Benchmarking Uncertainty Quantification for Protein Engineering

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

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

Background and Motivation

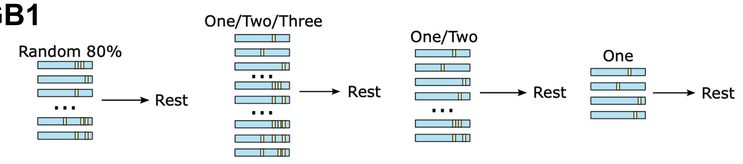
- Machine learning (ML) currently applied successfully in protein engineering (low-cost estimates to replace time- and resource-intensive experiments)
- ML model performance highly dependent on domain shift between training and testing data
- Domain shift common in protein engineering because of biased 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.

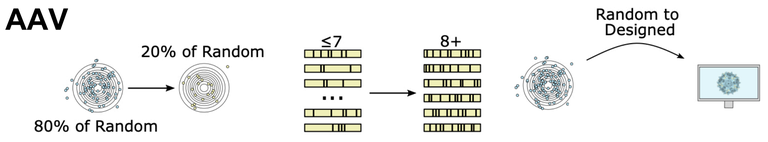
Datasets and Splits

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

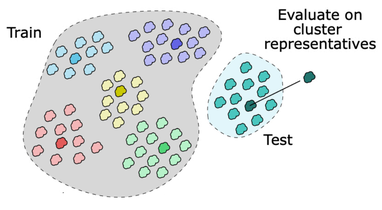
GB1



AAV

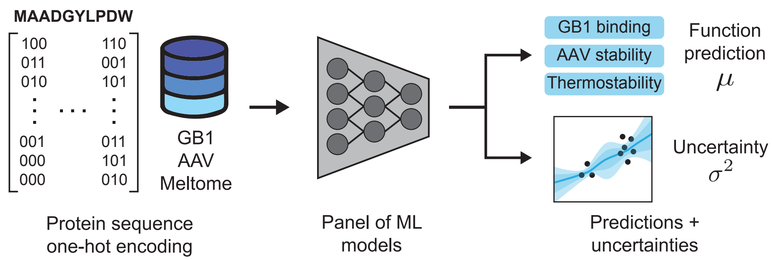


Meltome



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

Methods



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

Train models with 7 UQ methods on each of 8 dataset splits and compare performance

Evaluation Metrics

↓ RMSE

root mean square error of predicted values to true values

↓ MA

miscalibration area: area under the calibration error curve

↑ C

coverage: % of true values that fall within ±2σ of prediction

↓ 4σ/R

width of 95% confidence interval relative to training set range

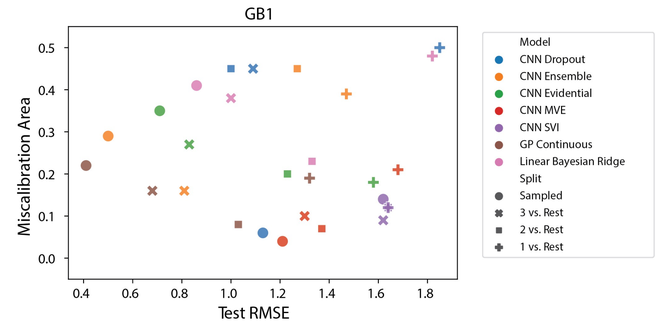
↑ ρ

rank correlation of predicted values to true values

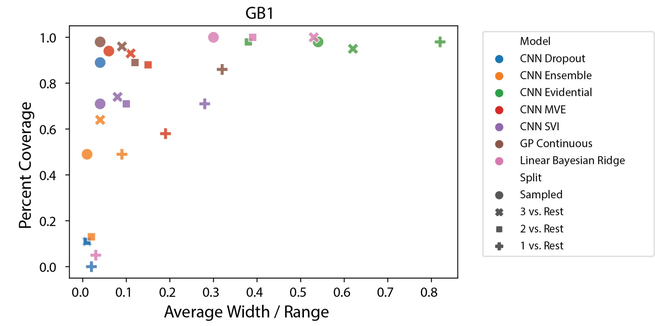
↑ ρ_{unc}

rank correlation of uncertainties to residuals

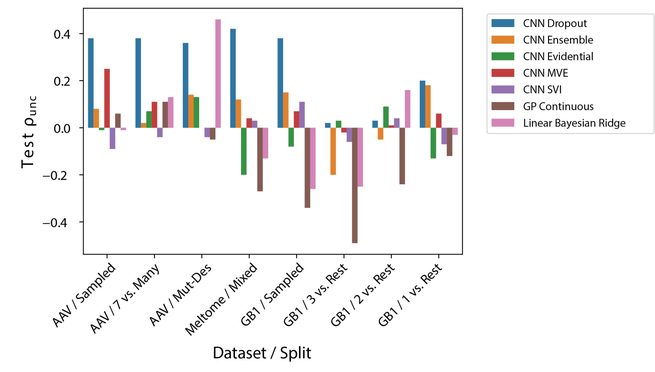
Results



- Some models highly calibrated on the highest-domain-shift splits, while others poorly calibrated even on random splits
- CNN ensemble is often one of the highest accuracy models, but also one of the most poorly calibrated



- Few methods perform well in both coverage and width
- GP among the best across all landscapes and splits



- Dropout has one of the highest and most consistent rank correlations across splits
- Many methods have rank correlations near zero for the most challenging splits

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