculations (Janssen et al., 2024). Despite its importance, UQ is often left out or treated as a post-processing afterthought when building machine learning models for chemistry/physics. At the same time, the underlying architectures of these ML models are advancing. Recently, researchers have started mapping molecular structures to Δ-corrections using Equivariant Graph Neural Networks (EGNNs). A notable example of this is the Δ-EGNN method proposed by Huo and Dong, which successfully speeds up the creation of machine learning potentials (Huo & Dong, 2025) by predicting the accuracy Δ between several properties from the ωB97x xc-functional to CCSD(T).

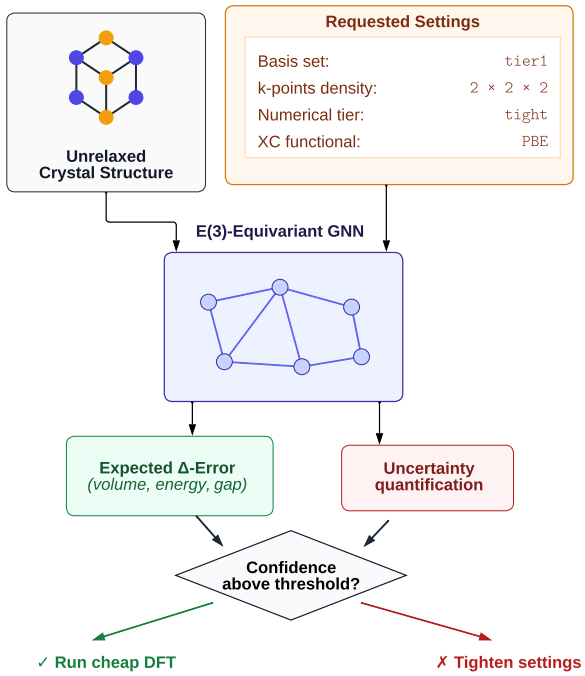


Figure 1. An $E(3)$-equivariant GNN operating as a recommender system. The network takes the unrelaxed crystal and the requested DFT settings as input, predicts the expected error, and reports an uncertainty estimate that flags structures it cannot reliably handle. In this way it can be used to find the cheapest DFT settings that meet a user-specified confidence threshold.

While Δ-EGNN focuses on accuracy corrections for specific molecular trajectories, our framework targets precision corrections for crystal structure relaxations across a diverse chemical space. Furthermore, whereas existing ML potentials are typically trained on small, localized trajectory datasets, our model is designed to tackle a wide range of binary semiconductors and make zero-shot predictions on unobserved structures. By conditioning directly on the computational settings and outputting a reliable measure of uncertainty, our approach ensures that researchers know exactly when the model's predictions can be trusted for both HPC convergence recommendations and database harmonization.

3. Proposed Method: Conditional Equivariant Δ-Learning

Standard graph neural networks are designed to predict absolute properties from a given structure. However, in our Δ-learning framework, the model must act as an error corrector or as a recommender system. To achieve this, we adapt two equivariant backbones which have been used successfully in materials science, EGNN (Satorras et al., 2021) and PaiNN (Schütt et al., 2021; Bechtel et al., 2025), into conditional architectures, which we term **CondEGNN** and **CondPaiNN**. We augment these backbones with physics-informed conditioning, custom loss functions for pathological data, and deep uncertainty extraction.

3.1. Architectural conditioning: FiLM and Attention

To ensure the CondEGNN and CondPaiNN models understand exactly which computational shortcuts were taken during the baseline calculations, we must feed the low-fidelity DFT settings directly into the network. Instead of merely appending these settings at the end of the model, which forces the message passing layers to operate blindly, we inject this prior using Feature-wise Linear Modulation (FiLM) (Perez et al., 2018).

FiLM applies an affine transformation to the network's internal features at every layer. We define a conditioning vector $\mathbf{c}_i$ containing the cheap DFT results (if available) and the specific precision settings. At layer l, a multi-layer perceptron projects this vector into a scaling factor $\boldsymbol{\gamma}_i^{(l)}$ and a shifting factor $\boldsymbol{\beta}_i^{(l)}$:

$$\boldsymbol{\gamma}_i^{(l)}, \boldsymbol{\beta}_i^{(l)} = \mathrm{MLP}^{(l)}(\mathbf{c}_i) \tag{1}$$

These factors then dynamically modulate the intermediate scalar node features $\mathbf{h}_i^{(l)}$ via element-wise multiplication ($\odot$) and addition. For the CondEGNN model, the update is:

$$\tilde{\mathbf{h}}_i^{(l)} = \boldsymbol{\gamma}_i^{(l)} \odot \mathbf{h}_i^{(l)} + \boldsymbol{\beta}_i^{(l)} \tag{2}$$

Equivariant Attention Mechanism: We assume not all atomic interactions contribute equally to the final property correction. We integrate a softmax-based, $E(3)$-equivariant attention mechanism into the message passing layers (Thölke & De Fabritiis, 2022) to allow the model to prioritize specific geometric bonds. For both CondEGNN and CondPaiNN, we compute a scalar score from the invariant inputs (e.g., Gaussian radial basis encodings and scalar node features) via a dedicated network ϕ_{att}, normalized across all neighbors of atom i:

$$\alpha_{ij} = \frac{\exp(\phi_{att}(\text{Invariant Inputs}))}{\sum_{k \in \mathcal{N}(i)} \exp(\phi_{att}(\text{Invariant Inputs}))}, \tag{3}$$

Since α_{ij} is derived entirely from invariant inputs, it safely modulates both the scalar feature aggregation and the equivariant coordinate updates without violating physical symmetries. Layer-wise formulations are detailed in Appendix A.

3.2. Handling Pathological Error Distributions

When predicting the difference between low- and high-fidelity relaxations, the error distributions appear pathological. As detailed in Appendix B, the target variables ($\Delta E, \Delta V, \Delta E_g$, which represent the deviation in total energy per atom, relaxed volume, and band gap, respectively) can be positive or negative and span multiple orders of magnitude. Furthermore, these distributions are strongly bimodal: they consist of either subtle deterministic convergence artifacts (when the cheap relaxation remains in the same local minimum) or larger structural shifts.

Traditional logarithmic metrics fail on signed data. Modifying the root-mean-squared-logarithmic error to operate on the absolute value of the target (i.e., Root-Mean-Squared-Logarithmic-Absolute-Error) handles scale well, but is completely sign-blind (i.e., predicting a $+1.0$ eV correction receives the exact same penalty as -1.0 eV) which can be catastrophic for thermodynamic modeling (Hyndman & Koehler, 2006; Hastie et al., 2009). The symmetric Mean Absolute Percentage Error (sMAPE) tracks relative error better, but suffers from numerical instability when highly converged baselines produce a Δ near zero (Makridakis, 1993). We mitigate this by adding $\epsilon =$ 1E-7 to the denominator of the sMAPE metric.

In order to optimize the network across all scales, we employ an inverse hyperbolic sine (asinh) transformed Mean Absolute Error (Burbidge et al., 1988). The asinh function smoothly passes through zero, naturally handles negative numbers, and behaves linearly for tiny numerical errors while applying logarithmic compression to larger structural outliers. We set the transition scale to $c =$ 1E-4:

$$\text{asinh-MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\text{asinh}\left(\frac{y_i}{c}\right) - \text{asinh}\left(\frac{\hat{y}_i}{c}\right)\right| \quad (4)$$

For many high-throughput applications, users simply want to filter data that has a precision error above a specific threshold (e.g., 10 meV/atom). For this use case, we adapt the concept of order-of-magnitude accuracy (Piskorz et al., 2026) into a custom metric we term Magnitude Accuracy (MagAcc). This tests whether the model correctly identifies the physical regime of the error by placing the absolute values of the predictions and targets into logarithmic bins $\mathcal{B} = \{10^{-3}, 10^{-2}, 10^{-1}, 1.0, 10.0\}$. We define MagAcc as the percentage of predictions that fall into the exact same order-of-magnitude bin B as the ground truth:

$$\text{MagAcc}(\%) = \frac{100}{N}\sum_{i=1}^{N}\mathbb{I}\left(B(|\hat{y}_i|, \mathcal{B}) = B(|y_i|, \mathcal{B})\right). \quad (5)$$

3.3. Extracting Uncertainty Estimates

A model that predicts an accurate Δ but is overconfident when encountering out-of-distribution (OOD) materials is dangerous for automated high-throughput screening. While CondEGNN and CondPaiNN are not intrinsically Bayesian, we extract rigorous UQ from them using two distinct branches.

Probabilistic UQ (Deep Ensembles and NLL): The most direct approach is to force the network to output a probability distribution rather than a point estimate. We modify the readout heads to output both a mean prediction μ and a variance σ^2, optimizing the network using the Gaussian Negative Log-Likelihood (NLL) loss (Nix & Weigend, 1994; Kendall & Gal, 2017):

$$\mathcal{L}_{\text{NLL}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\log(\sigma_i^2)}{2} + \frac{(y_i - \mu_i)^2}{2\sigma_i^2}\right)$$

It is critical to note that unlike experimental wet-lab data, DFT calculations are fundamentally deterministic. Therefore, the variance σ_i^2 learned via NLL does not represent true aleatoric (experimental) noise. Rather, it is intended to capture pseudo-aleatoric uncertainty, the distribution of the model's approximation error arising from its incomplete feature representation of the deterministic convergence artifacts.

To capture epistemic uncertainty (uncertainty about the model's own knowledge), we utilize Deep Ensembles (Lakshminarayanan et al., 2017). We train $M = 10$ models with different random initializations.

Distance-Based Latent UQ: Since NLL models are notoriously prone to variance collapse, where the network artificially inflates σ^2 to minimize loss on hard targets without actually learning the uncertainty, we also evaluate distance-based UQ metrics. Instead of relying on the regression head, we measure how different a new crystal structure and settings combination looks inside the model's latent space compared to the training data.

We extract the high-dimensional graph embeddings either just before message passing ($\mathbf{z}_{\text{pre_mp}}$) or at the final readout layer ($\mathbf{z}_{\text{readout}}$). We then calculate the Mahalanobis distance between a test sample embedding $\mathbf{z}_*$ and the training distribution (with mean $\boldsymbol{\mu}_{\text{train}}$ and covariance matrix $\boldsymbol{\Sigma}_{\text{train}}$) (Lee et al., 2018):

$$D_M(\mathbf{z}_*) = \sqrt{(\mathbf{z}_* - \boldsymbol{\mu}_{\text{train}})^T \boldsymbol{\Sigma}_{\text{train}}^{-1} (\mathbf{z}_* - \boldsymbol{\mu}_{\text{train}})}$$

As a non-parametric alternative, we also calculate the k-Nearest Neighbors (KNN) distance (Cover & Hart, 1967; Sun et al., 2022). If a structure is pushed far away from the

training manifold in the latent space, these metrics flag it as highly uncertain, regardless of what the final regression head predicts.

While the Mahalanobis and KNN distances are computationally cheap, they operate on globally pooled graph vectors, which inherently destroys the internal geometric distribution of the atoms. To evaluate whether preserving this internal 3D structure improves UQ, we also implemented rigorous node-level distance metrics using Maximum Mean Discrepancy (MMD) and Wasserstein-2 Optimal Transport (OT) (Gretton et al., 2012; Peyré & Cuturi, 2019). The full mathematical formulations, alongside the computational prefiltering strategies required to run these methods at scale, are detailed in Appendix C.

While continuous uncertainty scores (such as a latent KNN distance or ensemble variance) provide an excellent ranking of model confidence, deploying these networks in automated pipelines can often benefit from a binary decision: should the system trust this prediction or not? To transition our continuous metrics into an explicit Out-of-Distribution (OOD) detector, we can employ a strict, validation-based thresholding mechanism. The detection pipeline operates in three steps:

1. **Calibration:** After the model is fully trained, we perform a single forward pass to compute the chosen uncertainty score for every crystal structure in the validation set.
2. **Thresholding:** We establish a hard decision boundary by extracting the 95th percentile of this validation uncertainty distribution. This single scalar value defines the limit of the model's safe predictive regime.
3. **Inference and Rejection:** During deployment, the network calculates the uncertainty score for any incoming, unseen crystal. If this score exceeds the locked validation threshold, the system flags the input as OOD. Instead of failing silently, the pipeline rejects the neural network's Δ-prediction entirely and, in the case of the Recommender system, advises the user to run a convergence test.

4. Creating the multi-fidelity dataset

We constructed an automated high-throughput workflow to relax and calculate the properties of 4,220 binary semiconductors that exist experimentally in the Inorganic Crystal Structure Database (ICSD) and have previously been calculated with DFT in the AFLOW database (albeit with a single set of computational parameters) (Belsky et al., 2002; Curtarolo et al., 2012). Using the FHI-aims density functional theory (DFT) code, we performed full geometry relaxations followed by band-structure and density of states calculations (Blum et al., 2009). For every single material, we systematically varied the computational settings. We changed the basis-set size (from a very cheap *minimal* basis to a highly converged *tier2* basis), the k-point grid density (from 2 k-points per Å^{-1} to 8 k-points per Å^{-1}), and the numerical settings (*light*, *tight*, and *really tight*). In total, this generated roughly 65,000 unique data points mapping low-fidelity structural relaxations to their high-fidelity counterparts, establishing our target Δ-corrections for energy (ΔE), volume (ΔV), and band gap (ΔE_g). Note, our high-fidelity target data represents the numerical precision limit (converged basis-set size, numerical settings and k-point density) for the chosen xc functional. We intentionally fix the xc functional to isolate numerical convergence errors. Therefore, our target is not the absolute physical 'gold standard' for semiconductors (which would require computationally prohibitive methods like hybrid functionals or GW approximations), but rather the strict precision limit of standard high-throughput DFT.

The resulting data distribution is visualized in Appendix B and exhibits several pathologies. For example, the volume and energy errors form distinct bimodal distributions, with calculations run using the *minimal* basis set forming a peak of high errors. Furthermore, because changing the k-point mesh alters the integration grid without a variational guarantee, roughly 12% of the dataset contains negative ΔE values, and several volume changes jump by over 10 Å^3/atom, indicating the crystal's space group likely changed during the cheap relaxation.

Composition-Aware Data Splitting: To rigorously test our framework's ability to flag genuine out-of-distribution errors, we avoided random splitting, which often causes structural data leakage. Instead, we employed a strict composition-aware split based on source ICSD numbers, ensuring no two datapoints from the same crystal family appear across the train, validation, and test splits. However, while this prevents leakage, it does not cure the fundamental sampling bias inherent to materials databases (Sutton et al., 2019). Since standard gradient descent algorithms minimize the global error, models inevitably prioritize densely populated compositional regions, making explicit OOD detection critical for reliable evaluation.

5. Results

In order to rigorously evaluate our models, we use a strict validation lock-in protocol. For every experiment, the best model hyperparameters and checkpoint were selected exclusively on the validation set. The held-out test set was used only once per configuration, after lock-in, so the reported test metrics are not influenced by hyperparameter search.

Champion architectures and loss ablations: We first evaluated the models in "full" mode, where the network acts as an error-corrector by taking both the crystal structure and the low-fidelity DFT results as inputs. We compared our Graph Neural Networks against cross-validated tree-based models (Random Forest and Gradient Boosted Decision Trees) across three targets: total energy (ΔE), band gap (ΔE_g), and volume (ΔV). We ran an extensive 456-job ablation sweep (360 Single-Target and 96 Multi-Target configurations) to find the best graph architecture. We tested combinations of CondEGNN, CondPaiNN, equivariant attention, and various loss functions across multiple learning rates and hidden dimensions. The loss function was varied since some users will be more interested in reducing the mean error of the model (i.e., minimize the MAE) while others will want good performance across all scales (asinh-MAE, sMAPE). NLL-trained models substantially underperformed sMAPE and asinh-MAE-trained ones across all targets, likely because the heavy-tailed delta distributions violate the Gaussian assumption underlying NLL, and failed catastrophically when combined with the attention mechanism (test sMAPE $> 60\%$; Appendix C.3).

The best full-mode models (Table 1), those selected by having the highest validation metrics, achieved a 7.2% ΔE test sMAPE, for multi-target (MT) models trained on all targets simultaneously, and 8.1% for single-target (ST) energy models, beating the Random Forest baseline (9.4%). Among the sMAPE-trained models, no single architecture dominated all three targets. For ST prediction, the CondEGNN backbone with FiLM conditioning won both the energy and band gap corrections, while the CondPaiNN architecture with attention won the volume correction. For MT prediction, no single architecture dominated all three targets: CondEGNN with attention and FiLM achieved the best energy sMAPE (7.2%), while CondPaiNN with attention achieved the best band gap (30.9%) and volume (23.6%) sMAPE. We report the exhaustive per-architecture ablations in Appendix C.3 (MT), Appendix C.4 (ST full mode) and Appendix C.5 (ST recommender mode). The performance gap between the top architectures within each (paradigm, target) cell was generally small (under 2% val sMAPE in most cases). For CondPaiNN, removing the FiLM conditioning (which injects DFT settings via per-layer feature-wise linear modulation) raised energy sMAPE from 14.6% to 21.2% in MT mode under NLL loss (Appendix C.3). In this no-FiLM variant, DFT settings only enter the model through readout concatenation; the message-passing layers themselves see no conditioning. Attention helped substantially in MT mode but offered little benefit for ST energy prediction.

Zero-Cost recommender systems: While correcting a low-fidelity calculation is useful, a true recommender system must predict the error before any DFT calculation is run, relying solely on the initial unrelaxed crystal structure and the requested computational settings (e.g., basis set and k-point density). For the RF baseline (RF-PreCalc), we stripped the model of all results from the lower fidelity DFT calculation that were used as tabular features. The GNN recommender was similarly restricted, receiving only the 3D atomic coordinates, atomic numbers, and computational precision settings (k-point density, basis set size and numerical settings).

As shown in Table 2, the best recommender GNN (ST-GNN-sMAPE), a CondEGNN+FiLM trained with sMAPE loss, beats the RF-PreCalc baseline on sMAPE across all three targets, with the largest gap on energy (8.1% vs. 11.0%). The MAE comparison tells a different story: the single-model ST CondEGNN+FiLM ties RF-PreCalc on energy MAE (0.016 eV/atom in both) and loses to it on band gap (0.063 vs. 0.056 eV) and volume (0.162 vs. 0.145 $Å^3$). In terms of MagAcc, the GNN performs better by 1-2 percentage points across all three targets. The case for using a GNN over a tree baseline in the Recommender system mode therefore rests on relative error and order-of-magnitude binning, not on absolute error. The RF likely is competitive with the equivariant GNNs due to the way it naturally partitions discrete, non-linear categorical variables, such as the basis-set tier and numerical settings in FHI-aims.

Comparing the recommender models to the full mode error-correctors (Table 1) reveals an interesting pattern: the single-model ST-GNN-sMAPE recommender actually achieves a slightly lower test sMAPE on energy than its full-mode counterpart (8.1% recommender vs. 8.4% full mode). The MT ensemble in full mode remains the overall best (7.2% test sMAPE on energy), but the single-model comparison suggests that low-fidelity DFT energies do not provide as much signal as one might expect. We hypothesize that low-fidelity DFT energies contain systematic biases that the network can over-rely on; when forced to predict from geometry and settings alone, the model learns a more direct mapping. We do not see the same effect for band gap and volume, where the full-mode models retain a clear lead.

Uncertainty quantification (UQ): A model that confidently predicts an incorrect property for a novel chemistry is dangerous for automated high-throughput screening. We evaluated two branches of UQ methods: probabilistic methods (Deep Ensembles and NLL) and distance-based latent space methods (KNN and Mahalanobis). To quantify UQ quality, we measured the Spearman rank correlation (ρ) between the predicted uncertainty and the actual absolute error on the test set. Deep Ensembles performed best: in full mode, an ensemble of 10 models achieved $\rho = 0.898$ on energy. The NLL models also performed well ($\rho = 0.826$), while the distance-based methods underperformed both. Figure 2 shows the corresponding sparsification curves: as the most-uncertain predictions are progressively removed from the

Table 1. Performance of the best models for the full mode Δ-learning task, where the network has access to the low-fidelity DFT calculation results as input. Each row corresponds to one model variant; for ST-GNN rows, each cell reports the val-sMAPE-winning single-target model trained on that target with the indicated loss. The MT-GNN rows correspond to a single multi-target model. Ensembles are top-5 averages over the best models on the validation set.

	Energy (ΔE)			**Band gap (ΔE_g)**			**Volume (ΔV)**		
Model	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**
RF-Full	0.013	9.4%	95.8%	0.049	31.0%	86.9%	0.129	26.7%	89.4%
ST-GNN-sMAPE	0.017	8.4%	96.7%	0.054	29.4%	88.1%	0.144	24.8%	91.3%
MT-GNN-sMAPE	0.014	8.1%	96.6%	0.056	30.9%	87.1%	0.136	23.6%	**91.7%**
ST-GNN Ens Top-5	0.015	8.5%	96.4%	0.051	32.2%	87.3%	**0.123**	24.6%	90.8%
MT-GNN Ens Top-5	**0.011**	**7.2%**	**97.3%**	**0.046**	**29.5%**	**87.5%**	0.128	**23.4%**	90.8%

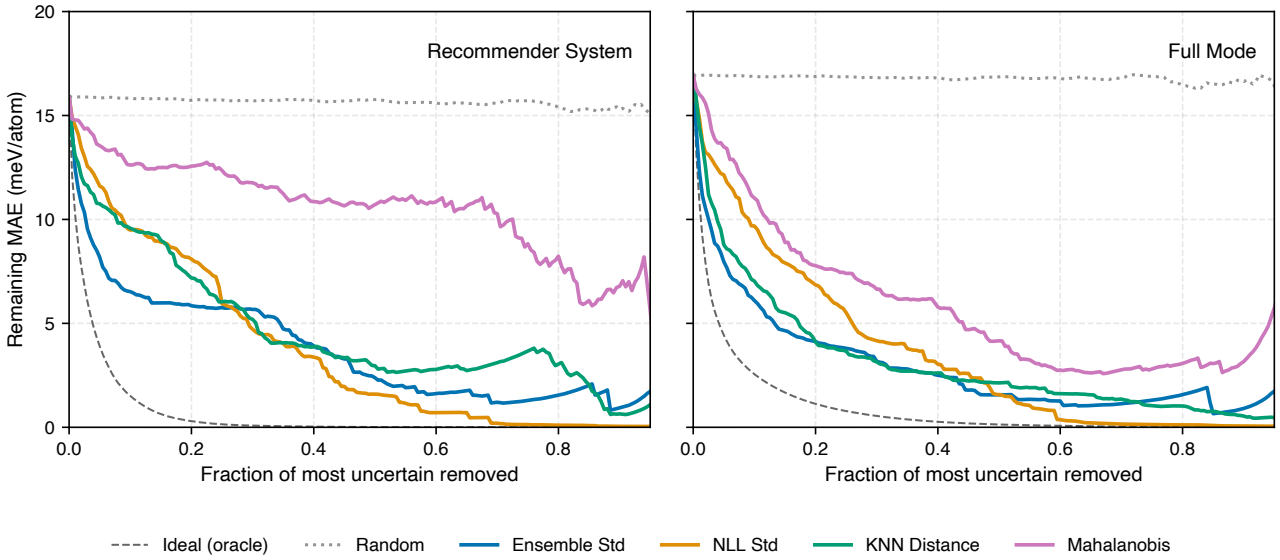


Figure 2. Sparsification curves for energy error prediction for the best ST recommender model (A) and ST full model (B). For different UQ methods, the most *uncertain* predictions are removed and the Mean Absolute Error (MAE) of the remaining dataset is shown.

test set, the remaining MAE drops toward the oracle baseline, with NLL tracking the oracle most closely in both full and recommender modes for low fractions. For higher fractions the NLL performs the best for the Recommender and full modes. Extended sparsification curves on sMAPE and an ablation over ensemble size are reported in Appendix F.3 and F.4 respectively.

Probing the latent space to extract uncertainty: While Deep Ensembles are highly accurate, running multiple graph networks for every prediction is computationally expensive. Therefore, we investigated distance-based UQ. By measuring how different a new crystal and DFT settings combination looks in the network's latent space compared to the training data, we can flag out-of-distribution structures in a single forward pass. The depth at which we extract the embeddings matters tremendously. For the recommender mode, the best UQ signal was found at the final *readout* layer using a k-Nearest Neighbors (KNN) distance ($\rho = 0.575$). However, when the model operates in full mode with access to continuous cheap DFT features, the latent space becomes more Gaussian, allowing the parametric Mahalanobis distance to dominate ($\rho = 0.619$). Figure 3 shows the UMAP projection of the Mahalanobis readout embeddings. The figure shows distinct clusters by basis-set size, confirming that FiLM successfully conditions the latent geometry on the computational settings prior. The quantitative impact of this conditioning is demonstrated in the ablation studies in Appendix C.3.

While Mahalanobis and KNN operate on globally pooled graph vectors, global pooling inherently destroys the internal geometric distribution of the atoms. We also tested node-level distance metrics, Maximum Mean Discrepancy (MMD) and Wasserstein-2 Optimal Transport (OT), to evaluate whether preserving the per-atom geometric distribution improves UQ. Beyond a marginal gain on volume prediction in full mode, neither method outperformed the sim-

Table 2. Performance of the best models for the Recommender system task, where the network must predict the Δ-correction from the unrelaxed crystal structure and the requested computational settings alone, without any low-fidelity DFT inputs. Conventions follow Table 1.

	Energy (ΔE)			Band gap (ΔE_g)			Volume (ΔV)		
Model	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**
RF-PreCalc	0.016	11.0%	95.2%	**0.056**	35.0%	85.0%	**0.145**	30.4%	88.1%
ST-GNN-sMAPE	0.016	**8.1%**	**96.7%**	0.063	**34.6%**	**86.1%**	0.162	28.7%	89.3%
MT-GNN-sMAPE	0.019	9.8%	96.3%	0.067	36.0%	85.7%	0.163	**27.5%**	**89.4%**
MT-GNN-AsinhL1	0.026	11.9%	95.3%	0.072	35.4%	85.4%	0.193	30.7%	88.7%
MT-GNN-L1	0.021	24.8%	92.7%	0.065	51.4%	78.4%	0.168	34.7%	87.2%
ST-GNN Ens Top-5	0.018	9.5%	96.2%	0.062	37.1%	84.7%	0.150	29.4%	89.1%
MT-GNN Ens Top-5	**0.015**	8.7%	96.4%	0.060	35.2%	84.9%	0.157	28.5%	89.1%

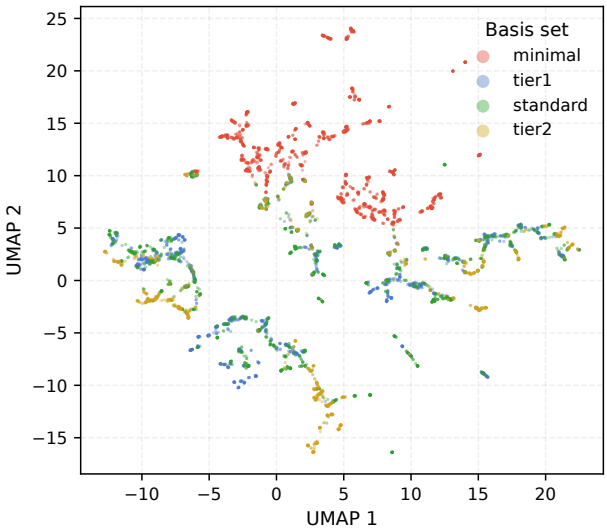


Figure 3. UMAP projection of the final readout embeddings, colored by the basis-set size used in the low-fidelity calculation.

pler graph-pooled distances (full results are given in Appendix F). The two-stage filtering pipeline required to scale these methods is reported in Appendix D.

Characterizing the OOD Rejections: In order to understand which structure and settings combinations the network classifies as uncertain, we characterized the populations of the accepted versus rejected datapoints (95th-percentile) using the readout node embeddings using the KNN distance since this approach had good correlation for both the full and recommender models. Across all models, the rejected bucket exhibited ground-truth Δ-corrections that were two to three times larger than the accepted bucket. This proves the UQ metric correctly flags the largest physical deviations. Furthermore, the rejections are heavily driven by computational and compositional extremes. In the recommender models, over 90% of rejected predictions were calculations run with the *minimal* basis set, the model correctly identifies that these calculations with a single basis function per occupied atomic orbital, are very crude approximations to the converged basis set values. In terms of the atomic composition of the crystals, the OOD detection acts as a strict epistemic guardrail, aggressively rejecting test structures containing rare transition metals (e.g., Ru, Os, Co) that comprised less than 1.5% of the training data. A comprehensive breakdown of OOD behavior across crystal systems is provided in Appendix H.

6. Discussion and Conclusion

The primary goal of this work was to determine if we could safely bypass the computationally expensive convergence testing required for high-fidelity Density Functional Theory (DFT) calculations. By framing this as a Δ-learning problem, we demonstrated that equivariant neural networks where the computational settings are passed in via FiLM can act as accurate, zero-cost error correctors and recommender systems. Our results show that the method used to integrate computational priors dictates model success. By injecting low-fidelity settings directly into the intermediate layers using FiLM, the networks learned to dynamically adapt to the computational settings, outperforming tabular baselines and naive conditioning approaches across all three targets on relative error metrics.

In high-throughput screening, a confident wrong answer is worse than no answer. Deep Ensembles ranked errors most reliably but they require ten forward passes and training ten models which can be prohibitively expensive computationally. We demonstrated that distance-based latent UQ offers a highly effective, single-pass alternative. Specifically, using a simple k-Nearest Neighbors (KNN) distance on the final readout layer provides an excellent flag for out-of-distribution (OOD) structures, offering a practical guardrail for automated screening pipelines.

To accelerate the development of robust error-correcting models, we are entirely open-sourcing the novel dataset

generated for this study, which comprises over 60,000 multi-fidelity relaxation targets across 4,220 binary semiconductors. By combining $E(3)$-equivariant networks with robust UQ, our framework approaches high-precision DFT while systematically flagging when it enters extrapolative regimes, allowing researchers to safely leverage large, heterogeneous mixed-fidelity data collections.

Acknowledgements

Work was carried out in part by the Max Planck Graduate Center for Quantum Materials. Partial funding is appreciated from the German Science Foundation (DFG) through the CRC FONDA (project 414984028) and the NIFD consortium FAIRmat (project 460197019), and the European Union's Horizon 2020 research and innovation program under the grant agreement No. 951786 (NOMAD CoE). D. S. acknowledges support by the IMPRS for Elementary Processes in Physical Chemistry. Computing time for the FHI-aims data was granted by the Resource Allocation Board and provided on the supercomputers Lise and Emmy at NHR@ZIB and NHR@Göttingen as part of the NHR infrastructure (bep00098). ML models were trained on the HPC system Raven at the Max Planck Computing and Data Facility. We thank Jonathan Godwin, Tim Bechtel, Luca Ghinghelli, Martin Kuban, Freja Nordsiek, and Sebastian Kehl for valuable discussions.

Software and Data

The dataset consisting of multi-fidelity DFT relaxation outputs and processed graph data is available at https://huggingface.co/datasets/dtsdts/multifidelity-dft-deltas. DFT workflow, parsing, model training, evaluation, analysis and plotting scripts are available at https://github.com/speckhard/multifidelity-dft-deltas.

Impact Statement

This paper presents work whose goal is to advance the intersection of machine learning and the physical sciences. The societal consequence of our framework and the models we provide is firstly to reduce the computational cost required to achieve high-fidelity DFT calculations and secondly to help ML practitioners use and repurpose DFT data calculated with heterogeneous settings. We open-source this multi-fidelity DFT dataset and model code to promote transparency and reproducibility in the AI for Physics community.

References

Bechtel, T., Speckhard, D. T., Godwin, J., and Draxl, C. Band-gap regression with architecture-optimized message-passing neural networks. *Chemistry of Materials*, 37(4):1358–1369, 2025.

Belsky, A., Hellenbrandt, M., Karen, V., and Luksch, P. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science*, 58(3):364–369, 2002.

Blum, V., Gehrke, R., Hanke, F., Havu, P., Havu, V., Ren, X., Reuter, K., and Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Computer Physics Communications*, 180(11):2175–2196, 2009.

Burbidge, J. B., Magee, L., and Robb, A. L. Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127, 1988.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012.

Draxl, C. and Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, 2019.

Giustino, F. *Materials modelling using density functional theory: properties and predictions*. Oxford University Press, 2014.

Godwin, J., Schaarschmidt, M., Gaunt, A. L., Sanchez-Gonzalez, A., Bapst, V., Kohli, P., and Battaglia, P. W. Simple gnn regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations*, 2021.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Grumet, M., von Scarpatetti, C., Bucko, T., and Egger, D. A. Delta machine learning for predicting dielectric properties and raman spectra. *The Journal of Physical Chemistry C*, 128(15):6464–6470, 2024.

Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Huo, J. and Dong, H. δ-egnn method accelerates the construction of machine learning potential. *The Journal of Physical Chemistry Letters*, 16(8):2080–2088, 2025.

Hyndman, R. J. and Koehler, A. B. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

Jain, A., Shin, Y., and Persson, K. A. Computational predictions of energy materials using density functional theory. *Nature Reviews Materials*, 1(1):1–13, 2016.

Janssen, J., Makarov, E., Hickel, T., Shapeev, A. V., and Neugebauer, J. Automated optimization and uncertainty quantification of convergence parameters in plane wave density functional theory calculations. *npj Computational Materials*, 10(1):263, 2024.

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

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, volume 30, 2017.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in neural information processing systems*, volume 31, 2018.

Lejaeghere, K., Bihlmayer, G., Björkman, T., Blaha, P., Blügel, S., Blum, V., Caliste, D., Castelli, I. E., Clark, S. J., Dal Corso, A., et al. Reproducibility in density functional theory calculations of solids. *Science*, 351 (6280):aad3000, 2016.

Makridakis, S. Accuracy measures: theoretical and practical concerns. *International journal of forecasting*, 9(4):527–529, 1993.

Marks, L. Predictive mixing for density functional theory (and other fixed-point problems). *Journal of Chemical Theory and Computation*, 17(9):5715–5732, 2021.

Nix, D. A. and Weigend, A. S. 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.

Pauletti, M., Rybkin, V. V., and Iannuzzi, M. Subsystem density functional theory augmented by a delta learning approach to achieve kohn–sham accuracy. *Journal of Chemical Theory and Computation*, 17(10):6423–6431, 2021.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Peyré, G. and Cuturi, M. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

Piskorz, J., Kobalczyk, K., and van der Schaar, M. Eliciting numerical predictive distributions of llms without autoregression. *arXiv preprint arXiv:2603.02913*, 2026.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: The δ-machine learning approach. *Journal of chemical theory and computation*, 11(5):2087–2096, 2015.

Ruth, M., Gerbig, D., and Schreiner, P. R. Machine learning of coupled cluster (t)-energy corrections via delta (δ)-learning. *Journal of Chemical Theory and Computation*, 18(8):4846–4855, 2022.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schunck, N., McDonnell, J., Higdon, D., Sarich, J., and Wild, S. Uncertainty quantification and propagation in nuclear density functional theory. *The European Physical Journal A*, 51(12):169, 2015.

Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

Speckhard, D. T., Carbogno, C., Ghiringhelli, L. M., Lubeck, S., Scheffler, M., and Draxl, C. Extrapolation to the complete basis-set limit in density-functional theory using statistical learning. *Physical Review Materials*, 9 (1):013801, 2025.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pp. 20827–20840. PMLR, 2022.

Sutton, C., Boley, M., Ghiringhelli, L. M., Rupp, M., Vreeken, J., and Scheffler, M. Identifying domains of applicability of machine learning models for materials science. *Nature communications*, 10(1):1204, 2019.

Tabor, D. P., Roch, L. M., Saikin, S. K., Kreisbeck, C., Sheberla, D., Montoya, J. H., Dwaraknath, S., Aykol, M., Ortiz, C., Tribukait, H., et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nature reviews materials*, 3(5):5–20, 2018.

Thölke, P. and De Fabritiis, G. Torchmd-net: Equivariant message passing neural network with property-aware pooling. In *International Conference on Learning Representations*, 2022.

A. Equivariant Attention in CondPaiNN

As discussed in the main text, adapting the $E(3)$-equivariant attention mechanism for the CondPaiNN architecture requires careful handling of its separate scalar and vector tracks. Since PaiNN does not compress interactions into a single invariant message vector like EGNN, we cannot compute attention post-message-creation.

Instead, we compute the attention logits directly from the raw invariant edge features and node scalars. We pass the concatenation of the Gaussian RBF distance encoding ($\mathbf{e}_{ij}^{\text{RBF}}$) and the scalar features of both the target and source atoms ($\mathbf{s}_i, \mathbf{s}_j$) into a two-layer Multi-Layer Perceptron (MLP) with a SiLU activation, denoted as ϕ_{att}:

$$\alpha_{ij} = \frac{\exp\left(\phi_{\text{att}}(\mathbf{e}_{ij}^{\text{RBF}}, \mathbf{s}_i, \mathbf{s}_j)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\phi_{\text{att}}(\mathbf{e}_{ik}^{\text{RBF}}, \mathbf{s}_i, \mathbf{s}_k)\right)} \tag{6}$$

These normalized, invariant scalar weights (α_{ij}) are then used to modulate the aggregation of both the scalar and the equivariant vector messages. The updated message passing equations for CondPaiNN become:

$$\Delta\mathbf{s}_i = \sum_{j\in\mathcal{N}(i)} \alpha_{ij}\Big(W_s(\mathbf{e}_{ij}^{\text{RBF}}) \odot \mathbf{s}_j\Big) \tag{7}$$

$$\Delta\mathbf{v}_i = \sum_{j\in\mathcal{N}(i)} \alpha_{ij}\left[W_{v_1}(\mathbf{e}_{ij}^{\text{RBF}}) \odot \mathbf{v}_j + \left(W_{v_2}(\mathbf{e}_{ij}^{\text{RBF}}) \odot \mathbf{s}_j\right)\hat{\mathbf{r}}_{ij}\right] \tag{8}$$

where $\hat{\mathbf{r}}_{ij}$ is the normalized directional unit vector between atom i and atom j, and W_s, W_{v_1}, W_{v_2} are radial filter networks. Since α_{ij} is strictly an invariant scalar, multiplying the equivariant vector messages $\Delta\mathbf{v}_i$ by α_{ij} preserves the $E(3)$ equivariance of the network. The subsequent PaiNN update block, including our novel FiLM conditioning and the scalar-vector interactions, remains mathematically unchanged.

B. Dataset target distributions

As discussed in the main text, our Δ-learning framework optimizes for the difference between low-fidelity and high-fidelity DFT calculations. To fully appreciate why we had to introduce loss functions like the asinh-MAE, it is necessary to look at the raw distributions of these target variables: the corrections for total energy (ΔE), band gap (ΔE_g), and volume (ΔV).The data is highly pathological. Since changing the k-point mesh alters the integration grid without a variational guarantee, we see a significant number of negative Δ values (where the cheap calculation was actually closer to the "true" absolute minimum than an intermediate step). Table 3 shows the unified distributions for all three targets. Notice that the errors are often distinctly bimodal or highly skewed. Figure 4 shows that one of the two peaks in both the band gap and energy delta distributions can be attributed to the minimal basis set size.

C. Hyperparameter Grid and Reproducibility

To ensure our findings are fully reproducible and to establish that our graph networks were not given an unfair advantage over the tree-based baselines, we utilized strict, deterministic hyperparameter grids for both the Single-Target (ST) and Multi-Target (MT) sweeps.

All graph neural networks were trained using the Adam optimizer with a batch size of 32 for a maximum of 800 epochs. Early stopping was enforced with a patience of 10 epochs on the validation loss. The fixed and swept hyperparameters for the graph models are detailed in Table 4.

C.1. Validation-Set Metrics for Champion Models

For completeness, Table 5 reports the validation-set metrics for the champion models presented in Tables 1 and 2. Models were selected on validation sMAPE; the val numbers shown here are those that drove that selection. The val–test gap is

Table 3. Unified distribution of the target Δ-corrections across the 4,220 binary semiconductors (59,235 total correction pairs). Values represent the percentage of the dataset falling into each bin.

Range	ΔE **(eV/atom)**	ΔE_g **(eV)**	ΔV **(Å^3/atom)**
$< -10^3$	0.01%	0.00%	0.00%
-10^3 to -10^2	0.00%	0.00%	0.10%
-10^2 to -10	0.01%	0.00%	0.21%
-10 to -1	0.00%	14.47%	5.53%
-1 to -10^{-1}	0.04%	15.16%	7.38%
-10^{-1} to -10^{-2}	0.19%	23.75%	5.63%
-10^{-2} to -10^{-3}	2.37%	10.37%	3.34%
-10^{-3} to -10^{-4}	6.22%	3.25%	0.57%
-10^{-4} to 0	1.66%	0.31%	0.07%
Negative (< 0)	**10.50%**	**67.31%**	**22.83%**
0 to 10^{-4}	0.62%	0.34%	0.40%
10^{-4} to 10^{-3}	1.68%	2.91%	1.23%
10^{-3} to 10^{-2}	13.19%	9.24%	10.09%
10^{-2} to 10^{-1}	31.78%	12.17%	27.97%
10^{-1} to 1	24.06%	6.41%	18.49%
1 to 10	18.17%	1.61%	18.48%
> 10	0.00%	0.00%	0.50%

generally small (< 2 percentage points on sMAPE for most rows), confirming that our composition-aware splitting and validation lock-in protocol prevent overfitting to the validation set.

C.2. Auxiliary regularization: Noisy Nodes

We investigated whether the Noisy Nodes denoising objective (Godwin et al., 2021) could regularize CondPaiNN training on our limited multi-fidelity dataset. Standard EGNN backbones natively update and predict atomic coordinates and so implicitly learn geometric gradients; PaiNN, however, relies on separate scalar and vector tracks and does not intrinsically update coordinates during message passing. We therefore injected Gaussian noise $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$ into the starting atomic coordinates and added a linear readout head $\mathbf{W}_{vec}$ that predicts the displacement from the final equivariant vector features:

$$\hat{\boldsymbol{\epsilon}}_i = \mathbf{W}_{vec}\mathbf{v}_i \tag{9}$$

The network was trained to minimize the mean squared error between $\boldsymbol{\epsilon}_i$ and $\hat{\boldsymbol{\epsilon}}_i$, optimized jointly with the primary Δ-learning targets. The denoising rows of Table 6 report the result: the auxiliary objective substantially degrades performance on all three targets, regardless of the primary loss. We hypothesize that while Noisy Nodes is a useful regularizer when learning absolute quantum-mechanical forces, in a Δ-learning framework it diverts capacity away from the highly specific, physically meaningful structural relaxation pathways the model is trying to predict.

C.3. Multi-Target (MT) Ablations

The MT ablation shown in Table 6 reveals four effects. First, FiLM conditioning helps substantially: CondPaiNN trained without FiLM achieves 21.2% energy sMAPE under NLL loss, while adding FiLM drops this to 14.6%; the same monotone improvement holds for band gap (53.9% to 38.3%) and volume (41.0% to 29.8%). Second, attention combined with sMAPE loss further reduces energy sMAPE from 14.6% (CondPaiNN + FiLM, NLL) to 8.4% (CondPaiNN + FiLM + Attention (Att), sMAPE), with the CondEGNN family showing the same pattern (10.1% NLL to 8.1% sMAPE+Att+FiLM). Third, the denoising auxiliary objective consistently degrades performance and we do not recommend it for Δ-learning on multi-fidelity DFT data. Fourth, sMAPE beats asinh-MAE within each (architecture, loss) cell on energy and is comparable on the other two targets, motivating the use of sMAPE as the default loss for our champion models. The dominant axis is attention +

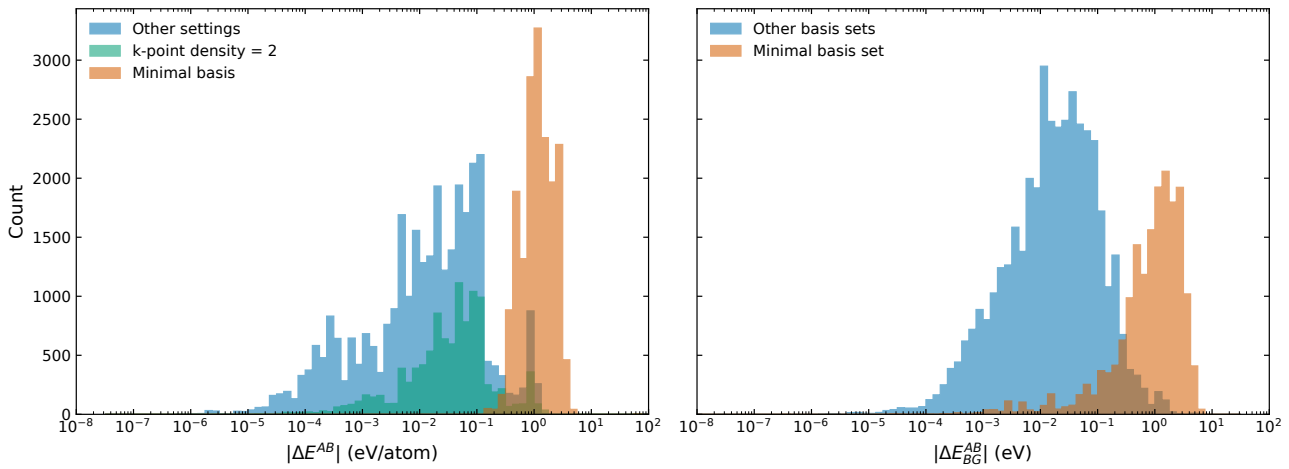


Figure 4. Logarithmic histograms of $|\Delta E|$ (left) and $|\Delta E_g|$ (right), color-coded by baseline precision. The distinct secondary peaks of large errors are dominated almost entirely by relaxations performed using the minimal basis set.

Table 4. Hyperparameter grid for the ST-GNN and MT-GNN ablation sweeps. The champion models were selected strictly based on validation sMAPE performance from this grid.

Hyperparameter	**Values / Search Space**
Fixed Architecture Settings	
Number of Message Passing Layers	3 (PaiNN variants), 4 (EGNN variants)
Batch Size	32
Maximum Epochs	800
Radial Cutoff	5.0 Å
Number of RBF Features	20
Grid Search Space (360 ST Jobs, 96 MT Jobs)	
Hidden Dimension Sizes	32, 64, 128, 256
Learning Rates	1E-3, 5E-4, 1E-4
Loss Functions	`smape`, `asinh_l1`, `l1`, `nll`
Random Seeds	42, 123, 7

sMAPE-loss together; other axes shift performance by 1–6 percentage points at most.

C.4. Single-Target (ST) Ablations: Full Mode

In contrast to the multi-target (MT) regime, where attention substantially improved volume prediction (Table 6), the single-target full mode regime tells a different story: attention worsens energy and band gap predictions when added on top of either backbone, while FiLM conditioning consistently helps. Table 7 reports the test-set metrics across six combinations of backbone and conditioning method each trained under three losses (sMAPE, AsinhL1, RMSLAE), with the val-sMAPE-winning hyperparameter configuration (learning rate, hidden dimension size, message passing steps), selected per cell. Note, the RMSLAE has very poor results for MAE and sMAPE metrics. The Root Mean Squared Logarithmic Absolute Error operates on the absolute value of signed targets and is therefore sign-blind, as discussed in Section 3.2. Test sMAPE exceeds 25% on energy, 75% on band gap, and 55% on volume across every architecture, confirming that it is only useful for order of magnitude binning.

The data supports three findings. First, FiLM conditioning consistently helps: CondEGNN base (no FiLM) achieves 8.8% energy sMAPE under sMAPE loss, while adding FiLM (CondEGNN + FiLM) drops this to 8.4%, with a comparable improvement on band gap (32.5% to 29.4%). The CondEGNN + FiLM with sMAPE loss is the best ST energy and band gap performer overall, winning four of nine metric cells. Second, attention hurts in the ST regime on the scalar targets: adding

Table 5. Validation-set metrics for the champion models from Tables 1 (full mode) and 2 (Recommender Mode). We omit val MagAcc since model selection used val sMAPE and val MAE only. Bold marks per-column best within each mode (ties bolded).

	Energy (ΔE)		**Band gap (ΔE_g)**		**Volume (ΔV)**	
Model	**val MAE**	**val sMAPE**	**val MAE**	**val sMAPE**	**val MAE**	**val sMAPE**
full mode						
RF-Full	0.009	8.1%	0.047	29.2%	**0.120**	26.3%
ST-GNN-sMAPE	0.013	6.4%	0.053	28.3%	0.141	25.4%
MT-GNN-sMAPE	0.010	6.9%	0.053	28.5%	0.145	24.1%
ST-GNN Ens Top-5	0.011	7.5%	0.049	30.0%	0.122	24.1%
MT-GNN Ens Top-5	**0.008**	**6.2%**	**0.044**	**27.2%**	0.127	**22.5%**
Recommender Mode						
RF-PreCalc	**0.012**	9.7%	**0.053**	33.3%	**0.135**	29.4%
ST-GNN-sMAPE	**0.012**	**7.8%**	0.063	32.5%	0.157	28.2%
MT-GNN-sMAPE	0.015	8.6%	0.060	33.2%	0.151	**26.1%**
MT-GNN-AsinhL1	0.017	10.1%	0.064	**31.5%**	0.180	28.1%
MT-GNN-L1	0.013	22.3%	**0.053**	48.3%	0.160	33.5%
ST-GNN Ens Top-5	0.013	8.2%	0.057	34.2%	0.140	27.5%
MT-GNN Ens Top-5	0.013	8.1%	0.054	32.0%	0.144	26.6%

attention to the EGNN backbone increases energy sMAPE from 8.8% to 12.4% (no FiLM) or to 15.6% (with FiLM), and a similar pattern holds for the PaiNN backbone. This is consistent with the multi-target results: attention's contribution depends on cross-target gradient flow that is absent when training a single target in isolation. Third, the volume target is an exception, the CondEGNN + Att + FiLM configuration trained on AsinhL1 produces the lowest test volume sMAPE (24.6%) and MAE (0.121), suggesting attention captures geometric structure useful for volume prediction even without cross-target signal.

The best ST models for ΔE and ΔV are shown in Figure 5. To intuitively visualize this performance across multiple orders of magnitude, the figure displays the logarithm of the absolute value of the target with a 1E-7 epsilon value added to the argument to avoid division by zero. Since the target Δ-corrections span several orders of magnitude, the tight clustering of predictions along the diagonal indicates the ability of the models to predict the correct order of magnitude of the error (MagAcc) and percentage error across the different scales.

C.5. Single-Target (ST) Ablations: Recommender Mode

The recommender variant of the ST regime trains models without the low-fidelity DFT outputs, predicting the Δ-correction from the unrelaxed crystal structure and the requested computational settings alone. Since this paradigm has fewer input features and a different model interface, fewer architecture and loss combinations were swept than in the full mode Table 8 reports the test-set metrics across the architecture and loss combinations we evaluated.

The recommender ST results closely mirror the full mode ST findings: CondEGNN + FiLM under sMAPE loss is the clear winner on energy and band gap, achieving the lowest test sMAPE in both (8.1% and 34.6%, respectively) and the highest MagAcc. As in full mode, adding attention degrades scalar-target performance: CondEGNN + Att (no FiLM) reaches 15.2% energy sMAPE and CondEGNN + Att + FiLM reaches 14.2%, both substantially worse than the no-attention CondEGNN + FiLM baseline at 8.1%. The volume target again partially benefits from attention, with CondPaiNN + Att + FiLM (AsinhL1) achieving the lowest volume MAE (0.155). Since the recommender mode lacks the low-fidelity DFT energy as an input, all sMAPE values are higher in absolute terms than the corresponding full mode entries, the GNN must predict the correction from geometry and settings alone, without an anchor.

D. Node-Level Uncertainty Quantification: MMD and Optimal Transport

Standard distance-based UQ metrics (like Mahalanobis and KNN) operate on the `readout` or `post_mp` embeddings after a global pooling operation (e.g., mean or sum pooling). While computationally efficient, global pooling washes away the

Table 6. Multi-target (MT) architecture and loss ablation on the test set, comprehensive view. Each row reports the best validation model (in terms of optimization metric) in terms of hyperparameter configuration for a given (architecture, conditioning) combination. The table includes catastrophic-failure rows (NLL paired with attention; RMSLE) explicitly so the reader can see which combinations break down.

		Energy (ΔE)			Band gap (ΔE_g)			Volume (ΔV)		
Architecture	**Loss**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**
CondPaiNN family										
CondPaiNN base	NLL	0.023	21.2%	93.2%	0.074	53.9%	77.6%	0.243	41.0%	84.4%
CondPaiNN + Denoise	NLL	0.030	35.9%	83.9%	0.104	70.9%	69.5%	0.277	52.3%	80.0%
CondPaiNN + FiLM	NLL	0.017	14.6%	95.7%	0.054	38.3%	85.3%	0.148	29.8%	88.5%
CondPaiNN + FiLM	sMAPE	0.028	14.6%	94.6%	0.066	42.2%	81.6%	0.174	35.4%	86.9%
CondPaiNN + FiLM	AsinhL1	0.021	21.9%	92.3%	0.051	40.7%	83.4%	0.141	31.3%	88.4%
CondPaiNN + FiLM	RMSLE	0.305	131.4%	44.8%	0.449	145.1%	31.6%	0.951	158.8%	40.4%
CondPaiNN + FiLM + Denoise	NLL	0.029	28.7%	88.6%	0.085	64.7%	72.0%	0.229	48.3%	81.6%
CondPaiNN + FiLM (scalar)	sMAPE	0.015	8.5%	**96.8%**	0.057	31.4%	87.3%	0.207	32.1%	89.2%
CondPaiNN + FiLM (scalar)	AsinhL1	0.020	10.2%	95.3%	0.051	**29.0%**	88.0%	0.130	24.9%	90.4%
CondPaiNN + FiLM (scalar)	L1	0.017	13.9%	95.0%	**0.048**	29.4%	87.6%	0.140	25.8%	90.3%
CondPaiNN + FiLM (scalar)	NLL	0.058	55.8%	72.5%	0.129	90.1%	61.5%	0.340	80.1%	69.3%
CondPaiNN + FiLM + Att	sMAPE	**0.012**	8.4%	96.4%	0.056	30.9%	87.1%	0.136	**23.6%**	**91.7%**
CondPaiNN + FiLM + Att	AsinhL1	0.023	10.3%	96.2%	0.060	30.6%	86.5%	0.152	25.4%	90.0%
CondPaiNN + FiLM + Att	L1	0.015	17.1%	94.6%	0.051	34.7%	85.7%	0.147	29.4%	89.4%
CondPaiNN + FiLM + Att	NLL	0.092	79.6%	63.6%	0.212	107.6%	53.6%	0.519	111.1%	54.2%
CondPaiNN + FiLM + Att (scalar)	sMAPE	0.015	11.3%	96.6%	0.071	36.8%	85.6%	0.154	27.3%	89.7%
CondPaiNN + FiLM + Att (scalar)	AsinhL1	0.019	12.4%	95.3%	0.059	32.7%	86.9%	0.152	27.4%	88.8%
CondPaiNN + FiLM + Att (scalar)	L1	**0.012**	11.6%	95.8%	0.049	29.5%	**88.5%**	**0.130**	24.1%	91.4%
CondPaiNN + FiLM + Att (scalar)	NLL	0.092	77.7%	65.7%	0.184	108.5%	54.1%	0.448	96.9%	61.8%
CondPaiNN + FiLM + Att + Denoise	sMAPE	0.029	35.8%	86.9%	0.088	65.6%	74.0%	0.241	49.3%	80.8%
CondPaiNN + FiLM + Att + Denoise	AsinhL1	0.048	43.8%	79.8%	0.122	67.2%	70.6%	0.242	49.1%	79.3%
CondPaiNN + FiLM + Att + Denoise	L1	0.016	29.6%	89.9%	0.056	50.3%	79.7%	0.144	34.7%	87.1%
CondPaiNN + FiLM + Att + Denoise	NLL	0.082	62.3%	72.1%	0.206	109.4%	54.6%	0.522	113.9%	52.1%
CondEGNN family										
CondEGNN + FiLM	NLL	0.015	10.1%	96.0%	0.048	31.5%	87.5%	0.230	39.0%	84.4%
CondEGNN + Att (no FiLM)	NLL	0.014	13.2%	95.0%	0.054	37.0%	85.0%	0.226	39.3%	84.8%
CondEGNN + Att + FiLM	sMAPE	0.014	**8.1%**	96.6%	0.058	34.0%	86.7%	0.153	24.4%	90.8%
CondEGNN + Att + FiLM	AsinhL1	0.022	10.6%	95.5%	0.056	30.3%	86.5%	0.144	25.3%	90.8%
CondEGNN + Att + FiLM	L1	0.014	23.2%	92.9%	0.052	42.4%	82.7%	0.136	29.3%	89.8%
CondEGNN + Att + FiLM	NLL	0.106	69.1%	65.9%	0.227	109.9%	53.9%	0.527	111.2%	55.1%

internal geometric distribution of the atoms within the crystal.

To determine if comparing the full distribution of atomic embeddings (point cloud vs. point cloud) provides a stronger uncertainty signal, we evaluated two advanced node-level distribution matching techniques: Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Optimal Transport (OT) (Peyré & Cuturi, 2019).

D.1. Computational Pre-Filtering

Computing the exact node-to-node distance between every single crystal in a training set of $N \approx 40,000$ structures requires roughly N^2 (over 1.6 billion) pairwise point cloud comparisons, which is computationally intractable for exact Optimal Transport. To solve this, we employed a two-stage evaluation pipeline. First, we used the computationally cheap graph-level KNN to find the $K = 50$ nearest neighbors for a given test structure. Second, we executed the heavy MMD and OT computations strictly on this reduced subset of 50 candidate neighbors. This reduced the problem space from hundreds of millions of pairs to roughly 277,000, making parallel CPU evaluation feasible.

D.2. Maximum Mean Discrepancy (MMD)

MMD measures the distance between two probability distributions by embedding them into a Reproducing Kernel Hilbert Space (RKHS). Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the unpooled atomic node embeddings of the test crystal, and $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$ be the node embeddings of a training crystal. The squared MMD distance is defined as:

Table 7. Single-target (ST) full mode architecture and loss ablation on the test set. Each row reports the validation-sMAPE-winning hyperparameter configuration for a given (backbone, conditioning, loss) combination. RMSLAE is reported alongside sMAPE and AsinhL1 to show that the sign-blind logarithmic metric fails in the ST regime as it does in MT (Table 6). Bold marks per-column best across the converged configurations only (RMSLAE rows excluded from the bolding pool).

		Energy (ΔE)			**Band gap (ΔE_g)**			**Volume (ΔV)**		
Architecture	**Loss**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**
CondEGNN family										
CondEGNN base	sMAPE	**0.015**	8.8%	**96.9%**	0.059	32.5%	86.0%	0.167	28.2%	89.9%
CondEGNN base	AsinhL1	0.019	9.8%	96.8%	0.070	36.9%	85.4%	0.153	28.3%	89.9%
CondEGNN base	RMSLAE	0.020	29.0%	96.0%	0.155	85.6%	83.6%	0.386	57.7%	91.3%
CondEGNN + FiLM	sMAPE	0.017	**8.4%**	96.7%	**0.054**	**29.4%**	**88.1%**	0.150	28.4%	89.8%
CondEGNN + FiLM	AsinhL1	0.021	12.1%	95.6%	0.062	33.8%	86.0%	0.151	26.0%	89.9%
CondEGNN + FiLM	RMSLAE	0.024	25.3%	95.1%	0.142	77.7%	88.3%	0.377	55.4%	92.5%
CondEGNN + Att	sMAPE	0.018	12.4%	95.9%	0.058	34.0%	86.5%	0.183	27.6%	90.6%
CondEGNN + Att	AsinhL1	0.029	16.9%	94.2%	0.054	30.7%	87.2%	0.149	27.7%	89.2%
CondEGNN + Att	RMSLAE	0.021	30.8%	95.3%	0.170	89.1%	81.9%	0.399	56.0%	91.4%
CondEGNN + Att + FiLM	sMAPE	0.018	15.6%	95.9%	0.058	31.7%	88.0%	0.146	26.8%	90.3%
CondEGNN + Att + FiLM	AsinhL1	0.022	16.7%	94.8%	0.058	34.0%	86.4%	**0.121**	**24.6%**	90.5%
CondEGNN + Att + FiLM	RMSLAE	0.027	32.2%	95.1%	0.741	111.5%	88.4%	0.375	55.2%	93.0%
CondPaiNN family										
CondPaiNN + FiLM	sMAPE	0.021	12.4%	95.8%	0.058	32.7%	86.6%	0.165	27.9%	90.3%
CondPaiNN + FiLM	AsinhL1	0.027	18.3%	94.5%	0.057	35.2%	85.9%	0.149	28.1%	89.8%
CondPaiNN + FiLM	RMSLAE	0.039	36.1%	92.5%	0.167	80.5%	88.1%	0.407	62.6%	89.1%
CondPaiNN + Att + FiLM	sMAPE	0.022	15.1%	95.5%	0.059	30.9%	87.7%	0.144	24.8%	**91.3%**
CondPaiNN + Att + FiLM	AsinhL1	0.025	13.9%	95.7%	0.072	38.3%	84.2%	0.139	26.1%	90.6%
CondPaiNN + Att + FiLM	RMSLAE	0.824	91.0%	90.4%	0.134	75.9%	88.9%	0.418	60.6%	90.8%

$$\mathrm{MMD}^2(X,Y) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(\mathbf{y}_i,\mathbf{y}_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(\mathbf{x}_i,\mathbf{y}_j)$$

For the kernel function $k(\cdot,\cdot)$, we utilized the standard Radial Basis Function (RBF) kernel:

$$k(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2}\right)$$

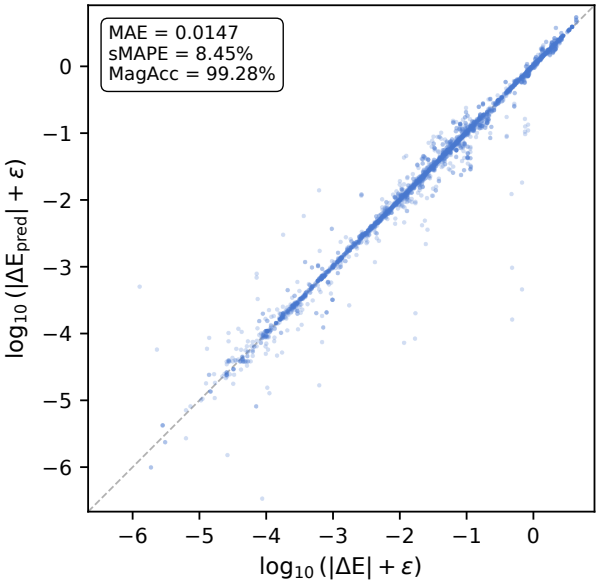
To ensure the kernel bandwidth σ dynamically scales with the density of the latent space at different layers of the network, we employed the median heuristic. For each pairwise crystal comparison, σ was set to the median pairwise Euclidean distance between all nodes in X and Y. Since MMD relies on a simple kernel matrix calculation, we implemented it as a fully vectorized, batched operation on the GPU.

D.3. Wasserstein-2 Optimal Transport (OT)

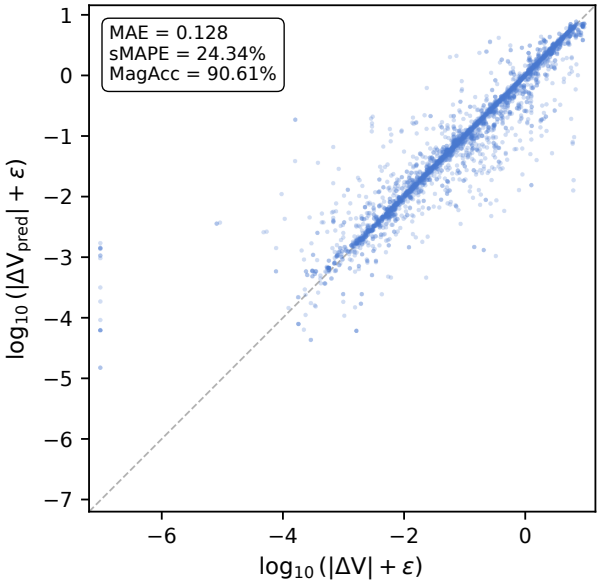
While MMD compares smoothed kernel densities, Optimal Transport directly computes the minimum "work" required to transform one atomic point cloud into the other, making it highly sensitive to discrete structural shifts. We computed the exact Earth-Mover's Distance, formally known as the Wasserstein-2 (W_2) distance (Peyré & Cuturi, 2019).

Using the same point clouds X and Y, we can define discrete uniform probability measures over the atoms, where each atom is assigned a mass of $1/n$ and $1/m$, respectively. The W_2 distance finds the optimal transport plan $\pi \in \Pi(X,Y)$ that minimizes the total geometric cost of moving the "mass" of the test crystal's atoms to match the training crystal's atoms:

$$W_2(X,Y) = \left(\min_{\pi\in\Pi(X,Y)} \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij}||\mathbf{x}_i-\mathbf{y}_j||^2\right)^{\frac{1}{2}}$$



(a) Energy (ΔE) ST-GNN Ensemble



(b) Volume (ΔV) MT-GNN

Figure 5. Logarithmic parity plots comparing the predicted vs. true Δ-corrections on the test set for the CondEGNN+FiLM model. The data is plotted as $\log_{10}(|x| + \epsilon)$ to accommodate negative values on the logarithmic scale.

where $\Pi(X, Y)$ is the set of all joint probability distributions with marginals $1/n$ and $1/m$. This exact linear program is highly sequential. As a result, we distributed the 277,000 OT pair computations across 72 CPU cores.

As discussed in the main text results, while Optimal Transport successfully captures granular geometric differences and outperforms MMD, its computational cost ($\sim$25 CPU-hours) only yielded a marginal UQ performance gain over simple graph-level Mahalanobis/KNN distances, and only specifically for predicting volume (ΔV) errors.

E. Feature-wise Linear Modulation (FiLM) Analysis

We demonstrated that incorporating the low-fidelity DFT settings into the network via FiLM conditioning drastically improved performance across all architectures. FiLM operates by projecting the computational settings into a scaling vector ($\boldsymbol{\gamma}$) and a shifting vector ($\boldsymbol{\beta}$), which are then applied to the intermediate node features at every layer.

To visualize exactly how the network uses this information, we extracted the raw $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ parameters at each layer for different input structures. Figure 6 shows the distribution of these modulation parameters. The network learns to aggressively scale and shift the latent representations based on the basis-set size, effectively warping its own internal geometry to account for the physical approximations made during the baseline DFT calculation.

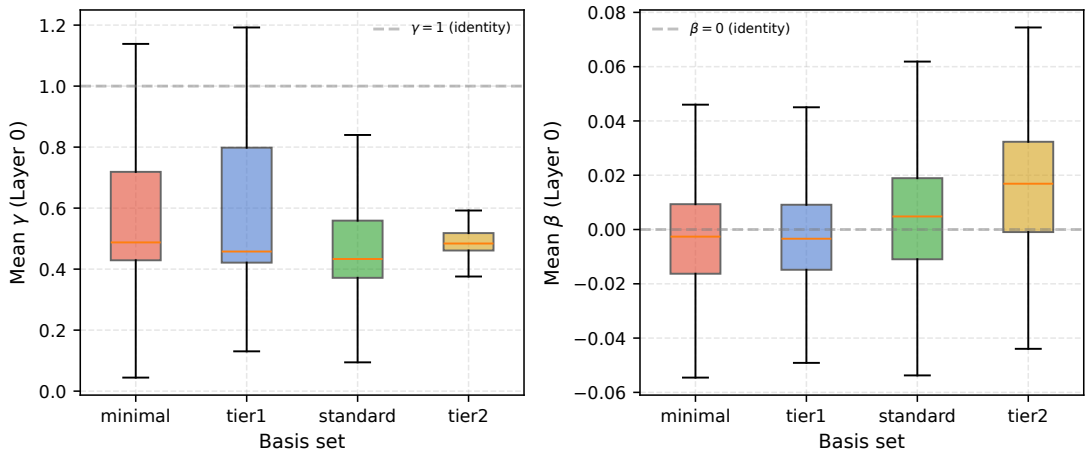
F. Comprehensive Uncertainty Quantification Metrics

In the main text, we established that extracting embeddings from different depths of the network dramatically impacts the quality of distance-based Uncertainty Quantification (UQ). Here, we provide the exhaustive layer-by-layer ablation, measuring the Spearman rank correlation (ρ) between the latent distances (k-Nearest Neighbors and Mahalanobis) and the actual absolute error.

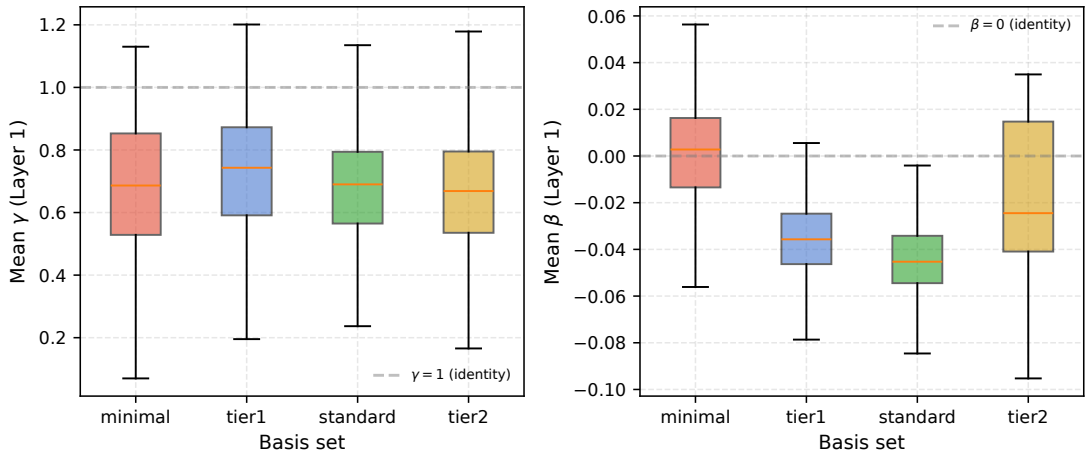
F.1. Embedding Extraction Depth

We extracted the high-dimensional graph embeddings at four distinct points in the network:

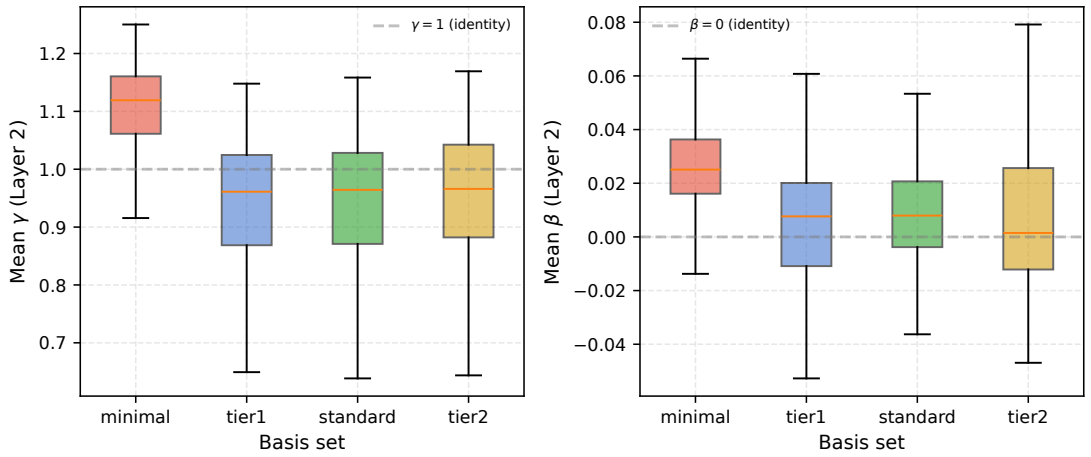
1. **`h_z`**: The pure atomic number embedding (no structural or computational settings information).



(a) Layer 0



(b) Layer 1



(c) Layer 2

Figure 6. Distribution of the FiLM scaling (γ) and shifting (β) parameters across the first three message passing layers. The identity transformations ($\gamma = 1, \beta = 0$) are marked with dashed lines. The network learns to apply highly divergent, layer-specific modulations depending on the provided baseline DFT settings, with the *minimal* basis set experiencing the most aggressive feature warping.

Table 8. Single-target (ST) recommender mode architecture and loss ablation on the test set. Each row reports the validation-sMAPE-winning hyperparameter configuration. The recommender mode receives only the unrelaxed crystal structure and computational settings as inputs, with no low-fidelity DFT features. Bold marks per-column best.

		Energy (ΔE)			Band gap (ΔE_g)			Volume (ΔV)		
Architecture	**Loss**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**	**MAE**	**sMAPE**	**MagAcc**
CondEGNN + FiLM	sMAPE	**0.016**	**8.1%**	**96.7%**	**0.063**	**34.6%**	**86.1%**	0.179	32.3%	88.2%
CondEGNN + FiLM	AsinhL1	0.022	11.7%	95.1%	0.064	38.1%	84.5%	0.166	30.7%	88.2%
CondEGNN + Att	sMAPE	0.025	15.2%	94.5%	0.082	41.0%	84.0%	0.219	34.2%	88.0%
CondEGNN + Att	AsinhL1	0.033	17.1%	93.7%	0.080	42.7%	83.4%	0.177	32.0%	88.0%
CondEGNN + Att + FiLM	sMAPE	0.026	14.2%	94.8%	0.076	41.2%	84.8%	0.179	32.8%	87.6%
CondEGNN + Att + FiLM	AsinhL1	0.028	15.8%	93.9%	0.065	35.4%	85.5%	0.178	33.4%	87.2%
CondPaiNN + FiLM	sMAPE	0.025	13.9%	94.3%	0.071	39.1%	83.6%	0.162	**28.7%**	89.3%
CondPaiNN + FiLM	AsinhL1	0.021	12.8%	94.4%	0.069	35.2%	84.4%	0.160	31.5%	89.3%
CondPaiNN + Att + FiLM	sMAPE	0.027	17.0%	94.1%	0.074	38.6%	85.6%	0.166	29.4%	90.1%
CondPaiNN + Att + FiLM	AsinhL1	0.026	14.9%	94.7%	0.071	39.4%	84.1%	**0.155**	28.9%	**89.6%**

2. **`pre_mp`**: After concatenating the atomic features with the low-fidelity DFT settings, but before message passing.
3. **`post_mp`**: After the full E(3)-equivariant message passing and FiLM modulation.
4. **`readout`**: Just before the final regression head, where the globally pooled graph vector is concatenated again with the DFT settings.

Tables 9 and 10 show the results for the energy target. The data clearly demonstrates that the raw atomic embedding contains very little UQ signal. However, concatenating the low-fidelity DFT settings (`pre_mp`) and performing message passing (`post_mp`) monotonically improves the network's ability to flag out-of-distribution structures.

Table 9. Spearman correlation (ρ) for distance-based UQ evaluated on the energy target in **Recommender Mode** (the network must predict the error without access to the cheap DFT results). The concatenated settings at the `readout` layer paired with KNN provides the best UQ signal.

Embedding Level	Dim	Distance Metric	Spearman ρ [95% CI]	AUSC-MAE
`h_z`	256	KNN	0.383 [0.358, 0.404]	0.00721
`h_z`	256	Mahalanobis	0.208 [0.181, 0.234]	0.01345
`pre_mp`	512	KNN	0.478 [0.456, 0.499]	0.00558
`pre_mp`	512	Mahalanobis	0.211 [0.184, 0.237]	0.01485
`post_mp`	512	KNN	0.514 [0.493, 0.534]	0.00455
`post_mp`	512	Mahalanobis	0.144 [0.119, 0.169]	0.01024
`readout`	**768**	**KNN**	**0.575 [0.555, 0.594]**	**0.00404**
`readout`	768	Mahalanobis	0.386 [0.362, 0.410]	0.00652

F.2. Node-Level vs. Graph-Level UQ

Graph-level UQ metrics pool the entire crystal structure into a single vector before computing the distance. We investigated whether retaining the individual atomic node embeddings as a point cloud and comparing them using Maximum Mean Discrepancy (MMD) or Wasserstein-2 Optimal Transport (OT) would yield a stronger uncertainty signal.

Computing the exact Earth-Mover's distance between thousands of atomic point clouds is computationally expensive. Therefore, we employed a two-stage approach: a rapid graph-level KNN pre-filter (K=50) followed by the exact node-level OT calculation on those neighbors.

As shown in Tables 11 and 12, Optimal Transport consistently beats MMD. However, for predicting scalar properties like energy and band gap, the simple graph-level distance (KNN or Mahalanobis) is still superior. The pooled graph representation, which includes the concatenated DFT settings, is simply more informative. The single exception is for volume (ΔV) in the full mode: here, the internal atomic distribution matters more than the global graph summary, allowing Optimal Transport to edge out the graph-level methods.

Table 10. Spearman correlation (ρ) for distance-based UQ evaluated on the energy target in full mode (the network is conditioned on the cheap DFT energy). Access to the continuous physics priors shapes the latent space such that the parametric Mahalanobis distance outperforms KNN.

Embedding Level	Dim	Distance Metric	Spearman ρ [95% CI]	AUSC-MAE
h_z	128	KNN	0.282 [0.256, 0.307]	0.00634
h_z	128	Mahalanobis	0.210 [0.185, 0.235]	0.01132
pre_mp	256	KNN	0.404 [0.382, 0.426]	0.00453
pre_mp	256	Mahalanobis	0.259 [0.233, 0.283]	0.00967
post_mp	256	KNN	0.533 [0.513, 0.553]	0.00301
post_mp	256	Mahalanobis	0.539 [0.519, 0.558]	0.00596
readout	396	KNN	0.535 [0.515, 0.556]	0.00311
readout	**396**	**Mahalanobis**	**0.619 [0.602, 0.634]**	**0.00411**

Table 11. Comparison of Node-Level (MMD, Optimal Transport) vs. Graph-Level (KNN) UQ metrics in **Recommender Mode**. Node-level methods operate on the unpooled post_mp representations. Graph-level KNN on the readout layer remains superior for energy and band gap.

Target	MMD ρ	Optimal Transport ρ	Best Graph-Level ρ (Reference)
Energy (ΔE)	0.345 [0.322, 0.368]	**0.414** [0.391, 0.435]	**0.575** (KNN readout)
Band gap (ΔE_g)	**0.502** [0.482, 0.522]	0.492 [0.472, 0.512]	**0.553** (KNN readout)
Volume (ΔV)	0.260 [0.232, 0.286]	**0.267** [0.240, 0.293]	0.263 (KNN readout)

F.3. Extended Sparsification Curves (sMAPE)

In the main text, we demonstrated the effectiveness of our Uncertainty Quantification (UQ) using sparsification curves based on the Mean Absolute Error (MAE). As we sequentially remove the predictions our model flags as most uncertain, the overall MAE of the remaining dataset drops, proving the UQ correlates with actual error.

However, because our targets ($\Delta E, \Delta V, \Delta E_g$) can span multiple orders of magnitude, optimizing solely for absolute error can mask relative performance. To ensure our UQ metrics also effectively filter out large relative percentage errors, we tracked the sparsification curves using the symmetric Mean Absolute Percentage Error (sMAPE).

As shown in Figure 7, the UQ metrics successfully sparsify relative errors as well. Notably, while Deep Ensembles perfectly track the oracle in both metrics, Mahalanobis and KNN distances behave differently depending on whether they are evaluated via MAE or sMAPE. This is because Mahalanobis excels at identifying geometric outliers (which heavily improves MAE), but sMAPE normalizes by the prediction magnitude, subtly altering the impact of outlier identification.

F.4. Deep Ensemble Size Ablation

In the main text, we established that Deep Ensembles provide the highest quality Uncertainty Quantification (UQ), achieving a near-perfect Spearman ρ rank correlation. However, maintaining and running an ensemble of graph neural networks is computationally expensive. To determine the optimal trade-off between compute cost and UQ quality, we performed an ablation study over the ensemble size ($N \in \{3, 5, 7, 10\}$).

As shown in Figure 8, the UQ quality jumps significantly when moving from 3 to 5 models, but the returns diminish rapidly afterward. For both the full mode (with low-fidelity DFT results) and the recommender mode (without), an ensemble size of $N = 5$ appears to be the "sweet spot" for practical applications, capturing the vast majority of the epistemic uncertainty without requiring the compute overhead of a 10-model ensemble.

G. Zero-Cost Recommender System Evaluation

In the main text, we introduced the concept of a "zero-cost" recommender system—a model that predicts the computational error of a specific DFT configuration before the calculation is even run. To prove that our E(3)-equivariant networks genuinely outperform tabular models on this task, we must ensure a fair comparison.

Table 12. Comparison of Node-Level (MMD, Optimal Transport) vs. Graph-Level (Mahalanobis) UQ metrics in full mode Optimal Transport provides the best UQ signal for predicting the uncertainty of structural volume changes.

Target	MMD ρ	Optimal Transport ρ	Best Graph-Level ρ (Reference)
Energy (ΔE)	0.509 [0.489, 0.529]	**0.527** [0.507, 0.547]	**0.619** (Mahalanobis `readout`)
Band gap (ΔE_g)	**0.507** [0.485, 0.526]	0.493 [0.470, 0.513]	**0.606** (Mahalanobis `post_mp`)
Volume (ΔV)	0.444 [0.422, 0.465]	**0.456** [0.434, 0.476]	0.445 (Mahalanobis `post_mp`)

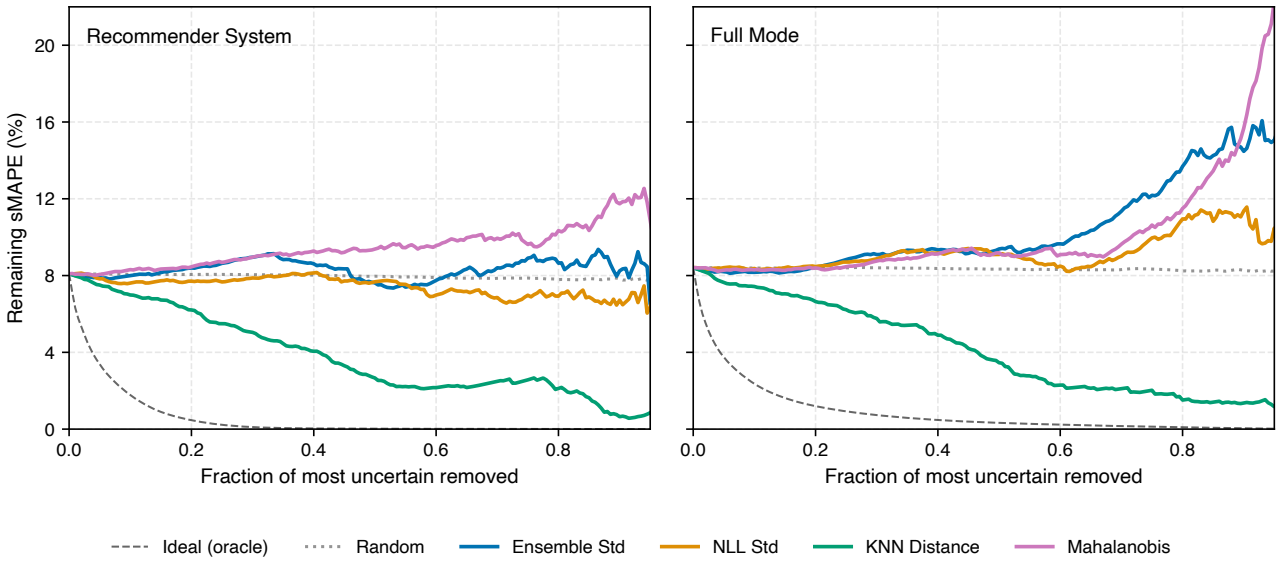


Figure 7. Sparsification curves for test ΔE sMAPE for the Δ-EGNN model. The left panel shows the model operating in recommender mode (no cheap DFT inputs), while the right panel shows full mode Random sample removal and the ideal (oracle) ranking are provided as baseline comparisons. The curves confirm that the UQ metrics effectively reject high relative-percentage errors alongside absolute errors.

A standard Random Forest (RF) trained on all available data has an unfair advantage: it relies heavily on tabular features derived directly from the converged baseline DFT calculations (e.g., the sum of eigenvalues, free energy components, and relaxed initial volumes). A true recommender system cannot use these features because they do not exist until the calculation finishes.

G.1. The Cost of Dropping High-Fidelity Features (RF Ablation)

To create a fair baseline, we stripped the Random Forest of all DFT-derived variables. We defined three feature sets:

- **RF-Full (102 features):** The upper bound. Includes all DFT-derived energy components and geometries.
- **RF-PreCalc (89 features):** The fair baseline. Excludes all DFT outputs. Relies only on initial ICSD crystal geometry, elemental properties, and pre-computed isolated-atom (monomer) energies.
- **RF-Minimal (57 features):** Excludes monomer features to test if isolated-atom data is strictly necessary.

As shown in Table 13, stripping away the cheap DFT features severely penalizes the Random Forest, costing it between 2 to 5 sMAPE points across all targets. This proves exactly why the recommender task is so difficult: without explicit quantum chemical descriptors, tabular models struggle to map basic elemental properties to complex energetic corrections.

G.2. Feature Importance: What are Tabular Models Learning?

To understand exactly *what* the Random Forest learns when the DFT outputs are removed, we performed a permutation feature importance analysis on the RF trained on the energy target. When granted access to the full dataset (RF-Full), the

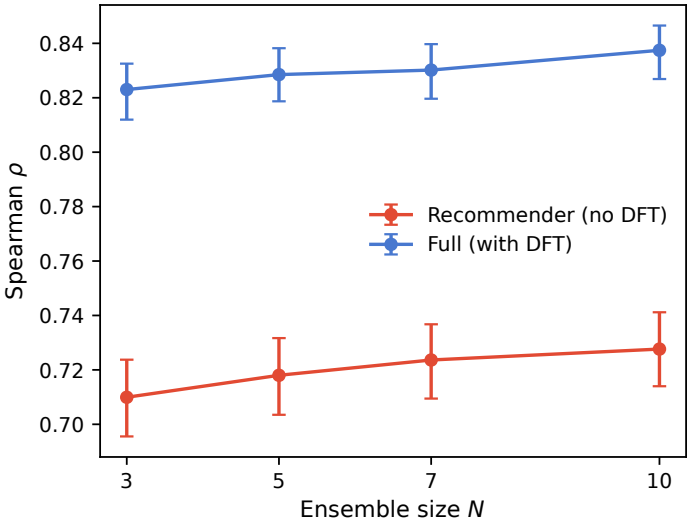


Figure 8. Spearman ρ rank correlation versus the ensemble size. The blue line represents the model operating in full mode (with low-fidelity DFT inputs), while the red line represents the recommender mode.

Table 13. Random Forest feature ablation on the Test Split. Removing the 13 DFT-derived outputs to create a fair recommender baseline (RF-PreCalc) heavily penalizes the model's accuracy compared to the upper-bound (RF-Full).

	Energy (ΔE)		**Band gap (ΔE_g)**		**Volume (ΔV)**	
Model (Feature Count)	**sMAPE**	**MagAcc**	**sMAPE**	**MagAcc**	**sMAPE**	**MagAcc**
RF-Full (102 features)	9.70%	95.8%	30.46%	88.4%	28.15%	88.8%
RF-PreCalc (89 features)	11.62%	95.2%	35.07%	85.4%	31.49%	87.4%
RF-Minimal (57 features)	12.45%	94.7%	34.57%	85.0%	32.18%	87.0%

model leverages post-DFT physical descriptors like the final structural volume to estimate the energy correction. However, when restricted strictly to pre-calculation features (RF-PreCalc), the Random Forest is forced to rely almost entirely on heuristic isolated-atom (monomer) energy statistics and an integer encoded variable for the basis set size and numerical settings level (*binary.precision*). These ranked variables for RF-PreCalc are shown in Table 14.

H. Characterization of Out-of-Distribution (OOD) Predictions

To understand the physical and computational drivers of model uncertainty, we performed a comprehensive characterization of the crystals rejected by our 95th-percentile k-Nearest Neighbors (KNN) distance threshold. By comparing the population of the accepted (in-distribution) versus rejected (out-of-distribution) test structures, several distinct physical patterns emerge.

Computational Settings Drive Uncertainty: The most dominant predictor of a rejected structure is the basis set size. While the *minimal* basis set accounts for roughly 28% of the training data, it comprises 45–70% of the rejected crystals in the Full-mode models. In recommender mode (where the network must guess the error without seeing the cheap DFT results), this effect becomes extreme: over 90% of rejected predictions for the band gap and energy recommender models use a minimal basis. The network correctly identifies that minimal-basis DFT produces qualitatively poor wavefunctions, pushing the resulting latent embeddings far from the reliable training manifold. Similarly, coarser k-point integration grids (e.g., 2 k-points per Å) are consistently over-represented in the rejected set by a factor of 1.5 to 2.

Compositional Epistemic Uncertainty: The network successfully flags epistemic uncertainty driven by training set imbalances. Elements with less than 1.5% presence in the training set act as primary triggers for OOD rejection. For example, rare transition metals like Osmium and Cobalt are enriched in the rejected set by factors of 50× and 65×, respectively. Ruthenium, which was entirely absent from the training set but appeared in the test holdout, was perfectly

Table 14. Top 12 tabular features used by the RF-PreCalc baseline for predicting energy (ΔE), ranked by permutation importance (ΔsMAPE). Without access to full DFT calculations, the tree models rely heavily on proxy isolated-atom (monomer) calculations.

Rank	Feature Name	Importance (ΔsMAPE)	Std. Dev.
1	delta_mean_monomer_total_energy_per_atom	59.15	$\pm$0.44
2	binary_precision	43.55	$\pm$0.52
3	delta_max_monomer_total_energy_per_atom	42.65	$\pm$0.54
4	delta_min_monomer_total_energy_per_atom	39.48	$\pm$0.63
5	delta_mad_monomer_total_energy_per_atom	21.91	$\pm$0.32
6	max_basis_functions	18.90	$\pm$0.51
7	mean_basis_functions	18.53	$\pm$0.41
8	mean_monomer_total_energy_per_atom	10.32	$\pm$0.19
9	mean_atomic_IP_delta	9.59	$\pm$0.11
10	delta_mean_monomer_gamma_gap	7.98	$\pm$0.16
11	max_atomic_EA_delta	7.72	$\pm$0.23
12	max_monomer_gamma_gap	7.70	$\pm$0.17

isolated into the rejected bucket.

Crystal System and Geometric Complexity: The OOD threshold also isolates specific crystallographic symmetries. As shown in Table 15, low-symmetry systems (triclinic and monoclinic) are consistently over-represented in the rejected bucket across all prediction targets. Conversely, high-symmetry trigonal structures are the most reliably in-distribution. Furthermore, for predicting ΔE, the network exhibits a strong sensitivity to unit cell size. The rejected bucket for the full ST-Energy model contained an average of 23.8 atoms per cell, compared to just 8.9 atoms for the accepted structures, indicating that larger unit cells naturally fall outside the learned interpolation domain.

Table 15. Enrichment ratio (Rejected % / Train %) of crystal systems in the OOD bucket. Values > 1.0 indicate the system is over-represented in the rejected set. Low-symmetry systems (triclinic, monoclinic) are consistently flagged as out-of-distribution.

		Full Mode (ST)			Recommender Mode (ST)		
Crystal System	**Train %**	**Energy (ΔE)**	**Band gap (ΔE_g)**	**Volume (ΔV)**	**Energy (ΔE)**	**Band gap (ΔE_g)**	**Volume (ΔV)**
Triclinic	9.9%	1.76	2.44	3.31	1.59	1.87	2.77
Monoclinic	6.7%	1.70	1.99	1.81	1.32	1.68	2.55
Orthorhombic	7.0%	1.32	1.21	1.40	1.17	0.94	1.12
Tetragonal	16.4%	0.77	0.50	0.46	0.94	0.68	0.78
Trigonal	31.1%	0.21	0.92	0.46	0.34	0.60	0.31
Hexagonal	21.6%	1.21	0.49	0.54	1.23	0.82	0.64
Cubic	7.3%	2.31	0.91	1.61	1.99	2.21	1.57

Evaluating the Reject Decision: Ultimately, the purpose of an explicit OOD threshold is to filter out the most inaccurate predictions. Across all evaluated models, the rejected structures exhibited a ground-truth absolute error ($|\Delta|$) that was 2 to 3 times larger than the accepted structures. For example, in the full ST-Volume model, the mean $|\Delta V|$ of the accepted structures was 0.66 Å^3/atom, while the rejected structures averaged 1.78 Å^3/atom. This confirms the UQ mechanism successfully acts as a guardrail against large structural shifts and computational failures.