Data-Efficient Multi-Property Prediction of Crystals Using Active Learning and Graph Neural Networks

Group 14
Chamakura Sai Kumar (ED25D402)
Department of Engineering Design
Indian Institute of Technology Madras (IITM)
Chennai, India
ed25d402@smail.iitm.ac.in

***Abstract*—The discovery of novel materials is often hampered by the high computational cost of simulations like Density Functional Theory (DFT), creating a data scarcity problem for machine learning models. This paper presents a comprehensive benchmarking framework to evaluate data-efficient learning strategies for predicting multiple material properties simultaneously. We focus on Active Learning (AL) and compare two prominent uncertainty quantification techniques—Monte Carlo (MC) Dropout and Deep Ensembles—against standard Random Sampling and Traditional Learning baselines. These strategies are applied to four different Graph Neural Network (GNN) architectures (a custom GCN, SchNet, CGCNN, and MEGNet) for the joint prediction of Formation Energy and Band Gap using a curated dataset of 6,996 inorganic crystals derived from the Materials Project. Our experiments demonstrate that AL strategies consistently outperform random sampling, achieving lower prediction errors with significantly less training data. Notably, the combination of a custom GNN model with a Deep Ensemble AL strategy proved to be the most data-efficient. Crucially, we show that this application-driven data curation is vital for model stability and performance, as parallel experiments on the unfiltered, heterogeneous dataset of 50,000 inorganic crystals led to catastrophic failures for several architectures. Analysis of the queried materials reveals that AL methods effectively explore the feature space by selecting more diverse and informative crystal structures from the tails of the property distributions. This work provides a systematic evaluation of AL techniques, offering valuable insights and a practical framework for accelerating materials discovery under data constraints.**

***Index Terms*—Active Learning, Graph Neural Networks, Materials Discovery, Data-Efficient Learning, Uncertainty Quantification, Multi-Property Prediction**

I. Introduction

The quest for novel materials with tailored properties is a cornerstone of technological advancement, driving innovation in fields ranging from energy storage to electronics [1]. This pursuit is central to initiatives like the Materials Genome Initiative, which aims to halve the time and cost of discovering, developing, and deploying new materials [2]. Traditionally, this process has relied on a combination of intuition-driven experimentation and computationally expensive first-principles simulations, such as Density Functional Theory (DFT). While powerful, these methods are resource-intensive, creating a significant bottleneck in the high-throughput screening of the vast chemical space of potential materials [3]. Furthermore, large public datasets are often highly heterogeneous, containing diverse material classes (e.g., metals, semiconductors, insulators) whose properties arise from different physical principles. Training a single model on such a broad and imbalanced dataset, especially in a data-scarce regime, presents a significant challenge to generalization and performance.

In recent years, machine learning (ML), particularly Graph Neural Networks (GNNs), has emerged as a promising paradigm for accelerating materials discovery. GNNs are well-suited for this domain as they can directly learn from the atomic structure of a crystal—represented as a graph where atoms are nodes and bonds are edges—to predict its properties, bypassing the need for handcrafted features or descriptors [4], [5]. However, the performance of these deep learning models is highly dependent on the availability of large, high-quality labeled datasets. This ”data scarcity” is a fundamental challenge in materials science, as each data point may require hours or days of supercomputer time to generate via DFT.

This challenge motivates the exploration of data-efficient learning paradigms. Active Learning (AL) is one such paradigm that aims to maximize model performance with a minimal number of labeled training samples. The core idea of AL is to iteratively and intelligently query the most informative data points from a large pool of unlabeled candidates, thereby guiding the model to learn more effectively and reduce the number of expensive DFT calculations required [6]. The selection of these points is typically driven by an uncertainty metric, where the model identifies samples for which its predictions are least certain [7], [8].

While AL has shown great promise in various scientific domains, its application and systematic evaluation in multi-property materials prediction remain an area of active research. Different GNN architectures and various uncertainty quantification (UQ) techniques exist, but a comprehensive comparison of their effectiveness within an AL framework for materials science is lacking. Furthermore, many real-world applications require the simultaneous optimization of multiple properties (e.g., low formation energy for stability and a specific band gap for optoelectronic applications).

This paper addresses this gap by presenting a comprehensive benchmarking framework to evaluate AL strategies for the simultaneous prediction of Formation Energy and Band Gap. We compare two popular UQ methods: Monte Carlo (MC)

Dropout [9] and Deep Ensembles [10]. Their performance is benchmarked against two baselines: a standard Random Sampling approach and a Traditional Learning setup. This evaluation is conducted across four distinct GNN architectures: a custom Graph Convolutional Network (GNN), SchNet [11], Crystal Graph Convolutional Neural Network (CGCNN) [4], and MEGNet [12]. Our primary contributions are:

- A systematic comparison of AL strategies (MC Dropout, Deep Ensemble) against baselines for multi-property prediction of crystals.
- A head-to-head evaluation of the performance and data efficiency of four different GNN architectures within this AL framework.
- A detailed analysis of the properties of materials selected by AL, demonstrating its ability to explore the design space more effectively than random selection.
- A reproducible workflow, from data acquisition and filtering to model training and evaluation, serving as a practical guide for researchers.

Our findings provide practical guidelines for researchers aiming to develop accurate and data-efficient models for accelerating the discovery of new materials under realistic data constraints.

II. Related Work

The intersection of materials science, deep learning, and data-efficient methods has become a vibrant area of research. Our work is situated within three key domains: foundational GNN architectures for materials representation, strategies for data-efficient learning, and the application of active learning to accelerate scientific discovery.

A. Foundational GNNs for Materials Science

The representation of crystals as graphs has paved the way for GNNs to become a primary tool in materials informatics. Foundational architectures form the basis of our study, including CGCNN, which learned universal representations from atomic connectivity [4]; SchNet, which introduced continuous-filter convolutions for quantum-chemical interactions [11]; and MEGNet, which incorporated global state attributes for state-of-the-art performance [12]. Our work benchmarks these diverse architectures within an active learning framework to understand their data efficiency and robustness in a low-data regime.

B. Strategies for Data-Efficient Learning

Given the high cost of data generation in materials science, methods that maximize model performance with limited data are critical. One approach is Multi-Task Learning (MTL), which improves data efficiency by learning multiple properties simultaneously, as demonstrated by Sanyal et al. with MT-CGCNN [13]. Our work adopts a multi-property goal but focuses on data acquisition. A second approach is Active Learning (AL), which intelligently selects the most informative data points to label next, guiding expensive DFT simulations to where they are most needed [7], [8]. The effectiveness of different AL uncertainty quantification (UQ) methods can vary significantly, motivating a broad benchmark [6]. Our paper builds on this body of work by providing a broad, head-to-head benchmark of two of the most popular UQ methods for AL—MC Dropout [9] and Deep Ensembles [10]—across multiple foundational GNN architectures for a multi-property prediction task in crystalline solids.

III. Methodology

Our methodology is designed to systematically evaluate the data efficiency of different learning strategies through a reproducible pipeline. It encompasses dataset preparation, model architecture definition, the implementation of learning frameworks, and the evaluation protocol. A high-level overview of the experimental workflow and the specific learning strategy loops are depicted in Fig. 6.

A. Dataset Acquisition and Preprocessing

The dataset was programmatically sourced from the Materials Project database using its official API [14]. An initial query downloaded data for 50,000 inorganic crystal structures, including their atomic coordinates, formation energy per atom (FE), and band gap (BG). The initial distributions of these properties, shown in Fig. 1, reveal a wide range of values, numerous outliers, and a strong skew in the band gap data, motivating the need for a targeted data curation strategy.In addition to constructing a filtered, application-specific dataset, the original unfiltered dataset was also preserved to enable a controlled comparison of model performance under both data regimes.

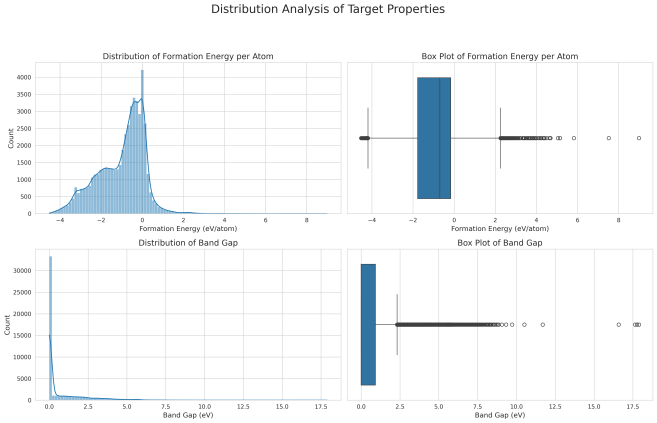


Fig. 1. Distribution analysis of the two target properties in the initial, unfiltered dataset of materials with complete entries. TOP: Formation energy distribution. BOTTOM: Band gap distribution.

a) Application-Driven Data Curation: A crucial step in our methodology was the curation of the initial dataset through a physically motivated, application-driven filtering strategy. Rather than performing simple statistical outlier removal, our goal was to define a specific, high-value problem space: **the discovery of novel, synthesizable semiconductors for energy and optoelectronic applications.** This focus is essential

for training a specialized and effective model. As shown in Fig. 2, we applied a 2D filter to retain materials that satisfy:

- Formation Energy: $-3.5 \leq \text{FE} \leq 0.5$ eV/atom
- Band Gap: $0.01 \leq \text{BG} \leq 1.2$ eV

The scientific justification for these specific bounds is detailed below:

b) Justification for the Band Gap (BG) Filter: This filter isolates technologically significant non-metallic materials. A lower bound (BG $\geq$ 0.01 eV) excludes metals, whose physics differ from semiconductors. An upper bound (BG $\leq$ 1.2 eV) focuses the model on the relevant range for applications like photovoltaics, avoiding wide-bandgap insulators.

c) Justification for the Formation Energy (FE) Filter: This filter ensures the model trains on thermodynamically plausible materials. An upper bound (FE $\leq$ 0.5 eV/atom) excludes computationally predicted but physically unrealistic structures. A lower bound (FE $\geq$ -3.5 eV/atom) focuses the model on more chemically complex regions where novel discovery is more likely, beyond simple, well-known ionic compounds.

In summary, this data curation process transforms a vast, heterogeneous dataset into a focused, high-quality collection of 6,996 materials. Importantly, throughout the study we retained both the filtered dataset and the original unfiltered dataset (50,000 materials) so that parallel experiments could be performed.

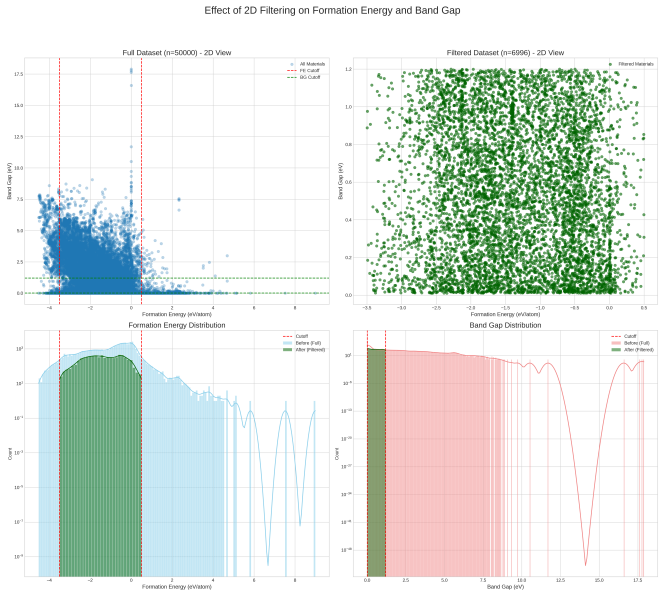


Fig. 2. Effect of 2D filtering on the dataset. The top-left plot shows the full dataset with red/green lines indicating the cutoffs. The top-right plot shows the filtered dataset (n=6996). The bottom plots show the 1D distributions before and after filtering on a log scale.

d) Dataset Splitting and Normalization: The curated dataset of 6,996 graphs was split into a training pool of 3,498 materials, a validation set of 1,749, and a fixed, held-out test set of 1,749. The validation set was used for hyperparameter tuning. The distributions of the target properties for both the filtered and unfiltered datasets, shown in Fig. 3, are consistent across the splits, ensuring an unbiased evaluation. The two target properties (FE and BG) were normalized using a 'StandardScaler' to have zero mean and unit variance.

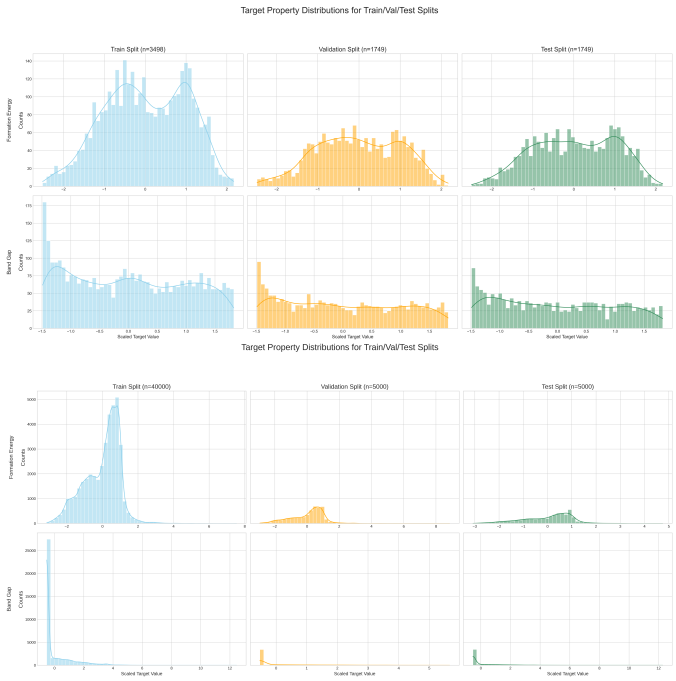


Fig. 3. Comparison of the scaled target property distributions across the Train, Validation, and Test splits for the filtered dataset (top) and the original unfiltered dataset (bottom). The similarity in the distributions confirms that both splitting procedures produce consistent, unbiased partitions suitable for model training and evaluation.

B. Graph Representation of Crystals

Each crystal structure in the curated set was converted into a graph representation compatible with PyTorch Geometric [15], as illustrated in Fig. 4. In this framework, atoms are treated as nodes in the graph, and their primary feature is their atomic number, which is passed through an embedding layer in the GNN. The relationships between atoms are represented as edges. We define an edge between any two atoms within a cutoff radius of 5.0 Å. The continuous interatomic distance is used as the primary edge attribute, providing crucial geometric information to the model. This graph-based representation allows the GNN to learn structure-property relationships directly from the atomic topology and chemistry, without requiring manual feature engineering.

C. Model Architectures

To ensure our findings are generalizable, we benchmarked four different GNN architectures, each adapted for multi-target regression. Architectural details and diagrams are provided in Fig. 5 and Table I.

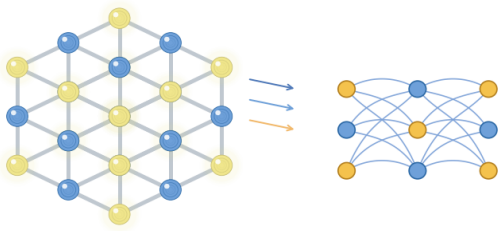


Fig. 4. Conceptual overview of the crystal-to-graph conversion process. Left: A 3D crystal lattice. Middle: Atoms are mapped to numerical feature vectors encoding their properties. Right: The atoms become nodes and their neighborhood relationships become edges in a graph, which serves as the input to the GNN.

a) GNN (Custom): Our baseline GNN is a simple Graph Convolutional Network (GCN) built with standard 'GCNConv' layers from PyTorch Geometric. It consists of an embedding layer, three GCN layers with ReLU activations, and a global mean pooling layer. This is passed through two fully-connected layers with dropout for regularization to produce the final 2D output.

b) SchNet: SchNet is a physics-inspired architecture designed to model quantum-chemical interactions [11]. It employs continuous-filter convolutional (cfconv) layers that operate on interatomic distances, expanded using a radial basis function (RBF) to learn complex, continuous interaction functions.

c) CGCNN: The Crystal Graph Convolutional Neural Network (CGCNN) is specifically designed for crystalline materials [4]. Its core is a gated convolutional layer that concatenates feature vectors of a central atom, a neighboring atom, and their connecting bond, allowing the model to learn the importance of different bond interactions dynamically.

d) MEGNet (GAT-based): Our implementation of a MEGNet-like architecture utilizes Graph Attention (GATv2) layers [12], which introduce a self-attention mechanism. This allows the model to assign different weights (attentions) to neighboring nodes, making it more flexible than fixed-weighting schemes.

D. Training Workflow and Learning Frameworks

The core of our investigation is the comparison of four learning strategies. For all iterative strategies, we start with an initial labeled set of 200 samples and add 100 new samples for 10 cycles, growing the training set to 1,200 samples. The workflows are visualized in Fig. 6 and detailed in Algorithms 1 and 2.

a) Active Learning (MC Dropout): The model is trained on the current labeled set. Uncertainty for each sample in the unlabeled pool is estimated by performing 20 stochastic forward passes with dropout enabled. The variances of the predictions for FE and BG are summed to create a single uncertainty score. The 100 samples with the highest scores are selected (Algorithm 2).

Algorithm 1 Iterative Learning Loop (Active and Random)

Require: Labeled set $\mathcal{L}$, Unlabeled pool $\mathcal{U}$, Query size K, Number of cycles N_{cycles}, Epochs per cycle E_{cycle}, Query Strategy Q

Initialize history $\mathcal{H} \leftarrow []$
for $c = 0$ to N_{cycles} **do**
  Train model(s) $\mathcal{M}$ on $\mathcal{L}$ for E_{cycle} epochs
  Evaluate $\mathcal{M}$ on test set $\mathcal{D}_{test}$ to get metrics $\mathcal{M}_c$
  Append $(\text{size}(\mathcal{L}), \mathcal{M}_c)$ to $\mathcal{H}$
  if $c < N_{cycles}$ **then**
    $\mathcal{S} \leftarrow \text{QueryBatch}(\mathcal{M}, \mathcal{U}, K, Q)$
    $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{S}$ ▷ Move queried samples
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{S}$
  end if
end for
return $\mathcal{H}$

b) Active Learning (Deep Ensemble): An ensemble of 5 identical but independently initialized models is trained on the current labeled set. For each unlabeled sample, the variance across the ensemble's predictions is calculated for FE and BG and then summed. The 100 most uncertain samples are selected (Algorithm 2).

c) Random Sampling (Baseline): In each cycle, 100 samples are chosen uniformly at random from the unlabeled pool and added to the training set.

d) Traditional Learning (Baseline): A single model is trained on a fixed dataset of 1,200 samples, which is sampled once from the training pool.

Algorithm 2 QueryBatch Function

function QUERYBATCH($\mathcal{M}, \mathcal{U}, K$, Strategy)
  if Strategy is Random **then**
    return Randomly select K samples from $\mathcal{U}$
  else if Strategy is MC Dropout **then**
    For each sample $x \in \mathcal{U}$, compute predictions $\{\hat{y}_t(x)\}_{t=1}^{T}$ by T stochastic forward passes of $\mathcal{M}$.
    Compute uncertainty $u(x) = \text{Var}(\{\hat{y}_t(x)\})$
    return Top K samples from $\mathcal{U}$ with highest $u(x)$
  else if Strategy is Ensemble **then**
    Let $\mathcal{M}_{ens} = \{\mathcal{M}_1, ..., \mathcal{M}_M\}$ be the ensemble.
    For each sample $x \in \mathcal{U}$, compute predictions $\{\hat{y}_m(x)\}_{m=1}^{M}$ from each model in $\mathcal{M}_{ens}$.
    Compute uncertainty $u(x) = \text{Var}(\{\hat{y}_m(x)\})$
    return Top K samples from $\mathcal{U}$ with highest $u(x)$
  end if
end function

E. Evaluation Metrics

Model performance was assessed using three standard regression metrics for N test samples with true values y_i and predicted values $\hat{y}_i$: Mean Absolute Error (MAE), which measures the average magnitude of errors; Root Mean Squared Error (RMSE), which penalizes larger errors more heavily;

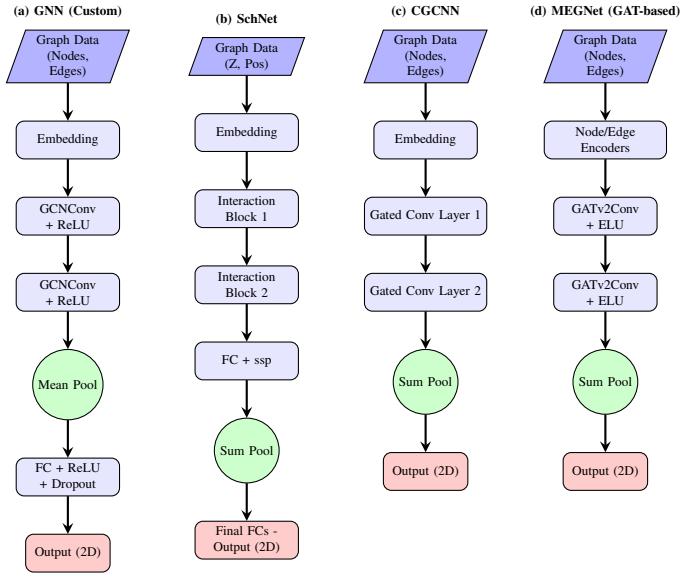


Fig. 5. Simplified architectural diagrams for the four GNN models used in this study. Each is adapted to output a 2D vector for multi-property prediction.

TABLE I
DETAILED COMPARISON OF GNN ARCHITECTURAL DIFFERENCES

Aspect	GNN (Custom GCN)	SchNet	CGCNN	MEGNet (GAT-based)
Core Convolution	GCNConv Layer	Continuous-Filter Conv (cfconv)	Custom Gated Convolution	Graph Attention Conv (GATv2)
Message Passing	Normalized sum of neighbor features.	Filter-generating network based on distances.	Concatenation of node and edge features.	Weighted sum of neighbors based on learned attention scores.
Use of Edge Attributes	Only defines graph connectivity. Not used in update rule.	Distances passed through radial basis function (RBF) expansion.	Distances concatenated with node features before convolution.	Edge features used to compute attention coefficients.
Activation Function	ReLU	Shifted Softplus (ssp)	Sigmoid (gating) and Softplus	Leaky ReLU or ELU
Graph Pooling	Global Mean Pooling	Global Sum Pooling	Global Sum Pooling	Global Sum Pooling
Final Prediction Head	Two fully-connected layers with Dropout.	Two fully-connected layers.	One fully-connected layer.	Two fully-connected layers.
Design Philosophy	Simplicity and scalability. A strong baseline for graph-level tasks.	Physics-inspired, focuses on modeling local atomic environments with continuous functions.	Specifically designed for crystals with an interpretable gating mechanism to weight bonds.	Highly flexible; learns the importance of specific neighbors via self-attention for complex interactions.

and the Coefficient of Determination (R^2), which represents the proportion of variance explained by the model, where 1 indicates a perfect prediction.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{1}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

where $\bar{y}$ is the mean of the true values, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

IV. RESULTS AND DISCUSSION

A. Learning Curve Analysis: Data Efficiency on Curated and Unfiltered Data

The learning curves in Fig. 7 provide the primary evidence for the data efficiency of the tested strategies across both the curated and unfiltered datasets. A strategy is considered more data-efficient if it achieves a lower prediction error for a given amount of training data.

On the **curated dataset**, the AL strategies (MC Dropout and Deep Ensemble) consistently outperform Random Sampling across both target properties. Their error curves show a noticeably steeper descent, indicating that the models learn more rapidly from the actively selected samples. The Deep Ensemble strategy in particular demonstrates a strong advantage in the low-data regime (200–500 samples), suggesting that ensemble-based uncertainty estimates provide more stable

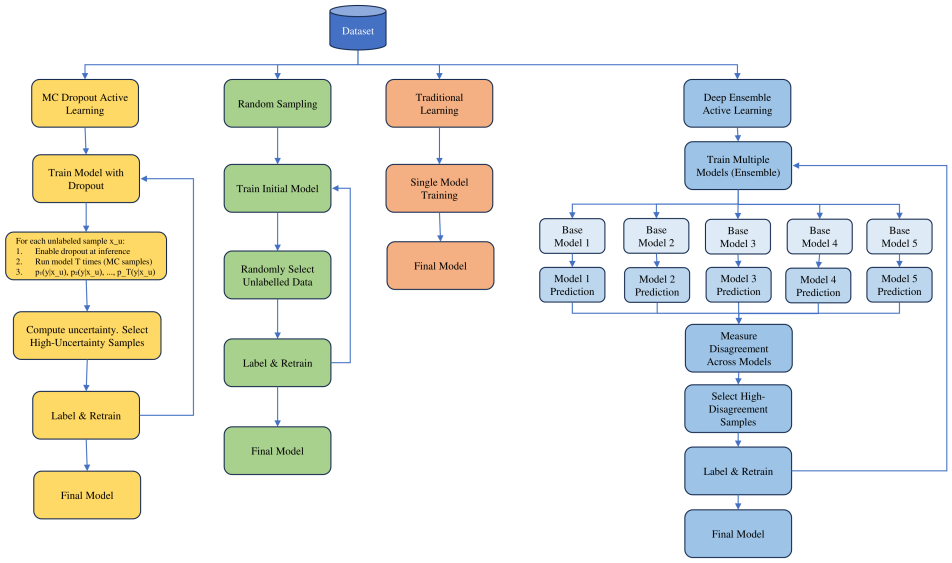


Fig. 6. Diagram of the four learning strategies benchmarked. (a) Active learning with MC Dropout iteratively adds samples where a single model with dropout is most uncertain. (b) Random sampling iteratively adds randomly selected samples. (c) Traditional learning uses a fixed, large dataset. (d) Active learning with Deep Ensembles iteratively adds samples where an ensemble of models disagrees the most.

acquisition signals. In several cases, the AL strategies reach the final performance level of the Traditional Learning baseline using only 600–800 samples, corresponding to a 33–50% reduction in required data.

For the **unfiltered dataset**, the general trends remain consistent, although the scale and heterogeneity of the larger dataset naturally result in higher overall error levels and slower convergence. Nevertheless, the relative ordering of strategies is preserved: AL continues to outperform Random Sampling and maintains its data-efficiency advantage. The larger chemical diversity and increased noise in the unfiltered data amplify the benefits of informed acquisition, highlighting the robustness of AL even when the problem space is less controlled.

Together, these results demonstrate that active learning improves data efficiency in both problem settings. However, its impact is more pronounced on the curated dataset, where the physically motivated filtering yields a cleaner, more coherent learning domain, enabling AL to extract maximal value from each acquired sample.

B. Final Model Performance Evaluation

Table II summarizes the final test set performance across all models and strategies for both the unfiltered and filtered datasets. Across nearly all metrics, the custom GNN architecture remains the strongest and most robust model.

On the **filtered dataset**, the best overall performance for Formation Energy is achieved by the **GNN + Active (Ensemble)** strategy, obtaining an MAE of **0.2675** eV/atom, RMSE of **0.3933** eV/atom, and an R^2 score of **0.8795**. This represents the lowest error and highest explained variance among all filtered FE models.

On the **unfiltered dataset**, the **GNN + Active (Ensemble)** strategy again provides the best performance overall, exhibiting strong generalization despite the increased noise and chemical diversity. It achieves an FE MAE of **0.1581**, FE RMSE of **0.3413**, and FE R^2 of **0.9084**. For Band Gap, GNN also achieves the best unfiltered results, with BG MAE of **0.3366**, BG RMSE of **0.6560**, and BG R^2 of **0.8192**.

A notable finding across all experiments is the significant disparity in predictive performance between the two target properties. While the models, particularly the custom GNN, achieved high accuracy for Formation Energy (R^2 up to 0.8795 on the curated dataset), they consistently struggled to learn the Band Gap, as evidenced by near-zero or even negative R^2 scores.

1) Catastrophic Failures in SchNet, CGCNN, and MEGNet: A striking observation from Table II is the catastrophic failure of several complex architectures on the unfiltered dataset.

SchNet collapses completely, producing FE MAE values as large as 10^{11}–10^{13} and astronomically negative R^2 scores. This confirms that SchNet's physics-inspired inductive bias does not transfer to large, compositionally diverse crystalline datasets with limited supervision. Notably, SchNet performs reasonably on the filtered dataset (FE MAE ≈ 0.60), showing the failure is triggered by dataset heterogeneity.

CGCNN also diverges severely on the unfiltered dataset, with FE MAE values up to 10^6. Even on the filtered dataset, CGCNN performs poorly (FE MAE ≈ 0.92–1.39), suggesting its gating mechanisms require significantly larger datasets to stabilize.

MEGNet does not collapse catastrophically but consistently underperforms the GNN. It delivers the best filtered BG MAE but with only marginal improvement and still near-zero R^2.

2) Key Insight: Inductive Bias Matters More Than Architectural Complexity: A central takeaway from this benchmark is that *simpler, well-aligned inductive biases outperform more*

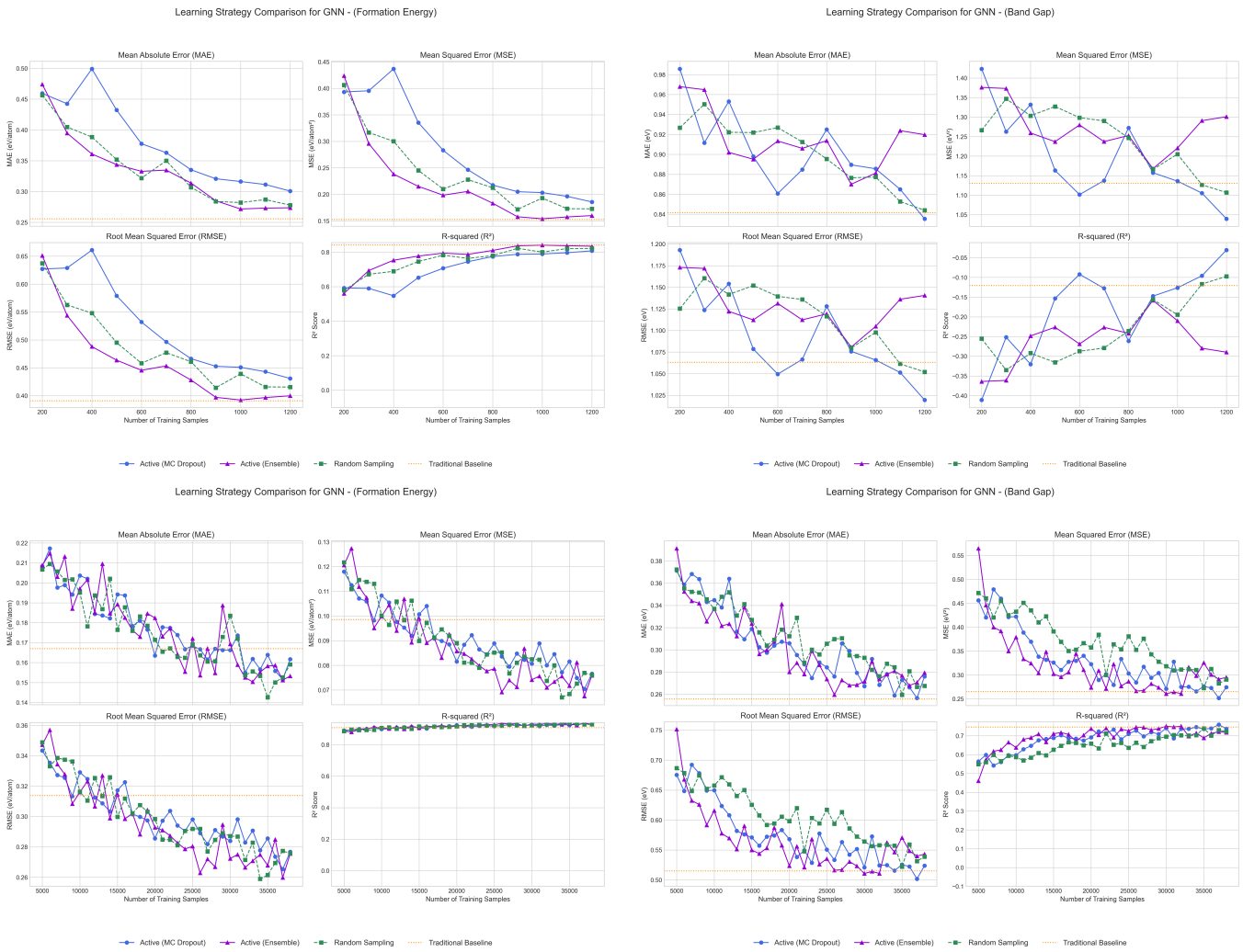


Fig. 7. Comparison of learning strategies for the GNN model on both datasets. Top row: filtered dataset (FE, BG). Bottom row: unfiltered dataset (FE, BG).

complex architectures in data-scarce materials discovery scenarios. Despite being less complex than SchNet, CGCNN, and MEGNet, the custom GNN consistently achieves the best performance across both datasets, both target properties, and all acquisition strategies. This demonstrates that advanced architectures are not automatically superior. Their success depends critically on the match between architectural design and task characteristics. In practical, low-data materials discovery settings, a simpler and more robust GNN model can substantially outperform specialized, high-capacity architectures.

C. The Critical Impact of Data Curation: Filtered vs. Unfiltered

To quantify the importance of our application-driven data curation, we present a direct comparison of model performance on the curated (filtered) dataset versus the full, unfiltered dataset (Table II). The findings are striking and provide conclusive evidence for our curation approach.

The most immediate observation is the catastrophic failure of the SchNet and CGCNN architectures on the unfiltered data. For instance, the SchNet model under traditional learning yields a physically meaningless Formation Energy MAE of 1.62×10^{13} eV/atom. This is not an outlier but a systemic failure to learn a generalizable representation from the highly heterogeneous and imbalanced raw dataset.

This failure arises from two primary issues. First, the unfiltered dataset is heavily skewed toward metals (zero or near-zero band gap), as shown in Fig. 1. Forcing a model to learn the distinct physics of metals, semiconductors, and insulators simultaneously from a limited training set leads to an ill-posed learning problem and instability. Second, the architectural inductive biases of models such as SchNet—designed for smooth potential energy surfaces—are fundamentally mismatched with the disjointed, multi-modal property space of the unfiltered data.

In contrast, our simpler custom GNN demonstrates greater robustness, although its performance is still significantly degraded on the unfiltered dataset compared to its results on the focused, curated set. This highlights that even for a more

robust architecture, focusing the learning task on a well-defined, physically coherent subset of materials is essential for achieving high accuracy and reliability.

D. Analysis of Queried Materials

Fig. 8 provides a direct comparison of the queried materials from the filtered and unfiltered datasets, illustrating how the underlying data quality fundamentally shapes the behavior of active learning. In the curated, filtered dataset (Fig. 8a), active learning consistently selects materials spanning broad, diverse, and often extreme regions of the property space. This wide exploratory behavior reflects the well-balanced nature of the curated dataset, where meaningful diversity exists for AL to exploit. In contrast, the unfiltered dataset (Fig. 8b) exhibits far narrower and more skewed queried-property distributions. Here, the dominance of metallic and structurally repetitive compounds constrains exploration. This comparison demonstrates that while AL inherently seeks diversity, its effectiveness is tightly coupled to dataset quality: curated data enables rich exploration, whereas unfiltered data limits AL's ability to discover new regions of the chemical space.

E. Case Study: Prediction on a New Material

To demonstrate the difference in predictive stability, we compare model predictions on two unseen materials, each representative of its respective dataset's chemical space, in Table III. The models trained on the filtered dataset were evaluated on mp-726207, a semiconductor within the curated property range. Conversely, the models trained on the unfiltered dataset were evaluated on mp-2767524. Using distinct materials is necessary as the models are specialized to their training domains. This case study illustrates the profound difference in model stability that stems from the quality of the training data. The filtered dataset yields stable, reasonable predictions across all models, whereas the unfiltered dataset produces extreme instabilities—especially in SchNet and sometimes CGCNN—demonstrating how noisy data severely disrupts model reliability.

F. Parity Plot Visualization

The parity plots in Fig. 9 visually confirm the quantitative results. For the GNN model on the curated dataset, the plots show strong correlation between true and predicted values, with points tightly clustered around the ideal y=x line. The poor performance of SchNet and CGCNN is evident from highly scattered point clouds. When trained on unfiltered data, their plots (not shown) degrade into completely uncorrelated clouds, visually confirming their catastrophic failure.

V. Conclusion and Future Work

This work presented a comprehensive benchmark of active learning strategies for the data-efficient, multi-property prediction of inorganic crystals using Graph Neural Networks. Our results robustly demonstrate the superiority of active learning over passive random sampling. By intelligently selecting the most uncertain data points, AL can train more accurate models with significantly less data, directly addressing the data scarcity bottleneck in computational materials science.

Our key findings are:

1) **Active Learning is Effective:** Both MC Dropout and Deep Ensemble AL strategies consistently outperform Random Sampling, leading to faster convergence. Deep Ensembles generally provided the highest data efficiency.
2) **Architecture Matters:** A relatively simple custom GNN model proved more effective and robust for this task than more complex architectures like SchNet and CGCNN, which showed instability.
3) **AL Explores the Design Space:** Analysis of the queried materials confirmed that AL succeeds by selecting diverse and structurally complex materials from the edges of the data distribution, maximizing the information gained from each expensive label.
4) **Application-Driven Data Curation is Paramount:** Our comparative analysis provides definitive evidence that defining a focused, physically coherent problem space is critical for model success. Models trained on the vast, unfiltered dataset exhibited instability and catastrophic failure, proving that thoughtful data curation is a prerequisite for developing reliable predictive models in materials science.

Future work could explore more sophisticated query strategies that balance uncertainty with diversity. Additionally, extending this framework to a multi-fidelity setting, where AL can choose to query cheap, low-accuracy calculations or expensive, high-accuracy DFT simulations, would be a promising direction for optimizing the cost-accuracy trade-off in real-world materials discovery campaigns.

Author Contributions

This work was completed as a solo term project. The sole author, Chamakura Sai Kumar, was responsible for all aspects of the research, including the initial project conception, literature review, data acquisition and curation from the Materials Project, and the design of the experimental framework. The author implemented all four Graph Neural Network architectures, developed the active learning and baseline training loops, conducted the computational experiments, and performed the subsequent data analysis. The writing of the manuscript, creation of all figures and tables, and final preparation of the paper were also carried out entirely by the author.

Acknowledgment

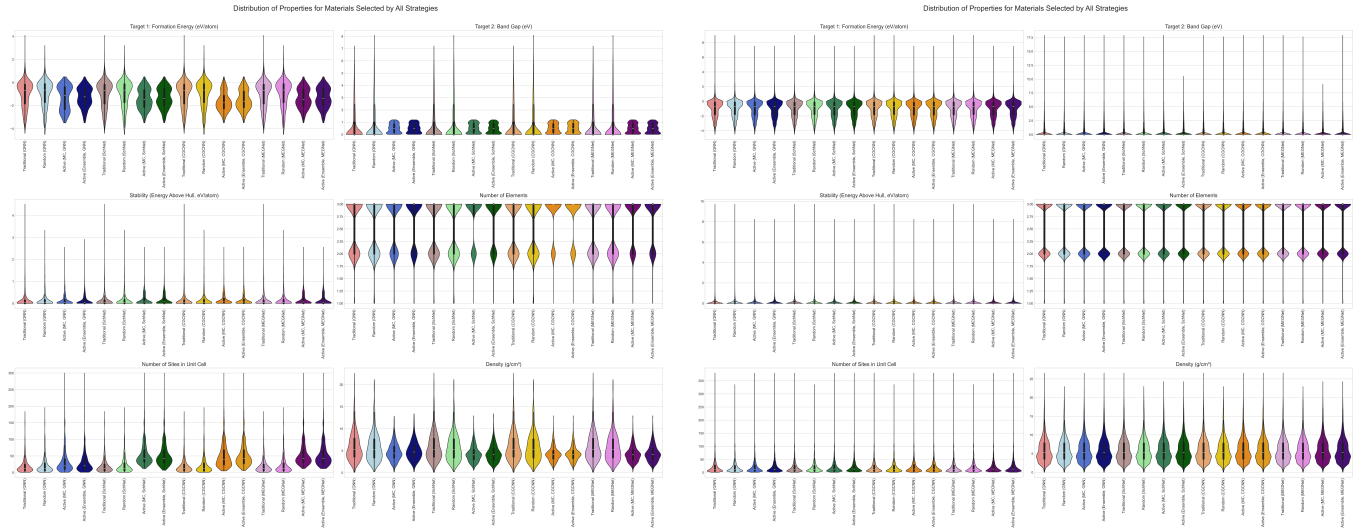
The author acknowledges the Materials Project for providing the public data used in this study and the developers of the open-source libraries PyTorch, PyTorch Geometric, and pymatgen that made this research possible.

References

[1] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011002, 2013.

TABLE II
COMPARISON OF UNFILTERED AND FILTERED FINAL TEST SET METRICS.

Model	Strategy	FE_MAE		FE_RMSE		FE_R2		BG_MAE		BG_RMSE		BG_R2	
		Unf	Filt	Unf	Filt	Unf	Filt	Unf	Filt	Unf	Filt	Unf	Filt
GNN	Active (MC)	0.1965	0.3398	0.3628	0.4852	0.8964	0.8166	0.3786	1.1648	0.7210	1.4146	0.7817	-0.0835
	Active (Ensemble)	**0.1581**	0.2675	**0.3413**	0.3933	**0.9084**	0.8795	**0.3366**	1.1687	**0.6560**	1.4377	**0.8192**	-0.1192
	Random Sampling	0.2009	0.3254	0.3773	0.4708	0.8880	0.8273	0.3551	1.2017	0.6892	1.4876	0.8005	-0.1982
	Traditional Learning	0.1876	0.2913	0.3682	0.4316	0.8933	0.8549	0.3628	1.2330	0.7247	1.5285	0.7794	-0.2651
SchNet	Active (MC)	1.51e11	0.6828	7.40e11	0.8689	-4.31e23	0.4119	1.95e11	1.2409	9.58e11	1.5126	-3.85e23	-0.2389
	Active (Ensemble)	3.51e11	0.6042	1.03e12	0.7870	-8.38e23	0.5175	4.23e11	1.1451	1.24e12	1.3761	-6.53e23	-0.0254
	Random Sampling	365.7101	0.7164	540.4358	1.1516	-229705.9	-0.0331	405.7150	1.3240	600.3611	1.7638	-151389.6	-0.6846
	Traditional Learning	1.62e13	59.5005	6.33e13	84.0677	-3.15e28	-5504.8027	1.95e13	8.4183	7.65e13	11.9050	-2.45e27	-75.7430
CGCNN	Active (MC)	5.19e05	0.9243	1.63e07	1.1259	-2.11e14	0.0125	3.01e05	1.3533	9.52e06	1.7270	-3.81e13	-0.6149
	Active (Ensemble)	1.30e06	0.9546	3.96e07	1.1570	-1.23e15	-0.0429	1.25e05	1.2132	3.60e06	1.4741	-5.47e12	-0.1766
	Random Sampling	3.26e08	1.3920	1.03e10	2.2149	-8.38e17	-2.8217	2.41e07	2.5532	7.62e08	3.7735	-2.44e17	-6.7103
	Traditional Learning	7.77e04	1.2059	2.40e06	2.7731	-4.55e12	-4.9911	9.33e04	1.5933	2.84e06	2.6688	-3.43e12	-2.8566
MEGNet	Active (MC)	0.6313	0.7451	0.8506	0.9273	0.4308	0.3302	0.8453	1.2105	1.2975	1.4265	0.2929	-0.1019
	Active (Ensemble)	0.5940	0.7017	0.8089	0.8624	0.4853	0.4206	0.8055	1.1448	1.2563	1.3443	0.3371	0.0214
	Random Sampling	6.1570	0.8264	8.2990	1.1675	4.5820	-0.0619	0.7965	1.2736	1.2682	1.5734	0.3244	-0.3405
	Traditional Learning	0.5703	0.7819	0.8070	1.3461	0.4876	-0.4116	0.7650	1.3059	1.1448	1.7701	0.4495	-0.6966



(a) Filtered Dataset (b) Unfiltered Dataset

Fig. 8. Comparison of queried-material distributions across filtered and unfiltered datasets. Active learning explores a wider and more diverse property space in the curated dataset, while the unfiltered dataset constrains exploration due to its heavily skewed and repetitive composition.

TABLE III
SIDE-BY-SIDE COMPARISON OF PREDICTED FORMATION ENERGY (FE) AND BAND GAP (BG) VALUES FOR FILTERED AND UNFILTERED DATASETS.

Filtered Dataset (mp-726207)			
Model	**Strategy**	**FE Pred**	**BG Pred**
DFT Value	–	**-1.237**	**1.165**
GNN	Active (MC)	-1.683	0.828
	Active (Ens.)	-1.481	0.575
	Random	-1.469	0.517
	Traditional	-0.953	0.655
SchNet	Active (MC)	-0.582	0.453
	Active (Ens.)	-0.799	0.460
	Random	-1.135	0.211
	Traditional	-61.236	3.425
CGCNN	Active (MC)	-1.919	0.666
	Active (Ens.)	-0.896	0.378
	Random	-2.140	-0.095
	Traditional	-1.648	0.401
MEGNet	Active (MC)	-1.372	0.487
	Active (Ens.)	-1.403	0.508
	Random	-1.870	0.544
	Traditional	-1.760	0.850

Unfiltered Dataset (mp-2767524)			
Model	**Strategy**	**FE Pred**	**BG Pred**
DFT Value	–	**-1.8809**	**0.8233**
GNN	Active (MC)	-2.0025	0.7855
	Active (Ens.)	-2.0037	1.5667
	Random	-1.9505	0.6434
	Traditional	-2.4096	2.3242
SchNet	Active (MC)	1.77e11	-2.29e11
	Active (Ens.)	7.19e11	-8.68e11
	Random	-543.03	603.06
	Traditional	-5.23e13	6.32e13
CGCNN	Active (MC)	-3.7675	3.4060
	Active (Ens.)	-2.2990	0.3765
	Random	-2.5586	0.8940
	Traditional	-1.9348	1.3482
MEGNet	Active (MC)	-1.1691	0.4464
	Active (Ens.)	-1.2617	0.3347
	Random	-1.2175	0.3167
	Traditional	-1.4623	0.3466

[2] T. Kalil and C. Wadia, "Materials Genome Initiative for Global Competitiveness," Office of Science and Technology Policy, June 2011.

[3] K. Alberi *et al.*, "The 2019 materials by design roadmap," *Journal of Physics D: Applied Physics*, vol. 52, no. 1, p. 013001, 2018.

[4] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," *Physical Review Letters*, vol. 120, no. 14, p. 145301, 2018.

[5] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," *npj Computational Materials*, vol. 7, no. 1, p. 185, 2021.

[6] T. Yin, G. Panapitiya, E. D. Coda, and E. G. Saldanha, "Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction," *Journal of Cheminformatics*, vol. 15, no. 1, p. 105, 2023.

[7] G. S. Jung, J. Y. Choi, and S. M. Lee, "Active learning of neural network potentials for rare events," *Digital Discovery*, vol. 3, no. 3, pp. 514–527, 2024.

[8] S. Thaler, F. Mayr, S. Thomas, A. Gagliardi, and J. Zavadlav, "Active learning graph neural networks for partial charge prediction of metal-organic frameworks via dropout Monte Carlo," *npj Computational Materials*, vol. 10, no. 1, p. 86, 2024.

[9] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation:

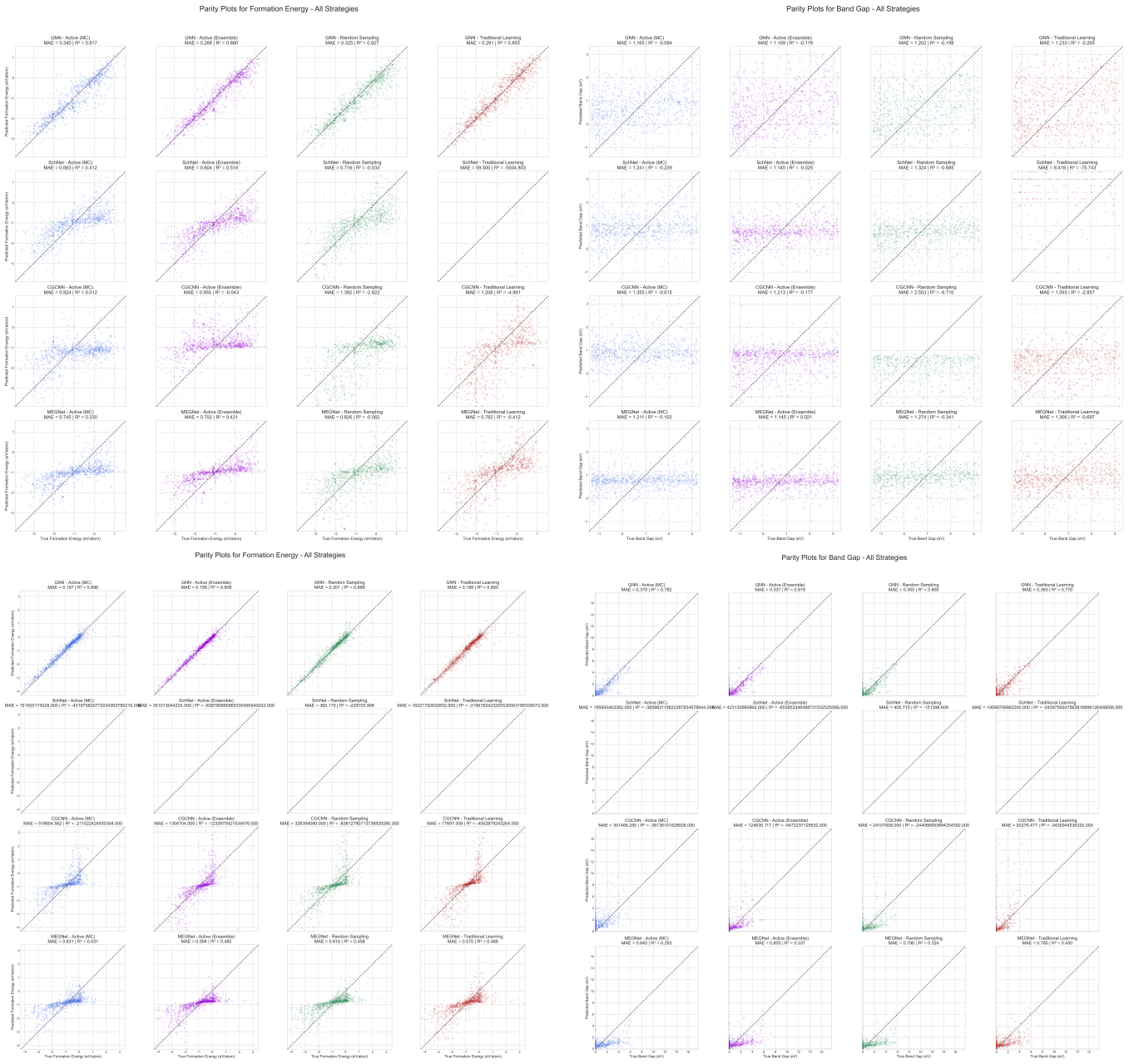


Fig. 9. Parity plots for Formation Energy (FE) and Band Gap (BG) predictions. The top row shows results on the filtered dataset, while the bottom row presents the corresponding parity plots on the unfiltered dataset.

Representing model uncertainty in deep learning," in *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, ser. PMLR, 2016, pp. 1050–1059.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 6402–6413.

[11] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017, pp. 992–1002.

[12] C. Chen and S. P. Ong, "A universal graph deep learning interatomic potential for the periodic table," *Nature Computational Science*, vol. 2, no. 11, pp. 718–728, 2022.

[13] S. Sanyal, J. Balachandran, N. Yadati, A. Kumar, S. Sanyal, P. Talukdar, and P. Rajagopalan, "MT-CGCNN: Integrating Crystal Graph Convolutional Neural Network with Multitask Learning for Material Property Prediction," arXiv preprint arXiv:1811.05660, 2018.

[14] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis," *Computational Materials Science*, vol. 68, pp. 314–319, 2013.

[15] M. Fey and J. E. Lenssen, "Fast graph representation learning with pytorch geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019. [Online]. Available: https://arxiv.org/abs/1903.02428