Benchmarking probabilistic machine learning in protein fitness landscape predictions

Ningning Chen[1], Wenkai Han[2], Sai T Reddy[3]
[1]ETH Zurich, Department of Biosystems Science and Engineering, Basel 4056, Switzerland; [2]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center

1 Introduction

- Uncertainty quantification enhances the trustworthiness of model predictions by indicating reliability and guide the next-step experiment design
- Machine learning show great potential of predicting the fitness of proteins not captured by experiments and extrapolating higher-order mutations
- A comprehensive framework is needed to evaluate the potential benefits of different types of uncertainties.

2 Method Overview

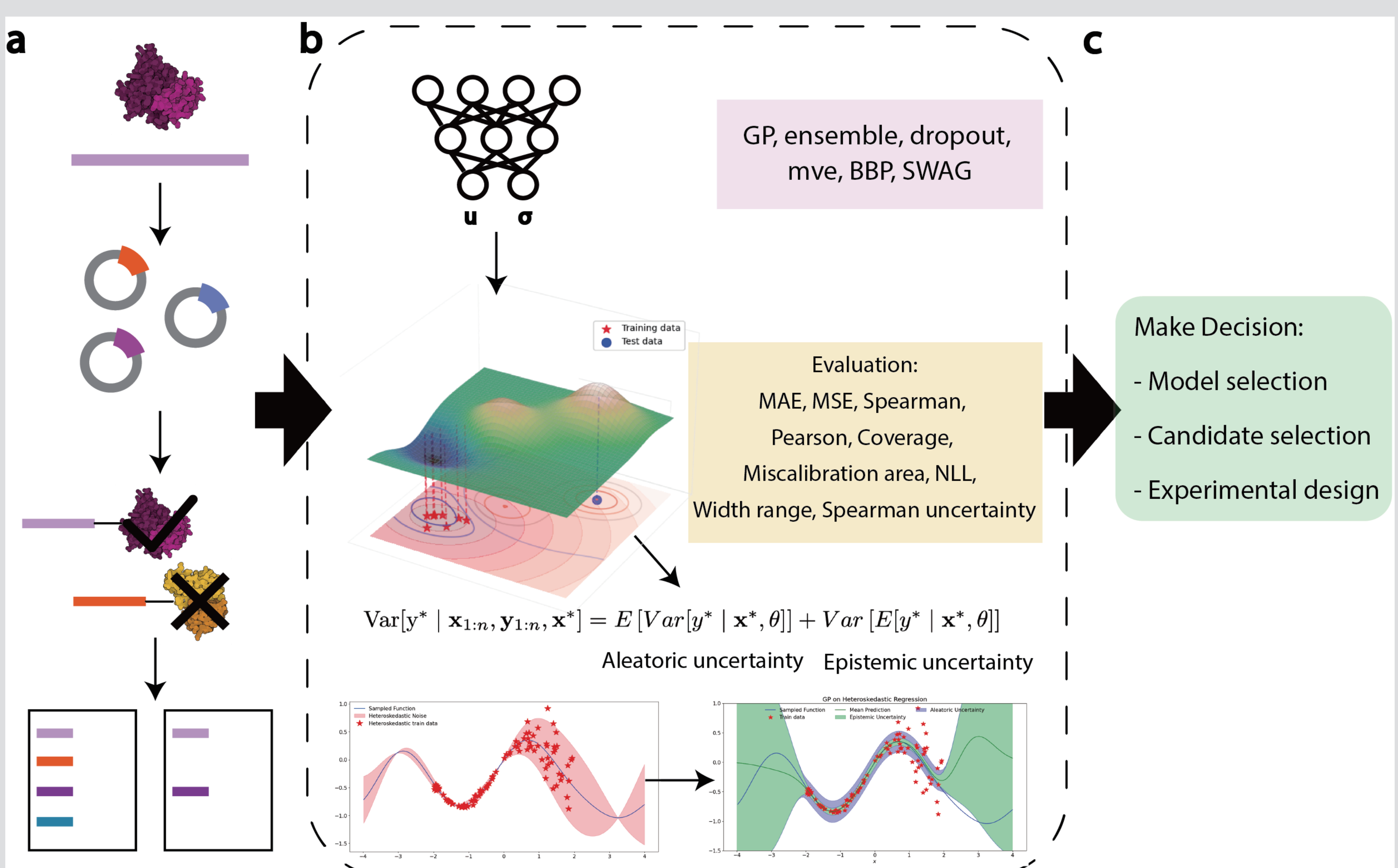
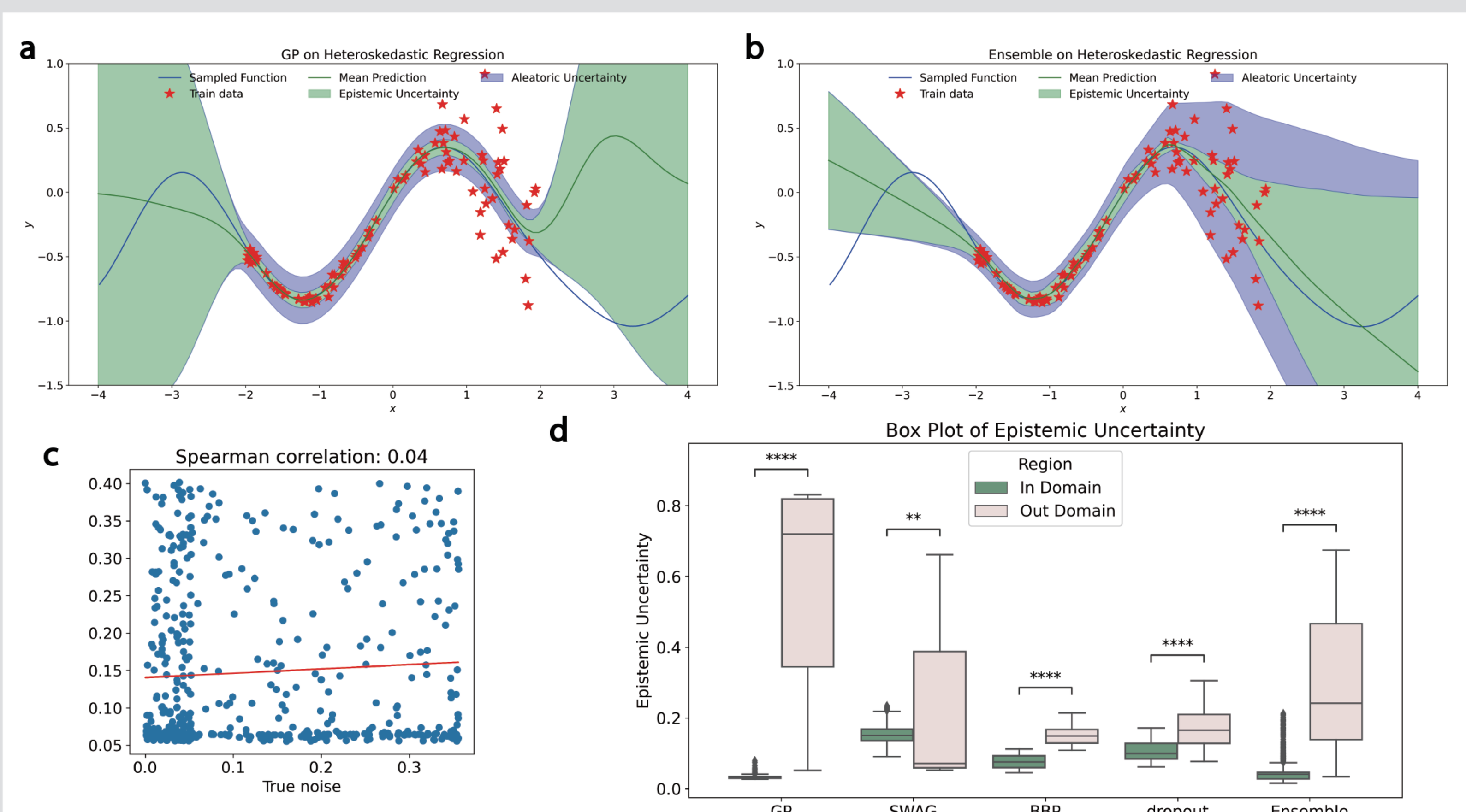


Illustration of the pipeline for leveraging the probabilistic modeling in protein fitness landscape prediction

– Probabilistic ML: Gaussian Process (GP), Bayes by Backprop (BBP), Mean-Variance Estimation (MVE), Deep Ensemble, MC Dropout, Stochastic Weight Averaging (SWAG)

3 Synthetic dataset

– Evaluate the probabilistic MLs on 1-D synthetic dataset in in-distribution(ID) and out-of-distribution(OOD) settings.



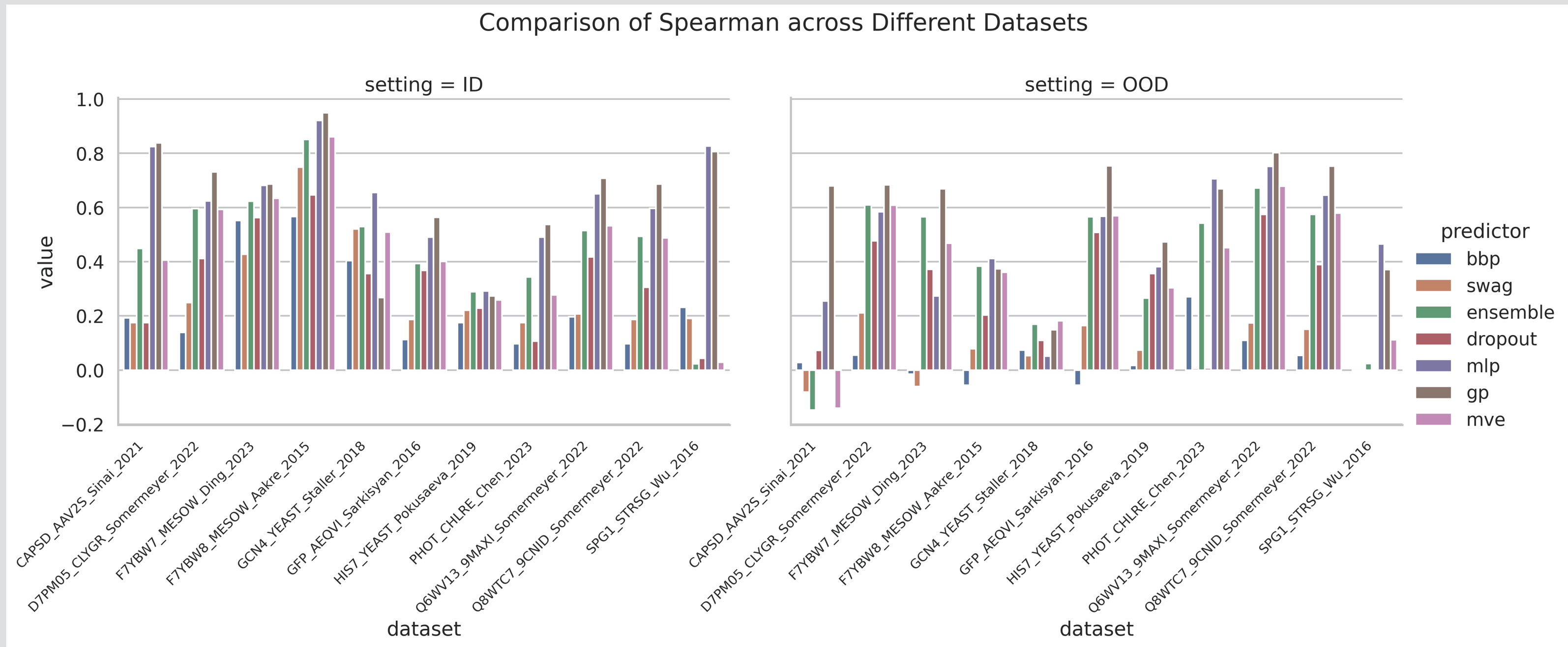
Method	MAE	MSE	Pearson	SPCC	Per_Co	AVR	SPCC_unc	NLL	Mis_area
In Distribution									
swag	0.140	0.050	0.910	0.920	0.990	0.290	0.530	0.450	0.130
dropout	0.120	0.040	0.910	0.930	**1.000**	0.290	0.670	0.470	0.210
ensemble	0.120	0.040	0.910	0.930	0.980	0.200	0.620	0.840	0.100
gp	**0.110**	0.040	0.920	0.940	0.800	**0.110**	0.070	-0.660	0.110
mlp	0.180	0.050	0.880	0.830	-	-	-	-	-
mve	**0.110**	**0.030**	**0.920**	0.930	0.970	0.180	**0.750**	**0.980**	0.080
bbp	0.280	0.110	0.740	0.650	**1.000**	0.470	0.420	-0.280	**0.050**
Out of Distribution									
swag	0.450	0.240	-0.860	-0.670	0.660	**0.320**	0.390	-2.280	0.320
dropout	0.440	0.230	-0.320	-0.270	0.700	0.350	0.370	-1.100	0.320
ensemble	**0.250**	**0.110**	**0.750**	**0.720**	0.730	0.480	-0.040	**-0.900**	**0.090**
gp	0.630	0.630	-0.760	-0.790	**0.790**	0.730	**0.420**	-1.150	0.120
mlp	0.540	0.360	-0.480	-0.540	-	-	-	-	-
mve	0.310	0.180	**0.750**	**0.720**	0.510	0.380	-0.130	-448.68	0.160
bbp	0.880	0.820	-0.790	-0.720	0.150	0.360	-0.350	-5.080	0.460

– Ensemble shows the best and robust performance
– Aleatoric uncertainty can only be roughly estimated
– Epistemic uncertainty is captured between ID and OOD settings

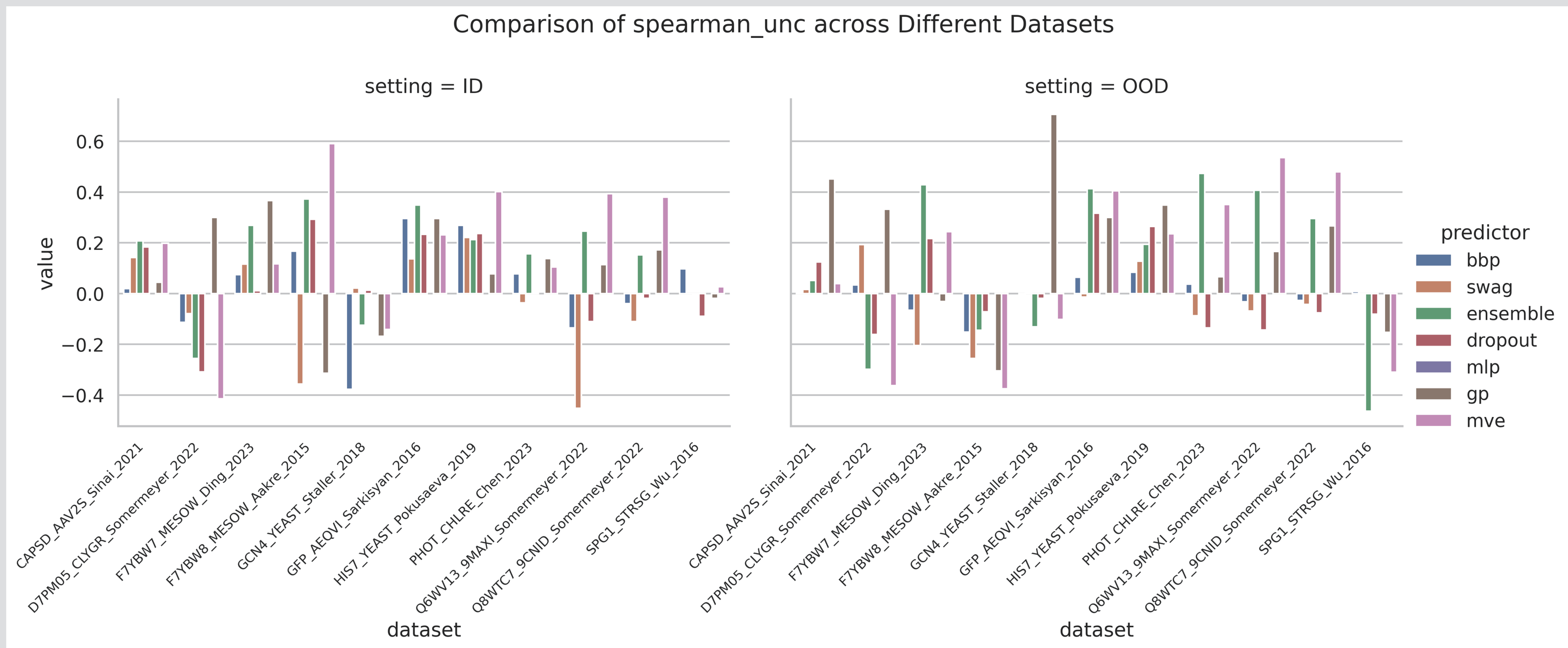
4 Protein fitness dataset

Then we benchmark 11 Deep mutational scanning(DMS) datasets targeting various sequence lengths and functions of proteins, including properties such as binding affinity and fluorescence of GFP

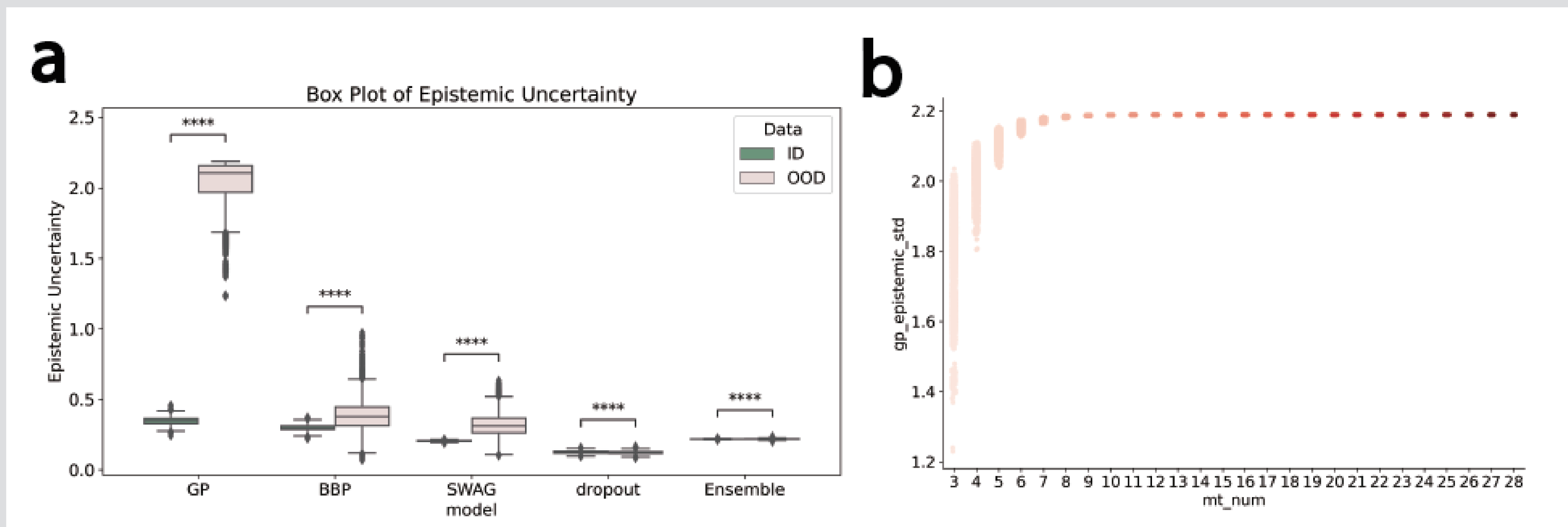
Prediction evaluation



Uncertainty evaluation



Epistemic uncertainty



5 Conclusion & Discussion

– GP generally outperforms other models in both prediction performance and uncertainty quality.
– MLP performs well in ID but fails in some OOD settings, probabilistic ML outperform MLP in these scenarios.
– Different metrics for evaluating uncertainty quality are not always consistent, metric selection should align with the specific goals of the modeling task.
– Epistemic uncertainty can be captured by probabilstic ML, while aleatoric uncertainty need to be further investiagted.

Laboratory for Systems and Synthetic Immunology
Contact:
ningning.chen@bsse.ethz.ch