Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d6dd00020g

Deep graph kernel learning for material & atomic level uncertainty quantification in adsorption energy prediction

Osman Mamun, [a] Chenlu Yang [b] and Shuwen Yue *[a]

Graph neural networks (GNNs) have emerged as powerful and efficient surrogates for computationally intensive density functional theory calculations and have greatly accelerated catalytic material discovery. However, their practical utility is often constrained by poor reliability and, critically, generalization to OOD chemical space. To address these challenges, we develop a deep graph kernel learning (DGKL) model, a scalable and versatile framework that integrates GNN backbones with the rigorous uncertainty quantification (UQ) of sparse variational Gaussian processes (SVGPs) for adsorption energy prediction. Compared to ensemble, evidential, and Monte Carlo dropout methods, DGKL consistently delivers better-calibrated uncertainties with competitive accuracy and fast inference. Across two benchmarks (CatHub and OC20) and two GNN backbones (SchNet and PaiNN), DGKL achieves superior calibration metrics (expected normalized calibration error: 0.06–0.10, miscalibration area: 0.04–0.07), strong error–uncertainty correlation (Spearman coefficient up to 0.51), and stable probabilistic fits. Furthermore, we present DGKL-atomic, which provides atomic level UQ critical for controlling active learning sampling towards desired region of the chemical space. DGKL-atomic shows excellent performance in detecting out-of-distribution (OOD) samples (ROC–AUC: 0.84–0.88), and its atomic uncertainties correlate well with local structural novelty. Together, these methods enable calibrated, granular, and computationally efficient UQ to enhance active learning workflows and accelerating catalyst discovery.

Received 17th January 2026
Accepted 23rd March 2026

DOI: 10.1039/d6dd00020g

rsc.li/digitaldiscovery

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1 Introduction

Machine learning (ML) models trained on density functional theory (DFT)-generated datasets have emerged as promising alternatives to direct DFT calculations for high-throughput materials screening.[1] Among these, Graph Neural Networks (GNNs) have established themselves as state-of-the-art approaches for accelerating computational materials discovery.[2,3] While GNNs demonstrate excellent performance on test data that conform to the training data distributions, their reliability often diminishes when applied to out-of-distribution (OOD) 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.[4] Ensemble approaches estimate uncertainty by aggregating predictions from multiple independently trained models, with increasing disagreement serving as a proxy for uncertainty. Monte Carlo dropout samples network architectures by stochastically deactivating network weights during training. Evidential models, based on subjective logic and Dempster–Shafer theory, directly predict probability distribution parameters, enabling second-order uncertainty estimation that distinguishes epistemic from aleatoric uncertainty without multiple forward passes.[5]

Gaussian processes (GPs) represent another established approach for Bayesian modeling, offering interpretability, reliable uncertainty estimates, and a principled probabilistic framework.[6] However, traditional GP implementations face several limitations when applied to molecular and materials science dataset. Firstly, they lack native capabilities to process molecular graph data and therefore rely on molecular representation algorithms such as atom center symmetry function (ACSF) or smooth overlap of atomic orbitals (SOAP). Second, their high computational and memory costs typically restrict their application to relatively small datasets, which limits their applicability to existing large-scale molecular datasets. To build an end-to-end solution without requiring an intermediate representation algorithm, deep graph kernel learning (DGKL) has been applied successfully to several scientific studies. In DGKL, a GNN learns meaningful latent representations directly, and GP operates on this learned embedding space to provide probabilistic predictions and uncertainty estimates. Even though DGKL solves the representation learning problem, the

[a]R. F. Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, USA. E-mail: shuwen.yue@cornell.edu
[b]Department of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853, USA

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

GP formulation still requires storing the entire latent space to be in memory to construct the kernel matrix, resulting in prohibitive memory costs at scale. Furthermore, during end-to-end training, the latent representation from the GNN evolves continuously, which complicates the kernel construction and can disrupt gradient backpropagation through the GP component. Sparse variational Gaussian process (SVGP)[7,8] addresses this issue by approximating the latent space with a smaller, compact 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, training DGKL models remains challenging in practice, primarily due to potential mode collapse and conflicting optimization dynamics between the GNN backbone and SVGP components. These challenges can be partially 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 DGKL model (Fig. 1) for uncertainty quantification of adsorption energy prediction by integrating GNN architectures with SVGP. 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 for enhanced catalyst discovery. 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, and in particular DGKL-atomic, maintains well-calibrated uncertainty estimates even for OOD predictions, whereas traditional methods such as Ensemble and Monte Carlo dropout often exhibit inconsistent uncertainty calibration in challenging scenarios with novel catalyst materials and adsorbate combinations. DGKL-atomic can be promising for enhancing 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.

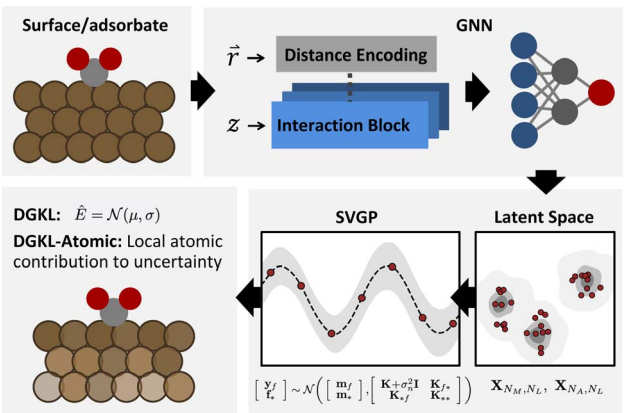

Fig. 1 Schematic workflow of the deep graph kernel learning framework showing both standard DGKL and DGKL-atomic approaches for adsorption energy prediction and uncertainty quantification. (top left) The adsorbate/slab catalytic system provides atomic coordinates ($\vec{r}$) and atomic numbers (z) as input to the framework. (top right) The GNN backbone processes these inputs through interaction blocks and distance encoding to extract latent representations with two pathways: material-level latent features ($X_{NM,NL}$) for standard DGKL and atom-level latent features ($X_{NA,NL}$) for DGKL-atomic. (bottom right) The SVGP processes the latent representations to predict energy contributions and uncertainties, yielding $\hat{E} = \mathcal{N}(\mu, \sigma)$. (bottom left) For DGKL-atomic, atomic uncertainties are aggregated back to the material level using $\mu_{material} = \sum \mu_{atomic,i}$ and ${\sigma_{material}}^2 \approx \sum {\sigma_{atomic,i}}^2$ to produce the final adsorption energy prediction with quantified uncertainty. N_M, N_A, and N_L represent the number of material structures, the total number of atoms, and the latent space dimension, respectively. Atom level uncertainty magnitude is represented by the color gradient on each atom.

2 Results & discussion

Here we present a comprehensive comparative evaluation of both DGKL and DGKL-atomic against established UQ methods for adsorption energy uncertainty prediction across three key dimensions: (i) predictive performance, (ii) uncertainty calibration and reliability, and (iii) computational efficiency. The evaluation proceeds in three stages: we establish the benchmarking framework, analyze comparative performance across methods along the three aforementioned axes, and demonstrate DGKL-atomic's enhanced out-of-distribution detection capabilities.

2.1 Benchmarking framework

2.1.1 Databases. We benchmark DGKL on two complementary open-access datasets spanning a broad chemical space relevant for hydrocarbon conversion catalysis:

- Catalysis-Hub (CatHub)[9] provides $\sim 3.7 \times 10^4$ high-fidelity DFT data points for six adsorbates (CH, CH_2, CH_3, OH, NH, SH) on $\sim 2 \times 10^3$ mono- and bimetallic surfaces. Its narrow chemical scope but rich surface-site diversity makes it ideal for in-distribution (ID) evaluation where test configurations share the training data's reaction family.
- Open Catalyst 2020 (OC20) subset[10] contains $\sim 4.6 \times 10^4$ adsorption configurations across 20 transition metals and three adsorbate motifs (C, H, O), broadening the compositional space with complex adsorbate-surface orientations. Its validation set enables ID assessment analogous to CatHub, while its OOD test set—featuring catalyst compositions and adsorbates absent from training data—provides rigorous evaluation of OOD detection capabilities.

2.1.2 GNN backbones. We employ SchNet[11] as our baseline encoder for its data efficiency and permutation invariance. On CatHub, we additionally deploy PaiNN,[12] whose orientation-equivariant message passing captures directional effects at higher computational cost. Identical latent dimensionality and training protocols ensure that performance differences reflect

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

architectural capability rather than hyperparameter choices (architectural details in SI S3).

2.1.3 Evaluation metrics. We assess each dataset–backbone combination using quantitative error metrics (MAE, RMSE, R^2) and interval-quality diagnostics (NLL, ENCE, miscalibration area, RMSE–RMV parity). While MAE and R^2 measure predictive accuracy, they cannot guide uncertainty-driven workflows like active learning or risk-aware screening, which require well-calibrated confidence intervals that statistically contain true values at expected rates. We therefore emphasize probabilistic calibration (NLL, ENCE, miscalibration area), error–uncertainty correlation (Spearman's coefficient, ROC–AUC), and computational cost. This multidimensional evaluation rigorously assesses both raw performance and practical utility for catalyst discovery.

A more detailed discussion of the dataset, GNN backbones, and evaluation metrics are provided in the SI (Section S2, S3 and S7).

2.2 Comparative evaluation of predictive accuracy and uncertainty quantification

Here we benchmark the performance of DGKL against established UQ methods, focusing on three central questions: (i) how accurately do they predict adsorption energies, (ii) how reliably do they quantify and calibrate predictive uncertainty, and (iii) the computational cost of obtaining these metrics. DGKL and DGKL-atomic are compared against Ensemble, Evidential, and Monte Carlo dropout (MCD) across three dataset–GNN backbone combinations: CatHub-SchNet, CatHub-PaiNN, and OC20-SchNet.

2.2.1 Predictive accuracy. While DGKL is primarily designed for well-calibrated uncertainty estimates, it maintains competitive predictive accuracy. Table 1 summarizes MAE and R^2 performance across methods. As expected, Ensemble and Evidential approaches achieve the lowest MAE, given their accuracy-focused training objectives (Huber or MSE loss). On the CatHub-PaiNN dataset, Evidential and Ensemble models yield MAEs of 0.13 ± 0.02 eV and 0.15 ± 0.09 eV, respectively, *versus* 0.31 ± 0.07 eV for DGKL and 0.30 ± 0.02 eV for DGKL-atomic. Similar trends appear for SchNet and the OC20 dataset.

This performance gap reflects fundamental differences in optimization objectives. Ensemble and Evidential methods directly minimize point prediction error through MSE or Huber loss—well-suited for accuracy but not uncertainty calibration. Conversely, DGKL employs probabilistic objectives like the Evidence Lower Bound (ELBO) (Section 4.1.1), which incorporates KL divergence regularization to encourage smoother function approximations.[6] This probabilistic framework trades slightly higher point errors for more faithful uncertainty representations, a well-documented trade-off.[13,14] These accuracy differences carry practical implications. The observed MAE gaps between DGKL and point predictors (∼0.15–0.2 eV on CatHub, ∼0.1–0.2 eV on OC20), when translated to activation barriers *via* Brønsted–Evans–Polanyi relationships,[15] could alter predicted reaction rates by orders of magnitude at typical catalytic temperatures. For applications requiring maximum point accuracy on well-characterized, in-domain data without uncertainty quantification, accuracy-optimized models may be preferable.

However, in molecular discovery workflows—particularly computational catalyst screening—prioritizing calibrated uncertainty over marginal accuracy gains is often more advantageous. In active learning, slightly higher MAE becomes inconsequential if well-calibrated uncertainties effectively guide acquisition functions toward informative DFT calculations or optimal catalysts.[2,16,17] When screening novel candidates like single-atom catalysts, reliable uncertainty estimates that flag promising-yet-uncertain materials for validation outweigh marginal accuracy improvements, particularly when avoiding overconfident misclassification of breakthrough materials.[18]

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[a]

	Method	MAE (eV) [↓]	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

[a] Notes: [↑]: higher is better. [↓]: lower is better. Bold values indicate the best method for each dataset-GNN combination. [*] SCC = Spearman Correlation Coefficient.

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

The computational cost of additional validation is insignificant compared to discovering transformative catalysts that circumvent scaling relation limitations.

2.2.2 Error-based metrics: probabilistic fit and correlation. Error-based metrics directly assess UQ quality by comparing predicted uncertainty against actual prediction error. Well-calibrated models exhibit low, stable Negative Log-Likelihood (NLL) and strong positive correlation between error magnitudes and predicted uncertainty, quantified by Spearman's rank correlation (ρ_{SCC}) and ROC–AUC scores.

The NLL (Table 1) measures how well predicted Gaussian distributions ($\mathcal{N}(\mu, \sigma^2)$) match true observations. On CatHub, while Ensemble achieves low average NLL, its substantial standard deviations compromise reliability. Owing to the construction of ensemble model based on top performing models during hyperparameter sweep, the model is sensitive to (i) which architectures are selected and (ii) how differently they respond to a given data partition. This sensitivity results in either homogeneous or heterogeneous ensemble members, leading to a large standard deviations observed in our experiments. DGKL demonstrates superior stability with minimal variance. This advantage becomes pronounced on the complex OC20-SchNet dataset, where DGKL-atomic (1.07 ± 0.03) and DGKL (1.30 ± 0.02) achieve the lowest, most stable NLL scores—particularly notable given that prior benchmarks typically report higher positive NLLs for less calibrated models on complex datasets.[19,20] For computational catalysis, stable low NLL ensures consistently trustworthy uncertainty estimates. Conversely, Ensemble's NLL degrades substantially on OC20, while MCD performs poorly. Notably, superior NLL alone does not guarantee overall uncertainty quality, as the metric is sensitive to both accuracy and calibration.[21]

Reliable UQ methods must also produce uncertainties that correlate with prediction errors. On CatHub-SchNet, DGKL achieves the highest correlation ($\rho_{SCC} = 0.51 \pm 0.02$) and best error discrimination (ROC–AUC $= 0.74 \pm 0.01$), indicating a 74% probability of correctly ranking high-error predictions as more uncertain. On OC20, DGKL-atomic demonstrates superior performance ($\rho_{SCC} = 0.34 \pm 0.02$, ROC–AUC $= 0.66 \pm 0.01$), surpassing standard DGKL. This error–uncertainty correlation optimizes computational resources by prioritizing expensive DFT calculations for promising yet uncertain candidates.[16] Fig. 2 visualizes this correlation between predicted uncertainty and actual error magnitudes. DGKL and DGKL-atomic (Fig. 2(a and b)) exhibit ideal behavior with predictions clustering around the 1σ line and contained within 2σ bands, demonstrating the model "knows what it doesn't know." Conversely, Ensemble (Fig. 2(c)) shows slight overconfidence by assigning low uncertainty to large errors. Evidential and MCD methods (Fig. 2(d and e)) display poor calibration with uncertainties clustered near zero, severely limiting their utility for risk assessment. This analysis confirms that DGKL-based methods provide more informative and reliable uncertainty estimates than conventional approaches.[22]

2.2.3 Interval-based metrics: statistical reliability and calibration quality. Interval-based metrics evaluate the statistical integrity of predicted uncertainty distributions. Well-calibrated models produce statistically reliable confidence intervals—the X% CI is the range that has an X% probability that the true value lies within that range. This property is assessed through Expected Normalized Calibration Error (ENCE), Miscalibration Area, calibration plots,[23,24] and reliability diagrams.[25]

The DGKL framework consistently achieves superior calibration across all dataset-GNN combinations (Table 1). On CatHub-SchNet, DGKL yields a ENCE of 0.07 ± 0.04 and a Miscalibration Area of 0.04 ± 0.02, significantly outperforming competitors. ENCE values below 0.1 indicate that binned RMSE deviates from binned Root Model Variance (RMV) by less than 10%, demonstrating excellent calibration. This advantage is amplified on OC20-SchNet, where DGKL and DGKL-Atomic maintain excellent calibration (ENCE 0.06 ± 0.04 and 0.08 ± 0.05), while other methods degrade substantially. Poor performance indicates fundamental issues with either scale or distributional assumptions of uncertainty estimates, consistent with documented calibration challenges of evidential methods under distribution shifts[26,27] and ensemble degradation on OOD data.[28] Fig. 3 visualizes this superior performance. DGKL and DGKL-atomic calibration curves (top row) closely track the ideal diagonal, confirming statistically reliable confidence intervals. Other methods show significant deviations indicating overconfidence (Ensemble, MCD) or underconfidence (Evidential). Reliability diagrams (bottom row) reinforce this: DGKL variants follow the ideal $y = x$ line, meaning predicted uncertainty (RMV) correctly matches actual error (RMSE), while other methods deviate substantially on OC20.

Quantitative reliability assessment (Table 2) provides further confirmation. Beyond R^2 values, the slope and intercept of RMSE *vs.* RMV linear fits offer direct calibration insights. DGKL achieves slopes near the ideal value of 1, particularly on OC20-SchNet (1.02 ± 0.11), with intercepts consistently near 0. Other methods exhibit substantial positive intercepts on OC20, indicating systematic uncertainty underestimation for the most reliable predictions—a critical failure mode. Empirical coverage rates underscore these differences. DGKL and DGKL-atomic achieve 1σ and 2σ coverage remarkably close to theoretical Gaussian values (68.27% and 95.45%). For instance, DGKL on CatHub-SchNet yields 0.68 ± 0.02 (1σ) and 0.95 ± 0.01 (2σ) coverage. Conversely, MCD exhibits severe under-coverage (0.16 ± 0.03 at 1σ on OC20), revealing misleadingly optimistic error bars. As emphasized previously,[29] empirical coverage evaluation is essential for understanding model reliability.

These interval-based metrics demonstrate that DGKL produces statistically sound, well-calibrated confidence intervals critical for reliable decision-making in materials discovery. DGKL-based approaches offer the most balanced combination of calibration quality, scalability, and interpretability, evidenced by superior ENCE, miscalibration area, reliability characteristics, and empirical coverage matching theoretical expectations. Despite slightly higher MAE, DGKL's uncertainty estimates correlate better with actual errors and provide more stable probabilistic fit, particularly on challenging datasets like OC20.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

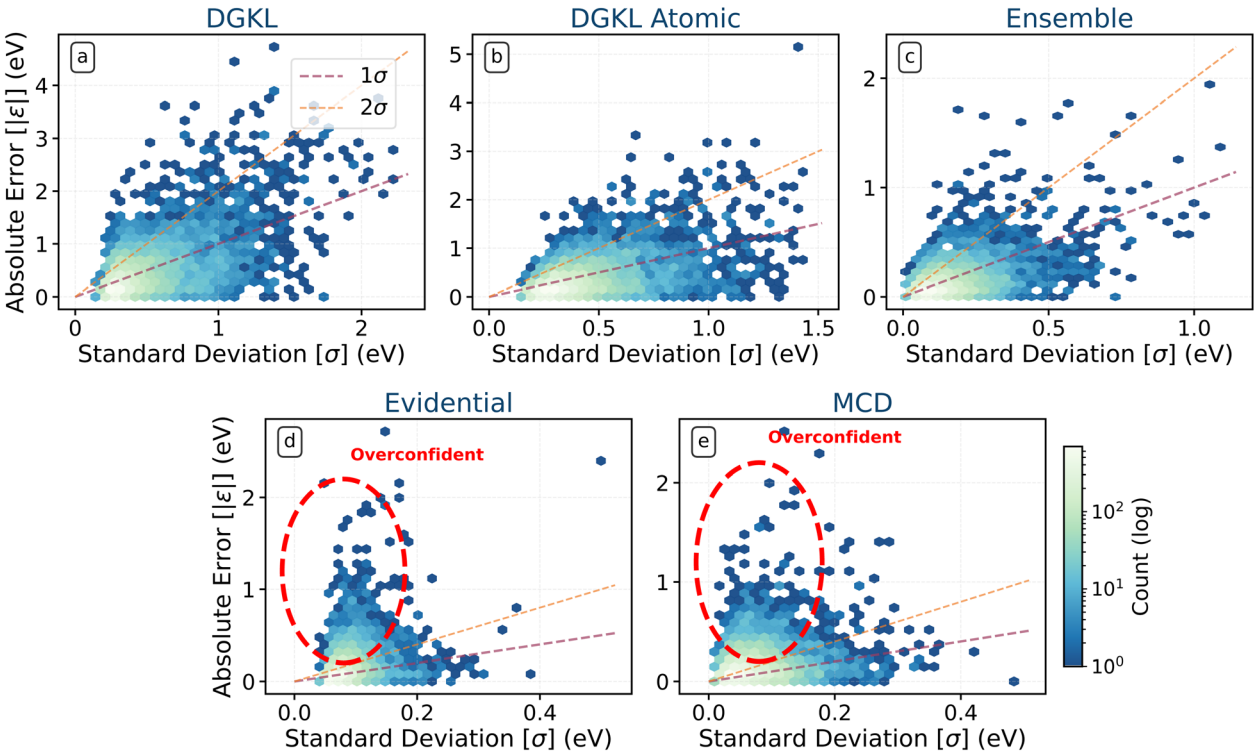

Fig. 2 Uncertainty calibration analysis for different uncertainty quantification methods using the SchNet GNN backbone on the CatHub dataset test set. Each subplot shows the relationship between the absolute prediction error ($|\varepsilon|$) and the predicted uncertainty (standard deviation, σ) for a specific method: (a) DGKL, (b) DGKL-atomic, (c) Ensemble, (d) Evidential, and (e) Monte Carlo Dropout (MCD). The dashed purple and orange lines represent the ideal 1σ and 2σ calibration targets, respectively, where absolute error would equal the predicted standard deviation. The color intensity indicates the density of data points on a logarithmic scale, highlighting regions with higher concentrations of predictions. Methods exhibiting points clustered closer to the dashed lines, particularly the 1σ line, demonstrate better calibration between predicted uncertainty and observed error magnitude. We also add a red ellipsoid to (d) & (e) to illustrate the overconfident region where the predicted error severely underestimates the true error.

2.3 Performance on out-of-distribution datasets

Evaluating uncertainty estimation on OOD data is crucial for assessing model robustness in real-world applications. The OC20 benchmark provides three OOD test sets curated by Meta FAIR: OOD-cat (unseen catalyst compositions), OOD-ads (unseen adsorbates), and OOD-both (unseen combinations), representing progressively challenging generalization scenarios. Fig. 4 analyzes uncertainty distributions across ID and OOD test sets using OC20-SchNet. Well-calibrated models should predict lower uncertainty for ID examples and higher uncertainty for OOD examples, where unfamiliar chemistry may reduce prediction reliability. DGKL (Fig. 4(a)) exhibits consistent uncertainty estimates across all datasets, with mean values increasing modestly from ID (0.7 eV) to OOD scenarios (∼0.9 eV). While this stability indicates strong calibration, the limited ID-OOD separation suggests only moderate sensitivity to distributional shifts. The Ensemble method (Fig. 4(b)) shows more pronounced ID-OOD separation than DGKL, with median uncertainty increasing from 0.2 eV (ID) to 0.6 eV (OOD-cat), 1.0 eV (OOD-ads), and 0.5 eV (OOD-both). DGKL-atomic decomposes uncertainty into atomic contributions, enabling assessment at both molecular (Fig. 4(c)) and atomic scales (Fig. 4(d)). At the molecular scale, median uncertainty increases from 0.7 eV (ID) to 1.25 eV (OOD), exceeding both DGKL and Ensemble. At the atomic scale, ID uncertainties center around 0.01 eV, while OOD datasets exhibit higher medians (∼0.02 eV) and broader distributions. The consistency of atomic uncertainties across OOD types suggests they capture local chemical environment novelty, providing robust signals for active learning. Table 3 quantifies OOD detection capability using ROC–AUC scores, where predicted uncertainties serve as classification scores for distinguishing ID from OOD data. DGKL shows modest performance (0.566–0.603), while Ensemble achieves improved detection (0.732–0.790). DGKL-atomic demonstrates superior performance (0.837–0.876), additionally per-atom uncertainties also achieving strong detection (0.782–0.791). These results underscore the importance of well-calibrated uncertainty estimates for OOD detection. DGKL-atomic's consistent and physically meaningful uncertainty behavior across diverse scenarios makes it particularly suitable for uncertainty-driven workflows in computational catalysis.

2.4 DGKL-atomic: atom-resolved uncertainty and applications

DGKL-atomic uniquely quantifies uncertainty at the atomic level, enabling granular insights into prediction reliability and

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

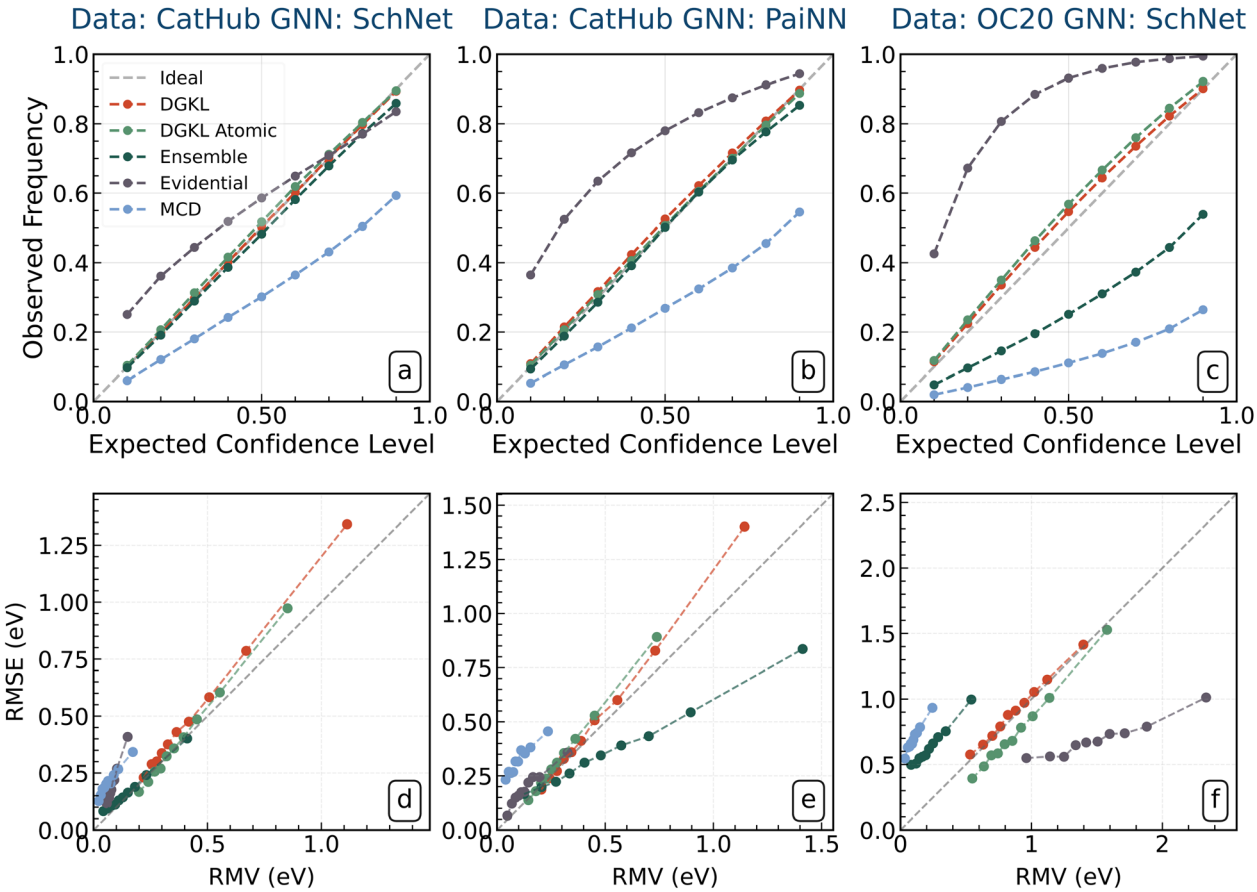

Fig. 3 Interval-based uncertainty calibration metrics for different UQ methods across dataset-GNN combinations. (Top row) (a–c) Calibration plots showing observed frequency *versus* expected confidence level. The dashed diagonal line represents perfect calibration, where the observed frequency matches the expected confidence. Curves above the diagonal indicate underconfidence, while curves below indicate overconfidence. (Bottom row) (d–f) Reliability diagrams plotting Root Mean Squared Error (RMSE) against Root Mean Variance (RMV) calculated over uncertainty bins. The dashed diagonal line represents ideal calibration where RMSE equals RMV. Points above the line indicate underconfidence (uncertainty estimate is larger than the error), while points below indicate overconfidence (uncertainty estimate is smaller than the error). Results are shown for: (a and d) CatHub dataset with SchNet backbone, (b and e) CatHub dataset with PaiNN backbone, and (c and f) OC20 dataset with SchNet backbone. Methods whose curves closely follow the ideal diagonal lines in both plot types demonstrate superior calibration.

Table 2 Calibration metrics for RMSE *vs.* RMV plot along with empirical coverage at 1 and 2σ level[a]

	Method	R^2 [↑]	Slope [→1]	Intercept [→0]	1σ coverage [→0.68]	2σ coverage [→0.95]
	DGKL	**1.00 ± 0.00**	1.22 ± 0.18	−0.06 ± 0.06	0.70 ± 0.03	**0.94 ± 0.02**
CatHub	DGKL atomic	**1.00 ± 0.00**	1.28 ± 0.05	−0.06 ± 0.02	0.69 ± 0.03	0.94 ± 0.01
PaiNN	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
	DGKL	**1.00 ± 0.00**	1.05 ± 0.10	**0.01 ± 0.02**	**0.68 ± 0.02**	**0.95 ± 0.01**
CatHub	DGKL atomic	**1.00 ± 0.00**	1.32 ± 0.20	−0.08 ± 0.02	0.69 ± 0.03	0.94 ± 0.01
SchNet	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
	DGKL	0.99 ± 0.01	**1.02 ± 0.11**	**0.01 ± 0.05**	**0.72 ± 0.03**	**0.94 ± 0.01**
OC20	DGKL atomic	**1.00 ± 0.00**	1.14 ± 0.08	−0.14 ± 0.07	0.74 ± 0.04	0.96 ± 0.01
SchNet	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

[a] 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. Bold values indicate the best method(s) for each dataset-GNN combination.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

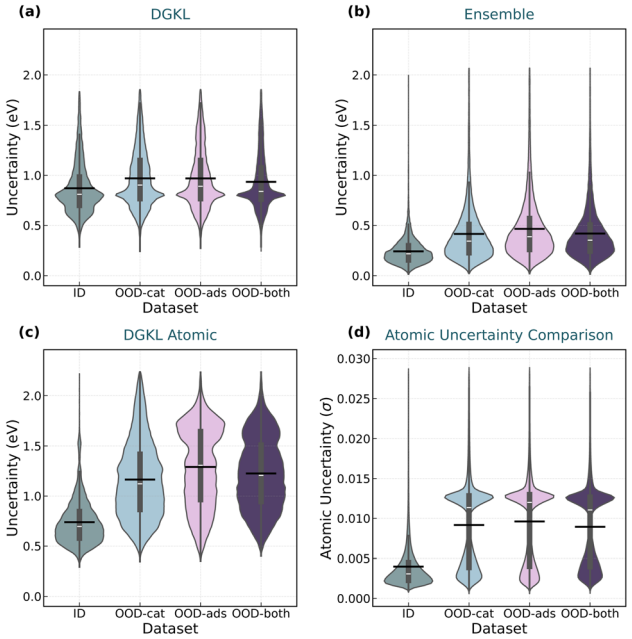

Fig. 4 Comparison of uncertainty measurements (σ) across different datasets and methods. Panel (a) shows uncertainty distributions for the DGKL method, (b) shows results for the Ensemble method, (c) presents DGKL Atomic uncertainty measurements, and (d) displays atomic uncertainty comparisons for ID 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. For panel (b) and (d), we remove the top 2% outliers which obscures the violin plot shape by skewing the distribution to the top.

Table 3 Out-of-distribution detection performance measured by ROC–AUC scores for different uncertainty quantification methods across three OOD scenarios. The table compares DGKL, Ensemble, DGKL Atomic, and DGKL Atomic (per-atom) methods on their ability to discriminate between ID and OOD samples using predicted uncertainty values as classification scores. OOD-cat represents novel catalyst compositions, OOD-ads represents novel adsorbates, and OOD-both represents novel combinations of both catalysts and adsorbates. ROC–AUC scores range from 0.5 (random discrimination) to 1.0 (perfect discrimination), with higher values indicating superior OOD detection capability

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 (per-atom)	0.791	0.787	0.782

targeted chemical space exploration. In the absence of ground truth labels for atomic adsorption energy contributions, validation of atomic uncertainty in terms of error-based and interval-based metrics was not feasible. To assess the performance of atom-resolved uncertainty, we rely only on relative magnitude of uncertainty between OOD and ID atoms/materials. As a result, atom-level uncertainty serves as a interpretability tools for materials discovery pipeline, rather than a direct validation for risk assessment. Fig. 4(d) reveals a bimodal distribution in per-atom uncertainty for OOD samples. We hypothesize that this reflects two populations: atoms in genuinely novel environments (high uncertainty) and atoms in familiar local environments despite belonging to OOD materials (low uncertainty). To validate this, we analyzed structural similarity using SOAP descriptors,[30,31] which provides an asymptotically complete and smooth representation of atomic local environment. The cosine similarity between OOD and nearest ID atomic environments shows strong negative correlation with predicted uncertainty ($r \approx -0.8$), confirming the model differentiates familiar from unfamiliar local structures.

To isolate the model's performance on truly novel atoms, we removed those OOD atoms that were highly similar to ID atoms (cosine similarity >0.8) and re-evaluated OOD detection (SI S11). This filtering led to a substantial increase in ROC–AUC scores, achieving values of 0.90–0.91. Even though the magnitude of difference between ID and OOD uncertainties on the atomic scale is small compared to molecular scale, it still provides a strong signal to isolate truly OOD atoms from the ID atoms. These results indicate that DGKL-atomic's uncertainty estimates are strongly grounded in local structure: the model reliably flags novel atomic environments with high uncertainty, while appropriately assigning low uncertainty to familiar ones-even when they appear in OOD materials. This structural awareness explains the bimodal distribution observed in Fig. 4(d), where one peak corresponds to truly novel atoms and the other to atoms whose local environments resemble those seen during training. This also clarifies why the per-atom ROC–AUC (Table 3, fourth row) is slightly lower than the material-level score (third row) – while the evaluation metric assumes all atoms in OOD materials should appear uncertain, DGKL-atomic appropriately assigns low uncertainty to atoms with familiar local environments, which is a behavior that reflects the model's structural sensitivity but lowers the apparent performance under this metric. Beyond enabling robust OOD detection, atomic-level uncertainty provides practical guidance for materials discovery. In catalytic systems, atoms with high uncertainty often coincide with active sites or chemically sensitive regions, making them natural candidates for further sampling or validation. Coupling these uncertainty estimates with structural descriptors like SOAP can help pinpoint which local environments most limit model confidence. This fine-grained interpretability supports targeted data acquisition and improved active learning strategies that prioritize the most informative, underrepresented structures. We note that, while DGKL-atomic's per-atom uncertainties correlate strongly with SOAP-based measures of local structural novelty and show improved OOD detection performance, we do not perform controlled atomic-level ablation studies in this work, such as withholding specific element types or chemically distinct coordination environments during training and reintroducing them only at test time. Consequently, our analysis of atomic-level uncertainty remains correlational rather than fully causal, and atomic uncertainties should be interpreted as structure-aware indicators of local novelty and model confidence, rather than

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

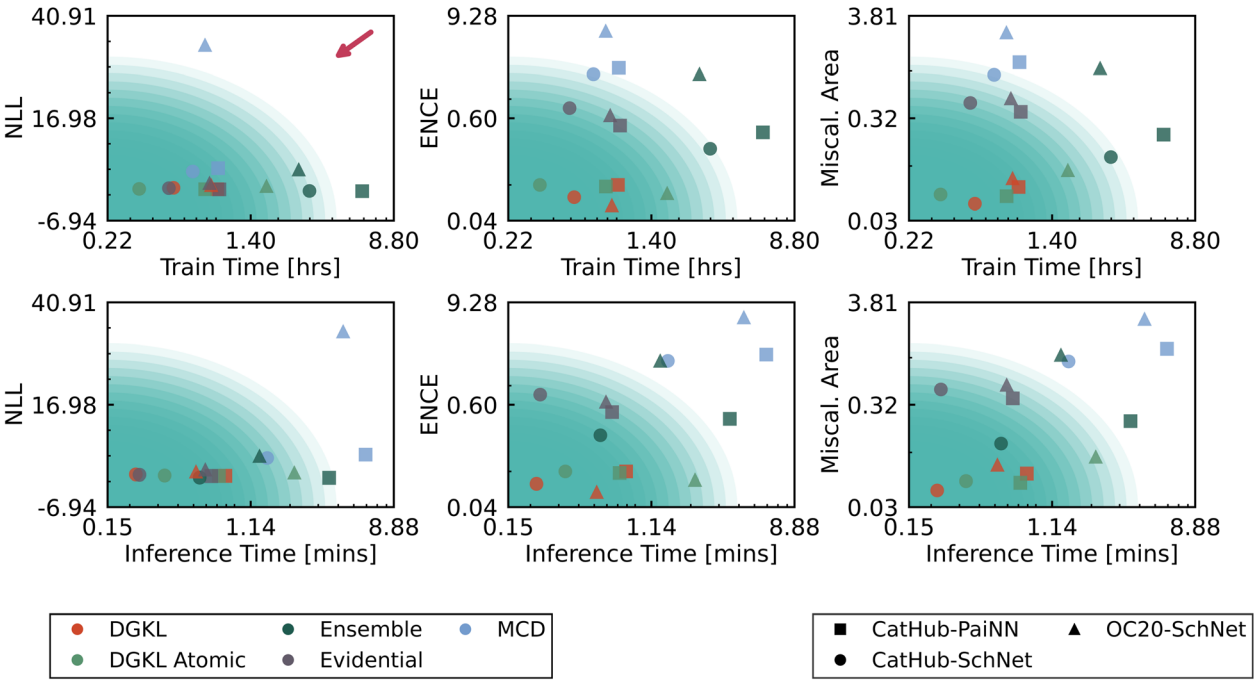

Fig. 5 Pareto analysis comparing UQ performance metrics against computational cost for different UQ methods and dataset-GNN combinations. Each point represents the average performance of a specific UQ method on a particular dataset-GNN setup over five runs. Colors indicate the UQ method: DGKL (Red), DGKL atomic (Green), Ensemble (Dark Green), Evidential (Gray), MCD (Blue). Marker shapes indicate the dataset-GNN combination: CatHub-SchNet (Circle), CatHub-PaiNN (Square), OC20-SchNet (Triangle). (Top row) UQ metrics (NLL, ENCE, miscalibration area; lower is better) plotted against total training time (hours, log scale). (Bottom row) The same UQ metrics plotted against inference time for the entire test set (min, log scale). The Pareto optimal region (bottom-left corner) is emphasized by concentric circles with decreasing opacity, where methods achieve the best trade-off between computational efficiency and UQ performance. Methods appearing closer to the center of these circles offer superior efficiency-performance balance.

as definitive probes of unseen chemistry. Designing targeted ablation protocols to isolate the impact of missing species or local environments is an important direction that we leave for future work.

2.5 Performance *vs.* computational cost

Practical deployment of UQ methods in computational catalysis requires not only accurate predictions but also computational efficiency.[4,20] Fig. 5 presents Pareto plots comparing key UQ metrics (NLL, ENCE, Miscalibration Area) against training time (top row) and inference time (bottom row) on log scales. Ideal methods occupy the lower-left region, achieving high UQ performance at low computational cost. For training, Ensemble methods are inherently expensive, requiring independent training of multiple models (>3 hours in our experiments; see SI Section S1.1). Single-model training (DGKL, DGKL-Atomic, Evidential, MCD) is significantly faster (<1.5 hours). While DGKL optimizes both GNN and SVGP components, its training time still remains competitive with Evidential and MCD methods (details in SI Section S5.5). DGKL and Evidential methods require only one forward pass, enabling rapid inference (<0.3 minutes per test set)—a critical advantage for high-throughput screening and machine learning interatomic potential (MLIP) driven simulations. Conversely, Ensemble ($N_{\text{forward pass}} = 5$) and MCD ($N_{\text{forward pass}} = 10$) methods require multiple passes, substantially increasing inference time (0.5–1 minute).[32] Overall, DGKL and DGKL-atomic achieve Pareto-optimal performance for ENCE and miscalibration area, combining superior calibration with low training and inference costs. While Ensembles exhibit competitive NLL, their computational expense limits scalability. MCD shows poor cost-performance trade-offs, and Evidential methods suffer from miscalibration despite computational efficiency. These results establish DGKL-based frameworks as computationally practical solutions for uncertainty quantification in catalyst discovery and materials simulation.

3 Conclusion

In this work, we report the development of a deep graph kernel learning (DGKL) framework for adsorption energy prediction. By combining the representational power of graph neural network with the Sparse Variational Gaussian Processes, DGKL consistently delivers well-calibrated uncertainty estimates. Our comprehensive benchmarks across the CatHub and OC20 datasets demonstrate that the DGKL approach, in its standard form, achieves superior performance in critical calibration metrics. Specifically, DGKL yielded excellent ENCE values (0.06–0.10) and miscalibration areas (0.04–0.07), robust error–uncertainty correlation (Spearman ρ_{SCC} up to 0.51, ROC–AUC up to 0.74), and near-ideal reliability diagram characteristics (RMSE *vs.* RMV $R^2 \sim 1.00$, slope ~1.0), as well as empirical coverage at

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

1σ and 2σ levels. This performance marks a significant improvement over conventional UQ methods such as Ensemble, Evidential regression, and Monte Carlo dropout, particularly in terms of calibration consistency and computational efficiency. DGKL's superior performance can facilitate the screening of high-throughput catalytic materials based on adsorption energy prediction for hydrocarbon conversion reactions within single- or multi-metallic alloy systems. Looking forward, a natural extension is coupling DGKL with more expressive equivariant and foundation-model backbones (*e.g.*, GemNet, MACE, Equiformer) while scaling to larger, more diverse benchmarks (OC20/OC22, mixed-fidelity corpora) and exploring richer GP approximations and structured covariance to maintain calibration. A complementary direction is applying these calibrated atomic-level uncertainties to complex catalytic systems—multicomponent alloys, oxides, and single-atom catalysts—where reliably quantifying model confidence in heterogeneous, strongly cooperative environments is essential for breaking traditional scaling relations and identifying novel active sites.

DGKL-atomic, a variant that extends UQ to the individual atom level by delaying the final pooling operation, allows for fine-grained insight into local spatial prediction reliability. Atomic uncertainty can be a very powerful interpretability tool, but it's not a principled way of assessing risk because we don't have the ground truth labels to validate our risk assessment capacity. Consequently, it should be used for exploratory analysis, such as active learning-guided discovery, rather than as a predictive uncertainty to facilitate critical decision-making analysis. DGKL-atomic demonstrated exceptional robustness and discriminative power for OOD data, as evidenced by its superior ROC–AUC scores (0.84–0.88) on the challenging OC20 OOD test dataset, implying consistent uncertainty calibration for novel catalyst and/or adsorbate scenarios. While standard DGKL provides excellent calibration for in-distribution data typical of the CatHub dataset, DGKL-atomic excelled in navigating the complexities of diverse and out-of-distribution chemical space of the OC20 dataset. This makes DGKL-atomic an especially promising tool for exploratory research and discovery of novel catalytic materials where encountering OOD samples is inevitable, such as single atom or high entropy alloy catalysts. The ability to identify specific, uncertain atomic environments can help refine active learning strategies by guiding data acquisition towards the most informative and scientifically interesting regions of chemical and structural space.

4 Methods

4.1 Uncertainty quantification models

4.1.1 Deep graph kernel learning (DGKL). Deep Graph Kernel Learning (DGKL) integrates a Graph Neural Network (GNN) feature extractor with Gaussian Process (GP) regression, training both components end-to-end. We employ SchNet and PaiNN as GNN backbones, as discussed previously. Fig. 1 illustrates the DGKL architecture: atomic positions and chemical information from adsorbate/slab systems are encoded through the GNN into a latent space representation with inducing points, then processed *via* Sparse Variational Gaussian Process (SVGP) regression to yield adsorption energy predictions as normal distributions $\Delta E_{\text{ads}} = \mathcal{N}(\mu, \sigma^2)$.

4.1.1.1 Gaussian process component. GPs are non-parametric methods that learn distributions over function spaces *via* kernel functions, providing both predictions and uncertainty estimates.[33–35] A GP is defined by a mean function $\mu(\cdot)$ and covariance function $K(\cdot,\cdot)$. However, exact GP inference requires inverting the Gram matrix $K(M, M)$, yielding $\mathcal{O}(N^3)$ complexity that becomes prohibitive for large datasets.

To address this limitation, we employ Sparse Variational GPs (SVGPs), which approximate the full GP using a learned subset of inducing points U in the latent space rather than all training points M.[36,37] This reduces computational cost while maintaining predictive accuracy. The SVGP learns both inducing point locations and kernel parameters through variational inference.

4.1.1.2 DGKL architecture. In DGKL, the GNN extracts molecular-level features *via* global pooling (sum or mean) of per-atom representations. These latent features are fed to the SVGP, which predicts adsorption energies with associated uncertainties. We systematically optimized the following SVGP hyperparameters alongside GNN backbone parameters:

(1) Latent space dimension
(2) Kernel functions (Matérn, RBF)
(3) Variational distributions (Cholesky, mean-field)
(4) Variational strategies (standard, decoupled)
(5) Objective functions (ELBO, PLL)

4.1.1.3 Training challenges and solutions. End-to-end DGKL training is challenging due to mismatched optimization dynamics between GNN and SVGP components, potentially causing mode collapse and numerical instabilities.[6,38] We implemented four key strategies to ensure robust training:

(1) Feature normalization: layer and feature normalization stabilize latent space representations.

(2) Differential learning rates: GNN learning rates were set two orders of magnitude lower than SVGP rates to maintain stable inducing points.

(3) Early stopping: training halted after five epochs without validation improvement to prevent overfitting.

(4) Adaptive recovery: upon detecting numerical instabilities (*e.g.*, NaN values), we rolled back to previous weights and halved the learning rate.

These techniques enabled stable end-to-end DGKL training across all configurations tested. A comprehensive discussion of Ensemble, Monte-Carlo Dropout, and Evidential models is provided in SI S4 and S8.

4.1.2 Deep graph kernel learning with atomic uncertainty (DGKL-atomic). DGKL-atomic extends the DGKL framework by learning latent representations at the atomic level rather than the material level. This is achieved by deferring global pooling until after SVGP inference, allowing the model to predict both energy contributions and uncertainties for individual atoms. These atomic-level quantities are then aggregated to yield material-level predictions.

4.1.2.1 Mathematical formulation. The aggregation from atomic to material-level predictions follows:

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

$$\mu_{\text{material}} = \sum_{i \in \text{graph}} \mu_{\text{atomic},i}, \quad (1)$$

$$\sigma_{\text{material}}{}^2 = \sum_{i \in \text{graph}} \sigma_{\text{atomic},i}{}^2 + 2\sum_{i<j} \text{Cov}(E_i, E_j) \approx \sum_{i \in \text{graph}} \sigma_{\text{atomic},i}{}^2, \quad (2)$$

where μ_{material} and $\sigma_{\text{material}}{}^2$ are the predicted adsorption energy and variance for the full adsorbate–surface system; $\mu_{\text{atomic},i}$ and $\sigma_{\text{atomic},i}{}^2$ are the mean energy contribution and predictive variance for atom i; and $\text{Cov}(E_i, E_j)$ represents the covariance between atoms i and j. Prior studies have shown that the off-diagonal covariances can be non-negligible.[39] Here we disregard them in eqn (2) because we lack atom-level target properties that would enable the model to accurately condition itself to produce well-calibrated atomic covariances. Without such atom-level target properties, such as forces, the optimization algorithm has considerable freedom to manipulate the atomic covariance to minimize loss without actually learning the underlying distribution. SI S3.4 presents a comparison between models conditioned on the whole covariance matrix *vs.* the only the diagonal elements, which shows that incorporating the full covariance structure leads to worse predictive performance. Consequently, we emphasize that the atomic uncertainty should be utilized as an interpretability and exploratory tool, rather than as a proper risk assessment tool.

4.1.2.2 Advantages and interpretability. DGKL-atomic offers distinct advantages over material-level UQ methods. By modeling uncertainty at the atomic level, it captures local electronic and structural contributions to adsorption energetics with higher fidelity, identifying specific sites that contribute most to prediction uncertainty. This granularity provides actionable insights for targeted catalyst design.

The atomic-level framework enables spatial visualization of uncertainty distributions across material structures, revealing patterns related to coordination environments, bonding configurations, and surface features that correlate with prediction reliability. Such detailed uncertainty characterization is particularly valuable for complex heterogeneous catalysts and nanostructured materials where local atomic environments vary significantly. This interpretability bridges the gap between model predictions and chemical intuition, facilitating rational design of catalytic materials.

Author contributions

O. M.: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft. C. Y.: formal analysis, writing – review & editing. S. Y.: funding acquisition, investigation, methodology, project administration, resources, supervision, visualization, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

This study was carried out using adsorption energy data from the Open Catalyst 2020 (OC20) database (https://opencatalystproject.org, Chanussot *et al.*, *ACS Catal.* 2021, DOI: https://doi.org/10.1021/acscatal.0c04525) and Catalysis-Hub (CatHub) database (https://www.catalysis-hub.org). All code needed to reproduce the analysis is available in the GitHub repository https://github.com/YueGroup/DGKL. The source code for the DGKL framework is available under the MIT license and archived on Zenodo (https://doi.org/10.5281/zenodo.18809988).

Supplementary information (SI): additional details regarding the methods used and SI figures are provided in the SI document. See DOI: https://doi.org/10.1039/d6dd00020g.

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.

Notes and references

1 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nat. Commun.*, 2017, **8**, 14621.
2 Z. Jiao, Y. Liu and Z. Wang, *J. Chem. Phys.*, 2024, **161**, 171001.
3 Z. Jiao, Y. Mao, R. Lu, Y. Liu, L. Guo and Z. Wang, *J. Chem. Theory Comput.*, 2025, **21**(6), 3176–3186.
4 A. R. Tan, S. Urata, S. Goldman, J. C. Dietschreit and R. Gómez-Bombarelli, *npj Comput. Mater.*, 2023, **9**, 225.
5 Y.-W. Du and J.-J. Zhong, *Inf. Sci.*, 2021, **547**, 1201–1232.
6 S. W. Ober, C. E. Rasmussen and M. van der Wilk, The Promises and Pitfalls of Deep Kernel Learning, *arXiv*, 2021, preprint, arXiv:2102.12108, DOI: 10.48550/arXiv.2102.12108.
7 D. Burt, C. E. Rasmussen and M. Van Der Wilk, *International Conference on Machine Learning*, 2019, pp. 862–871.
8 F. Leibfried, V. Dutordoir, S. John and N. Durrande, *arXiv*, 2020, preprint, arXiv:2012.13962, DOI: 10.48550/arXiv.2012.13962.
9 K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich and T. Bligaard, *Sci. Data*, 2019, **6**, 75.
10 L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick and Z. Ulissi, *ACS Catal.*, 2021, **11**, 6059–6072.
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, *arXiv*, 2017, preprint, arXiv:1706.08566, DOI: 10.48550/arXiv.1706.08566.

© 2026 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 24 March 2026. Downloaded on 4/20/2026 4:19:29 PM. This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

12 K. T. Schütt, O. T. Unke and M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, *arXiv*, 2021, preprint, arXiv:2102.03150, DOI: 10.48550/arXiv.2102.03150.
13 C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, On Calibration of Modern Neural Networks, *arXiv*, 2017, preprint, arXiv:1706.04599, DOI: 10.48550/arXiv.1706.04599.
14 B. Lakshminarayanan, A. Pritzel and C. Blundell, *Advances in neural information processing systems 30*, 2017.
15 J. K. Norskov, T. Bligaard, A. Logadottir, S. Bahn, L. B. Hansen, M. Bollinger, H. Bengaard, B. Hammer and Z. Sljivancanin, *J. Catal.*, 2002, **209**, 275–278.
16 T. Yin, G. Panapitiya, E. D. Coda and E. G. Saldanha, *J. Cheminf.*, 2023, **15**, 105.
17 Y. Zhang, C. Chen, M.-C. Li, T.-H. Wu, A. Gopakrishnan, C.-S. Lee, A. Agrawal, W.-k. Liao, A. Choudhary, W. Chen and C. Wolverton, *npj Comput. Mater.*, 2019, **5**, 83.
18 C. J. Gruich, V. Madhavan, Y. Wang and B. R. Goldsmith, *Mach. Learn.: Sci. Technol.*, 2023, **4**, 025019.
19 D. Varivoda, R. Dong, S. Sadeed, S. M. N. Omee and J. Hu, *Applied Physics Reviews*, 2023, **10**, 021409.
20 Y. Li, L. Kong, Y. Du, Y. Yu, Y. Zhuang, W. Mu and C. Zhang, MUBen: Benchmarking the Uncertainty of Molecular Representation Models, *arXiv*, 2024, preprint, arXiv:2306.10060, DOI: 10.48550/arXiv.2306.10060.
21 M. H. Rasmussen, C. Duan, H. J. Kulik and J. H. Jensen, *J. Cheminf.*, 2023, **15**, 121.
22 K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing and Z. W. Ulissi, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 025006.
23 T. Gneiting and M. Katzfuss, *Annu. Rev. Stat. Appl.*, 2014, **1**, 125–151.
24 V. Kuleshov, N. Fenner and S. Ermon, *International conference on machine learning*, 2018, pp. 2796–2804.
25 D. Levi, L. Gispan, N. Giladi and E. Fetaya, *Sensors*, 2022, **22**(15), 5540.
26 T. Wollschläger, N. Gao, B. Charpentier, M. A. Ketata and S. Günnemann, Uncertainty Estimation for Molecules: Desiderata and Methods, *arXiv*, 2023, preprint, arXiv:2306.14916, DOI: 10.48550/arXiv.2306.14916.
27 A. P. Soleimany, A. Amini, S. Goldman, D. Rus, S. N. Bhatia and C. W. Coley, *ACS Cent. Sci.*, 2021, **7**, 1356–1367.
28 J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler and X. X. Zhu, A Survey of Uncertainty in Deep Neural Networks, *arXiv*, 2022, preprint, arXiv:2107.03342, DOI: 10.48550/arXiv.2107.03342.
29 B. Kompa, J. Snoek and A. L. Beam, *Entropy*, 2021, **23**(12), 1608.
30 S. De, A. P. Bartók, G. Csányi and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754–13769.
31 A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
32 M. Wen and E. B. Tadmor, *npj Comput. Mater.*, 2020, **6**, 124.
33 C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, Springer, 2006, vol. 4.
34 K. Jakkala, Deep Gaussian Processes: A Survey, *arXiv*, 2021, preprint, arXiv:2106.12135, DOI: 10.48550/arXiv.2106.12135.
35 C. E. Rasmussen, C. K. Williams, *et al.*, *Gaussian processes for machine learning*, 2006, vol. 1.
36 M. Titsias, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, Hilton Clearwater Beach Resort*, Clearwater Beach, Florida USA, 2009, pp. 567–574.
37 J. Hensman, N. Fusi and N. D. Lawrence, Gaussian Processes for Big Data, *arXiv*, 2013, preprint, arXiv:1309.6835, DOI: 10.48550/arXiv.1309.6835.
38 S. I. Allec and M. Ziatdinov, *Digital Discovery*, 2025, **4**, 1284–1297.
39 C.-I. Yang and Y.-P. Li, *J. Cheminf.*, 2023, **15**, 13.

© 2026 The Author(s). Published by the Royal Society of Chemistry