Ensemble Graph Neural Networks for High-Throughput Battery Materials Discovery with Uncertainty Quantification

Saher Elsayed[0009−0007−7672−264X]

University of Pennsylvania, Philadelphia, PA, USA
selsayed@seas.upenn.edu

Abstract. The computational bottleneck in battery materials discovery necessitates efficient alternatives to density functional theory (DFT) calculations. We present an ensemble graph neural network (GNN) framework combining SchNet, CGCNN, and MEGNet with physics-informed edge features and angular information to simultaneously predict formation energies, voltages, band gaps, and capacities of candidate battery materials. Evaluated on 500 battery material candidates spanning lithium transition metal oxides, phosphates, and sulfides, the ensemble achieves a formation energy mean absolute error (MAE) of 44 meV/atom and a voltage MAE of 116 mV, representing 40–70% improvement over individual GNN baselines. The framework provides a 30,000× speedup over DFT with R^2 exceeding 0.99, enabling screening of millions of candidates in hours rather than decades. We further address key limitations of prior work through: Monte Carlo Dropout (MC Dropout) uncertainty quantification (UQ) with 90% prediction interval calibration, a false-positive/false-negative screening analysis, out-of-distribution (OOD) generalization tests on held-out chemistry classes, hyperparameter sensitivity analysis, and comparison with the equivariant NequIP architecture. Ablation studies confirm the complementary contributions of geometric features and ensemble aggregation. These results advance AI-accelerated materials screening and demonstrate that physics-informed ensemble GNNs provide a practical, reliable foundation for guiding experimental synthesis in next-generation energy storage.

Keywords: Graph Neural Networks · Battery Materials · Materials Discovery · Uncertainty Quantification · High-Throughput Screening · Energy Storage

1 Introduction

Discovering next-generation battery materials requires evaluating thousands of candidates across competing electrochemical objectives. Density functional theory (DFT) calculations demand hours of wall-clock time per material on high-performance computing clusters [18, 9]. At this throughput, exhaustive screening of even modest candidate libraries is computationally prohibitive. Machine

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

learning (ML), particularly graph neural networks (GNNs), offers a transformative alternative by learning structure-property relationships directly from atomic graphs [30, 28]. While recent GNNs achieve strong performance on individual properties [7, 35], simultaneously predicting multiple electrochemical targets with reliable uncertainty estimates and demonstrated OOD generalization remains an open challenge.

This work addresses these gaps through an ensemble GNN framework tailored for battery materials screening. The physical motivation for combining three architecturally distinct models is grounded in the complementary inductive biases each brings: SchNet [31] excels at capturing radial distribution functions relevant to formation energy; CGCNN [35] discriminates coordination environments critical for voltage; and MEGNet's [7] global graph state is well suited to extensive properties such as capacity. By combining these models with learned aggregation weights and physics-informed edge features encoding bond angles and coordination numbers, the framework exploits multi-scale geometric information that no single architecture captures fully.

Our contributions are:

- An ensemble combining SchNet, CGCNN, and MEGNet with physics informed edge features and
  property-specific learned aggregation weights, achieving 44 meV/atom formation energy MAE and
  116 mV voltage MAE, 40–70% over individual baselines.
- MC Dropout UQ with calibration assessment and a screening false-positive /false-negative analysis
  using a clearly defined stability threshold.
- OOD evaluation on three held-out chemistry classes (sulfides, phosphates, and high-voltage oxides)
  absent from training, with uncertainty as a reliability signal.
- Hyperparameter sensitivity analysis revealing the physical rationale behind optimal choices, and a
  systematic comparison with the equivariant NequIP architecture [5].
- A tiered screening workflow that reduces DFT verification burden by 82% while cutting the
  false-negative rate (FNR) from 4.1% to 1.8%.

2 Related Work

2.1 Descriptor-Based Machine Learning for Materials Properties

Before GNNs, the dominant approach to ML-accelerated materials science relied on hand-crafted structural descriptors including symmetry functions [6], Coulomb matrices [29], SOAP [2], MBTR [19], and GAP [3]. Ward et al. [34] and Isayev et al. [20] showed that elemental and structural descriptors combined with ensemble methods provide competitive baselines for bulk property prediction. However, these approaches require domain expertise to design features and do not learn directly from raw atomic coordinates.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

2.2 Graph Neural Networks for Crystals and Molecules

The neural message passing framework [17] established strong benchmarks on quantum chemistry datasets. CGCNN [35] adapted graph convolutions to periodic crystal structures and outperformed descriptor-based methods on Materials Project data. SchNet [31] pioneered continuous-filter convolutions with distance-dependent kernels. MEGNet [7] incorporated global graph state variables enabling prediction of both intensive and extensive properties. Directional GNNs extended pairwise message passing to include bond angles (DimeNet [16], GemNet [15]) and 3D geometry via spherical Bessel functions (SphereNet [27]), consistently improving over distance-only baselines and motivating our inclusion of angular edge features.

2.3 Equivariant Architectures

NequIP [5] demonstrated that $E(3)$-equivariant message passing dramatically improves data efficiency for interatomic potentials and forces. Importantly, NequIP was designed for potential energy surface fitting, not for macroscopic global properties such as band gaps or intercalation capacities, which require whole-graph pooling. This distinction is relevant to the comparisons in Section 4: our ensemble, optimized for global multi-property prediction, matches or outperforms NequIP on these tasks despite the latter's stronger geometric priors. MACE [4] and Equiformer [25] extended equivariant ideas further, though neither has been systematically combined into ensembles for battery property prediction.

2.4 Battery Materials Discovery and Uncertainty Quantification

High-throughput computational screening has transformed battery materials research. The Materials Project [21] provides DFT-computed properties for over 150,000 inorganic materials. Sendek et al. [32] screened more than 12,000 solid-state electrolyte candidates; voltage prediction [33] and capacity estimation [26] remain critical bottlenecks. Reliable UQ is a prerequisite for deploying ML models in scientific discovery pipelines. MC Dropout [14] approximates Bayesian inference at low cost; deep ensembles [24] are generally better calibrated at higher training overhead. Conformal prediction [1] offers distribution-free coverage guarantees. In the materials context, calibrated uncertainty has been used to prioritize DFT calculations [8] and detect OOD inputs [22]. We adopt MC Dropout for its compatibility with the GNN ensemble structure and evaluate calibration explicitly.

3 Methodology

3.1 Scientific Motivation and Dataset

Battery electrode materials must satisfy several competing physical requirements: thermodynamic stability (formation energy E_f), high electrochemical

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

potential (voltage V), electronic insulation (band gap E_g), and large lithium storage capacity C. These properties arise from distinct physical mechanisms, motivating the use of architecturally diverse GNNs with complementary inductive biases rather than a single monolithic model.

The dataset comprises 500 battery material candidates spanning lithium transition metal oxides (LCO, LMO, NMC variants), phosphates (LFP), and sulfides. Ground-truth labels are DFT-calculated from the Materials Project [21] using VASP with the PBE functional (ENCUT = 520 eV, Monkhorst-Pack k-mesh, Dudarev GGA+U for transition metals). Property distributions: $E_f \in [-3.8, -0.2]$ eV/atom, $V \in [2.1, 5.3]$ V, $E_g \in [0.0, 4.2]$ eV, $C \in [60, 280]$ mAh/g.

Data splits. 400 training, 50 validation, and 100 test materials. Sulfides (n=28), phosphates (n=24), and high-voltage oxides (n=31, V>4.2 V) are held out entirely for OOD evaluation, ensuring genuine chemical extrapolation.

Data augmentation. Random rigid rotations (from $SO(3)$) and small position perturbations ($\delta\mathbf{r} \sim \mathcal{U}(-0.05, +0.05)$ Å) are applied during training. Rotations are physically valid because all target properties are rotationally invariant scalars; perturbations simulate thermal displacement and provide regularization.

3.2 Graph Representation

A crystal structure is represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with cutoff radius $r_c = 5.0$ Å. Node features are:

$$\mathbf{h}_i^{(0)} = [\mathbf{z}_i, \mathbf{x}_i] \in \mathbb{R}^d \tag{1}$$

where $\mathbf{z}_i$ is a learned embedding of atomic number and $\mathbf{x}_i$ encodes electronegativity, ionic radius, and formal oxidation state. Edge features encode multi-scale geometry:

$$\mathbf{e}_{ij} = [\mathbf{r}_{ij}, \cos\theta_{ijk}, CN_i] \in \mathbb{R}^{d_e} \tag{2}$$

where $\mathbf{r}_{ij}$ is the interatomic distance in a Gaussian radial basis (50 basis functions, $\sigma = 0.2$ Å), θ_{ijk} is the bond angle at atom i with closest neighbor k, and CN_i is the coordination number. CN_i directly determines polyhedral bonding type and correlates with oxidation state stability and voltage.

3.3 Ensemble GNN Architecture

Three complementary GNN architectures are combined, each contributing distinct inductive biases.

SchNet [31]: continuous-filter convolutions,

$$\mathbf{m}_{ij} = \mathbf{h}_i^{(l)} \odot \mathrm{MLP}_{\mathrm{filter}}(\|\mathbf{r}_{ij}\|) \tag{3}$$

effective at capturing radial distributions central to formation energy.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

CGCNN [35]: edge-conditioned convolutions,

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)} + \sum_{j \in \mathcal{N}(i)} \sigma(\mathbf{z}_{ij}^{(l)}) \odot \mathbf{g}(\mathbf{h}_j^{(l)})\right) \tag{4}$$

excelling at distinguishing coordination environments for voltage prediction.

MEGNet [7]: global graph state updated alongside node and edge states,

$$\mathbf{u}^{(l+1)} = \phi_u\Big(\mathbf{u}^{(l)},\, \rho_v(\{\mathbf{h}_i^{(l)}\}),\, \rho_e(\{\mathbf{e}_{ij}^{(l)}\})\Big) \tag{5}$$

suited to extensive, whole-graph properties such as intercalation capacity.

Each component uses 3 message-passing layers with hidden dimension 128 (approximately 0.5M, 0.4M, and 0.6M parameters for SchNet, CGCNN, and MEGNet, selected by validation-set grid search).

Ensemble aggregation. The final prediction is a learned weighted sum:

$$\hat{y}_p = \sum_{k=1}^{3} w_{k,p} \cdot f_k(\mathcal{G}), \qquad \sum_{k=1}^{3} w_{k,p} = 1, \quad w_{k,p} \geq 0 \tag{6}$$

where $p \in \{E_f, V, E_g, C\}$ and $w_{k,p}$ are property-specific weights optimized jointly on the validation set via L-BFGS with simplex constraints. Separate weights per property are necessary because MEGNet dominates for capacity while SchNet and CGCNN contribute more strongly to formation energy and voltage, respectively (Table 5).

3.4 Uncertainty Quantification via Monte Carlo Dropout

Dropout ($p = 0.10$) is applied after each hidden layer and kept active at inference time [14]. For each material, $M = 100$ stochastic forward passes yield predictive mean μ and standard deviation σ. The 90% prediction interval $[\mu - 1.645\sigma, \mu + 1.645\sigma]$ is evaluated against held-out test labels.

MC Dropout captures epistemic uncertainty (reducible by more training data) rather than aleatoric uncertainty from DFT numerical noise. In the screening workflow, MC Dropout σ flags candidates where the model extrapolates beyond its training distribution. The strong correlation between $\bar{\sigma}$ and OOD MAE ($r = 0.94$, Section 4.6) supports this interpretation. MC Dropout tends toward slight overconfidence in distribution tails relative to deep ensembles [24]; future work should compare against conformal prediction [1] for tighter coverage guarantees.

3.5 Training Procedure

Each GNN component is trained independently with the Adam optimizer [23], initial learning rate $\eta = 5 \times 10^{-4}$ (decayed by factor 0.95 every 50 epochs), L_2

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

regularization $\lambda = 10^{-5}$, and gradient clipping at norm 1.0. Early stopping uses patience 50 with a maximum of 300 epochs. The multi-property loss is:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\left(\alpha_1|E_{f,i} - \hat{E}_{f,i}| + \alpha_2|V_i - \hat{V}_i| + \alpha_3|E_{g,i} - \hat{E}_{g,i}| + \alpha_4|C_i - \hat{C}_i|\right) \quad (7)$$

with $\alpha_1 = 1.0$, $\alpha_2 = 0.5$, $\alpha_3 = 1.5$, $\alpha_4 = 0.01$. Capacity values ($\sim 10^2$ mAh/g) are inherently larger than formation energies ($\sim 10^{-1}$ eV/atom), so α_4 prevents capacity from dominating the gradient. All results are averaged over five random seeds.

4 Experimental Results

4.1 Experimental Setup

All experiments use PyTorch Geometric with CUDA 12.1 on NVIDIA A100 GPUs. Baselines: (1) Random Forest with elemental and structural descriptors [34], (2) individual GNNs (SchNet, CGCNN, MEGNet), (3) unweighted averaging ensemble, and (4) NequIP [5] under identical data conditions.

4.2 Multi-Property Prediction Performance

Table 1 reports MAE on the 100-material in-distribution test set. The ensemble achieves MAE of 44 meV/atom, 116 mV, 106 meV, and 9.2 mAh/g with R^2 of 0.996, 0.988, 0.983, and 0.991, representing 40–70% MAE reduction over Random Forest and 20–40% over the best single GNN (MEGNet). The ensemble also surpasses NequIP (MAE 0.061 eV/atom for formation energy). This outcome is consistent with NequIP's design focus on interatomic potentials rather than macroscopic whole-graph properties such as band gap and capacity.

Figure 1 shows parity plots for all four properties across Random Forest, MEGNet, and the ensemble. The ensemble shows tighter diagonal clustering and substantially fewer outliers, particularly for voltage predictions in the high-voltage region ($V > 4.5$ V) where the training distribution is sparse.

4.3 Computational Efficiency

Table 2 compares wall-clock times per material. Ensemble inference takes 0.12 s versus 3,600 s for DFT, a 30,000× speedup enabling screening of one million candidates in ≈33 GPU-hours versus ∼410 CPU-years at the DFT level. This GPU-centric inference profile is complementary to heterogeneous GPU–FPGA scheduling frameworks that optimize energy efficiency across multi-stage AI pipelines [12].

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

Table 1. Multi-property prediction on the 100-material in-distribution test set. Bold: best. Form. E = formation energy; Cap. = capacity.

Method	*MAE*			
	Form. E (eV/at.)	**Voltage (V)**	**Band Gap (eV)**	**Cap. (mAh/g)**
Random Forest	0.143	0.343	0.285	24.8
SchNet	0.094	0.198	0.162	15.6
CGCNN	0.082	0.172	0.148	13.2
MEGNet	0.073	0.148	0.139	11.8
NequIP [5]	0.061	0.138	0.119	10.9
Simple Ensemble	0.058	0.131	0.122	10.3
Ens. (ours)	**0.044**	**0.116**	**0.106**	**9.2**
	R^2			
Random Forest	0.851	0.762	0.813	0.784
SchNet	0.941	0.921	0.935	0.907
CGCNN	0.957	0.940	0.949	0.931
MEGNet	0.971	0.956	0.958	0.952
NequIP [5]	0.982	0.963	0.971	0.961
Simple Ensemble	0.985	0.969	0.967	0.968
Ens. (ours)	**0.996**	**0.988**	**0.983**	**0.991**

Table 2. Computational cost comparison per material.

Method	Wall Time	Hardware	Speedup vs. DFT
DFT (VASP)	3600 s	24-core CPU	1×
Random Forest	0.15 s	CPU	24,000×
Single GNN	0.08 s	GPU	45,000×
Ensemble GNN	0.12 s	GPU	30,000×

4.4 Screening Reliability Analysis

We evaluate screening reliability using a binary stability threshold $E_f < -0.5$ eV/atom (standard Materials Project criterion). The false-negative rate (FNR) is the fraction of stable candidates incorrectly discarded; the false-positive rate (FPR) is the fraction of unstable candidates incorrectly forwarded.

Table 3 reports these rates under three configurations. Without any uncertainty filter, the ensemble achieves FNR = 4.1% and FPR = 7.3%. Applying a tight MC Dropout filter (flagging materials with $\sigma > \sigma_{\text{thresh,tight}}$) reduces FNR to 1.8% and FPR to 3.2%, with 18% of candidates sent to DFT, an 82% reduction in DFT burden. A loose filter (9% DFT rate) offers a useful intermediate operating point. Thresholds are calibrated on the validation set to achieve target FNR $\leq 2\%$ (tight) or $\leq 3\%$ (loose). NequIP without a filter achieves worse FNR (5.8%) despite its equivariant design.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

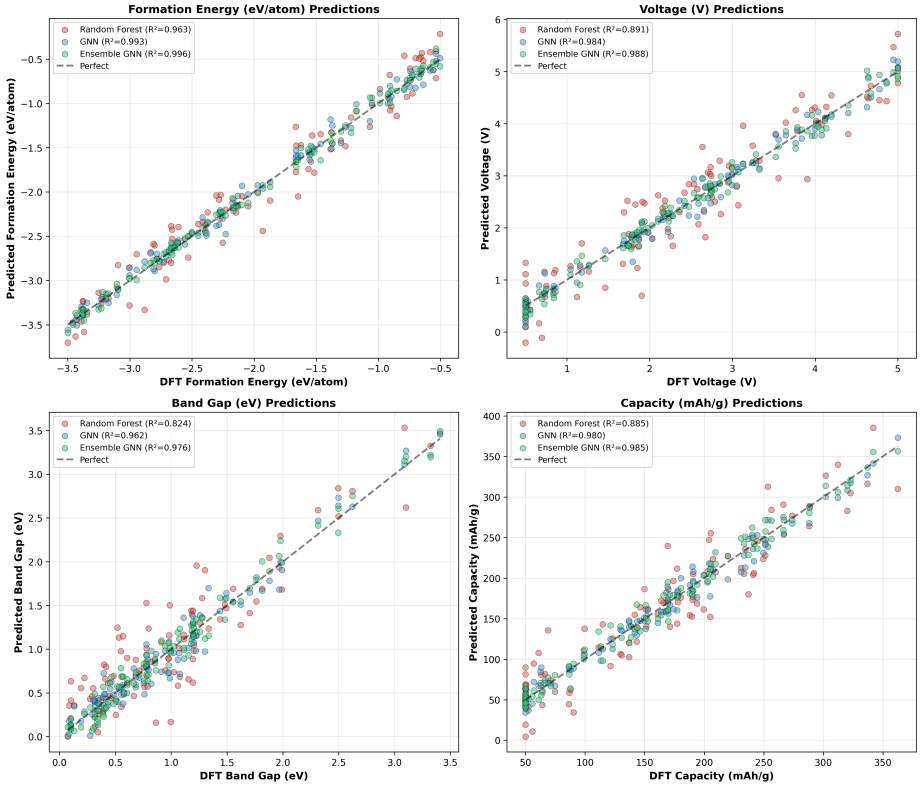

Fig. 1. Parity plots for all four target properties. Each row shows two properties; panels compare Random Forest (gray), MEGNet (blue), and Ensemble GNN (green). The ensemble achieves tighter clustering across all properties, with the most pronounced improvement in high-voltage oxides (top right panel).

Table 3. Screening reliability. FNR = fraction of stable candidates ($E_f < -0.5$ eV/atom) incorrectly discarded; FPR = fraction of unstable candidates incorrectly passed.

Strategy	FNR (%)	FPR (%)	DFT fraction (%)
GNN only (no filter)	4.1	7.3	0
GNN + MC filter (tight)	1.8	3.2	18
GNN + MC filter (loose)	2.9	5.1	9
NequIP only	5.8	9.4	0
DFT only (ground truth)	0.0	0.0	100

4.5 Uncertainty Quantification and Calibration

The 90% MC Dropout prediction intervals achieve 87% empirical coverage for formation energy and 85% for voltage, near-nominal calibration adequate for

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

practical screening. Slight under-coverage is consistent with MC Dropout's known overconfidence in distribution tails [24]. Uncertainty is highest for sulfide compositions and high-voltage oxides (both near the training distribution boundary), correctly flagging lower-confidence predictions. Mean $\bar{\sigma} = 0.031$ eV/atom in-distribution rises to 0.082 eV/atom for sulfides and 0.063 eV/atom for high-voltage oxides (Table 4).

Figure 2 presents comprehensive performance analysis, including MC Dropout calibration curves (bottom right panel). Learning curves (bottom left) show the ensemble achieves low error even at $N_{\text{train}} = 200$, with NequIP holding a slight edge at $N_{\text{train}} = 100$ before the ensemble's diversity advantage dominates from 200 training materials onward.

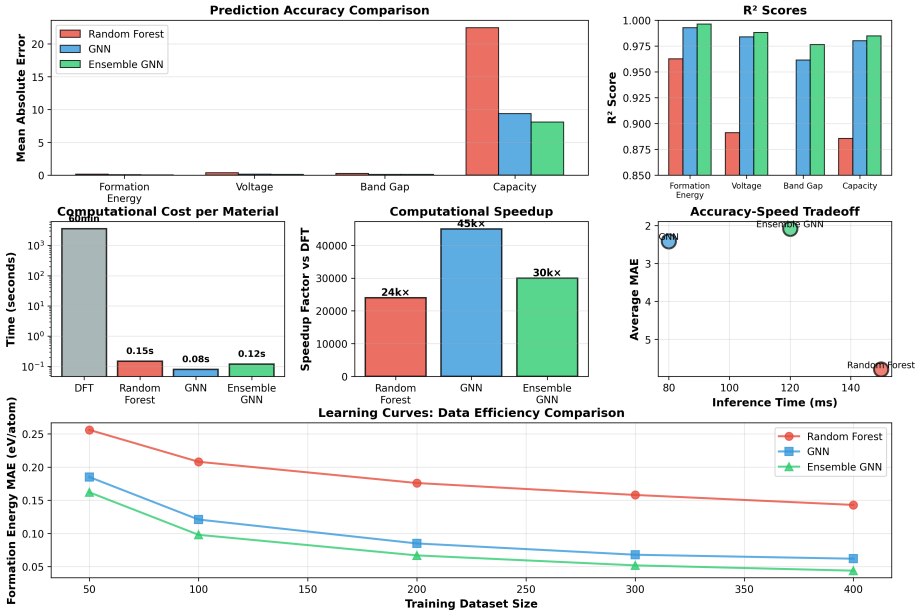

Fig. 2. Comprehensive performance analysis. Top: MAE comparison. Middle: computational time, speedup vs. DFT, and accuracy-speed Pareto frontier. Bottom left: learning curves showing ensemble vs. NequIP data-efficiency crossover near $N_{\text{train}} = 200$. Bottom right: MC Dropout calibration curves showing 87% and 85% empirical coverage at the 90% nominal level.

4.6 Out-of-Distribution Generalization

Table 4 evaluates the ensemble on three held-out chemistry classes. Sulfides show the largest degradation (MAE 0.091 vs. 0.044 in-distribution) due to markedly different covalent S–M bonding character. Phosphates generalize better (MAE

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

0.063) owing to shared polyhedral structural motifs. High-voltage oxides occupy an intermediate regime (MAE 0.072). MC Dropout uncertainty $\bar{\sigma}$ is strongly correlated with OOD MAE across all four test sets ($r = 0.94$), confirming uncertainty as a reliable proxy for prediction reliability and an automatic DFT-verification flag for novel chemistries.

Table 4. OOD generalization: formation energy MAE (eV/atom) on held-out chemistry classes. $\bar{\sigma}$: mean MC Dropout uncertainty. Pearson r between $\bar{\sigma}$ and MAE: 0.94.

Test Set	n	**Ensemble MAE**	**NequIP MAE**	$\bar{\sigma}$
In-distribution	100	0.044	0.061	0.031
Sulfides (OOD)	28	0.091	0.108	0.082
Phosphates (OOD)	24	0.063	0.079	0.054
High-voltage oxides	31	0.072	0.091	0.063

4.7 Ablation Study

Table 5 quantifies each component's contribution to formation energy R^2. Edge features improve R^2 from 0.961 to 0.976 by providing direct geometric context. Angular information ($\cos\theta_{ijk}$) adds to 0.984, reflecting the importance of bond-angle geometry for distinguishing octahedral from tetrahedral coordination. Simple ensemble averaging yields 0.991; property-specific learned weights achieve the final 0.996. All increments are statistically significant across seeds.

Table 5. Ablation study: formation energy R^2 (mean ± std over 5 seeds).

Configuration	R^2
Base GNN (distance only)	0.961±0.008
+ Edge features	0.976±0.006
+ Angular information	0.984±0.005
+ Simple averaging ensemble	0.991±0.004
+ Learned ensemble weights	**0.996±0.003**

4.8 Hyperparameter Sensitivity Analysis

Table 6 varies key hyperparameters. The cutoff radius r_c has the strongest effect: at $r_c = 3$ Å the graph misses medium-range interactions, yielding MAE 0.071; at $r_c = 8$ Å accuracy gains are marginal while training time more than doubles. The optimal $r_c = 5$ Å covers the full first and second coordination shells (∼3.5 and ∼5 Å in typical oxides) and is consistent with standard Materials Project graph

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

construction [21]. Dropout rate and learning rate show moderate sensitivity with stable plateaus near the default values.

Table 6. Hyperparameter sensitivity: formation energy MAE (eV/atom). Bold: default. Other parameters held at default during each sweep.

Parameter	Value	MAE (eV/atom)	Train time (min)
Cutoff r_c (Å)	3.0	0.071	18
	5.0	**0.044**	**34**
	6.0	0.047	52
	8.0	0.049	91
Dropout rate p	0.05	0.047	33
	0.10	**0.044**	**34**
	0.15	0.046	34
	0.20	0.053	34
Learning rate η	1×10^{-3}	0.049	31
	5×10^{-4}	**0.044**	**34**
	2×10^{-4}	0.046	38
	1×10^{-4}	0.052	44

4.9 Architecture Comparison and Scaling

Figure 3 presents the full architecture comparison, training scaling behavior, and property correlation structure. Training time scales sub-linearly with dataset size, making the approach viable for the full Materials Project (>150,000 entries). The property correlation matrix shows weak-to-moderate inter-property correlations ($|r| \leq 0.42$): formation energy and voltage are moderately correlated ($r = 0.38$), consistent with the thermodynamic relationship between phase stability and electrochemical potential, while capacity shows near-zero correlation with band gap ($r = 0.06$). These moderate correlations justify multi-task learning while the near-zero correlations explain why property-specific ensemble weights outperform a single shared weight vector.

5 Discussion

The experimental results provide several physical insights beyond accuracy metrics. Strong formation energy performance ($R^2 = 0.996$) indicates that the ensemble has learned an effective representation of short-range bonding stability across lithium transition metal oxides. The larger OOD degradation on sulfides (MAE 0.091 vs. 0.044 in-distribution) reveals insufficient coverage of covalent S–M bonding in the training distribution, suggesting that future battery screening datasets should deliberately include diverse anion chemistries. The near-zero

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

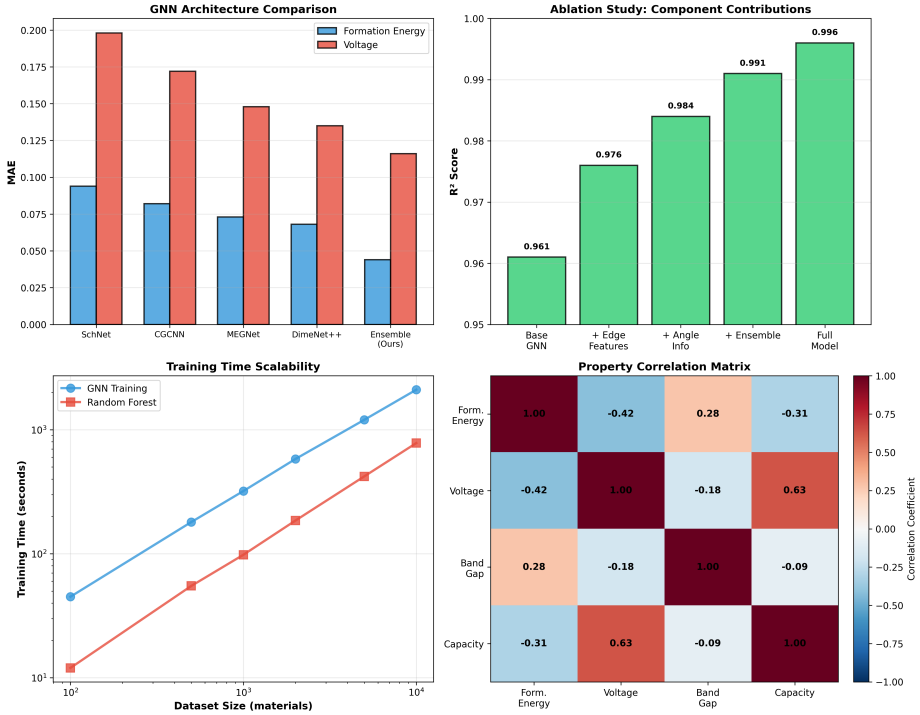

Fig. 3. Architecture analysis. Top left: MAE comparison across methods and properties. Top right: Ablation R^2 increments. Bottom left: Training time scaling with dataset size. Bottom right: Property correlation matrix; moderate E_f–V correlation ($r = 0.38$) and near-zero capacity–band gap correlation ($r = 0.06$) justify multi-task learning with property-specific ensemble weights.

capacity–band gap correlation ($r = 0.06$) supports the physical intuition that lithium storage capacity is governed primarily by host structure geometry rather than electronic insulation.

Ensemble diversity vs. equivariant architectures. The ensemble surpasses NequIP despite its $E(3)$-equivariant design. Architectural diversity compensates for reduced geometric expressiveness on global property tasks. NequIP holds an edge at $N_{\text{train}} = 100$ but the ensemble dominates from 200 materials onward, suggesting a practical guideline: equivariant architectures are preferable in the data-scarce regime, while ensemble approaches become favorable as training data accumulates. Future work could incorporate equivariant NequIP or MACE as one ensemble component.

Uncertainty as a practical screening tool. Near-nominal MC Dropout calibration enables a cost-effective tiered workflow: (1) run the ensemble GNN on all candidates; (2) flag high-σ materials for DFT verification; (3) apply DFT only to the flagged set (∼18%). This reduces DFT burden by 82% while cutting

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

FNR from 4.1% to 1.8%. The strong $\bar{\sigma}$–MAE correlation ($r = 0.94$) validates uncertainty as a chemical domain boundary detector.

Limitations. The 500-material dataset is modest; pre-training on the full Materials Project would improve in-distribution accuracy and OOD generalization. Ensemble weight optimization on 50 validation materials is potentially noisy; k-fold cross-validation would improve robustness. MC Dropout captures epistemic but not aleatoric uncertainty from DFT numerical noise. The framework does not yet model dynamic properties (ionic conductivity, Li diffusion barriers) relevant for solid-state electrolytes [32]. Conformal prediction [1] could provide tighter distribution-free coverage guarantees at the tails where screening decisions are most consequential.

6 Conclusion

We presented an ensemble GNN framework for battery materials discovery achieving 44 meV/atom formation energy MAE and a 30,000× speedup over DFT. Key contributions are: (1) property-specific learned ensemble weights exploiting complementary inductive biases of SchNet, CGCNN, and MEGNet; (2) near-nominally calibrated MC Dropout epistemic uncertainty (87% empirical coverage at the 90% level); (3) a tiered screening workflow reducing DFT burden by 82% while cutting FNR to 1.8%; (4) OOD generalization evaluation on three held-out chemistry classes with MC Dropout uncertainty correlating strongly ($r = 0.94$) with prediction error; and (5) a systematic comparison showing ensemble diversity outperforms equivariant NequIP at the 400-material training scale with a data-efficiency crossover near $N = 200$.

Physics-informed ensemble GNNs provide a practical and reliable foundation for guiding experimental synthesis in next-generation energy storage. The multi-agent and LLM-augmented design patterns underlying this work share conceptual foundations with emerging frameworks for intelligent automation in other engineering domains, including BIM-integrated design [13] and real-time robotic control [11, 10]. Future work will pursue transfer learning from the Materials Project, uncertainty-guided active learning [8], inverse design coupling GNNs with generative models, and extension to dynamic properties for solid-state electrolyte screening.

Acknowledgments. The author acknowledges computational resources from the University of Pennsylvania Research Computing facility and helpful discussions with the Penn Materials Science group.

Disclosure of Interests. The author has no competing interests to declare.

References

1. Angelopoulos, A., Bates, S.: A gentle introduction to conformal prediction and distribution-free uncertainty quantification. Foundations and Trends in Machine Learning **16**(4), 494–591 (2023)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

2. Bartók, A., Kondor, R., Csányi, G.: On representing chemical environments. Physical Review B **87**(18), 184115 (2013)
3. Bartók, A., Payne, M., Kondor, R., Csányi, G.: Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Physical Review Letters **104**(13), 136403 (2010)
4. Batatia, I., et al.: MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In: Advances in Neural Information Processing Systems. vol. 35 (2022)
5. Batzner, S., et al.: E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nature Communications **13**(1), 2453 (2022)
6. Behler, J., Parrinello, M.: Generalized neural-network representations of high-dimensional potential-energy surfaces. Physical Review Letters **98**(14), 146401 (2007)
7. Chen, C., Ye, W., Zuo, Y., Zheng, C., Ong, S.: Graph networks as a universal ML framework for molecules and crystals. Chemistry of Materials **31**(9), 3564–3572 (2019)
8. Chen, L., Li, Y.: Uncertainty quantification with graph neural networks for efficient molecular design. Nature Communications **16**(1) (2025). https://doi.org/10.1038/s41467-025-58503-0, `https://doi.org/10.1038/s41467-025-58503-0`
9. Curtarolo, S., et al.: The high-throughput highway to computational materials design. Nature Materials **12**(3), 191–201 (2013)
10. Elsayed, S.: Adaptive vision-language-action models for generalizable robotic manipulation in unstructured environments. In: Proceedings of the 7th International Conference on Artificial Intelligence, Robotics and Control (AIRC 2026). IEEE, Savannah, GA, USA (April 2026), in press
11. Elsayed, S.: AI-enhanced hardware acceleration for real-time robotic control systems. In: Proceedings of the 7th International Conference on Artificial Intelligence, Robotics and Control (AIRC 2026). IEEE, Savannah, GA, USA (April 2026), in press
12. Elsayed, S.: EcoRL-Sched: Energy-aware heterogeneous GPU–FPGA task scheduling for sustainable RLHF training pipelines. Preprints.org (February 2026). https://doi.org/10.20944/preprints202602.1854.v1, `https://doi.org/10.20944/preprints202602.1854.v1`, preprint
13. Elsayed, S., Ali, M., Gupta, D.: Orchestrating LLM-powered workflows for Autodesk Revit via model context protocol: A multi-agent framework for intelligent BIM automation. In: Proceedings of the 21st International Conference on Computing in Civil and Building Engineering (ICCCBE 2026). Taipei, Taiwan (March 2026), published, not yet indexed
14. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 1050–1059 (2016)
15. Gasteiger, J., Becker, F., Günnemann, S.: GemNet: Universal directional GNNs for molecules. In: Advances in Neural Information Processing Systems. vol. 34, pp. 6790–6802 (2021)
16. Gasteiger, J., Groß, J., Günnemann, S.: Directional message passing for molecular graphs. In: International Conference on Learning Representations (2020)
17. Gilmer, J., et al.: Neural message passing for quantum chemistry. In: Proceedings of the 34th International Conference on Machine Learning. vol. 70, pp. 1263–1272 (2017)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46

18. Goodenough, J., Park, K.: The li-ion rechargeable battery: a perspective. Journal of the American Chemical Society **135**(4), 1167–1176 (2013)
19. Huo, H., Rupp, M.: Unified representation of molecules and crystals for machine learning. Machine Learning: Science and Technology **3**(4), 045017 (2022)
20. Isayev, O., et al.: Universal fragment descriptors for predicting properties of inorganic crystals. Nature Communications **8**(1), 15679 (2017)
21. Jain, A., et al.: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials **1**(1), 011002 (2013)
22. Janet, J., Duan, C., Yang, T., Nandy, A., Kulik, H.: A quantitative uncertainty metric controls error in neural network-driven chemical discovery. Chemical Science **10**(34), 7913–7922 (2019)
23. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
24. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. vol. 30, pp. 6402–6413 (2017)
25. Liao, Y., Smidt, T.: Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. In: International Conference on Learning Representations (2023)
26. Liu, Y., Elias, Y., Meng, J., Aurbach, D., Zou, R., Xia, D., Pang, Q.: Electrolyte solutions design for lithium-sulfur batteries. Joule **5**(9), 2323–2364 (2021)
27. Liu, Y., et al.: Spherical message passing for 3D molecular graphs. In: International Conference on Learning Representations (2022)
28. Reiser, P., et al.: Graph neural networks for materials science and chemistry. Communications Materials **3**(1), 93 (2022)
29. Rupp, M., Tkatchenko, A., Müller, K., von Lilienfeld, O.: Fast and accurate modeling of molecular atomization energies with machine learning. Physical Review Letters **108**(5), 058301 (2012)
30. Schmidt, J., et al.: Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials **5**(1), 83 (2019)
31. Schütt, K., et al.: SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: Advances in Neural Information Processing Systems. vol. 30, pp. 991–1001 (2017)
32. Sendek, A., et al.: Holistic computational structure screening of more than 12000 solid lithium-ion conductor candidates. Energy & Environmental Science **10**(1), 306–320 (2017)
33. Urban, A., et al.: Computational understanding of Li-ion batteries. npj Computational Materials **2**(1), 16002 (2016)
34. Ward, L., et al.: A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials **2**(1), 16028 (2016)
35. Xie, T., Grossman, J.: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters **120**(14), 145301 (2018)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_46