MUBen: Benchmarking the Uncertainty of Molecular Representation Models

Yinghao Li[1], **Lingkai Kong**[1], **Yuanqi Du**[2], **Yue Yu**[1], **Yuchen Zhuang**[1], **Wenhao Mu**[1], **Chao Zhang**[1]
[1]{*yinghaoli, lkkong, yueyu, yczhuang, wmu30, chaozhang*}*@gatech.edu;* [2]*yd392@cornell.edu*
[1]*Georgia Institute of Technology;* [2]*Cornell University*

Reviewed on OpenReview: *https://openreview.net/forum?id=qYceFeHgm4*

Abstract

Large molecular representation models pre-trained on massive unlabeled data have shown great success in predicting molecular properties. However, these models may tend to overfit the fine-tuning data, resulting in over-confident predictions on test data that fall outside of the training distribution. To address this issue, uncertainty quantification (UQ) methods can be used to improve the models' calibration of predictions. Although many UQ approaches exist, not all of them lead to improved performance. While some studies have included UQ to improve molecular pre-trained models, the process of selecting suitable backbone and UQ methods for reliable molecular uncertainty estimation remains underexplored. To address this gap, we present MUBen, which evaluates different UQ methods for state-of-the-art backbone molecular representation models to investigate their capabilities. By fine-tuning various backbones using different molecular descriptors as inputs with UQ methods from different categories, we assess the influence of architectural decisions and training strategies on property prediction and uncertainty estimation. Our study offers insights for selecting UQ for backbone models, which can facilitate research on uncertainty-critical applications in fields such as materials science and drug discovery.

1 Introduction

The task of molecular representation learning is pivotal in scientific domains and bears the potential to facilitate research in fields such as chemistry, biology, and materials science (Walters & Barzilay, 2021). Nonetheless, supervised training typically requires vast quantities of data, which are challenging to acquire due to the need for expensive laboratory experiments (Eyke et al., 2020). With the advent of self-supervised Transformers (Vaswani et al., 2017) pioneered by BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), there has been a surge of interest in creating large-scale pre-trained molecular representation models from unlabeled datasets (Wang et al., 2019; You et al., 2020; Rong et al., 2020; Liu et al., 2022; Zhou et al., 2023; Stärk et al., 2022). Such models have demonstrated impressive representational capabilities, achieving state-of-the-art (SOTA) performance on a variety of molecular property prediction tasks (Wu et al., 2018).

However, in many applications, there is a pressing need for reliable predictions that are not only precise but also *uncertainty-aware*. The provision of calibrated uncertainty estimates allows us to distinguish "noisy" predictions to improve model robustness (Geifman & El-Yaniv, 2017), or estimate data distributions to enhance downstream tasks such as high throughput screening (Noh et al., 2020; Graff et al., 2021), activity cliff identification (van Tilborg et al., 2022), or wet-lab experimental design (Eyke et al., 2020; Soleimany et al., 2021). Unfortunately, large-scale models can easily overfit the fine-tuning data and exhibit misplaced confidence in their predictions (Guo et al., 2017; Kong et al., 2020). Several works have introduced various uncertainty quantification (UQ) methods in molecular property prediction (Ryu et al., 2019; Zhang & Lee, 2019; Scalia et al., 2020; Hwang et al., 2020; Mamun et al., 2020; Hie et al., 2020; Busk et al., 2021; Soleimany et al., 2021; Deshwal et al., 2021; Gruich et al., 2023; Wollschläger et al., 2023) and in neighboring research

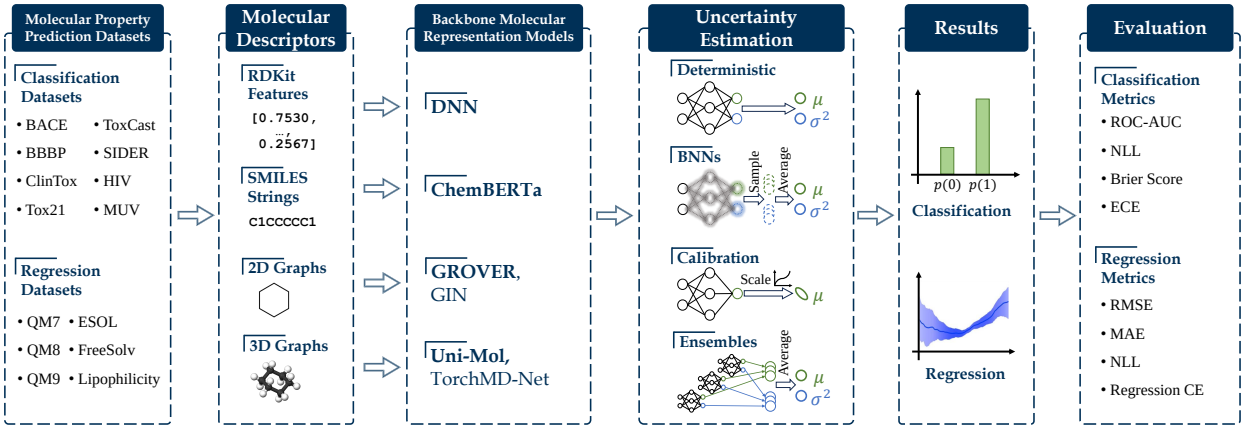


Figure 1: MUBen's pipeline with datasets, backbones, UQ methods and metrics enumerated.

areas such as protein engineering (Greenman et al., 2022). For instance, Busk et al. (2021) extend a message passing neural network (MPNN, Gilmer et al., 2017) with post-hoc recalibration (Guo et al., 2017) to address the overconfidence in molecular property prediction; Soleimany et al. (2021) apply evidential message passing networks (Sensoy et al., 2018; Amini et al., 2020) to quantitative structure-activity relationship regression tasks; Deshwal et al. (2021) utilized Bayesian optimization to conduct reliable predictions of Nanoporous material properties and to direct experiments.

Despite their contributions, these studies exhibit several constraints: **1)** the variety of uncertainty estimation methods considered is limited with some effective strategies being overlooked; **2)** each research focuses on a limited range of properties such as quantum mechanics, failing to explore a broader spectrum of tasks; and **3)** none embraces the power of recent pre-trained backbone models, which have shown remarkable performance in prediction but may lead to different results in UQ. To the best of our knowledge, a comprehensive evaluation of UQ methods applied to pre-trained molecular representation models is currently lacking and warrants further investigation.

We present MUBen, a benchmark designed to assess the performance of UQ methods applied to the molecular representation models for property prediction on various metrics, as illustrated in Figure 1. It encompasses UQ methods from different categories, including deterministic prediction, Bayesian neural networks, post-hoc calibration, and ensembles (§ 4.2), on top of a set of molecular representation (backbone) models, each relies on a molecular descriptor from a distinct perspective (§ 4.1). MUBen delivers intriguing results and insights that may guide the selection of backbone model and/or UQ methods in practice (§ 5). We structure our code, available at https://github.com/Yinghao-Li/MUBen, to be user-friendly and easily transferrable and extendable, with the hope that this work will promote the future development of UQ methods, pre-trained models, or applications within the domains of materials science and drug discovery.

2 Related Works

2.1 Pre-Training Molecular Representation Models

Existing pre-trained models for molecular representation learning can be distinguished into three categories based on the type of molecular descriptors. The first category includes ChemBERTa (Chithrananda et al., 2020), SMILES-BERT (Wang et al., 2019), Molformer (Ross et al., 2022), GPT-MolBERTa (Balaji et al., 2023), *etc.*, which leverage string-based input descriptors such as SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020), Group SELFIES (Cheng et al., 2023), or SAFE (Noutahi et al., 2023). Such models are generally built upon the Transformer encoder and pre-trained with masked language modeling (MLM) objectives (Devlin et al., 2019), which masks out a portion of tokens and trains the model to predict them using the context information.

Given that molecules can be represented as 2D topological graphs with atoms as vertices and chemical bonds as edges, the second category of pre-training strategies employs techniques such as context prediction (Hu et al., 2020; Rong et al., 2020), contrastive learning (Sun et al., 2020; You et al., 2020; Wang et al., 2022), and replaced component detection (Li et al., 2021; Kim et al., 2022) to learn molecular representation models with the awareness of the 2D atomic relationships. Graph neural networks (GNNs, Scarselli et al., 2009) as well as Transformer encoders are commonly selected as the backbone architectures (Xia et al., 2023).

Recent advancements in the development of 3D atomistic models have garnered significant attention, particularly due to their capability to capture the intricate three-dimensional nature of molecular structures. The body of research in this area is primarily divided into two categories: invariant models and equivariant models. Invariant models, including SchNet (Schütt et al., 2017), DimeNet (Klicpera et al., 2020), SphereNet (Liu et al., 2021), GemNet (Gasteiger et al., 2021), and ClofNet (Du et al., 2022), utilize scalar quantities that remain unchanged under Euclidean transformations such as rotations, translations, and reflections. These quantities, which include pairwise distances, angles, and dihedrals, serve as features for constructing message-passing neural networks. On the other hand, equivariant models like TFN (Thomas et al., 2018), Equiformer (Liao & Smidt, 2023), NequIP (Batzner et al., 2022), EGNN (Satorras et al., 2021), and LEFTNet (Du et al., 2023), implement message-passing mechanisms designed to maintain equivariance—ensuring that operations such as linear transformations or tensor products do not compromise the model's ability to handle geometric changes. Further insights and comprehensive analyses of these models are discussed in (Duval et al., 2023). Additionally, the pre-training objectives of these models emphasize the importance of 3D molecular properties. For instance, GEM (Fang et al., 2022) introduces bond angles and lengths as edge attributes to enhance 3D molecular insights, while Uni-Mol (Zhou et al., 2023) employs a unique combination of a 3D position denoising loss and a masked atom prediction loss, facilitating the effective learning of 3D spatial representations of molecules.

Beyond property prediction, pre-trained molecular representation models are leveraged for additional tasks such as chemical design (Taylor et al., 2022; Ghugare et al., 2023; Li et al., 2023a). For instance, Ghugare et al. (2023) delve into the potential of merging text-based generative models with reinforcement learning to synthesize molecules with targeted properties. This trend further emphasize the value of MUBen.

2.2 Uncertainty Quantification

Uncertainty estimation plays a crucial role in stimulating deep neural networks to acknowledge what they don't know, which has been applied to various domains such as object detection (He et al., 2019; Chen et al., 2021), decision making (Amiri et al., 2020; Kong et al., 2022), and text classification and information extraction (Chowdhury et al., 2011; Mukherjee & Awadallah, 2020; Yu et al., 2022; Li et al., 2022). UQ on neural networks generally takes two directions: Bayesian and non-Bayesian. Bayesian neural networks (BNNs) quantify predictive uncertainty by imposing probability distributions over model parameters instead of using point estimates. While BNNs provide a principled way of UQ, exact inference of parameter posteriors is often intractable. To circumvent this difficulty, existing works adopt Monte Carlo dropout (Gal & Ghahramani, 2016), variational inference (Blundell et al., 2015; Kingma et al., 2015) and Markov chain Monte Carlo (MCMC) (Welling & Teh, 2011; Maddox et al., 2019) for posterior approximation. Along the other direction, non-Bayesian methods estimate uncertainty by assembling multiple networks (Lakshminarayanan et al., 2017; Fort et al., 2019), adopting specialized output distribution (Sensoy et al., 2018; Amini et al., 2020; Soleimany et al., 2021) or training scheme (Lin et al., 2017; Mukhoti et al., 2020), or calibrating model outputs through additional network layers during validation and test (Platt et al., 1999; Guo et al., 2017; Song et al., 2019; Angelopoulos & Bates, 2021; Angelopoulos et al., 2021).

Recently, there have been several UQ tools (Ghosh et al., 2021; Chung et al., 2021; Detommaso et al., 2023) and benchmarks from the general domain (Nado et al., 2021; Schmähling et al., 2023), biomedical engineering (Mehrtens et al., 2023; Greenman et al., 2022) and materials science (Hirschfeld et al., 2020; Hwang et al., 2020; Scalia et al., 2020; Varivoda et al., 2023; Wollschläger et al., 2023). MUBen also targets the last domain but covers a more diverse suite of tasks, stronger pre-trained molecular representation models, and a more comprehensive set of UQ methods. We hope MUBen can serve as a standard benchmark to faithfully evaluate the ability of UQ approaches.

3 Problem Setup

Let $\boldsymbol{x}$ denote a descriptor of a molecule, such as SMILES or a 2D topological graph. Variable y is the label with domain $y \in \{1, \ldots, K\}$ for K-class classification or $y \in \mathbb{R}$ for regression. $\boldsymbol{\theta}$ denotes all parameters of the molecular representation network and the task-specific layer. We assume that a training dataset $\mathbb{D}$ consists of N i.i.d. samples $\mathbb{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. Our goal is to fine-tune the parameters $\boldsymbol{\theta}$ on $\mathbb{D}$ in order to develop a predictive model that accurately estimates the probability distribution $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_n)$. The UQ metrics are briefly illustrated below with more details being provided in appendix D.

- Negative Log-Likelihood (**NLL**) is often utilized to assess the quality of model uncertainty on holdout sets for both classification and regression tasks. Despite being a proper scoring rule as per Gneiting's framework (Gneiting & Raftery, 2007), certain limitations such as overemphasizing tail probabilities (Quiñonero-Candela et al., 2006) make it inadequate to serve as the only UQ metric.
- The Brier Score (**BS**, Brier, 1950) quantifies the mean squared error between predicted probabilities and actual labels, and is recognized as a proper scoring rule for classification tasks. The formula for the Brier Score is given by:

$$\mathrm{BS} = \frac{1}{N} \sum_{n=1}^{N} \sum_{y=1}^{K} \left(p_{\boldsymbol{\theta}}(y|\boldsymbol{x}_n) - \mathbb{I}(y = y_n)\right)^2,$$

  where $\mathbb{I}(\cdot)$ is the indicator function. Specifically, $\mathbb{I}(a = b) = 1$ if $a = b$; otherwise, $\mathbb{I}(a = b) = 0$.
- Calibration Errors measure the correspondence between predicted probabilities and empirical accuracy. For classification, we use Expected Calibration Error (**ECE**, Naeini et al., 2015), which first divides the predicted probabilities $\{p_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{x}_n)\}_{n=1}^N$ into S bins $\mathbb{B}_s = \{n \in \{1, \ldots, N\} \mid p_{\boldsymbol{\theta}}(y_n \mid \boldsymbol{x}_n) \in (\rho_s, \rho_{s+1}]\}$ with $s \in \{1, \ldots, S\}$, each with size N_s, and then compute the L_1 loss between the accuracy and predicted probabilities within each bin:

$$\mathrm{ECE} = \sum_{s=1}^{S} \frac{N_s}{N} \left| \frac{1}{N_s} \sum_{n \in \mathbb{B}_s} \mathbb{I}(y_n = \hat{y}_n) - \frac{1}{N_s} \sum_{n \in \mathbb{B}_s} p_{\boldsymbol{\theta}}(\hat{y}_n \mid \boldsymbol{x}_n) \right|,$$

  where $\hat{y}_n = \arg\max_y p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x}_n)$ is the n-th prediction.
  For regression tasks, the Regression Calibration Error (**CE**, Kuleshov et al., 2018) measures the accuracy of prediction intervals. It calculates the true frequency of the predicted points falling within each confidence interval against the predicted fraction of points in that interval:

$$\mathrm{CE} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{s}{S} - \frac{1}{N} \left| \{n \in \{1, \ldots, N\} \mid F_{\boldsymbol{\theta}}(y_n; \boldsymbol{x}_n) \leq \frac{s}{S}\} \right| \right)^2,$$

  where $\frac{s}{S}$ represents the expected quantile. $F_{\boldsymbol{\theta}}(y_n; \boldsymbol{x}_n)$ denotes the predicted quantile value for y_n, such as a Gaussian cumulative distribution function $\Phi(y_n; \hat{y}_n, \hat{\sigma}_n)$, which is parameterized by the estimated mean $\hat{y}_n$ and variance $\hat{\sigma}_n$ assuming a Gaussian distribution for the labels.

4 Experiment Setup

4.1 Backbone Models

Molecular descriptors, such as fingerprints, SMILES strings, and graphs, package the structural data of a molecule into a format recognizable for computational algorithms. When paired with model architectures carrying divergent inductive biases, these descriptors can offer varied benefits depending on the specific task. Therefore, we select 4 primary backbone models that accept distinct descriptors as input. We also include 2 supplementary backbones to cover other aspects such as model architecture and pre-training objectives.

For primary backbone models, we select various pre-trained Transformer encoder-based models, each is SOTA or close to SOTA for their respective input format: **1) ChemBERTa** (Chithrananda et al., 2020; Ahmad et al., 2022), which accepts SMILES strings as input; **2) GROVER** (Rong et al., 2020), which pre-trains Transformer encoders augmented with the dynamic message-passing GNN on 2D molecular graphs ; and

3) Uni-Mol (Zhou et al., 2023), which incorporates 3D molecular conformations into its input, and pre-trains specialized Transformer architectures to encode the invariant spatial positions of atoms and represent the edges between atom pairs. We also implement a **4)** fully connected deep neural network (**DNN**) that uses 200-dimensional RDKit features which are regarded as fixed outputs of a hand-crafted backbone feature generator. This simple model aims to highlight the performance difference between heuristic feature generators and the automatic counterparts that draw upon self-supervised learning processes for knowledge acquisition. In addition, we have two other backbones: **5) TorchMD-NET**, as introduced by Thölke & Fabritiis (2022) and pre-trained according to Zaidi et al. (2023). This model is designed to handle 3D inputs and leverages an equivariant Transformer architecture aimed at quantum mechanical properties. **6) GIN** from Xu et al. (2019), which processes 2D molecular graphs. GIN is theoretically among the most powerful GNNs available, offering an additional randomly initialized backbone baseline with less sophisticated input features. Compared with the primary one, these models are constrained in pre-training: TorchMD-NET is exclusively pre-trained on quantum mechanical data, whereas GIN does not involve any pre-processed features. Please refer to appendix B for more detailed introduction on backbone architectures and hyperparameters.

4.2 Uncertainty Quantification Methods

In MUBen, we choose popular UQ methods from different categories. We provide a high-level sketch here of each method and leave the details to appendix C.

Deterministic Uncertainty Prediction In standard practice, deep classification networks utilize SoftMax or Sigmoid outputs to distribute prediction weights among target classes. Such weights serve as an estimate of the model's uncertainty. In regression tasks, rather than one single output value, models often parameterize an independent Gaussian distribution with predicted means and variances for each data point, and the variance magnitude acts as an estimate of the model's uncertainty.

An alternate loss function for classification is the **Focal Loss** (Lin et al., 2017; Mukhoti et al., 2020), which minimizes a regularised KL divergence between the predicted values and the true labels. The regularization increases the entropy of the predicted distribution while the KL divergence is minimized, mitigating the model overconfidence and improving the uncertainty representation.

Bayesian Learning and Inference BNNs estimate the probability distribution over the parameters of the network. Specifically, BNNs operate on the assumption that the network layer weights follow a specific distribution that is optimized through maximum a posteriori estimation during training. During inference, multiple network instances are randomly sampled from the learned distributions. Each instance then makes independent predictions, and the prediction distribution inherently captures the desired uncertainty information. Some notable methods that employ this approach include:

- Bayes by Backprop (**BBP**, Blundell et al., 2015) first introduced Monte Carlo gradients, an extension of the Gaussian reparameterization trick (Kingma & Welling, 2014), to learn the posterior distribution of network weights directly through backpropagation.
- Stochastic Gradient Langevin Dynamics (**SGLD**, Welling & Teh, 2011) applies Langevin dynamics to infuse noise into the stochastic gradient descent training process, thereby mitigating the overfitting of network parameters. The samples generated via Langevin dynamics can be used to form Monte Carlo estimates of posterior expectations during the inference stage.
- **MC Dropout** (Gal & Ghahramani, 2016) demonstrates that applying dropout is equivalent to a Bayesian approximation of the deep Gaussian process (Damianou & Lawrence, 2013). Uncertainty is derived from an ensemble of multiple stochastic forward passes with dropout enabled.
- SWA-Gaussian (**SWAG**, Maddox et al., 2019) provides an efficient way of estimating Gaussian posteriors over network weights by utilizing stochastic weight averaging (SWA, Izmailov et al., 2018) combined with low-rank & diagonal Gaussian covariance approximations.

Post-Hoc Calibration Post-hoc calibration is proposed to address the over-confidence issue of deterministic classification models by adjusting the output logits after training. The most popular method is **Temperature Scaling** (Platt et al., 1999; Guo et al., 2017), which adds a learned scaling factor to the Sigmoid or SoftMax

output activation to control their "spikiness". Guo et al. (2017) shows that this simple step improves model calibration in general, yielding better UQ performance.

Deep Ensembles The ensemble method has long been used in machine learning to improve model performance (Dietterich, 2000) but was first adopted to estimate uncertainty in (Lakshminarayanan et al., 2017). It trains a deterministic network multiple times with different random seeds and combines their predictions at inference. The ensembles can explore training modes thoroughly in the loss landscape and thus are robust to noisy and out-of-distribution (OOD) examples (Fort et al., 2019).

4.3 Datasets

We carry out our experiments on MoleculeNet (Wu et al., 2018), a collection of widely utilized datasets covering molecular properties such as quantum mechanics, solubility, and toxicity. For classification, MUBen incorporates BBBP, ClinTox, Tox21, ToxCast, SIDER, BACE, HIV, and MUV with all labels being binary; the first 5 concern **physiological** properties and the last 3 **biophysics**. For regression, we select ESOL, FreeSolv, Lipophilicity, QM7, QM8, and QM9, of which the first 3 contain **physical chemistry** properties and the rest fall into the **quantum mechanics** category.

In line with previous studies (Fang et al., 2022; Zhou et al., 2023), MUBen divides all datasets with scaffold splitting to minimize the impact of dataset randomness and, consequently, enhances the reliability of the evaluation. Moreover, scaffold splitting alienates the molecular features in each dataset split, inherently creating a challenging OOD setup that better reflects the real-world scenario. For comparison, we also briefly report the results of random splitting. Please check appendix A for detailed descriptions.

4.4 Training and Evaluation Protocols

We use Sigmoid output activation and binary cross-entropy (BCE) loss function on all classification models unless the UQ method specifies a different objective. For regression, we standardize the labels by mapping the training labels to a standard Gaussian distribution before model training and evaluation to mitigate the impact of label magnitude on multi-task datasets. During inference, the predicted values are converted back to their original distribution according to the label mean and variance from the training set before computing the metrics. The numbers of tasks are available in Table 4. A classification model features one output head for each prediction task, while regression models have two—one for predicting mean $\hat{y}$ and the other for variance $\hat{\sigma}$, which is guaranteed to be positive by applying a SoftPlus activation upon the output logits (Lakshminarayanan et al., 2017). Regression models are trained with Gaussian NLL objective. We universally apply validation-based early stopping unless the UQ method has other requirements.

Consistent with recommendations from Wu et al. (2018) and other previous works (Fang et al., 2022; Zhou et al., 2023), we report **ROC-AUC** (area under the receiver operating characteristic curve) as the metric for classification prediction and **RMSE** (root-mean-square error) and **MAE** (mean absolute error) for regression. In quantifying classification uncertainty, we use **ECE**, **NLL**, and **Brier Score** as introduced in § 3. For regression, we compute the **Gaussian NLL** and Regression **CE**. We compute the metrics for each task individually before calculating their macro average. Each reported metric is the average of 3 individual training-test runs with random seeds 0, 1, and 2. This differs from Deep Ensembles, which aggregate model predictions prior to metric calculation.

5 Results and Analysis

Tables 1 and 2 present the macro-averaged rankings of the primary backbone-UQ combinations across all datasets, as well as the metric scores on 4 representative datasets, two for classification and two for regression. Figure 2 provides further insights into primary backbone model performance by visualizing the Mean Reciprocal Ranks (MRRs) with averaged across UQ methods. Compared to the ranks in the tables, MRRs accentuate top results, offering a complementary perspective on the backbone performance. Our analysis initiates with how the UQ methods contribute to the models' property prediction, followed by an exploration of the uncertainty estimation performance.

Table 1: Classification results. "↑" and "↓" imply that better performance is indicated by a larger or smaller value, respectively. Text in bold signifies the top-performing UQ method within each backbone model, and cells in blue indicate the best performance across all backbone-UQ combinations. "ROC" represents ROC-AUC. Deep Ensembles consistently outperforms other UQ methods for both property prediction and uncertainty quantification; MC Dropout and Temperature Scaling also consistently improve UQ performance.

	Tox21[a]				ToxCast[a]				Average Ranking[b]			
	ROC↑	ECE↓	NLL↓	BS↓	ROC↑	ECE↓	NLL↓	BS↓	ROC↓	ECE↓	NLL↓	BS↓
DNN-RDKit												
Deterministic	0.7386	0.0417	0.2771	0.0779	0.6222	0.1168	0.4436	0.1397	25.75	17.25	24.13	22.88
Temperature	0.7386	**0.0342**	0.2723	0.0773	0.6220	0.1114	0.4882	0.1398	25.75	15.38	21.25	19.88
Focal Loss	0.7374	0.1058	0.3161	0.0871	0.6289	0.1264	0.4389	0.1396	25.88	24.38	22.38	24.00
MC Dropout	0.7376	0.0356	0.2727	0.0763	0.6248	0.1093	0.4319	0.1358	26.50	13.00	19.38	18.63
SWAG	0.7364	0.0438	0.2793	0.0790	0.6207	0.1175	0.4441	0.1400	26.38	18.50	25.63	23.50
BBP	0.7243	0.0422	0.2847	0.0814	0.6020	0.1443	0.4673	0.1510	22.75	19.13	19.38	21.38
SGLD	0.7257	0.1192	0.3455	0.0978	0.5319	0.3054	0.6685	0.2378	27.75	26.00	28.88	27.88
Ensembles	**0.7540**	0.0344	**0.2648**	**0.0746**	**0.6486**	**0.0900**	**0.4008**	**0.1292**	**20.00**	**7.13**	**11.75**	**13.13**
ChemBERTa												
Deterministic	0.7542	0.0571	0.2962	0.0812	0.6554	0.1209	0.4313	0.1330	15.63	17.38	18.88	19.38
Temperature	0.7542	0.0424	0.2744	0.0792	0.6540	0.1067	0.4817	0.1313	15.88	12.00	13.88	15.38
Focal Loss	0.7523	0.0969	0.3052	0.0845	0.6442	0.1197	0.4243	0.1346	17.63	20.13	17.88	20.63
MC Dropout	0.7641	0.0423	0.2697	**0.0744**	0.6624	0.1069	0.4070	0.1276	12.50	**10.75**	**10.63**	**10.00**
SWAG	0.7538	0.0592	0.3008	0.0818	0.6556	0.1202	0.4305	0.1327	16.13	19.50	21.38	20.38
BBP	0.7433	0.0459	0.2765	0.0780	0.5814	0.1276	0.4545	0.1469	20.88	19.00	16.00	19.50
SGLD	0.7475	0.0504	0.2784	0.0795	0.5436	0.2238	0.5602	0.1881	21.13	19.88	19.13	18.63
Ensembles	**0.7681**	**0.0440**	**0.2679**	0.0750	**0.6733**	**0.1037**	**0.3986**	**0.1258**	**12.38**	13.00	11.63	12.25
GROVER												
Deterministic	0.7808	0.0358	0.2473	0.0694	0.6587	0.1043	0.4091	0.1298	11.63	11.50	8.88	9.88
Temperature	0.7810	**0.0291**	0.2439	0.0686	0.6496	0.1424	0.4612	0.1424	12.63	8.88	7.50	9.25
Focal Loss	0.7779	0.1148	0.3052	0.0811	0.6359	0.1221	0.4365	0.1383	15.00	23.38	21.50	22.25
MC Dropout	0.7817	0.0346	0.2455	0.0689	0.6615	**0.1009**	**0.4042**	**0.1288**	11.25	10.50	7.63	8.63
SWAG	0.7837	0.0360	0.2482	0.0689	0.6603	0.1060	0.4114	0.1301	9.13	11.63	8.88	8.38
BBP	0.7697	0.0438	0.2552	0.0711	0.5995	0.1731	0.5090	0.1660	16.75	22.00	14.75	15.88
SGLD	0.7635	0.0402	0.2558	0.0716	0.5542	0.2712	0.6194	0.2139	18.75	18.63	15.25	15.63
Ensembles	**0.7876**	0.0316	**0.2411**	**0.0675**	**0.6646**	0.1034	0.4061	0.1290	**8.50**	**8.13**	**5.38**	**6.88**
Uni-Mol												
Deterministic	0.7895	0.0454	0.2601	0.0716	0.6734	0.1020	0.3983	0.1274	11.50	15.50	16.13	13.88
Temperature	0.7896	0.0346	0.2483	0.0704	**0.7028**	0.1456	0.4566	0.1355	11.00	12.00	13.13	12.75
Focal Loss	0.7904	0.0972	0.2899	0.0785	0.6934	0.1227	0.4079	0.1284	10.00	23.50	19.38	20.50
MC Dropout	0.7891	0.0480	0.2628	0.0726	0.6833	0.1074	0.4015	0.1274	12.88	19.88	19.25	17.25
SWAG	0.7842	0.0593	0.2994	0.0728	0.6870	0.1085	0.4005	0.1271	10.13	22.13	22.25	18.25
BBP	0.7932	0.0396	0.2520	0.0703	0.6273	0.1296	0.4522	0.1456	12.13	17.00	15.50	16.25
SGLD	0.7887	0.0433	0.2569	0.0684	0.5700	0.1953	0.5207	0.1717	14.75	18.13	18.88	14.50
Ensembles	**0.8052**	**0.0332**	**0.2389**	**0.0662**	0.6841	**0.0953**	**0.3877**	**0.1247**	**5.13**	**8.88**	**7.63**	**6.50**
TorchMD-NET[c]	0.7793	0.0409	0.2614	0.0708	0.6540	0.1546	0.4424	0.1396	-	-	-	-
GIN[c]	0.6829	0.0634	0.3268	0.0840	0.5752	0.1381	0.4835	0.1477	-	-	-	-

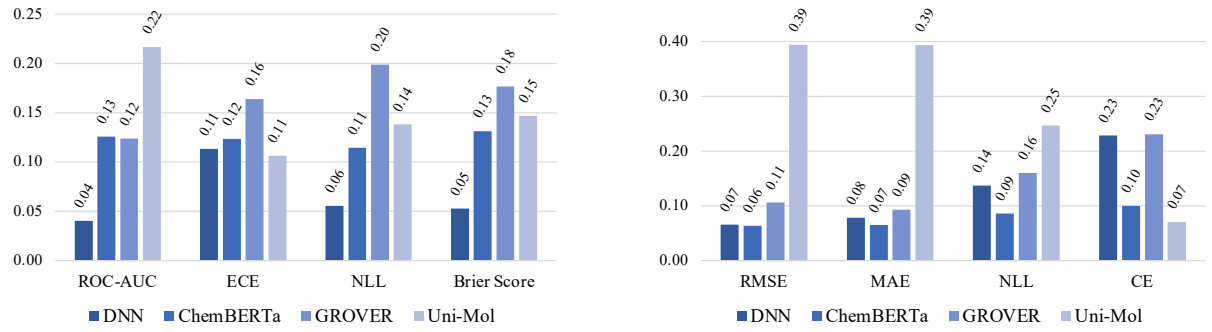
[a] The "Tox21" and "ToxCast" columns present metric scores on representative exemplar datasets, highlighting the trends observable across all datasets.
[b] The "Average Ranking" columns provide the rank of each model's UQ metrics against all other backbone-UQ combinations averaged from all classification datasets; smaller number indicates better performance.
[c] We report the results from the best-performing UQ method—Deep Ensembles for TorchMD-NET and GIN. TorchMD-NET and GIN are not ranked together with the primary backbones while reporting the "Average Ranking".

Property Prediction Performance Theoretically, inserting UQ methods into the training pipeline does not guarantee better property prediction on datasets containing i.i.d. instances. However, since we ensure OOD test points with scaffold splitting, UQ methods may mitigate the distribution gap, yielding better test results. The columns ROC-AUC, RMSE, and MAE in Tables 1, 2, and Figure 2 illustrate the predictive performance of each method. Examining from the lens of UQ methods, none provides a consistent guarantee of performance improvement over the deterministic baseline, except for Deep Ensembles, and MC Dropout for regression. The randomness in the initialization and training trajectory of Deep Ensembles explores a broader range of loss landscapes, which partially addresses the distribution shift issue, as observed in previous works (Garipov et al., 2018; Fort et al., 2019; Fang et al., 2023). MC Dropout samples multiple sets of model parameters from the Gaussian distributions centered at the fine-tuned deterministic net weights, which may flatten extreme regression abnormality triggered by OOD features. This phenomenon is less pronounced for classification due to the $(0, 1)$ output domain. However, other BNNs do not exhibit the same advantage. SWAG, while similar to MC Dropout *w.r.t.* training, might intensify training data overfitting due to the additional steps taken to fit the parameter distribution. While mitigating overfitting, the stochastic sampling methods used in BBP and SGLD can hinder network convergence, leading to insufficient exploration of

Table 2: The regression outcomes for property prediction and uncertainty quantification, where lower scores are preferable, cover two example datasets along with the average ranking. Similar to classification, Deep Ensembles consistently outperforms other methods. Specifically within the context of regression, BBP and SGLD demonstrate exceptional capabilities in estimating uncertainty, despite not consistently improving property prediction outcomes.

	Lipophilicity				QM9				Average Ranking			
	RMSE	MAE	NLL	CE	RMSE	MAE	NLL	CE	RMSE	MAE	NLL	CE
DNN-RDKit												
Deterministic	0.7575	0.5793	0.6154	0.0293	0.01511	0.01012	-3.379	0.04419	16.67	15.67	11.33	10.50
MC Dropout	0.7559	0.5773	0.9071	0.0341	0.01480	0.01000	-3.526	0.04327	15.17	15.17	11.33	10.67
SWAG	0.7572	0.5823	0.7191	0.0308	0.01524	0.01019	-3.284	0.04495	18.00	17.67	14.00	12.50
BBP	0.7730	0.5938	0.7578	0.0305	0.01534	0.01025	-3.347	0.04452	21.17	20.50	10.33	7.33
SGLD	0.7468	0.5743	**0.2152**	**0.0090**	0.01958	0.01437	-3.335	**0.00702**	19.83	19.33	6.83	**1.83**
Ensembles	**0.7172**	**0.5490**	0.6165	0.0322	**0.01430**	**0.00956**	**-3.602**	0.04362	**11.50**	**11.17**	**6.33**	8.67
ChemBERTa												
Deterministic	0.7553	0.5910	1.2368	0.0362	0.01464	0.00916	-2.410	0.05468	17.83	16.67	18.83	14.67
MC Dropout	**0.7142**	**0.5601**	0.8178	0.0349	0.01412	0.00880	-3.150	0.05133	**14.33**	**13.83**	15.00	13.67
SWAG	0.7672	0.5992	1.5809	0.0395	0.01477	0.00925	-2.170	0.05535	19.33	18.50	20.33	15.83
BBP	0.7542	0.5869	**0.4419**	**0.0279**	0.01443	0.00928	-2.593	0.05399	17.67	18.50	10.83	9.00
SGLD	0.7622	0.5982	0.8719	0.0355	0.01530	0.01012	**-3.758**	**0.03378**	19.83	20.50	**9.50**	**8.50**
Ensembles	0.7367	0.5763	0.9756	0.0360	**0.01397**	**0.00868**	-2.876	0.05425	14.83	**13.83**	14.67	12.67
GROVER												
Deterministic	0.6316	0.4747	2.1512	0.0478	0.01148	0.00678	-0.787	0.06206	10.67	11.67	17.67	16.00
MC Dropout	0.6293	0.4740	2.0526	0.0476	0.01140	0.00676	-1.100	0.06161	9.33	11.17	16.00	14.33
SWAG	0.6317	0.4750	2.3980	0.0485	0.01156	0.00678	-0.477	0.06252	12.33	13.33	20.33	17.83
BBP	0.6481	0.5058	0.0789	0.0196	0.01179	0.00700	-1.885	0.05909	14.17	15.67	7.67	4.00
SGLD	0.6360	0.4984	**0.0544**	**0.0215**	0.01359	0.00878	**-3.785**	**0.02911**	14.83	15.17	**4.67**	**2.67**
Ensembles	**0.6250**	**0.4693**	1.6046	0.0460	**0.01143**	**0.00667**	-1.028	0.06199	**8.17**	**8.67**	13.50	14.00
Uni-Mol												
Deterministic	0.6079	0.4509	0.8975	0.0425	0.00962	0.00538	0.014	0.06637	5.83	4.83	16.67	21.17
MC Dropout	0.5983	0.4438	1.3663	0.0440	0.00961	0.00535	-0.251	0.06615	4.00	3.17	16.67	21.33
SWAG	0.6026	0.4476	1.0101	0.0453	0.00969	0.00541	-0.462	0.06597	6.33	6.17	19.67	22.67
BBP	0.6044	0.4469	**0.0679**	**0.0306**	0.00952	0.00544	-2.959	0.06179	3.17	3.00	4.33	10.33
SGLD	0.6040	0.4554	0.1565	0.0329	0.00950	0.00546	**-4.209**	**0.04593**	**2.50**	4.00	**2.50**	**9.17**
Ensembles	**0.5809**	**0.4266**	0.6450	0.0438	**0.00948**	**0.00526**	-0.319	0.06629	**2.50**	**1.83**	11.00	20.67
TorchMD-NET[(a)]	1.0313	0.8196	0.8619	0.0195	0.00860	0.00464	2.262	0.06868	-	-	-	-
GIN[(a)]	0.8071	0.6515	0.3241	0.0020	0.01295	0.00814	-3.521	0.04997	-	-	-	-

[(a)] We report the results from the Deep Ensembles UQ method for TorchMD-NET and GIN.



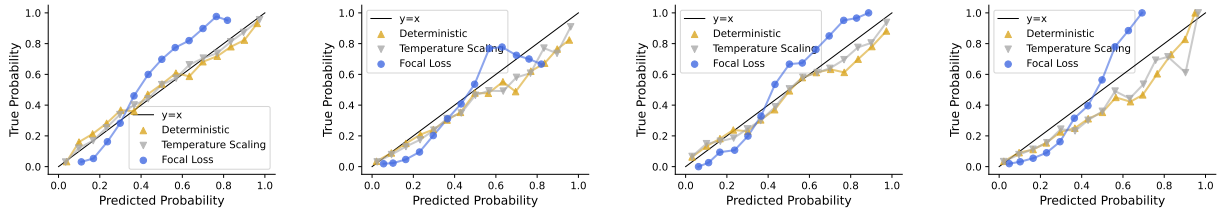
(a) Classification MRR↑

(b) Regression MRR↑

Figure 2: MRR of the backbone models; each metric is macro-averaged from the reciprocal ranks of all corresponding UQ methods across all datasets. Uni-Mol stands out by consistently surpassing its counterparts in property prediction accuracy for both classification and regression tasks. Conversely, GROVER exhibits a more balanced performance in both prediction accuracy and UQ across various tasks.

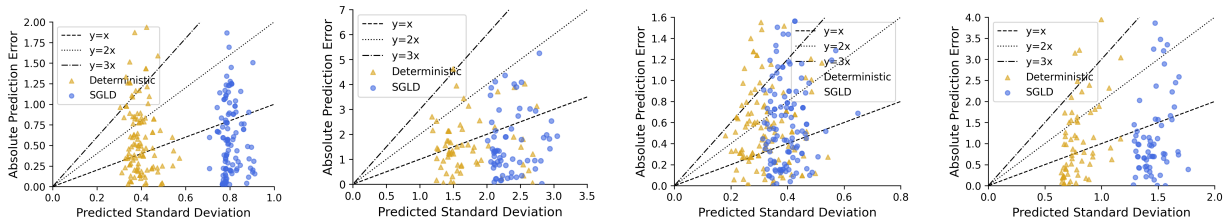
the feature and loss spaces, and consequently, adversely affecting prediction performance. This impact is particularly acute in smaller models, which are inherently less expressive, as illustrated in Figure 10.

Looking from the perspective of primary backbones, Uni-Mol secures the best prediction performance for both classification and regression. The superior molecular representation capability of Uni-Mol is attributed to the large network size, the various pre-training data and tasks, and the integration of results from different conformations of the same molecule (Zhou et al., 2023). When contrasting DNN, ChemBERTa, and GROVER,



(a) Uni-Mol on SIDER (b) Uni-Mol on Tox21 (c) DNN on SIDER (d) DNN on Tox21

Figure 3: The calibration plot of UQ methods. Here, "true probability" refers to the empirical prediction accuracy, and "predicted probability" refers to the average predicted probability within each bin (§ 3). These results reveal that Temperature Scaling typically enhances calibration effectiveness in the majority of scenarios, whereas Focal Loss tends to lead to over-calibration of the model, producing under-confident results.



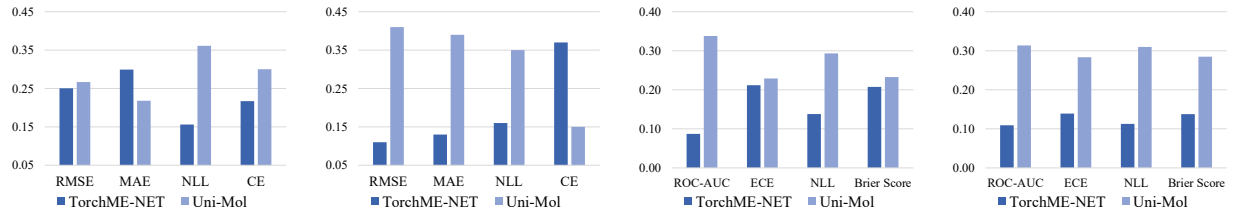
(a) DNN on Lipo (b) DNN on FreeSolv (c) ChemBERTa on Lipo (d) Uni-Mol on FreeSolv

Figure 4: The absolute error between the model-predicted mean and true labels against the predicted standard deviation. We compare the performance of SGLD with the deterministic prediction on different backbones and datasets. The "$y = kx$" lines indicate whether the true labels lie within the k-std range of the predicted Gaussian. Also, a model is perfectly calibrated when its output points are arranged on an "$y = kx$" line for an arbitrary k. Notably, SGLD is observed to generate a larger variance for OOD samples, which tends to correspond more closely with the prediction errors on average.

it becomes apparent that the expressiveness of the molecular descriptors varies for different molecules/tasks. Moreover, pre-trained models do not invariably surpass heuristic features when integrated with UQ methods, which aligns with the findings of Deng et al. (2023). Selecting a model attuned to the specific task is advised over an indiscriminate reliance on "deep and complex" networks.

Uncertainty Quantification Performance Tables 1, 2 and Figure 2 depict the uncertainty estimation performances via ECE, NLL, Brier Score, and CE columns. One discernible trend is the consistent performance enhancement from Deep Ensembles even when the number of ensembles is limited, such as the QM9 case (appendix C). MC Dropout also exhibits a similar trend, albeit less pronounced. Despite a marginal compromise in prediction accuracy, Temperature Scaling emerges as another method that almost invariably improves calibration, as evidenced in the average ranking columns—a finding that aligns with Guo et al. (2017). Figure 3 shows that deterministic prediction tends to be over-confident, and Temperature Scaling mitigates this issue. However, it is susceptible to calibration-test alignment. If the correlation is weak, Temperature Scaling may worsen the calibration, as presented in the ToxCast dataset. In contrast, Focal Loss does not work as impressively for binary classification, which is also observed by Mukhoti et al. (2020). It is hypothesized that while the Sigmoid function sufficiently captures model confidence in binary cases, Focal Loss could over-regularize parameters, diminishing the prediction sharpness to an unreasonable value, a conjecture verified by the "S"-shaped calibration curves in Figure 3.

Although limited in classification efficacy, both BBP and SGLD deliver commendable performance in predicting regression uncertainty, capturing 7 out of 8 top ranks for 4 backbones on two metrics, narrowly missing out to Deep Ensembles for the remaining one (Table 2). Yet, their inconsistent improvement of RMSE and



(a) Quantum Mechanics (b) Physical Chemistry (c) Biophysics (d) Physiology

Figure 5: MRRs of TorchMDNet and Uni-Mol on datasets grouped by dataset property categories. MRR calculations are confined to results from these two backbones. Only relative values matter. TorchMDNet is comparible to Uni-Mol on Quantum Mechanics properties, where it is pre-trained on, but is outperformed by Uni-Mol on all other property categories.

MAE implies a greater influence on variance prediction than the mean. Figure 4 reveals SGLD's tendency to "play safe" by predicting larger variances, while the deterministic method is prone to over-confident by ascribing small variances even to its inaccurate predictions. In addition, we do not observe a better correlation between SGLD's error and variance. We assume that the noisy training trajectory prevents SGLD and BBP from sufficiently minimizing the gap between the predicted mean and true labels, thus encouraging them to maintain larger variances to compensate for the error. Please refer to appendix C.2 and appendix E.1 for more analysis on UQ performance.

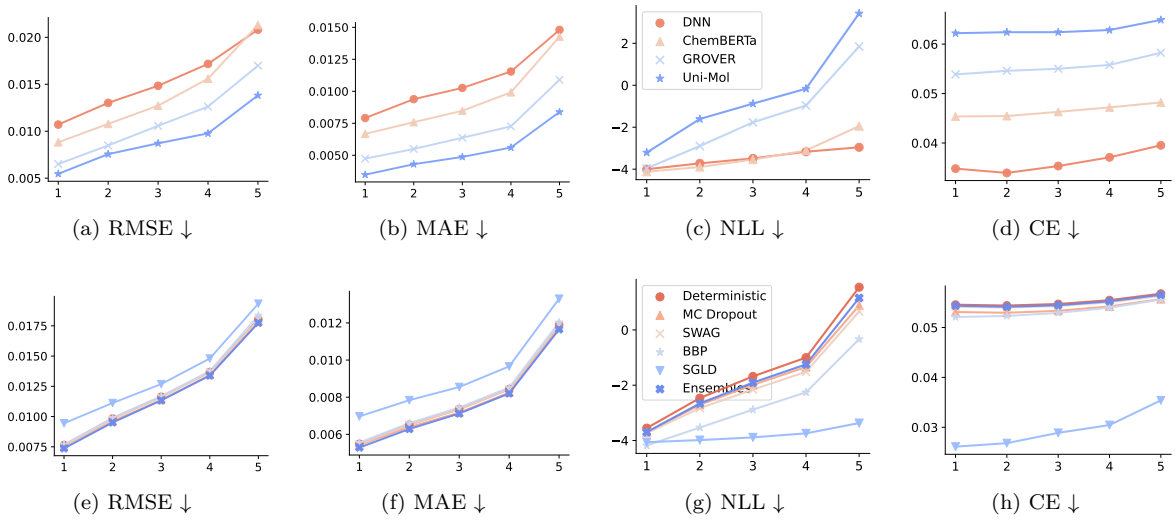
Figure 2 indicates that Uni-Mol exhibits subpar calibration, particularly for regression. Comparing Uni-Mol, ChemBERTa, and DNN in Figure 4, we notice that larger models such as Uni-Mol are more confident in their results, as illustrated by their smaller variances and larger portion of (std, error) points exceeding $y = kx$. Apart from the inherent susceptibility of larger models to overfitting (Ying, 2019), such phenomenon for Uni-Mol could also be attributed to shared structural features in 3D conformations in the training and test molecules that remain unobserved in simpler descriptors (Zhou et al., 2023). While this similarity benefits property prediction, it also potentially misleads the model into considering data points as in-distribution, thereby erroneously assigning high confidence. Overall, our findings broadly support the notion that models with lower expressiveness tend to exhibit better calibration. Therefore, the selection of an appropriate UQ method is more critical for enhancing the calibration of larger models. This is particularly evident from the larger discrepancies in calibration errors between the Deterministic approach and the most effective UQ method for Uni-Mol compared to DNN, as demonstrated in Tables 1, 2, and the following discussion.

TorchMD-NET and GIN Our analysis also encompasses TorchMD-NET and GIN, two additional backbone models excluded from the primary benchmark due to their limited capabilities. As presented in the tables and Figure 5, TorchMD-NET's performance is on par with Uni-Mol when predicting quantum mechanical properties but falls short in others. This outcome aligns with expectations, given that TorchMD-NET's architecture is tailored specifically for predicting quantum mechanical properties (Thölke & Fabritiis, 2022). Moreover, it is pre-trained on the relatively niche dataset PCQM4Mv2 (Hu et al., 2021) with only the denoising objective, which might be suitable for molecular dynamics but limited for other properties. In contrast, Uni-Mol stands out as a versatile model, benefiting from diverse pre-training objectives that ensure superiority across various tasks. On the other hand, GIN's performance is consistently inferior to other models including DNN, with examples presented in Tables 1 and 2, likely due to the limited expressiveness of 2D graphs and the GNN architecture when pre-training is absent. Appendix E presents the complete results.

Frozen Backbone and Randomly Split Datasets Table 3 demonstrates a notable drop in prediction performance when backbone weights are fixed. Also, models evaluated on random splits outperform scaffold splits. This is consistent with intuition: if backbone models serve solely as feature extractors instead of a part of the trainable predictors, they are less expressive for downstream tasks. Additionally, in-distribution features tend to be more predictable. An interesting exception emerges in regression calibration error, where frozen backbones perform better and random splits score lower. Upon examining the predicted values, we

Table 3: Performance with frozen backbone weights and on random split datasets compared with the original scores in Table 1 and 2. The result is calculated as (new − original)/original and is macro-averaged over all datasets, backbones and UQ methods. The performance of the frozen backbone is significantly worse than the original, while the model performs better on the randomly split (in-domain) test set.

	Classification				Regression			
	ROC-AUC (%)↑	ECE (%)↓	NLL (%)↓	Brier Score (%)↓	RMSE (%)↓	MAE (%)↓	NLL (%)↓	CE (%)↓
Frozen Backbone	-24.07	145.13	53.43	88.98	78.60	92.40	58.18	-46.06
Random Split	13.25	-35.19	-33.87	-37.35	-26.34	-31.36	-53.87	19.06



(a) RMSE ↓ (b) MAE ↓ (c) NLL ↓ (d) CE ↓

(e) RMSE ↓ (f) MAE ↓ (g) NLL ↓ (h) CE ↓

Figure 6: Performance of backbone and UQ methods according to the similarity between test data points and training scaffolds in the QM9 dataset. (a)–(d) illustrate the performance of backbone models averaged across all UQ methods, while (e)–(h) display the UQ performance with backbone averaged. The x-axis categorizes test data points into 5 bins based on their average Tanimoto similarity to the training scaffolds, starting from the most similar (leftmost bin). The average Tanimoto similarities for bins 1–5 are 0.116, 0.103, 0.093, 0.084, and 0.066, respectively. Lower values indicate better performance. The complete results are in Figure 10.

note that predictions for random splits exhibit a sharper distribution, *i.e.*, smaller $\hat{\sigma}$. This suggests that the models are more confident in regressing in-distribution data, aligning with our previous observation for Uni-Mol. Conversely, frozen backbones are less prone to overfitting due to their constrained expressiveness. This behavior underscores the original models' capability to distinguish between in-distribution and OOD features and assign confidence scores with precision.

Impact of Training-Test Distribution Shift Here we delve into the effects of data distribution on model performance within the QM9 dataset. We classify the test data into 5 bins according to their average Tanimoto similarities to the training scaffolds, from the most to the least similar. Please refer to appendix A for more details on the dataset preparation. As depicted in Figure 6, there is a clear trend of worsening model performance, as evidenced by a nearly linear increase in RMSE and MAE, when the test data features gradually diverge from the training scaffolds. Conversely, the calibration error remains comparatively stable across these bins, suggesting that the estimated uncertainties are more robust to shifts in distribution, which reinforces our earlier findings regarding the behavior of backbone and UQ methods.

6 Conclusion, Limitation, and Future Works

We present MUBen, a benchmark designed to evaluate a variety of UQ methods across different categories, using backbone models that employ various descriptors for multiple molecular property prediction tasks.

Our findings indicate that Deep Ensembles consistently enhance performance compared to the deterministic baseline, albeit at a substantial computational cost. For classification tasks, Temperature Scaling and MC Dropout are straightforward yet effective approaches. Conversely, for regression, BBP and SGLD appear more suitable in estimating uncertainty, although they may lead to a decrease in prediction accuracy, particularly with smaller backbone models. Among the different backbones, Uni-Mol stands out due to its effective utilization of 3D molecular conformations, which, while highly expressive, is also prone to overconfidence. Other backbone models, leveraging different descriptors, offer advantages under varying conditions and should be chosen based on the specific requirements of the use case.

Given the rapid advancements in molecular representation learning and UQ methods, MUBen cannot encompass all possible combinations, forcing us to focus on a curated selection of representative methods. Furthermore, we use coarse-grained hyperparameter grids to maintain experimental feasibility, which makes MUBen, while indicative of trends, might not present the best possible results. We remain committed to refining MUBen and welcome contributions from the broader community to enhance its inclusivity and utility for research in this field and related domains.

Acknowledgments

This work was supported in part by ONR MURI N00014-17-1-2656, NSF IIS-2008334, IIS-2106961, and CAREER IIS-2144338.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. *CoRR*, abs/2209.01712, 2022. doi: 10.48550/arXiv.2209.01712. URL https://doi.org/10.48550/arXiv.2209.01712.

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

Saeid Amiri, Mohammad Shokrolah Shirazi, and Shiqi Zhang. Learning and reasoning for robot sequential decision making under uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2726–2733, 2020.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=eNdiU_DbM9.

Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, and Amir Barati Farimani. Gpt-molberta: GPT molecular features language model for molecular property prediction. *CoRR*, abs/2310.03030, 2023. doi: 10.48550/ARXIV.2310.03030. URL https://doi.org/10.48550/arXiv.2310.03030.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *CoRR*, abs/1505.05424, 2015. URL http://arxiv.org/abs/1505.05424.

Glemm W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, January 1950. doi: 10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2. URL https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2.

Jonas Busk, Peter Bjørn Jørgensen, Arghya Bhowmik, Mikkel N Schmidt, Ole Winther, and Tejs Vegge. Calibrated uncertainty for molecular property prediction using ensembles of message passing neural networks. *Machine Learning: Science and Technology*, 3(1):015012, 2021.

Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10379–10388, 2021.

Austin H Cheng, Andy Cai, Santiago Miret, Gustavo Malkomes, Mariano Phielipp, and Alán Aspuru-Guzik. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2023.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020. URL https://arxiv.org/abs/2010.09885.

Sudatta Chowdhury, Forbes Gibb, and Monica Landoni. Uncertainty in information seeking and retrieval: A study in an academic environment. *Inf. Process. Manag.*, 47(2):157–175, 2011. doi: 10.1016/J.IPM.2010.09.006. URL https://doi.org/10.1016/j.ipm.2010.09.006.

Youngseog Chung, Ian Char, Han Guo, Jeff Schneider, and Willie Neiswanger. Uncertainty toolbox: an open-source library for assessing, visualizing, and improving uncertainty quantification. *CoRR*, abs/2109.10254, 2021. URL https://arxiv.org/abs/2109.10254.

Andreas C. Damianou and Neil D. Lawrence. Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, volume 31 of *JMLR Workshop and Conference Proceedings*, pp. 207–215. JMLR.org, 2013. URL http://proceedings.mlr.press/v31/damianou13a.html.

Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395, 2023.

Aryan Deshwal, Cory M Simon, and Janardhan Rao Doppa. Bayesian optimization of nanoporous materials. *Molecular Systems Design & Engineering*, 6(12):1066–1086, 2021.

Gianluca Detommaso, Alberto Gasparin, Michele Donini, Matthias W. Seeger, Andrew Gordon Wilson, and Cédric Archambeau. Fortuna: A library for uncertainty quantification in deep learning. *CoRR*, abs/2302.04019, 2023. doi: 10.48550/ARXIV.2302.04019. URL https://doi.org/10.48550/arXiv.2302.04019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Thomas G. Dietterich. Ensemble methods in machine learning. In Josef Kittler and Fabio Roli (eds.), *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1. URL https://doi.org/10.1007/3-540-45014-9_1.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. SE(3) equivariant graph neural networks with complete local frames. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5583–5608. PMLR, 2022. URL https://proceedings.mlr.press/v162/du22e.html.

Weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3d equivariant graph neural networks. *CoRR*, abs/2304.04757, 2023. doi: 10.48550/ARXIV.2304.04757. URL https://doi.org/10.48550/arXiv.2304.04757.

Alexandre Duval, Simon V. Mathis, Chaitanya K. Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D. Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael M. Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *CoRR*, abs/2312.07511, 2023. doi: 10.48550/ARXIV.2312.07511. URL https://doi.org/10.48550/arXiv.2312.07511.

Natalie S. Eyke, William H. Green, and Klavs F. Jensen. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *React. Chem. Eng.*, 5:1963–1972, 2020. doi: 10.1039/D0RE00232A. URL http://dx.doi.org/10.1039/D0RE00232A.

Kun Fang, Qinghua Tao, Xiaolin Huang, and Jie Yang. Revisiting deep ensemble for out-of-distribution detection: A loss landscape perspective. *CoRR*, abs/2310.14227, 2023. doi: 10.48550/ARXIV.2310.14227. URL https://doi.org/10.48550/arXiv.2310.14227.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.*, 4(2):127–134, 2022. doi: 10.1038/s42256-021-00438-4. URL https://doi.org/10.1038/s42256-021-00438-4.

Matthias Fey and Jan E. Lenssen. Fast graph representation learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *CoRR*, abs/1912.02757, 2019. URL http://arxiv.org/abs/1912.02757.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria-Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 1050–1059. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/gal16.html.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P. Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 8803–8812, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/be3087e74e9100d4bc4c6268cdbe8456-Abstract.html.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 6790–6802, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/35cf8659cfcb13224cbd47863a34fc58-Abstract.html.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342, 2021. URL https://arxiv.org/abs/2107.03342.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4878–4887, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Soumya Ghosh, Q. Vera Liao, Karthikeyan Natesan Ramamurthy, Jirí Navrátil, Prasanna Sattigeri, Kush R. Varshney, and Yunfeng Zhang. Uncertainty quantification 360: A holistic toolkit for quantifying and communicating the uncertainty of AI. *CoRR*, abs/2106.01410, 2021. URL https://arxiv.org/abs/2106.01410.

Raj Ghugare, Santiago Miret, Adriana Hugessen, Mariano Phielipp, and Glen Berseth. Searching for high-value molecules using reinforcement learning and transformers. *CoRR*, abs/2310.02902, 2023. doi: 10.48550/ARXIV.2310.02902. URL https://doi.org/10.48550/arXiv.2310.02902.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL http://proceedings.mlr.press/v70/gilmer17a.html.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.*, 12:7866–7881, 2021. doi: 10.1039/D0SC06805E. URL http://dx.doi.org/10.1039/D0SC06805E.

Kevin P. Greenman, Ava Soleimany, and Kevin K Yang. Benchmarking uncertainty quantification for protein engineering. In *ICLR2022 Machine Learning for Drug Discovery*, 2022. URL https://openreview.net/forum?id=G0vuqNwxaeA.

Cameron J Gruich, Varun Madhavan, Yixin Wang, and Bryan R Goldsmith. Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. *Machine Learning: Science and Technology*, 4(2):025019, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL http://proceedings.mlr.press/v70/guo17a.html.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *The Twelfth International Conference on Learning Representations*, 2023.

Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2888–2897, 2019.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016.

Brian Hie, Bryan D Bryson, and Bonnie Berger. Leveraging uncertainty in machine learning accelerates biological discovery and design. *Cell systems*, 11(5):461–477, 2020.

Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W. Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60 (8):3770–3780, 2020. doi: 10.1021/acs.jcim.0c00502. URL https://doi.org/10.1021/acs.jcim.0c00502. PMID: 32702986.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Weihua Hu, Matthias Fey, Hongyu Ren, Maho Nakata, Yuxiao Dong, and Jure Leskovec. OGB-LSC: A large-scale challenge for machine learning on graphs. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/db8e1af0cb3aca1ae2d0018624204529-Abstract-round2.html.

Doyeong Hwang, Grace Lee, Hanseok Jo, Seyoul Yoon, and Seongok Ryu. A benchmark study on reliable molecular supervised learning via bayesian learning. *CoRR*, abs/2006.07021, 2020. URL https://arxiv.org/abs/2006.07021.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pp. 876–885. AUAI Press, 2018. URL http://auai.org/uai2018/proceedings/papers/313.pdf.

Harshavardhan Kamarthi, Lingkai Kong, Alexander Rodríguez, Chao Zhang, and B. Aditya Prakash. When in doubt: Neural non-parametric uncertainty quantification for epidemic forecasting. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 19796–19807, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/a4a1108bbcc329a70efa93d7bf060914-Abstract.html.

Dongki Kim, Jinheon Baek, and Sung Ju Hwang. Graph self-supervised learning with accurate discrepancy learning. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/5b175f9e93873e3a10a6ce43dbb82e05-Abstract-Conference.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Diederik P. Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2575–2583, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/bc7316929fe1545bf0b98d114ee3ecb8-Abstract.html.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1eWbxStPH.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in- and out-of-distribution data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 1326–1340. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.102. URL https://doi.org/10.18653/v1/2020.emnlp-main.102.

Lingkai Kong, Jiaming Cui, Yuchen Zhuang, Rui Feng, B. Aditya Prakash, and Chao Zhang. End-to-end stochastic optimization with energy-based model. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/49cf35ff2298c10452db99d08036805b-Abstract-Conference.html.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1 (4):45024, 2020. doi: 10.1088/2632-2153/ABA947. URL https://doi.org/10.1088/2632-2153/aba947.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6402–6413, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1): 7568, 2023a.

Pengyong Li, Jun Wang, Yixuan Qiao, Hao Chen, Yihuan Yu, Xiaojun Yao, Peng Gao, Guotong Xie, and Sen Song. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Briefings in Bioinformatics*, 22(6):bbab109, 2021.

Yinghao Li, Le Song, and Chao Zhang. Sparse conditional hidden markov model for weakly supervised named entity recognition. In Aidong Zhang and Huzefa Rangwala (eds.), *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pp. 978–988. ACM, 2022. doi: 10.1145/3534678.3539247. URL https://doi.org/10.1145/3534678.3539247.

Ziyue Li, Kan Ren, Yifan Yang, Xinyang Jiang, Yuqing Yang, and Dongsheng Li. Towards inference efficient deep ensemble learning. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 8711–8719. AAAI Press, 2023b. doi: 10.1609/aaai.v37i7.26048. URL https://doi.org/10.1609/aaai.v37i7.26048.

Yi-Lun Liao and Tess E. Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=KwmPfARgOTD.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2999–3007. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.324. URL https://doi.org/10.1109/ICCV.2017.324.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=xQUe1pOKPam.

Yi Liu, Limei Wang, Meng Liu, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3d graph networks. *CoRR*, abs/2102.05013, 2021. URL https://arxiv.org/abs/2102.05013.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Wesley J. Maddox, Pavel Izmailov, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13132–13143, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/118921efba23fc329e6560b27861f0c2-Abstract.html.

Osman Mamun, Kirsten T Winther, Jacob R Boes, and Thomas Bligaard. A bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *npj Computational Materials*, 6(1):177, 2020.

Hendrik A. Mehrtens, Alexander Kurz, Tabea-Clara Bucher, and Titus J. Brinker. Benchmarking common uncertainty estimation methods with histopathological images under domain shift and label noise. *CoRR*, abs/2301.01054, 2023. doi: 10.48550/arXiv.2301.01054. URL https://doi.org/10.48550/arXiv.2301.01054.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018.

Subhabrata Mukherjee and Ahmed Hassan Awadallah. Uncertainty-aware self-training for few-shot text classification. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/f23d125da1e29e34c552f448610ff25f-Abstract.html.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html.

Allan H Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600, 1973.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael W. Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Z. Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim G. J. Rudner, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty baselines: Benchmarks for uncertainty & robustness in deep learning. *CoRR*, abs/2106.04015, 2021. URL https://arxiv.org/abs/2106.04015.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. *In AAAI Conference on Artificial Intelligence*, 29(1), February 2015. doi: 10.1609/aaai.v29i1.9602. URL https://doi.org/10.1609/aaai.v29i1.9602.

Maho Nakata and Tomomi Shimazaki. Pubchemqc project: A large-scale first-principles electronic structure database for data-driven chemistry. *J. Chem. Inf. Model.*, 57(6):1300–1308, 2017. doi: 10.1021/acs.jcim.7b00083. URL https://doi.org/10.1021/acs.jcim.7b00083.

Juhwan Noh, Geun Ho Gu, Sungwon Kim, and Yousung Jung. Uncertainty-quantified hybrid machine learning/density functional theory high throughput screening method for crystals. *Journal of Chemical Information and Modeling*, 60(4):1996–2003, 2020.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C. Lim, and Prudencio Tossou. Gotta be SAFE: A new framework for molecular design. *CoRR*, abs/2310.10773, 2023. doi: 10.48550/ARXIV.2310.10773. URL https://doi.org/10.48550/arXiv.2310.10773.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Joaquin Quiñonero-Candela, Carl Edward Rasmussen, Fabian Sinz, Olivier Bousquet, and Bernhard Schölkopf. Evaluating predictive uncertainty challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 1–27. Springer Berlin Heidelberg, 2006. doi: 10.1007/11736790_1. URL https://doi.org/10.1007/11736790_1.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/94aef38441efa3380a3bed3faf1f9d5d-Abstract.html.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mac. Intell.*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7. URL https://doi.org/10.1038/s42256-022-00580-7.

Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. A bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chem. Sci.*, 10:8438–8446, 2019. doi: 10.1039/C9SC01992H. URL http://dx.doi.org/10.1039/C9SC01992H.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9323–9332. PMLR, 2021. URL http://proceedings.mlr.press/v139/satorras21a.html.

Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2697–2717, 2020. doi: 10.1021/acs.jcim.9b00975. URL https://doi.org/10.1021/acs.jcim.9b00975. PMID: 32243154.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Trans. Neural Networks*, 20(1):61–80, 2009. doi: 10.1109/TNN.2008.2005605. URL https://doi.org/10.1109/TNN.2008.2005605.

Franko Schmähling, Jörg Martin, and Clemens Elster. A framework for benchmarking uncertainty in deep regression. *Appl. Intell.*, 53(8):9499–9512, 2023. doi: 10.1007/s10489-022-03908-3. URL https://doi.org/10.1007/s10489-022-03908-3.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 991–1001, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3183–3193, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Ava P. Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7(8):1356–1367, 2021. doi: 10.1021/acscentsci.1c00546. URL https://doi.org/10.1021/acscentsci.1c00546.

Hao Song, Tom Diethe, Meelis Kull, and Peter A. Flach. Distribution calibration for regression. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5897–5906. PMLR, 2019. URL http://proceedings.mlr.press/v97/song19a.html.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014. doi: 10.5555/2627435.2670313. URL https://dl.acm.org/doi/10.5555/2627435.2670313.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 3d infomax improves gnns for molecular property prediction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502. PMLR, 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Teague Sterling and John J. Irwin. Zinc 15 - ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559. URL https://doi.org/10.1021/acs.jcim.5b00559. PMID: 26479676.

Fan-Yun Sun, Jordan Hoffmann, Vikas Verma, and Jian Tang. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1lfF2NYvH.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *CoRR*, abs/2211.09085, 2022. doi: 10.48550/ARXIV.2211.09085. URL https://doi.org/10.48550/arXiv.2211.09085.

Philipp Thölke and Gianni De Fabritiis. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=zNHzqZ9wrRB.

Nathaniel Thomas, Tess E. Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *CoRR*, abs/1802.08219, 2018. URL http://arxiv.org/abs/1802.08219.

Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, 2022. doi: 10.1021/acs.jcim.2c01073. URL https://doi.org/10.1021/acs.jcim.2c01073. PMID: 36456532.

Daniel Varivoda, Rongzhi Dong, Sadman Sadeed Omee, and Jianjun Hu. Materials property prediction with uncertainty quantification: A benchmark study. *Applied Physics Reviews*, 10(2), 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

W. Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, 54(2):263–270, 2021. doi: 10.1021/acs.accounts.0c00699. URL https://doi.org/10.1021/acs.accounts.0c00699. PMID: 33370107.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In Xinghua Mindy Shi, Michael Buck, Jian Ma, and Pierangelo Veltri (eds.), *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019, Niagara Falls, NY, USA, September 7-10, 2019*, pp. 429–436. ACM, 2019. doi: 10.1145/3307339.3342186. URL https://doi.org/10.1145/3307339.3342186.

Yanli Wang, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang, and Stephen H. Bryant. Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, 37(Web-Server-Issue): 623–633, 2009. doi: 10.1093/nar/gkp456. URL https://doi.org/10.1093/nar/gkp456.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 681–688. Omnipress, 2011. URL https://icml.cc/2011/papers/398_icmlpaper.pdf.

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=Sklf1yrYDr.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771, 2019. URL http://arxiv.org/abs/1910.03771.

Tom Wollschläger, Nicholas Gao, Bertrand Charpentier, Mohamed Amine Ketata, and Stephan Günnemann. Uncertainty estimation for molecules: Desiderata and methods. *CoRR*, abs/2306.14916, 2023. doi: 10.48550/arXiv.2306.14916. URL https://doi.org/10.48550/arXiv.2306.14916.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. ISSN 2041-6539. doi: 10.1039/c7sc02664a. URL http://dx.doi.org/10.1039/c7sc02664a.

Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z. Li. A systematic survey of chemical pre-trained models. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 6787–6795. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/760. URL https://doi.org/10.24963/ijcai.2023/760. Survey Track.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Correction to analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(12):5304–5305, 2019. doi: 10.1021/acs.jcim.9b01076. URL https://doi.org/10.1021/acs.jcim.9b01076. PMID: 31814400.

Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, pp. 022022. IOP Publishing, 2019.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. Actune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 1422–1436. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.102. URL https://doi.org/10.18653/v1/2022.naacl-main.102.

Sheheryar Zaidi, Michael Schaarschmidt, James Martens, Hyunjik Kim, Yee Whye Teh, Alvaro Sanchez-Gonzalez, Peter W. Battaglia, Razvan Pascanu, and Jonathan Godwin. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=tYIMtogyee.

Yao Zhang and Alpha A. Lee. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.*, 10:8154–8163, 2019. doi: 10.1039/C9SC00616H. URL http://dx.doi.org/10.1039/C9SC00616H.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

Table 4: Dataset statistics in MUBen.

Property Category	Dataset	# Compounds	# Tasks	Average LIR[(a)]	Max LIR[(a)]
		Classification			
Physiology	BBBP	2,039	1	0.7651	0.7651
	ClinTox	1,478	2	0.9303	0.9364
	Tox21	7,831	12	0.9225	0.9712
	ToxCast	8,575	617	0.8336	0.9972
	SIDER	1,427	27	0.7485	0.9846
Biophysics	BACE	1,513	1	0.5433	0.5433
	HIV	41,127	1	0.9649	0.9649
	MUV	93,087	17	0.9980	0.9984
		Regression			
Physical Chemistry	ESOL	1,128	1	-	-
	FreeSolv	642	1	-	-
	Lipophilicity	4,200	1	-	-
Quantum Mechanics	QM7	7,160	1	-	-
	QM8	21,786	12	-	-
	QM9	133,885	3[(b)]	-	-

[(a)] LIR stands for "label imbalance ratio", which is calculated as $\text{LIR}_k \in [0.5, 1] = \max\{p_{\text{pos}}, p_{\text{neg}}\}$ for a task indexed by $k \in \{1, \ldots, K\}$, where p_{pos} and p_{neg} are the proportions of positive and negative samples in the dataset, respectively. LIR = 0.5 indicates a balanced dataset and LIR = 1 indicates a completely imbalanced one, with either $y = 0$ or $y = 1$ being absent. Average LIR $= \sum_{k=1}^{K} \text{LIR}_k$, and Max LIR $= \max_k\{\text{LIR}_k\}$. LIR is calculated on the entire dataset, and the number could slightly vary for each partition.
[(b)] QM9 dataset contains 12 tasks, but we follow Fang et al. (2022); Zhou et al. (2023) and use only 3 most popular ones (homo, lumo, and gap).

A Dataset Details

In this section, we provide details about the datasets used in our study. We follow the approach of previous works (Fang et al., 2022; Zhou et al., 2023) and select a subset of widely used and publicly available datasets from the MoleculeNet benchmark (Wu et al., 2018). The datasets cover both classification and regression tasks from 4 property categories, including Physiology, Biophysics, Physical Chemistry and Quantum Mechanics. The classification datasets include:

- **BACE** provides binary binding properties for a set of inhibitors of human β-secretase 1 (BACE-1) from experimental values from the published papers;
- **BBBP (Blood-Brain Barrier Penetration)** studies the classification of molecules by their permeability of the blood-brain barrier;
- **ClinTox** consists of two classification tasks for drugs: whether they are absent of clinical toxicity and whether they are approved by the FDA;
- **Tox21 (Toxicology in the 21st Century)** consists of qualitative toxicity measurements of 8,014 compounds on 12 different targets;
- **ToxCast** provides toxicology data of 8,576 compounds on 617 different targets;
- **SIDER (Side Effect Resource)** contains marketed drugs and adverse drug reactions (ADR) extracted from package inserts;
- **HIV**, introduced by the Drug Therapeutics Program (DTP) AIDS Antiviral Screen, contains compounds that are either active or inactive against HIV replication;
- **MUV (Maximum Unbiased Validation)**, a challenging benchmark dataset selected from PubChem BioAssay for validation of virtual screening techniques, contains 93,087 compounds tested for activity against 17 different targets.

The regression datasets include:

- **ESOL** provides experimental values for water solubility data for 1,128 compounds;
- **FreeSolv (Free Solvation Database)** contains experimental and calculated hydration free energies for 643 small molecules;
- **Lipophilicity** contains experimental results of octanol/water distribution coefficient for 4,200 compounds;
- **QM7/8/9** are subsets of GDB-13 and GDB-17 databases containing quantum mechanical calculations of energies, enthalpies, and dipole moments for organic molecules with up to 7/8/9 "heavy" atoms, respectively. For QM9, we follow previous works and select 3 tasks that predict properties (homo, lumo, and gap) with a similar quantitative range out of the total 12 (Fang et al., 2022; Zhou et al., 2023).

A summary of the dataset statistics is provided in Table 4. It is worth noting that some datasets, such as HIV and MUV, exhibit a high degree of class imbalance. This characteristic adds further challenges to the tasks of molecular property prediction and uncertainty quantification.

Scaffold Splitting MUBen primarily utilizes a dataset split based on molecule scaffolds, effectively segregating training, validation, and test sets to maximize feature separation. This approach generates OOD test sets, crucial for assessing the model's ability to handle patterns not encountered during the fine-tuning process. We adhere to the standard 8:1:1 ratio for training, validation, and test splits across all datasets. The raw datasets can be accessed on the MoleculeNet website.[1] Additionally, we employ the pre-processed versions provided by Zhou et al. (2023), using the identical dataset splits outlined in their study.[2]

Random Splitting To investigate the impact of dataset splitting methods, we conducted experiments on 4 classification datasets (BACE, BBBP, ClinTox, Tox21) and 4 regression datasets (ESOL, FreeSolv, Lipophilicity, QM7) using random splitting. Despite their relatively small sizes, these datasets provide a reliable representation of model performance. Each dataset was divided into three separate training-validation-test sets with an 8:1:1 ratio, utilizing random seeds of 0, 1, and 2. We trained each backbone-UQ combination once per split and averaged the results across the three runs to establish the final metrics for each dataset. Deep Ensembles was the only exception; it underwent three training cycles per dataset split, using different seeds, resulting in a total of 9 training cycles per dataset. All other training configurations remained consistent with those used in the scaffold-split experiments.

Binning Test Data by Similarity to Training Scaffolds Our experimental analysis is conducted exclusively on the QM9 dataset due to its ample volume of test data, ensuring a statistically significant number of samples within each defined bin for reliable outcomes. Initially, the dataset is partitioned into training, validation, and test sets using scaffold splits. We then compute the average Tanimoto similarity between each test data point and the unique training scaffolds (1,096 in total). Test data points are sorted and grouped into 5 bins based on descending similarity scores, with the most similar points placed in the first bin. The average Tanimoto similarities for these 5 bins are 0.116, 0.103, 0.093, 0.084, and 0.066, respectively. During the experiments, each combination of backbone and UQ model is trained once on the training dataset and then tested separately across each of the bins. This process is repeated three times using random seeds 0, 1, and 2, with the results averaged across these runs. For the Deep Ensembles method, these three runs are integrated into a single ensemble, considered equivalent to a single run.

B Backbone Models and Implementation Details

Our experimental code is designed on the PyTorch framework (Paszke et al., 2019). The experiments are conducted on a single NVIDIA A100 Tensor Core GPU with a memory capacity of 80GB. The backbone models are fine-tuned using the AdamW optimizer (Loshchilov & Hutter, 2019) with a weight decay rate of 0.01 using full-precision floating-point numbers (FP32) for maximum compatibility. We apply different learning rates, numbers of training epochs and batch sizes for the backbone models, as specified in the following paragraphs. We adopt early stopping to select the best-performed checkpoint on the validation set, and all models have achieved their peak validation performance before the training ends. ROC-AUC is selected to assess classification validation performance. For regression, we follow Wu et al. (2018) and use

[1] https://moleculenet.org/
[2] https://github.com/dptech-corp/Uni-Mol/tree/main/unimol

Table 5: Model statistics.

Model	# Parameters (M)	Average Time per Training Step (ms)[(a)]
DNN	0.158	5.39
ChemBERTa	3.43	30.18
GROVER	48.71	334.47
Uni-Mol	47.59	392.55
TorchMD-NET	7.23	217.29
GIN	0.26	7.21

[(a)] All models are evaluated on the BBBP dataset with a batch size of 128. We only measure the time for forward passing, backward passing, and parameter updating. We train the model for 6 epochs and take the average of the last 5 to reduce the impact of GPU initialization.

RMSE for Physical Chemistry properties and MAE for Quantum Mechanics. Notice that these metrics only concern prediction, and do not take the uncertainty into account during validation steps (appendix D.1). Table 5 presents the number of parameters and average time per training step for each backbone model. The training time is calculated for each model update step instead of each epoch based on our implementation, which might be slower than the models' original realizations. Below, we offer a detailed description of each backbone model's architecture and its implementation.

ChemBERTa ChemBERTa (Chithrananda et al., 2020; Ahmad et al., 2022) leverages the RoBERTa model architecture and pre-training strategy (Liu et al., 2019) but with fewer layers, attention heads, and smaller hidden dimensionalities. Unlike language models which process sentences in natural language, ChemBERTa uses Simplified Molecular-Input Line-Entry System (SMILES, Weininger, 1988) strings as input. This representation is a compact and linear textual depiction of a molecule's structure that's frequently employed in cheminformatics. ChemBERTa pre-training adopts a corpus of 77M SMILES strings from PubChem (Wang et al., 2009), along with the masked language modeling (MLM) objective (Devlin et al., 2019).

ChemBERTa is built with the HuggingFace Transformers library (Wolf et al., 2019), and its pre-trained parameters are shared through the Huggingface's model hub. We retain the default model architecture and use the `DeepChem/ChemBERTa-77M-MLM` checkpoint for ChemBERTa's weight initialization.[3] We employ the last-layer hidden state corresponding to the token `[CLS]` to represent the input SMILES sequence and attach the output heads specified by the tasks/UQ methods on top of it for property prediction. On all datasets, ChemBERTa is fine-tuned with a learning rate of 10^{-5}, a batch size of 128, and for 200 epochs. A tolerance of 40 epochs for early stopping is adopted.

GROVER GROVER (Rong et al., 2020) is pre-trained on 11 million unlabeled molecules represented as 2D molecular graphs, sourced from ZINC15 (Sterling & Irwin, 2015) and ChEMBL (Gaulton et al., 2012). This model enhances pre-training by integrating Message Passing Neural Networks (MPNNs, Gilmer et al., 2017), a GNN-based method, into Transformer encoder architectures. GROVER employs self-supervised learning objectives that reflect different levels of molecular structural complexity to capture detailed structural and semantic information. Specifically, it replaces the linear layers that map the input query, key, and value vectors in each multi-head attention module of the Transformer encoder with a dynamic MPNN, which aggregates latent features from k-hop neighboring nodes, where k is a random integer chosen from a predefined uniform or Gaussian distribution. In essence, GROVER integrates GNN layers before each head's self-attention process to more effectively encode the molecular graph structure. To aid learning, GROVER introduces training objectives such as contextual property prediction, which estimates the statistical properties of masked subgraphs of varying sizes, and graph-level motif prediction, which identifies the classes of masked functional groups. A readout function is used to aggregate node features, which are subsequently processed by linear layers for property prediction.

[3] https://huggingface.co/DeepChem/ChemBERTa-77M-MLM.

For implementation, we use the GROVER-base checkpoint for our model initialization.[4] We incorporate the "node-view" branch as discussed above and disregard the "edge-view" architecture detailed in the appendix of Rong et al. (2020), in accordance with the default settings in their GitHub repository. Under this configuration, the model generates 2 sets of node embeddings (but no edge embeddings), one from the node hidden states and another from the edge hidden states (Rong et al., 2020). Each set of embeddings is passed into 2 linear layers with GELU (Hendrycks & Gimpel, 2016) activation and a dropout ratio of 0.1 post-readout layer, which simply averages the embeddings in the default configuration. Each set of embeddings corresponds to an individual output branch, predicting the properties independently. In line with the original implementation, we compute the loss for each branch individually during fine-tuning and apply a squared Euclidean distance regularization with a coefficient of 0.1 between the two. During inference, we average the logits from the 2 branches to generate the final output logits. Our implementation of GROVER does not incorporate external RDKit or Morgan fingerprints, diverging from the authors' original implementation.

In our experiments, we configure the fine-tuning batch size at 256, and the number of epochs at 100 with a tolerance of 40 epochs for early stopping. The learning rate is set at 10^{-4}, and the entire model has a dropout ratio of 0.1. We substitute the original Noam learning rate scheduler (Vaswani et al., 2017) with a linear learning rate scheduler with a 0.1 warm-up ratio for easier implementation. No substantial differences in model performance were observed between the two learning rate schedulers.

Uni-Mol Uni-Mol (Zhou et al., 2023) is a universal molecular representation framework that enhances representational capacity and broadens applicability by incorporating 3D molecular structures as model input. For the property prediction task, Uni-Mol undergoes pre-training on 209M 3D conformations of organic molecules gathered from ZINC15 (Sterling & Irwin, 2015), ChEMBL (Gaulton et al., 2012), and a database comprising 12M purchasable molecules (Zhou et al., 2023). It portrays atoms as tokens and utilizes pair-type aware Gaussian kernels to encode the positional information in the 3D space, thereby ensuring rotational and translational invariance. Furthermore, Uni-Mol introduces a pair-level representation by orchestrating an "atom-to-pair" communication—updating positional encodings with query-key products—and a "pair-to-atom" communication—adding pair representation as bias terms in the self-attention atom update procedure. For pre-training, Uni-Mol employs masked atom prediction, akin to BERT's MLM, corrupting 3D positional encodings with random noise at a 15% ratio. Additionally, the model is tasked with restoring the corrupted Euclidean distances between atoms and the coordinates for atoms.

Our codebase is developed atop the publicly accessible Uni-Mol repository,[2] and their pre-trained checkpoint for molecular prediction serves as our model initialization. During fine-tuning, Uni-Mol generates 10 sets of 3D conformations for each molecule, supplemented with an additional 2D molecular graph. Thereafter Uni-Mol samples one from these 11 molecular representations for each molecule at the beginning of every training epoch as the input feature. For inference, we average the logits from all 11 representations to generate the final output. We utilize the conformations prepared by the Uni-Mol repository in our implementation.

We configure the fine-tuning batch size at 128, the number of epochs at 100, and employ early stopping with a tolerance of 40 epochs. We use a linear learning rate scheduler with a 0.1 warm-up ratio and a peak learning rate of 5×10^{-5}. The model is trained with a dropout ratio of 0.1. Although the Uni-Mol repository provides a set of recommended hyperparameters, we observe no discernible improvement in model performance with these settings.

DNN The Deep Neural Network (DNN) serves as a simple, randomly initialized baseline model designed to explore how heuristic descriptors like Morgan fingerprints (Morgan, 1965) or RDKit features perform for molecular property prediction. DNN enables us to compare the pre-trained models, which learn the molecular representation automatically through self-learning, with heuristic molecular features, which are constructed manually, and investigate whether or under what circumstances the heuristic features can achieve comparable results. For the descriptor, we adopt the approach of previous work (Wu et al., 2018; Yang et al., 2019; Rong et al., 2020) and extract 200-dimensional molecule-level features using RDKit for each molecule, which are then used as DNN input.[5] The DNN consists of 8 fully connected 128-dimensional hidden layers with GELU

[4] https://github.com/tencent-ailab/grover

[5] We use the `rdkit2dnormalized` descriptor in `DescriptaStorus`, available at https://github.com/bp-kelley/descriptastorus.

Table 6: Impact of the number and dimensionality of hidden layers across various datasets. The selected hyper-parameters generally achieve reasonable performance, although they may not be optimal for all datasets.

	Number of Layers (Dimension: 128)				Layer Dimensionality (Number of Layers: 5)			
Dataset: Metric	3 Layers	4 Layers	5 Layers	6 Layers	64 Dim	128 Dim	256 Dim	512 Dim
BBBP: ROC-AUC ↑	0.5736	0.6233	0.6284	0.6223	0.5876	0.6284	0.6255	0.5977
Tox21: ROC-AUC ↑	0.6644	0.6774	0.6832	0.6901	0.6584	0.6832	0.6924	0.6962
ESOL: RMSE ↓	1.6741	1.664	1.6017	1.6383	1.9951	1.6017	1.6167	1.5868
FreeSolv: RMSE ↓	2.3296	2.2348	2.1362	2.5788	4.0666	2.1362	2.8263	3.1005

activation and an intervening dropout ratio of 0.1. We find no performance gain from deeper or wider DNNs and thus assume that our model is fully capable of harnessing the expressivity of RDKit features. The model is trained with a batch size of 256 and a constant learning rate of 2×10^{-4} for 400 epochs with an early stopping tolerance of 50 epochs.

TorchMD-NET The architectural design of TorchMD-NET is detailed in (Thölke & Fabritiis, 2022), while its pre-training methodology is discussed in a separate study (Zaidi et al., 2023). TorchMD-NET is an equivariant Transformer tailored for the prediction of quantum mechanical properties of 3D conformers. Unique elements of its architecture include a specialized embedding layer—encoding not just atomic numbers but also the interatomic distances influenced by neighboring atoms, a modified multi-head attention mechanism that integrates edge data, and an equivariant update layer computing atomic interactions. The model undergoes pre-training on the 3.4M PCQM4Mv2 dataset (Nakata & Shimazaki, 2017; Hu et al., 2021), leveraging a denoising auto-encoder objective. This entails predicting Gaussian noise disturbances on atomic positions, mirroring techniques seen in prevalent diffusion models in computer vision (Song et al., 2021).

To implement TorchMD-NET, we sourced the code and model checkpoint from (Zaidi et al., 2023).[6] We made minor architectural adjustments, replacing their single-head output block with our adaptive multi-head output layers. Consequently, we omitted the denoising objective during the fine-tuning process due to compatibility concerns. Our fine-tuning regimen for TorchMD-NET entails a batch size of 128 over 100 epochs, adopting an early stopping mechanism with the patience of 40 epochs. The learning rate peaks at 2×10^{-4}, coupled with a linear scheduler with a 0.1 warm-up ratio, and the model trains with a dropout ratio of 0.1.

GIN Graph Isomorphism Network GIN (Xu et al., 2019) is a randomly initialized model with 2D graph structures as input. Compared with graph convolutional networks (GCN, Kipf & Welling, 2017), GIN mainly differs in that within the neighboring nodes message aggregation process, GIN adds a weight to each node's self-looping, which is either trainable or pre-defined. In addition, GIN substitutes the one-layer feed-forward network within each GCN layer with a multi-layer perceptron (MLP). It has been proved in theory that these minor changes make GIN among the most powerful graph neural networks (Xu et al., 2019).

We use the Pytorch Geometric (Fey & Lenssen, 2019) to realize GIN. Our implementation contains 5 GIN layers with 128 hidden units and 0.1 dropout ratio. A study of the hyperparameters of GIN is presented in Table 6. The model is trained with a batch size of 128 for 200 epochs with an early stopping tolerance of 50 epochs, at a constant learning rate of 10^{-4}.

C Uncertainty Quantification

C.1 Method and Implementation Details

Focal Loss First proposed by Lin et al. (2017), Focal Loss is designed to address the class imbalance issue for dense object detection in computer vision, where the number of negative samples (background) far exceeds the number of positive ones (objects). It is adopted for uncertainty estimation and model calibration later by Mukhoti et al. (2020). The idea is to add a modulating factor to the standard cross-entropy loss to down-weight the contribution from easy examples and thus focus more on hard examples. In the binary

[6] https://github.com/shehzaidi/pre-training-via-denoising

classification case, it adds a modulating factor $|y_n - \hat{p}_n|^\gamma$ to the standard cross-entropy loss, where $y_n \in \{0, 1\}$ is the ground truth label, $\hat{p}_n \in (0, 1)$ is the predicted Sigmoid probability for the n-th example, and $\gamma \geq 0$ is a focusing parameter:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N} \sum_{n=1}^{N} \left[y_n (1 - \hat{p}_n)^\gamma \log \hat{p}_n + (1 - y_n) \hat{p}_n^\gamma \log(1 - \hat{p}_n) \right]. \tag{1}$$

The focusing parameter γ smoothly adjusts the rate at which easy examples are down-weighted. When $\gamma = 0$, Focal Loss is equivalent to the cross-entropy loss. As γ increases, the effect of the modulating factor increases likewise. In our implementation, we take advantage of the realization in the `torchvision` library and use their default hyperparameters for all experiments.[7]

Bayes by Backprop Bayes by Backprop (BBP, Blundell et al., 2015; Kingma et al., 2015) is an algorithm for training BNNs, where weights are not point estimates but distributions. The idea is to replace the deterministic network weights with Gaussian a porteriori learned from the data, which allows quantifying the uncertainty in the predictions by assembling the predictions from random networks sampled from the posterior distribution of the weights:

$$\boldsymbol{W}^{\text{MAP}} = \arg\max_{\boldsymbol{W}} \log p(\boldsymbol{W}|\mathbb{D}) = \arg\max_{\boldsymbol{W}} \left(\log p(\mathbb{D}|\boldsymbol{W}) + \log p(\boldsymbol{W}) \right), \tag{2}$$

where $p(\boldsymbol{W})$ is the prior distribution of the weights, which are also Gaussian in our realization.

However, the true posterior is generally intractable for neural networks and can only be approximated with variational inference $q(\boldsymbol{W}|\boldsymbol{\theta})$ (Kingma & Welling, 2014), where $\boldsymbol{\theta}$ are variational parameters, which consist of the multivariate Gaussian mean and variance in our case. The learning then involves finding the $\boldsymbol{\theta}$ that minimizes the Kullback-Leibler (KL) divergence between the true posterior and the variational distribution. The loss can be written as

$$\mathcal{L} = D_{\text{KL}} \left(q(\boldsymbol{W}|\boldsymbol{\theta}) || p(\boldsymbol{W}) \right) - \mathbb{E}_{q(\boldsymbol{W}|\boldsymbol{\theta})} [\log p(\mathbb{D}|\boldsymbol{W})]. \tag{3}$$

For each training step i, we first draw a sample from a porteriori $\boldsymbol{W}_i \sim q(\boldsymbol{W}|\boldsymbol{\theta})$ and then compute the Monte Carlo estimation of the loss:

$$\mathcal{L}_i \approx \log q(\boldsymbol{W}_i|\boldsymbol{\theta}) - \log p(\boldsymbol{W}_i) - \log p(\mathbb{D}|\boldsymbol{W}_i). \tag{4}$$

During backpropagation, the gradient can be pushed through the sampling process with the reparameterization trick (Kingma & Welling, 2014). Specifically, we adopt the local reparameterization trick (Kingma et al., 2015), which samples the pre-activation $\boldsymbol{a}_i$ directly from the distribution $q(\boldsymbol{a}_i|\boldsymbol{x}_i)$ parameterized by the input feature $\boldsymbol{x}_i$, instead of computing the is as $\boldsymbol{a}_i = \boldsymbol{W}_i \boldsymbol{x}_i$ using the sampled network weights. This has parameters that are deterministic functions of $\boldsymbol{x}_i$ and $\boldsymbol{\theta}$, which reduces the variance of the gradient estimates and can improve the efficiency of the learning process. Following the common practice (Harrison et al., 2023), we only apply BBP to the models' output layer to reduce computational costs and optimization difficulty for large pre-trained backbone models. In addition, The last layer often directly relates to the task's uncertainty. Modeling uncertainty in this layer can provide practical benefits in decision-making processes, allowing for an estimation of confidence in the predictions. During inference, we sample 30 networks from the posterior distribution, generating 30 sets of prediction results, and computing the mean of the predictions as the final output.

SGLD Stochastic gradient Langevin dynamics (SGLD, Welling & Teh, 2011) combines the efficiency of stochastic gradient descent (SGD) with Langevin diffusion which introduces the ability to estimate parameter a posteriori. The update rule for SGLD is given by:

$$\Delta\boldsymbol{\theta} = -\frac{\eta_t}{2} \nabla \mathcal{L}(\boldsymbol{\theta}) + \sqrt{\eta_t} \boldsymbol{\epsilon}, \tag{5}$$

[7] https://pytorch.org/vision/main/generated/torchvision.ops.sigmoid_focal_loss.html.

where $\boldsymbol{\theta}$ is the network parameters, η_t is the learning rate at time t, and $\boldsymbol{\epsilon}_t$ is the standard Gaussian noise. With learning rate $\boldsymbol{\theta}$ or weight gradient $\nabla\mathcal{L}(\boldsymbol{\theta})$ decreasing to small values, the update rule can transit from network optimization to posterior estimation. After sufficient optimization, subsequent samples of parameters can be seen as drawing from the posterior distribution of the model parameters given the data.

In our implementation, we follow the previous implementation and use a constant learning rate.[8] Similar to BBP, we only apply SGLD to the last layer of the model. We first train the model until its performance has stopped improving on the validation set, and then continue training it for another 30 epochs, resulting in 30 networks sampled from the Langevin dynamics. This generates 30 sets of prediction results during the test, and we compute the mean of the predictions as the final output.

MC Dropout Compared to other Bayesian networks, Monte-Carlo Dropout (MC Dropout, Gal & Ghahramani, 2016) is a simple and efficient for modeling the network uncertainty. Dropout is proposed to prevent overfitting by randomly setting some neurons' outputs to zero during training (Srivastava et al., 2014). At test time, dropout is deactivated and the weights are scaled down by the dropout rate to simulate the presence of all neurons. In contrast, MC Dropout proposes to keep dropout active during testing and make predictions with dropout turned on. By running several (*e.g.*, 30 in our experiments) forward passes with random dropout masks, we effectively obtain a Monte Carlo estimation of the predictive distribution.

SWAG Stochastic Weight Averaging-Gaussian (SWAG, Maddox et al., 2019) is an extension of the Stochastic Weight Averaging (SWA, Izmailov et al., 2018) method, a technique used for finding wider optima in the loss landscape and leads to improved generalization. SWAG fits a Gaussian distribution with a low rank plus diagonal covariance derived from the SGD iterates to approximate the posterior distribution over neural network weights. In SWAG, the model keeps tracking the weights encountered during the last T steps of the stochastic gradient descent updates and computes the Gaussian mean and covariance:

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\theta}} &= \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}_t; \\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}} &= \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{\boldsymbol{\theta}})(\boldsymbol{\theta}_t - \boldsymbol{\mu}_{\boldsymbol{\theta}})^{\mathsf{T}}. \end{aligned} \tag{6}$$

However, such computation requires storing all the model weights in the last T steps, which is expensive for large models. To address this issue, Maddox et al. (2019) propose to approximate the covariance matrix with a low-rank plus diagonal matrix, and compute the mean and covariance iteratively. Specifically, at update step $t \in \{1, \ldots, T\}$,

$$\bar{\boldsymbol{\theta}}_t = \frac{t\bar{\boldsymbol{\theta}}_{t-1} + \boldsymbol{\theta}_t}{t+1}; \quad \overline{\boldsymbol{\theta}^2}_t = \frac{t\overline{\boldsymbol{\theta}^2}_{t-1} + \boldsymbol{\theta}_t^2}{t+1}; \quad \widehat{D}_{:,t} = \boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}}_t, \tag{7}$$

where $\bar{\boldsymbol{\theta}}_0$ is the best parameter weights found during training prior to the SWA session, $\overline{\boldsymbol{\theta}^2}_T - \bar{\boldsymbol{\theta}}_T^2$ is the covariance diagonal, and $\frac{1}{T-1}\widehat{D}\widehat{D}^{\mathsf{T}} \in \mathbb{R}^{d,d}$ is the low-rank approximation of the covariance matrix with d being the parameter dimensionality. We can write the Gaussian weight posterior as

$$\boldsymbol{\theta}_{\text{SWAG}} \sim \mathcal{N}_{\text{SWAG}}\left(\bar{\boldsymbol{\theta}}_T, \frac{1}{2}\left(\overline{\boldsymbol{\theta}^2}_T - \bar{\boldsymbol{\theta}}_T^2 + \frac{1}{T-1}\widehat{D}\widehat{D}^{\mathsf{T}}\right)\right). \tag{8}$$

Notice that $\hat{D}$ has a different rank $K <= T$ in (Maddox et al., 2019), but we set $K = T < d$ for simplicity. Uncertainty in SWAG is estimated by drawing weight samples from $\mathcal{N}_{\text{SWAG}}$ and running these through the network. We set $T = 20$, and draw 30 samples during the test in our experiments.

Temperature Scaling Temperature Scaling (Platt et al., 1999; Guo et al., 2017) is a simple and effective post-hoc method for calibrating the confidence of a neural network. Post-hoc methods calibrate the output probabilities of a pre-trained model without updating the fine-tuned network parameters. The core idea behind Temperature Scaling is to add a learnable parameter h (the temperature) to adjust the output

[8] https://github.com/JavierAntoran/Bayesian-Neural-Networks/.

Table 7: Computation resources required by different uncertainty estimation methods. We assume that we have already trained a deterministic backbone model for property prediction, and would like to build up a UQ method on top of it.

UQ Method	Training Starting Checkpoint	Additional Cost[(a)]
Deterministic	-	0
Temperature	from fine-tuned backbone	$(T_{\text{infer}} + T_{\text{train-FFN}}) \times M_{\text{train-extra}}$
Focal Loss	from scratch	$T_{\text{train}} \times M_{\text{train}}$
MC Dropout	no training	$T_{\text{infer}} \times M_{\text{infer}}$
SWAG	from fine-tuned backbone	$T_{\text{train}} \times M_{\text{train-extra}} + T_{\text{infer}} \times M_{\text{infer}}$
BBP	from scratch	$T_{\text{train}} \times M_{\text{train}} + T_{\text{infer}} \times M_{\text{infer}}$
SGLD	from scratch	$T_{\text{train}} \times (M_{\text{train}} + M_{\text{train-extra}}) + T_{\text{infer}} \times M_{\text{infer}}$
Ensembles	from scratch	$T_{\text{train}} \times M_{\text{train}} \times (N_{\text{ensembles}} - 1)$

[(a)] T_{train} and T_{infer} are the time for one epoch of training and inference of the backbone model, respectively. In general, $T_{\text{train}} \gg T_{\text{infer}}$. M_{train}, $M_{\text{train-extra}}$, and M_{infer} are the number of training epochs, additional training epochs, and inference epochs, respectively (appendix C.1). In general, $M_{\text{train}} \gg M_{\text{train-extra}}$. Different backbones and UQ methods have different Ts and Ms, but we use the same symbols nonetheless for simplicity. The result is a rough estimation without considering the additional inference time or the early stopping if a model is retrained.

probability of the model. For a trained binary classification model, Temperature Scaling scales the logits z with

$$z' = \frac{z}{h} \tag{9}$$

before feeding z' into the Sigmoid output activation function. The temperature h is learned by minimizing the negative log-likelihood of the training data with other network parameters frozen. For multi-task classification such as Tox21, we assign an individual temperature to each task.

In precise terms, "Temperature Scaling" is introduced for multi-class classification utilizing SoftMax output activation (Guo et al., 2017). For binary classification in our study, we implement Platt Scaling, excluding the bias term (Platt et al., 1999). Nonetheless, we continue using "Temperature Scaling" for its widespread recognition.

Deep Ensembles Deep Ensembles (Lakshminarayanan et al., 2017) is a technique where multiple deep learning models are independently trained from different initializations, and their predictions are combined to make a final prediction. This approach exploits the idea that different models will make different types of errors, which can be reduced by averaging model predictions, leading to better overall performance and more robust uncertainty estimates (Fort et al., 2019). Formally, given M models in the ensemble, each with parameters $\boldsymbol{\theta}_m, m \in \{1, \ldots, M\}$, the ensemble prediction for an input data point $\boldsymbol{x}$ is given by:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m = \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{x}; \boldsymbol{\theta}_m) \tag{10}$$

where f represents the model architecture, and $\hat{y}_m$ is the post-activation result of the m-th model. We set $M = 3$ for QM8, QM9 and MUV to reduce computational consumption, and $M = 10$ for other scaffold split datasets. For random split, we uniformly use $M = 3$, as mentioned in appendix A.

For regression tasks, Lakshminarayanan et al. (2017) aggregate the variances of different network predictions through parameterizing a Gaussian mixture model. In contrast, we take a simpler approach by computing the mean of the variances as the final output variance.

C.2 Resource Analysis

Table 7 summarizes the additional training cost to apply each UQ method to a backbone model already fine-tuned for property prediction. From the table, we can see that post-hoc calibration and MC Dropout are

the most efficient methods, while Deep Ensembles is undoubtedly the most expensive one, even though it performs the best most of the time. Several works aim to reduce the computational cost (Wen et al., 2020; Li et al., 2023b), but we do not consider them in MUBen and leave them to future works.

D Metrics

In this section, we elaborate on the metrics utilized in our research for both property prediction and uncertainty estimation.

D.1 Property Prediction Metrics

ROC-AUC The receiver operating characteristic area under the curve (ROC-AUC) is widely used for binary classification tasks. The ROC curve plots the true positive rate (TPR), or *recall*, against the false positive rate (FPR) at various decision thresholds $t \in (0, 1)$. Given a set of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the TPR and FPR are computed as $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ and $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ respectively. The AUC signifies the likelihood of a randomly selected positive instance being ranked above a randomly chosen negative instance. This integral under the ROC curve is calculated as

$$\text{ROC-AUC} = \int_0^1 \text{TPR}(t) \frac{d}{dt} \text{FPR}(t) dt \tag{11}$$

and can be approximated using numerical methods.[9]

RMSE For regression tasks, the root mean square error (RMSE) quantifies the average discrepancy between predicted values $\hat{y}_n \in \mathbb{R}$ and actual values $y_n \in \mathbb{R}$ for N data points, given by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2}. \tag{12}$$

MAE The mean absolute error (MAE) is another regression metric, measuring the average absolute deviation between predicted and actual values. It is calculated as:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_n - y_n|. \tag{13}$$

D.2 Uncertainty Quantification Metrics

NLL Negative Log-Likelihood (NLL) quantifies the mean deviation between predicted and actual values in logarithmic space and is commonly used to evaluate UQ performance (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017; Gawlikowski et al., 2021).

For binary classification with Sigmoid output activation, NLL is given by:

$$\text{NLL} = -\frac{1}{N} \sum_{n=1}^{N} \left[y_n \log \hat{p}_n + (1 - y_n) \log(1 - \hat{p}_n) \right], \tag{14}$$

where $y_n \in \{0, 1\}$ is the true label and $\hat{p}_n \in (0, 1) = p_{\boldsymbol{\theta}}(\hat{y}_n \mid \boldsymbol{x}_n)$ is the predicted probability.

For regression, the Gaussian NLL is calculated as:

$$\text{NLL} = -\frac{1}{N} \sum_{n=1}^{N} \log \mathcal{N}(y_n; \hat{y}_n, \hat{\sigma}_n) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2} \left[\log(2\pi\hat{\sigma}_n) + \frac{(y_n - \hat{y}_n)^2}{\hat{\sigma}_n} \right], \tag{15}$$

where $\hat{y}_n \in \mathbb{R}$ is the predicted mean and $\hat{\sigma}_n \in \mathbb{R}_+$ is the predicted variance regularized by the SoftPlus activation.

[9] We use scikit-learn's roc_auc_score function in our implementation.

Brier Score In classification, the Brier Score (BS) is a proper scoring rule for measuring the accuracy of predicted probabilities. Similar to mean square error (MSE) in regression, it measures the mean squared difference between predicted probabilities and actual binary outcomes. For binary classification, the Brier Score is calculated as

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_n - y_n)^2, \tag{16}$$

where $y_n \in \{0, 1\}$ and $\hat{p}_n \in (0, 1)$.

The Brier score provides an insightful decomposition as BS = Uncertainty − Resolution + Reliability (Murphy, 1973). "Uncertainty" corresponds to the inherent variability over labels; "Resolution" quantifies the deviation of individual predictions from the average; and "Reliability" gauges the extent to which predicted probabilities align with the actual probabilities.

ECE Expected Calibration Error (ECE) measures the average difference between predicted probabilities and their corresponding empirical probabilities, whose calculation is presented in § 3 and rewritten here specifically for the binary classification case.

Binary ECE puts the predicted probabilities $\{\hat{p}_n\}_{n=1}^N$ into S (which is set to $S = 15$ in our experiments) equal-width bins $\mathbb{B}_s = \{n \in \{1, \dots, N\} \mid \hat{p}_n \in (\rho_s, \rho_{s+1}]\}$ with $s \in \{1, \dots, S\}$. For binary classification, ECE is calculated as

$$\begin{aligned} \text{ECE} &= \sum_{s=1}^{S} \frac{|\mathbb{B}_s|}{N} \left|\text{acc}(\mathbb{B}_s) - \text{conf}(\mathbb{B}_s)\right|; \\ \text{acc}(\mathbb{B}_s) = \frac{1}{|\mathbb{B}_s|}\sum_{n\in\mathbb{B}_s} y_n; &\qquad \text{conf}(\mathbb{B}_s) = \frac{1}{|\mathbb{B}_s|}\sum_{n\in\mathbb{B}_s} \hat{p}_n. \end{aligned} \tag{17}$$

Here, acc$(\cdot)$ is the predicion accuracy and conf$(\cdot)$ is the confidence level within each bin. Notice that $|\cdot|$ represents the size when used with a set, and the absolute value when its input is a real number.

Regression CE Regression calibration is less studied than classification calibration, and Regression Calibration Error (CE), proposed by Kuleshov et al. (2018), is adopted in only a few works (Song et al., 2019; Kamarthi et al., 2021). Similar to ECE, CE measures the average difference between the observed confidence level and the expected confidence level within each bin. Specifically, a confidence level of α is measured as the fraction of data points whose true value falls within $\frac{1-\alpha}{2}$th quantile and $(1 - \frac{\alpha}{2})$th quantile of the predicted distribution. For example, 90% confidence level contains data points with true values falling within the 5%th quantile and 95%th quantile of the predicted distribution. With this established, we can calculate the confidence levels with the predicted quantile function, in our case, the cumulative distribution function (CDF) of the predicted Gaussian. Specifically,

$$\text{CE} = \frac{1}{S}\sum_{s=1}^{S}\left(\rho_s - \frac{1}{N}\left|\{n \in \{1, \dots, N\} \mid \Phi(y_n; \hat{y}_n, \hat{\sigma}_n) \le \rho_s\}\right|\right)^2 \tag{18}$$

where $\rho_s = \frac{s}{S}$ is the expected quantile, Φ is the Gaussian CDF:

$$\begin{aligned} \Phi(y_n; \hat{y}_n, \hat{\sigma}_n) &= \frac{1}{2}\left(1 + \text{erf}\left(\frac{y_n - \hat{y}_n}{\sqrt{2}\hat{\sigma}_n}\right)\right); \\ \text{erf}(z) &= \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2} dt. \end{aligned} \tag{19}$$

$\hat{p}_s = \frac{1}{N}\left|\{n \in \{1, \dots, N\} \mid \Phi(y_n; \hat{y}_n, \hat{\sigma}_n) \le \rho_s\}\right|$ is the empirical frequency of the predicted values falling into the ρ_sth quantile. This is slightly different from the confidence level-based calculation, where

$$\text{CE} = \frac{2}{S}\sum_{s=1}^{\frac{S}{2}}\left((\rho_{S-s} - \rho_s) - (\hat{p}_{S-s} - \hat{p}_s)\right)^2, \tag{20}$$

but it reveals the same trend. We use $S = 20$ in our experiments.

E Additional Results

In this section, we supplement the discussions from § 5 with further results and analyses. Given the extensive volume of outcomes from our experiments, we have consolidated them into tables at the end of this article. For enhanced accessibility and offline analysis, we have also made these results available as CSV files in our GitHub repository.[10]

E.1 More Visualization of Uncertainty Estimation Performance

Table 8 displays the average rankings of UQ methods, complemented by Figure 7 which offers a visual representation. As elaborated in § 5, both Deep Ensembles and MC Dropout generally enhance prediction and uncertainty estimation across tasks. However, the performance boost is more noticeable with Deep Ensembles compared to MC Dropout, while DC Dropout is much cheaper to apply. For classification tasks, Temperature Scaling offers moderate efficacy. Given the minimal computational cost of implementing Temperature Scaling, it emerges as a favorable option for calibration enhancement when necessary. In regression tasks, both BBP and SGLD stand out as reliable choices irrespective of the backbone model, particularly for uncertainty estimation. Nonetheless, they mandate a full re-training. Conversely, SWAG and Focal Loss are underwhelming and warrant reconsideration in selection processes while dealing with molecular property prediction.

E.2 Comprehensive Fine-Tuning Results

Tables 10–17 illustrate scores associated with each backbone-UQ pairing across the scaffold split classification datasets, while Tables 18–23 detail the scores regarding regression metrics. Scores are represented in a "mean (standard deviation)" structure, where both mean and standard deviation values are derived from 3 experiment runs conducted with random seeds 0, 1, and 2. As Deep Ensembles aggregates network predictions prior to metric calculation, standard deviation values are not generated and metrics are calculated only once.

Challenges are encountered during training on the notably imbalanced HIV and MUV dataset, depicted in Table 4. SGLD at times proves unsuccessful due to weak directional cues from the true labels, overwhelmed by the random noise embedded in the Langevin dynamics. With TorchMD-NET backbone, SGLD even fails training and constantly produces NaN values regardless of the hyperparameters we choose. Conversely, the application of Focal Loss demonstrates effective performance on the MUV dataset, aligning with assertions in Lin et al. (2017). ECE also may fail to compute on these datasets as predicted probabilities $\hat{p}$ are excessively minimal, leading to unpopulated bins and subsequent "division by zero" errors.

In terms of the majority of datasets, the performance trends closely adhere to our previous discussions. However, an anomaly is observed with the BBBP and ClinTox datasets, where ChemBERTa produces higher performance than the more powerful models, deviating from prior findings (Ahmad et al., 2022; Zhou et al., 2023). Conceptually, the ChemBERTa checkpoint we use is solely pre-trained on the MLM objective, which is unrelated to downstream tasks and should not include label information. Consequently, we hypothesize that either the SMILES strings are potentially more expressive than other descriptors with respect to the physiological properties given that the ChemBERTa's architectures is not considered as more expressive than GROVER or Uni-Mol.

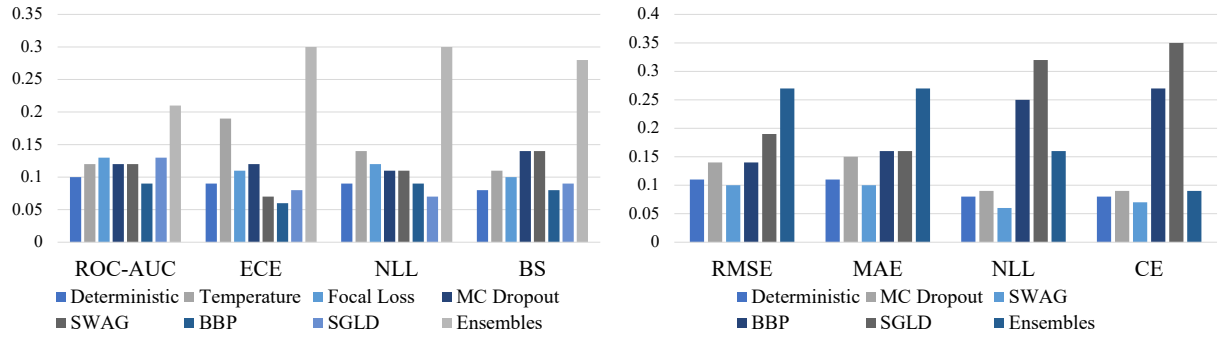
E.3 Results on Random Split Datasets

Tables 24 to 31 illustrate the efficacy of primary backbone models across datasets with random divisions. Mirroring the discussion in appendix E.1, Table 9 and Figure 8 depict the average (reciprocal) rankings of UQ methods. For the most part, the comparative performance among the backbone models and UQ methods is consistent with those observed in scaffold split datasets. Nonetheless, certain exceptions emerge, especially concerning BBP, SGLD, and SWAG. On randomly split classification datasets, both BBP and SGLD continue to underperform, while a marked decline is noted in their regression NLL. This phenomenon might be attributed to the balanced nature of random split datasets, intensifying the predictive indications

[10] Access them here: https://github.com/Yinghao-Li/MUBen/tree/main/reports.

Table 8: Averaged ranks of UQ methods across all scaffold-split datasets and backbone models. Lower value indicates better resuls. Temperature Scaling and Focal Loss are not applicable to regression tasks.

	Classification				Regression			
	ROC-AUC	ECE	NLL	BS	RMSE	MAE	NLL	CE
Deterministic	16.12	15.41	17.03	16.53	12.75	12.21	16.12	15.58
Temperature	16.31	12.06	13.97	14.38	-	-	-	-
Focal Loss	17.12	22.84	20.41	21.97	-	-	-	-
MC Dropout	15.78	13.53	14.28	13.62	10.71	10.83	14.75	15.0
SWAG	15.47	17.94	19.62	17.69	14.0	13.92	18.58	17.21
BBP	18.97	20.28	17.0	18.88	14.04	14.42	8.29	7.67
SGLD	20.72	20.66	20.59	19.22	14.25	14.75	5.87	5.54
Ensembles	11.5	9.28	9.09	9.72	9.25	8.88	11.38	14.0



(a) Classification MRR↑

(b) Regression MRR↑

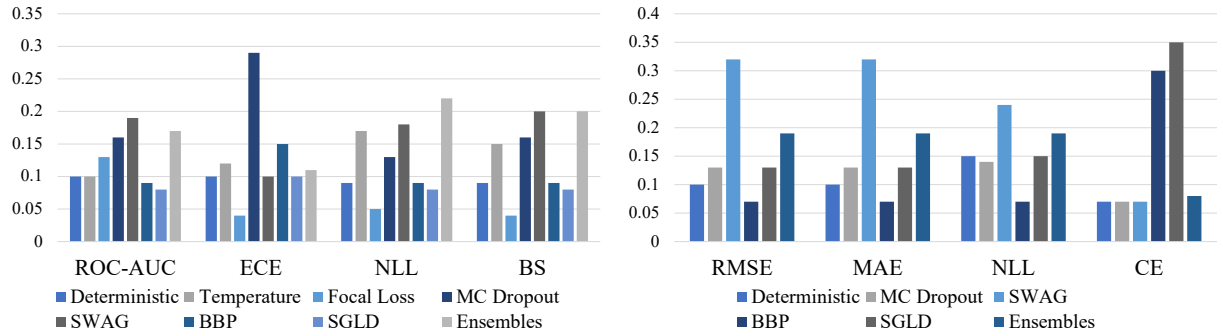
Figure 7: MRR of the UQ methods for each metric, each is macro-averaged from the reciprocal ranks of the results of all corresponding backbone models on all datasets.

from actual labels. Given this environment, the noise from Langevin dynamics or Gaussian samples, which usually aids generalization, might become detrimental. This revelation highlights the capability of traditional Bayesian methods in aptly navigating OOD data (*i.e.*, addressing epistemic uncertainty), mainly through curbing overfitting. However, they may struggle when dealing with in-distribution attributes (*i.e.*, addressing aleatoric uncertainty) having relatively stable labels in the context of molecular property prediction. On the other hand, SWAG, another approximation of the Bayesian network that is almost useless on OOD data, delivers a much more impressive performance on random split datasets. Deep Ensembles, MC Dropout, and Temperature Scaling still emerge as the most reliable UQ methods, although Deep Ensembles no longer dominates the benchmark in the in-distribution case.

Upon comparing Figure 9 with Figure 2, it's evident that the relative positions of the backbone models remain largely unchanged for regression datasets across both splitting methods; however, this consistency falters for classification datasets. Notably, Uni-Mol displays a marked performance decline in the randomly split datasets, positioning it below the more straightforward and cost-efficient backbones, ChemBERTa and GROVER. While definitive explanations for this observed behavior remain elusive, we hypothesize that the molecular descriptor or the pre-training data inherent to ChemBERTa and GROVER could render them better equipped to adapt to Physiological and Biophysical features. Such potential advantages might be obscured by their model architectures' suboptimal generalization capabilities when scaffold splitting is employed.

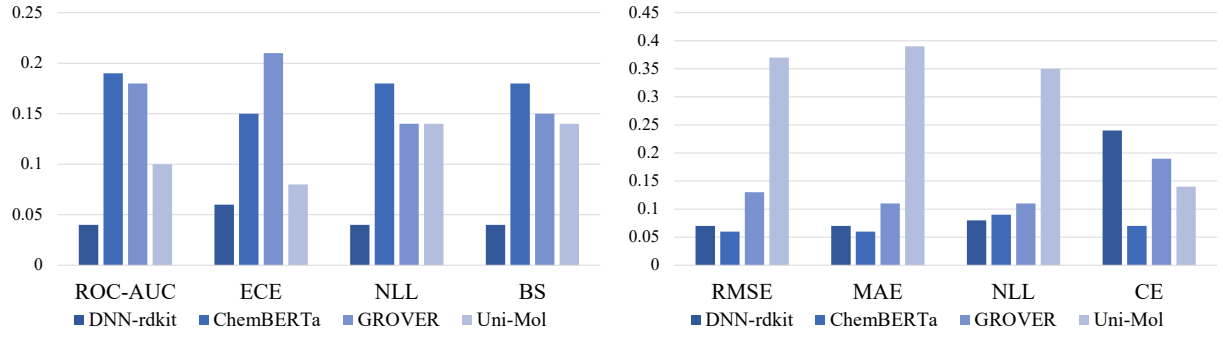
Table 9: Averaged ranks of UQ methods across all randomly split datasets. Lower rank indicates better performance.

	Classification				Regression			
	ROC-AUC	ECE	NLL	BS	RMSE	MAE	NLL	CE
Deterministic	16.44	16.31	17.06	16.5	13.25	12.69	12.62	15.88
Temperature	16.88	14.0	13.06	14.5	-	-	-	-
Focal Loss	17.31	28.75	23.62	26.56	-	-	-	-
MC Dropout	14.88	9.56	12.88	13.75	10.81	10.94	10.75	15.56
SWAG	13.81	15.0	15.06	14.0	10.44	10.31	11.19	16.44
BBP	19.94	16.12	18.12	17.31	16.56	17.5	15.94	5.44
SGLD	18.44	18.81	19.0	17.56	14.12	14.12	14.19	6.69
Ensembles	14.31	13.44	13.19	11.81	9.81	9.44	10.31	15.0



(a) Classification MRR↑ (b) Regression MRR↑

Figure 8: MRR of the UQ methods for each metric, each is macro-averaged from the reciprocal ranks of the results of all corresponding backbone models on all randomly split datasets.

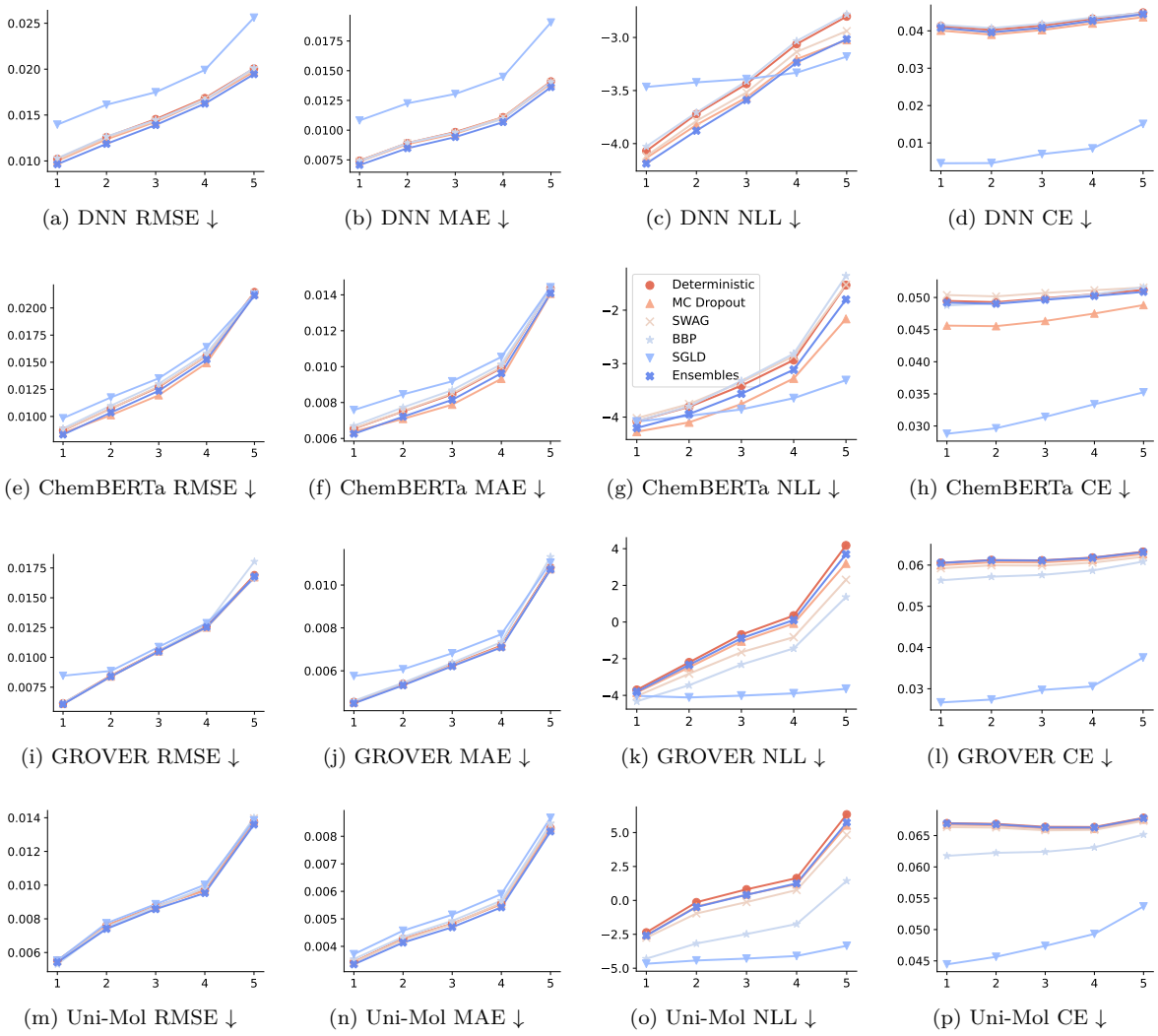


(a) Classification MRR↑ (b) Regression MRR↑

Figure 9: MRR of the UQ methods for each backbone model, each is macro-averaged from the reciprocal ranks of the results of all corresponding UQ methods on all randomly split datasets.

E.4 Results with Frozen Backbone Models

Analysis of Tables 32 through 39 reveals that when the network parameters of the backbone models are fixed, all models underperform significantly. Among the three primary backbone models, Uni-Mol experiences the most drastic decline, emerging as the least effective in the majority of evaluations. These findings highlight



(a) DNN RMSE ↓ (b) DNN MAE ↓ (c) DNN NLL ↓ (d) DNN CE ↓

(e) ChemBERTa RMSE ↓ (f) ChemBERTa MAE ↓ (g) ChemBERTa NLL ↓ (h) ChemBERTa CE ↓

(i) GROVER RMSE ↓ (j) GROVER MAE ↓ (k) GROVER NLL ↓ (l) GROVER CE ↓

(m) Uni-Mol RMSE ↓ (n) Uni-Mol MAE ↓ (o) Uni-Mol NLL ↓ (p) Uni-Mol CE ↓

Figure 10: Performance of backbone and UQ methods according to the similarity between test data points and training scaffolds in the QM9 dataset.

that these backbone models are not inherently optimized for mere feature extraction. The fine-tuning process is pivotal for ensuring optimal model performance, irrespective of the use of UQ methods.

Table 10: Test results on BACE in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.8185 ± 0.0164	0.2827 ± 0.0155	0.8946 ± 0.0797	0.2810 ± 0.0125
	Temperature	0.8185 ± 0.0164	0.2537 ± 0.0144	0.7412 ± 0.0472	0.2553 ± 0.0111
	Focal Loss	0.8190 ± 0.0112	0.2543 ± 0.0011	0.6908 ± 0.0099	0.2488 ± 0.0027
	MC Dropout	0.8168 ± 0.0121	0.2648 ± 0.0323	0.7667 ± 0.0446	0.2589 ± 0.0157
	SWAG	0.8173 ± 0.0119	0.2728 ± 0.0068	0.8730 ± 0.0643	0.2730 ± 0.0085
	BBP	0.8288 ± 0.0039	0.2711 ± 0.0193	0.7943 ± 0.0235	0.2535 ± 0.0075
	SGLD	0.8181 ± 0.0160	0.2713 ± 0.0024	0.9023 ± 0.0416	0.2734 ± 0.0073
	Ensembles	0.8207 ± -	0.2550 ± -	0.7985 ± -	0.2557 ± -
ChemBERTa	Deterministic	0.7223 ± 0.0089	0.2407 ± 0.0135	0.7567 ± 0.0203	0.2720 ± 0.0081
	Temperature	0.7223 ± 0.0089	0.1758 ± 0.0059	0.6937 ± 0.0118	0.2498 ± 0.0052
	Focal Loss	0.6985 ± 0.0062	0.2140 ± 0.0070	0.7322 ± 0.0051	0.2679 ± 0.0023
	MC Dropout	0.7613 ± 0.0216	0.2391 ± 0.0262	0.7464 ± 0.0257	0.2673 ± 0.0070
	SWAG	0.7580 ± 0.0139	0.2539 ± 0.0091	0.7707 ± 0.0374	0.2654 ± 0.0139
	BBP	0.7372 ± 0.0233	0.2767 ± 0.0032	0.8325 ± 0.0190	0.2878 ± 0.0046
	SGLD	0.7775 ± 0.0205	0.2492 ± 0.0217	0.7922 ± 0.0283	0.2588 ± 0.0069
	Ensembles	0.7350 ± -	0.2352 ± -	0.7457 ± -	0.2683 ± -
GROVER	Deterministic	0.8308 ± 0.0012	0.2304 ± 0.0658	0.6921 ± 0.1114	0.2342 ± 0.0386
	Temperature	0.8305 ± 0.0017	0.2083 ± 0.0521	0.6251 ± 0.0671	0.2184 ± 0.0283
	Focal Loss	0.8509 ± 0.0072	0.2334 ± 0.0302	0.6397 ± 0.0300	0.2281 ± 0.0141
	MC Dropout	0.8278 ± 0.0011	0.2366 ± 0.0629	0.6899 ± 0.1094	0.2368 ± 0.0394
	SWAG	0.8604 ± 0.0048	0.2239 ± 0.0016	0.6656 ± 0.0169	0.2148 ± 0.0030
	BBP	0.8322 ± 0.0099	0.2570 ± 0.0331	0.7001 ± 0.0583	0.2450 ± 0.0228
	SGLD	0.8494 ± 0.0028	0.2281 ± 0.0066	0.6769 ± 0.0144	0.2204 ± 0.0038
	Ensembles	0.8388 ± -	0.2288 ± -	0.6571 ± -	0.2252 ± -
Uni-Mol	Deterministic	0.8080 ± 0.0266	0.2419 ± 0.0409	0.7928 ± 0.1639	0.2517 ± 0.0335
	Temperature	0.8087 ± 0.0277	0.2125 ± 0.0259	0.6782 ± 0.0662	0.2352 ± 0.0212
	Focal Loss	0.8208 ± 0.0259	0.2445 ± 0.0287	0.6721 ± 0.0482	0.2406 ± 0.0198
	MC Dropout	0.8268 ± 0.0269	0.2483 ± 0.0333	0.8065 ± 0.1383	0.2512 ± 0.0331
	SWAG	0.8193 ± 0.0106	0.2677 ± 0.0075	0.9415 ± 0.0623	0.2625 ± 0.0058
	BBP	0.8386 ± 0.0135	0.2673 ± 0.0473	0.8158 ± 0.1584	0.2588 ± 0.0375
	SGLD	0.8699 ± 0.0118	0.2437 ± 0.0131	0.7354 ± 0.0502	0.2244 ± 0.0150
	Ensembles	0.8505 ± -	0.2176 ± -	0.6301 ± -	0.2184 ± -
TorchMD-NET	Deterministic	0.7945 ± 0.0232	0.2193 ± 0.0374	1.1663 ± 0.6182	0.2272 ± 0.0201
	Temperature	0.7945 ± 0.0231	0.1808 ± 0.0425	0.7151 ± 0.1720	0.2085 ± 0.0187
	Focal Loss	0.7602 ± 0.0089	0.2141 ± 0.0249	0.7328 ± 0.0376	0.2420 ± 0.0200
	MC Dropout	0.8018 ± 0.0260	0.2017 ± 0.0297	0.8454 ± 0.2767	0.2178 ± 0.0199
	SWAG	0.7457 ± 0.0208	0.3287 ± 0.0164	1.9159 ± 0.3026	0.3172 ± 0.0122
	BBP	0.7638 ± 0.0374	0.3130 ± 0.0457	2.8823 ± 1.2929	0.3035 ± 0.0425
	SGLD	0.7787 ± 0.0173	0.2793 ± 0.0199	1.1648 ± 0.2135	0.2731 ± 0.0226
	Ensembles	0.8169 ± -	0.2220 ± -	0.6703 ± -	0.2197 ± -
GIN	Deterministic	0.7312 ± 0.0152	0.4024 ± 0.0362	1.5262 ± 0.4443	0.3946 ± 0.0340
	Temperature	0.7312 ± 0.0152	0.3494 ± 0.0392	1.0850 ± 0.1950	0.3482 ± 0.0285
	Focal Loss	0.7272 ± 0.0374	0.3034 ± 0.1213	1.2375 ± 0.6434	0.3264 ± 0.0917
	MC Dropout	0.7312 ± 0.0152	0.4024 ± 0.0362	1.5262 ± 0.4443	0.3946 ± 0.0340
	SWAG	0.5554 ± 0.0075	0.1929 ± 0.0326	0.7817 ± 0.0442	0.2839 ± 0.0156
	BBP	0.6775 ± 0.0533	0.2567 ± 0.0644	1.3036 ± 0.7463	0.2842 ± 0.0546
	SGLD	0.6682 ± 0.0302	0.2271 ± 0.1028	0.8664 ± 0.2399	0.2887 ± 0.0569
	Ensembles	0.7767 ± -	0.2051 ± -	0.7157 ± -	0.2469 ± -

Table 11: Test results on BBBP in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.6721 ± 0.0129	0.3292 ± 0.0517	3.8147 ± 4.0782	0.3423 ± 0.0378
	Temperature	0.6719 ± 0.0129	0.2890 ± 0.0742	2.1338 ± 1.8963	0.3186 ± 0.0489
	Focal Loss	0.6693 ± 0.0047	0.1689 ± 0.1056	0.8925 ± 0.3424	0.2623 ± 0.0470
	MC Dropout	0.6669 ± 0.0151	0.3287 ± 0.0402	2.1699 ± 1.7478	0.3391 ± 0.0296
	SWAG	0.6777 ± 0.0136	0.3020 ± 0.0694	4.8415 ± 5.5407	0.3268 ± 0.0465
	BBP	0.6856 ± 0.0075	0.3392 ± 0.0506	1.8126 ± 0.6030	0.3463 ± 0.0357
	SGLD	0.6922 ± 0.0125	0.3134 ± 0.0487	1.7569 ± 1.0592	0.3247 ± 0.0396
	Ensembles	0.7018 ± -	0.2615 ± -	0.9043 ± -	0.2938 ± -
ChemBERTa	Deterministic	0.7407 ± 0.0101	0.2259 ± 0.0206	1.0306 ± 0.1002	0.2559 ± 0.0090
	Temperature	0.7407 ± 0.0101	0.2107 ± 0.0196	0.8539 ± 0.0266	0.2453 ± 0.0064
	Focal Loss	0.7410 ± 0.0122	0.1639 ± 0.0118	0.6995 ± 0.0246	0.2313 ± 0.0087
	MC Dropout	0.7451 ± 0.0067	0.2082 ± 0.0035	0.8814 ± 0.0292	0.2418 ± 0.0032
	SWAG	0.7408 ± 0.0112	0.2373 ± 0.0202	1.1435 ± 0.0810	0.2629 ± 0.0109
	BBP	0.7245 ± 0.0112	0.2415 ± 0.0293	0.9653 ± 0.1062	0.2652 ± 0.0120
	SGLD	0.7359 ± 0.0104	0.2398 ± 0.0256	1.0692 ± 0.1940	0.2580 ± 0.0056
	Ensembles	0.7399 ± -	0.2290 ± -	0.9672 ± -	0.2506 ± -
GROVER	Deterministic	0.7075 ± 0.0061	0.2902 ± 0.0111	0.9457 ± 0.0234	0.2998 ± 0.0060
	Temperature	0.7072 ± 0.0059	0.2685 ± 0.0147	0.8607 ± 0.0258	0.2865 ± 0.0069
	Focal Loss	0.7146 ± 0.0059	0.1340 ± 0.0177	0.6505 ± 0.0070	0.2284 ± 0.0040
	MC Dropout	0.7066 ± 0.0071	0.2901 ± 0.0147	0.9267 ± 0.0285	0.2996 ± 0.0071
	SWAG	0.7160 ± 0.0032	0.3314 ± 0.0032	1.1105 ± 0.0142	0.3251 ± 0.0025
	BBP	0.6968 ± 0.0152	0.2735 ± 0.0209	0.9146 ± 0.0326	0.2986 ± 0.0055
	SGLD	0.7266 ± 0.0037	0.3127 ± 0.0065	1.0187 ± 0.0202	0.3149 ± 0.0019
	Ensembles	0.7083 ± -	0.2890 ± -	0.9338 ± -	0.3029 ± -
Uni-Mol	Deterministic	0.7103 ± 0.0075	0.3311 ± 0.0506	1.4209 ± 0.3000	0.3354 ± 0.0397
	Temperature	0.7093 ± 0.0083	0.2922 ± 0.0418	1.0023 ± 0.1064	0.3084 ± 0.0315
	Focal Loss	0.7102 ± 0.0045	0.2031 ± 0.0816	0.7646 ± 0.0923	0.2632 ± 0.0275
	MC Dropout	0.7029 ± 0.0142	0.3328 ± 0.0483	1.4086 ± 0.2828	0.3382 ± 0.0372
	SWAG	0.7247 ± 0.0062	0.3547 ± 0.0136	1.8530 ± 0.0824	0.3456 ± 0.0078
	BBP	0.6902 ± 0.0286	0.3420 ± 0.0318	1.3548 ± 0.0713	0.3463 ± 0.0275
	SGLD	0.7147 ± 0.0118	0.3336 ± 0.0127	1.4174 ± 0.0475	0.3359 ± 0.0059
	Ensembles	0.7253 ± -	0.3035 ± -	1.0889 ± -	0.3035 ± -
TorchMD-NET	Deterministic	0.6753 ± 0.0060	0.3207 ± 0.0249	1.7352 ± 0.2714	0.3299 ± 0.0218
	Temperature	0.6753 ± 0.0060	0.2739 ± 0.0251	1.1382 ± 0.1017	0.2995 ± 0.0188
	Focal Loss	0.7060 ± 0.0206	0.1910 ± 0.0200	0.8688 ± 0.0498	0.2487 ± 0.0119
	MC Dropout	0.6767 ± 0.0057	0.3098 ± 0.0236	1.4864 ± 0.1962	0.3190 ± 0.0177
	SWAG	0.6720 ± 0.0131	0.3601 ± 0.0121	3.1976 ± 0.5027	0.3632 ± 0.0090
	BBP	0.6895 ± 0.0166	0.3019 ± 0.0598	3.1956 ± 1.7451	0.3159 ± 0.0343
	SGLD	0.6708 ± 0.0058	0.3629 ± 0.0083	2.3208 ± 0.5515	0.3617 ± 0.0068
	Ensembles	0.6844 ± -	0.2700 ± -	1.1018 ± -	0.2934 ± -
GIN	Deterministic	0.6282 ± 0.0025	0.2277 ± 0.0302	1.0566 ± 0.0130	0.2920 ± 0.0078
	Temperature	0.6281 ± 0.0025	0.1987 ± 0.0164	0.9092 ± 0.0166	0.2778 ± 0.0051
	Focal Loss	0.6234 ± 0.0081	0.1550 ± 0.0116	0.7581 ± 0.0143	0.2611 ± 0.0043
	MC Dropout	0.6282 ± 0.0025	0.2277 ± 0.0302	1.0566 ± 0.0130	0.2920 ± 0.0078
	SWAG	0.6067 ± 0.0131	0.3162 ± 0.0021	1.2223 ± 0.0235	0.3353 ± 0.0056
	BBP	0.6100 ± 0.0095	0.2738 ± 0.0056	1.1617 ± 0.0529	0.3119 ± 0.0058
	SGLD	0.5577 ± 0.0497	0.3493 ± 0.0217	1.4064 ± 0.1006	0.3681 ± 0.0218
	Ensembles	0.6298 ± -	0.2112 ± -	0.9696 ± -	0.2802 ± -

Table 12: Test results on ClinTox in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.8301 ± 0.0226	0.0637 ± 0.0126	0.2719 ± 0.0481	0.0597 ± 0.0052
	Temperature	0.8300 ± 0.0224	0.0547 ± 0.0109	0.2371 ± 0.0059	0.0572 ± 0.0039
	Focal Loss	0.8507 ± 0.0165	0.0812 ± 0.0368	0.2284 ± 0.0305	0.0639 ± 0.0114
	MC Dropout	0.8194 ± 0.0217	0.0445 ± 0.0055	0.2149 ± 0.0307	0.0534 ± 0.0005
	SWAG	0.8290 ± 0.0163	0.0590 ± 0.0117	0.2964 ± 0.0617	0.0591 ± 0.0054
	BBP	0.7541 ± 0.0102	0.0574 ± 0.0121	0.2439 ± 0.0276	0.0609 ± 0.0046
	SGLD	0.8335 ± 0.0109	0.0567 ± 0.0125	0.2769 ± 0.0452	0.0553 ± 0.0063
	Ensembles	0.8154 ± -	0.0283 ± -	0.1860 ± -	0.0484 ± -
ChemBERTa	Deterministic	0.9856 ± 0.0028	0.0215 ± 0.0035	0.1206 ± 0.0318	0.0225 ± 0.0016
	Temperature	0.9856 ± 0.0028	0.0199 ± 0.0026	0.0964 ± 0.0147	0.0222 ± 0.0013
	Focal Loss	0.9832 ± 0.0027	0.0897 ± 0.0378	0.1367 ± 0.0350	0.0273 ± 0.0045
	MC Dropout	0.9848 ± 0.0021	0.0212 ± 0.0019	0.1083 ± 0.0247	0.0213 ± 0.0021
	SWAG	0.9850 ± 0.0031	0.0222 ± 0.0025	0.1327 ± 0.0317	0.0232 ± 0.0010
	BBP	0.9816 ± 0.0079	0.0325 ± 0.0212	0.0972 ± 0.0156	0.0226 ± 0.0024
	SGLD	0.9869 ± 0.0012	0.0183 ± 0.0006	0.0893 ± 0.0050	0.0214 ± 0.0005
	Ensembles	0.9812 ± -	0.0157 ± -	0.0810 ± -	0.0208 ± -
GROVER	Deterministic	0.9398 ± 0.0152	0.0635 ± 0.0114	0.1471 ± 0.0106	0.0367 ± 0.0031
	Temperature	0.9409 ± 0.0130	0.0449 ± 0.0055	0.1260 ± 0.0045	0.0333 ± 0.0034
	Focal Loss	0.9377 ± 0.0042	0.2158 ± 0.0034	0.2963 ± 0.0018	0.0731 ± 0.0011
	MC Dropout	0.9347 ± 0.0152	0.0706 ± 0.0122	0.1475 ± 0.0121	0.0366 ± 0.0036
	SWAG	0.9623 ± 0.0060	0.0460 ± 0.0117	0.1119 ± 0.0046	0.0280 ± 0.0022
	BBP	0.9085 ± 0.0109	0.0940 ± 0.0104	0.1976 ± 0.0136	0.0488 ± 0.0033
	SGLD	0.9507 ± 0.0058	0.0595 ± 0.0065	0.1287 ± 0.0063	0.0306 ± 0.0011
	Ensembles	0.9429 ± -	0.0743 ± -	0.1453 ± -	0.0344 ± -
Uni-Mol	Deterministic	0.8814 ± 0.0379	0.0522 ± 0.0120	0.1822 ± 0.0381	0.0484 ± 0.0124
	Temperature	0.8830 ± 0.0375	0.0417 ± 0.0158	0.1754 ± 0.0358	0.0477 ± 0.0121
	Focal Loss	0.8836 ± 0.0202	0.1651 ± 0.0556	0.2972 ± 0.0587	0.0827 ± 0.0207
	MC Dropout	0.8745 ± 0.0438	0.0560 ± 0.0131	0.1889 ± 0.0406	0.0521 ± 0.0138
	SWAG	0.9055 ± 0.0027	0.0311 ± 0.0034	0.1450 ± 0.0035	0.0337 ± 0.0009
	BBP	0.9257 ± 0.0287	0.0379 ± 0.0026	0.1468 ± 0.0047	0.0398 ± 0.0032
	SGLD	0.8598 ± 0.0209	0.0327 ± 0.0029	0.1619 ± 0.0059	0.0389 ± 0.0014
	Ensembles	0.9088 ± -	0.0389 ± -	0.1476 ± -	0.0386 ± -
TorchMD-NET	Deterministic	0.8938 ± 0.0178	0.0567 ± 0.0058	0.1905 ± 0.0083	0.0470 ± 0.0014
	Temperature	0.8938 ± 0.0178	0.0608 ± 0.0100	0.1869 ± 0.0076	0.0487 ± 0.0016
	Focal Loss	0.8952 ± 0.0195	0.1260 ± 0.0242	0.2582 ± 0.0218	0.0683 ± 0.0050
	MC Dropout	0.8898 ± 0.0192	0.0505 ± 0.0029	0.1711 ± 0.0054	0.0460 ± 0.0013
	SWAG	0.8995 ± 0.0185	0.0517 ± 0.0082	0.1969 ± 0.0065	0.0486 ± 0.0027
	BBP	0.8801 ± 0.0278	0.0667 ± 0.0139	0.2483 ± 0.0463	0.0610 ± 0.0077
	SGLD	0.8855 ± 0.0177	0.0475 ± 0.0045	0.1776 ± 0.0123	0.0475 ± 0.0045
	Ensembles	0.8944 ± -	0.0569 ± -	0.1886 ± -	0.0485 ± -
GIN	Deterministic	0.6965 ± 0.0184	0.1042 ± 0.0000	0.3398 ± 0.0260	0.0798 ± 0.0044
	Temperature	0.6965 ± 0.0184	0.0876 ± 0.0000	0.3595 ± 0.0346	0.0789 ± 0.0044
	Focal Loss	0.6777 ± 0.0168	0.2690 ± 0.0232	0.4610 ± 0.0314	0.1448 ± 0.0148
	MC Dropout	0.6965 ± 0.0184	0.1042 ± 0.0000	0.3398 ± 0.0260	0.0798 ± 0.0044
	SWAG	0.6552 ± 0.0197	0.1000 ± 0.0000	0.4573 ± 0.1485	0.0799 ± 0.0031
	BBP	0.6122 ± 0.0078	- ± -	0.5270 ± 0.1577	0.0888 ± 0.0131
	SGLD	0.6539 ± 0.0312	0.1256 ± 0.0000	0.4745 ± 0.1417	0.0827 ± 0.0046
	Ensembles	0.6872 ± -	0.1170 ± -	0.3222 ± -	0.0868 ± -

Table 13: Test results on Tox21 in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.7386 ± 0.0061	0.0417 ± 0.0028	0.2771 ± 0.0024	0.0779 ± 0.0013
	Temperature	0.7386 ± 0.0061	0.0342 ± 0.0027	0.2723 ± 0.0028	0.0773 ± 0.0013
	Focal Loss	0.7374 ± 0.0038	0.1058 ± 0.0060	0.3161 ± 0.0078	0.0871 ± 0.0024
	MC Dropout	0.7376 ± 0.0039	0.0356 ± 0.0016	0.2727 ± 0.0016	0.0763 ± 0.0013
	SWAG	0.7364 ± 0.0045	0.0438 ± 0.0010	0.2793 ± 0.0037	0.0790 ± 0.0017
	BBP	0.7243 ± 0.0036	0.0422 ± 0.0019	0.2847 ± 0.0010	0.0814 ± 0.0002
	SGLD	0.7257 ± 0.0018	0.1192 ± 0.0332	0.3455 ± 0.0315	0.0978 ± 0.0100
	Ensembles	0.7540 ± -	0.0344 ± -	0.2648 ± -	0.0746 ± -
ChemBERTa	Deterministic	0.7542 ± 0.0009	0.0571 ± 0.0020	0.2962 ± 0.0067	0.0812 ± 0.0009
	Temperature	0.7542 ± 0.0009	0.0424 ± 0.0040	0.2744 ± 0.0027	0.0792 ± 0.0011
	Focal Loss	0.7523 ± 0.0022	0.0969 ± 0.0012	0.3052 ± 0.0018	0.0845 ± 0.0004
	MC Dropout	0.7641 ± 0.0032	0.0423 ± 0.0023	0.2697 ± 0.0032	0.0744 ± 0.0005
	SWAG	0.7538 ± 0.0021	0.0592 ± 0.0022	0.3008 ± 0.0079	0.0818 ± 0.0011
	BBP	0.7433 ± 0.0055	0.0459 ± 0.0028	0.2765 ± 0.0035	0.0780 ± 0.0014
	SGLD	0.7475 ± 0.0036	0.0504 ± 0.0038	0.2784 ± 0.0030	0.0795 ± 0.0014
	Ensembles	0.7681 ± -	0.0440 ± -	0.2679 ± -	0.0750 ± -
GROVER	Deterministic	0.7808 ± 0.0017	0.0358 ± 0.0038	0.2473 ± 0.0022	0.0694 ± 0.0005
	Temperature	0.7810 ± 0.0016	0.0291 ± 0.0016	0.2439 ± 0.0008	0.0686 ± 0.0003
	Focal Loss	0.7779 ± 0.0028	0.1148 ± 0.0102	0.3052 ± 0.0094	0.0811 ± 0.0028
	MC Dropout	0.7817 ± 0.0018	0.0346 ± 0.0015	0.2455 ± 0.0016	0.0689 ± 0.0005
	SWAG	0.7837 ± 0.0017	0.0359 ± 0.0009	0.2482 ± 0.0009	0.0689 ± 0.0001
	BBP	0.7697 ± 0.0031	0.0438 ± 0.0017	0.2552 ± 0.0016	0.0711 ± 0.0003
	SGLD	0.7635 ± 0.0009	0.0402 ± 0.0024	0.2558 ± 0.0010	0.0716 ± 0.0002
	Ensembles	0.7876 ± -	0.0316 ± -	0.2411 ± -	0.0675 ± -
Uni-Mol	Deterministic	0.7895 ± 0.0017	0.0454 ± 0.0056	0.2601 ± 0.0072	0.0716 ± 0.0022
	Temperature	0.7896 ± 0.0016	0.0346 ± 0.0027	0.2483 ± 0.0035	0.0704 ± 0.0014
	Focal Loss	0.7904 ± 0.0040	0.0972 ± 0.0045	0.2899 ± 0.0009	0.0785 ± 0.0001
	MC Dropout	0.7891 ± 0.0021	0.0480 ± 0.0066	0.2628 ± 0.0065	0.0726 ± 0.0021
	SWAG	0.7842 ± 0.0048	0.0593 ± 0.0025	0.2994 ± 0.0054	0.0728 ± 0.0012
	BBP	0.7932 ± 0.0043	0.0396 ± 0.0048	0.2520 ± 0.0039	0.0703 ± 0.0010
	SGLD	0.7887 ± 0.0059	0.0433 ± 0.0023	0.2569 ± 0.0055	0.0684 ± 0.0014
	Ensembles	0.8052 ± -	0.0332 ± -	0.2389 ± -	0.0662 ± -
TorchMD-NET	Deterministic	0.7342 ± 0.0143	0.0738 ± 0.0146	0.4523 ± 0.1083	0.0857 ± 0.0044
	Temperature	0.7342 ± 0.0143	0.0515 ± 0.0047	0.2944 ± 0.0115	0.0801 ± 0.0014
	Focal Loss	0.7403 ± 0.0032	0.0639 ± 0.0044	0.3102 ± 0.0140	0.0808 ± 0.0027
	MC Dropout	0.7369 ± 0.0144	0.0685 ± 0.0131	0.3913 ± 0.0759	0.0837 ± 0.0033
	SWAG	0.7347 ± 0.0093	0.0740 ± 0.0040	0.4567 ± 0.0576	0.0832 ± 0.0018
	BBP	0.7459 ± 0.0067	0.0808 ± 0.0075	0.5355 ± 0.1830	0.0888 ± 0.0016
	SGLD	0.7387 ± 0.0051	0.0427 ± 0.0019	0.2761 ± 0.0029	0.0757 ± 0.0004
	Ensembles	0.7793 ± -	0.0409 ± -	0.2614 ± -	0.0708 ± -
GIN	Deterministic	0.6706 ± 0.0032	0.0789 ± 0.0172	0.3573 ± 0.0052	0.0916 ± 0.0070
	Temperature	0.6706 ± 0.0032	0.0914 ± 0.0158	0.3393 ± 0.0150	0.0941 ± 0.0068
	Focal Loss	0.6571 ± 0.0061	0.1040 ± 0.0073	0.3361 ± 0.0036	0.0930 ± 0.0014
	MC Dropout	0.6706 ± 0.0032	0.0789 ± 0.0172	0.3573 ± 0.0052	0.0916 ± 0.0070
	SWAG	0.6619 ± 0.0103	0.0671 ± 0.0032	0.3536 ± 0.0029	0.0867 ± 0.0026
	BBP	0.6425 ± 0.0034	0.0851 ± 0.0084	0.3632 ± 0.0039	0.0949 ± 0.0033
	SGLD	0.6190 ± 0.0048	0.0967 ± 0.0135	0.3817 ± 0.0275	0.0995 ± 0.0056
	Ensembles	0.6829 ± -	0.0634 ± -	0.3268 ± -	0.0840 ± -

Table 14: Test results on ToxCast in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.6222 ± 0.0042	0.1168 ± 0.0096	0.4436 ± 0.0146	0.1397 ± 0.0021
	Temperature	0.6220 ± 0.0046	0.1114 ± 0.0050	0.4882 ± 0.0170	0.1398 ± 0.0015
	Focal Loss	0.6289 ± 0.0077	0.1264 ± 0.0034	0.4389 ± 0.0084	0.1396 ± 0.0024
	MC Dropout	0.6248 ± 0.0030	0.1093 ± 0.0075	0.4319 ± 0.0096	0.1358 ± 0.0015
	SWAG	0.6207 ± 0.0073	0.1175 ± 0.0098	0.4440 ± 0.0150	0.1400 ± 0.0023
	BBP	0.6020 ± 0.0028	0.1443 ± 0.0060	0.4673 ± 0.0059	0.1510 ± 0.0020
	SGLD	0.5319 ± 0.0040	0.3054 ± 0.0034	0.6685 ± 0.0035	0.2378 ± 0.0017
	Ensembles	0.6486 ± -	0.0900 ± -	0.4008 ± -	0.1292 ± -
ChemBERTa	Deterministic	0.6554 ± 0.0037	0.1209 ± 0.0026	0.4313 ± 0.0054	0.1330 ± 0.0008
	Temperature	0.6540 ± 0.0042	0.1067 ± 0.0019	0.4817 ± 0.0261	0.1313 ± 0.0004
	Focal Loss	0.6442 ± 0.0170	0.1197 ± 0.0037	0.4243 ± 0.0153	0.1346 ± 0.0051
	MC Dropout	0.6624 ± 0.0024	0.1069 ± 0.0036	0.4070 ± 0.0052	0.1276 ± 0.0015
	SWAG	0.6556 ± 0.0033	0.1202 ± 0.0036	0.4305 ± 0.0064	0.1327 ± 0.0010
	BBP	0.5814 ± 0.0075	0.1276 ± 0.0015	0.4545 ± 0.0011	0.1469 ± 0.0003
	SGLD	0.5436 ± 0.0056	0.2238 ± 0.0019	0.5602 ± 0.0022	0.1881 ± 0.0011
	Ensembles	0.6733 ± -	0.1037 ± -	0.3986 ± -	0.1258 ± -
GROVER	Deterministic	0.6587 ± 0.0018	0.1043 ± 0.0006	0.4091 ± 0.0023	0.1298 ± 0.0004
	Temperature	0.6496 ± 0.0023	0.1424 ± 0.0026	0.4612 ± 0.0045	0.1424 ± 0.0013
	Focal Loss	0.6359 ± 0.0026	0.1221 ± 0.0005	0.4365 ± 0.0012	0.1383 ± 0.0004
	MC Dropout	0.6615 ± 0.0014	0.1009 ± 0.0010	0.4042 ± 0.0022	0.1288 ± 0.0004
	SWAG	0.6603 ± 0.0020	0.1060 ± 0.0007	0.4114 ± 0.0015	0.1301 ± 0.0001
	BBP	0.5995 ± 0.0026	0.1731 ± 0.0041	0.5090 ± 0.0040	0.1660 ± 0.0015
	SGLD	0.5542 ± 0.0069	0.2712 ± 0.0008	0.6194 ± 0.0006	0.2139 ± 0.0003
	Ensembles	0.6646 ± -	0.1034 ± -	0.4061 ± -	0.1290 ± -
Uni-Mol	Deterministic	0.6734 ± 0.0101	0.1020 ± 0.0023	0.3983 ± 0.0075	0.1274 ± 0.0025
	Temperature	0.7028 ± 0.0101	0.1456 ± 0.0045	0.4566 ± 0.0104	0.1355 ± 0.0029
	Focal Loss	0.6934 ± 0.0035	0.1227 ± 0.0011	0.4079 ± 0.0016	0.1284 ± 0.0003
	MC Dropout	0.6833 ± 0.0117	0.1074 ± 0.0090	0.4015 ± 0.0112	0.1274 ± 0.0025
	SWAG	0.6870 ± 0.0075	0.1085 ± 0.0034	0.4005 ± 0.0055	0.1271 ± 0.0019
	BBP	0.6273 ± 0.0066	0.1296 ± 0.0065	0.4522 ± 0.0068	0.1456 ± 0.0021
	SGLD	0.5700 ± 0.0005	0.1953 ± 0.0033	0.5207 ± 0.0029	0.1717 ± 0.0012
	Ensembles	0.6841 ± -	0.0953 ± -	0.3877 ± -	0.1247 ± -
TorchMD-NET	Deterministic	0.6361 ± 0.0001	0.1734 ± 0.0053	0.4870 ± 0.0164	0.1486 ± 0.0027
	Temperature	0.6384 ± 0.0020	- ± -	0.6053 ± 0.0984	0.1466 ± 0.0009
	Focal Loss	0.6443 ± 0.0073	0.1543 ± 0.0032	0.4337 ± 0.0011	0.1395 ± 0.0006
	MC Dropout	0.6339 ± 0.0042	0.1707 ± 0.0048	0.4778 ± 0.0141	0.1477 ± 0.0027
	SWAG	0.6457 ± 0.0064	0.1648 ± 0.0012	0.4617 ± 0.0025	0.1446 ± 0.0004
	BBP	0.6103 ± 0.0182	0.1471 ± 0.0049	0.4539 ± 0.0171	0.1434 ± 0.0026
	SGLD	0.5224 ± 0.0048	0.2976 ± 0.0020	0.6678 ± 0.0020	0.2374 ± 0.0010
	Ensembles	0.6540 ± -	0.1546 ± -	0.4424 ± -	0.1396 ± -
GIN	Deterministic	0.5667 ± 0.0032	0.1443 ± 0.0018	0.5142 ± 0.0048	0.1506 ± 0.0015
	Temperature	0.5669 ± 0.0028	0.1402 ± 0.0013	0.5333 ± 0.0192	0.1476 ± 0.0014
	Focal Loss	0.5483 ± 0.0028	0.1446 ± 0.0027	0.4615 ± 0.0009	0.1468 ± 0.0006
	MC Dropout	0.5667 ± 0.0032	0.1443 ± 0.0018	0.5142 ± 0.0048	0.1506 ± 0.0015
	SWAG	0.5633 ± 0.0059	0.1418 ± 0.0018	0.5125 ± 0.0052	0.1495 ± 0.0014
	BBP	0.5096 ± 0.0059	0.2243 ± 0.0007	0.6405 ± 0.0128	0.1935 ± 0.0014
	SGLD	0.4872 ± 0.0050	0.3166 ± 0.0019	0.8083 ± 0.0245	0.2668 ± 0.0047
	Ensembles	0.5752 ± -	0.1381 ± -	0.4835 ± -	0.1477 ± -

Table 15: Test results on SIDER in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.5981 ± 0.0152	0.1188 ± 0.0344	0.5462 ± 0.0469	0.1760 ± 0.0124
	Temperature	0.5981 ± 0.0152	0.1033 ± 0.0271	0.5186 ± 0.0270	0.1709 ± 0.0094
	Focal Loss	0.5864 ± 0.0071	0.1663 ± 0.0217	0.5605 ± 0.0193	0.1893 ± 0.0075
	MC Dropout	0.5977 ± 0.0164	0.1071 ± 0.0271	0.5245 ± 0.0304	0.1705 ± 0.0084
	SWAG	0.5988 ± 0.0210	0.1206 ± 0.0333	0.5477 ± 0.0491	0.1764 ± 0.0125
	BBP	0.5738 ± 0.0272	0.1215 ± 0.0046	0.5241 ± 0.0077	0.1733 ± 0.0027
	SGLD	0.5895 ± 0.0087	0.1559 ± 0.0119	0.5653 ± 0.0119	0.1900 ± 0.0051
	Ensembles	0.6158 ± -	0.0954 ± -	0.4950 ± -	0.1631 ± -
ChemBERTa	Deterministic	0.6146 ± 0.0078	0.1483 ± 0.0102	0.5740 ± 0.0217	0.1804 ± 0.0055
	Temperature	0.6146 ± 0.0078	0.1275 ± 0.0104	0.5256 ± 0.0138	0.1721 ± 0.0043
	Focal Loss	0.6119 ± 0.0105	0.1498 ± 0.0136	0.5346 ± 0.0129	0.1777 ± 0.0046
	MC Dropout	0.6166 ± 0.0022	0.1136 ± 0.0143	0.5191 ± 0.0131	0.1672 ± 0.0037
	SWAG	0.6149 ± 0.0095	0.1527 ± 0.0090	0.5836 ± 0.0230	0.1825 ± 0.0056
	BBP	0.6067 ± 0.0108	0.1184 ± 0.0018	0.5127 ± 0.0043	0.1686 ± 0.0013
	SGLD	0.6127 ± 0.0110	0.1299 ± 0.0118	0.5255 ± 0.0157	0.1725 ± 0.0053
	Ensembles	0.6193 ± -	0.1305 ± -	0.5396 ± -	0.1718 ± -
GROVER	Deterministic	0.6213 ± 0.0027	0.1034 ± 0.0061	0.5023 ± 0.0071	0.1637 ± 0.0021
	Temperature	0.6209 ± 0.0030	0.0939 ± 0.0040	0.4935 ± 0.0046	0.1618 ± 0.0018
	Focal Loss	0.6059 ± 0.0165	0.1817 ± 0.0019	0.5666 ± 0.0027	0.1903 ± 0.0011
	MC Dropout	0.6248 ± 0.0060	0.0969 ± 0.0035	0.4942 ± 0.0057	0.1616 ± 0.0018
	SWAG	0.6133 ± 0.0024	0.1183 ± 0.0016	0.5280 ± 0.0066	0.1702 ± 0.0019
	BBP	0.5926 ± 0.0052	0.1347 ± 0.0014	0.5367 ± 0.0023	0.1762 ± 0.0009
	SGLD	0.5989 ± 0.0129	0.1012 ± 0.0040	0.5017 ± 0.0051	0.1637 ± 0.0019
	Ensembles	0.6237 ± -	0.0878 ± -	0.4917 ± -	0.1604 ± -
Uni-Mol	Deterministic	0.6214 ± 0.0073	0.1054 ± 0.0130	0.5047 ± 0.0184	0.1661 ± 0.0065
	Temperature	0.6210 ± 0.0075	0.0950 ± 0.0074	0.4937 ± 0.0101	0.1625 ± 0.0040
	Focal Loss	0.6200 ± 0.0118	0.1660 ± 0.0096	0.5471 ± 0.0090	0.1834 ± 0.0038
	MC Dropout	0.6125 ± 0.0060	0.1269 ± 0.0178	0.5324 ± 0.0270	0.1739 ± 0.0071
	SWAG	0.6246 ± 0.0090	0.1327 ± 0.0084	0.5397 ± 0.0125	0.1753 ± 0.0033
	BBP	0.6301 ± 0.0260	0.0991 ± 0.0130	0.4928 ± 0.0119	0.1612 ± 0.0049
	SGLD	0.6323 ± 0.0108	0.1286 ± 0.0093	0.5408 ± 0.0137	0.1725 ± 0.0030
	Ensembles	0.6378 ± -	0.0872 ± -	0.4836 ± -	0.1585 ± -
TorchMD-NET	Deterministic	0.5881 ± 0.0060	0.1305 ± 0.0239	0.5988 ± 0.0541	0.1789 ± 0.0072
	Temperature	0.5881 ± 0.0060	0.1142 ± 0.0109	0.5380 ± 0.0147	0.1721 ± 0.0038
	Focal Loss	0.5909 ± 0.0048	0.1515 ± 0.0034	0.5470 ± 0.0058	0.1848 ± 0.0031
	MC Dropout	0.5862 ± 0.0059	0.1230 ± 0.0205	0.5551 ± 0.0327	0.1756 ± 0.0061
	SWAG	0.6081 ± 0.0085	0.1281 ± 0.0152	0.5596 ± 0.0326	0.1751 ± 0.0059
	BBP	0.6072 ± 0.0150	0.1293 ± 0.0005	0.5619 ± 0.0059	0.1753 ± 0.0008
	SGLD	0.6026 ± 0.0089	0.1143 ± 0.0076	0.5162 ± 0.0054	0.1693 ± 0.0017
	Ensembles	0.6329 ± -	0.1093 ± -	0.5175 ± -	0.1660 ± -
GIN	Deterministic	0.6159 ± 0.0016	0.0900 ± 0.0020	0.5385 ± 0.0084	0.1627 ± 0.0009
	Temperature	0.6159 ± 0.0017	0.0916 ± 0.0031	0.5219 ± 0.0125	0.1625 ± 0.0009
	Focal Loss	0.5892 ± 0.0046	0.1863 ± 0.0011	0.5764 ± 0.0034	0.1953 ± 0.0013
	MC Dropout	0.6159 ± 0.0016	0.0900 ± 0.0020	0.5385 ± 0.0084	0.1627 ± 0.0009
	SWAG	0.6054 ± 0.0034	0.0959 ± 0.0042	0.5520 ± 0.0021	0.1642 ± 0.0011
	BBP	0.5499 ± 0.0132	0.1477 ± 0.0071	0.5666 ± 0.0079	0.1876 ± 0.0037
	SGLD	0.5792 ± 0.0074	0.1121 ± 0.0028	0.6447 ± 0.0421	0.1723 ± 0.0018
	Ensembles	0.6232 ± -	0.0874 ± -	0.5238 ± -	0.1584 ± -

Table 16: Test results on HIV in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.7345 ± 0.0065	0.0077 ± 0.0010	0.1236 ± 0.0016	0.0276 ± 0.0006
	Temperature	0.7345 ± 0.0065	0.0136 ± 0.0005	0.1277 ± 0.0021	0.0277 ± 0.0007
	Focal Loss	0.7338 ± 0.0078	0.1205 ± 0.0179	0.2059 ± 0.0179	0.0429 ± 0.0042
	MC Dropout	0.7365 ± 0.0088	0.0066 ± 0.0010	0.1219 ± 0.0016	0.0270 ± 0.0004
	SWAG	0.7439 ± 0.0014	0.0105 ± 0.0025	0.1240 ± 0.0013	0.0279 ± 0.0005
	BBP	0.7541 ± 0.0120	0.0181 ± 0.0065	0.1235 ± 0.0036	0.0277 ± 0.0011
	SGLD	0.7414 ± 0.0019	0.0160 ± 0.0048	0.1313 ± 0.0061	0.0281 ± 0.0007
	Ensembles	0.7613 ± -	0.0074 ± -	0.1160 ± -	0.0260 ± -
ChemBERTa	Deterministic	0.7681 ± 0.0061	0.0214 ± 0.0033	0.1247 ± 0.0058	0.0292 ± 0.0019
	Temperature	0.7681 ± 0.0061	0.0221 ± 0.0039	0.1259 ± 0.0067	0.0289 ± 0.0018
	Focal Loss	0.7509 ± 0.0106	0.0651 ± 0.0351	0.1599 ± 0.0205	0.0343 ± 0.0025
	MC Dropout	0.7717 ± 0.0042	0.0079 ± 0.0014	0.1105 ± 0.0011	0.0247 ± 0.0005
	SWAG	0.7474 ± 0.0123	0.0263 ± 0.0045	0.1384 ± 0.0063	0.0319 ± 0.0017
	BBP	0.7672 ± 0.0031	0.0217 ± 0.0043	0.1241 ± 0.0071	0.0289 ± 0.0027
	SGLD	0.7224 ± 0.0095	0.0374 ± 0.0011	0.1602 ± 0.0106	0.0376 ± 0.0002
	Ensembles	0.7885 ± -	0.0205 ± -	0.1161 ± -	0.0267 ± -
GROVER	Deterministic	0.7979 ± 0.0045	0.0078 ± 0.0032	0.1133 ± 0.0019	0.0263 ± 0.0005
	Temperature	0.7977 ± 0.0046	0.0120 ± 0.0011	0.1154 ± 0.0022	0.0265 ± 0.0004
	Focal Loss	0.7746 ± 0.0113	0.1286 ± 0.0223	0.2156 ± 0.0259	0.0444 ± 0.0066
	MC Dropout	0.8002 ± 0.0018	0.0077 ± 0.0028	0.1126 ± 0.0006	0.0261 ± 0.0001
	SWAG	0.8021 ± 0.0033	0.0090 ± 0.0019	0.1118 ± 0.0018	0.0258 ± 0.0005
	BBP	0.7928 ± 0.0116	0.0162 ± 0.0062	0.1151 ± 0.0034	0.0258 ± 0.0002
	SGLD	0.7420 ± 0.0131	0.0200 ± 0.0012	0.1231 ± 0.0001	0.0272 ± 0.0000
	Ensembles	0.8103 ± -	0.0035 ± -	0.1081 ± -	0.0251 ± -
Uni-Mol	Deterministic	0.7913 ± 0.0042	0.0189 ± 0.0063	0.1335 ± 0.0225	0.0273 ± 0.0014
	Temperature	0.7915 ± 0.0041	0.0151 ± 0.0017	0.1190 ± 0.0040	0.0269 ± 0.0009
	Focal Loss	0.7785 ± 0.0073	0.0713 ± 0.0251	0.1561 ± 0.0205	0.0316 ± 0.0032
	MC Dropout	0.7858 ± 0.0044	0.0229 ± 0.0091	0.1557 ± 0.0339	0.0294 ± 0.0033
	SWAG	0.7786 ± 0.0061	0.0264 ± 0.0015	0.1676 ± 0.0160	0.0286 ± 0.0011
	BBP	0.7781 ± 0.0128	0.0191 ± 0.0071	0.1428 ± 0.0327	0.0284 ± 0.0016
	SGLD	0.7777 ± 0.0121	0.0205 ± 0.0006	0.1373 ± 0.0087	0.0263 ± 0.0002
	Ensembles	0.8020 ± -	0.0181 ± -	0.1181 ± -	0.0265 ± -
TorchMD-NET	Deterministic	0.7221 ± 0.0133	0.0270 ± 0.0031	0.2340 ± 0.0668	0.0315 ± 0.0007
	Temperature	0.7221 ± 0.0133	0.0266 ± 0.0025	0.1532 ± 0.0024	0.0317 ± 0.0008
	Focal Loss	0.7350 ± 0.0170	0.0673 ± 0.0312	0.1585 ± 0.0212	0.0332 ± 0.0036
	MC Dropout	0.7225 ± 0.0098	0.0265 ± 0.0036	0.2191 ± 0.0590	0.0313 ± 0.0009
	SWAG	0.7555 ± 0.0132	0.0253 ± 0.0007	0.2110 ± 0.0153	0.0288 ± 0.0006
	BBP	0.7521 ± 0.0078	0.0195 ± 0.0114	0.1384 ± 0.0251	0.0290 ± 0.0036
	SGLD	- ± -	- ± -	- ± -	- ± -
	Ensembles	0.7370 ± -	0.0218 ± -	0.1537 ± -	0.0284 ± -
GIN	Deterministic	0.6893 ± 0.0214	- ± -	0.1715 ± 0.0040	0.0396 ± 0.0021
	Temperature	0.6893 ± 0.0214	- ± -	0.1708 ± 0.0032	0.0378 ± 0.0024
	Focal Loss	0.6703 ± 0.0204	- ± -	0.2283 ± 0.0151	0.0515 ± 0.0046
	MC Dropout	0.6893 ± 0.0214	- ± -	0.1715 ± 0.0040	0.0396 ± 0.0021
	SWAG	0.6863 ± 0.0305	- ± -	0.1496 ± 0.0091	0.0318 ± 0.0007
	BBP	0.6947 ± 0.0106	- ± -	0.1625 ± 0.0055	0.0360 ± 0.0021
	SGLD	0.6845 ± 0.0063	- ± -	0.1954 ± 0.0322	0.0362 ± 0.0026
	Ensembles	0.7209 ± -	- ± -	0.1492 ± -	0.0336 ± -

Table 17: Test results on MUV in the format of "metric mean ± standard deviation".

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.6029 ± 0.0576	0.0015 ± 0.0003	0.0326 ± 0.0196	0.0026 ± 0.0001
	Temperature	0.6059 ± 0.0538	- ± -	0.0321 ± 0.0145	0.0027 ± 0.0003
	Focal Loss	0.6261 ± 0.0322	0.0780 ± 0.0033	0.0720 ± 0.0262	0.0076 ± 0.0032
	MC Dropout	0.6282 ± 0.0530	0.0015 ± 0.0003	0.0256 ± 0.0114	0.0025 ± 0.0000
	SWAG	0.5830 ± 0.0392	0.0025 ± 0.0006	0.0393 ± 0.0254	0.0027 ± 0.0001
	BBP	0.5974 ± 0.0287	- ± -	0.0251 ± 0.0068	0.0027 ± 0.0001
	SGLD	0.4955 ± 0.0509	0.2380 ± 0.0156	0.2798 ± 0.0201	0.0610 ± 0.0071
	Ensembles	0.6736 ± -	0.0013 ± -	0.0175 ± -	0.0025 ± -
ChemBERTa	Deterministic	0.6695 ± 0.0343	0.0030 ± 0.0001	0.0241 ± 0.0011	0.0029 ± 0.0000
	Temperature	0.6695 ± 0.0343	0.0026 ± 0.0003	0.0196 ± 0.0010	0.0029 ± 0.0000
	Focal Loss	0.7150 ± 0.0038	0.0208 ± 0.0119	0.0333 ± 0.0100	0.0040 ± 0.0007
	MC Dropout	0.7168 ± 0.0100	0.0026 ± 0.0003	0.0230 ± 0.0027	0.0026 ± 0.0001
	SWAG	0.6672 ± 0.0342	0.0031 ± 0.0000	0.0261 ± 0.0015	0.0029 ± 0.0000
	BBP	0.5970 ± 0.0622	0.0026 ± 0.0010	0.0198 ± 0.0018	0.0030 ± 0.0004
	SGLD	0.5030 ± 0.0346	0.0074 ± 0.0003	0.0217 ± 0.0002	0.0026 ± 0.0000
	Ensembles	0.7224 ± -	0.0025 ± -	0.0197 ± -	0.0027 ± -
GROVER	Deterministic	0.7157 ± 0.0330	0.0016 ± 0.0004	0.0169 ± 0.0004	0.0025 ± 0.0000
	Temperature	0.7155 ± 0.0331	0.0014 ± 0.0001	0.0187 ± 0.0004	0.0025 ± 0.0000
	Focal Loss	0.6825 ± 0.0127	0.0869 ± 0.0033	0.0995 ± 0.0036	0.0101 ± 0.0006
	MC Dropout	0.7163 ± 0.0363	0.0016 ± 0.0004	0.0170 ± 0.0004	0.0025 ± 0.0000
	SWAG	0.7281 ± 0.0226	0.0014 ± 0.0003	0.0162 ± 0.0002	0.0025 ± 0.0000
	BBP	0.7490 ± 0.0068	0.0052 ± 0.0015	0.0192 ± 0.0011	0.0025 ± 0.0000
	SGLD	0.5532 ± 0.0255	0.0375 ± 0.0009	0.0488 ± 0.0008	0.0039 ± 0.0001
	Ensembles	0.7268 ± -	0.0014 ± -	0.0166 ± -	0.0025 ± -
Uni-Mol	Deterministic	0.7718 ± 0.0200	0.0024 ± 0.0002	0.0225 ± 0.0027	0.0025 ± 0.0000
	Temperature	0.7625 ± 0.0098	0.0017 ± 0.0000	0.0233 ± 0.0039	0.0025 ± 0.0001
	Focal Loss	0.8239 ± 0.0140	0.0250 ± 0.0126	0.0354 ± 0.0116	0.0036 ± 0.0006
	MC Dropout	0.7781 ± 0.0201	0.0026 ± 0.0001	0.0253 ± 0.0017	0.0025 ± 0.0000
	SWAG	0.7429 ± 0.0442	0.0028 ± 0.0001	0.0276 ± 0.0016	0.0027 ± 0.0001
	BBP	0.7394 ± 0.0140	0.0016 ± 0.0002	0.0169 ± 0.0002	0.0025 ± 0.0000
	SGLD	0.5901 ± 0.0200	0.0014 ± 0.0000	0.0178 ± 0.0000	0.0025 ± 0.0000
	Ensembles	0.7904 ± -	0.0023 ± -	0.0195 ± -	0.0025 ± -
TorchMD-NET	Deterministic	0.7043 ± 0.0027	- ± -	0.0553 ± 0.0092	0.0027 ± 0.0001
	Temperature	0.7073 ± 0.0016	- ± -	0.0318 ± 0.0059	0.0027 ± 0.0001
	Focal Loss	0.7227 ± 0.0117	- ± -	0.0225 ± 0.0034	0.0028 ± 0.0002
	MC Dropout	0.7003 ± 0.0057	- ± -	0.0455 ± 0.0074	0.0026 ± 0.0001
	SWAG	0.7285 ± 0.0094	- ± -	0.0457 ± 0.0011	0.0029 ± 0.0001
	BBP	0.7330 ± 0.0044	- ± -	0.0407 ± 0.0070	0.0029 ± 0.0002
	SGLD	- ± -	- ± -	- ± -	- ± -
	Ensembles	0.7045 ± -	- ± -	0.0437 ± -	0.0025 ± -
GIN	Deterministic	0.6996 ± 0.0094	0.0024 ± 0.0006	0.0224 ± 0.0034	0.0028 ± 0.0001
	Temperature	0.6885 ± 0.0072	- ± -	0.0284 ± 0.0037	0.0028 ± 0.0002
	Focal Loss	0.6964 ± 0.0121	0.0146 ± 0.0053	0.0280 ± 0.0042	0.0037 ± 0.0004
	MC Dropout	0.6996 ± 0.0094	0.0024 ± 0.0006	0.0224 ± 0.0034	0.0028 ± 0.0001
	SWAG	0.6898 ± 0.0132	0.0021 ± 0.0003	0.0226 ± 0.0030	0.0026 ± 0.0000
	BBP	0.6425 ± 0.0114	0.0022 ± 0.0002	0.0229 ± 0.0009	0.0027 ± 0.0001
	SGLD	0.5553 ± 0.0110	0.0063 ± 0.0007	0.0400 ± 0.0058	0.0029 ± 0.0001
	Ensembles	0.7155 ± -	0.0021 ± -	0.0191 ± -	0.0026 ± -

Table 18: Test results on ESOL in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.9048 ± 0.0289	0.6747 ± 0.0059	1.0033 ± 0.3765	0.0465 ± 0.0052
	MC Dropout	0.9129 ± 0.0212	0.6853 ± 0.0049	1.0246 ± 0.3698	0.0495 ± 0.0051
	SWAG	0.9165 ± 0.0244	0.6827 ± 0.0074	0.9461 ± 0.3428	0.0459 ± 0.0048
	BBP	0.9557 ± 0.0298	0.7070 ± 0.0237	0.4306 ± 0.0325	0.0325 ± 0.0038
	SGLD	0.9630 ± 0.0425	0.7175 ± 0.0302	0.5101 ± 0.0380	0.0162 ± 0.0017
	Ensembles	0.8690 ± -	0.6455 ± -	0.5674 ± -	0.0453 ± -
ChemBERTa	Deterministic	1.0044 ± 0.0358	0.7989 ± 0.0216	2.2803 ± 0.1218	0.0500 ± 0.0007
	MC Dropout	0.9497 ± 0.0271	0.7620 ± 0.0176	1.4258 ± 0.1905	0.0481 ± 0.0023
	SWAG	0.9966 ± 0.0334	0.7940 ± 0.0206	2.5815 ± 0.2470	0.0513 ± 0.0013
	BBP	1.0125 ± 0.0113	0.8139 ± 0.0110	0.5990 ± 0.0499	0.0337 ± 0.0013
	SGLD	0.9978 ± 0.0340	0.8083 ± 0.0225	1.2727 ± 0.0711	0.0464 ± 0.0013
	Ensembles	0.9752 ± -	0.7763 ± -	1.4981 ± -	0.0470 ± -
GROVER	Deterministic	0.9355 ± 0.0252	0.7396 ± 0.0193	0.9888 ± 0.0395	0.0451 ± 0.0006
	MC Dropout	0.9437 ± 0.0246	0.7491 ± 0.0216	1.0065 ± 0.0351	0.0449 ± 0.0007
	SWAG	0.9443 ± 0.0219	0.7420 ± 0.0189	1.2996 ± 0.1883	0.0470 ± 0.0011
	BBP	0.9892 ± 0.0430	0.7821 ± 0.0327	0.5470 ± 0.0378	0.0153 ± 0.0028
	SGLD	0.9955 ± 0.0104	0.7942 ± 0.0153	0.4786 ± 0.0140	0.0243 ± 0.0008
	Ensembles	0.9110 ± -	0.7209 ± -	0.5854 ± -	0.0394 ± -
Uni-Mol	Deterministic	0.8530 ± 0.0224	0.6809 ± 0.0164	1.0739 ± 0.1136	0.0527 ± 0.0004
	MC Dropout	0.8367 ± 0.0237	0.6648 ± 0.0209	1.1098 ± 0.2625	0.0527 ± 0.0012
	SWAG	0.8644 ± 0.0097	0.6950 ± 0.0042	1.3868 ± 0.1522	0.0540 ± 0.0003
	BBP	0.8441 ± 0.0332	0.6719 ± 0.0279	0.3151 ± 0.0292	0.0329 ± 0.0015
	SGLD	0.8037 ± 0.0121	0.6444 ± 0.0112	0.2685 ± 0.0187	0.0343 ± 0.0018
	Ensembles	0.8181 ± -	0.6516 ± -	0.8372 ± -	0.0512 ± -
TorchMD-NET	Deterministic	0.8772 ± 0.0638	0.6688 ± 0.0517	0.6817 ± 0.1947	0.0409 ± 0.0059
	MC Dropout	0.8720 ± 0.0645	0.6660 ± 0.0510	0.6166 ± 0.1780	0.0402 ± 0.0063
	SWAG	0.8569 ± 0.0581	0.6642 ± 0.0401	0.4144 ± 0.1136	0.0372 ± 0.0053
	BBP	2.4684 ± 0.5232	1.9253 ± 0.3748	1.5283 ± 0.0753	0.0191 ± 0.0173
	SGLD	0.9691 ± 0.0579	0.7843 ± 0.0439	0.5284 ± 0.0243	0.0143 ± 0.0016
	Ensembles	1.0512 ± -	0.8394 ± -	0.9756 ± -	0.0005 ± -
GIN	Deterministic	1.5513 ± 0.0639	1.2268 ± 0.0618	1.0705 ± 0.1566	0.0122 ± 0.0072
	MC Dropout	1.5513 ± 0.0639	1.2268 ± 0.0618	1.0705 ± 0.1566	0.0122 ± 0.0072
	SWAG	1.8390 ± 0.2814	1.4650 ± 0.2128	1.4527 ± 0.2933	0.0166 ± 0.0089
	BBP	1.9791 ± 0.0920	1.5041 ± 0.0290	1.4189 ± 0.2794	0.0048 ± 0.0013
	SGLD	3.3157 ± 2.1193	2.7240 ± 1.8086	1.6595 ± 0.4615	0.0424 ± 0.0276
	Ensembles	1.3952 ± -	1.0846 ± -	0.9598 ± -	0.0025 ± -

Table 19: Test results on FreeSolv in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	2.2403 ± 0.1366	1.6768 ± 0.1007	1.3370 ± 0.1120	0.0272 ± 0.0026
	MC Dropout	2.2316 ± 0.1070	1.6469 ± 0.1122	1.3357 ± 0.1257	0.0247 ± 0.0015
	SWAG	2.4516 ± 0.1188	1.7853 ± 0.1478	1.8461 ± 0.2920	0.0358 ± 0.0043
	BBP	2.8456 ± 0.1416	2.1049 ± 0.1037	1.5285 ± 0.0469	0.0173 ± 0.0056
	SGLD	2.5503 ± 0.1010	1.8549 ± 0.1092	1.4260 ± 0.0095	0.0137 ± 0.0033
	Ensembles	1.9627 ± -	1.4359 ± -	1.0149 ± -	0.0240 ± -
ChemBERTa	Deterministic	2.5267 ± 0.0864	1.7760 ± 0.0318	1.7859 ± 0.1916	0.0364 ± 0.0034
	MC Dropout	2.5042 ± 0.0981	1.7804 ± 0.0456	1.7810 ± 0.1526	0.0359 ± 0.0033
	SWAG	2.6269 ± 0.1627	1.8562 ± 0.1055	2.2239 ± 0.2879	0.0426 ± 0.0003
	BBP	2.6340 ± 0.1368	1.9072 ± 0.1600	1.6143 ± 0.1930	0.0322 ± 0.0066
	SGLD	2.6564 ± 0.0357	1.8633 ± 0.0344	1.7627 ± 0.0809	0.0347 ± 0.0053
	Ensembles	2.3727 ± -	1.6904 ± -	1.5142 ± -	0.0341 ± -
GROVER	Deterministic	2.1352 ± 0.0193	1.4770 ± 0.0303	2.8244 ± 0.8307	0.0428 ± 0.0034
	MC Dropout	2.0981 ± 0.0069	1.4660 ± 0.0231	2.6575 ± 0.7926	0.0415 ± 0.0040
	SWAG	2.2321 ± 0.0902	1.5041 ± 0.0460	3.3021 ± 0.0530	0.0469 ± 0.0008
	BBP	2.6133 ± 0.0662	1.8288 ± 0.0843	1.4699 ± 0.0598	0.0126 ± 0.0014
	SGLD	2.1117 ± 0.0985	1.3784 ± 0.0459	1.2791 ± 0.1411	0.0238 ± 0.0020
	Ensembles	2.0238 ± -	1.3539 ± -	1.9474 ± -	0.0393 ± -
Uni-Mol	Deterministic	1.7871 ± 0.1265	1.2906 ± 0.0483	2.5450 ± 1.2435	0.0521 ± 0.0018
	MC Dropout	1.7585 ± 0.1322	1.2652 ± 0.0559	2.3425 ± 1.2327	0.0515 ± 0.0018
	SWAG	1.8524 ± 0.1449	1.3481 ± 0.1063	2.9865 ± 1.1583	0.0545 ± 0.0013
	BBP	1.7464 ± 0.0246	1.2424 ± 0.0049	0.9940 ± 0.0065	0.0289 ± 0.0027
	SGLD	1.7462 ± 0.0908	1.2963 ± 0.0610	1.0358 ± 0.0881	0.0296 ± 0.0048
	Ensembles	1.6716 ± -	1.1780 ± -	1.4306 ± -	0.0504 ± -
TorchMD-NET	Deterministic	2.6436 ± 0.0969	1.9379 ± 0.1236	1.9251 ± 0.2705	0.0219 ± 0.0060
	MC Dropout	2.6360 ± 0.0986	1.9475 ± 0.1212	1.8397 ± 0.2050	0.0212 ± 0.0052
	SWAG	2.6997 ± 0.1009	1.8466 ± 0.0863	2.5681 ± 0.4816	0.0399 ± 0.0026
	BBP	3.6709 ± 0.7927	2.7858 ± 0.6062	1.8484 ± 0.2682	0.0227 ± 0.0055
	SGLD	2.7769 ± 0.1054	1.9696 ± 0.1725	1.5583 ± 0.1031	0.0155 ± 0.0025
	Ensembles	2.9347 ± -	2.0826 ± -	1.6610 ± -	0.0138 ± -
GIN	Deterministic	2.2009 ± 0.2750	1.7292 ± 0.2873	1.3182 ± 0.0981	0.0142 ± 0.0039
	MC Dropout	2.2009 ± 0.2750	1.7292 ± 0.2873	1.3182 ± 0.0981	0.0142 ± 0.0039
	SWAG	2.3232 ± 0.0818	1.8154 ± 0.1109	1.5179 ± 0.1044	0.0198 ± 0.0106
	BBP	3.1290 ± 0.6024	2.4897 ± 0.5111	1.8594 ± 0.2348	0.0092 ± 0.0048
	SGLD	2.7563 ± 0.6927	2.2125 ± 0.6421	1.6046 ± 0.2001	0.0116 ± 0.0093
	Ensembles	2.0088 ± -	1.5676 ± -	1.5171 ± -	0.0005 ± -

Table 20: Test results on Lipophilicity in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.7575 ± 0.0116	0.5793 ± 0.0138	0.6154 ± 0.2016	0.0293 ± 0.0030
	MC Dropout	0.7559 ± 0.0110	0.5773 ± 0.0132	0.9071 ± 0.3721	0.0341 ± 0.0052
	SWAG	0.7572 ± 0.0101	0.5823 ± 0.0129	0.7191 ± 0.2267	0.0308 ± 0.0027
	BBP	0.7730 ± 0.0253	0.5938 ± 0.0245	0.7578 ± 0.1627	0.0305 ± 0.0046
	SGLD	0.7468 ± 0.0018	0.5743 ± 0.0016	0.2152 ± 0.0097	0.0090 ± 0.0009
	Ensembles	0.7172 ± -	0.5490 ± -	0.6165 ± -	0.0322 ± -
ChemBERTa	Deterministic	0.7553 ± 0.0074	0.5910 ± 0.0065	1.2368 ± 0.3994	0.0362 ± 0.0032
	MC Dropout	0.7142 ± 0.0063	0.5601 ± 0.0029	0.8178 ± 0.2701	0.0349 ± 0.0033
	SWAG	0.7672 ± 0.0101	0.5992 ± 0.0080	1.5809 ± 0.4168	0.0395 ± 0.0021
	BBP	0.7542 ± 0.0050	0.5869 ± 0.0087	0.4419 ± 0.0775	0.0279 ± 0.0023
	SGLD	0.7622 ± 0.0045	0.5982 ± 0.0070	0.8719 ± 0.1359	0.0355 ± 0.0020
	Ensembles	0.7367 ± -	0.5763 ± -	0.9756 ± -	0.0360 ± -
GROVER	Deterministic	0.6316 ± 0.0066	0.4747 ± 0.0040	2.1512 ± 0.2203	0.0478 ± 0.0009
	MC Dropout	0.6293 ± 0.0067	0.4740 ± 0.0028	2.0526 ± 0.2139	0.0476 ± 0.0009
	SWAG	0.6317 ± 0.0075	0.4750 ± 0.0041	2.3980 ± 0.3597	0.0485 ± 0.0007
	BBP	0.6481 ± 0.0063	0.5058 ± 0.0041	0.0789 ± 0.0206	0.0196 ± 0.0024
	SGLD	0.6360 ± 0.0041	0.4984 ± 0.0017	0.0544 ± 0.0199	0.0215 ± 0.0010
	Ensembles	0.6250 ± -	0.4693 ± -	1.6046 ± -	0.0460 ± -
Uni-Mol	Deterministic	0.6079 ± 0.0032	0.4509 ± 0.0044	0.8975 ± 0.5565	0.0425 ± 0.0061
	MC Dropout	0.5983 ± 0.0099	0.4438 ± 0.0117	1.3663 ± 1.2187	0.0440 ± 0.0084
	SWAG	0.6026 ± 0.0004	0.4476 ± 0.0022	1.0101 ± 0.1639	0.0453 ± 0.0012
	BBP	0.6044 ± 0.0016	0.4469 ± 0.0025	0.0679 ± 0.0218	0.0306 ± 0.0010
	SGLD	0.6040 ± 0.0031	0.4554 ± 0.0042	0.1565 ± 0.0573	0.0329 ± 0.0017
	Ensembles	0.5809 ± -	0.4266 ± -	0.6450 ± -	0.0438 ± -
TorchMD-NET	Deterministic	0.8235 ± 0.2218	0.6462 ± 0.1853	4.3336 ± 2.6567	0.0447 ± 0.0127
	MC Dropout	0.8887 ± 0.3142	0.7001 ± 0.2635	3.9796 ± 2.2502	0.0439 ± 0.0130
	SWAG	0.6819 ± 0.0271	0.5288 ± 0.0225	2.2934 ± 1.5338	0.0394 ± 0.0144
	BBP	2.8400 ± 2.2383	2.5087 ± 2.1497	1.6533 ± 0.9853	0.1008 ± 0.1219
	SGLD	0.8608 ± 0.0679	0.7002 ± 0.0502	0.9818 ± 0.4045	0.0183 ± 0.0128
	Ensembles	1.0313 ± -	0.8196 ± -	0.8619 ± -	0.0195 ± -
GIN	Deterministic	0.9286 ± 0.1113	0.7455 ± 0.1056	0.4542 ± 0.1096	0.0048 ± 0.0055
	MC Dropout	0.9286 ± 0.1113	0.7455 ± 0.1056	0.4542 ± 0.1096	0.0048 ± 0.0055
	SWAG	1.1105 ± 0.2448	0.9010 ± 0.2179	0.6805 ± 0.2447	0.0147 ± 0.0090
	BBP	1.0863 ± 0.0704	0.8606 ± 0.0492	0.6095 ± 0.0841	0.0018 ± 0.0010
	SGLD	1.3804 ± 0.4327	1.1187 ± 0.3763	0.9884 ± 0.4025	0.0229 ± 0.0137
	Ensembles	0.8071 ± -	0.6515 ± -	0.3241 ± -	0.0020 ± -

Table 21: Test results on QM7 in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	142.9279 ± 2.9477	95.7200 ± 3.2512	10.2976 ± 1.0848	0.0393 ± 0.0036
	MC Dropout	138.6817 ± 2.5873	90.7039 ± 3.1531	7.4627 ± 0.4138	0.0339 ± 0.0020
	SWAG	143.6922 ± 1.9420	97.2590 ± 1.9750	11.3649 ± 1.4063	0.0418 ± 0.0041
	BBP	141.5333 ± 4.8179	99.0430 ± 3.9287	6.8897 ± 0.8805	0.0330 ± 0.0064
	SGLD	139.8857 ± 5.4764	99.4873 ± 2.0159	5.4071 ± 0.0338	0.0098 ± 0.0019
	Ensembles	128.7182 ± -	84.7269 ± -	5.4568 ± -	0.0281 ± -
ChemBERTa	Deterministic	114.1954 ± 1.9473	66.6504 ± 3.6637	6.9619 ± 1.4000	0.0251 ± 0.0059
	MC Dropout	112.4931 ± 1.2013	64.7828 ± 3.4349	6.1942 ± 0.7998	0.0309 ± 0.0061
	SWAG	114.2023 ± 1.5302	68.0992 ± 3.5122	6.9684 ± 1.3236	0.0236 ± 0.0052
	BBP	116.2333 ± 1.8548	69.9783 ± 0.3732	5.5954 ± 0.2659	0.0262 ± 0.0034
	SGLD	116.9667 ± 0.3664	74.6714 ± 1.3652	5.0906 ± 0.0281	0.0147 ± 0.0019
	Ensembles	111.2764 ± -	63.8752 ± -	5.3896 ± -	0.0224 ± -
GROVER	Deterministic	139.5969 ± 3.1922	98.6424 ± 4.8571	5.4936 ± 0.0870	0.0121 ± 0.0012
	MC Dropout	136.1780 ± 2.9204	95.4940 ± 4.4117	5.4272 ± 0.0697	0.0107 ± 0.0014
	SWAG	142.6412 ± 1.3203	107.1310 ± 1.6574	5.7041 ± 0.1256	0.0153 ± 0.0018
	BBP	136.8576 ± 3.2036	99.8460 ± 2.2180	5.4555 ± 0.0135	0.0035 ± 0.0007
	SGLD	159.7489 ± 2.9839	126.6406 ± 2.8867	5.5587 ± 0.0308	0.0072 ± 0.0009
	Ensembles	131.9411 ± -	89.4504 ± -	5.3241 ± -	0.0090 ± -
Uni-Mol	Deterministic	103.9043 ± 1.3754	56.6287 ± 2.3360	4.6198 ± 0.2317	0.0338 ± 0.0065
	MC Dropout	103.5853 ± 1.3201	54.8976 ± 1.3061	4.5158 ± 0.2036	0.0386 ± 0.0019
	SWAG	107.8128 ± 0.1969	63.3542 ± 2.2114	5.5588 ± 0.6515	0.0412 ± 0.0026
	BBP	101.5479 ± 1.1382	51.9550 ± 0.7905	4.5153 ± 0.0535	0.0263 ± 0.0022
	SGLD	105.1573 ± 2.6264	57.0601 ± 4.5219	4.5642 ± 0.0453	0.0290 ± 0.0014
	Ensembles	101.1805 ± -	52.0409 ± -	4.3732 ± -	0.0360 ± -
TorchMD-NET	Deterministic	104.1467 ± 0.3144	49.6743 ± 0.7443	11.6606 ± 0.7198	0.0619 ± 0.0004
	MC Dropout	104.1265 ± 0.3415	49.8274 ± 0.5136	10.8641 ± 0.6233	0.0609 ± 0.0003
	SWAG	103.6753 ± 0.6387	50.1259 ± 1.4952	7.4456 ± 0.2718	0.0580 ± 0.0005
	BBP	104.4771 ± 0.4558	51.7268 ± 0.6870	9.5890 ± 0.4097	0.0568 ± 0.0002
	SGLD	112.7406 ± 3.9319	69.4583 ± 2.8874	5.1299 ± 0.0314	0.0212 ± 0.0022
	Ensembles	102.0535 ± -	48.2234 ± -	10.6292 ± -	0.0618 ± -
GIN	Deterministic	122.4303 ± 0.0690	76.2333 ± 1.0882	5.2490 ± 0.0500	0.0136 ± 0.0003
	MC Dropout	122.4303 ± 0.0690	76.2333 ± 1.0882	5.2490 ± 0.0500	0.0136 ± 0.0003
	SWAG	125.4810 ± 0.9210	83.3207 ± 1.3398	5.3423 ± 0.0619	0.0129 ± 0.0001
	BBP	125.2453 ± 0.6326	85.2772 ± 1.4557	5.3348 ± 0.0059	0.0077 ± 0.0014
	SGLD	132.1431 ± 6.9275	97.6270 ± 11.0966	5.3910 ± 0.0487	0.0126 ± 0.0011
	Ensembles	118.0550 ± -	71.2196 ± -	5.1922 ± -	0.0105 ± -

Table 22: Test results on QM8 in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.0356 ± 0.0004	0.0209 ± 0.0001	-0.7811 ± 2.1740	0.0282 ± 0.0008
	MC Dropout	0.0352 ± 0.0002	0.0209 ± 0.0002	-2.7716 ± 0.1922	0.0267 ± 0.0001
	SWAG	0.0360 ± 0.0002	0.0212 ± 0.0001	-0.7143 ± 1.5808	0.0289 ± 0.0005
	BBP	0.0370 ± 0.0002	0.0214 ± 0.0001	-0.9807 ± 0.5485	0.0269 ± 0.0004
	SGLD	0.0446 ± 0.0008	0.0294 ± 0.0005	-2.7053 ± 0.0184	0.0121 ± 0.0001
	Ensembles	0.0345 ± -	0.0202 ± -	-3.0118 ± -	0.0279 ± -
ChemBERTa	Deterministic	0.0377 ± 0.0001	0.0208 ± 0.0000	3.8214 ± 1.1599	0.0356 ± 0.0001
	MC Dropout	0.0370 ± 0.0002	0.0204 ± 0.0001	-0.6884 ± 0.1974	0.0332 ± 0.0001
	SWAG	0.0379 ± 0.0002	0.0210 ± 0.0000	5.4548 ± 1.9400	0.0360 ± 0.0002
	BBP	0.0358 ± 0.0003	0.0206 ± 0.0001	-1.5642 ± 0.0700	0.0316 ± 0.0005
	SGLD	0.0362 ± 0.0002	0.0212 ± 0.0004	-3.1380 ± 0.0611	0.0169 ± 0.0007
	Ensembles	0.0373 ± -	0.0205 ± -	0.9180 ± -	0.0349 ± -
GROVER	Deterministic	0.0319 ± 0.0007	0.0174 ± 0.0003	0.0201 ± 1.4388	0.0386 ± 0.0020
	MC Dropout	0.0318 ± 0.0006	0.0174 ± 0.0003	-0.8810 ± 1.1057	0.0381 ± 0.0021
	SWAG	0.0323 ± 0.0004	0.0176 ± 0.0003	0.6655 ± 0.9582	0.0409 ± 0.0013
	BBP	0.0324 ± 0.0004	0.0183 ± 0.0003	-3.3037 ± 0.0260	0.0199 ± 0.0005
	SGLD	0.0340 ± 0.0007	0.0205 ± 0.0007	-3.1826 ± 0.0378	0.0157 ± 0.0006
	Ensembles	0.0315 ± -	0.0171 ± -	-1.1036 ± -	0.0383 ± -
Uni-Mol	Deterministic	0.0334 ± 0.0006	0.0172 ± 0.0000	2.7759 ± 2.9972	0.0432 ± 0.0020
	MC Dropout	0.0331 ± 0.0006	0.0171 ± 0.0001	0.8170 ± 2.8226	0.0412 ± 0.0023
	SWAG	0.0338 ± 0.0004	0.0173 ± 0.0001	3.5568 ± 2.5309	0.0440 ± 0.0013
	BBP	0.0309 ± 0.0007	0.0164 ± 0.0002	-2.4581 ± 0.4546	0.0316 ± 0.0023
	SGLD	0.0300 ± 0.0003	0.0163 ± 0.0001	-3.2399 ± 0.0908	0.0210 ± 0.0004
	Ensembles	0.0331 ± -	0.0169 ± -	-0.0541 ± -	0.0421 ± -
TorchMD-NET	Deterministic	0.0324 ± 0.0006	0.0178 ± 0.0006	1.4661 ± 2.0676	0.0439 ± 0.0045
	MC Dropout	0.0324 ± 0.0007	0.0179 ± 0.0006	0.2637 ± 1.5600	0.0419 ± 0.0042
	SWAG	0.0319 ± 0.0005	0.0177 ± 0.0005	-1.1581 ± 0.9675	0.0403 ± 0.0019
	BBP	0.0332 ± 0.0002	0.0176 ± 0.0001	4.4554 ± 0.7641	0.0469 ± 0.0007
	SGLD	0.0354 ± 0.0003	0.0217 ± 0.0001	-2.9735 ± 0.0091	0.0091 ± 0.0003
	Ensembles	0.0315 ± -	0.0173 ± -	-1.9286 ± -	0.0419 ± -
GIN	Deterministic	0.0344 ± 0.0004	0.0206 ± 0.0003	-3.3711 ± 0.0959	0.0271 ± 0.0008
	MC Dropout	0.0344 ± 0.0004	0.0206 ± 0.0003	-3.3711 ± 0.0959	0.0271 ± 0.0008
	SWAG	0.0352 ± 0.0005	0.0215 ± 0.0004	-3.3091 ± 0.0569	0.0277 ± 0.0003
	BBP	0.0352 ± 0.0003	0.0220 ± 0.0001	-3.2994 ± 0.0296	0.0238 ± 0.0007
	SGLD	0.0428 ± 0.0011	0.0290 ± 0.0009	-2.8866 ± 0.0191	0.0189 ± 0.0016
	Ensembles	0.0341 ± -	0.0205 ± -	-3.4906 ± -	0.0271 ± -

Table 23: Test results on QM9 in the format of "metric mean ± standard deviation".

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.0151 ± 0.0001	0.0101 ± 0.0001	-3.3798 ± 0.0592	0.0442 ± 0.0003
	MC Dropout	0.0148 ± 0.0001	0.0100 ± 0.0001	-3.5269 ± 0.0258	0.0433 ± 0.0003
	SWAG	0.0152 ± 0.0002	0.0102 ± 0.0001	-3.2845 ± 0.0814	0.0449 ± 0.0003
	BBP	0.0153 ± 0.0002	0.0103 ± 0.0001	-3.3470 ± 0.0280	0.0445 ± 0.0004
	SGLD	0.0196 ± 0.0003	0.0144 ± 0.0002	-3.3352 ± 0.0052	0.0070 ± 0.0003
	Ensembles	0.0143 ± -	0.0096 ± -	-3.6023 ± -	0.0436 ± -
ChemBERTa	Deterministic	0.0146 ± 0.0004	0.0092 ± 0.0002	-2.4104 ± 0.0698	0.0547 ± 0.0002
	MC Dropout	0.0141 ± 0.0002	0.0088 ± 0.0001	-3.1501 ± 0.1094	0.0513 ± 0.0005
	SWAG	0.0148 ± 0.0003	0.0092 ± 0.0002	-2.1703 ± 0.0205	0.0553 ± 0.0001
	BBP	0.0144 ± 0.0001	0.0093 ± 0.0001	-2.5932 ± 0.0665	0.0540 ± 0.0002
	SGLD	0.0153 ± 0.0001	0.0101 ± 0.0002	-3.7586 ± 0.0106	0.0338 ± 0.0007
	Ensembles	0.0140 ± -	0.0087 ± -	-2.8766 ± -	0.0543 ± -
GROVER	Deterministic	0.0115 ± 0.0000	0.0068 ± 0.0000	-0.7874 ± 0.5323	0.0621 ± 0.0008
	MC Dropout	0.0114 ± 0.0000	0.0068 ± 0.0000	-1.1002 ± 0.5003	0.0616 ± 0.0008
	SWAG	0.0116 ± 0.0000	0.0068 ± 0.0000	-0.4778 ± 0.1923	0.0625 ± 0.0002
	BBP	0.0118 ± 0.0000	0.0070 ± 0.0000	-1.8855 ± 0.1189	0.0591 ± 0.0004
	SGLD	0.0136 ± 0.0006	0.0088 ± 0.0003	-3.7855 ± 0.0411	0.0291 ± 0.0017
	Ensembles	0.0114 ± -	0.0067 ± -	-1.0286 ± -	0.0620 ± -
Uni-Mol	Deterministic	0.0096 ± 0.0001	0.0054 ± 0.0001	0.0143 ± 0.4195	0.0664 ± 0.0003
	MC Dropout	0.0096 ± 0.0001	0.0054 ± 0.0001	-0.2513 ± 0.3961	0.0661 ± 0.0003
	SWAG	0.0097 ± 0.0000	0.0054 ± 0.0000	-0.4624 ± 0.0619	0.0660 ± 0.0001
	BBP	0.0095 ± 0.0002	0.0054 ± 0.0000	-2.9595 ± 0.1250	0.0618 ± 0.0003
	SGLD	0.0095 ± 0.0001	0.0055 ± 0.0001	-4.2093 ± 0.0049	0.0459 ± 0.0002
	Ensembles	0.0095 ± -	0.0053 ± -	-0.3192 ± -	0.0663 ± -
TorchMD-NET	Deterministic	0.0087 ± 0.0001	0.0048 ± 0.0000	3.6334 ± 1.0190	0.0690 ± 0.0003
	MC Dropout	0.0087 ± 0.0001	0.0048 ± 0.0000	3.1525 ± 1.0291	0.0687 ± 0.0003
	SWAG	0.0087 ± 0.0000	0.0048 ± 0.0000	-0.7161 ± 0.1242	0.0670 ± 0.0001
	BBP	0.0083 ± 0.0002	0.0046 ± 0.0001	1.7970 ± 1.8061	0.0685 ± 0.0010
	SGLD	0.0113 ± 0.0006	0.0075 ± 0.0004	-3.6560 ± 0.0469	0.0132 ± 0.0013
	Ensembles	0.0086 ± -	0.0046 ± -	2.2619 ± -	0.0687 ± -
GIN	Deterministic	0.0132 ± 0.0001	0.0084 ± 0.0001	-3.4286 ± 0.0389	0.0497 ± 0.0003
	MC Dropout	0.0132 ± 0.0001	0.0084 ± 0.0001	-3.4286 ± 0.0389	0.0497 ± 0.0003
	SWAG	0.0136 ± 0.0000	0.0088 ± 0.0001	-3.3355 ± 0.0589	0.0495 ± 0.0005
	BBP	0.0131 ± 0.0000	0.0084 ± 0.0000	-3.3469 ± 0.0650	0.0499 ± 0.0003
	SGLD	0.0161 ± 0.0003	0.0115 ± 0.0004	-3.5925 ± 0.0198	0.0211 ± 0.0031
	Ensembles	0.0130 ± -	0.0081 ± -	-3.5207 ± -	0.0500 ± -

Table 24: Test results on the randomly split BACE dataset.

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.8850 ± 0.0220	0.1123 ± 0.0233	0.4820 ± 0.0723	0.1380 ± 0.0266
	Temperature	0.8850 ± 0.0220	0.1022 ± 0.0019	0.4475 ± 0.0624	0.1359 ± 0.0231
	Focal Loss	0.8884 ± 0.0167	0.1351 ± 0.0228	0.4623 ± 0.0239	0.1471 ± 0.0072
	MC Dropout	0.8910 ± 0.0245	0.0969 ± 0.0134	0.4328 ± 0.0594	0.1344 ± 0.0215
	SWAG	0.8898 ± 0.0220	0.1019 ± 0.0070	0.4346 ± 0.0525	0.1346 ± 0.0211
	BBP	0.8784 ± 0.0133	0.1101 ± 0.0256	0.4954 ± 0.0755	0.1414 ± 0.0156
	SGLD	0.8830 ± 0.0214	0.1119 ± 0.0294	0.4706 ± 0.0793	0.1360 ± 0.0244
	Ensembles	0.8872 ± 0.0223	0.0972 ± 0.0101	0.4572 ± 0.0712	0.1308 ± 0.0252
ChemBERTa	Deterministic	0.8835 ± 0.0044	0.1240 ± 0.0226	0.5182 ± 0.0812	0.1492 ± 0.0072
	Temperature	0.8835 ± 0.0044	0.1077 ± 0.0201	0.4566 ± 0.0274	0.1427 ± 0.0029
	Focal Loss	0.8884 ± 0.0096	0.1371 ± 0.0357	0.4490 ± 0.0328	0.1437 ± 0.0110
	MC Dropout	0.8825 ± 0.0124	0.0978 ± 0.0099	0.4544 ± 0.0082	0.1447 ± 0.0073
	SWAG	0.8868 ± 0.0059	0.1248 ± 0.0265	0.5286 ± 0.0964	0.1479 ± 0.0103
	BBP	0.8918 ± 0.0057	0.1230 ± 0.0362	0.5192 ± 0.0661	0.1468 ± 0.0117
	SGLD	0.8846 ± 0.0092	0.1362 ± 0.0046	0.5406 ± 0.0774	0.1438 ± 0.0073
	Ensembles	0.8959 ± 0.0058	0.1219 ± 0.0141	0.4715 ± 0.0441	0.1370 ± 0.0071
GROVER	Deterministic	0.9067 ± 0.0148	0.1003 ± 0.0139	0.4230 ± 0.0372	0.1254 ± 0.0114
	Temperature	0.9067 ± 0.0148	0.0884 ± 0.0130	0.4049 ± 0.0351	0.1237 ± 0.0113
	Focal Loss	0.9084 ± 0.0154	0.1600 ± 0.0069	0.4651 ± 0.0323	0.1492 ± 0.0137
	MC Dropout	0.9078 ± 0.0168	0.0802 ± 0.0069	0.4073 ± 0.0394	0.1238 ± 0.0119
	SWAG	0.9112 ± 0.0141	0.0887 ± 0.0260	0.3916 ± 0.0351	0.1203 ± 0.0112
	BBP	0.9089 ± 0.0065	0.0807 ± 0.0112	0.3946 ± 0.0178	0.1209 ± 0.0103
	SGLD	0.9066 ± 0.0145	0.0852 ± 0.0101	0.4029 ± 0.0313	0.1231 ± 0.0097
	Ensembles	0.9046 ± 0.0133	0.0915 ± 0.0054	0.4207 ± 0.0289	0.1248 ± 0.0104
Uni-Mol	Deterministic	0.9056 ± 0.0204	0.0894 ± 0.0129	0.4081 ± 0.0529	0.1257 ± 0.0191
	Temperature	0.9059 ± 0.0205	0.0854 ± 0.0082	0.4060 ± 0.0491	0.1261 ± 0.0180
	Focal Loss	0.8939 ± 0.0206	0.1634 ± 0.0148	0.5027 ± 0.0139	0.1655 ± 0.0065
	MC Dropout	0.9061 ± 0.0199	0.0864 ± 0.0086	0.4069 ± 0.0514	0.1256 ± 0.0188
	SWAG	0.9105 ± 0.0142	0.0988 ± 0.0122	0.3822 ± 0.0340	0.1179 ± 0.0131
	BBP	0.8990 ± 0.0154	0.0867 ± 0.0206	0.4303 ± 0.0513	0.1309 ± 0.0184
	SGLD	0.9080 ± 0.0104	0.0834 ± 0.0139	0.3949 ± 0.0317	0.1223 ± 0.0138
	Ensembles	0.9023 ± 0.0189	0.1020 ± 0.0140	0.4040 ± 0.0491	0.1272 ± 0.0220

Table 25: Test results on the randomly split BBBP dataset.

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.9215 ± 0.0254	0.0838 ± 0.0047	0.3722 ± 0.0275	0.0918 ± 0.0052
	Temperature	0.9215 ± 0.0254	0.0711 ± 0.0035	0.3131 ± 0.0141	0.0878 ± 0.0063
	Focal Loss	0.9180 ± 0.0142	0.1214 ± 0.0528	0.3506 ± 0.0599	0.1038 ± 0.0188
	MC Dropout	0.9226 ± 0.0244	0.0683 ± 0.0007	0.3156 ± 0.0151	0.0864 ± 0.0079
	SWAG	0.9205 ± 0.0234	0.0816 ± 0.0065	0.3358 ± 0.0247	0.0876 ± 0.0081
	BBP	0.9238 ± 0.0149	0.0846 ± 0.0188	0.3752 ± 0.0675	0.0932 ± 0.0113
	SGLD	0.9172 ± 0.0284	0.0865 ± 0.0125	0.3930 ± 0.0735	0.0905 ± 0.0072
	Ensembles	0.9226 ± 0.0088	0.0700 ± 0.0120	0.3324 ± 0.0220	0.0856 ± 0.0066
ChemBERTa	Deterministic	0.9605 ± 0.0034	0.0686 ± 0.0066	0.2494 ± 0.0089	0.0731 ± 0.0030
	Temperature	0.9605 ± 0.0034	0.0591 ± 0.0013	0.2275 ± 0.0111	0.0697 ± 0.0040
	Focal Loss	0.9587 ± 0.0027	0.0816 ± 0.0262	0.2645 ± 0.0437	0.0795 ± 0.0133
	MC Dropout	0.9607 ± 0.0027	0.0615 ± 0.0094	0.2438 ± 0.0193	0.0717 ± 0.0018
	SWAG	0.9588 ± 0.0018	0.0737 ± 0.0041	0.2856 ± 0.0097	0.0780 ± 0.0039
	BBP	0.9590 ± 0.0011	0.0702 ± 0.0125	0.2979 ± 0.0891	0.0766 ± 0.0052
	SGLD	0.9580 ± 0.0033	0.0854 ± 0.0057	0.3513 ± 0.0324	0.0826 ± 0.0055
	Ensembles	0.9604 ± 0.0018	0.0691 ± 0.0124	0.2480 ± 0.0045	0.0716 ± 0.0028
GROVER	Deterministic	0.9300 ± 0.0172	0.0551 ± 0.0138	0.2746 ± 0.0256	0.0742 ± 0.0075
	Temperature	0.9300 ± 0.0171	0.0575 ± 0.0185	0.2689 ± 0.0252	0.0740 ± 0.0074
	Focal Loss	0.9294 ± 0.0198	0.1944 ± 0.0271	0.3963 ± 0.0258	0.1153 ± 0.0102
	MC Dropout	0.9311 ± 0.0171	0.0549 ± 0.0115	0.2701 ± 0.0242	0.0742 ± 0.0072
	SWAG	0.9324 ± 0.0219	0.0601 ± 0.0112	0.2736 ± 0.0300	0.0741 ± 0.0097
	BBP	0.9196 ± 0.0210	0.0549 ± 0.0103	0.2712 ± 0.0225	0.0751 ± 0.0084
	SGLD	0.9306 ± 0.0182	0.0588 ± 0.0140	0.2750 ± 0.0304	0.0757 ± 0.0092
	Ensembles	0.9320 ± 0.0141	0.0633 ± 0.0124	0.2739 ± 0.0305	0.0738 ± 0.0092
Uni-Mol	Deterministic	0.9285 ± 0.0126	0.0785 ± 0.0103	0.3165 ± 0.0332	0.0810 ± 0.0108
	Temperature	0.9272 ± 0.0121	0.0687 ± 0.0094	0.2838 ± 0.0231	0.0790 ± 0.0099
	Focal Loss	0.9242 ± 0.0086	0.1001 ± 0.0261	0.3201 ± 0.0332	0.0936 ± 0.0109
	MC Dropout	0.9284 ± 0.0125	0.0750 ± 0.0087	0.3124 ± 0.0322	0.0807 ± 0.0107
	SWAG	0.9359 ± 0.0048	0.0797 ± 0.0086	0.3233 ± 0.0266	0.0792 ± 0.0091
	BBP	0.9175 ± 0.0121	0.0706 ± 0.0047	0.3256 ± 0.0262	0.0809 ± 0.0094
	SGLD	0.9234 ± 0.0069	0.0869 ± 0.0053	0.3671 ± 0.0262	0.0846 ± 0.0109
	Ensembles	0.9341 ± 0.0117	0.0652 ± 0.0140	0.2925 ± 0.0429	0.0766 ± 0.0107

Table 26: Test results on the randomly split ClinTox dataset.

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.8659 ± 0.0483	0.0754 ± 0.0129	0.5589 ± 0.4190	0.0672 ± 0.0104
	Temperature	0.8628 ± 0.0300	0.0719 ± 0.0087	0.4434 ± 0.1979	0.0645 ± 0.0068
	Focal Loss	0.8813 ± 0.0564	0.0652 ± 0.0054	0.1980 ± 0.0696	0.0523 ± 0.0093
	MC Dropout	0.8757 ± 0.0382	0.0616 ± 0.0093	0.2560 ± 0.0963	0.0566 ± 0.0067
	SWAG	0.8819 ± 0.0424	0.0619 ± 0.0070	0.2574 ± 0.1058	0.0572 ± 0.0066
	BBP	0.8358 ± 0.0778	0.0547 ± 0.0198	0.2679 ± 0.1254	0.0534 ± 0.0083
	SGLD	0.8745 ± 0.0506	0.0666 ± 0.0085	0.2488 ± 0.0714	0.0572 ± 0.0068
	Ensembles	0.8241 ± 0.0587	0.0646 ± 0.0088	0.4562 ± 0.3799	0.0561 ± 0.0077
ChemBERTa	Deterministic	0.9862 ± 0.0063	0.0193 ± 0.0026	0.0703 ± 0.0063	0.0168 ± 0.0015
	Temperature	0.9862 ± 0.0063	0.0181 ± 0.0045	0.0639 ± 0.0039	0.0163 ± 0.0013
	Focal Loss	0.9824 ± 0.0038	0.0493 ± 0.0183	0.0907 ± 0.0132	0.0193 ± 0.0025
	MC Dropout	0.9870 ± 0.0051	0.0162 ± 0.0029	0.0640 ± 0.0087	0.0148 ± 0.0025
	SWAG	0.9863 ± 0.0060	0.0216 ± 0.0036	0.0751 ± 0.0084	0.0179 ± 0.0019
	BBP	0.9855 ± 0.0077	0.0268 ± 0.0124	0.0687 ± 0.0095	0.0165 ± 0.0021
	SGLD	0.9851 ± 0.0053	0.0201 ± 0.0058	0.0713 ± 0.0020	0.0178 ± 0.0035
	Ensembles	0.9885 ± 0.0034	0.0242 ± 0.0108	0.0633 ± 0.0056	0.0154 ± 0.0015
GROVER	Deterministic	0.8823 ± 0.0347	0.0392 ± 0.0059	0.1550 ± 0.0289	0.0409 ± 0.0069
	Temperature	0.8818 ± 0.0341	0.0387 ± 0.0041	0.1575 ± 0.0294	0.0412 ± 0.0069
	Focal Loss	0.8554 ± 0.0542	0.1687 ± 0.0024	0.2767 ± 0.0073	0.0682 ± 0.0026
	MC Dropout	0.8840 ± 0.0332	0.0324 ± 0.0015	0.1505 ± 0.0242	0.0392 ± 0.0058
	SWAG	0.8897 ± 0.0337	0.0400 ± 0.0033	0.1490 ± 0.0267	0.0397 ± 0.0060
	BBP	0.8466 ± 0.0496	0.0684 ± 0.0059	0.1748 ± 0.0226	0.0419 ± 0.0064
	SGLD	0.8904 ± 0.0285	0.0424 ± 0.0021	0.1545 ± 0.0297	0.0414 ± 0.0086
	Ensembles	0.8854 ± 0.0371	0.0342 ± 0.0102	0.1470 ± 0.0295	0.0385 ± 0.0071
Uni-Mol	Deterministic	0.8391 ± 0.0620	0.0625 ± 0.0119	0.1949 ± 0.0291	0.0531 ± 0.0078
	Temperature	0.8433 ± 0.0611	0.0633 ± 0.0119	0.1880 ± 0.0236	0.0526 ± 0.0073
	Focal Loss	0.8416 ± 0.0705	0.0961 ± 0.0199	0.2153 ± 0.0136	0.0586 ± 0.0050
	MC Dropout	0.8394 ± 0.0616	0.0627 ± 0.0112	0.1931 ± 0.0281	0.0528 ± 0.0077
	SWAG	0.8211 ± 0.0871	0.0515 ± 0.0131	0.1974 ± 0.0588	0.0484 ± 0.0097
	BBP	0.8389 ± 0.0345	0.0593 ± 0.0234	0.1788 ± 0.0387	0.0472 ± 0.0132
	SGLD	0.8367 ± 0.0855	0.0462 ± 0.0092	0.1867 ± 0.0655	0.0475 ± 0.0110
	Ensembles	0.8387 ± 0.0463	0.0671 ± 0.0197	0.1803 ± 0.0254	0.0469 ± 0.0032

Table 27: Test results on the randomly split Tox21 dataset.

		ROC-AUC	ECE	NLL	BS
DNN-rdkit	Deterministic	0.8451 ± 0.0162	0.0335 ± 0.0012	0.1999 ± 0.0041	0.0537 ± 0.0005
	Temperature	0.8451 ± 0.0162	0.0338 ± 0.0023	0.1945 ± 0.0016	0.0534 ± 0.0005
	Focal Loss	0.8422 ± 0.0120	0.0829 ± 0.0065	0.2265 ± 0.0055	0.0587 ± 0.0010
	MC Dropout	0.8494 ± 0.0137	0.0259 ± 0.0009	0.1885 ± 0.0012	0.0510 ± 0.0002
	SWAG	0.8496 ± 0.0129	0.0256 ± 0.0009	0.1882 ± 0.0008	0.0509 ± 0.0003
	BBP	0.8216 ± 0.0033	0.0293 ± 0.0018	0.2052 ± 0.0054	0.0565 ± 0.0011
	SGLD	0.8114 ± 0.0067	0.0864 ± 0.0038	0.2568 ± 0.0084	0.0685 ± 0.0028
	Ensembles	0.8544 ± 0.0101	0.0271 ± 0.0025	0.1849 ± 0.0018	0.0506 ± 0.0005
ChemBERTa	Deterministic	0.8373 ± 0.0140	0.0345 ± 0.0046	0.1993 ± 0.0110	0.0545 ± 0.0027
	Temperature	0.8373 ± 0.0140	0.0360 ± 0.0020	0.1973 ± 0.0066	0.0543 ± 0.0022
	Focal Loss	0.8339 ± 0.0128	0.0939 ± 0.0078	0.2361 ± 0.0024	0.0605 ± 0.0007
	MC Dropout	0.8430 ± 0.0101	0.0240 ± 0.0017	0.1859 ± 0.0069	0.0506 ± 0.0019
	SWAG	0.8369 ± 0.0149	0.0370 ± 0.0025	0.2027 ± 0.0110	0.0553 ± 0.0022
	BBP	0.8301 ± 0.0131	0.0403 ± 0.0038	0.2052 ± 0.0077	0.0575 ± 0.0025
	SGLD	0.8314 ± 0.0127	0.0376 ± 0.0018	0.2002 ± 0.0102	0.0552 ± 0.0031
	Ensembles	0.8509 ± 0.0082	0.0304 ± 0.0027	0.1890 ± 0.0100	0.0519 ± 0.0025
GROVER	Deterministic	0.8754 ± 0.0155	0.0259 ± 0.0018	0.1690 ± 0.0067	0.0457 ± 0.0015
	Temperature	0.8754 ± 0.0154	0.0281 ± 0.0030	0.1700 ± 0.0060	0.0458 ± 0.0015
	Focal Loss	0.8785 ± 0.0099	0.1337 ± 0.0012	0.2623 ± 0.0046	0.0636 ± 0.0017
	MC Dropout	0.8760 ± 0.0149	0.0279 ± 0.0015	0.1697 ± 0.0062	0.0457 ± 0.0015
	SWAG	0.8776 ± 0.0135	0.0247 ± 0.0014	0.1674 ± 0.0059	0.0452 ± 0.0014
	BBP	0.8667 ± 0.0135	0.0439 ± 0.0038	0.1850 ± 0.0049	0.0483 ± 0.0016
	SGLD	0.8724 ± 0.0123	0.0390 ± 0.0023	0.1811 ± 0.0061	0.0473 ± 0.0019
	Ensembles	0.8775 ± 0.0137	0.0244 ± 0.0010	0.1680 ± 0.0055	0.0453 ± 0.0013
Uni-Mol	Deterministic	0.8666 ± 0.0138	0.0302 ± 0.0016	0.1773 ± 0.0052	0.0466 ± 0.0012
	Temperature	0.8661 ± 0.0136	0.0295 ± 0.0008	0.1727 ± 0.0037	0.0464 ± 0.0010
	Focal Loss	0.8631 ± 0.0071	0.0824 ± 0.0093	0.2154 ± 0.0097	0.0543 ± 0.0028
	MC Dropout	0.8667 ± 0.0139	0.0292 ± 0.0021	0.1758 ± 0.0050	0.0465 ± 0.0012
	SWAG	0.8666 ± 0.0170	0.0308 ± 0.0018	0.1779 ± 0.0042	0.0452 ± 0.0010
	BBP	0.8731 ± 0.0107	0.0259 ± 0.0031	0.1731 ± 0.0046	0.0469 ± 0.0013
	SGLD	0.8663 ± 0.0149	0.0333 ± 0.0027	0.1807 ± 0.0058	0.0470 ± 0.0011
	Ensembles	0.8778 ± 0.0090	0.0262 ± 0.0024	0.1660 ± 0.0043	0.0440 ± 0.0008

Table 28: Test results on the randomly split ESOL dataset.

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.6530 ± 0.0615	0.4788 ± 0.0279	0.1067 ± 0.0716	0.0419 ± 0.0034
	MC Dropout	0.6597 ± 0.0666	0.4820 ± 0.0351	0.1212 ± 0.0925	0.0436 ± 0.0037
	SWAG	0.6542 ± 0.0766	0.4752 ± 0.0392	0.0977 ± 0.0884	0.0435 ± 0.0034
	BBP	0.7060 ± 0.0604	0.5294 ± 0.0458	0.2153 ± 0.1182	0.0263 ± 0.0050
	SGLD	0.7033 ± 0.0648	0.5235 ± 0.0273	0.3825 ± 0.0019	0.0163 ± 0.0019
	Ensembles	0.6455 ± 0.0518	0.4735 ± 0.0310	0.0791 ± 0.0905	0.0405 ± 0.0031
ChemBERTa	Deterministic	0.7760 ± 0.0921	0.5920 ± 0.0465	0.3768 ± 0.1070	0.0464 ± 0.0042
	MC Dropout	0.6919 ± 0.0947	0.5192 ± 0.0619	0.1908 ± 0.1299	0.0448 ± 0.0043
	SWAG	0.7704 ± 0.0784	0.5921 ± 0.0350	0.4094 ± 0.1176	0.0475 ± 0.0046
	BBP	0.7648 ± 0.0328	0.5923 ± 0.0176	0.2208 ± 0.0289	0.0326 ± 0.0022
	SGLD	0.7282 ± 0.0764	0.5539 ± 0.0290	0.2455 ± 0.0835	0.0456 ± 0.0030
	Ensembles	0.7318 ± 0.0778	0.5641 ± 0.0416	0.3073 ± 0.1107	0.0470 ± 0.0032
GROVER	Deterministic	0.6137 ± 0.0392	0.4730 ± 0.0173	0.0299 ± 0.0935	0.0377 ± 0.0063
	MC Dropout	0.5997 ± 0.0256	0.4621 ± 0.0110	0.0124 ± 0.0890	0.0376 ± 0.0065
	SWAG	0.5895 ± 0.0251	0.4476 ± 0.0102	-0.0342 ± 0.0624	0.0419 ± 0.0035
	BBP	0.6449 ± 0.0517	0.4892 ± 0.0286	0.3780 ± 0.0474	0.0152 ± 0.0018
	SGLD	0.5888 ± 0.0379	0.4542 ± 0.0252	0.0748 ± 0.0683	0.0299 ± 0.0010
	Ensembles	0.5847 ± 0.0326	0.4526 ± 0.0210	-0.0377 ± 0.0725	0.0384 ± 0.0031
Uni-Mol	Deterministic	0.6253 ± 0.0500	0.4691 ± 0.0291	0.0024 ± 0.0693	0.0388 ± 0.0008
	MC Dropout	0.6224 ± 0.0501	0.4640 ± 0.0299	0.0058 ± 0.0647	0.0375 ± 0.0007
	SWAG	0.5642 ± 0.0564	0.4030 ± 0.0342	-0.0603 ± 0.1092	0.0361 ± 0.0028
	BBP	0.6762 ± 0.0766	0.4884 ± 0.0544	0.2844 ± 0.0276	0.0235 ± 0.0024
	SGLD	0.6063 ± 0.0281	0.4388 ± 0.0098	0.1947 ± 0.2311	0.0270 ± 0.0096
	Ensembles	0.5930 ± 0.0374	0.4311 ± 0.0129	-0.0273 ± 0.0238	0.0375 ± 0.0015

Table 29: Test results on the randomly split FreeSolv dataset.

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	1.1867 ± 0.0369	0.8582 ± 0.0521	0.7262 ± 0.0311	0.0252 ± 0.0038
	MC Dropout	1.1893 ± 0.0716	0.8576 ± 0.0283	0.7085 ± 0.0277	0.0263 ± 0.0038
	SWAG	1.1183 ± 0.0960	0.7817 ± 0.0602	0.6026 ± 0.0144	0.0305 ± 0.0047
	BBP	1.8405 ± 0.2306	1.2685 ± 0.0533	1.2661 ± 0.0458	0.0060 ± 0.0013
	SGLD	1.5395 ± 0.1830	1.0855 ± 0.0925	1.2214 ± 0.0395	0.0045 ± 0.0020
	Ensembles	1.1690 ± 0.0630	0.8500 ± 0.0239	0.7444 ± 0.0502	0.0240 ± 0.0042
ChemBERTa	Deterministic	1.4125 ± 0.1890	1.0323 ± 0.1206	0.8274 ± 0.1999	0.0387 ± 0.0059
	MC Dropout	1.3675 ± 0.2456	1.0278 ± 0.1732	0.6970 ± 0.1730	0.0357 ± 0.0067
	SWAG	1.3789 ± 0.2128	1.0024 ± 0.1570	0.8477 ± 0.2851	0.0408 ± 0.0041
	BBP	1.4355 ± 0.2128	1.0643 ± 0.1312	0.8143 ± 0.0637	0.0270 ± 0.0032
	SGLD	1.3864 ± 0.2747	0.9852 ± 0.1613	0.6654 ± 0.1475	0.0361 ± 0.0023
	Ensembles	1.2276 ± 0.2395	0.8745 ± 0.1727	0.5995 ± 0.1546	0.0399 ± 0.0013
GROVER	Deterministic	1.3823 ± 0.1108	1.0223 ± 0.0942	0.9105 ± 0.0684	0.0188 ± 0.0018
	MC Dropout	1.3570 ± 0.1082	1.0093 ± 0.0940	0.9247 ± 0.0680	0.0177 ± 0.0017
	SWAG	1.3181 ± 0.0898	0.9858 ± 0.0727	0.8386 ± 0.0352	0.0218 ± 0.0042
	BBP	1.4832 ± 0.1322	1.1451 ± 0.0893	1.1459 ± 0.0166	0.0060 ± 0.0019
	SGLD	1.4965 ± 0.2643	1.1682 ± 0.2289	1.1154 ± 0.1317	0.0072 ± 0.0016
	Ensembles	1.2826 ± 0.1700	0.9424 ± 0.1316	0.9476 ± 0.0558	0.0140 ± 0.0020
Uni-Mol	Deterministic	1.0429 ± 0.0170	0.7345 ± 0.0308	0.5741 ± 0.0637	0.0322 ± 0.0054
	MC Dropout	1.0381 ± 0.0249	0.7280 ± 0.0394	0.5872 ± 0.0636	0.0310 ± 0.0055
	SWAG	0.9959 ± 0.0902	0.6842 ± 0.1248	0.6050 ± 0.0705	0.0271 ± 0.0051
	BBP	1.1541 ± 0.1779	0.8681 ± 0.1550	0.9795 ± 0.0621	0.0154 ± 0.0027
	SGLD	1.0486 ± 0.1898	0.7543 ± 0.1538	1.0065 ± 0.1283	0.0126 ± 0.0016
	Ensembles	1.0009 ± 0.0416	0.7201 ± 0.0378	0.6766 ± 0.0365	0.0270 ± 0.0045

Table 30: Test results on the randomly split Lipophilicity dataset.

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	0.6790 ± 0.0415	0.5007 ± 0.0253	0.6158 ± 0.1691	0.0365 ± 0.0049
	MC Dropout	0.6732 ± 0.0398	0.4971 ± 0.0286	0.5558 ± 0.1208	0.0366 ± 0.0048
	SWAG	0.6699 ± 0.0441	0.4926 ± 0.0277	0.6208 ± 0.1477	0.0370 ± 0.0052
	BBP	0.6833 ± 0.0444	0.5092 ± 0.0222	0.2681 ± 0.2026	0.0281 ± 0.0040
	SGLD	0.6819 ± 0.0427	0.5090 ± 0.0222	0.1218 ± 0.0554	0.0134 ± 0.0010
	Ensembles	0.6399 ± 0.0402	0.4690 ± 0.0199	0.3242 ± 0.1517	0.0349 ± 0.0004
ChemBERTa	Deterministic	0.7033 ± 0.0521	0.5282 ± 0.0234	1.4475 ± 1.1982	0.0418 ± 0.0064
	MC Dropout	0.6583 ± 0.0330	0.4882 ± 0.0095	0.8234 ± 0.7527	0.0380 ± 0.0056
	SWAG	0.7114 ± 0.0502	0.5363 ± 0.0196	1.7413 ± 1.2705	0.0441 ± 0.0059
	BBP	0.7103 ± 0.0508	0.5377 ± 0.0212	0.5764 ± 0.3161	0.0368 ± 0.0043
	SGLD	0.6988 ± 0.0540	0.5271 ± 0.0249	0.7978 ± 0.2380	0.0414 ± 0.0016
	Ensembles	0.6763 ± 0.0534	0.5044 ± 0.0232	1.4618 ± 0.8428	0.0451 ± 0.0033
GROVER	Deterministic	0.5717 ± 0.0320	0.3957 ± 0.0144	0.6373 ± 0.0281	0.0446 ± 0.0016
	MC Dropout	0.5633 ± 0.0349	0.3900 ± 0.0163	0.5493 ± 0.0500	0.0441 ± 0.0018
	SWAG	0.5599 ± 0.0340	0.3851 ± 0.0117	0.5192 ± 0.1369	0.0440 ± 0.0002
	BBP	0.5638 ± 0.0314	0.4021 ± 0.0088	-0.0783 ± 0.0532	0.0232 ± 0.0010
	SGLD	0.5597 ± 0.0409	0.3958 ± 0.0171	-0.0915 ± 0.0856	0.0271 ± 0.0006
	Ensembles	0.5539 ± 0.0314	0.3833 ± 0.0134	0.4942 ± 0.1247	0.0447 ± 0.0010
Uni-Mol	Deterministic	0.5387 ± 0.0439	0.3824 ± 0.0203	-0.0690 ± 0.1128	0.0362 ± 0.0024
	MC Dropout	0.5367 ± 0.0437	0.3812 ± 0.0204	-0.0859 ± 0.1088	0.0355 ± 0.0024
	SWAG	0.5077 ± 0.0446	0.3598 ± 0.0178	-0.1168 ± 0.1292	0.0369 ± 0.0018
	BBP	0.5632 ± 0.0509	0.4139 ± 0.0207	-0.0682 ± 0.0724	0.0248 ± 0.0039
	SGLD	0.5238 ± 0.0357	0.3718 ± 0.0160	-0.1525 ± 0.0618	0.0265 ± 0.0011
	Ensembles	0.5255 ± 0.0434	0.3706 ± 0.0187	-0.1311 ± 0.1122	0.0347 ± 0.0010

Table 31: Test results on the randomly split QM7 dataset.

		RMSE	MAE	NLL	CE
DNN-rdkit	Deterministic	107.9546 ± 8.1501	57.5611 ± 4.8507	5.5231 ± 0.0677	0.0449 ± 0.0026
	MC Dropout	106.5277 ± 8.4035	56.8402 ± 4.4571	4.9801 ± 0.0805	0.0434 ± 0.0025
	SWAG	105.3930 ± 8.3626	56.1532 ± 3.9650	4.9387 ± 0.1316	0.0429 ± 0.0018
	BBP	111.6410 ± 8.9230	65.9097 ± 3.7377	5.1715 ± 0.2092	0.0388 ± 0.0020
	SGLD	111.1149 ± 7.8189	75.5816 ± 4.4419	5.2277 ± 0.0341	0.0127 ± 0.0004
	Ensembles	104.0890 ± 9.9660	55.6554 ± 4.1624	5.0123 ± 0.1890	0.0438 ± 0.0020
ChemBERTa	Deterministic	97.4000 ± 8.8093	47.3906 ± 3.6007	4.5928 ± 0.1364	0.0402 ± 0.0021
	MC Dropout	97.1763 ± 9.0430	43.4817 ± 4.8435	4.3224 ± 0.3558	0.0458 ± 0.0017
	SWAG	97.5653 ± 8.4876	49.3942 ± 4.0995	4.7928 ± 0.1495	0.0397 ± 0.0019
	BBP	100.8003 ± 8.9576	53.0419 ± 4.4522	4.6938 ± 0.2195	0.0347 ± 0.0027
	SGLD	99.6487 ± 7.9623	54.8976 ± 2.4214	4.6314 ± 0.1117	0.0308 ± 0.0015
	Ensembles	97.5500 ± 8.6923	47.2016 ± 3.7146	4.5436 ± 0.2688	0.0392 ± 0.0020
GROVER	Deterministic	98.7533 ± 9.8780	42.7787 ± 4.8753	4.7424 ± 0.3523	0.0512 ± 0.0003
	MC Dropout	96.6699 ± 9.6047	42.6411 ± 4.6449	4.7324 ± 0.3130	0.0511 ± 0.0003
	SWAG	97.1293 ± 10.0641	43.3197 ± 5.1696	4.6304 ± 0.2678	0.0495 ± 0.0008
	BBP	99.0196 ± 8.1011	56.8246 ± 3.9466	4.9841 ± 0.0630	0.0223 ± 0.0003
	SGLD	101.0303 ± 6.6971	57.6752 ± 3.7755	4.9718 ± 0.0556	0.0237 ± 0.0009
	Ensembles	98.0631 ± 9.7788	42.8305 ± 4.6214	4.6874 ± 0.3069	0.0498 ± 0.0006
Uni-Mol	Deterministic	71.4475 ± 2.3163	36.7968 ± 2.3826	4.1470 ± 0.1635	0.0411 ± 0.0032
	MC Dropout	70.8934 ± 2.2622	35.5291 ± 2.4115	4.1425 ± 0.1709	0.0406 ± 0.0032
	SWAG	68.6536 ± 1.3320	29.1093 ± 2.1596	3.9460 ± 0.0134	0.0433 ± 0.0011
	BBP	73.0523 ± 1.1094	39.2194 ± 1.8792	4.5952 ± 0.0996	0.0297 ± 0.0042
	SGLD	71.2770 ± 1.6947	32.4820 ± 3.3687	4.4407 ± 0.0538	0.0321 ± 0.0017
	Ensembles	69.6610 ± 0.9264	33.8478 ± 0.0577	4.0436 ± 0.0870	0.0429 ± 0.0020

Table 32: Test results with frozen backbone model weights on BACE.

		ROC-AUC	ECE	NLL	BS
ChemBERTa	Deterministic	0.6289 ± 0.0090	0.2096 ± 0.0037	0.7570 ± 0.0073	0.2751 ± 0.0028
	Temperature	0.6289 ± 0.0090	0.1554 ± 0.0103	0.7134 ± 0.0032	0.2587 ± 0.0014
	Focal Loss	0.6474 ± 0.0051	0.1964 ± 0.0035	0.7450 ± 0.0023	0.2741 ± 0.0011
	MC Dropout	0.6515 ± 0.0164	0.1931 ± 0.0045	0.7393 ± 0.0029	0.2693 ± 0.0013
	SWAG	0.6305 ± 0.0081	0.2045 ± 0.0032	0.7559 ± 0.0077	0.2746 ± 0.0030
	BBP	0.5454 ± 0.0619	0.0998 ± 0.0235	0.7080 ± 0.0087	0.2573 ± 0.0042
	SGLD	0.5217 ± 0.0512	0.0531 ± 0.0114	0.6941 ± 0.0022	0.2505 ± 0.0011
	Ensembles	0.6308 ± -	0.1415 ± -	0.7081 ± -	0.2569 ± -
GROVER	Deterministic	0.7633 ± 0.0116	0.2083 ± 0.0100	0.6934 ± 0.0070	0.2514 ± 0.0030
	Temperature	0.7635 ± 0.0117	0.1915 ± 0.0109	0.6826 ± 0.0067	0.2461 ± 0.0029
	Focal Loss	0.7221 ± 0.0343	0.2125 ± 0.0066	0.7358 ± 0.0048	0.2710 ± 0.0020
	MC Dropout	0.7672 ± 0.0081	0.2221 ± 0.0111	0.7037 ± 0.0050	0.2561 ± 0.0020
	SWAG	0.7691 ± 0.0096	0.2232 ± 0.0131	0.7027 ± 0.0036	0.2559 ± 0.0013
	BBP	0.7425 ± 0.0089	0.1946 ± 0.0135	0.7046 ± 0.0038	0.2565 ± 0.0015
	SGLD	0.5485 ± 0.0612	0.0944 ± 0.0227	0.7007 ± 0.0135	0.2536 ± 0.0067
	Ensembles	0.7826 ± -	0.2278 ± -	0.6807 ± -	0.2462 ± -
Uni-Mol	Deterministic	0.3713 ± 0.1533	0.1999 ± 0.0357	0.7544 ± 0.0216	0.2795 ± 0.0101
	Temperature	0.3712 ± 0.1534	0.1811 ± 0.0340	0.7317 ± 0.0149	0.2689 ± 0.0072
	Focal Loss	0.3512 ± 0.1450	0.2720 ± 0.0414	0.7977 ± 0.0301	0.2985 ± 0.0131
	MC Dropout	0.3698 ± 0.1569	0.2052 ± 0.0513	0.7512 ± 0.0203	0.2780 ± 0.0095
	SWAG	0.4297 ± 0.1408	0.1513 ± 0.0237	0.7330 ± 0.0089	0.2695 ± 0.0042
	BBP	0.5958 ± 0.0671	0.1614 ± 0.0191	0.7361 ± 0.0096	0.2709 ± 0.0045
	SGLD	0.7189 ± 0.0291	0.1414 ± 0.0218	0.6737 ± 0.0091	0.2405 ± 0.0045
	Ensembles	0.6799 ± -	0.1543 ± -	0.7370 ± -	0.2714 ± -

Table 33: Test results with frozen backbone model weights on BBBP.

		ROC-AUC	ECE	NLL	BS
ChemBERTa	Deterministic	0.6651 ± 0.0075	0.1025 ± 0.0103	0.6534 ± 0.0072	0.2308 ± 0.0033
	Temperature	0.6651 ± 0.0075	0.1681 ± 0.0189	0.7366 ± 0.0315	0.2518 ± 0.0073
	Focal Loss	0.6600 ± 0.0082	0.1342 ± 0.0038	0.6782 ± 0.0017	0.2430 ± 0.0009
	MC Dropout	0.6592 ± 0.0084	0.1358 ± 0.0159	0.6813 ± 0.0058	0.2428 ± 0.0030
	SWAG	0.6650 ± 0.0079	0.1068 ± 0.0112	0.6537 ± 0.0074	0.2310 ± 0.0033
	BBP	0.6280 ± 0.0074	0.1684 ± 0.0089	0.7250 ± 0.0082	0.2645 ± 0.0035
	SGLD	0.4822 ± 0.0247	0.0346 ± 0.0091	0.6945 ± 0.0020	0.2507 ± 0.0010
	Ensembles	0.6545 ± -	0.1059 ± -	0.6539 ± -	0.2320 ± -
GROVER	Deterministic	0.6088 ± 0.0124	0.3048 ± 0.0071	0.9424 ± 0.0027	0.3303 ± 0.0024
	Temperature	0.6083 ± 0.0122	0.2738 ± 0.0073	0.8674 ± 0.0022	0.3119 ± 0.0019
	Focal Loss	0.5979 ± 0.0167	0.0867 ± 0.0234	0.6872 ± 0.0071	0.2469 ± 0.0034
	MC Dropout	0.6089 ± 0.0135	0.3073 ± 0.0114	0.9372 ± 0.0028	0.3297 ± 0.0021
	SWAG	0.6179 ± 0.0112	0.3087 ± 0.0038	0.9424 ± 0.0042	0.3300 ± 0.0022
	BBP	0.6096 ± 0.0069	0.3023 ± 0.0065	0.9211 ± 0.0038	0.3271 ± 0.0012
	SGLD	0.4628 ± 0.0682	0.0685 ± 0.0306	0.6979 ± 0.0108	0.2523 ± 0.0053
	Ensembles	0.6107 ± -	0.2982 ± -	0.9338 ± -	0.3285 ± -
Uni-Mol	Deterministic	0.5325 ± 0.0364	0.3534 ± 0.0074	1.0741 ± 0.0298	0.3726 ± 0.0064
	Temperature	0.5326 ± 0.0366	0.2867 ± 0.0074	0.9013 ± 0.0181	0.3297 ± 0.0056
	Focal Loss	0.5405 ± 0.0458	0.0522 ± 0.0042	0.6912 ± 0.0075	0.2490 ± 0.0037
	MC Dropout	0.5336 ± 0.0397	0.3449 ± 0.0077	1.0457 ± 0.0290	0.3666 ± 0.0066
	SWAG	0.5592 ± 0.0143	0.3008 ± 0.0042	0.9277 ± 0.0109	0.3374 ± 0.0033
	BBP	0.4626 ± 0.0239	0.2631 ± 0.0447	0.8784 ± 0.0809	0.3209 ± 0.0243
	SGLD	0.4590 ± 0.0456	0.1206 ± 0.0298	0.7238 ± 0.0136	0.2646 ± 0.0059
	Ensembles	0.5875 ± -	0.3370 ± -	1.0186 ± -	0.3614 ± -

Table 34: Test results with frozen backbone model weights on ClinTox.

		ROC-AUC	ECE	NLL	BS
ChemBERTa	Deterministic	0.6344 ± 0.2837	0.2913 ± 0.1651	0.4713 ± 0.2530	0.1618 ± 0.0952
	Temperature	0.6344 ± 0.2837	0.2389 ± 0.1655	0.4157 ± 0.2533	0.1398 ± 0.0964
	Focal Loss	0.6369 ± 0.2860	0.3267 ± 0.1183	0.5099 ± 0.2000	0.1688 ± 0.0859
	MC Dropout	0.6689 ± 0.2643	0.2886 ± 0.1675	0.4706 ± 0.2491	0.1612 ± 0.0934
	SWAG	0.6911 ± 0.2300	0.2245 ± 0.1211	0.3728 ± 0.1831	0.1142 ± 0.0607
	BBP	0.7242 ± 0.0410	0.2659 ± 0.1039	0.4515 ± 0.1399	0.1379 ± 0.0638
	SGLD	0.5489 ± 0.1696	0.4339 ± 0.0032	0.6898 ± 0.0048	0.2483 ± 0.0024
	Ensembles	0.9734 ± -	0.2885 ± -	0.4085 ± -	0.1139 ± -
GROVER	Deterministic	0.6437 ± 0.0235	0.0304 ± 0.0051	0.2357 ± 0.0041	0.0599 ± 0.0008
	Temperature	0.6426 ± 0.0240	0.0179 ± 0.0032	0.2337 ± 0.0050	0.0594 ± 0.0007
	Focal Loss	0.6328 ± 0.0220	0.2116 ± 0.0054	0.3897 ± 0.0064	0.1109 ± 0.0027
	MC Dropout	0.6534 ± 0.0201	0.0222 ± 0.0041	0.2329 ± 0.0029	0.0594 ± 0.0005
	SWAG	0.6838 ± 0.0124	0.0268 ± 0.0064	0.2306 ± 0.0031	0.0589 ± 0.0007
	BBP	0.6019 ± 0.0346	0.0339 ± 0.0028	0.2422 ± 0.0032	0.0609 ± 0.0005
	SGLD	0.4601 ± 0.0878	0.4365 ± 0.0355	0.7050 ± 0.0650	0.2558 ± 0.0321
	Ensembles	0.7039 ± -	0.0348 ± -	0.2311 ± -	0.0591 ± -
Uni-Mol	Deterministic	0.5750 ± 0.0185	0.1620 ± 0.0598	0.3423 ± 0.0569	0.0916 ± 0.0209
	Temperature	0.5752 ± 0.0182	0.0713 ± 0.0499	0.2694 ± 0.0319	0.0678 ± 0.0086
	Focal Loss	0.5696 ± 0.0199	0.2605 ± 0.0328	0.4397 ± 0.0359	0.1290 ± 0.0158
	MC Dropout	0.5797 ± 0.0259	0.1704 ± 0.0595	0.3499 ± 0.0577	0.0944 ± 0.0216
	SWAG	0.5887 ± 0.0228	0.1125 ± 0.0324	0.2948 ± 0.0250	0.0741 ± 0.0075
	BBP	0.5899 ± 0.0112	0.2754 ± 0.0510	0.4608 ± 0.0683	0.1391 ± 0.0313
	SGLD	0.5257 ± 0.0576	0.4285 ± 0.0442	0.7081 ± 0.0984	0.2556 ± 0.0463
	Ensembles	0.5551 ± -	0.1449 ± -	0.3199 ± -	0.0811 ± -

Table 35: Test results with frozen backbone model weights on Tox21.

		ROC-AUC	ECE	NLL	BS
ChemBERTa	Deterministic	0.6615 ± 0.0039	0.0402 ± 0.0008	0.2934 ± 0.0004	0.0826 ± 0.0002
	Temperature	0.6615 ± 0.0039	0.0307 ± 0.0015	0.2894 ± 0.0008	0.0817 ± 0.0001
	Focal Loss	0.6646 ± 0.0026	0.1183 ± 0.0025	0.3428 ± 0.0021	0.0949 ± 0.0007
	MC Dropout	0.6729 ± 0.0048	0.0315 ± 0.0016	0.2859 ± 0.0007	0.0807 ± 0.0002
	SWAG	0.6619 ± 0.0038	0.0397 ± 0.0012	0.2934 ± 0.0005	0.0826 ± 0.0002
	BBP	0.5426 ± 0.0080	0.1150 ± 0.0079	0.3629 ± 0.0059	0.1009 ± 0.0019
	SGLD	0.5062 ± 0.0212	0.3973 ± 0.0005	0.6903 ± 0.0008	0.2486 ± 0.0004
	Ensembles	0.6663 ± -	0.0383 ± -	0.2916 ± -	0.0822 ± -
GROVER	Deterministic	0.6897 ± 0.0102	0.0361 ± 0.0002	0.2838 ± 0.0011	0.0807 ± 0.0003
	Temperature	0.6897 ± 0.0102	0.0223 ± 0.0017	0.2789 ± 0.0011	0.0799 ± 0.0003
	Focal Loss	0.6632 ± 0.0139	0.1326 ± 0.0023	0.3593 ± 0.0005	0.0998 ± 0.0002
	MC Dropout	0.6874 ± 0.0093	0.0343 ± 0.0004	0.2842 ± 0.0009	0.0810 ± 0.0002
	SWAG	0.6923 ± 0.0075	0.0344 ± 0.0002	0.2830 ± 0.0008	0.0807 ± 0.0002
	BBP	0.6496 ± 0.0040	0.0364 ± 0.0005	0.2951 ± 0.0003	0.0836 ± 0.0001
	SGLD	0.5072 ± 0.0266	0.3898 ± 0.0018	0.6806 ± 0.0024	0.2438 ± 0.0012
	Ensembles	0.6911 ± -	0.0365 ± -	0.2839 ± -	0.0807 ± -
Uni-Mol	Deterministic	0.5702 ± 0.0065	0.0556 ± 0.0008	0.3352 ± 0.0009	0.0905 ± 0.0002
	Temperature	0.5702 ± 0.0066	0.0153 ± 0.0012	0.3036 ± 0.0001	0.0862 ± 0.0000
	Focal Loss	0.5620 ± 0.0067	0.0472 ± 0.0030	0.3183 ± 0.0020	0.0888 ± 0.0004
	MC Dropout	0.5696 ± 0.0056	0.0519 ± 0.0004	0.3297 ± 0.0007	0.0900 ± 0.0002
	SWAG	0.5936 ± 0.0056	0.0293 ± 0.0011	0.3070 ± 0.0005	0.0868 ± 0.0001
	BBP	0.5505 ± 0.0184	0.0472 ± 0.0021	0.3249 ± 0.0021	0.0896 ± 0.0002
	SGLD	0.4927 ± 0.0089	0.4022 ± 0.0118	0.7135 ± 0.0216	0.2599 ± 0.0104
	Ensembles	0.5866 ± -	0.0546 ± -	0.3335 ± -	0.0901 ± -

Table 36: Test results with frozen backbone model weights on ESOL.

		RMSE	MAE	NLL	CE
ChemBERTa	Deterministic	1.9543 ± 0.0830	1.4348 ± 0.0895	1.2260 ± 0.0454	0.0090 ± 0.0010
	MC Dropout	1.8954 ± 0.1132	1.4051 ± 0.1047	1.1497 ± 0.0815	0.0080 ± 0.0006
	SWAG	1.9708 ± 0.0865	1.4372 ± 0.0905	1.2561 ± 0.0495	0.0105 ± 0.0011
	BBP	2.2450 ± 0.0440	1.7152 ± 0.0346	1.3560 ± 0.0285	0.0155 ± 0.0026
	SGLD	2.3491 ± 0.0542	1.7899 ± 0.0472	1.4684 ± 0.0373	0.0180 ± 0.0055
	Ensembles	1.8844 ± -	1.3699 ± -	1.1914 ± -	0.0082 ± -
GROVER	Deterministic	1.8533 ± 0.1590	1.4171 ± 0.1281	1.1335 ± 0.0833	0.0097 ± 0.0024
	MC Dropout	1.8629 ± 0.1562	1.4131 ± 0.1234	1.1444 ± 0.0836	0.0110 ± 0.0031
	SWAG	1.8059 ± 0.1521	1.3651 ± 0.1102	1.0930 ± 0.0857	0.0120 ± 0.0029
	BBP	1.9032 ± 0.1318	1.4502 ± 0.1186	1.3243 ± 0.0225	0.0088 ± 0.0014
	SGLD	2.4699 ± 0.1371	1.8858 ± 0.1161	1.3697 ± 0.0510	0.0285 ± 0.0050
	Ensembles	1.8185 ± -	1.3161 ± -	1.1063 ± -	0.0098 ± -
Uni-Mol	Deterministic	2.2126 ± 0.0258	1.7134 ± 0.0482	1.3168 ± 0.0112	0.0078 ± 0.0054
	MC Dropout	2.2155 ± 0.0223	1.7118 ± 0.0426	1.3169 ± 0.0106	0.0078 ± 0.0053
	SWAG	2.1809 ± 0.0258	1.6573 ± 0.0294	1.2883 ± 0.0210	0.0085 ± 0.0035
	BBP	2.1964 ± 0.1226	1.7223 ± 0.0996	1.3008 ± 0.0625	0.0053 ± 0.0042
	SGLD	2.3596 ± 0.0599	1.8831 ± 0.0734	1.4158 ± 0.0808	0.0139 ± 0.0037
	Ensembles	2.1476 ± -	1.6524 ± -	1.2669 ± -	0.0053 ± -

Table 37: Test results with frozen backbone model weights on FreeSolv.

		RMSE	MAE	NLL	CE
ChemBERTa	Deterministic	3.7564 ± 0.0520	2.7933 ± 0.0770	2.0303 ± 0.0828	0.0378 ± 0.0078
	MC Dropout	3.7861 ± 0.0542	2.8493 ± 0.0801	2.0393 ± 0.0773	0.0429 ± 0.0065
	SWAG	3.7629 ± 0.0297	2.8017 ± 0.0563	2.0520 ± 0.0679	0.0395 ± 0.0062
	BBP	4.5822 ± 0.0444	3.4921 ± 0.0636	2.3843 ± 0.0507	0.0669 ± 0.0037
	SGLD	4.3996 ± 0.1248	3.3132 ± 0.1317	2.3018 ± 0.1113	0.0546 ± 0.0088
	Ensembles	3.7094 ± -	2.7309 ± -	2.0012 ± -	0.0370 ± -
GROVER	Deterministic	4.2407 ± 0.3174	3.1254 ± 0.4023	2.0701 ± 0.1023	0.0233 ± 0.0141
	MC Dropout	4.2045 ± 0.3199	3.1098 ± 0.3737	2.0422 ± 0.0929	0.0199 ± 0.0130
	SWAG	4.1854 ± 0.2899	3.1039 ± 0.3888	2.0604 ± 0.1230	0.0251 ± 0.0127
	BBP	4.4898 ± 0.2067	3.3533 ± 0.2186	2.0452 ± 0.0530	0.0378 ± 0.0136
	SGLD	4.6731 ± 0.5142	3.5858 ± 0.5165	2.1826 ± 0.0714	0.0529 ± 0.0178
	Ensembles	4.0551 ± -	3.0166 ± -	1.9352 ± -	0.0245 ± -
Uni-Mol	Deterministic	3.9034 ± 0.4092	2.8791 ± 0.3542	1.8779 ± 0.0965	0.0207 ± 0.0150
	MC Dropout	3.9103 ± 0.4142	2.8836 ± 0.3612	1.8813 ± 0.0990	0.0209 ± 0.0151
	SWAG	4.1746 ± 0.2749	3.1272 ± 0.2718	1.9582 ± 0.0786	0.0360 ± 0.0128
	BBP	4.0001 ± 0.2701	3.0110 ± 0.2507	1.9486 ± 0.1219	0.0276 ± 0.0196
	SGLD	4.1575 ± 0.0222	3.1164 ± 0.0627	2.0574 ± 0.1045	0.0321 ± 0.0066
	Ensembles	4.0416 ± -	2.9823 ± -	1.9718 ± -	0.0287 ± -

Table 38: Test results with frozen backbone model weights on Lipophilicity.

		RMSE	MAE	NLL	CE
ChemBERTa	Deterministic	0.9958 ± 0.0033	0.8083 ± 0.0047	0.4837 ± 0.0039	0.0016 ± 0.0001
	MC Dropout	0.9985 ± 0.0064	0.8185 ± 0.0061	0.4885 ± 0.0063	0.0041 ± 0.0001
	SWAG	0.9948 ± 0.0029	0.8070 ± 0.0043	0.4823 ± 0.0031	0.0016 ± 0.0001
	BBP	1.0477 ± 0.0017	0.8578 ± 0.0018	0.5537 ± 0.0033	0.0025 ± 0.0001
	SGLD	1.1189 ± 0.0033	0.9266 ± 0.0036	0.6225 ± 0.0038	0.0074 ± 0.0012
	Ensembles	0.9837 ± -	0.7981 ± -	0.4729 ± -	0.0016 ± -
GROVER	Deterministic	0.9297 ± 0.0182	0.7568 ± 0.0165	0.4238 ± 0.0178	0.0019 ± 0.0002
	MC Dropout	0.9366 ± 0.0146	0.7683 ± 0.0113	0.4348 ± 0.0148	0.0033 ± 0.0004
	SWAG	0.9290 ± 0.0148	0.7621 ± 0.0104	0.4223 ± 0.0144	0.0034 ± 0.0003
	BBP	0.9753 ± 0.0229	0.7950 ± 0.0192	0.6023 ± 0.0346	0.0025 ± 0.0007
	SGLD	1.1203 ± 0.0238	0.9301 ± 0.0243	0.6134 ± 0.0206	0.0052 ± 0.0031
	Ensembles	0.9160 ± -	0.7455 ± -	0.4134 ± -	0.0019 ± -
Uni-Mol	Deterministic	1.0917 ± 0.0210	0.9069 ± 0.0203	0.6446 ± 0.0189	0.0055 ± 0.0014
	MC Dropout	1.0929 ± 0.0216	0.9080 ± 0.0208	0.6442 ± 0.0192	0.0054 ± 0.0014
	SWAG	1.0822 ± 0.0114	0.8986 ± 0.0115	0.6144 ± 0.0121	0.0045 ± 0.0004
	BBP	1.1003 ± 0.0198	0.9112 ± 0.0160	0.8454 ± 0.0580	0.0124 ± 0.0022
	SGLD	1.1796 ± 0.0644	0.9567 ± 0.0190	0.6812 ± 0.0469	0.0135 ± 0.0051
	Ensembles	1.0863 ± -	0.9019 ± -	0.6421 ± -	0.0052 ± -

Table 39: Test results with frozen backbone model weights on QM7.

		RMSE	MAE	NLL	CE
ChemBERTa	Deterministic	137.8736 ± 2.2380	98.5749 ± 2.9618	5.4092 ± 0.0182	0.0049 ± 0.0001
	MC Dropout	133.8345 ± 1.8327	97.7798 ± 2.3986	5.4049 ± 0.0117	0.0028 ± 0.0001
	SWAG	138.2471 ± 2.2020	99.0168 ± 2.7806	5.4107 ± 0.0177	0.0048 ± 0.0001
	BBP	141.0033 ± 1.2990	104.8384 ± 1.5366	5.4635 ± 0.0100	0.0026 ± 0.0005
	SGLD	202.5163 ± 1.4235	165.5756 ± 1.1589	5.8131 ± 0.0087	0.0208 ± 0.0015
	Ensembles	136.6747 ± -	97.2405 ± -	5.4066 ± -	0.0045 ± -
GROVER	Deterministic	134.4406 ± 3.7901	95.5997 ± 4.3313	5.4077 ± 0.0717	0.0058 ± 0.0023
	MC Dropout	131.7638 ± 3.5591	93.2436 ± 4.3578	5.4005 ± 0.0761	0.0042 ± 0.0021
	SWAG	130.2656 ± 2.2246	92.1086 ± 3.1537	5.3670 ± 0.0446	0.0047 ± 0.0015
	BBP	137.9643 ± 6.4603	100.5829 ± 7.7898	5.8332 ± 0.1322	0.0055 ± 0.0030
	SGLD	191.3884 ± 3.4342	154.4454 ± 1.8662	5.8488 ± 0.0227	0.0158 ± 0.0033
	Ensembles	130.9773 ± -	93.3082 ± -	5.4027 ± -	0.0051 ± -
Uni-Mol	Deterministic	183.5096 ± 9.5881	148.3822 ± 9.8276	5.7908 ± 0.0353	0.0180 ± 0.0051
	MC Dropout	183.7110 ± 9.1987	148.5660 ± 9.3875	5.7908 ± 0.0336	0.0178 ± 0.0048
	SWAG	171.6691 ± 1.7586	137.1572 ± 2.1770	5.7147 ± 0.0078	0.0111 ± 0.0008
	BBP	178.0057 ± 3.9910	142.9159 ± 3.6271	6.0873 ± 0.0082	0.0193 ± 0.0007
	SGLD	226.7307 ± 21.2619	188.4545 ± 19.4483	6.0090 ± 0.1956	0.0458 ± 0.0322
	Ensembles	175.6095 ± -	140.6629 ± -	5.7658 ± -	0.0129 ± -