Deep Graph Kernel Learning for Material & Atomic Level Uncertainty Quantification in Adsorption Energy Prediction

Osman Mamun[1], Chenlu Yang[1], Shuwen Yue[1*]

[1]Robert F. Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY, 14853, USA.

*Corresponding author(s). E-mail(s): shuwen.yue@cornell.edu;
Contributing authors: om235@cornell.edu; cy533@cornell.edu;

Abstract

Graph Neural Networks (GNNs) provide an efficient surrogate for computationally intensive density functional theory calculations in catalytic material discovery. Yet, they often struggle with prediction reliability and out-of-domain generalization. These limitations necessitate reliable prediction uncertainty quantification for informed catalyst discovery. While Gaussian Processes (GPs) offer principled Bayesian uncertainty quantification, their cubic time complexity, high memory requirements, and inability to learn from graph structures limit their application in high-throughput discovery. We introduce Deep Graph Kernel Learning (DGKL), a scalable framework that combines a GNN backbone with a sparse variational Gaussian Process (SVGP) for uncertainty quantification in adsorption energy prediction. We benchmark DGKL against state-of-the-art methods, including ensemble/query-by-committee, evidential, and Monte-Carlo dropout approaches. DGKL consistently outperforms existing methods across ranking-based metrics (negative log-likelihood, expected normalized calibration error, miscalibration area) and error-based metrics (RMSE vs. RMV and error vs. standard deviation plots) while maintaining computational efficiency. Specifically, DGKL achieves the lowest expected normalized calibration error (0.06-0.10), lowest miscalibration area (0.04-0.07), and highest Spearman correlation coefficient (0.34-0.51) across diverse datasets and GNN backbone combinations. Qualitatively, DGKL's RMSE vs. RMV plots demonstrate superior calibration compared to competing methods. Additionally, we propose a DGKL variation capable of predicting atomic-level uncertainty - a feature absent in existing methods — offering fine-grained insights into out-of-domain data.

DGKL can be incorporated into active learning frameworks to efficiently explore catalytic material space, accelerating the discovery of novel catalysts.

Keywords: Adsorption Energy Prediction, Graph Kernel Learning, Uncertainty Quantification, Gaussian Processes

1 Introduction

Addressing climate change and energy scarcity necessitates the identification and development of renewable and sustainable energy processes [1]. Two critical domains that can significantly contribute to this endeavor are electricity generation via fuel cells and fuel production from renewable sources. In both areas, catalysts are essential for enhancing process efficiency and enabling novel catalytic reaction pathways [2, 3]. Given the time-intensive nature of catalyst discovery and optimization, it is imperative to integrate advanced computational techniques, particularly Density Functional Theory (DFT), into the materials discovery pipeline. DFT has substantially advanced our understanding of reaction mechanisms for specific catalytic materials, reconciling experimental observations and guiding the proposal of novel materials for experimental validation [4]. However, despite exponential growth in computational resources, the inherent time complexity of DFT calculations remains a significant bottleneck for high-throughput materials discovery applications.

Recently, Machine Learning (ML) models trained on DFT-generated datasets have emerged as promising alternatives to direct DFT calculations for high-throughput materials screening [5]. Among these, Graph Neural Networks (GNNs) have established themselves as state-of-the-art approaches for accelerating computational materials discovery[6, 7]. While GNNs demonstrate excellent performance on test data that conform to the training data distribution, their efficacy often diminishes when applied to out-of-distribution samples. This performance degradation underscores the necessity for robust uncertainty quantification (UQ) methods that can accurately assess the reliability of model predictions.

Conventional UQ approaches for GNNs include ensemble methods, Monte Carlo dropout, evidential methods, and mean-variance estimation techniques[8]. Ensemble models leverage the consensus among multiple models trained on different data subsets or with varying architectures. As test data diverges from the training distribution, independently initialized and trained models with different architectures or data should yield increasingly divergent predictions. Similarly, Monte Carlo dropout exploits architectural sampling by stochastically deactivating network weights during training. Evidential methods, inspired by subjective logic and Dempster-Shafer theory[9], learn to predict the parameters of probability distributions directly, allowing for second-order uncertainty estimation that distinguishes between epistemic and aleatoric uncertainty without requiring multiple forward passes.

Gaussian Processes (GPs) represent another established approach for Bayesian modeling, valued for their interpretability and reliable uncertainty estimates [10]. However, traditional GP implementations face several limitations: (1) popular kernel

functions typically allow the model to learn only limited degrees of freedom with sufficient smoothing, (2) they incur substantial computational and memory costs, and (3) they lack native capabilities for processing graph-structured data, restricting their application to molecular and materials systems.

Given the sophisticated representation learning capabilities of GNNs, integrating learned representations with GP-based uncertainty quantification presents an opportunity to synergize the advantages of both approaches. This combined methodology, termed Deep Graph Kernel Learning (DGKL), maps input graphs onto an intermediate feature space using a GNN, which subsequently serves as the input domain for a GP. Within this framework, GNN parameters effectively function as hyperparameters of the kernel function, enabling end-to-end training where both components are optimized simultaneously to enhance model performance metrics such as the evidence lower bound (ELBO) or negative log-likelihood (NLL).

However, implementing DGKL presents several practical challenges. Conventional GP formulations require maintaining the entire training dataset in memory to construct the kernel matrix, which becomes problematic as the intermediate latent space evolves during GNN training. To address this constraint, we employ Sparse Variational Gaussian Processes (SVGP)[11, 12], wherein the kernel matrix is approximated using a smaller set of inducing points. During training, the positions of these inducing points are optimized concurrently, establishing continuity between the GNN backbone and the SVGP model. Nevertheless, DGKL models remain challenging, primarily due to potential mode collapse and conflicting optimization dynamics between the GNN backbone and SVGP components. These challenges can be mitigated through various numerical and implementation strategies, including feature representation normalization at each layer and careful calibration of component-specific learning rates.

In this work, we develop a comprehensive uncertainty quantification method by integrating GNN architectures with SVGP models. We demonstrate that through judicious parameter selection and appropriate numerical techniques, DGKL models can be reliably trained for robust uncertainty quantification. We further extend our approach to develop DGKL-Atomic, a variant that provides granular uncertainty estimates at the atomic level rather than only at the molecular level. This capability is particularly valuable for identifying specific atomic environments that contribute most significantly to prediction uncertainty, enabling more targeted refinement of models and selective data acquisition strategies.

We evaluate DGKL and DGKL-Atomic against established UQ methods on adsorption energy prediction tasks for both in-distribution (ID) and out-of-distribution (OOD) datasets. The OOD evaluation is crucial for catalysis applications, where models often encounter novel catalyst compositions or adsorbates not represented in training data. Our benchmarks demonstrate that DGKL approaches maintain well-calibrated uncertainty estimates even for OOD predictions, where traditional methods such as ensembles and Monte Carlo dropout often exhibit inconsistent uncertainty calibration. This performance distinction is especially pronounced in challenging scenarios with novel catalyst materials and adsorbate combinations.

The atomic-level uncertainty quantification provided by DGKL-Atomic shows particular promise for guiding active learning workflows in computational catalysis. By identifying specific atomic sites with high uncertainty, this approach enables more efficient exploration of vast chemical spaces by prioritizing the addition of maximally informative structures to the training dataset. This targeted data acquisition strategy could substantially accelerate the discovery of novel catalytic materials by focusing computational and experimental resources on the most promising and uncertain regions of chemical space.

2 Results & Discussion

To rigorously evaluate the performance of Deep Gaussian Kernel Learning (DGKL) for adsorption energy prediction, we strategically selected two datasets with significant relevance to hydrocarbon conversion reactions: the Catalysis Hub (CatHub) repository and a carefully curated subset of the Open Catalyst 2020 (OC20) dataset. Comprehensive details regarding dataset composition, preprocessing protocols, and selection criteria are thoroughly documented in the Methods section. For our Graph Neural Network (GNN) architecture, we systematically implemented SchNet as the backbone across both datasets, while the more computationally intensive PaiNN architecture was exclusively deployed for the OC20 dataset to capture complex molecular interactions. In the following analysis, we first present a detailed comparative assessment of various dataset-GNN backbone combinations through the lens of both quantitative error-based metrics and qualitative interval-based evaluation approaches; the mathematical formulations and statistical interpretations of these complementary methodologies are precisely delineated in the Methods section. Subsequently, we critically examine the model's out-of-distribution generalization capabilities and the discriminative power of its predictive uncertainty estimates—a fundamental requirement for implementing robust uncertainty-driven active learning discovery protocols. The final component of our analysis focuses on atomic-level uncertainty quantification and its potential applications for establishing more sophisticated and granular control mechanisms in active learning discovery workflows for novel catalytic systems, effectively leveraging knowledge transferred from well-characterized catalytic databases as initialization points.

2.1 Performance of Different Dataset-GNN Combinations for Uncertainty Quantification

Throughout our analysis of various dataset-GNN architecture combinations, we provide a comprehensive reference to Tables 1 and 2, which present detailed quantitative comparisons of error-based and interval-based uncertainty metrics, respectively. To complement these numerical assessments, we direct readers to Figures 2–5, which offer qualitative visualizations and quantitative performance comparisons across the examined configurations. These visualizations illuminate important trends in predictive accuracy and uncertainty calibration that may not be immediately apparent from tabulated metrics alone.

2.1.1 CatHub Dataset with SchNet GNN Backbone

Table 1 presents uncertainty quantification (UQ) metrics for different methods evaluated on the CatHub hold-out test dataset using SchNet as the graph neural network (GNN) backbone. Figure 2 illustrates the distribution of absolute errors against predicted uncertainty, while Figure 3 shows the root mean squared error (RMSE) versus root mean-variance (RMV) relationships.

The predictive performance metrics (R^2 and MAE) indicate that DGKL and DGKL-Atomic exhibit slightly lower accuracy compared to ensemble, evidential, and MCD methods. However, our primary focus is developing well-calibrated uncertainty quantification rather than maximizing prediction accuracy alone.

Examining the UQ metrics, we observe that the ensemble model achieved an NLL of -0.06 ± 0.37, while DGKL and DGKL-Atomic produced values of 0.68 ± 0.09 and 0.42 ± 0.16, respectively. Although lower NLL values are generally preferable, the high standard deviation associated with the ensemble model (± 0.37) suggests inconsistent performance across different datasets and architectures. Also, having a lower MAE or RMSE facilitates lower NLL, as discussed in the methods section, which does not necessarily mean the UQ is generally better with a lower NLL. In contrast, both DGKL variants demonstrate more robust performance with lower standard deviations in their NLL values, indicating better reliability across different experimental conditions.

For the Expected Normalized Calibration Error (ENCE), all methods except evidential & MCD exhibited relatively low standard deviations, confirming their stability. The ENCE performance ranking is: DGKL (0.07 ± 0.04) > DGKL-Atomic (0.10 ± 0.05) > Ensemble (0.26 ± 0.10) > Evidential (0.79 ± 0.43) > MCD (1.94 ± 0.59). A similar trend is observed for the miscalibration area, with values of 0.04 ± 0.02, 0.05 ± 0.02, 0.12 ± 0.05, 0.46 ± 0.22, and 0.91 ± 0.24, respectively, further confirming the superior calibration of DGKL & DGKL Atomic methods.

The Spearman correlation coefficient between absolute error and predicted standard deviation also follows this pattern, with DGKL achieving the highest value (0.51 ± 0.02), followed by DGKL-Atomic (0.46 ± 0.01), Ensemble (0.41 ± 0.10), Evidential (0.30 ± 0.08), and MCD (0.18 ± 0.02). This indicates that DGKL variants better capture the relationship between model uncertainty and actual prediction errors.

A critical metric for UQ performance is the calibration R^2 of the RMSE vs. RMV plot, which ideally should approach 1.0. This metric measures explicitly how well the model's predicted uncertainty (RMV) correlates with the actual error (RMSE), not predictive accuracy. DGKL and DGKL-Atomic achieve near-perfect calibration with R^2 values of 1.00 ± 0.00 and 1.00 ± 0.00, respectively, while Ensemble (0.99 ± 0.01), Evidential (0.98 ± 0.01), and MCD (0.95 ± 0.02) show progressively poorer calibration. The slope & intercept of the reliability diagram (RMSE vs. RMV) also shows DGKL variants to be more well-calibrated than the other models, barring the ensemble, which is on par with the DGKL and DGKL Atomic. For empirical coverage within 1σ and 2σ standard deviations, the theoretical expectations are 68.27% and 95.45%, respectively. DGKL achieves nearly ideal coverage of 0.68 ± 0.02 and 0.95 ± 0.01, followed closely by DGKL-Atomic with 0.69 ± 0.03 and 0.94 ± 0.01.

The ROC-AUC metric, which quantifies a model's ability to distinguish between low and high-error predictions based on uncertainty estimates, shows values of $0.74 \pm$

0.01 for DGKL, 0.72 ± 0.01 for DGKL-Atomic, 0.70 ± 0.05 for Ensemble, 0.64 ± 0.04 for Evidential, and 0.59 ± 0.01 for MCD.

Figure 2 reveals that the distribution of absolute errors correlates better with predicted uncertainty for DGKL and DGKL-Atomic compared to other methods, as evidenced by the more consistent spread of errors at higher uncertainty values. Figure 3 further confirms that DGKL achieves nearly perfect calibration, with data points closely following the ideal line across the entire range. DGKL-Atomic shows slight deviation only at higher RMV values, while the Ensemble model exhibits more significant deviations at both extremes of the plot. As expected, evidential and MCD perform poorly, with substantial deviations from the ideal calibration line.

These comprehensive evaluations demonstrate that DGKL provides well-calibrated uncertainty estimates while maintaining computational efficiency compared to ensemble methods. DGKL-Atomic delivers comparable performance with the added benefit of atomic-level uncertainty quantification, accepting a minor accuracy trade-off. In contrast, evidential & MCD consistently underperform across all UQ metrics, indicating its limitations for robust uncertainty quantification in this domain. The apparent contradiction between predictive accuracy and well-calibrated uncertainty can be explained by the fact that the non-DGKL models are optimized with Huber loss (for ensemble and MCD) and NIG loss (for evidential) which resulted in higher accuracy but overconfident standard deviations. From Figure 5, we can see the predictive uncertainty is narrowly distributed for ensemble, evidential, and MCD while the error outside 2σ envelope suggests the predicted standard deviation should be higher. While for DGKL and DGKL Atomic, the predictions are more scattered, the points outside the 2σ envelope also show higher predictive uncertainty.

2.1.2 CatHub Dataset with PaiNN GNN Backbone

Extending our analysis to the CatHub dataset with PaiNN as the GNN backbone, we observe several notable differences in performance patterns. Table 1 shows that the Evidential and Ensemble methods achieve the lowest MAE values of 0.13 ± 0.02 and 0.15 ± 0.09, respectively, outperforming DGKL (0.31 ± 0.07) and DGKL-Atomic (0.30 ± 0.02). Overall, the MAE values are lower compared to the SchNet models, thanks to the rotational equivariance of PaiNN models, resulting in a more accurate model because of enhanced representation of graph structure.

The Ensemble method again demonstrates the best NLL performance at -0.11 ± 0.31, followed by DGKL-Atomic (0.34 ± 0.07), DGKL (0.36 ± 0.18), and Evidential (0.31 ± 0.70). However, the substantial standard deviations for the Ensemble and particularly the Evidential method indicate less consistent performance across different data subsets, similar to what we observed with SchNet.

For ENCE, both DGKL variants perform identically well ($0.10 \pm 0.03/0.05$), significantly outperforming Ensemble (0.41 ± 0.21), Evidential (0.49 ± 0.26), and MCD (2.30 ± 0.51). This pattern extends to the miscalibration area metric, where DGKL-Atomic (0.05 ± 0.02) and DGKL (0.06 ± 0.03) demonstrate superior calibration compared to Ensemble (0.21 ± 0.13), Evidential (0.37 ± 0.22), and MCD (1.23 ± 0.13).

The Spearman correlation coefficient between absolute error and predicted uncertainty shows DGKL and Ensemble methods performing similarly (0.48 ± 0.03 and

0.47 ± 0.07), followed by DGKL-Atomic (0.44 ± 0.02), with Evidential (0.26 ± 0.05) and MCD (0.15 ± 0.02) demonstrating weaker correlations. For ROC-AUC, both DGKL and Ensemble methods achieve the highest scores ($0.73 \pm 0.02/0.03$), closely followed by DGKL-Atomic (0.71 ± 0.01).

The calibration R^2 of the RMSE versus RMV plot shows perfect calibration for DGKL, DGKL-Atomic, and Ensemble (all at 1.00 ± 0.00), while Evidential (0.93 ± 0.02) and MCD (0.92 ± 0.01) exhibit poorer performance. For the slope, which ideally should be 1.0, the Ensemble and Evidential methods perform best (both at 0.95, with different standard deviations), while DGKL (1.22 ± 0.18) and DGKL-Atomic (1.28 ± 0.05) show a tendency to overestimate uncertainty at higher variance regions.

Regarding empirical coverage, the Ensemble method achieves nearly ideal 1σ coverage (0.68 ± 0.11), while DGKL-Atomic (0.69 ± 0.03) and DGKL (0.70 ± 0.03) slightly overestimate uncertainty. The Evidential method significantly overestimates uncertainty (0.87 ± 0.13), and MCD underestimates it (0.37 ± 0.05). At the 2σ level, DGKL, DGKL-Atomic, and Evidential methods achieve near-ideal coverage (0.94 ± 0.02, 0.94 ± 0.01, and 0.96 ± 0.04, respectively).

Examining the error distribution plots reveals that both DGKL variants maintain a consistent correlation between absolute errors and predicted uncertainty. The Evidential method, despite its strong predictive accuracy, demonstrates poor alignment between error magnitude and uncertainty estimates, with numerous high-error predictions receiving low uncertainty scores.

In summary, with the PaiNN backbone, the Evidential and Ensemble methods achieve better predictive accuracy, but the DGKL variants continue to demonstrate superior uncertainty calibration across most metrics. The Evidential method, in particular, despite its competitive MAE, suffers from inconsistent uncertainty estimation and poor calibration properties. Such underperformance of UQ metrics of the evidential method is well-reported in various literature[13–15].

2.1.3 OC20 Dataset with SchNet GNN Backbone

Moving to the larger and more complex OC20 dataset with SchNet as the GNN backbone, we observe several notable shifts in performance trends. Table 1 reveals that the Ensemble method achieves the best MAE (0.48 ± 0.03), followed by Evidential (0.50 ± 0.01), MCD (0.51 ± 0.01), DGKL-Atomic (0.57 ± 0.02), and DGKL (0.69 ± 0.01). This represents a more substantial gap in predictive accuracy between ensemble-based and DGKL methods compared to the CatHub dataset. Due to the objective function (predictive log-likelihood) for DGKL variants, their predictive performance is poor, a trade-off the training algorithm makes to accommodate better uncertainty quantification.

For NLL, DGKL-Atomic demonstrates the best performance (1.07 ± 0.03), followed by DGKL (1.30 ± 0.02), Evidential (1.78 ± 0.30), Ensemble (4.96 ± 1.19), and MCD (34.07 ± 16.18). The dramatically higher NLL values and standard deviations for Ensemble and particularly MCD indicate severe miscalibration on this more challenging dataset, while both DGKL variants maintain relatively consistent performance.

A similar pattern emerges for ENCE, with DGKL achieving the best result (0.06 ± 0.04), closely followed by DGKL-Atomic (0.08 ± 0.05), and with Evidential (0.65 ± 0.10), Ensemble (1.95 ± 0.34), and MCD (6.19 ± 1.69) showing progressively poorer calibration. The miscalibration area metric confirms this trend, with DGKL (0.07 ± 0.02) and DGKL-Atomic (0.09 ± 0.05) substantially outperforming Evidential (0.52 ± 0.08), Ensemble (1.07 ± 0.24), and MCD (2.54 ± 1.26).

For the Spearman correlation coefficient, DGKL-Atomic achieves the highest value (0.34 ± 0.02), followed by DGKL (0.25 ± 0.02), Ensemble (0.19 ± 0.01), MCD (0.15 ± 0.02), and Evidential (0.08 ± 0.14). The ROC-AUC metric shows a similar ranking, with DGKL-Atomic leading (0.66 ± 0.01), followed by DGKL (0.62 ± 0.01), Ensemble (0.59 ± 0.01), MCD (0.57 ± 0.01), and Evidential (0.54 ± 0.07).

The calibration R^2 of the RMSE versus RMV plot reveals excellent performance for DGKL-Atomic (1.00 ± 0.00) and DGKL (0.99 ± 0.01), with Ensemble (0.99 ± 0.01) and MCD (0.98 ± 0.01) also performing well, while Evidential (0.59 ± 0.33) shows remarkably poor calibration with high variability. However, examining the slope values reveals significant deviations from the ideal value of 1.0 for MCD (2.23 ± 1.17) and extremely poor performance for Evidential (0.07 ± 0.20), indicating severe systematic miscalibration despite reasonable R^2 values in some cases.

The most striking disparities appear in the empirical coverage metrics. At the 1σ level, DGKL-Atomic (0.74 ± 0.04) and DGKL (0.72 ± 0.03) slightly overestimate uncertainty, while Evidential significantly overestimates (0.97 ± 0.02), and Ensemble (0.36 ± 0.04) and MCD (0.16 ± 0.03) severely underestimate it. At the 2σ level, DGKL-Atomic (0.96 ± 0.01) and DGKL (0.94 ± 0.01) maintain near-ideal coverage, Evidential reaches perfect but meaningless coverage (1.00 ± 0.00) due to overestimation, while Ensemble (0.62 ± 0.05) and MCD (0.31 ± 0.06) continue to underestimate uncertainty significantly.

The error distribution plots further illustrate these findings, showing that DGKL and DGKL-Atomic maintain consistent relationships between error magnitude and predicted uncertainty even on this challenging dataset. In contrast, Ensemble, Evidential, and MCD methods show poor alignment between errors and uncertainty estimates, with particularly problematic calibration at high uncertainty levels.

These results highlight a crucial strength of the DGKL methodology: while its predictive accuracy may lag behind ensemble methods on complex datasets, its uncertainty calibration remains robust. The dramatic deterioration in calibration quality for the Ensemble, Evidential, and MCD methods on OC20 suggests that these approaches are less generalizable to challenging data distributions. DGKL-Atomic emerges as particularly valuable in this context, offering better predictive accuracy than standard DGKL while maintaining excellent calibration and providing atom-level uncertainty information.

2.2 Performance on Out-of-Domain Datasets

Evaluating model performance and uncertainty estimation on out-of-distribution (OOD) data is crucial for assessing robustness and generalizability in real-world applications. Figure 6 illustrates the uncertainty distributions across different OOD scenarios for the OC20 dataset using the SchNet backbone.

The OC20 benchmark provides three distinct OOD test sets: OOD-cat (new catalyst compositions not seen during training), OOD-ads (new adsorbates not seen during training), and OOD-both (new combinations of both catalysts and adsorbates). These datasets represent progressively more challenging generalization scenarios for energy prediction models.

As shown in Figure 6(a), the DGKL method demonstrates relatively consistent uncertainty estimates across all datasets, with similar median values and distributions for ID and OOD scenarios. This consistency indicates that DGKL maintains reliable calibration even when faced with novel data distributions. We observe slight increases in the mean uncertainty values from ID (0.7 eV) to OOD-cat and OOD-ads (approximately 0.9 eV), with similar distributions for OOD-both, which means the model can discriminate between ID and OOD data, albeit the discriminative power is not as significant as we would expect from a capable confidence aware model.

In contrast, the Ensemble method (Figure 6(b)) shows dramatically different behavior, with substantially higher uncertainty estimates for OOD data compared to ID data. The median uncertainty increases from approximately 0.2 eV for ID data to 0.6 eV for OOD-cat, 1.0 eV for OOD-ads, and 0.5 eV for OOD-both. The much wider distribution and higher values for OOD-ads suggest that the ensemble method is susceptible to novel adsorbates. Interestingly, the uncertainty for OOD-both is lower than for OOD-ads alone, indicating a potential inconsistency in how the ensemble method captures uncertainties across different distribution shifts. However, as seen from the violin plot, the uncertainty is not as well-behaved, indicated by the low-lying uncertainties of the OOD data in the near-zero region.

Figure 6(c) shows that DGKL-Atomic exhibits superior behavior than standard DGKL and the Ensemble method. It maintains more consistent uncertainty distributions than the Ensemble method while differentiating between ID and OOD data. The median uncertainty increases from approximately 0.7 eV for ID data to 1.25 eV for all three OOD datasets, with slightly wider distributions for the OOD cases. However, the error distribution is much more physically meaningful, as the low-lying uncertainties are higher than the low-lying uncertainties of the ID data.

These findings indicate that while the Ensemble method exhibits sensitivity in detecting distributional shifts, its performance exhibits inconsistency across various distributional shifts. Consequently, the error estimate derived from this method may be unreliable for out-of-distribution (OOD) data. In contrast, DGKL-Atomic offers a more stable and consistent uncertainty estimation across diverse data distributions, potentially enhancing the reliability of decision-making processes in practical applications, such as active learning for novel catalytic applications.

To quantitatively assess each method's ability to discriminate between in-distribution (ID) and out-of-distribution (OOD) data, we computed the Receiver Operating Characteristic Area Under Curve (ROC-AUC) scores shown in Table 3. For this analysis, we assigned binary labels to the data points (ID samples as 0, OOD samples as 1) and used the predicted standard deviation values as classification scores. The resulting ROC-AUC represents the probability that the model assigns a higher uncertainty to a random OOD example than a random ID example. It is noteworthy that an ROC-AUC score of 0.5 indicates the absence of any distinction between the

ID and OOD uncertainties. In contrast, a score of 1.0 signifies the presence of a clear and distinct separating line between ID and OOD uncertainties.

The standard DGKL method shows modest discriminative power, with ROC-AUC scores ranging from 0.566 to 0.603 across the different OOD scenarios, confirming our observations from the uncertainty distributions. The Ensemble method demonstrates substantially improved OOD detection performance (ROC-AUC between 0.732 and 0.790), consistent with its more distinct uncertainty distributions for ID and OOD data. DGKL Atomic notably achieves the highest ROC-AUC scores across all OOD scenarios (0.837-0.876), indicating superior OOD detection capability. Interestingly, even when using purely atomic-level uncertainties without molecular aggregation (denoted as "DGKL Atomic (Atomic)" in the table), we still achieve strong OOD detection performance (ROC-AUC scores of 0.782-0.791), demonstrating the fundamental discriminative power of atomic-level uncertainty estimates.

2.3 Atomic-level Uncertainty

A unique advantage of the DGKL Atomic method is its ability to quantify uncertainty at the atomic level, providing more granular insights into prediction reliability. Figure 6(d) compares the distributions of atomic-level uncertainties across ID and OOD datasets.

For the ID test set, we observe a relatively narrow distribution of atomic uncertainties centered around 0.01σ, with occasional higher values up to approximately 0.04σ. As we move to the OOD scenarios, several notable patterns emerge:

1. All OOD datasets show higher median atomic uncertainties compared to the ID dataset, with values increasing from approximately 0.01σ for ID to 0.02σ for each of the OOD datasets.

2. The distribution width increases for all OOD datasets, indicating more significant variability in atomic-level uncertainty when encountering novel structures or compositions.

3. Unlike the molecule-level uncertainties shown in previous plots, the atomic-level uncertainties show relatively similar distributions across all three OOD scenarios (OOD-cat, OOD-ads, and OOD-both), suggesting that atomic uncertainties may capture local chemical environment changes more consistently, regardless of whether they arise from novel catalysts or adsorbates.

This atomic-level uncertainty information provides valuable insights for materials design and discovery applications. For instance, identifying specific atoms with high uncertainty can guide targeted data acquisition or experimental validation efforts. In catalysis applications, atoms with high uncertainty often correspond to active sites or regions experiencing significant electronic structure changes during reactions. Specifically, if it is coupled with atomic environment analysis tools, such as the smooth overlap of atomic positions (SOAP)[16, 17], it can identify the local atomic environment that has the highest uncertainty. In subsequent iterations, the addition of materials with those particular local atomic environments will offer better exploratory or exploitative control during the active learning cycle.

The consistency in atomic uncertainty distributions across different OOD scenarios suggests that DGKL-Atomic captures fundamental aspects of prediction reliability

at the atomic level that transcend the specific nature of the distribution shift. This property makes it particularly valuable for guiding active learning strategies in materials discovery, where identifying the most informative structures for additional data collection is crucial for the efficient exploration of vast chemical spaces.

Furthermore, atomic-level uncertainties enable more interpretable model outputs by highlighting specific structural regions where predictions may be less reliable, enhancing trust in the model and providing mechanistic insights that are not available from molecule-level uncertainty estimates alone.

3 Methods

3.1 Graph Models

For the graph neural network (GNN) base models, we employ SchNet [18] and PaiNN [19]. Both are message-passing neural networks (MPNNs) designed specifically for predicting molecular and material properties. Below, we briefly describe the architectural features of these two models.

3.1.1 SchNet

SchNet is an MPNN that maintains invariance to atom indexing and translation, thereby conforming to the physical principle that intensive properties of materials, such as adsorption energy, should remain unaffected by permutation of identical atoms, reordering of atomic positions, or translation of atomic coordinates. The energy predicted by SchNet is also rotationally invariant, adhering to fundamental physical symmetry constraints.

In the SchNet architecture, atoms' nuclear charges are initially embedded through an atom-type embedding layer. These embeddings are subsequently processed through a series of interaction blocks designed to model interatomic interactions. A key innovation in this architecture is the utilization of continuous filters rather than discrete ones to capture the radial dependency of interactions. After processing through multiple interaction blocks, the resulting representation passes through several dense layers, and finally, the energy is predicted using an invariant operation such as sum pooling or mean pooling.

3.1.2 PaiNN

While SchNet demonstrates excellent performance in various material property prediction tasks, it can be less data-efficient compared to fully equivariant methods. A primary limitation of SchNet is its handling of rotational symmetry, specifically the lack of equivariant representation, which necessitates learning rotational equivariance explicitly through data augmentation. The Polarizable Atom Interaction Neural Network (PaiNN) addresses this limitation by explicitly leveraging the directional properties of relative positions between atoms.

Similar to SchNet, PaiNN processes atomic embeddings through several layers of message and update blocks. However, PaiNN computes and updates both scalar and vector properties in a manner that preserves rotational symmetry. Subsequently,

these embeddings are processed through dense layers followed by a pooling operation, analogous to the approach in SchNet, to predict the target properties.

3.2 Uncertainty Quantification Models

3.2.1 Ensemble

The ensemble method is a widely used uncertainty quantification (UQ) technique across various molecular and material property prediction applications. It represents a simple yet reliable approach to UQ, as demonstrated by several benchmark studies [20–23]. Due to its consistent performance in ensembling/query-by-committee frameworks, it is frequently referred to as the 'gold standard' for uncertainty quantification.

Ensemble methods capture model uncertainty (epistemic uncertainty) by employing varied model architectures or different parameter initializations, allowing models to converge to distinct local minima on the energy landscape. The average prediction across these trained models can be conceptualized through a statistical mechanics analogy to ensemble averaging. The standard deviation among predictions from this ensemble provides a straightforward and effective means of quantifying uncertainty. Through full or partial exploration of the loss landscape, ensemble methods demonstrate robustness to noisy and out-of-distribution test samples [24].

While this approach offers simplicity and effectiveness, it presents two notable limitations: (1) the computational cost scales linearly with the number of constituent models, resulting in compute-intensive training, and (2) inference requires each test sample to be evaluated by every model in the ensemble, reducing inference speed.

For our ensemble of graph models, we conducted hyperparameter optimization across various model configurations, including learning rate, batch size, number of hidden channels, and number of layers, using the open-source Ray Tune library [25]. The five best-performing model configurations were selected for the final ensemble. Importantly, we used identical training and validation sets for hyperparameter optimization and final model training to prevent data leakage to the test set.

3.2.2 Monte Carlo Dropout

Monte Carlo (MC) dropout represents a computationally efficient alternative to traditional ensemble approaches, enabling uncertainty quantification through a single model within a probabilistic framework. MC dropout was first formalized by Gal & Ghahramani [26] in their seminal 2016 paper, which established that applying dropout—probabilistically deactivating connections to different nodes during neural network training—can be interpreted as approximate Bayesian inference in deep Gaussian processes.

This training approach serves as an effective regularization technique, as it prevents the model from relying on any specific set of connections for making predictions. During standard inference, the weighted average of node predictions based on their assigned probabilities provides the mean estimation. However, for uncertainty quantification, one can generate multiple predictions by maintaining active dropout during inference time. These stochastic forward passes yield a distribution of predictions from which standard deviation can be calculated, providing a measure of model uncertainty.

Conceptually, this process parallels the ensemble approach, as each stochastic forward pass effectively samples from a slightly different neural network architecture. The key advantage lies in computational efficiency: training requires only the resources needed for a single graph model, though inference still necessitates multiple forward passes to generate the prediction distribution.

For our implementation, we performed hyperparameter optimization over the same parameter space as used for the ensemble models, with the addition of dropout probability as an optimization parameter. After training the optimal model, we executed 10 stochastic forward passes per input to obtain robust uncertainty estimations.

3.2.3 Evidential Regression

In contrast to sampling-based methods such as ensemble and Monte Carlo dropout, evidential regression represents a sampling-free approach to uncertainty quantification. Rather than directly parameterizing the target variable, evidential regression parameterizes an evidential distribution over the target. For our adsorption energy prediction task, the network is trained to learn the parameters of a Normal-Inverse-Gamma (NIG) distribution [27–29].

The loss function for evidential regression comprises two components: (1) the NIG loss, which encourages accurate prediction of the target distribution, and (2) a regularization loss that penalizes overconfident predictions, particularly for out-of-distribution samples. Given the learned NIG distribution parameters, we compute the aleatoric uncertainty as:

$$\sigma^2_{\text{aleatoric}} = \frac{\beta}{\alpha - 1} \tag{1}$$

where β and α are NIG distribution parameters learned during training. The parameter α represents the degrees of freedom (shape parameter) and must satisfy $\alpha > 1$ for the variance to be properly defined, while $\beta > 0$ serves as the scale parameter of the distribution.

Consistent with our approach for ensemble methods, we performed hyperparameter optimization for evidential regression using the same parameter space as applied to the ensemble models to ensure fair comparison across methods.

3.2.4 Deep Graph Kernel Learning (DGKL)

Deep Graph Kernel Learning (DGKL) comprises two principal components: a Graph Neural Network (GNN) that encodes data into a learned representation via a low-dimensional latent space and a Gaussian Process (GP) that learns the kernel function from this latent space. The parameters of both the GNN and GP are optimized simultaneously in an end-to-end manner. For the GNN backbone to learn meaningful representations, we employ SchNet and PaiNN models, as discussed previously. Below, we elaborate on the GP component of DGKL. In Figure 1, we illustrate the DGKL architecture for adsorption energy prediction and uncertainty quantification. The workflow shows how atomic positions and chemical information from adsorbate/slab models flow through SchNet/PaiNN neural networks into a latent space representation

with inducing points. This representation is then processed through Sparse Variational Gaussian Process (SVGP) regression to yield adsorption energy predictions with quantified uncertainty in the form of a normal distribution $\Delta E_{\text{ads}} = \mathcal{N}(\mu, \sigma^2)$.

Gaussian Processes are highly effective non-parametric methods that learn distributions over function spaces via kernel functions, which dictate how data points correlate with each other [30, 31]. They are ubiquitously employed in machine learning and statistics for their flexibility and expressiveness in modeling complex relationships between variables probabilistically, providing both mean predictions and associated uncertainty estimates. Following the notation of Rasmussen and Williams [32], a GP prior can be defined as:

$$f(M) \sim \mathcal{GP}(\mu(M), K(M, M')) \tag{2}$$

where $\mu(\cdot)$ is the mean function and $K(\cdot, \cdot)$ is the covariance function. By conditioning the model parameters on observed labels y_i, we can predict values for new inputs $M_\star$ via the predictive distribution:

$$\mathbb{E}[f_\star] = \mu(M_\star) + K(M_\star, M)[K(M, M) + \sigma^2 I]^{-1} y \tag{3}$$

$$\text{Cov}[f_\star] = K(M_\star, M_\star) - K(M_\star, M)[K(M, M) + \sigma^2 I]^{-1} K(M, M_\star) \tag{4}$$

where σ^2 is the noise variance and I is the identity matrix. M_* and M are the gram matrix for the unknown and known points, respectively.

Due to the inversion of the Gram matrix, the prediction time complexity of GP is $\mathcal{O}(N^3)$, where N is the number of training samples, necessitating high memory requirements and extended prediction times. To address this limitation, several approaches have been proposed, such as sparse approximations and variational inference [33, 34]. A GP combining sparse and variational approximations is known as a Sparse Variational GP (SVGP). Rather than learning the kernel function over all training samples, SVGP selects a subset of data points in the latent space and learns the kernel function over these selected points. Mathematically, it replaces M, the known training points in Equation 3 & 4, with U, the inducing points in the SVGP framework. To efficiently approximate the training data distribution, SVGP learns the inducing points' locations as part of the optimization process.

In the DGKL formalism, the GNN acts as a feature extractor, and SVGP is employed to circumvent the high computational demands of exact GP inference. The feature extractor is trained to produce latent space representations at the molecular or material level, as the per-atom learned representations are aggregated with a global pooling operation (i.e., sum pooling or mean pooling). The resulting latent space representation, when trained with SVGP to predict adsorption energy, provides not only the predicted energy values but also the uncertainty associated with each prediction.

In this work, we systematically explored the following parameters to identify the optimal SVGP model, in addition to the GNN backbone parameters used in the ensemble models:

1. Latent space dimension
2. Kernel functions: Matérn and RBF
3. Distribution types for SVGP: Cholesky and meanfield distributions

4. Variational strategies: standard and decoupled approaches
5. Objective functions: Evidence Lower Bound Optimization (ELBO) and Predictive Log Likelihood (PLL)

It is well-established that training DGKL end-to-end presents significant challenges due to the uneven dynamics between GNN and SVGP components. Moreover, the latent space representations are not directly fed into the SVGP but are approximated using inducing points, potentially leading to mode collapse and other numerical instabilities (e.g., vanishing or exploding gradients, unstable Cholesky decomposition of the Gram matrix). To ensure robust training of such networks, we implemented several strategies:

1. Normalization: We incorporated layer normalization and feature normalization techniques to condition the features, ensuring they remain smooth and numerically well-behaved.
2. Learning rate management: The learning rate for the GNN backbone was set at least two orders of magnitude lower than that of the SVGP to maintain relatively stationary inducing point positions between successive epochs.
3. Early stopping: Training was halted if validation loss failed to improve over five consecutive epochs, preventing overfitting to the training data—a standard practice in deep learning.
4. Adaptive learning rate adjustments: To prevent numerical instabilities (e.g., NaN values) that could prematurely terminate training, we implemented a mechanism to roll back network states to weights from before the latest update and reduce the learning rate by a factor of two when such issues were detected.

By implementing these techniques, we successfully trained the DGKL model end-to-end with robust stability across various configurations.

3.2.5 Deep Graph Kernel Learning with Atomic Uncertainty (DGKL-Atomic)

In the DGKL-Atomic formalism, we learn the latent space representation at the atomic level rather than the material level. This is achieved by deferring the global pooling operation until after the SVGP inference. Consequently, SVGP provides both the energy of each atom and the uncertainty associated with the prediction. These atomic-level uncertainties are then aggregated to yield the predicted adsorption energy and its corresponding uncertainty at the material level.

For the aggregation process, we sum the mean predictions at the atomic level to obtain the adsorption energy, while the elements of the covariance matrix are combined to determine the uncertainty of the adsorption energy. This can be mathematically expressed as:

$$\mathbb{E}[E_{\text{ads}}] = \sum_i \mathbb{E}[E_{\text{atomic}}]_i \tag{5}$$

$$\text{Cov}(E_{\text{ads}}) = \sum_{i,j} \text{Cov}(E_{\text{atomic}})_{i,j} \tag{6}$$

where $\mathbb{E}[E_{\text{ads}}]$ represents the expected adsorption energy, $\mathbb{E}[E_{\text{atomic}}]_i$ denotes the expected energy contribution from atom i, and $\text{Cov}(E_{\text{atomic}})_{i,j}$ represents the covariance between atoms i and j.

The DGKL-Atomic approach offers several advantages over traditional material-level uncertainty quantification methods. By modeling uncertainty at the atomic level, we can capture local electronic and structural contributions to adsorption energetics with greater fidelity. This granularity enables the identification of specific atomic sites that contribute most significantly to overall prediction uncertainty, providing valuable insights for targeted material design and optimization.

Furthermore, the atomic-level uncertainty quantification facilitates interpretability by allowing researchers to visualize uncertainty distributions across the material structure. This spatial uncertainty mapping can reveal patterns related to atomic coordination environments, bond configurations, and surface features that may correlate with prediction reliability. Such detailed uncertainty characterization is particularly valuable for complex heterogeneous catalysts and nanostructured materials where local atomic environments vary considerably.

The computational implementation of DGKL-Atomic leverages the inherent graph structure of the material representation, enabling efficient propagation of both feature information and uncertainty through the network. This approach maintains computational tractability while providing a more comprehensive uncertainty profile than conventional methods operating at the material level.

3.3 Datasets

To evaluate the performance of our developed DGKL framework, we utilize two datasets: Catalysis Hub [35, 36], and Open Catalyst 2020 (OC20) [37], both of which were developed to facilitate high-throughput catalytic materials research.

The Catalysis Hub dataset contains approximately 37,000 entries pertaining to adsorption energies of CH, CH_2, CH_3, OH, NH, and SH on approximately 2,035 distinct bimetallic combinations of transition metals. This dataset provides a comprehensive benchmark for evaluating catalytic performance across a diverse range of metal combinations.

In contrast, the OC20 dataset comprises approximately 470,000 adsorption energy data points across a broad spectrum of transition metal and adsorbate combinations. Due to the computational intensity required to train models on the complete dataset, we employ a selected subset of 46,000 adsorption energy data points. This subset was constructed by selecting structures containing catalytically important transition metals: Ru, Re, Pt, Pd, Cu, Ni, Fe, Co, Rh, Ir, Mo, W, Au, Ag, Cr, Mn, Zn, Al, Ti, Zr, and V. We also refined our selection to include only structures with adsorbates containing carbon, hydrogen, and oxygen atoms.

For all model training procedures throughout this study, we consistently employed a 70:10:20 ratio for train, validation, and test sets, respectively, ensuring robust evaluation and comparison of model performance.

3.4 Evaluation Metrics

To evaluate the performance of uncertainty quantification methods, several metrics have been developed in the literature. These metrics often provide different and sometimes inconsistent insights about the performance of uncertainty quantification models [38]. Broadly, the evaluation metrics for UQ can be divided into two classes: (1) error-based approaches and (2) interval-based approaches.

Error-based approaches directly compare the predicted uncertainty to the observed error, providing a direct assessment of how well the uncertainty estimates correspond to actual prediction errors. In contrast, interval-based approaches examine how the observed errors are distributed within intervals of predicted uncertainties compared to ideal statistical behavior, such as evaluating whether approximately 68% of observations fall within $\pm 1\sigma$ of predictions as would be expected for a Gaussian distribution.

Below, we briefly discuss the specific evaluation metrics used in this work to assess the performance of our uncertainty quantification methods comprehensively.

3.4.1 Negative Log-Likelihood (NLL)

Negative log-likelihood–an error-based approach–is a standard metric used to evaluate the performance of uncertainty quantification models, which can also be used as a loss function to minimize during training. It measures the likelihood of observing the true values given the predicted distribution, where the underlying assumption is that the prediction errors are described by a Gaussian distribution with predicted variances (σ^2). NLL accounts for both the error and the variance to measure the quality of uncertainty quantification. A lower NLL indicates better uncertainty quantification performance.

$$\text{NLL} = \frac{1}{2N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \left[\ln(2\pi) + \ln(\sigma_i^2) + \frac{(y_i - \mu_i)^2}{\sigma_i^2} \right] \tag{7}$$

where μ_i and σ_i are the predicted mean and variance, respectively, and y_i is the true value.

For distributions with identical error magnitudes but different predicted variances, the uncertainties more representative of the actual error distribution will yield a lower NLL. It is important to note that a more accurate model (i.e., one having a lower mean absolute error) will generally result in a lower NLL, even if the uncertainties are poorly calibrated [39]. Furthermore, NLL is a relative metric, meaning its value is most meaningful when compared across different models. Due to these considerations, while a lower NLL indicates better uncertainty quantification performance, it is not sufficient in isolation to conclusively establish that a model is superior at quantifying uncertainty.

3.4.2 Expected Normalized Calibration Error (ENCE)

The Expected Normalized Calibration Error (ENCE)—an interval-based approach—quantifies the deviation of a model's calibration from ideal behavior in

the Root Mean Square Error (RMSE) versus Root Model Variance (RMV) plot. In traditional error-based calibration approaches, comparing models based on the area between the ideal calibration curve and the model's calibration curve can lead to inconsistent comparisons, as different models often exhibit varying ranges of RMSE and RMV values that are not bounded within a standardized interval such as [0, 1].

To facilitate meaningful cross-model comparison, ENCE calculates the relative deviation by taking the mean absolute difference between the RMSE and RMV for each bin, normalized by the RMV [40]:

$$\text{ENCE} = \frac{1}{N_{\text{bin}}} \sum_{i=1}^{N_{\text{bin}}} \frac{|\text{RMSE}_i - \text{RMV}_i|}{\text{RMV}_i} \tag{8}$$

where RMSE_i and RMV_i are the RMSE and RMV for the i-th bin, respectively, and N_{bin} is the total number of bins used in the calibration analysis.

When comparing multiple uncertainty quantification models, the one with lower ENCE is preferred as it demonstrates smaller deviations from the ideal calibration behavior in the RMSE versus RMV relationship. A perfectly calibrated model would have an ENCE of 0, indicating that the predicted variances (RMV) precisely match the observed squared errors (RMSE) across all uncertainty bins.

3.4.3 Miscalibration Area

The miscalibration area, another interval-based approach, quantifies the deviation between the expected and observed distributions of errors within confidence intervals. Specifically, it assesses how much the expected fraction of errors within each confidence interval differs from the observed fraction of errors that actually fall within those intervals.

This metric is computed as the absolute difference between the expected fraction of errors and the observed fraction of errors, effectively measuring the area between the diagonal line (representing perfect calibration) and the empirical calibration curve. The interpretation of this deviation has directional meaning: if the model's observed fraction of errors lies above the diagonal line, it indicates underconfident predictions (the model overestimates its uncertainty); conversely if the observed fraction lies below the diagonal, it indicates overconfident predictions (the model underestimates its uncertainty).

$$\text{Miscalibration Area} = \frac{1}{N_{\text{bin}}} \sum_{i=1}^{N_{\text{bin}}} |\text{Expected Fraction}_i - \text{Observed Fraction}_i| \tag{9}$$

where $\text{Expected Fraction}_i$ and $\text{Observed Fraction}_i$ are the expected and observed fractions of errors for the i-th confidence interval, respectively. This discrete summation approximates the continuous integral of differences across all possible confidence levels.

A smaller miscalibration area is desirable, as it indicates minimal deviation from ideal calibration. A perfectly calibrated model would have a miscalibration area of 0,

signifying that across all confidence intervals, the expected fractions precisely match the observed fractions of errors.

3.4.4 Spearman's Correlation Coefficient

Spearman's correlation coefficient (ρ) measures the monotonic relationship between predicted uncertainty and observed prediction errors. This error-based metric evaluates the model's ability to rank errors correctly from low to high values, regardless of the absolute magnitudes of the uncertainties.

The underlying premise is that data points with lower predicted uncertainties should generally correspond to lower observed errors, and conversely, higher predicted uncertainties should correlate with larger errors. Mathematically, Spearman's correlation coefficient is computed as the Pearson correlation between the rank variables of predicted uncertainties and observed absolute errors:

$$\rho = \frac{\text{cov}(R(|\hat{y} - y|), R(\sigma))}{\sigma_{R(|\hat{y}-y|)}\sigma_{R(\sigma)}} \tag{10}$$

where $R(|\hat{y}-y|)$ represents the ranks of absolute prediction errors, $R(\sigma)$ represents the ranks of predicted standard deviations, cov is the covariance, and $\sigma_{R(\cdot)}$ denotes the standard deviations of the rank variables.

While a higher ρ value (closer to 1) indicates better uncertainty quantification, it is important to note that perfect correlation ($\rho = 1$) is highly unlikely in practice. This is because even predictions with high uncertainty estimates can occasionally yield small errors due to the inherent stochasticity in the prediction process. A value of ρ substantially greater than zero indicates that the uncertainty estimates provide meaningful information about the expected magnitude of errors.

3.4.5 Coverage

Coverage at different confidence intervals quantifies the proportion of true values that fall within specified prediction intervals. This interval-based metric is computed as:

$$\text{Coverage} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \mathbf{1}_{y_i \in \text{CI}(x_i)} \tag{11}$$

where $\mathbf{1}_{y_i \in \text{CI}(x_i)}$ is an indicator function that equals one if the true value y_i falls within the confidence interval $\text{CI}(x_i)$ for input x_i, and zero otherwise. N_{test} represents the total number of test samples.

Coverage measures the model's empirical reliability by indicating the fraction of true values captured within the predicted confidence intervals [41]. For a well-calibrated model assuming Gaussian error distributions, approximately 68% of the data should fall within $\pm 1\sigma$ (one standard deviation) intervals, and approximately 95% should fall within $\pm 2\sigma$ intervals.

The deviation between the theoretical and observed coverage rates provides valuable insights into whether a model is overconfident (observed coverage lower than theoretical) or underconfident (observed coverage higher than theoretical). Ideally,

the observed coverage should closely match the theoretical coverage across various confidence levels, indicating proper uncertainty calibration.

3.4.6 ROC-AUC

The Area Under the Receiver Operating Characteristic curve (ROC-AUC) is an error-based approach that evaluates how well the predicted uncertainties discriminate between high and low prediction errors. This metric converts the continuous prediction errors into binary classifications based on a threshold value—in this work, we use the median of the observed errors as the threshold.

The ROC curve is constructed by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at various uncertainty thresholds. Specifically, for each possible threshold on the predicted uncertainty:

- Predictions with uncertainties above the threshold with errors above the error median are counted as true positives.
- Predictions with uncertainties above the threshold with errors below the error median are counted as false positives.

The AUC (Area Under Curve) value ranges from 0 to 1, where a value of 0.5 indicates random performance (uncertainties have no discriminative power for identifying large errors), and a value of 1 indicates perfect discrimination (uncertainties perfectly separate high and low errors). The ROC-AUC thus quantifies the model's ability to use predicted uncertainties as reliable indicators of potential prediction errors.

3.4.7 Reliability Diagram

A well-calibrated uncertainty quantification model should exhibit a one-to-one relationship between predicted variances and observed squared errors, meaning higher prediction errors should be reflected by correspondingly higher uncertainty estimates [42]. To obtain localized information about this relationship, the predicted uncertainties are ordered and divided into equal-sized bins. For each bin, the root mean square error (RMSE) and root mean variance (RMV) are computed as:

$$\text{RMSE} = \sqrt{\frac{1}{N_{\text{bin},i}} \sum_{j \in \text{bin}_i} (y_j - \mu_j)^2} \tag{12}$$

$$\text{RMV} = \sqrt{\frac{1}{N_{\text{bin},i}} \sum_{j \in \text{bin}_i} \sigma_j^2} \tag{13}$$

where $N_{\text{bin},i}$ represents the number of samples in the i-th bin, y_j is the true value, μ_j is the predicted mean, and σ_j is the predicted standard deviation for the j-th sample in that bin.

The reliability diagram plots RMSE against RMV for each bin. For a perfectly calibrated model, this plot should follow a straight line with a slope of 1 and an intercept of 0, indicating that the predicted uncertainties across all uncertainty levels accurately capture the magnitude of prediction errors.

3.4.8 Calibration Plot

The calibration plot provides a visual representation of how well the uncertainty estimates constitute the true correctness likelihood of predictions [43]. This plot compares the observed fraction of errors against the expected fraction of errors within specified confidence intervals.

To construct this plot, prediction errors are first normalized into z-scores using the predicted uncertainties:

$$z_i = \frac{y_i - \mu_i}{\sigma_i} \tag{14}$$

where y_i is the true value, μ_i is the predicted mean, and σ_i is the predicted standard deviation for the i-th sample.

The expected fraction of errors is then computed as the cumulative distribution function (CDF) of the standard normal distribution evaluated at different confidence levels:

$$\text{Expected Fraction of Errors} = \Phi(z) \tag{15}$$

Where Φ represents the CDF of the standard normal distribution.

For a well-calibrated model, the calibration plot should closely follow the diagonal line, indicating that the observed fractions of errors align with the theoretically expected fractions across all confidence levels. Deviations above the diagonal suggest underconfidence (the model overestimates uncertainty), while deviations below indicate overconfidence (the model underestimates uncertainty).

4 Conclusion

In summary, we have developed a scalable and efficient uncertainty quantification framework for adsorption energy prediction by combining the representation power of graph kernels with the uncertainty quantification capabilities of Sparse Variational Gaussian Processes. Our comprehensive evaluation across multiple model architectures (PaiNN, SchNet) and datasets (CatHub, OC20) demonstrates that DGKL consistently provides well-calibrated uncertainty estimates, as evidenced by superior performance in calibration metrics. For instance, DGKL achieved the lowest ENCE (0.07–0.10) and miscalibration area (0.04–0.07) across all tested configurations while maintaining strong ranking capabilities (Spearman correlation coefficient up to 0.51 and ROC AUC up to 0.74). The exceptional R^2 values ($\sim$1.00) for RMSE vs RMV plots, with slopes closely approaching the ideal value of 1.0 (e.g., 1.05 ± 0.10 for CatHub-SchNet and 1.02 ± 0.11 for OC20-SchNet), along with empirical coverage rates closely matching theoretical expectations (68.27% and 95.45% for 1σ and 2σ levels, respectively), further confirm DGKL's reliability.

In comparison to alternative methods, DGKL offers more consistent uncertainty quantification while maintaining competitive computational efficiency. Although ensemble/evidential/mcd methods - owing to the underlying graph backbone and accuracy-focused objective/loss function - achieve lower MAE, they generally show

Table 1: Comparison of Predictive Performance and Uncertainty Quantification Metrics Across Different Methods and Model Architectures, Averaged over five different runs with different subsets of data for training and testing. Reported metrics are for test set predictions.

	Method	MAE↓ [↓]	NLL↓ [↓]	ENCE↓ [↓]	Miscal. Area↓ [↓]	ρ_{SCC}↑ [↑]	ROC AUC↑ [↑]
CatHub PaiNN	DGKL	0.31 ± 0.07	0.36 ± 0.18	0.10 ± 0.05	0.06 ± 0.03	**0.48 ± 0.03**	**0.73 ± 0.02**
	DGKL Atomic	0.30 ± 0.02	0.34 ± 0.07	**0.10 ± 0.03**	**0.05 ± 0.02**	0.44 ± 0.02	0.71 ± 0.01
	Ensemble	0.15 ± 0.09	**-0.11 ± 0.31**	0.41 ± 0.21	0.21 ± 0.13	0.47 ± 0.07	0.73 ± 0.03
	Evidential	**0.13 ± 0.02**	0.31 ± 0.70	0.49 ± 0.26	0.37 ± 0.22	0.26 ± 0.05	0.62 ± 0.03
	MCD	0.16 ± 0.04	5.29 ± 2.17	2.30 ± 0.51	1.23 ± 0.13	0.15 ± 0.02	0.58 ± 0.01
CatHub SchNet	DGKL	0.43 ± 0.04	0.68 ± 0.09	**0.07 ± 0.04**	**0.04 ± 0.02**	**0.51 ± 0.02**	**0.74 ± 0.01**
	DGKL Atomic	0.33 ± 0.06	0.42 ± 0.16	0.10 ± 0.05	0.05 ± 0.02	0.46 ± 0.01	0.72 ± 0.01
	Ensemble	0.16 ± 0.07	**-0.06 ± 0.37**	0.26 ± 0.10	0.12 ± 0.05	0.41 ± 0.10	0.70 ± 0.05
	Evidential	**0.13 ± 0.01**	0.59 ± 0.68	0.79 ± 0.43	0.46 ± 0.22	0.30 ± 0.08	0.64 ± 0.04
	MCD	0.16 ± 0.05	4.50 ± 3.10	1.94 ± 0.59	0.91 ± 0.24	0.18 ± 0.02	0.59 ± 0.01
OC20 SchNet	DGKL	0.69 ± 0.01	1.30 ± 0.02	**0.06 ± 0.04**	**0.07 ± 0.02**	0.25 ± 0.02	0.62 ± 0.01
	DGKL Atomic	0.57 ± 0.02	**1.07 ± 0.03**	0.08 ± 0.05	0.09 ± 0.05	**0.34 ± 0.02**	**0.66 ± 0.01**
	Ensemble	**0.48 ± 0.03**	4.96 ± 1.19	1.95 ± 0.34	1.07 ± 0.24	0.19 ± 0.01	0.59 ± 0.01
	Evidential	0.50 ± 0.01	1.78 ± 0.30	0.65 ± 0.10	0.52 ± 0.08	0.08 ± 0.14	0.54 ± 0.07
	MCD	0.51 ± 0.01	34.07 ± 16.18	6.19 ± 1.69	2.54 ± 1.26	0.15 ± 0.02	0.57 ± 0.01

Notes:

[↑]: Higher is better. [↓]: Lower is better.

Bold values indicate the best method for each Dataset-GNN combination.

[*] SCC = Spearman Correlation Coefficient

Table 2: Calibration Metrics for RMSE vs. RMV Plot Along with Empirical Coverage at 1σ and 2σ Level, Averaged over five different runs with different subsets of data for training and testing. Reported metrics are for test set predictions.

	Method	$\mathbf{R^2}$ [↑]	**Slope** [→ 1]	**Intercept** [→ 0]	$\mathbf{1\sigma}$ **Coverage** [→ 0.68]	$\mathbf{2\sigma}$ **Coverage** [→ 0.95]
CatHub PaiNN	DGKL	1.00 ± 0.00	1.22 ± 0.18	-0.06 ± 0.06	0.70 ± 0.03	0.94 ± 0.02
	DGKL Atomic	1.00 ± 0.00	1.28 ± 0.05	-0.06 ± 0.02	0.69 ± 0.03	0.94 ± 0.01
	Ensemble	1.00 ± 0.00	0.95 ± 0.41	0.04 ± 0.03	0.68 ± 0.11	0.90 ± 0.06
	Evidential	0.93 ± 0.02	0.95 ± 0.91	-0.04 ± 0.18	0.87 ± 0.13	0.96 ± 0.04
	MCD	0.92 ± 0.01	0.76 ± 0.31	0.16 ± 0.03	0.37 ± 0.05	0.62 ± 0.04
CatHub SchNet	DGKL	1.00 ± 0.00	1.05 ± 0.10	0.01 ± 0.02	0.68 ± 0.02	0.95 ± 0.01
	DGKL Atomic	1.00 ± 0.00	1.32 ± 0.20	-0.08 ± 0.02	0.69 ± 0.03	0.94 ± 0.01
	Ensemble	0.99 ± 0.01	0.88 ± 0.14	0.04 ± 0.03	0.66 ± 0.02	0.91 ± 0.03
	Evidential	0.98 ± 0.01	1.70 ± 1.51	-0.01 ± 0.09	0.70 ± 0.22	0.88 ± 0.10
	MCD	0.95 ± 0.02	1.00 ± 0.28	0.14 ± 0.05	0.42 ± 0.05	0.66 ± 0.06
OC20 SchNet	DGKL	0.99 ± 0.01	1.02 ± 0.11	0.01 ± 0.05	0.72 ± 0.03	0.94 ± 0.01
	DGKL Atomic	1.00 ± 0.00	1.14 ± 0.08	-0.14 ± 0.07	0.74 ± 0.04	0.96 ± 0.01
	Ensemble	0.99 ± 0.01	1.12 ± 0.18	0.37 ± 0.04	0.36 ± 0.04	0.62 ± 0.05
	Evidential	0.59 ± 0.33	0.07 ± 0.20	0.58 ± 0.39	0.97 ± 0.02	1.00 ± 0.00
	MCD	0.98 ± 0.01	2.23 ± 1.17	0.45 ± 0.08	0.16 ± 0.03	0.31 ± 0.06

Notes:

[↑]: Higher is better. [↓]: Lower is better.

[→ 1]: Closer to 1 is better. [→ 0]: Closer to 0 is better.

[→ 0.68]/[→ 0.95]: Closer to theoretical coverage (68.27%/95.45%) is better.

Model	OOD-cat	OOD-ads	OOD-both
DGKL	0.600	0.566	0.603
Ensemble	0.790	0.753	0.732
DGKL Atomic	0.876	0.876	0.837
DGKL Atomic (Atomic)	0.791	0.787	0.782

Table 3: ROC-AUC scores for OOD detection using uncertainty values. Higher values indicate better OOD detection performance.

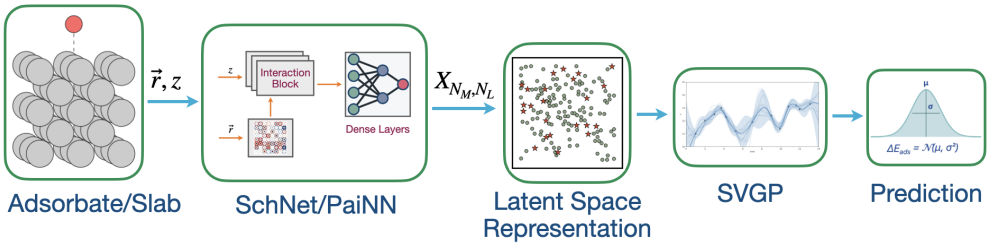

Fig. 1: Adsorption energy prediction and uncertainty quantification using deep graph kernel learning (DGKL). The workflow starts with an adsorbate/slab model, processes the atomic structure through SchNet/PaiNN neural networks, creates a latent space representation with inducing points, applies Sparse Variational Gaussian Process (SVGP) regression, and produces final energy predictions as a normal distribution $\Delta E_{\text{ads}} = \mathcal{N}(\mu, \sigma^2)$. For standard DGKL, the SVGP input tensor dimension is $N_M \times N_L$ where N_M is the number of materials samples, and N_L is the latent space dimension. In contrast, for DGKL Atomic, the input tensor dimension is N_A (number of atoms), which are then pooled after SVGP processing. In the latent space representation, the olive green circles show the training samples, and the red star denotes the location of the inducing point. Note that the visualization of the latent space is provided for illustrative purposes only and is a simplified 2D projection of the high-dimensional latent space.

poorer calibration metrics, suggesting their uncertainty estimates are less reliable. Furthermore, our DGKL Atomic variant provides unprecedented atom-level uncertainty quantification, offering deeper insights into prediction reliability at the molecular structural level while maintaining performance comparable to global DGKL.

Due to its simplicity and scalability, DGKL can be readily extended to other applications in materials property prediction, design, and optimization. Particularly, DGKL will be valuable for active learning workflows in catalyst discovery, where its reliable uncertainty estimates enable rational decision-making in high-throughput screening, potentially reducing experimental validation costs and accelerating the identification of promising catalyst candidates. Our work establishes DGKL as a robust and practical

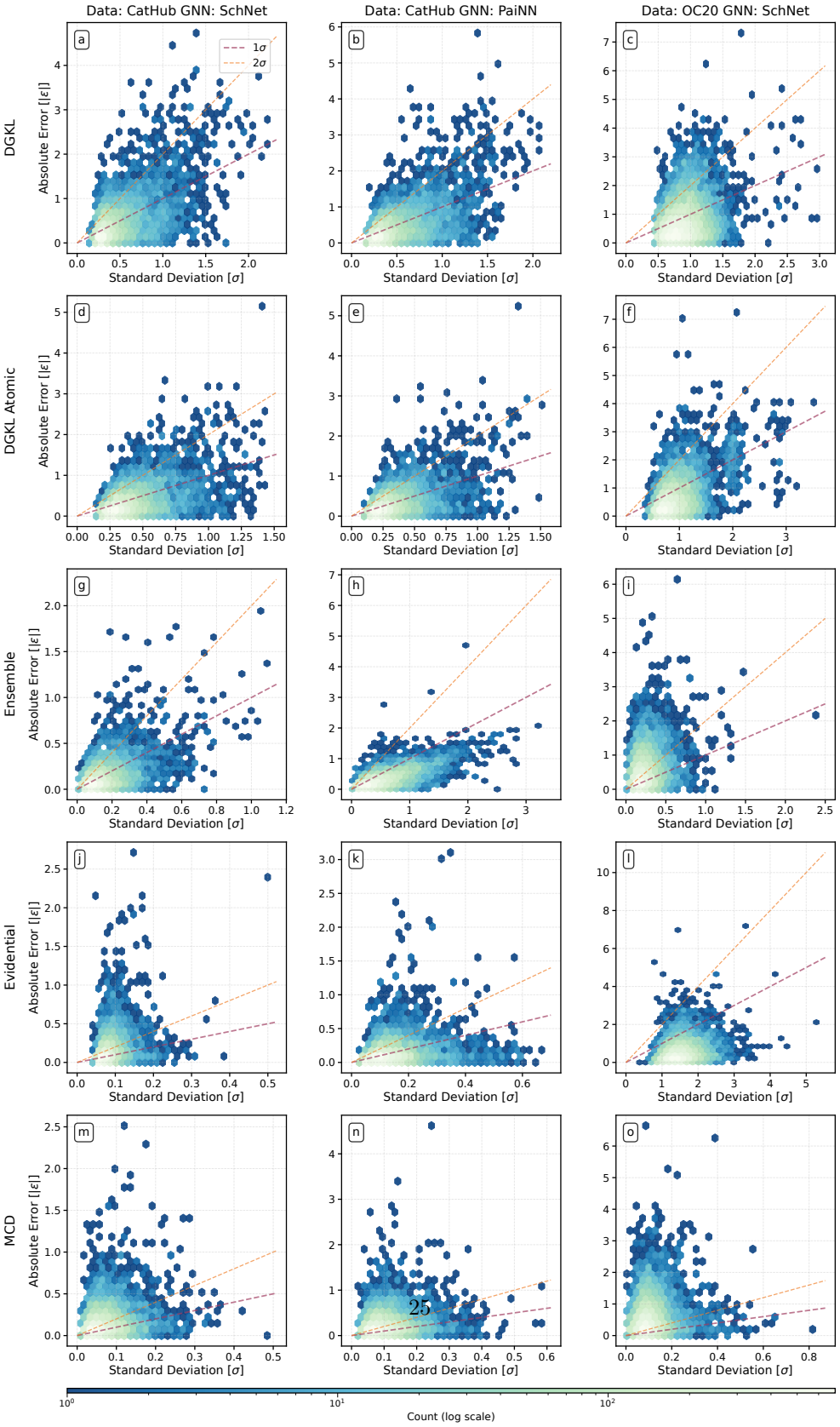

Fig. 2: Uncertainty calibration analysis for different Graph Neural Network models and uncertainty quantification methods. Each subplot shows the relationship between the standard deviation of model predictions (σ) and absolute error ($|\varepsilon|$). The analysis compares four uncertainty quantification methods (rows): DGKL, DGKL Atomic, Ensemble, Evidential, and MCD (Monte Carlo Dropout) across three different datasets/model combinations (columns): CatHub with SchNet, CatHub with PaiNN, and OC20 with SchNet. The dashed purple and orange lines represent the 1σ and 2σ calibration targets, respectively. The color intensity indicates the data point density on a logarithmic scale.

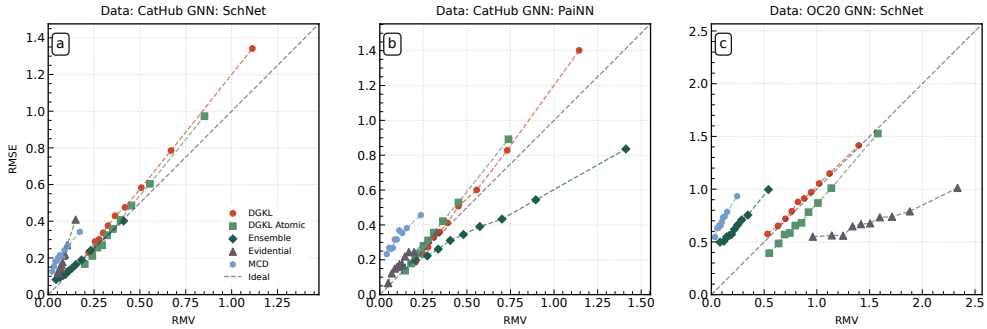

Fig. 3: Comparison of uncertainty estimation methods across different Graph Neural Network architectures and datasets. The plots show the relationship between Root Mean-Variance (RMV) and Root Mean Square Error (RMSE) for five uncertainty quantification methods: DGKL, DGKL Atomic, Ensemble, Evidential, and MCD (Monte Carlo Dropout). Results are presented for three dataset-model combinations: (a) CatHub with SchNet, (b) CatHub with PaiNN, and (c) OC20 with SchNet. The dashed gray diagonal line indicates the ideal calibration where RMSE equals RMV. Points above the line represent underconfident predictions, while points below indicate overconfident predictions. Methods with curves closer to the ideal line demonstrate better uncertainty calibration.

framework for uncertainty quantification in computational catalysis and materials science, addressing a critical need for reliable confidence estimates alongside predictions at reasonable computational cost.

5 Acknowledgements

Financial support for this publication results from Scialog grant #SA-SM3-2024-059b from the Research Corporation for Science Advancement and Alfred P. Sloan Foundation. This work was performed using compute resources from the Cornell University Center for Advanced Computing (CAC) and San Diego Supercomputer Center through allocation CHM220019 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program.

6 Ethics Declarations

6.1 Competing interests

The authors declare no competing interests.

6.2 Data availability

The datasets used in this study are available from the Catalysis-Hub.org and opencatalystproject.org.

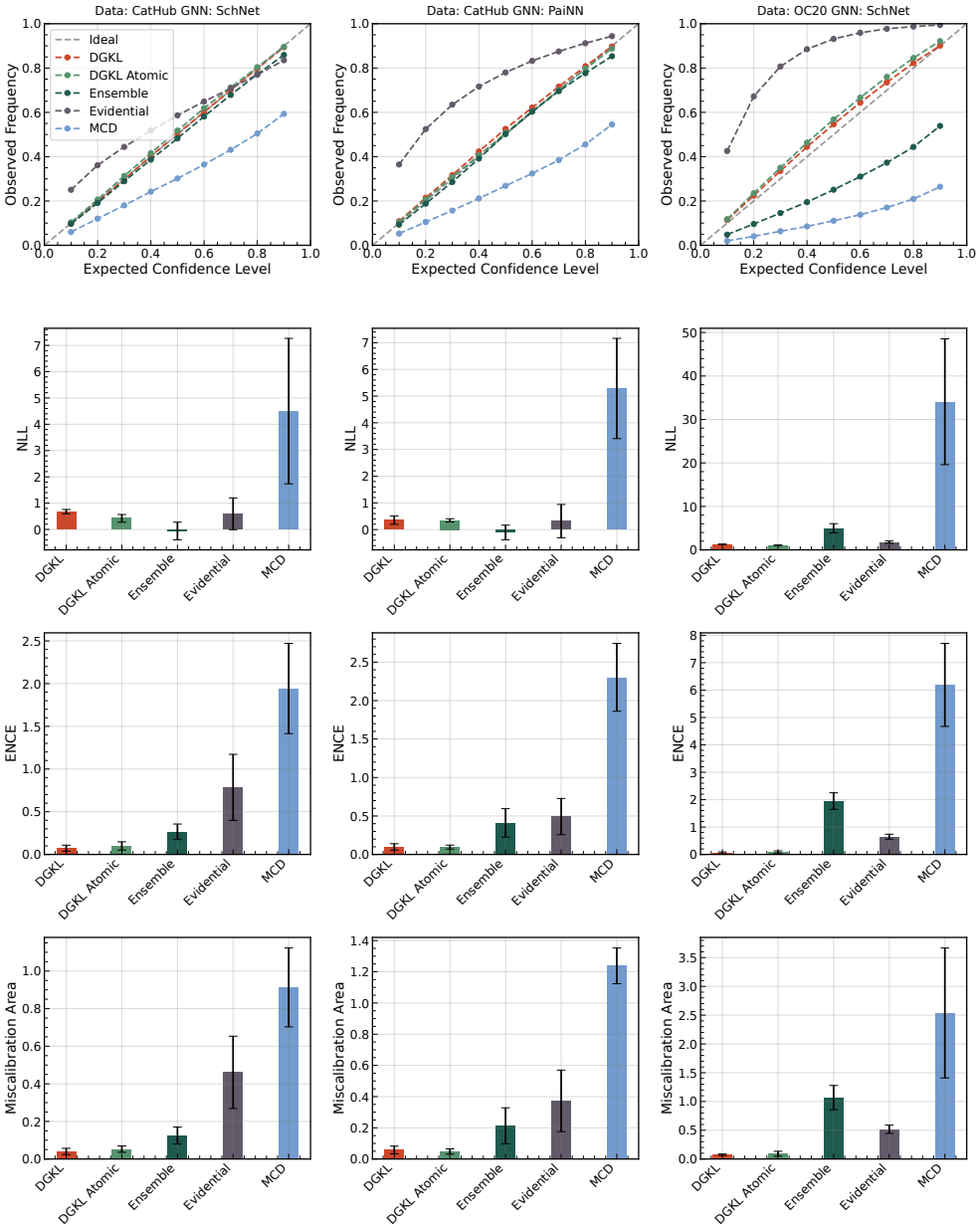

Fig. 4: Comprehensive evaluation of uncertainty calibration across five methods (DGKL, DGKL Atomic, Ensemble, Evidential, and MCD) for three dataset-model combinations (CatHub-SchNet, CatHub-PaiNN, and OC20-SchNet). **First row:** Calibration curves showing observed frequency vs. expected confidence level. The diagonal dashed line represents perfect calibration. Curves above the line indicate underconfidence, while curves below indicate overconfidence. **Second row:** Negative Log-Likelihood (NLL) scores with error bars. Lower values indicate better probabilistic predictions. **Third row:** Expected Normalized Calibration Error (ENCE) with error bars. Lower values indicate better calibration of uncertainty estimates. **Fourth row:** Miscalibration Area with error bars, measuring the area between the calibration curve and the ideal diagonal. Lower values indicate better calibration. Across all metrics and datasets, DGKL and DGKL Atomic consistently show superior calibration performance, while MCD typically exhibits the poorest calibration with highest error metrics and significant underconfidence.

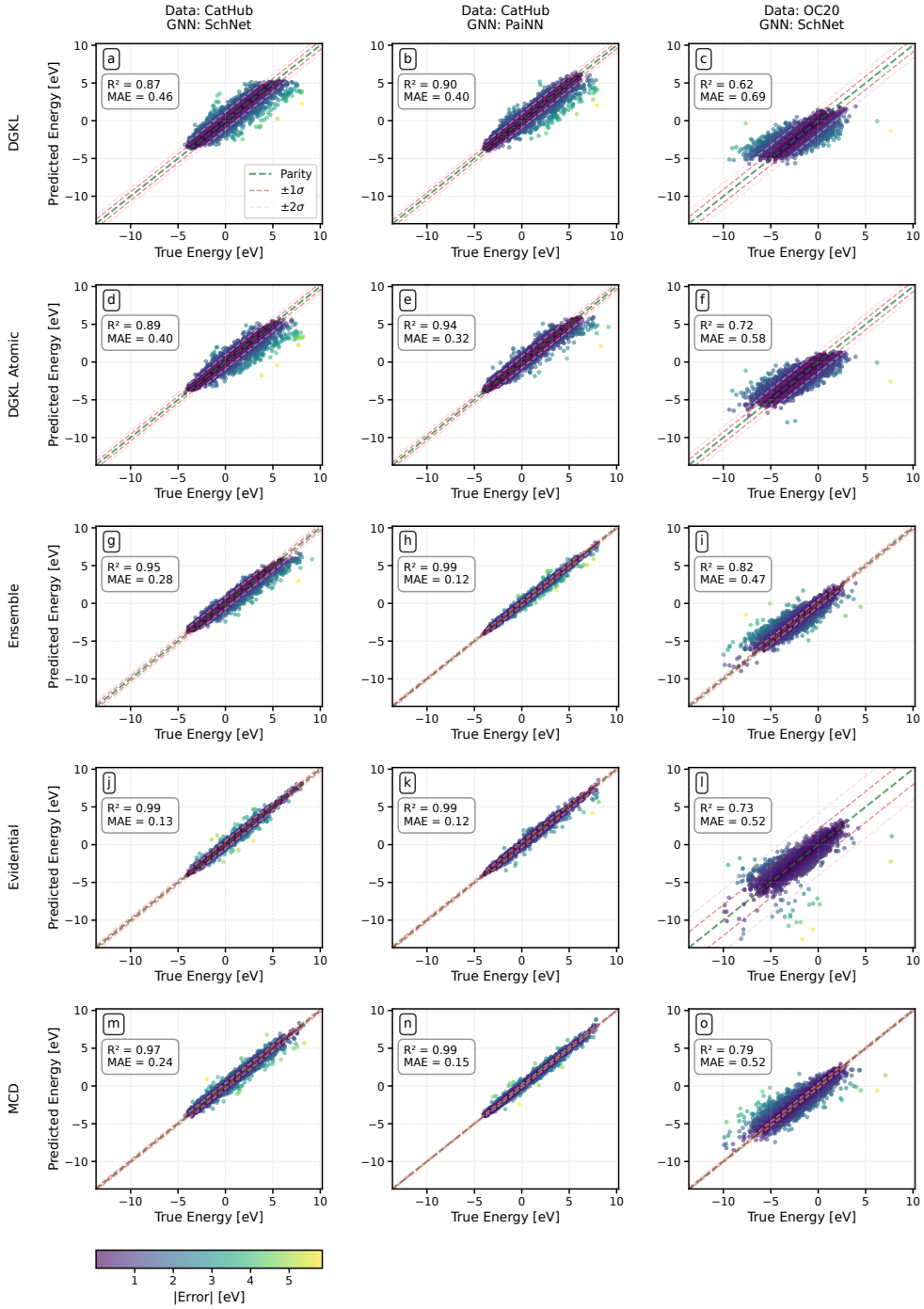

Fig. 5: Performance comparison of different uncertainty quantification methods for Graph Neural Networks (GNNs) in predicting adsorption energies. The figure shows parity plots of predicted vs. true energies (in eV) across different datasets (CatHub and OC20) and GNN architectures (SchNet and PaiNN). Each row represents a different uncertainty quantification approach: (a-c) standard DGKL, (d-f) DGKL Atomic, (g-i) Ensemble method, (j-l) Evidential approach, and (m-o) Monte Carlo Dropout (MCD). Performance metrics include the coefficient of determination (R^2) and mean absolute error (MAE). Dashed lines indicate parity (perfect prediction) and $\pm 1\sigma$, $\pm 2\sigma$ confidence intervals. The color bar at the bottom represents the absolute error magnitude, with darker colors indicating lower errors.

6.3 Code availability

The code used in this study is available from the GitHub repository: https://github.com/mamunm/DGKL

References

[1] Zitnick, C.L., Chanussot, L., Das, A., Goyal, S., Heras-Domingo, J., Ho, C., Hu, W., Lavril, T., Palizhati, A., Riviere, M., Shuaibi, M., Sriram, A., Tran, K., Wood, B., Yoon, J., Parikh, D., Ulissi, Z.: An Introduction to Electrocatalyst Design using Machine Learning for Renewable Energy Storage (2020). https://arxiv.org/abs/2010.09435

[2] Mamun, O., Winther, K.T., Boes, J.R., Bligaard, T.: A bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. npj Computational Materials **6**(1), 177 (2020) https://doi.org/10.1038/s41524-020-00447-8

[3] Nørskov, J.K., Bligaard, T.: The catalyst genome. Angewandte Chemie International Edition **52**(3), 776–777 (2013) https://doi.org/10.1002/anie.201208487 https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201208487

[4] Sholl, D.S., Steckel, J.A.: Density Functional Theory: a Practical Introduction. John Wiley & Sons, ??? (2022)

[5] Ulissi, Z.W., Medford, A.J., Bligaard, T., Nørskov, J.K.: To address surface reaction network complexity using scaling relations machine learning and dft calculations. Nature Communications **8**(1), 14621 (2017) https://doi.org/10.1038/ncomms14621

[6] Jiao, Z., Liu, Y., Wang, Z.: Application of graph neural network in computational heterogeneous catalysis. The Journal of Chemical Physics **161**(17) (2024)

[7] Jiao, Z., Mao, Y., Lu, R., Liu, Y., Guo, L., Wang, Z.: Fine-tuning graph neural networks via active learning: Unlocking the potential of graph neural networks trained on nonaqueous systems for aqueous co2 reduction. Journal of Chemical Theory and Computation

[8] Tan, A.R., Urata, S., Goldman, S., Dietschreit, J.C., Gómez-Bombarelli, R.: Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. npj Computational Materials **9**(1), 225 (2023)

[9] Du, Y.-W., Zhong, J.-J.: Generalized combination rule for evidential reasoning approach and dempster–shafer theory of evidence. Information Sciences **547**, 1201–1232 (2021) https://doi.org/10.1016/j.ins.2020.07.072

[10] Ober, S.W., Rasmussen, C.E., Wilk, M.: The Promises and Pitfalls of Deep Kernel

Learning (2021). https://arxiv.org/abs/2102.12108

[11] Burt, D., Rasmussen, C.E., Van Der Wilk, M.: Rates of convergence for sparse variational gaussian process regression. In: International Conference on Machine Learning, pp. 862–871 (2019). PMLR

[12] Leibfried, F., Dutordoir, V., John, S., Durrande, N.: A tutorial on sparse gaussian processes and variational inference. arXiv preprint arXiv:2012.13962 (2020)

[13] Charpentier, B., Zügner, D., Günnemann, S.: Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts (2020). https://arxiv.org/abs/2006.09239

[14] Kopetzki, A.-K., Charpentier, B., Zügner, D., Giri, S., Günnemann, S.: Evaluating Robustness of Predictive Uncertainty Estimation: Are Dirichlet-based Models Reliable? (2021). https://arxiv.org/abs/2010.14986

[15] Wollschläger, T., Gao, N., Charpentier, B., Ketata, M.A., Günnemann, S.: Uncertainty Estimation for Molecules: Desiderata and Methods (2023). https://arxiv.org/abs/2306.14916

[16] De, S., Bartók, A.P., Csányi, G., Ceriotti, M.: Comparing molecules and solids across structural and alchemical space. Phys. Chem. Chem. Phys. **18**, 13754–13769 (2016) https://doi.org/10.1039/C6CP00415F

[17] Bartók, A.P., Kondor, R., Csányi, G.: On representing chemical environments. Phys. Rev. B **87**, 184115 (2013) https://doi.org/10.1103/PhysRevB.87.184115

[18] Schütt, K.T., Kindermans, P.-J., Sauceda, H.E., Chmiela, S., Tkatchenko, A., Müller, K.-R.: SchNet: A continuous-filter convolutional neural network for modeling quantum interactions (2017). https://arxiv.org/abs/1706.08566

[19] Schütt, K.T., Unke, O.T., Gastegger, M.: Equivariant message passing for the prediction of tensorial properties and molecular spectra (2021). https://arxiv.org/abs/2102.03150

[20] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J.V., Lakshminarayanan, B., Snoek, J.: Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift (2019). https://arxiv.org/abs/1906.02530

[21] Li, Y., Kong, L., Du, Y., Yu, Y., Zhuang, Y., Mu, W., Zhang, C.: MUBen: Benchmarking the Uncertainty of Molecular Representation Models (2024). https://arxiv.org/abs/2306.10060

[22] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In:

Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc., ??? (2017). https://proceedings.neurips.cc/paper$_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf$

[23] Dietterich, T.G.: Ensemble Methods in Machine Learning, pp. 1–15. Springer, Berlin, Heidelberg (2000)

[24] Fort, S., Hu, H., Lakshminarayanan, B.: Deep Ensembles: A Loss Landscape Perspective (2020). https://arxiv.org/abs/1912.02757

[25] Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A Research Platform for Distributed Model Selection and Training (2018). https://arxiv.org/abs/1807.05118

[26] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (2016). https://proceedings.mlr.press/v48/gal16.html

[27] Soleimany, A.P., Amini, A., Goldman, S., Rus, D., Bhatia, S.N., Coley, C.W.: Evidential deep learning for guided molecular property prediction and discovery. ACS Central Science **7**(8), 1356–1367 (2021) https://doi.org/10.1021/acscentsci.1c00546

[28] Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep Evidential Uncertainty (2020). https://openreview.net/forum?id=S1eSoeSYwr

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

[30] Bishop, C.M., Nasrabadi, N.M.: Pattern Recognition and Machine Learning vol. 4. Springer, ??? (2006)

[31] Jakkala, K.: Deep Gaussian Processes: A Survey (2021). https://arxiv.org/abs/2106.12135

[32] Rasmussen, C.E., Williams, C.K., et al.: Gaussian processes for machine learning, vol. 1. MIT press Cambridge MA (2006)

[33] Titsias, M.: Variational learning of inducing variables in sparse gaussian processes. In: Dyk, D., Welling, M. (eds.) Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 5, pp. 567–574. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA (2009). https://proceedings.mlr.press/v5/titsias09a.html

[34] Hensman, J., Fusi, N., Lawrence, N.D.: Gaussian Processes for Big Data (2013). https://arxiv.org/abs/1309.6835

[35] Winther, K.T., Hoffmann, M.J., Boes, J.R., Mamun, O., Bajdich, M., Bligaard, T.: Catalysis-hub.org, an open electronic structure database for surface reactions. Scientific Data **6**(1), 75 (2019) https://doi.org/10.1038/s41597-019-0081-y

[36] Mamun, O., Winther, K.T., Boes, J.R., Bligaard, T.: High-throughput calculations of catalytic properties of bimetallic alloy surfaces. Scientific Data **6**(1), 76 (2019) https://doi.org/10.1038/s41597-019-0080-z

[37] Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C.L., Ulissi, Z.: Open catalyst 2020 (oc20) dataset and community challenges. ACS Catalysis **11**(10), 6059–6072 (2021) https://doi.org/10.1021/acscatal.0c04525

[38] Gruich, C.J., Madhavan, V., Wang, Y., Goldsmith, B.R.: Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. Machine Learning: Science and Technology **4**(2), 025019 (2023) https://doi.org/10.1088/2632-2153/accace

[39] Rasmussen, M.H., Duan, C., Kulik, H.J., Jensen, J.H.: Uncertain of uncertainties? a comparison of uncertainty quantification metrics for chemical data sets. Journal of Cheminformatics **15**(1), 121 (2023) https://doi.org/10.1186/s13321-023-00790-0

[40] Dai, J., Adhikari, S., Wen, M.: Uncertainty quantification and propagation in atomistic machine learning. Reviews in Chemical Engineering (2024) https://doi.org/10.1515/revce-2024-0028

[41] Kompa, B., Snoek, J., Beam, A.L.: Empirical frequentist coverage of deep learning uncertainty quantification procedures. Entropy **23**(12) (2021) https://doi.org/10.3390/e23121608

[42] Levi, D., Gispan, L., Giladi, N., Fetaya, E.: Evaluating and calibrating uncertainty prediction in regression tasks. Sensors **22**(15) (2022) https://doi.org/10.3390/s22155540

[43] Gneiting, T., Katzfuss, M.: Probabilistic forecasting. Annual Review of Statistics and Its Application **1**(Volume 1, 2014), 125–151 (2014) https://doi.org/10.1146/annurev-statistics-062713-085831

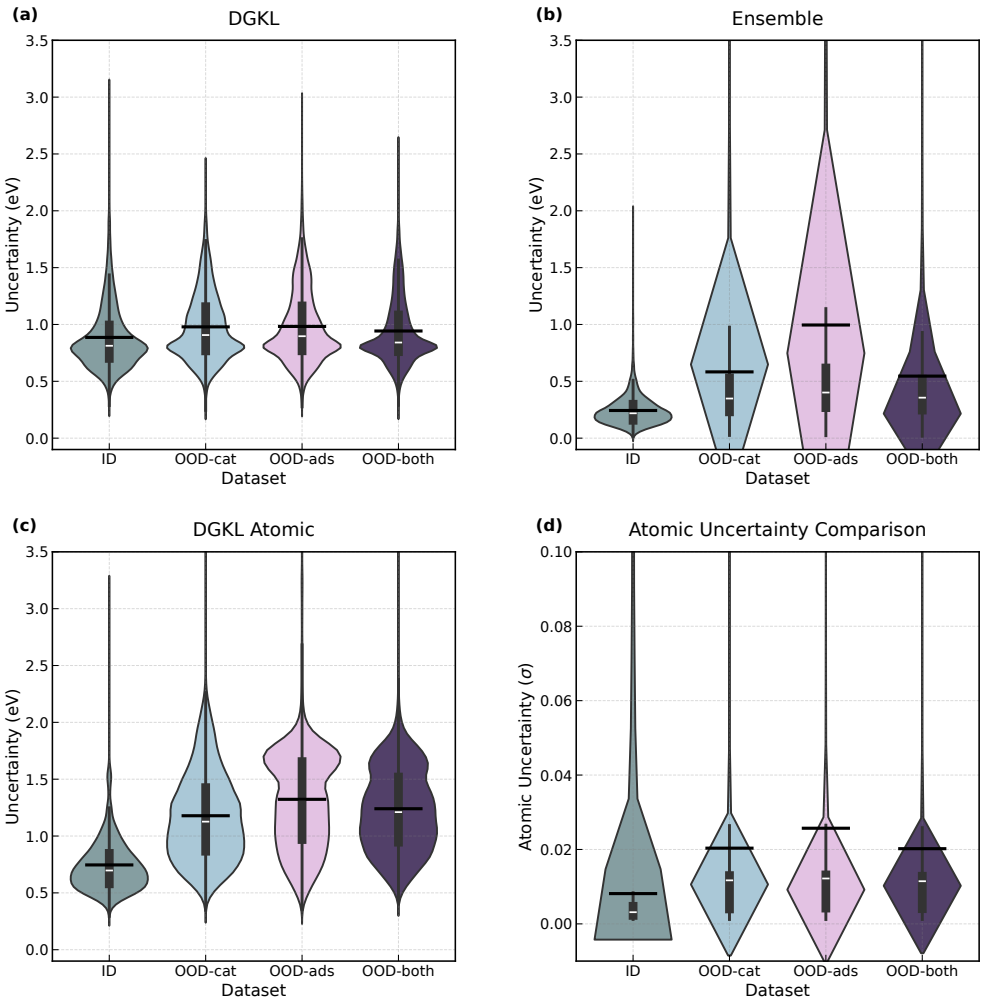

Fig. 6: Comparison of uncertainty measurements (σ) across different datasets and methods. Panel (a) shows uncertainty distributions for the DGKL method, (b) shows results for the Ensemble method, (c) presents DGKL Atomic uncertainty measurements, and (d) displays atomic uncertainty comparisons for ID and OOD data. The violin plots illustrate the distribution of uncertainty values, with box plots overlaid to highlight median, quartiles, and whiskers. We also add a black horizontal line to denote the mean of the distribution.