Uncertainty quantification of molecular property prediction using Bayesian neural network models

Seongok Ryu[1], Yongchan Kwon[2], Woo Youn Kim[1,3]
[1]Korea Advanced Institute of Science and Technology (KAIST), [2]Seoul National University,
[3]KAIST Institute for Artificial Intelligence

Why do we need Bayesian deep learning for molecular applications?

Despite successes of deep learening models in molecular applications, several limitations of the models remain: **i) very data hungry**, **ii) poor at representing uncertainty**, and **iii) hard to interpret and to trust results**. For example, the number of samples in Tox21 dataset is less than 15,000 for all toxicities, and in DUD-E dataset is almost 35,000 for EGFR target. Moreover, these datasets are biased, with most samples having negative labels. So **"how can we build more reliable deep neural network models for successful molecular applications?"** Our approach is to use Bayesian deep learning to quantify the uncertainty in molecular property predictions, in the spirit of **"Knowing what we don't know is as important as knowing itself"**.

Bayesian deep learning and uncertainty quantification

- **Maximum-a-posteriori (MAP) estimation** $\hat{\mathrm{w}} = \underset{\mathrm{w}\in\mathcal{W}}{\operatorname{argmax}}\, p(\mathrm{w}|\mathrm{X},\mathrm{Y}) \longrightarrow p(\mathrm{y}^*|\mathrm{x}^*,\mathrm{X},\mathrm{Y}) = \mathrm{f}^{\hat{\mathrm{w}}}(\mathrm{x}^*)$

MAP model estimation chooses a single model to best explain the given data distribution, which is equal to the mode of the posterior. Inferred output y^* to given input x^* is a single deterministic value $\mathrm{f}^{\hat{\mathrm{w}}}(\mathrm{x}^*)$.

- **Bayesian inference** $p(\mathbf{w}|\mathbf{X},\mathbf{Y}) = \dfrac{p(\mathbf{X},\mathbf{Y}|\mathbf{w}) \times p(\mathbf{w})}{p(\mathbf{X},\mathbf{Y})}$ $\quad p(\mathrm{y}^*|\mathrm{x}^*,\mathrm{X},\mathrm{Y}) = \int_{w\in\mathcal{W}} p(\mathrm{y}^*|\mathrm{x}^*,\mathrm{w})p(\mathrm{w}|\mathrm{X},\mathrm{Y})d\mathrm{w}$

Bayesian considers all quantities uncertain, except for the given data. Thus, model parameters and posterior have a distribution instead of single deterministic value. Therefore, **we can measure the uncertainty in our prediction as the variance of output distribution**. The uncertainty can be divided into two categories:
i) epistemic (model) uncertainty $Var(\mathbb{E}[y^*|x^*])$ and ii) aleatoric (data-inherent) uncertainty $\mathbb{E}[Var(y^*|x^*)]$.

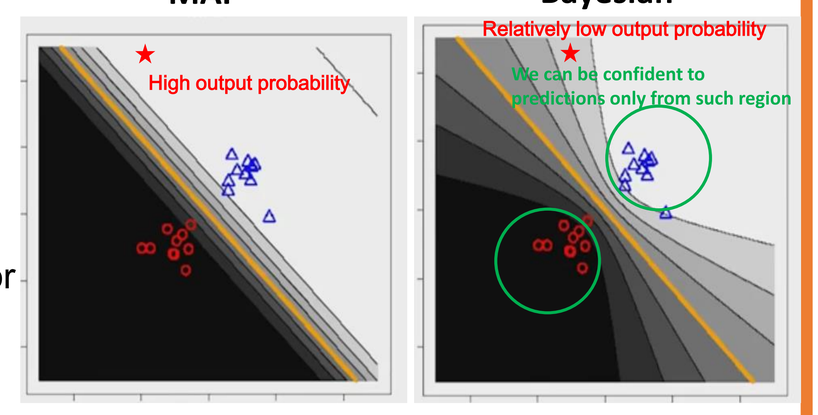


Figure source: https://www.youtube.com/watch?v=1ClZhMSHeBA

Graph convolutional network for molecular property predictions

We represented molecular structures by graph structure $\mathrm{G}(\mathbf{A},\mathbf{X})$ and used graph convolutional network (GCN) to update their node and graph features. The vanilla GCN updates node features according to (a)-*left*. We improved it by adopting the **attention mechanism** and **gated skip-connection** in (a)-*right* and (b)-*right,* respectively.
In order for the model parameters to be stochastic, we applied dropouts at every hidden layer, which is referred to as a Monte Carlo dropout network.

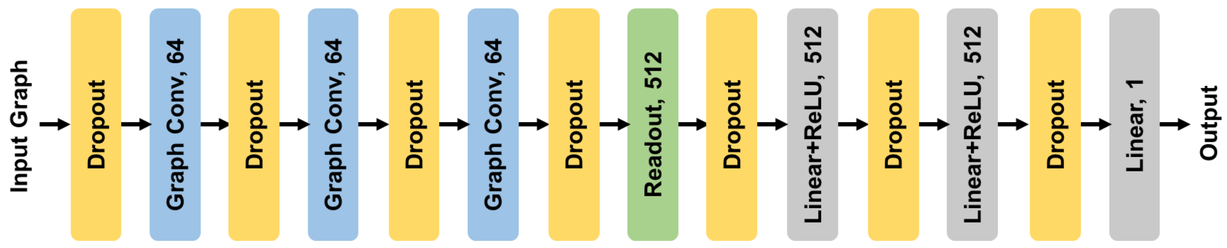
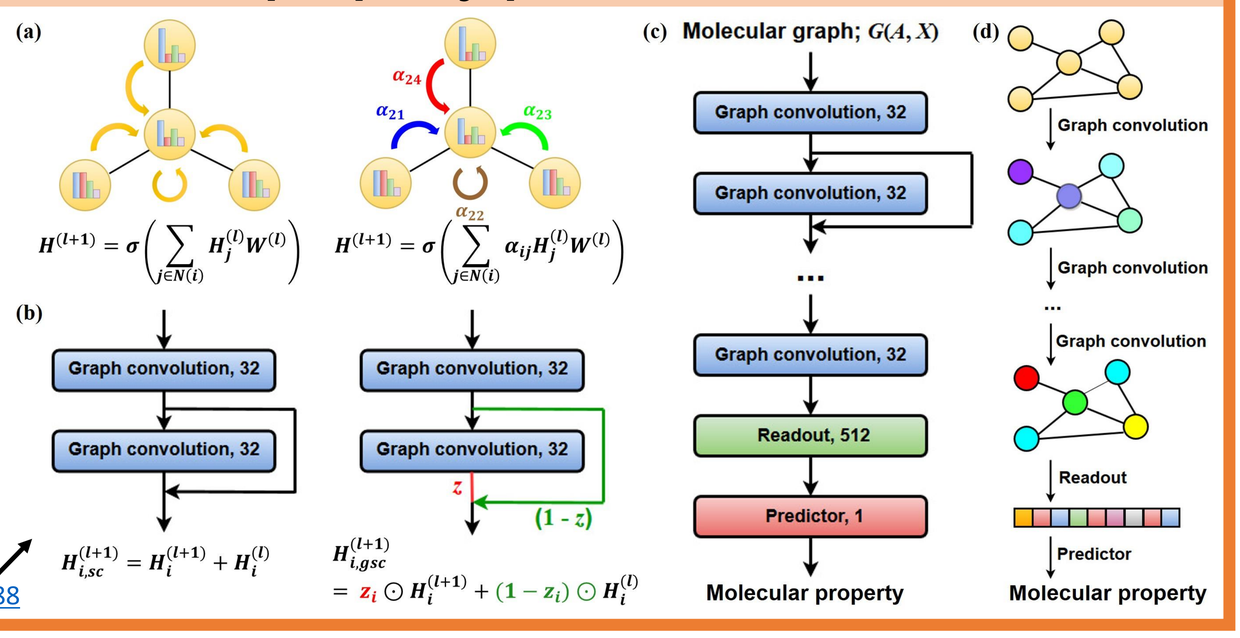
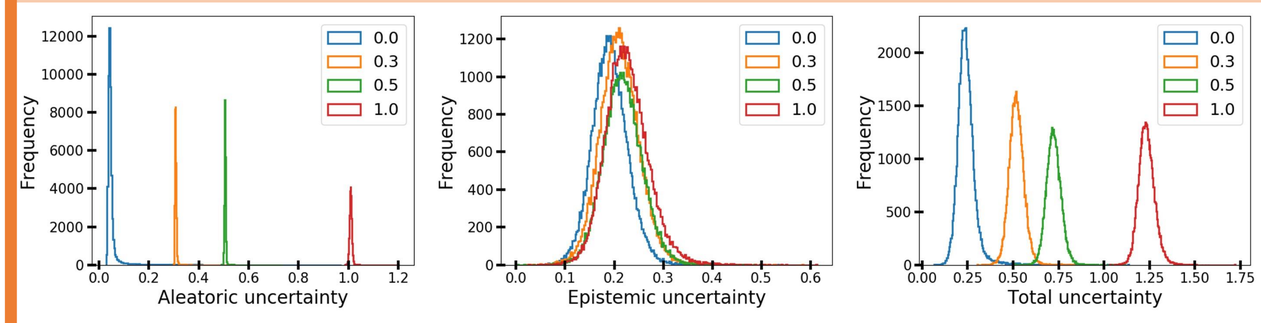


Figure source: S. Ryu et.al. arXiv:1805.10988

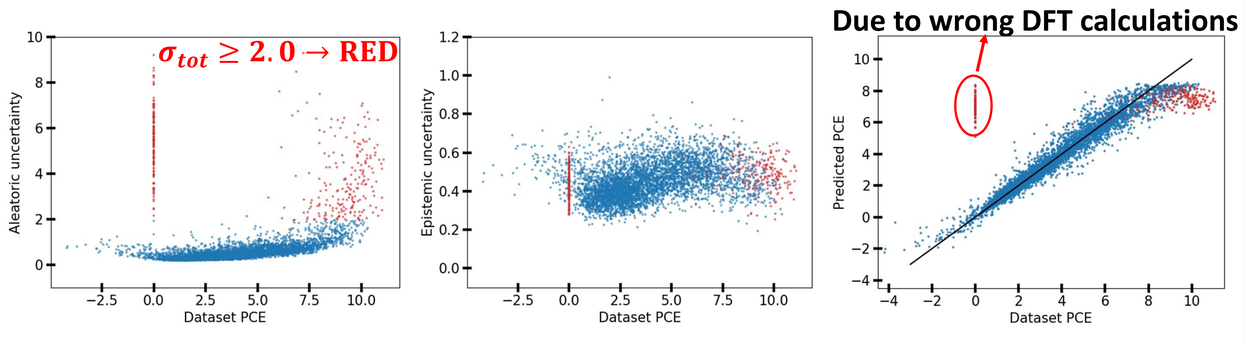


Exp 1. Aleatoric uncertainty due to data quality



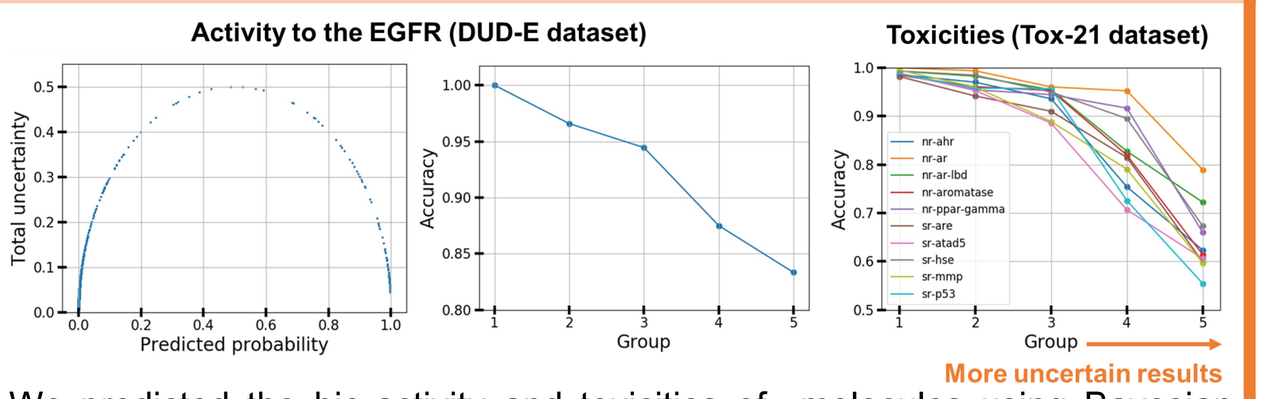
In the training regime, we added random noise $\epsilon \sim \mathcal{N}(0,\sigma^2)$ to logP. **As σ^2 increases, the aleatoric uncertainty grows, but the epistemic uncertainty remains the same.**

Exp 2. Evaluating quality of synthetic data using UQ

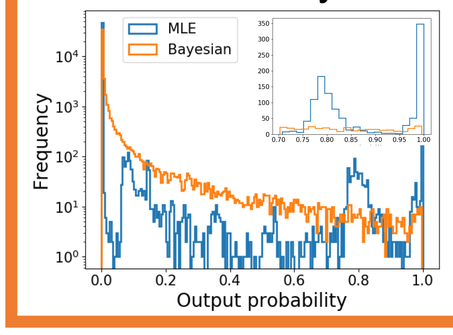


Many samples in Harvard Clean-Energy Project (CEP) dataset (J. Hachmann *et.al.*, 2011) have exactly zero PCE values. Questioning whether the given ground truth values are correct or not, we applied uncertainty quantification to evaluate synthetic data. The samples show large aleatoric uncertainties. We further investigated how PCE values were obtained. The DFT calculation gave infinitesimal values of J_{SC} when working poorly. Such erroneous approximation then resulted in zero $\mathrm{PCE}(\propto V_{OC} \times FF \times J_{SC})$ values.

Exp 3. Uncertainties in classification tasks



We predicted the bio-activity and toxicities of molecules using Bayesian NNs. As a result, the less uncertain results are more likely to have ground truth label zero or one. We divided the group in increasing order of uncertainty measured prediction accuracy for each group. **The groups with smaller uncertainty show more accurate results**.



We screened the ZINC dataset to find active molecules to the EGFR using MLE and Bayesian models. **It seems that the MLE model is over-confident, while the BNN isn't.** We plan to evaluate binding affinity of screened molecules via MD simulations and *in vitro* assays.

What deep learning approaches do we need for reliable and practical molecular applications?

- Bayesian deep learning
- Semi-supervised learning
- Pre-training, multi-task/transfer learning like BERT (J Devlin *et. al.*, 2018)

My future research directions
– towards low data drug discovery