Benchmarking Uncertainty Quantification for Protein Engineering

Kevin P. Greenman[1], Ava P. Amini[2], Kevin K. Yang[2]

[1] Chemical Engineering, MIT; [2] Microsoft Research New England

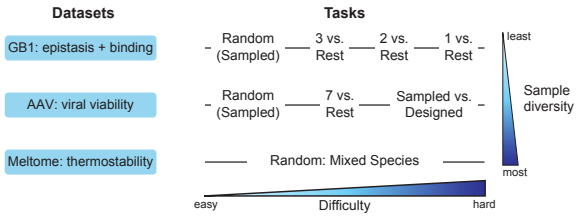
Background and Motivation

- Machine learning (ML) currently applied successfully in protein engineering (low-cost estimates to replace time- and resource-intensive experiments)
- ML model performance highly dependent on domain shift between training and testing data
- Domain shift common in protein engineering because of biased data collection
- Uncertainty quantification (UQ) benchmarked in other fields (e.g., chemistry and materials science) to understand effect of domain shift on model reliability
- No such benchmark has been done on protein datasets

We benchmark a panel of UQ methods on standardized datasets to assess the effect of distributional shift and provide recommendations for use in active learning.

Datasets and Splits

8 splits across 3 protein landscapes from FLIP[1] cover varied levels of distributional shift between train and test

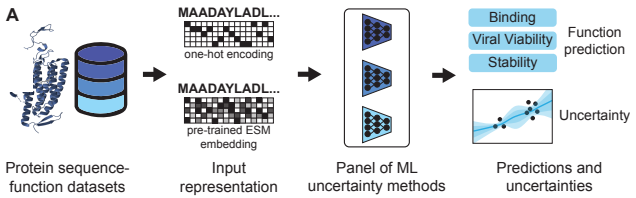


1. Dallago, et al. NeurIPS Datasets and Benchmarks Track (2021).

Uncertainty Evaluation Metrics

↓ **RMSE**	↓ **MA**	↑ **C**
root mean square error of predicted values to true values	miscalibration area: area under the calibration error curve	coverage: % of true values that fall within ±2σ of prediction
↑ **ρ**	↓ **4σ/R**	↑ **ρ_{unc}**
rank correlation of predicted values to true values	width of 95% confidence interval relative to training set range	rank correlation of uncertainties to residuals

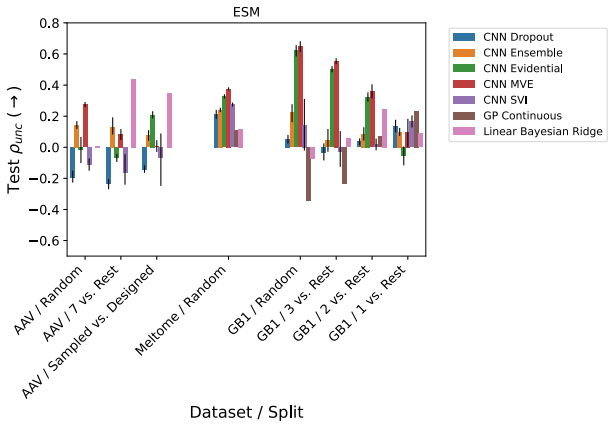
Models and Uncertainty Methods



CNN Methods		Other Methods
Dropout	MVE	Gaussian Process (GP)
Ensemble	SVI	Linear Bayesian Ridge
Evidential		

Trained models with 7 UQ methods on each of 8 dataset splits and compared performance

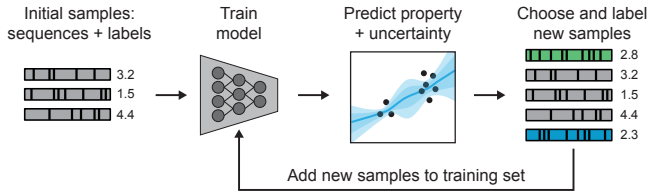
Uncertainty Results



- MVE and evidential uncertainty methods are most performant in ρ_{unc} for most cases of low to moderate domain shift
- Most methods have ρ_{unc} near zero for the most challenging splits.

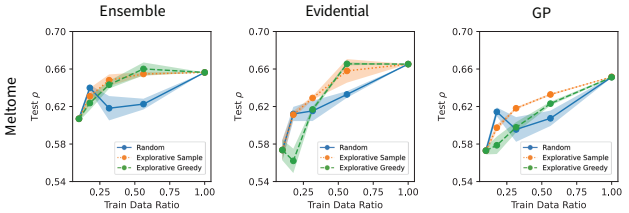
No single method performs consistently well across all metrics, landscapes, and splits

Active Learning and Bayesian Optimization



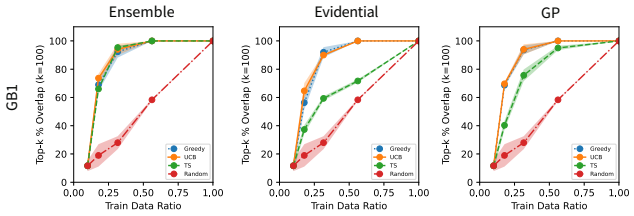
Uncertainty can be leveraged for model improvement or sequence optimization

Active Learning



- Uncertainty-based sampling can outperform random sampling in some cases of active learning
- Random sampling can perform better early on, before the model has enough training data to make good uncertainty predictions

Bayesian Optimization



- Uncertainty-based strategies typically perform better in optimization than random sampling
- Greedy sampling often performs as well or better than uncertainty-based strategies

Paper

K.P. Greenman, A.P. Amini, and K.K. Yang, "Benchmarking Uncertainty Quantification for Protein Engineering", *bioRxiv* (2023). https://doi.org/10.1101/2023.04.17.536962