Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Bayesian uncertainty quantification of graph neural networks using stochastic gradient Hamiltonian Monte Carlo

Qinghe Gao[a], Daniel C. Miedema[a], Yidong Zhao[b], Jana M. Weber[c], Qian Tao[b], and Artur M. Schweidtmann[a]*

[a] Process Intelligence Research Team, Department of Chemical Engineering, Delft University of Technology, Van der Maasweg 9, Delft 2629 HZ, The Netherlands
[b] Department of Imaging Physics, Delft University of Technology, Delft, the Netherlands
[c] Pattern Recognition and Bioinformatics, Department of Intelligent Systems, Delft University of Technology, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands
* Corresponding Author: a.schweidtmann@tudelft.nl.

ABSTRACT

Graph neural networks (GNNs) have proven state-of-the-art performance in molecular property prediction tasks. However, a significant challenge with GNNs is the reliability of their predictions, particularly in critical domains where quantifying model confidence is essential. Therefore, assessing uncertainty in GNN predictions is crucial to improving their robustness. Existing uncertainty quantification methods, such as Deep ensembles and Monte Carlo Dropout, have been applied to GNNs with some success, but these methods are limited to approximate the full posterior distribution. In this work, we propose a novel approach for scalable uncertainty quantification in molecular property prediction using Stochastic Gradient Hamiltonian Monte Carlo (SGHMC). Additionally, we utilize a cyclical learning rate to facilitate sampling from multiple posterior modes which improves posterior exploration within a single training round. Moreover, we compare the proposed methods with Monte Carlo Dropout and Deep ensembles, focusing on error analysis, calibration, and sharpness, considering both epistemic and aleatoric uncertainties. Our experimental results demonstrate that the proposed parallel-SGHMC approach significantly outperforms Monte Carlo Dropout and Deep ensembles in terms of calibration and sharpness. Specifically, parallel-SGHMC reduces the sum of squared errors by 99.4% and 75%, respectively, when compared to Monte Carlo Dropout and Deep Ensembles. These findings suggest that parallel-SGHMC is a promising method for uncertainty quantification in GNN-based molecular property prediction.

Keywords: Uncertainty quantification, graph neural networks, property prediction

INTRODUCTION

Graph neural networks (GNNs) have demonstrated promising potential in accelerating the computational estimation of molecular properties [1-3]. However, their reliability remains a significant challenge, e.g., in safety-critical applications where understanding model confidence is crucial. Accurately assessing the uncertainty of GNN predictions is essential for enhancing their robustness, particularly when predicting complex molecular behaviors.

Uncertainty in machine learning can generally be divided into two types: aleatoric and epistemic uncertainty [4]. Aleatoric uncertainty arises from inherent noise in the data generation process, such as experimental measurement errors or environmental variations, and is considered irreducible. In contrast, epistemic uncertainty stems from the lack of knowledge of a model, often due to limited data or an incomplete understanding of the process, and can be reduced through additional data collection or improved model design. For instance, in modeling complex chemical reactions, using a more advanced model architecture—such as incorporating reaction kinetics and thermodynamic principles—can reduce

epistemic uncertainty by capturing the process dynamics more accurately. Estimating both types of uncertainty is critical in domains like molecular property prediction, where errors can have substantial consequences.

Two popular methods for jointly uncertainty quantification of machine learning models are Deep ensembles and Monte Carlo Dropout (MC-dropout). MC-dropout [5] applies dropout during both training and inference, generating multiple stochastic forward passes to approximate the posterior distribution of the model. Deep ensembles [6] involves training multiple independent models and averaging their predictions, with the variance among them reflecting the model uncertainty. Although both methods have been widely adopted for scalable uncertainty estimation in molecular property prediction [7], they have some limitations. Deep ensembles primarily focus on the maximum a posteriori (MAP) estimate, neglecting broader regions of the posterior distribution. Conversely, MC-dropout samples from a single mode, overlooking the possibility of multiple modes in the posterior.

We propose a novel approach using Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) [8] to quantify the uncertainty of GNNs for molecular property prediction task. Specifically, we incorporate a cyclical learning rate, which combines the strengths of both sampling strategies and has been successfully utilized in medical imaging domain [9]. We benchmark the proposed SGHMC method against Deep ensembles and MC-dropout on molecular property prediction tasks. Specifically, we evaluate the models with three different metrics: Error analysis, calibration, and sharpness.

METHODS

Preliminary

To quantify the uncertainty of GNNs in molecular property prediction tasks, the aim is to transform a single prediction (aleatoric uncertainty) and a deterministic weight (epistemic uncertainty) into a distribution, respectively [10]. For aleatoric uncertainty, the mean-variance estimation is often considered as the most practical approach. The mean-variance approach is simply splitting the last layers of neural networks to predict both mean $\mu(x)$ and variance $\sigma^2(x)$ of Gaussian distribution $N(y;\mu(x),\sigma^2(x))$. Therefore, during the inference time, the test case $y_i = \mu(x_i) + \epsilon(x_i)$ where $\epsilon \sim N\left(0,\sigma^2(x_i)\right)$. Additionally, the corresponding objective function (Eq. 1) becomes minimizing the Gaussian negative log-likelihood (NLL) loss function:

$$NLL(w) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left[\frac{1}{\sigma_w^2}(x_i)\left(y_i - \mu_w(x_i)\right)^2 + \log\left(2\pi\sigma_w^2(x_i)\right)\right] \quad (1)$$

Moreover, the Bayesian approach is used to get the epistemic uncertainty, which essentially approximates the prior distribution $p(w|D) \propto p(w)p(D|w)$. Therefore, during the inference time on a test case: x_i :

$$p(y_i|x_i, D) = \int_w p(w|D)p(y_i|x_i, w)\, dw \quad (2)$$

However, in reality, the prior distribution and the integral are intractable. Therefore, they are commonly approximated with MC integration over samples w_i drawn from an approximate posterior distribution $q(w)$:

$$p(y_i|x_i, D) = \frac{1}{M}\sum_{j=1}^{M} p(y_i|x_i, w_j),\ w_j \sim q(w) \quad (3)$$

SGHMC

The proposed SGHMC [9] algorithm samples from the posterior distribution based on the Hamiltonian equation of motion. The Hamiltonian function is given by:

$$H(q,r) = U(q) + K(r) \quad (4)$$

where an object moves on a potential energy surface $U(q)$ with kinetic energy $K(r)$ where q represents the positions and r represents the momenta. Using the canonical distribution (probability of the states based on the energy), the joint probability distribution over positions and momenta can be defined as:

$$p(q,r) \propto e^{-H(q,r)} = e^{-U(q)}\, e^{-K(r)} \quad (5)$$

When the potential energy $U(q)$ is expressed as the NLL of the model combined with a prior, the canonical distribution with respect to $U(q)$ corresponds to posterior distribution of the model parameters w. Minimizing $U(w)$ is equivalent to finding the mode of the posterior distribution:

$$U(w) = -\log[p(w|D)] = NLL(w) + \lambda\left|\left|w\right|\right|^2 \quad (6)$$

The choice of distribution over momenta r can be arbitrary. Here, we chose a quadratic form for the kinetic energy:

$$K(r) = \frac{1}{2} r^T M^{-1} r \quad (7)$$

resulting in Gaussian distribution over r. The mass matrix M, typically set to the identity matrix for simplification [11].

Therefore, the dynamics in the end are governed by equations:

$$dw = M^{-1} r\, dt \quad (8)$$

$$dr = -\nabla U(w)\, dt \quad (9)$$

Additionally, to address the computational challenge of evaluating the full gradient $\nabla U(w)$ in large datasets, SGHMC uses a noisy gradient estimate $\nabla \widetilde{U}(w)$ based on mini-batches. The introduced noise can be mitigated by a friction term A in the momentum update:

$$r_i = r_{i-1} - \nabla \widetilde{U}(w_{i+1})h - A r_{i-1} h + \sqrt{2AhT}N(0, I) \quad (10)$$

This formulation is similar to SGD with momentum but includes an additional Gaussian noise term, controlled by temperature T. The temperature modulates the sensitivity of the posterior distribution to the potential energy: $T = 1$ yields the standard Bayesian posterior, $T < 1$ (cold posterior) increases sensitivity to low-loss regions, while $T > 1$ broadens exploration at potential cost of accuracy.

Evaluation metric

The quality of uncertainty estimates in predictive models can be assessed using three key metrics: error analysis, calibration, and sharpness. For the metric error analysis, we utilize the root mean square error (RMSE) and the coefficient of determination (R2). Calibration methods assess how well predicted uncertainties align with observed errors. Here, we utilize two metrics: NLL and sum of squared errors (SSE). Rather than only using the NLL as a loss function during training, it is also used as a quantitative measure to assess the quality of the predictive distribution after training. SSE measures the total squared error between the predicted quantiles and the observed empirical quantiles. If the SSE is small, it means that the predictions of the models are well-calibrated with respect to the observed data. Sharpness measures the informativeness of uncertainty estimates with root mean predicted variance (RMV) and coefficient of variation C_v. Specifically, RMV is denoted as:

$$RMV = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \widehat{\sigma_i^2}} \tag{11}$$

where $\widehat{\sigma_i^2}$ is predicted variance or uncertainty. A low RMV indicates the model on average predicts low uncertainty and thus low expected error. Additionally, C_v is denoted as:

$$C_v = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\sigma}_i - \sigma_*)}}{\sigma_*} \tag{12}$$

where $\hat{\sigma}_i$ is the predicted standard deviation (uncertainty) of instance i and σ_* is the mean predicted standard deviation. A high C_v suggests significant dispersion (heteroscedasticity), indicating that uncertainty estimates are highly dependent on the input. In contrast, a C_v of zero indicates constant (homoscedastic) uncertainty, meaning the estimates are uninformative and do not vary with the input [10].

Experiment design

The dataset used is the QM9 benchmark [12], consisting of 132,481 small organic molecules (CHONF) with 13 target properties, evaluated at the B3LYP/6-31G(2df,p) level. An 80:10:10 train/validation/test split was applied, using heat capacity at 298.15 K as the target property. The Directed Message Passing Neural Network (D-MPNN) is used as a base model implemented in Chemprop [13], with a mean-variance output and heteroscedastic noise model. The models contain three message passing layers and two feed-forward layers, each with a hidden size of 300.

All uncertainty methods were optimized using the Gaussian NLL loss function with a regularization parameter of λ = 0.0001. The initial learning rate was set to 0.01, with an exponential decay (γ = 0.95) and a batch size of 50. MC-Dropout models are trained for 216 epochs with the SGD optimizer and a dropout rate of p = 0.1, and 24 predictions are collected for the ensemble with different dropouts. Deep Ensemble models are trained for 650 epochs with the SGD optimizer (momentum m = 0.9), with three ensemble predictions from different models. Parallel-SGHMC models are trained for 650 epochs with SGHMC (momentum m = 0.9), with noise injected at T = 0.01 after 60% of training. Posterior samples were collected every ten epochs after noise injection. 24 ensemble predictions are collected from posterior samples. To explore multiple local modes of the posterior, we additionally propose Cyclical-SGHMC model. It uses four cycles within a 650-epoch budget, with a reset learning rate of 0.00025 after the first cycle. Noise was injected with T = 0.01 after 60% of each cycle. Similarly, 24 ensemble predictions are collected from posterior samples. All simulations ran on a Windows server with a 3.5 GHz 24-core Intel(R) Xeon(R) W-2265 CPU, an NVIDIA GeForce RTX 3090 GPU, and 64 GB of memory.

Figure 1 provides a conceptual representation of various posterior sampling strategies employed during a single training process. The x-axis represents the model parameter space www, while the y-axis shows the posterior probability distribution $P(w|D)$ given the data D. The black curve represents the posterior distribution with multiple modes. Deep Ensembles typically sample from a single local mode within a training cycle, limiting their ability to explore the broader posterior distribution. In contrast, both MC-Dropout and parallel-SGHMC exhibit similar sampling behaviours, generating multiple samples from the local regions surrounding the current mode. Notably, cyclical-SGHMC offers a more robust exploration of the posterior by sampling from multiple regions across different modes, facilitated by its cyclical learning rate, which enables escape from local minima and exploration of diverse posterior landscapes.

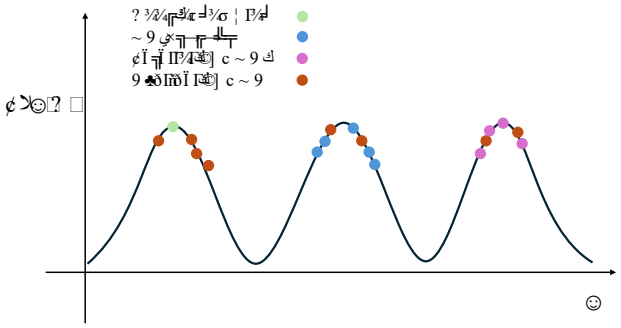

Figure 1. Conceptual representation of the different

Table 1: Uncertainty quantification results from four different models. RMSE stands for the root mean square error, R2 for coefficient of determination, SSE for the sum of squared errors, NLL for the negative log likelihood, RMV for the root mean predicted variance, and CV for coefficient of variation.

Models	Error		Calibration		Sharpness	
	RMSE↓	R^2↑	SSE↓	NLL↓	RMV↓	C_V↑
Parallel-SGHMC	0.6650	0.9735	0.0003	-0.4921	0.5137	2.0178
Cyclical-SGHMC	0.6156	0.9773	0.0405	-0.3522	0.6393	1.6555
MC-dropout	0.7227	0.9687	0.0500	0.4678	0.9167	0.9897
Deep ensemble	0.7399	0.9672	0.0012	0.4357	0.7905	1.0112

posterior sampling strategies for one training process employed by Deep ensembles, MC-dropout, parallel-SGHMC, and cyclical-SGHMC.

RESULTS AND DISCUSSION

Table 1 presents the performance of four uncertainty quantification models - parallel-SGHMC, cyclical-SGHMC, MC-Dropout, and Deep ensemble - on key evaluation metrics, including error, calibration, and sharpness.

In terms of error analysis, SGHMC-based models demonstrate superior performance with respect to predictive error. Specifically, cyclical-SGHMC achieves the lowest RMSE (0.6156) and the highest R^2 (0.9773), indicating that it provides the most accurate predictions. The potential reason is that cyclical-SGHMC samples a richer posterior, capturing parameter uncertainty more effectively than methods like MC-Dropout.

For the calibration category, parallel-SGHMC outperforms the other models with the lowest SSE of 0.0003 and the lowest NLL score of -0.4921. This indicates that its predicted uncertainties are highly aligned with the actual errors, suggesting that this model provides the most reliable uncertainty estimates. Cyclical-SGHMC, with an SSE of 0.0405 and an NLL of -0.3522, also performs well, although its uncertainty estimates are slightly less aligned with the observed errors compared to parallel-SGHMC. By contrast, MC-Dropout and Deep ensemble exhibit poorer calibration performance. Their SSE values of 0.0500 and 0.0012, respectively, and their higher NLL values (0.4678 and 0.4357) suggest that these models tend to overestimate uncertainty. This misalignment between predicted uncertainties and actual errors could lead to less reliable uncertainty estimates in practice, making these models less desirable for applications where accurate uncertainty quantification is important.

In terms of sharpness, parallel-SGHMC performs better than the other models, achieving the lowest RMV of 0.5137 and the highest C_V of 2.0178. This suggests that its uncertainty intervals are narrow and focused, indicating high confidence in its predictions. Cyclical-SGHMC is the next best model with an RMV of 0.6393 and a C_V of 1.6555, providing moderately sharp uncertainty intervals, which reflect a balance between confidence and calibration. On the other hand, MC-Dropout and Deep ensemble exhibit significantly less sharp uncertainty estimates, with RMV values of 0.9167 and 0.7905, respectively. Their lower C_V values (0.9897 for MC-Dropout and 1.0112 for Deep Ensemble) further indicate that these models generate broader uncertainty intervals, reflecting lower confidence in their predictions. While broader intervals can be beneficial in risk-averse or highly uncertain environments, they are less desirable in scenarios where tight confidence bounds are required.

In summary, the parallel-SGHMC model is recommended for tasks requiring accurate predictions, well-calibrated uncertainties, and narrow confidence intervals. Cyclical-SGHMC provides a balanced alternative, while MC-Dropout and Deep ensemble may be more appropriate for scenarios where broader, conservative uncertainty estimates are desired.

CONCLUSIONS

We proposed a novel approach to perform scalable uncertainty quantification for molecular property prediction with GNNs. Specifically, SGHMC with cyclical learning rate was researched for scalable uncertainty quantification in molecular property prediction with the D-MPNN architecture with mean variance. A comparative study was performed among SGHMC, Deep Ensemble and MC-dropout with the QM9 dataset with the heat capacity target property. Our results indicate that the proposed parallel-SGHMC approach outperforms MC-dropout and Deep ensembles in terms of calibration and sharpness. Specifically, parallel-SGHMC reduces the sum of squared errors (SSE) by 99.4% and 75%, respectively. These findings suggest that parallel-SGHMC is particularly suited for applications requiring narrow confidence intervals, where high precision in uncertainty quantification is essential, such as in safety-critical predictions or regulatory settings in pharmaceutical applications. While Deep ensemble and MC-dropout may offer advantages in scenarios where broader uncertainty bounds are required, such as in early-stage exploratory research.

REFERENCES

1. Schweidtmann AM, Rittig JG, Konig A, Grohe M, Mitsos A, Dahmen M. Graph neural networks for prediction of fuel ignition quality. Energy & Fuels 34:11395-11407 (2020) https://doi.org/10.1021/acs.energyfuels.0c01533
2. Wieder O, Kohlbacher S, Kuenemann M, Garon A, Ducrot P, Seidel T, Langer T. A compact review of molecular property prediction with graph neural networks. Drug Discov Today: Technol 37:1-12 (2020) https://doi.org/10.1016/j.ddtec.2020.11.009
3. Rittig JG, Mitsos A. Thermodynamics-consistent graph neural networks. Chem Sci 15(44):18504-18512 (2024) https://doi.org/10.1039/D4SC04554H
4. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. Inf Fusion 76:243-297 (2021) https://doi.org/10.1016/j.inffus.2021.05.008
5. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. Int Conf Mach Learn 1050-1059 (2016) https://doi.org/10.48550/arXiv.1506.02142
6. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst 30 (2017) https://doi.org/10.48550/arXiv.1612.01474
7. Wang F, Liu Y, Liu K, Wang Y, Medya S, Yu PS. Uncertainty in graph neural networks: a survey. arXiv preprint (2024) https://doi.org/10.48550/arXiv.2403.07185
8. Chen T, Fox E, Guestrin C. Stochastic gradient Hamiltonian Monte Carlo. Proc 31st Int Conf Mach Learn, PMLR 32(2):1683-1691 (2014) https://doi.org/10.48550/arXiv.1402.4102
9. Zhao Y, Yang C, Schweidtmann A, Tao Q. Efficient Bayesian uncertainty estimation for nnU-Net. Int Conf Med Image Comput Comput Assist Interv 535-544 (2022) https://doi.org/10.1007/978-3-031-16452-1_51
10. Busk J, Jørgensen PB, Bhowmik A, Schmidt MN, Winther O, Vegge T. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. Mach Learn Sci Technol 3(1):015012 (2021) https://doi.org/10.1088/2632-2153/ac3eb3
11. Neal RM. MCMC using Hamiltonian dynamics. In: Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC 113-162 (2011) https://doi.org/10.1201/b10905
12. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 1:1-7 (2014) https://doi.org/10.1038/sdata.2014.22
13. Heid E, Greenman KP, Chung Y, Li SC, Graff DE, Vermeire FH, et al. Chemprop: a machine learning package for chemical property prediction. J Chem Inf Model 64:9-17 (2023) https://doi.org/10.1021/acs.jcim.3c01250