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] Georgia Institute of Technology; [2] Cornell University

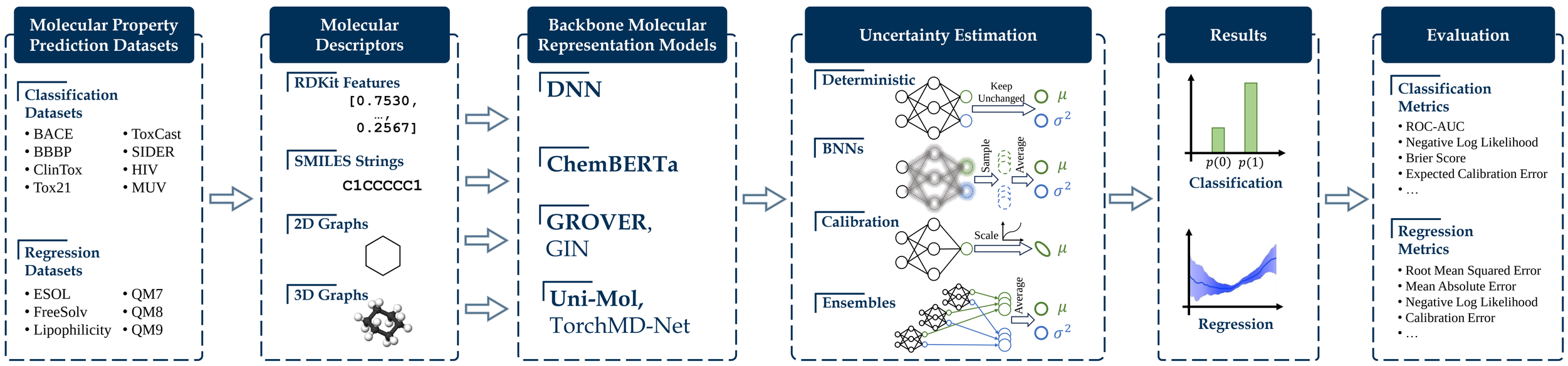

Figure 1. An overview of MUBen with datasets, backbone models, UQ methods, and metrics enumerated.

Motivation

- Pre-trained molecular representation models have demonstrated impressive representational capabilities, achieving SOTA performance on a variety of property prediction tasks.
- It is desirable for predictions to be precise and *uncertainty-aware*
  - To allows us to distinguish noisy predictions and improve model robustness or estimate data distributions.
  - Downstream tasks/applications: active learning; high throughput screening; wet-lab experimental design.
- MUBen: **U**ncertainty **Ben**chmark for **M**olecular Properties
  - Combines various uncertainty quantification (UQ) methods with representative molecular representation backbones.
  - Evaluates both property prediction & uncertainty estimation on various MoleculeNet tasks with different metrics.
  - The most comprehensive molecular UQ evaluation so far.

Backbone Models

Categorized by input molecular descriptors

- **RDKit Features** (normazlied, heuristic features)
  - Feed-Forward Deep Neural Network (DNN, not pre-trained)
- **SMILES Strings**
  - ChemBERTa (*Chithrananda et al., 2020; Ahmad et al., 2022*)
- **2D Graph**
  - GROVER (*Rong et al., 2020*)
  - GIN (not pre-trained, *Xu et al., 2019*)
- **3D Graph**
  - Uni-Mol (Zhou et al., 2023)
  - TorchMD-NET (*Thölke & Fabritiis, 2022; Zaidi et al., 2023*)

Uncertainty Quantification Methods

- **Deterministic Prediction**: one-time learning &inference
  - Post-activation probability (classification) & mean-variance (regression)
  - Focal Loss (classification-only, *Lin et al., 2017; Mukhoti et al., 2020*) increases the entropy of the predicted distribution by minimizing a regularized KL divergence between the predicted values and the true labels.
- **Bayesian Learning and Inference**: distribution over parameters
  - Bayes by Backprop (BBP, *Blundell et al., 2015; Kingma et al., 2015*), an algorithm for training Bayesian Neural Networks (BNNs) with Monte Carlo loss estimation and (local) reparameterization trick for gradient backpropagation.
  - Stochastic Gradient Langevin Dynamics (SGLD, *Welling & Teh, 2011*) applies Langevin dynamics to infuse noise into the stochastic gradient descent training process.
  - MC Dropout (*Gal & Ghahramani, 2016*) derives uncertainties from an ensemble of multiple stochastic forward passes with dropout enabled.
  - SWA-Gaussian (SWAG, *Maddox et al., 2019*) estimates Gaussian posteriors over weights with low-rank stochastic weight averaging (SWA, *Izmailov et al., 2018*).
- **Post-Hoc Calibration**: adjusting the model outputs after training
  - Temperature Scaling (classification-only, *Platt et al., 1999; Guo et al., 2017*) adds a learned scaling factor to the Sigmoid or SoftMax output activation to control the output spikiness.
- **Deep Ensembles**
  - Trains a deterministic network multiple times with different random seeds and combines their predictions at inference (*Lakshminarayanan et al., 2017*).

Datasets

We use a subset from MoleculeNet Benchmark (*Wu et al., 2018*)

	Category	Dataset	# Compounds	# Tasks	Average LIR	Max LIR
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	-	-

LIR: Label Imbalance Ratio: $\mathrm{LIR}_k \in [0.5, 1] = \max\{p_{\mathrm{pos}}, 1 - p_{\mathrm{pos}}\};\ p_{\mathrm{pos}} = \sum_{i=1}^{N} \mathbb{I}(l_i = 1)/N$

Results & Observation

Comparions of Backbone Models

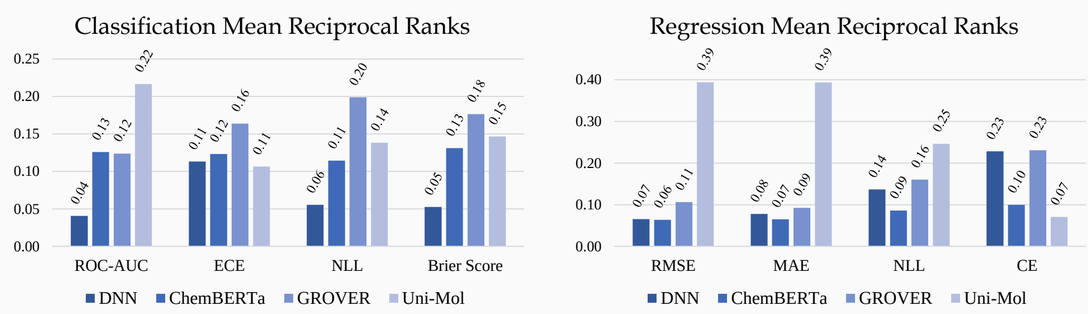

Figure 2. MRR of DNN, ChemBERTa, GROVER and Uni-Mol, each is macro-averaged from the reciprocal ranks of the results of all corresponding UQ methods on all datasets. GIN consistently underperforms other backbones.

- Uni-Mol performs the best for property prediction (ROC-AUC, RMSE and MAE), but tend to be over-confident, yielding sub-optimal calibration (ECE and CE).
- GROVER is a safer choice when both prediction and UQ performance are required.
- Pre-trained models do not invariably surpass heuristic features, as shown in the comparison between DNN & ChemBERTa for regression.

Comparisons of Uncertainty Quantification Methods

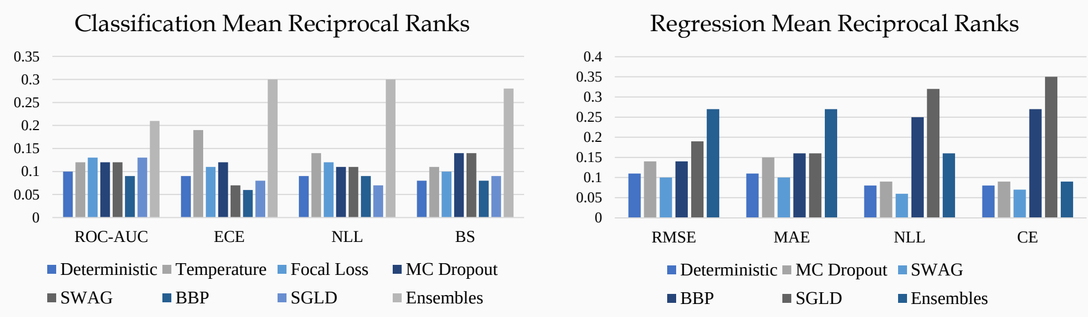

Figure 3. MRR of all UQ methods, macro-averaged of DNN, ChemBERTa, GROVER and Uni-Mol on all datasets.

- Most UQ methods enhance both value prediction and uncertainty estimation.
- BBP and SGLD fail on classification but deliver the greatest improvement on regression.
- Deep Ensembles guarantees to improve the prediction and UQ results, but at a cost of heavy computational consumption.
- MC Dropout is cheap to adopt and theoretically does not risk model performance under any circumstances, making it a first-pick when computation resource is limited.
- Temperature Scaling is also cheap for classification calibration, but it may fail when the held-out calibration dataset has a distribution different from the test set.

Case Studies