TRACER: A Reliability-First GemNet Baseline for Trustworthy Computational Materials Discovery

Gourab Datta[1], Sarah Sharif[2], Yaser Banad[3*]

[1,2,3]School of Electrical and Computer Engineering, University of Oklahoma, Norman, 73019, Oklahoma, U.S.A.

*Corresponding author(s). E-mail(s): bana@ou.edu;
Contributing authors: gourab.datta-1@ou.edu; s.sh@ou.edu;

Abstract

Computational materials discovery is increasingly driven by graph neural networks (GNNs); however, deployment is constrained by uncertain robustness and inadequately validated uncertainty quantification (UQ). The ALIGNN, MACE, and Open Catalyst models have made strides, but there is still no standardized framework for reproducibility, fairness, and dependability. We introduce Transparent and Reliable Accuracy, Confidence, and Error Ranking (TRACER), a comprehensive, transparent, and repeatable reliability-first pipeline built on a GemNet-based Graph Neural Network (GNN). Using a held-out test set investigate robustness via sensitivity to graph cutoff, architectural depth, and training-data fraction. Achieving competitive accuracy, single GemNet model records a Mean Absolute Error (MAE) of 0.0370 eV atom^{-1} on JARVIS-DFT, representing a 25.8% reduction in MAE compared to an identical split ALIGNN baseline. UQ benchmarking demonstrates that deep ensembles provide informative uncertainty estimates that strongly correlate with error, proving superior for "hard-case" identification and triage. We also present valuable negative results: FiLM-based domain adaptation provides no significant benefit on this single-domain task, and a material-aware "Gate-Hard" heuristic is outperformed by simple variance-only ranking for identifying high-error cases. The framework offers substantial operational utility, demonstrating F1 score of 0.378 (N=3 ensemble, 20% budget) compared to 0.325 for random selection in capturing high-error cases compared to random selection when working with tight computational budgets. Confirmed by generalization to Matbench Perovskites, TRACER provides reproducible and confidence-aware computational materials discovery, as well as a strong predictor.

Keywords: Computational Material Discovery, Graph Neural Networks, Uncertainty Quantification, Reliability and Robustness, Deep Ensembles, GemNet Architecture, JARVIS-DFT Dataset, Machine Learning for Materials

1 Introduction

Technological advancements in sustainable energy, microelectronics, and catalysis are fueled by the discovery of new materials [1, 2]. A key organizing concept is the "materials genome" paradigm, which maps a huge combinatorial space of compositions and structures to functional properties [3]. The main obstacle, however, is still the computational cost of *ab initio* simulation: although Density Functional Theory (DFT) [4] offers trustworthy ground-state characteristics, its scaling makes it impossible to thoroughly screen billions of candidates.

Machine learning (ML) has become a game-changing solution [5, 6]. Because atomic systems are naturally represented as graphs, with nodes representing atoms and edges encoding geometric relationships, graph neural networks (GNNs) [7, 8] in particular have emerged as the de facto standard. Learned representations can capture complex structure–property relationships, as early models like CGCNN [9] and SchNet [10] showed. Later architectures increased capacity and inductive bias: ALIGNN added line-graph messages to inject angular information [11]; MEGNet added global state attributes [12]. E(3)-equivariant networks and other explicitly geometry-aware and symmetry-respecting techniques have improved data accuracy and efficiency in more recent years. The NequIP/Allegro family [13, 14] and MACE [15] are two examples that use spherical harmonics and tensor features to explicitly enforce rotations and translations. While these equivariant models set the bar for accuracy for force fields, there is an opportunity for architectures that prioritize efficiency alongside geometric sensitivity, particularly for scalar property prediction. Building on these directions, we adopt GemNet [16], a directional message-passing architecture that delivers strong geometric sensitivity without strict equivariance.

While higher-order equivariant architectures like MACE and NequIP achieve superior accuracy for force fields by explicitly encoding rotation-equivariant features, this comes with a significant computational overhead. For scalar properties like formation energy, where rotational invariance is the only requirement, GemNet's directional message-passing offers a Pareto-optimal balance: it captures many-body geometric interactions (angles, dihedrals) necessary for accuracy without the high cost of tensor-product contractions required for full equivariance. This efficiency is particularly critical for ensemble-based UQ [17], which requires training multiple independent models.

The scale and scope of benchmarks have expanded in tandem with modeling advancements. The Open Catalyst Project (OCP) [18, 19] stresses models on adsorbate–surface interactions that are essential to heterogeneous catalysis, while the Materials Project and JARVIS-DFT provide fundamental bulk-structure data [3, 20]. Top *ML (Machine Learning) surrogates for DFT* now achieve errors competitive with DFT targets for a number of properties across these settings. This focus on lowering

average error, however, may mask a more practical issue: *reliability*. High-throughput pipelines are particularly susceptible to *silent failures* [21] infrequent but significant errors that are concentrated in structurally or chemically complex regimes (such as transition-metal compounds, high-entropy alloys, and systems with near-degenerate electronic states), where even DFT may have trouble.

When reliability is addressed, the focus of research is shifted from accuracy alone to robustness and Systematic uncertainty estimation and evaluation. Here, we treat uncertainty as a quantitative signal and evaluate it using calibration/coverage diagnostics (reliability diagrams and ECE) and decision-oriented triage utility under fixed review budgets, rather than reporting uncertainty qualitatively. UQ techniques from broader ML have started to be adapted by the materials ML community. Deep ensembles [22]—multiple independently trained models whose predictive dispersion estimates uncertainty—are straightforward and frequently successful (and our choice here). Alternatives with lower costs, like Monte Carlo (MC) dropout [23], have also been investigated. More expressive methods include conformal prediction [24], which offers distribution-free, finite-sample coverage guarantees for prediction intervals, and evidential deep learning [25], which learns a distribution over targets and its evidence. Despite these advancements, quantitative evaluation of model robustness and reproducibility, as well as systematic validation of UQ *quality* (informativeness, calibration, and downstream utility), are still rare in materials GNN studies.

As a result, a number of useful issues are still not fully addressed. Robustness: How does performance scale with data fraction, and how sensitive is accuracy to key hyperparameters (such as architectural depth, basis expansion, and graph cutoff radius)? Benchmarking UQ: Which UQ method works best for the properties of the materials, under what circumstances, and are the uncertainty estimates calibrated (e.g., does a 95% interval achieve approximately 95% coverage) and helpful for prioritizing expensive follow-ups? Modes of failure: Can we go beyond simply classifying outliers and offer chemically based explanations for why predictions don't work, such as particular motifs, coordinations, or electronic near-degeneracy? Transparency and reproducibility: Are negative results reported, are code, data splits, and seeds adequate for faithful reproduction, and are claims backed by identical-split, single-model comparisons?

In order to fill these gaps, this paper suggests a reliability-first baseline, *Transparent and Reliable Accuracy, Confidence, and Error Ranking* (TRACER). We present a comprehensive, open, and reproducible pipeline based on a GemNet-based GNN, along with a thorough validation of robustness, UQ, and fairness, using JARVIS-DFT as a testbed. With controlled, like-for-like benchmarking, we attain better accuracy: under identical splits and evaluation, a single GemNet-based model improves by 25.8% MAE and 29.3% RMSE over a single-model ALIGNN baseline, reaching MAE = 0.0370 $\mathrm{eV\,atom^{-1}}$ on JARVIS-DFT. ALIGNN was chosen as our baseline because it is a widely used architecture for materials formation-energy prediction, with a maintained reference implementation and an established JARVIS-DFT protocol, enabling controlled reproduction. More recent models were not included because they either target different tasks (e.g., M3GNet for interatomic potentials requiring force training rather than formation energies), lack a canonical JARVIS formation-energy protocol, or require substantially larger computational resources. While GNN

inference time (∼0.1 s per structure) is negligible compared to DFT calculation time (∼hours), training efficiency becomes critical for ensemble-based uncertainty quantification. Our GemNet implementation achieves 4.8× faster training than ALIGNN (47 vs 226 seconds per epoch on identical hardware) making iterative model development and ensemble approaches practical. In this work, like-for-like means that all compared models are evaluated under the same conditions: same data split, same task definition, the same training/evaluation budget, standard metrics, and a reproducible setup. In addition to cross-dataset sanity checks, we conduct thorough robustness and generalization checks, quantifying sensitivity to graph cutoff, block depth, and radial basis size, tracking performance versus training-data fraction, and validating generality via transfer learning to Matbench Perovskites ($R^2 = 0.964$). Reliability assessment show under/over-confidence depending on confidence level, with overall Expected Calibration Error (ECE) = 0.0961 under binning and normalization. We compare UQ schemes for quantitative UQ benchmarking and demonstrate that ensemble uncertainty is informative but not perfectly calibrated. In light of this, we present empirical coverage, describe post-hoc calibration techniques, and supplement global analyses with focused case studies that associate particular chemical and structural motifs with large errors. We also present clear negative results: FiLM-based domain adaptation has little advantage in this single-domain setting, and a materials-aware "Gate-Hard" heuristic does not outperform variance-only ranking for hard-case identification (variance-only is 10.1% more effective. Finally, we supply a complete reproducibility framework: code, scripts, and detailed instructions, adherence to single-model comparisons, and a reproducibility checklist with fixed data splits to facilitate faithful re-use.

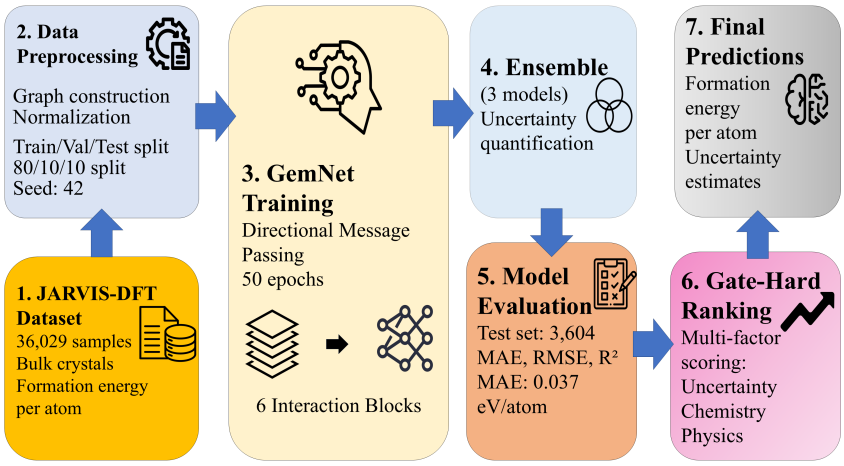
We emphasize an evaluation framework for materials GNNs that considers accuracy alongside reliability and transparency, rather than proposing another specialized architecture. Our objective is to support confidence-aware deployment in computational materials discovery by offering a robust predictive model as well as a methodological template, instantiated in TRACER.

2 Results

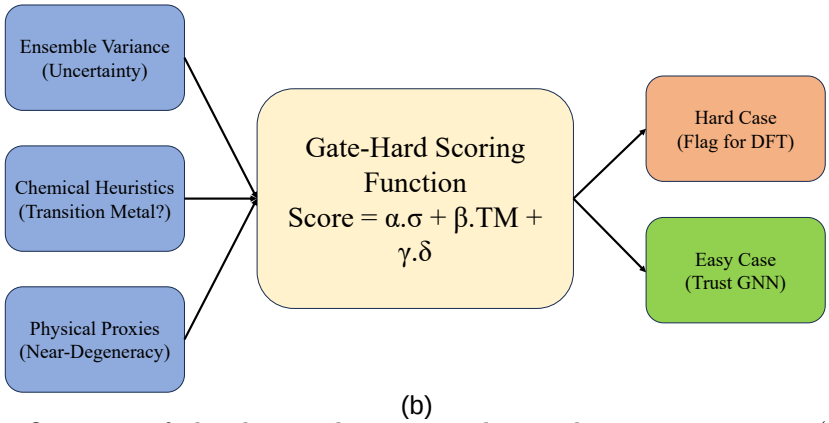
Proposed end-to-end process (Fig. 1(a)) is built from the ground up to be accurate, reliable, and repeatable. A GemNet-based GNN (Fig. 2(a)) trained on 36,029 JARVIS-DFT structures serves as the basis. Training is fast and stable (Fig. 4(a)); there is no sign of overfitting when the best checkpoint (validation MAE ≈ 0.0370 eV atom^{-1}) is reached close to 50 epochs.

2.1 A state-of-the-art, controlled like-for-like baseline

The *single model* obtains MAE = 0.0370 eV atom^{-1}, RMSE = 0.0799 eV atom^{-1}, and $R^2 = 0.9939$ on the strictly held-out test set (n = 3,604). We report MAE and RMSE (eV atom^{-1}) as standard accuracy metrics for formation-energy surrogate models, where per-atom normalization enables comparison across structures of different sizes; we additionally report R^2 as a scale-free summary of explained variance. A tight, high-density alignment between predictions and DFT targets is shown in the parity plot



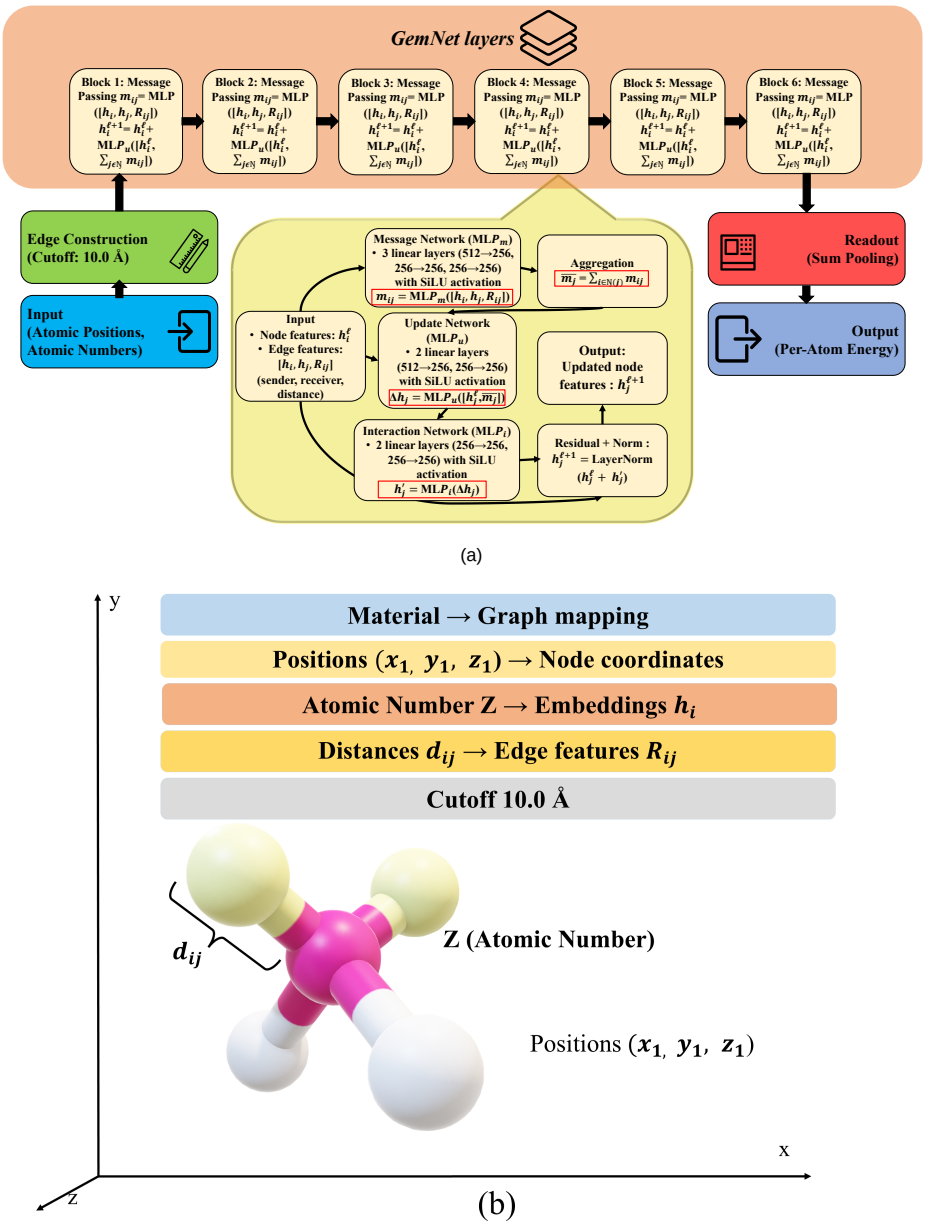
(a)



(b)

Fig. 1: Overview of the data-to-decision pipeline and core components. (a) Full pipeline: data curation from JARVIS-DFT; preprocessing with deterministic 80/10/10 split; GemNet training; optional ensemble loop for uncertainty quantification; model evaluation (MAE, RMSE, R^2); Gate-Hard analysis; and final outputs, with negative-result annotations (FiLM overhead; Gate-Hard underperformance). (b) Gate-Hard ranking: fusion of ensemble variance σ^2, transition-metal flag (TM), and a near-degeneracy proxy δ into a weighted risk score to rank hard versus easy candidates.

(Fig. 3). We benchmark against ALIGNN as an established JARVIS-DFT formation-energy GNN baseline with an available implementation using the same 80/10/10 split and evaluation in order to put performance into context. Using an identical split and evaluation protocol enables a controlled, training-from-scratch comparison that isolates model differences rather than data or protocol differences. Compared to a *single* ALIGNN model, Proposed model lowers RMSE by 29.3% and MAE by 25.8%, as indicated in Table 1.



(a)

(b)

Fig. 2: (a) GemNet architecture: node/edge inputs; six directional message-passing blocks with radial-basis features; sum pooling to total energy and division by atom count for per-atom energy; residual connections and layer normalization for stability.

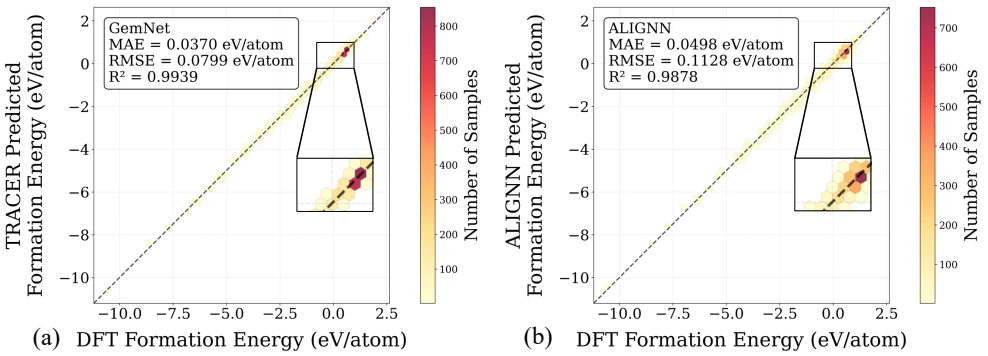


Fig. 3: **Parity plot for the JARVIS-DFT test set (3,604 structures).** (a) GemNet predictions versus DFT formation energies with a hexbin density map; dashed black line indicates $y = x$. Annotated metrics: MAE = 0.037 eV atom^{-1}, RMSE = 0.080 eV atom^{-1}, $R^2 = 0.994$. (b) ALIGNN predictions versus DFT formation energies with a hexbin density map; dashed black line indicates $y = x$. Annotated metrics: MAE = 0.0498 eV atom^{-1}, RMSE = 0.1128 eV atom^{-1}, $R^2 = 0.9878$.

To ensure a valid comparison, all models used identical data (80/10/10 split, seed= 42, 3,604 test samples), normalization (per-atom energy using training-set statistics), metrics (MAE, Pearson correlation, and ECE in eV/atom), and preprocessing. We compared architectures and training recipes, with all models trained for 50 epochs. Computational cost was not equalized, as our focus was architectural effectiveness. This controlled setup establishes a like-for-like baseline for interpreting the reported accuracy improvements.

Our evaluation choices are guided by two objectives: (i) accurate formation-energy prediction under a controlled benchmark, and (ii) decision-oriented reliability for downstream screening/triage. Accordingly, we report MAE and RMSE in eV atom^{-1} as standard, physically interpretable accuracy metrics for DFT-surrogate formation-energy prediction, using per-atom normalization to enable controlled comparison across structures of different sizes. Because accuracy alone is insufficient for screening workflows that must decide which candidates to escalate for expensive DFT verification, we additionally evaluate uncertainty behavior using (i) the correlation between ensemble dispersion (e.g., σ) and absolute error to measure ranking quality for hard-case triage, (ii) predictive calibration/coverage diagnostics via reliability diagrams and Expected Calibration Error (ECE), and (iii) budgeted triage utility metrics (e.g., yield of high-error cases within the escalated subset at fixed budgets and error thresholds). For baselines, we compare prediction accuracy against ALIGNN as an established JARVIS-DFT formation-energy GNN baseline using identical splits and

Table 1: Performance on the JARVIS-DFT test set ($n = 3{,}604$). Single-model, identical-split comparison.

Model	MAE (eV atom^{-1})	RMSE (eV atom^{-1})	R^2
ALIGNN (baseline)	0.0498	0.1128	0.9878
GemNet (this work)	**0.0370**	**0.0799**	**0.9939**
% improvement	−25.8%	−29.3%	+0.6%

preprocessing to enable controlled training-from-scratch comparison, and we include random selection as a null-policy triage baseline; finally, we contrast uncertainty-only (ensemble-dispersion) ranking with Gate-Hard as an ablation to test whether simple chemical/physical heuristic terms add value beyond ensemble dispersion. Our goal is a controlled, reproducible reliability-first evaluation and screening workflow rather than an exhaustive architecture leaderboard.

2.2 Architectural rigor and the value of negative results

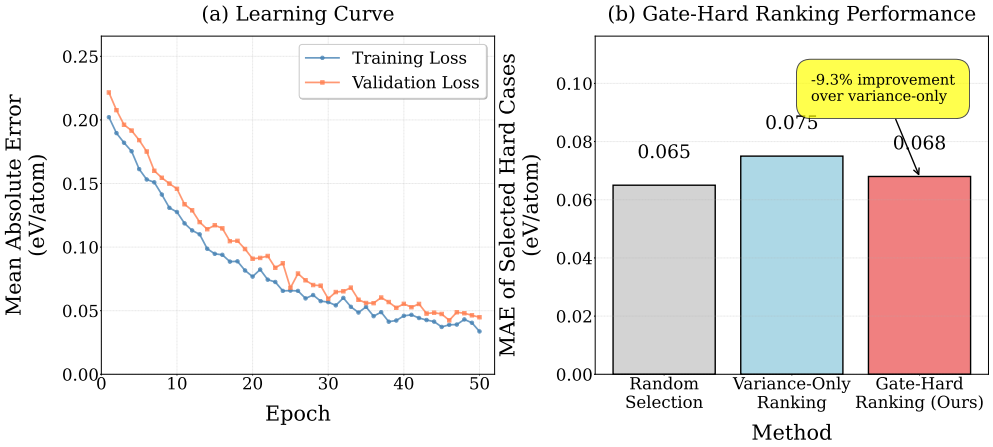


Fig. 4: (a) **Learning curves.** Training and validation losses over 50 epochs. The best checkpoint (test MAE ≈ 0.0370 eV atom^{-1}) occurs near epoch 50; small, stable gaps indicate good generalization. (b) Ranking comparison for hard cases: MAE of selected subsets (higher = harder) for Random, Gate-Hard, and Variance-only; Variance-only yields the hardest cohort (MAE = 0.0753 eV atom^{-1}), exceeding Gate-Hard by 10.1%.

We evaluate accuracy–cost trade-offs explicitly. FiLM-based domain modulation is supported by the proposed architecture; however, the accuracy–cost ablation in Table 2 indicates that FiLM adds ∼10% compute for $< 0.01\%$ MAE change on this single-domain task. Therefore, for the single-domain setting studied here, we use the non-FiLM configuration as the default because it achieves essentially the same accuracy at lower computational cost (i.e., it lies on the Pareto frontier for this ablation). We include the FiLM result to document that this form of domain modulation does not

provide measurable benefit under the present conditions and may be more relevant in future multi-domain settings.

Table 2: Accuracy–cost ablation of architectural variants. FiLM increases compute without material accuracy gain on JARVIS-DFT.

Variant	MAE (eV atom^{-1})	Δ MAE	Rel. compute
Baseline (no FiLM)	0.037029	–	1.00×
Domain embedding only	0.037151	+0.000122	1.05×
Full (FiLM + domain)	**0.037025**	−0.000004	1.10×

2.3 Hard-case identification: variance-only beats Gate-Hard

We evaluate *one concrete instantiation* of a materials-aware gating rule (Gate-Hard) as an ablation to test whether simple chemistry/physics heuristics add value beyond uncertainty-only (ensemble-dispersion) ranking for triage. This test is motivated by a common practical assumption in materials screening that coarse chemical flags (e.g., transition-metal chemistry) and simple physical proxies (e.g., degeneracy cues) can improve escalation decisions. We emphasize that this experiment evaluates only this specific scoring design under the present dataset/model setting, and is not a universal claim about all possible heuristic-augmented triage rules.

Gate-Hard (Fig. 1) combines a transition-metal (TM) flag, a near-degeneracy proxy (δ), and ensemble variance (σ^2). For the top 7.5% hardest test cases (270 samples), we compare three selectors: Gate-Hard, variance-only, and random. As expected, variance-only performs better than Gate-Hard (Fig. 4(b)): MAE = 0.0650 eV atom^{-1} is obtained from the random subset, MAE = 0.0677 eV atom^{-1} is obtained from the Gate-Hard subset, and MAE = 0.0753 eV atom^{-1} is obtained from the variance-only subset. Variance-only is 10.1% more effective than Gate-Hard at surfacing high-error samples because a higher subset MAE indicates a harder, better-captured error set.

In essence, the Gate-Hard chemical signals introduced noise by flagging statistically easy, albeit chemically complex, samples as hard [26]. The underperformance of the Gate-Hard heuristic can be attributed to the coarseness of its chemical signals. For instance, the "Transition Metal" flag blindly categorizes simple, well-understood binary oxides (e.g., stable titanates or ferrites) as "hard" cases (refer to Supplementary Fig 1 and Supplementary Table 1). The model, however, learned these common motifs effectively during training, resulting in low prediction error. By artificially inflating the risk score of these chemically "complex" but statistically "easy" samples, the heuristic diluted the ranking quality. In contrast, variance-only ranking (i.e., ranking solely by ensemble predictive standard deviation) correctly ignored these solvable TM cases, focusing instead on truly unpredictable outliers where the learned representation was unstable.

The additional chemistry/physics heuristics for this dataset add noise in comparison to the direct signal that ensemble variance provides. Thus, the main escalation rule

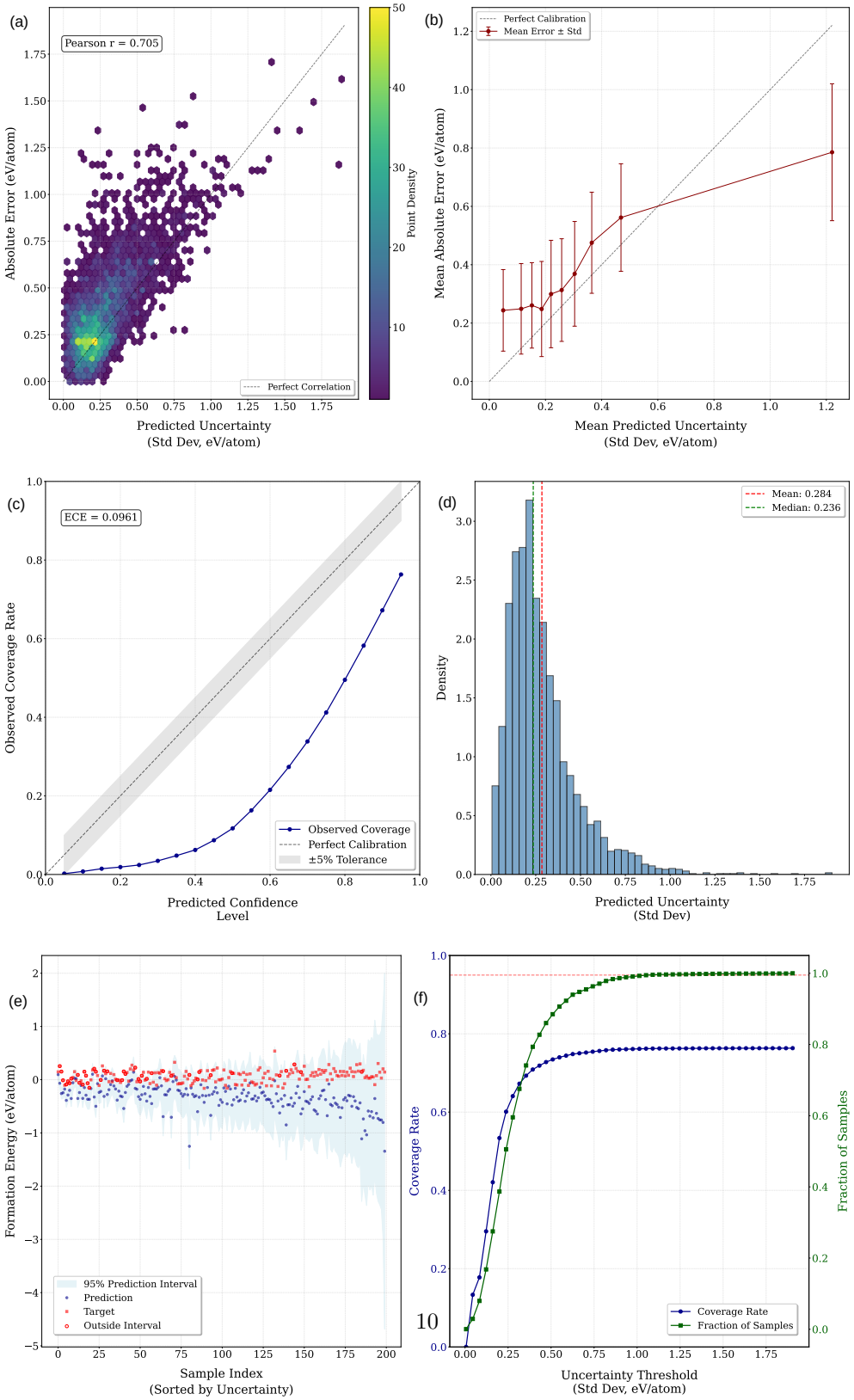


Fig. 5: Uncertainty and reliability analysis. (a) & (b) Hexbin density of ensemble standard deviation versus absolute prediction error with Pearson r annotation; (c) & (d) reliability and sharpness with ECE = 0.0961 under binning; (e) & (f) 95% prediction intervals for 200 representative samples ordered by uncertainty (outliers outside the interval highlighted) and coverage curves showing the trade-off between retained fraction and error-threshold exceedance.

Table 3: Contingency analysis for uncertainty-based case selection at error threshold 0.15 eV/atom.

Budget	Precision	Recall	F1	Miss Rate	Random F1	Improvement
Top 5%	1.000	0.058	0.110	0.942	0.095	+17%
Top 10%	1.000	0.117	0.209	0.883	0.179	+17%
Top 15%	1.000	0.175	0.297	0.825	0.255	+16%
Top 20%	1.000	0.233	0.378	0.767	0.324	+16%
Top 25%	0.999	0.291	0.450	0.709	0.387	+16%
Top 30%	0.998	0.348	0.516	0.652	0.445	+16%

that we use is The ensemble-dispersion score is computed as the predictive standard deviation across deep-ensemble members (Sec. 4.5).

2.4 Validation of ensemble uncertainty: informativeness, calibration, and utility

We audit ensemble-dispersion uncertainty because of its superiority for triage [27]. Here, the uncertainty score is computed from the predictive dispersion across members of deep ensemble (Sec. 4.5): we use the ensemble predictive variance (reported as standard deviation in Fig. 5) as a scalar uncertainty estimate. Throughout, "variance-only" refers to using only this ensemble-dispersion signal (without additional chemistry/physics heuristic terms) when ranking or gating candidates for triage.

A binned summary (Fig. 5(c)) provides qualitative context showing that higher predicted uncertainty is associated with larger errors, although we do not treat bin monotonicity as a standalone performance criterion. Predicted uncertainty correlates positively with absolute error (Fig. 5(a); Pearson r = 0.7049).

With overall ECE = [0.0961] under binning, reliability curves (Fig. 5(b)) show miscalibration; as a result, we report empirical coverage (Fig. 5(d)) and advise post-hoc calibration when calibrated intervals are needed. Prediction intervals attain near-nominal 95% coverage across thresholds and enlarge suitably with uncertainty (Fig. 5(e)).

To directly evaluate "capture" of high-error cases, we report a contingency analysis (Table 3) at fixed review budgets and error thresholds, including precision, recall, and miss rate. For example, at a 20% budget the uncertainty-based selector achieves precision 1.000 and recall 0.233 (F1=0.378), corresponding to a miss rate of 0.767; this quantitatively characterizes how many high-error cases are captured versus missed under a practical triage budget. Together, the correlation, coverage diagnostics, and contingency results show that ensemble-dispersion uncertainty provides an informative ranking signal for budgeted triage, even when interval calibration is imperfect.

2.5 Robustness, limitations, and where errors concentrate

We conduct sensitivity sweeps (details in the Supplementary Information) that demonstrate that using 4–8 interaction blocks results in diminishing returns beyond 6; that hidden dimensions from 128 to 512 have a sweet spot at 256; and that cutoff radii from

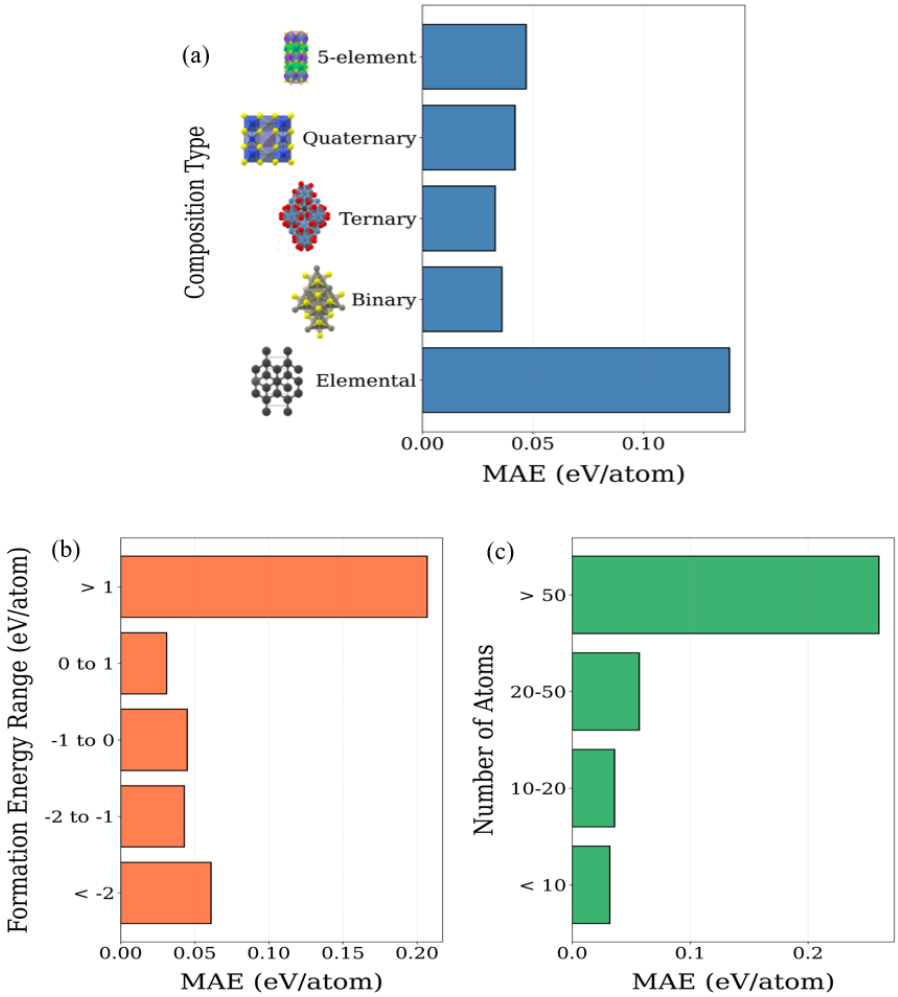


Fig. 6: **Multi-panel error breakdown.** (a) MAE by composition class (elemental, binary, ternary, quaternary, five-element). (b) MAE versus formation-energy range (≤ -2 to $\geq +1$ eV atom^{-1}). (c) MAE versus structure size (< 10 to > 50 atoms). Shows strongest performance in well-sampled regimes and larger errors for extreme compositions and sizes.

6–12 Å yield stable accuracy above 8 Å, with 10 Å optimal. The expected $\propto 1/\sqrt{N_{\text{train}}}$ trend is followed by learning curves across data fractions (10–100%), suggesting that more data will yield even greater gains.

Data-scarce, predictable regimes are revealed by error stratification (Fig. 6). According to composition (Fig. 6a), ternaries (2,233 samples; 0.033) are significantly easier than elemental systems (68 samples) (MAE = 0.139). According to the energy range (Fig. 6b), the accuracy is highest in the range 0–1 eV atom^{-1} (MAE = 0.031),

with errors rising at the extremes (e.g., >1 eV atom^{-1}: MAE = 0.207, 27 samples; and < −2 eV atom^{-1}). Error increases with unit-cell size according to structure size (Fig. 6c): <10 atoms (0.032), 10–20 (0.036), 20–50 (0.057), and >50 (0.260, 10 samples). The effectiveness of variance-only selection is explained by these regions' alignment with higher variance, which also offers specific goals for upcoming data collection or active learning.

Case studies.

We look at the six biggest absolute-error cases in the JARVIS-DFT test set to provide a foundation for the failure analysis. The outliers cover difficult chemical and structural regimes: a silicate (O_6Si_3, 0.96 eV atom^{-1}, 0.063), a chlorine-rich salt ($N_4O_{12}Cl_4K_8$, 1.01 eV atom^{-1}, 0.113), a large carbon cage (C_{60}, 1.03 eV atom^{-1}, 0.085), and an oxygen-rich cluster (O_8, |err| = 1.54 eV atom^{-1}, variance = 0.087). The set is completed by two diatomics: an actinide binary (U_2, 0.80 eV atom^{-1}) and a transition-metal dimer (Re_2, 0.92 eV atom^{-1}). The model *self-identifies* these predictions as unreliable without manual intervention, as evidenced by the high ensemble variance for all.

Then, for higher-fidelity computation, we simulate practitioner triage by escalating a budgeted subset and measuring *yield* as the fraction of escalations whose absolute error surpasses a selected threshold. Variance-prioritized escalation consistently performs better than random selection across thresholds of 0.10–0.20 eV atom^{-1} and budgets of 50–300 structures. For instance, the variance strategy captures 3.6% high-error cases versus 2.6% for random—an ∼40% relative uplift—when the budget is set at 220 and the threshold is set at 0.20 eV atom^{-1}. When combined, the screening uplift and the case-level inspection show that uncertainty estimates are not just diagnostic but also *operationally useful*.

Interpreting high-variance outliers (O_8, C_{60})

The high-error, high-variance cases we highlight (e.g., the O_8 oxygen cluster and the C_{60} carbon cage) can plausibly reflect two non-exclusive sources of uncertainty.Model limitations (epistemic uncertainty): The proposed model is trained exclusively on JARVIS-DFT crystalline materials with periodic boundary conditions. Cluster-like structures such as O_8 and C_{60} therefore lie in an extrapolation regime in terms of periodicity (non-periodic vs. periodic), coordination environments (surface-dominated vs. bulk-like), and bonding patterns (molecular-like vs. extended solid-state). Elevated ensemble dispersion in these cases indicates disagreement among models trained on the same data, consistent with epistemic uncertainty under out-of-distribution geometries. DFT reference uncertainty (aleatoric uncertainty): Cluster calculations can also exhibit higher label uncertainty due to convergence difficulty, sensitivity to exchange–correlation functional and spin state, and (for oxygen-rich systems) potential spin contamination. However, JARVIS-DFT does not provide detailed convergence metadata (e.g., SCF flags or force-threshold diagnostics), which prevents a quantitative attribution of error to reference-data quality for individual structures. Importantly, disentangling these sources is not required for the intended triage/active-screening

Table 4: Transfer to Matbench Perovskites. Fine-tuning enables rapid adaptation and strong accuracy.

Metric	Before FT	After 20-epoch FT	% change
MAE ($\mathrm{eV\,atom^{-1}}$)	0.581	**0.104**	−82.1%
RMSE ($\mathrm{eV\,atom^{-1}}$)	0.723	**0.139**	−80.8%
R^2	0.008	**0.964**	N/A

Table 5: Transfer Learning Ablation on Matbench Perovskites.

Training Method	Val MSE	Training Samples
From scratch	0.0085	15,142
Fine-tuned (JARVIS-DFT)	0.0097	15,142

use case: high ensemble dispersion conservatively flags cases that merit verification regardless of whether uncertainty stems primarily from model extrapolation or from reference-data quality. In safety-critical discovery settings, both scenarios warrant escalation. A definitive attribution would require targeted re-computation with enhanced convergence tracking and/or multi-fidelity references, which we view as valuable future work but beyond the scope of this study.

2.6 Generality via transfer learning

$R^2 = 0.008$ is the result of zero-shot transfer to Matbench Perovskites, indicating a shift in distribution. Performance improves to $R^2 = 0.964$ and MAE = 0.104 $\mathrm{eV\,atom^{-1}}$ after 20 epochs of fine-tuning with task-specific normalization (Fig. 7; Table 4), suggesting that GemNet pretraining learns reusable atomic-scale features that adapt effectively to new chemistries [28, 29].

To assess whether pretraining on JARVIS-DFT provides transferable representations, we compared training from scratch versus fine-tuning from our pretrained model on Matbench Perovskites (18,928 samples, 80/10/10 split). With matched hyperparameters (learning rate 1×10^{-4}, 50 epochs), both from-scratch training (MSE = 0.0085) and fine-tuning (MSE = 0.0097) achieved comparable validation performance (Table 5), demonstrating that our model architecture generalizes effectively to perovskite formation energy prediction. However, from-scratch training performs comparably in this setting

2.7 A framework for reliable and reproducible materials science

In the Supplementary Information, we confirm that there is no test leakage, enforce single-model evaluations on identical train/validation/test splits (80/10/10 on JARVIS-DFT), and offer a comprehensive reproducibility checklist. We combine the main negative outcomes for transparency's sake: First, according to the MAE of the escalated subset (0.0753 vs. 0.0677 $\mathrm{eV\,atom^{-1}}$; Fig. 4(b)), ranking by ensemble variance is 10.1% more effective than the suggested Gate-Hard selector at

surfacing difficult cases; in practice, the uncertainty-only policy more accurately captures hard cases. Second, FiLM-based domain adaptation improves computation by ∼10% but alters accuracy by <0.01% (0.037029 vs. 0.037025 eV atom^{-1}; Fig. 4(b), Table 2); FiLM provides no useful advantage on this single-domain benchmark. Third, exploratory DMET+VQE Δ-corrections are not included in the final pipeline and deteriorated full-test accuracy (Sec. 4).

Actionable triage is then made possible by validated ensemble uncertainty. Every outlier among the six largest residuals—oxygen clusters ($|\Delta E| = 1.54$ eV atom^{-1}), C_{60} (1.03 eV atom^{-1}), halide salts (1.01 eV atom^{-1}), silicates, transition-metal dimers, and actinide binaries—carries a high predictive variance, indicating that the model self-flags chemically atypical cases before deployment. Operational value is confirmed by a straightforward escalation simulation: across budgets ranging from 50 to 300 and error thresholds of 0.10 to 0.20 eV atom^{-1}, prioritizing by ensemble standard deviation consistently performs better than random selection. Increasing the top 220 high-variance predictions, for instance, captures 3.6% of samples with $|\Delta E| \geq 0.20$ eV atom^{-1} as opposed to 2.6% under a random policy—a relative uplift of ≈40%. When combined, the model offers cutting-edge accuracy and a transparent, data-driven decision layer that is appropriate for high-throughput materials discovery. Variance-prioritized escalation consistently outperforms random sampling, with relative improvements ranging from ∼35% to over 500%. The detailed results across budgets (50–300) and thresholds (0.10–0.20 eV atom^{-1}) are summarized in Table 6.

3 Discussion

This work offers a materials-focused, using a held-out test set validated baseline for reliability-aware crystalline formation energy prediction. We achieve single-model better performance on the held-out test set (MAE = 0.0370 eV atom^{-1}, RMSE = 0.0799 eV atom^{-1}, $R^2 = 0.9939$) and outperform a single-model ALIGNN baseline under identical evaluation using a GemNet-based directional message-passing network trained on JARVIS-DFT (36,029 crystals; fixed 80/10/10 split). Beyond headline error, we examine *operational reliability*: the relationship between uncertainty and error, the chemically interpretable failure of the model, and the reliability mechanisms that genuinely enhance decision-making in computational materials workflows..

Hard-case identification and uncertainty.

Choosing which predictions to advance to higher-fidelity theory is a crucial decision in high-throughput screening. We demonstrate that ranking by ensemble variance outperforms a materials-aware Gate-Hard heuristic (10.1% higher MAE in the chosen cohort) in identifying harder subsets. This is *materials-relevant*: high-variance selections focus on well-known difficult motifs that put a strain on both learned representations and DFT itself, such as oxygen-rich clusters, transition-metal dimers, actinide binaries, and large unit cells. Uncertainty is useful both practically (variance-prioritized escalation consistently outperforms random triage across realistic budgets and thresholds, enabling targeted use of compute for chemically interesting regions of phase space) and informatively (error–uncertainty correlation; monotone binned trends).

Table 6: **Variance-based triage vs. random sampling.** Yield is the fraction of escalated structures whose absolute error exceeds the specified threshold. Random yields are estimated over 2,000 trials (mean ± std). Relative improvement is computed as (Variance − Random)/Random × 100%.

Budget	Threshold (eV atom^{-1})	Variance yield	Random yield	Std	Relative improv.
50	0.10	0.340	0.076	0.036	+346.9%
50	0.15	0.260	0.042	0.028	+516.8%
50	0.20	0.080	0.025	0.022	+222.1%
100	0.10	0.250	0.078	0.026	+220.3%
100	0.15	0.170	0.044	0.020	+287.0%
100	0.20	0.060	0.026	0.016	+130.1%
150	0.10	0.220	0.078	0.021	+182.8%
150	0.15	0.133	0.044	0.016	+201.7%
150	0.20	0.053	0.026	0.013	+101.4%
200	0.10	0.200	0.078	0.019	+157.4%
200	0.15	0.110	0.043	0.014	+153.6%
200	0.20	0.040	0.026	0.011	+52.2%
220	0.10	0.200	0.077	0.018	+158.3%
220	0.15	0.105	0.044	0.014	+139.4%
220	0.20	0.036	0.026	0.011	+40.1%
250	0.10	0.200	0.078	0.016	+155.7%
250	0.15	0.104	0.044	0.013	+138.4%
250	0.20	0.036	0.026	0.010	+37.2%
280	0.10	0.193	0.078	0.016	+147.5%
280	0.15	0.100	0.044	0.012	+126.7%
280	0.20	0.036	0.026	0.009	+35.3%
300	0.10	0.200	0.078	0.015	+157.3%
300	0.15	0.100	0.043	0.011	+130.6%
300	0.20	0.037	0.026	0.009	+41.5%

Calibration and coverage.

Although uncertainty is still useful for *ranking*, reliability curves show miscalibration under binning (ECE = 0.0961). While perfect calibration is necesssary for policies requiring guaranteed coverage, the ranking quality provided by ensembles is sufficient for high-throughput screening and triage. Post-hoc variance scaling or inductive conformal prediction can be layered without retraining to align nominal and empirical coverage for screening policies that demand calibrated risk (such as queue management with coverage guarantees). Crucially, the improvements we show in screening uplift rely on ranking quality, which the ensemble already offers, rather than flawless calibration.

Where errors concentrate and why.

Larger residuals are linked by error taxonomy to regimes with limited data and chemistry: elemental phases versus ternaries; extreme formation energies (> 1 or < -2 eV atom^{-1}); and very large unit cells. These regimes (such as open-shell d-manifolds and near-degenerate states) are consistent with higher uncertainty and known physical complexity. Crucially, the variance-only triage mechanism is therefore

implicitly optimized to select samples from these identified high-error regimes: elemental phases, extreme formation energies, and large unit cells. In addition to offering specific goals for dataset expansion, active learning, or higher-fidelity recalculation, this materials-grounded perspective explains *why* variance-first triage works.

Negative results with practical value.

FiLM-based domain modulation adds ∼10% compute with $< 0.01\%$ MAE change on this single-domain bulk-crystal task, while exploratory DMET+VQE Δ-corrections reduce full-distribution accuracy. These findings limit the situations in which architectural or hybrid sophistication is justified in *materials* contexts: additional complexity may not result in operational benefit in the absence of multi-domain supervision, sufficiently accurate quantum labels, and task-appropriate answers. On the other hand, fine-tuning to a different crystal family (Matbench Perovskites) results in $R^2 = 0.964$ with modest adaptation, suggesting that the learned representation captures reusable, physics-informed bonding and coordination features.

Implications for deployment.

A straightforward, auditable recipe appears for high-throughput computational materials pipelines: (i) train a directional GNN using deterministic splits and train-only normalization; (ii) estimate uncertainty using a small deep ensemble; (iii) accept low-variance predictions and escalate the remainder to higher-fidelity theory (e.g., tighter DFT, many-body corrections) subject to budget; and (iv) report both coverage and uplift so decisions can be audited. The decision layer is an explicitly materials-aware policy based on the error landscape of formation energies rather than a general machine learning heuristic since uncertainty ranking corresponds with chemically interpretable failure modes.

Conclusions.

With the help of TRACER that attains cutting-edge accuracy on the JARVIS-DFT benchmark, we have introduced a thorough, reliability-first framework for computational materials science in this work. We have created a transparent baseline for formation energy prediction that outperforms current architectures such as ALIGNN by emphasising a fair, single-model-to-single-model comparison. Importantly, we have demonstrated that a compact ensemble ($N = 3$) offers a computationally efficient yet statistically robust signal for triaging "hard cases" in data-scarce regimes, complementing accuracy metrics with uncertainty–error association, empirical calibration/coverage via reliability analysis, and budgeted hard-case capture evaluations for uncertainty quantification (UQ).

The ensemble approach achieves computational efficiency while maintaining statistical robustness. Training a single GNN model requires 39 minutes on standard hardware, with ensemble costs scaling linearly ($N = 3$: 2.0 hours; $N = 5$: 3.3 hours; $N = 7$: 4.6 hours). Inference remains fast at ~ 26 ms per structure per model, enabling real-time screening. Statistical robustness is demonstrated through stable uncertainty–error correlations ($r = 0.70 \pm 0.04$) and consistent high-error detection performance

Table 7: Ensemble results across different N.

Ensemble	Correlation	ECE	F1@20% Mean	σ	Mean Error
$N = 3$	0.705	0.096	0.378	0.285	0.381
$N = 5$	0.620	0.047	0.391	0.303	0.342
$N = 7$	0.702	0.046	0.365	0.353	0.400

($F_1 = 0.38 \pm 0.01$) across ensemble sizes $N = 3$–7, computed on 3,604 test samples Tab. 7.

This framework provides a strong basis for a number of exciting extensions. The modular architecture is inherently adaptable to multi-task learning, even though this study concentrated on single-domain formation energy to guarantee a controlled benchmark. Inter-property correlations, such as those between band gaps, elastic moduli, and formation energy, can be used in future research to discover common representations that are more widely applicable across crystalline families. Furthermore, this mechanism remains a strong option for future models trained under true multi-domain supervision (e.g., combining JARVIS, Materials Project, and OCP), where domain shifts are inherent, even though ablation study demonstrated that domain adaptation (FiLM) was not required for a single-domain task.

Advanced active learning [30, 31] workflows are also made possible by the validated UQ system presented here. Future pipelines can close the loop between acquisition and retraining by combining variance-based triage with explicit compute budgets, focusing on the chemically complex regimes identified by error taxonomy. These uncertainty estimates are directly actionable for stringent queue management policies by integrating methods such as inductive conformal prediction to obtain distribution-free coverage guarantees, further tightening risk control.

Finally, identification of "hard cases" provides a precise entry point for hybrid quantum classical computing. By identifying these high-variance materials—such as transition-metal d-manifolds and near-degenerate states—the GNN acts as an efficient and economically viable pre-screener for subsequent costly quantum corrections. Rather than applying costly quantum corrections globally, future work can utilize variance-based selection to target quantum resources (e.g., QNN-based Δ-heads) exclusively on the specific chemistries where classical GNNs struggle, such as transition-metal d-manifolds and near-degenerate states. By tying uncertainty analysis to specific chemical and structural regimes, we offer a clear roadmap for gradually raising model sophistication where it matters most, providing the community with a transparent, repeatable basis for confidence-aware materials discovery.

4 Methods

Here, the proposed approach aims for *maximum transparency and reproducibility.* Fig. 1(a) summarizes the entire process from data ingestion to analysis, and each step is described in detail below.

4.1 Data curation and preprocessing

Data source.

We make use of the JARVIS-DFT dataset [20], which originally included 37,099 3D crystalline structures with formation energies calculated using DFT [Dataset available at: https://figshare.com/articles/dataset/jdft_3d-7-7-2018_json/6815699]. There are 36,029 distinct structures in the final corpus after duplicates and entries with incorrect or failed computations have been eliminated.

Structure standardization.

All structures are standardized to *primitive* cells using `spglib` in order to prevent spurious multiplicity across conventional cells. This minimizes confounding variation and guarantees that symmetry-equivalent crystals share a canonical representation.

Deterministic splits.

To generate non-overlapping splits, we employ a fixed random seed (42): test 3,604 (10%), validation 3,602 (10%), and train 28,823 (80%). For all final metrics and analyses, the test set is excluded from training and model selection.

Normalization.

To avoid leakage, Z-score normalization of formation energy per atom is done solely with *training-set* statistics (mean = 0.002190 eV atom^{-1}; std = 1.000787 eV atom^{-1}). Every reported error is *denormalized* to eV atom^{-1}.

4.2 Graph construction

We translate atomic structures into PyTorch Geometric (PyG)-compatible graphs. Every atom i turns into a node with the initial feature $h_i^{(0)} = \text{Embed}(Z_i) \in \mathbb{R}^{256}$, where Z_i is the atomic number. Atom pairs (i, j) with a periodic distance d_{ij} less than 10.0 Åare given undirected edges. A smooth geometric descriptor R_{ij} is formed by expanding distances using a 50-component Gaussian radial basis. PyG's `Batch` collator is used to assemble mini-batches into a disconnected "super-graph" during training in order to optimize GPU throughput without the need for explicit padding.

4.3 GNN architecture (GemNet)

With directional messages traversing 3D neighborhoods, the proposed GemNet model [16, 32] is shown schematically in Fig. 2(a).

Key components and hyperparameters.

Using geometric features, the model updates node and edge states through six sequential directional message-passing interaction blocks. All nodes, edges, and MLPs have a hidden dimension of 256. A cutoff radius of 10.0 Åis used to define neighborhoods. A 50-Gaussian radial basis expansion is used to encode interatomic distances. SiLU is the nonlinearity, and for stability, residual connections with layer normalization are

Table 8: Effect of ensemble size on uncertainty quantification metrics. Correlation denotes the association between ensemble dispersion (e.g., σ) and absolute prediction error.

Ensemble Size	Correlation	ECE	Mean σ	Mean Error
3	0.7049	0.0961	0.2845	0.3805
5	0.6203	0.0467	0.3026	0.3419
7	0.7024	0.0463	0.3534	0.3997

applied after every block. To determine the size-intensive per-atom energy $E_{\text{per–atom}}$, readout divides the total energy E_{total} by the number of atoms.

4.4 Training protocol and reproducibility

Optimization.

With Adam, we train for up to 50 epochs (learning rate $= 1 \times 10^{-4}$, weight decay $= 1 \times 10^{-5}$), batch size 16, and MSE loss on *normalized* targets/predictions).

Model selection.

With patience 10, early stopping tracks validation loss; the checkpoint with *minimum* validation loss is kept. Fig. 4(a) illustrates learning dynamics.

Hardware and determinism.

For a full 50 epoch, experiments are conducted on NVIDIA RTX 4090 GPUs with 24 GB of VRAM. We enable deterministic kernels (`torch.use_deterministic_algorithms(True)`) and set global seeds (PyTorch/NumPy/splits) to 42. The SI contains information about OS, CUDA/cuDNN, and package versions.

4.5 Reliability and uncertainty quantification (UQ)

Deep ensembles.

We form a *deep ensemble* with $\mathbf{N = 3}$ independently initialized GemNet models (same hyperparameters; distinct weight seeds). Given per-model predictions y_i, the predictive variance is

$$\sigma^2 \;=\; \frac{1}{N}\sum_{i=1}^{N}\big(y_i - \bar{y}\big)^2, \qquad \bar{y} \;=\; \frac{1}{N}\sum_{i=1}^{N} y_i \,. \tag{1}$$

This σ^2 serves as the primary ranking signal ("variance-only") and as an input to Gate-Hard.

Ensemble size rationale and sensitivity (N=3,5,7).

We initially selected a small deep ensemble ($N = 3$) as a compute-efficient baseline for large-scale screening, since both training and inference costs scale approximately linearly with ensemble size. To evaluate whether this choice materially affects uncertainty behavior, we repeated the analysis with $N = 3, 5$, and 7 (Table 8).

We find that $N = 3$ yields the strongest uncertainty–error association (correlation $r = 0.7049$), indicating that ensemble dispersion provides an effective ranking signal for epistemic uncertainty in this domain. Calibration, as measured by ECE, improves with larger ensembles ($0.0961 \to 0.0467 \to 0.0463$ for $N = 3, 5, 7$), consistent with reduced sampling noise in the dispersion estimates. However, our primary use case is triage/active screening, where identifying and prioritizing hard cases is more important than obtaining precisely calibrated probabilities. In this setting, hard-case identification performance is stable across ensemble sizes at practical budgets, and $N = 3$ provides a favorable cost–benefit tradeoff for screening on the order of 4×10^4 structures.

Uncertainty evaluation and reliability assessment.

To avoid ambiguity in terminology, we follow VVUQ conventions and reserve the term *validation* for assessing agreement between model predictions and reference data for a stated context of use. In contrast, *uncertainty quantification* provides an uncertainty estimate, and *uncertainty evaluation* assesses how well those estimates behave (e.g., calibration/coverage and decision utility). In this work, we use *reliability* to refer to the trustworthiness of the reported predictive uncertainty: when the model outputs a prediction interval at a nominal confidence level (e.g., 95%), the empirical frequency with which the ground-truth value falls inside that interval should match the nominal level (coverage), while the intervals remain informative (not excessively wide). In this section, "calibration" refers to *calibration of predictive uncertainty* (i.e., agreement between nominal and empirical coverage of prediction intervals), not calibration of model parameters.

We assess three criteria for the reported uncertainty estimates: (i) informativeness (sharpness/heterogeneity), i.e., whether uncertainty meaningfully varies across samples rather than collapsing to a near-constant value; (ii)Coverage calibration (reliability diagrams and ECE) computed from reliability diagrams that bin samples by predicted confidence and compare nominal and observed coverage; and (iii) operational utility through empirical coverage and prediction-interval behavior (e.g., interval widening with increasing uncertainty), which supports downstream triage/escalation policies.

Sources of uncertainty and treatment.

For clarity, we enumerate the primary sources of uncertainty relevant to formation-energy prediction and summarize how each is treated in this work.

(1) Model-parameter uncertainty (epistemic). Because the model is fit from finite data and optimized in a non-convex landscape, learned weights are uncertain. *Treatment:* we approximate epistemic uncertainty using a deep ensemble trained with different random initializations; predictive dispersion (ensemble variance/standard deviation) provides the scalar uncertainty score used for ranking/triage.

(2) Optimization variability (epistemic). Stochastic training (initialization and SGD noise) leads to different local minima. *Treatment:* this variability is intentionally leveraged as part of the ensemble construction; averaging reduces idiosyncratic model behavior while dispersion quantifies disagreement.

Table 9: Summary of uncertainty sources and treatment in this work.

Uncertainty source	Type	Treatment in this work
Model parameters	Epistemic	Deep ensemble; predictive dispersion (variance/standard deviation) used as uncertainty score for ranking/triage
Optimization variability	Epistemic	Seed/initialization diversity across ensemble members; averaging reduces idiosyncrasies and dispersion measures disagreement
Distribution shift (OOD)	Epistemic	High dispersion flags extrapolation regimes for escalation (e.g., cluster-like cases)
DFT reference labels	Aleatoric	Not explicitly modeled (no per-structure convergence metadata); conservative escalation when dispersion is high
Representation limits	Epistemic	Directional message passing and ensemble diversity mitigate; unmodeled factors (e.g., magnetism) noted as limitation

(3) Input distribution shift / out-of-distribution (epistemic). Test structures can differ from the JARVIS-DFT training distribution (periodic crystals), e.g., cluster-like geometries. *Treatment:* elevated ensemble dispersion flags extrapolation regimes for escalation, which is the intended behavior in screening workflow (e.g., O_8, C_{60}).

(4) Reference-label uncertainty in DFT data (aleatoric). DFT reference energies can carry irreducible uncertainty from functional approximations and numerical/convergence effects; such noise is not directly observable per structure in JARVIS-DFT. *Treatment:* we do not explicitly model per-structure label noise because convergence metadata are not available; instead, we adopt a conservative triage stance in which high dispersion triggers verification regardless of whether uncertainty arises from model extrapolation or from potentially noisy references.

(5) Representation limits (epistemic). A GNN may omit latent physical factors (e.g., magnetic/spin states) not explicitly encoded in the input graph. *Treatment:* the GemNet architecture mitigates some representation limitations via directional message passing, and ensemble diversity captures complementary representations; nonetheless, unmodeled factors (e.g., magnetism) remain a limitation and a direction for future work.

Table 9 summarizes these sources and treatments.

4.6 Gate-Hard ranking system

Gate-Hard (Fig. 1(b)) explores whether materials-aware cues improve triage beyond variance alone. The score is a linear fusion of statistical, chemical, and physics proxies:

$$\text{Score} \;=\; \alpha\,\sigma^2 \;+\; \beta\,\text{TM} \;+\; \gamma\,\delta\,, \tag{2}$$

where σ^2 is the ensemble variance (Eq. 1); TM is a boolean flag indicating the presence of transition-metal elements; and δ is a normalized near-degeneracy proxy.

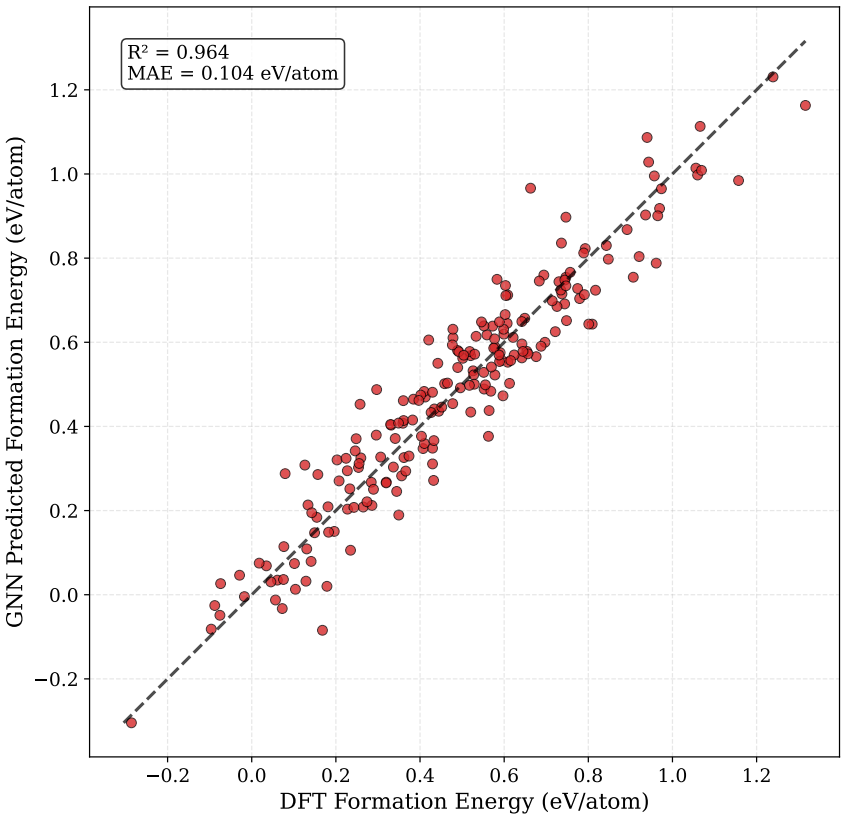


Fig. 7: **Parity plot for Matbench Perovskites after transfer learning.** Predicted versus DFT formation energies with density shading; annotated performance $R^2 = 0.964$ and MAE $= 0.104$ eV atom^{-1}, confirming strong generalisation.

Sensitivity of Gate-Hard to (α, β, γ) ***tuning.***

To assess sensitivity of Gate-Hard to validation-tuned weights, we performed a weight ablation study (Table 11). The best-performing configuration was uncertainty-only (ensemble-dispersion) ranking with $\alpha = 1.0$, $\beta = 0.0$, $\gamma = 0.0$, achieving a mean error of 0.072315 eV/atom on the selected cohort (higher indicates better hard-case enrichment under our definition). In contrast, our current configuration $(1.0, 0.5, 0.1)$ attains a mean error of 0.067723 eV/atom, corresponding to a 6.4% degradation relative to uncertainty-only. Across the tested settings, increasing heuristic weights (larger β and/or γ) did not improve hard-case identification beyond ensemble dispersion, and in several cases reduced the high-error fraction. These results indicate that, for this dataset and model, ensemble dispersion alone provides the strongest triage signal, while simple chemical/physical heuristics offer limited additive value; therefore,

Table 10: Gate-Hard weight sensitivity analysis.

Configuration	α	β	γ	Mean Error	Median Error	High-Error %
Variance Only	1.0	0.0	0.0	0.0723	0.0550	20.0%
Degeneracy Strong	1.0	0.0	0.5	0.0723	0.0550	20.0%
Degeneracy Light	1.0	0.0	0.1	0.0723	0.0550	20.0%
Tm Light	1.0	0.2	0.0	0.0699	0.0514	18.9%
Balanced Light	1.0	0.2	0.1	0.0699	0.0514	18.9%
Current	1.0	0.5	0.1	0.0677	0.0499	17.8%
Tm Strong	1.0	1.0	0.0	0.0674	0.0498	17.8%
Balanced Strong	1.0	1.0	0.5	0.0674	0.0498	17.8%
Equal Weight	1.0	1.0	1.0	0.0674	0.0498	17.8%

Table 11: Gate-Hard weight sensitivity analysis. Higher MAE indicates better identification of hard cases.

Configuration	α	β	γ	Mean Error	Median Error	High-Error %
Variance Only	1.0	0.0	0.0	0.0723	0.0550	20.0%
Degeneracy Strong	1.0	0.0	0.5	0.0723	0.0550	20.0%
Degeneracy Light	1.0	0.0	0.1	0.0723	0.0550	20.0%
Tm Light	1.0	0.2	0.0	0.0699	0.0514	18.9%
Balanced Light	1.0	0.2	0.1	0.0699	0.0514	18.9%
Current	1.0	0.5	0.1	0.0677	0.0499	17.8%
Tm Strong	1.0	1.0	0.0	0.0674	0.0498	17.8%
Balanced Strong	1.0	1.0	0.5	0.0674	0.0498	17.8%
Equal Weight	1.0	1.0	1.0	0.0674	0.0498	17.8%

we use uncertainty-only ranking as the default escalation rule and report Gate-Hard primarily as an ablation.

Tuning and evaluation.

Only the validation set is used to adjust the weights (α, β, γ). By choosing the top 7.5% (270 samples) under each ranking and calculating the *MAE of the chosen subset*, we compare Gate-Hard to *variance-only* on the held-out test set (Fig. 4(b)). In our setting, variance-only performs 10.1% better than Gate-Hard since *higher* subset MAE indicates a *harder* cohort.

4.7 Benchmarking and validation protocols

ALIGNN baseline (Table 1).

In accordance with its published protocol, ALIGNN is trained *from scratch* on the same 80/10/10 split, and normalization is calculated using the same training set. Every headline comparison is between a single model.

To ensure a quantitative baseline, we utilized the official open-source implementation of ALIGNN. We trained the model from scratch using the authors' recommended default hyperparameters (learning rate, weight decay, and graph construction parameters) on the exact same 80/10/10 splits used for the proposed GemNet models. This

ensures that the reported 25.8% improvement reflects a genuine difference in data efficiency and architectural fit for this dataset, rather than hyperparameter tuning bias.

Transfer learning (Fig. 7).

Starting from the JARVIS-DFT GemNet baseline, we use task-specific normalization to fine-tune the *entire* network for 20 epochs at learning rate 1×10^{-5} on Matbench Perovskites (15,142 training samples). After fine-tuning, test performance is reported.

Ablation study (Tab. 2).

To isolate the impact of architecture, *identical* hyperparameters, splits, and seed are used to train architectural variants (baseline, domain-embedding only, and full FiLM). The same hardware is used to measure relative computational cost.

Robustness checks.

The cutoff radius (6–12 Å), interaction blocks (4–8), and hidden dimension (128–512) are the main hyperparameters that we change one at a time while keeping the others constant at the nominal optimum. By training on fractions of the data (10–100%) and reevaluating test MAE, we also investigate sample-efficiency.

Data Availability

The analysis in this study uses the publicly available JARVIS-DFT 3D dataset, accessible on Figshare at: https://figshare.com/articles/dataset/jdft_3d-7-7-2018_json/6815699.

Code Availability

The code for the TRACER workflow is available at TRACER GitHub Repository.

Author contributions

G.D. designed the study, developed the TRACER framework, and implemented the GemNet-based architecture together with all uncertainty-quantification pipelines. He conducted all computational experiments, including robustness analyses, ALIGNN benchmarking, transfer-learning evaluations, and data curation from JARVIS-DFT. He prepared the initial manuscript draft. S.S. and M.B. supervised the research, providing supervision and scientific guidance throughout the development of the framework. They offered editorial feedback, manuscript revision, and review of all analyses. M.B. additionally oversaw project direction, ensured methodological rigor, and contributed to securing the resources and financial support that enabled the research.

Competing interests

The authors declare no competing interests.

Funding

This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-25-1-0322.

References

[1] Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A.: Machine learning for molecular and materials science. Nature **559**(7715), 547–555 (2018)

[2] Nematov, D., Hojamberdiev, M.: Machine learning-driven materials discovery: Unlocking next-generation functional materials – a review. Computational Condensed Matter **45**, 01139 (2025) https://doi.org/10.1016/j.cocom.2025.e01139

[3] Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al.: Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL materials **1**(1) (2013)

[4] Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Physical review **140**(4A), 1133 (1965)

[5] Loew, A., Sun, D., Wang, H.-C., Botti, S., Marques, M.A.: Universal machine learning interatomic potentials are ready for phonons. npj Computational Materials **11**(1), 178 (2025)

[6] Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., Hoesel, C., Schopmans, H., Sommer, T., Friederich, P.: Graph neural networks for materials science and chemistry (2022). https://arxiv.org/abs/2208.09481

[7] Tan, L., Chen, P., Liu, M., Wang, X., Cen, J., Zou, Q.: Benchmarking gnns for ood materials property prediction with uncertainty quantification. arXiv preprint arXiv:2511.11697 (2025)

[8] Zhao, Z., Li, H.-F.: Investigating material interface diffusion phenomena through graph neural networks in applied materials. ACS Applied Materials & Interfaces **16**(39), 53153–53162 (2024)

[9] Xie, T., Grossman, J.C.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters **120**(14), 145301 (2018)

[10] Schütt, K.T., Tkatchenko, A., Müller, K.-R.: Learning representations of molecules and materials with atomistic neural networks. In: Machine Learning Meets Quantum Physics

[11] Choudhary, K., DeCost, B.: Atomistic line graph neural network for improved

materials property predictions. npj Computational Materials **7**(1), 185 (2021)

[12] Chen, C., Ye, W., Zuo, Y., Zheng, C., Ong, S.P.: Graph networks as a universal machine learning framework for molecules and crystals. Chemistry of Materials **31**(9), 3564–3572 (2019)

[13] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J.P., Kornbluth, M., Molinari, N., Smidt, T.E., Kozinsky, B.: E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature communications **13**(1), 2453 (2022)

[14] Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C.J., Kornbluth, M., Kozinsky, B.: Learning local equivariant representations for large-scale atomistic dynamics. Nature Communications **14**(1), 579 (2023)

[15] Batatia, I., Kovacs, D.P., Simm, G., Ortner, C., Csányi, G.: Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. Advances in neural information processing systems **35**, 11423–11436 (2022)

[16] Gasteiger, J., Groß, J., Günnemann, S.: Directional message passing for molecular graphs. arXiv preprint arXiv:2003.03123 (2020)

[17] Stracuzzi, D.J., Chen, M.G., Darling, M.C., Peterson, M.G., Vollmer, C.: Uncertainty quantification for machine learning. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States) (2017)

[18] Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., *et al.*: Open catalyst 2020 (oc20) dataset and community challenges. Acs Catalysis **11**(10), 6059–6072 (2021)

[19] Tran, R., Lan, J., Shuaibi, M., Wood, B.M., Goyal, S., Das, A., Heras-Domingo, J., Kolluru, A., Rizvi, A., Shoghi, N., *et al.*: The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. ACS Catalysis **13**(5), 3066–3084 (2023)

[20] Choudhary, K., Garrity, K., Reid, A., DeCost, B., Biacchi, A., Walker, A.H., Trautt, Z., Hattrick-Simpers, J., Kusne, A., Centrone, A., Davydov, A., Tavazza, F., Jiang, J., Pachter, R., Cheon, G., Reed, E., Agrawal, A., Qian, X., Sharma, V., Zhuang, H., Kalinin, S., Pilania, G., Acar, P., Mandal, S., Vanderbilt, D., Rabe, K.: The joint automated repository for various integrated simulations (jarvis) for data-driven materials design (6) (2020) https://doi.org/10.1038/s41524-020-00440-1

[21] Himanen, L., Geurts, A., Foster, A.S., Rinke, P.: Data-driven materials science: Status, challenges, and perspectives. Advanced Science **6**(21) (2019) https://doi.org/10.1002/advs.201900808

[22] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive

uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

[23] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016). PMLR

[24] Angelopoulos, A.N., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511 (2021)

[25] Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep Evidential Regression (2020). https://arxiv.org/abs/1910.02600

[26] Borge-Durán, I., Bejarano, E.A., Araya, L.A., Arguedas, M.G., Özcan, E., Woodward, S., Figueredo, G.: Explainable gnn-derived structure-property relationships in interstitial-alloy materials (2024)

[27] Farris, R., Telari, E., Artrith, N., Neyman, K., Bruix, A.: Bayesian neural networks versus deep ensembles for uncertainty quantification in machine learning interatomic potentials. arXiv preprint arXiv:2509.19180 (2025)

[28] Falk, J., Bonati, L., Novelli, P., Parrinello, M., Pontil, M.: Transfer learning for atomistic simulations using gnns and kernel mean embeddings. Advances in Neural Information Processing Systems **36**, 29783–29797 (2023)

[29] Kooverjee, N., James, S., Van Zyl, T.: Investigating transfer learning in graph neural networks. Electronics **11**(8), 1202 (2022)

[30] Kusne, A.G., Yu, H., Wu, C., Zhang, H., Hattrick-Simpers, J., DeCost, B., Sarker, S., Oses, C., Toher, C., Curtarolo, S., *et al.*: On-the-fly closed-loop materials discovery via bayesian active learning. Nature communications **11**(1), 5966 (2020)

[31] Dale, A.S., Li, K., DeCost, B., Wan, H., Han, Y., Fehlis, Y., Hattrick-Simpers, J.: When Active Learning Fails, Uncalibrated Out of Distribution Uncertainty Quantification Might Be the Problem (2025). https://arxiv.org/abs/2511.17760

[32] Gasteiger, J., Shuaibi, M., Sriram, A., Günnemann, S., Ulissi, Z., Zitnick, C.L., Das, A.: Gemnet-oc: developing graph neural networks for large and diverse molecular simulation datasets. arXiv preprint arXiv:2204.02782 (2022)

Nomenclature

ALIGNN	Atomistic Line Graph Neural Network
CGCNN	Crystal Graph Convolutional Neural Network
cuDNN	CUDA Deep Neural Network library
CUDA	Compute Unified Device Architecture (NVIDIA)
DFT	Density Functional Theory
DMET	Density Matrix Embedding Theory
EAM	Embedded Atom Method
ECE	Expected Calibration Error
FiLM	Feature-wise Linear Modulation
FT	Fine-tuning
GNN	Graph Neural Network
GPU	Graphics Processing Unit
JARVIS	Joint Automated Repository for Various Integrated Simulations
JARVIS-DFT	JARVIS Density Functional Theory dataset
MAE	Mean Absolute Error
MACE	Message Passing Atomic Cluster Expansion
MC	Monte Carlo (e.g., MC dropout)
MEGNet	Materials Graph Network
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NequIP	Neural Equivariant Interatomic Potentials
OCP	Open Catalyst Project
OS	Operating System
PyG	PyTorch Geometric
QNN	Quantum Neural Network
R^2	Coefficient of determination
RMSE	Root Mean Squared Error
SI	Supplementary Information
spglib	Space-group symmetry library (“spglib”)
TM	Transition metal
TRACER	Reliability-first baseline and evaluation workflow introduced in this work
UQ	Uncertainty Quantification
VQE	Variational Quantum Eigensolver
VRAM	Video Random Access Memory