A Study of Bayesian Neural Networks for Bayesian Optimization

Yucen Lily Li* Tim G. J. Rudner Andrew Gordon Wilson

* Corresponding author: yucenli@nyu.edu

PAPER

Key Takeaways

We provide a comprehensive study of BNN surrogates over a wide range of model types and experiment settings.

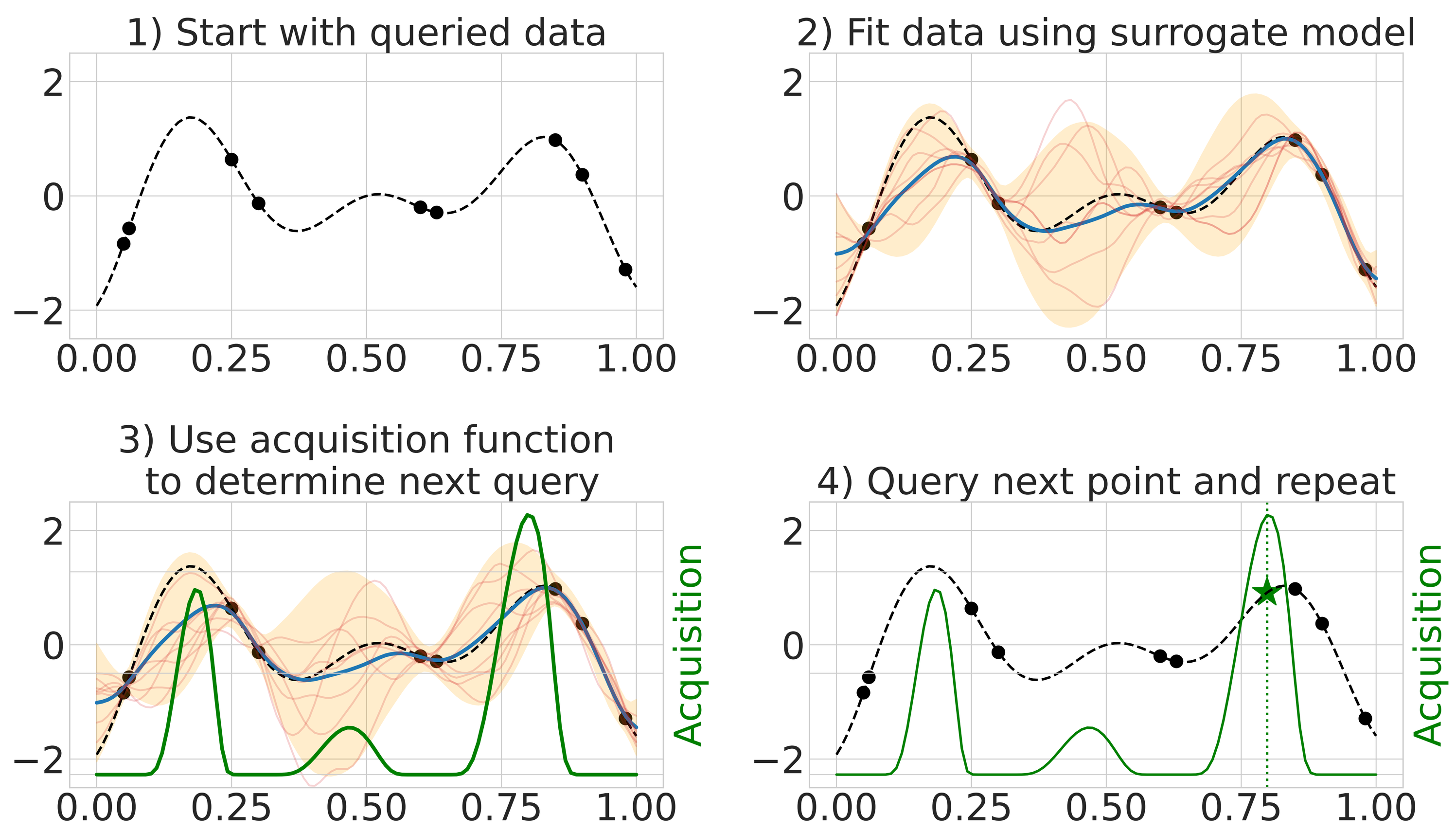
→ BNNs are competitive for Bayesian optimization and can outperform standard GPs on many problems.

→ The ranking of methods is highly problem-dependent, suggesting the need for problem-specific, tailored inductive biases.

→ Infinite-width BNNs perform well in high-dimensional spaces.

Background

Bayesian optimization is an efficient approach for finding the maximum of black-box functions which are expensive to query



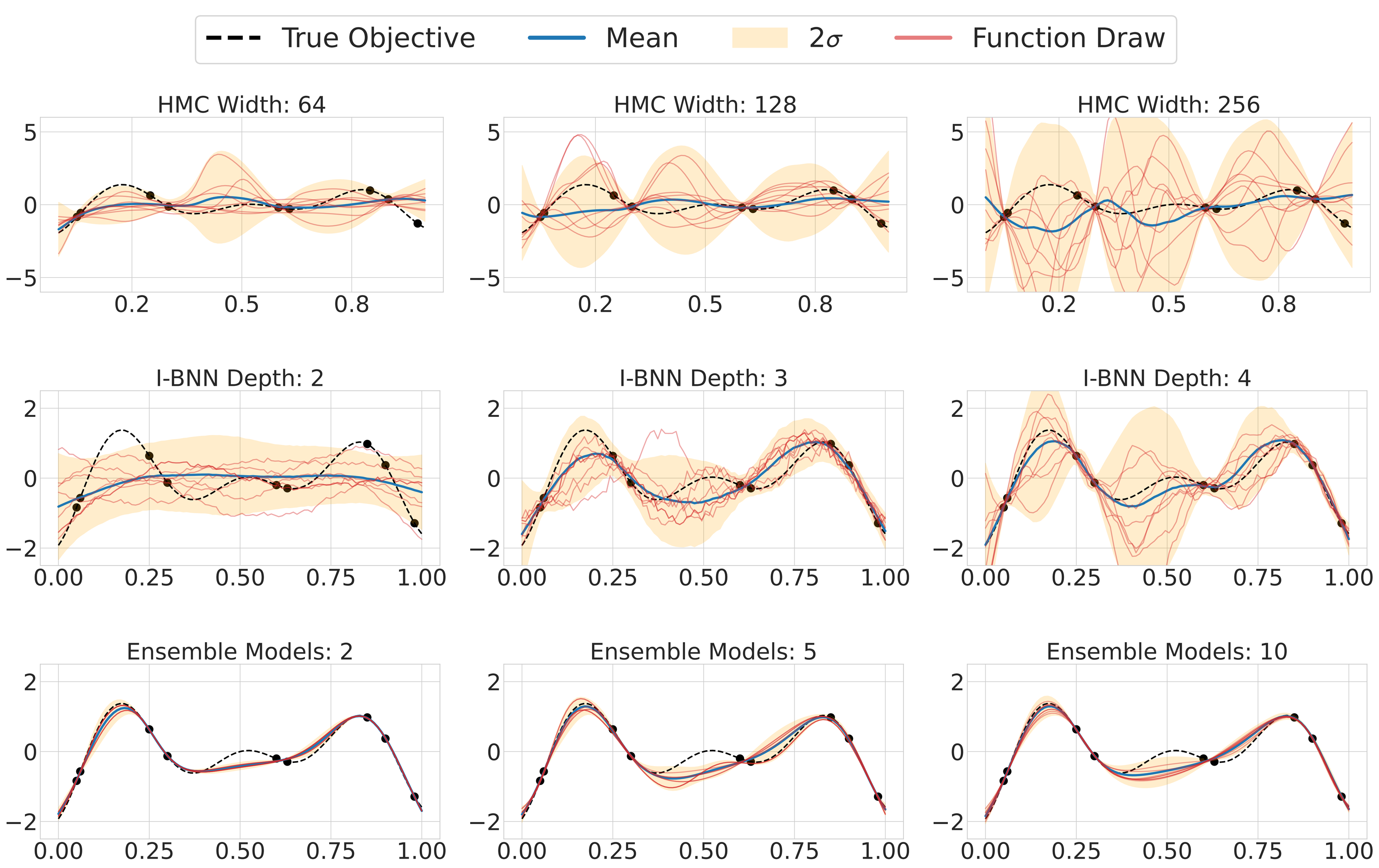
The majority of research on Bayesian optimization focuses on Gaussian process surrogates rather than Bayesian neural networks.

Surrogate Models

We benchmark seven different surrogate models:

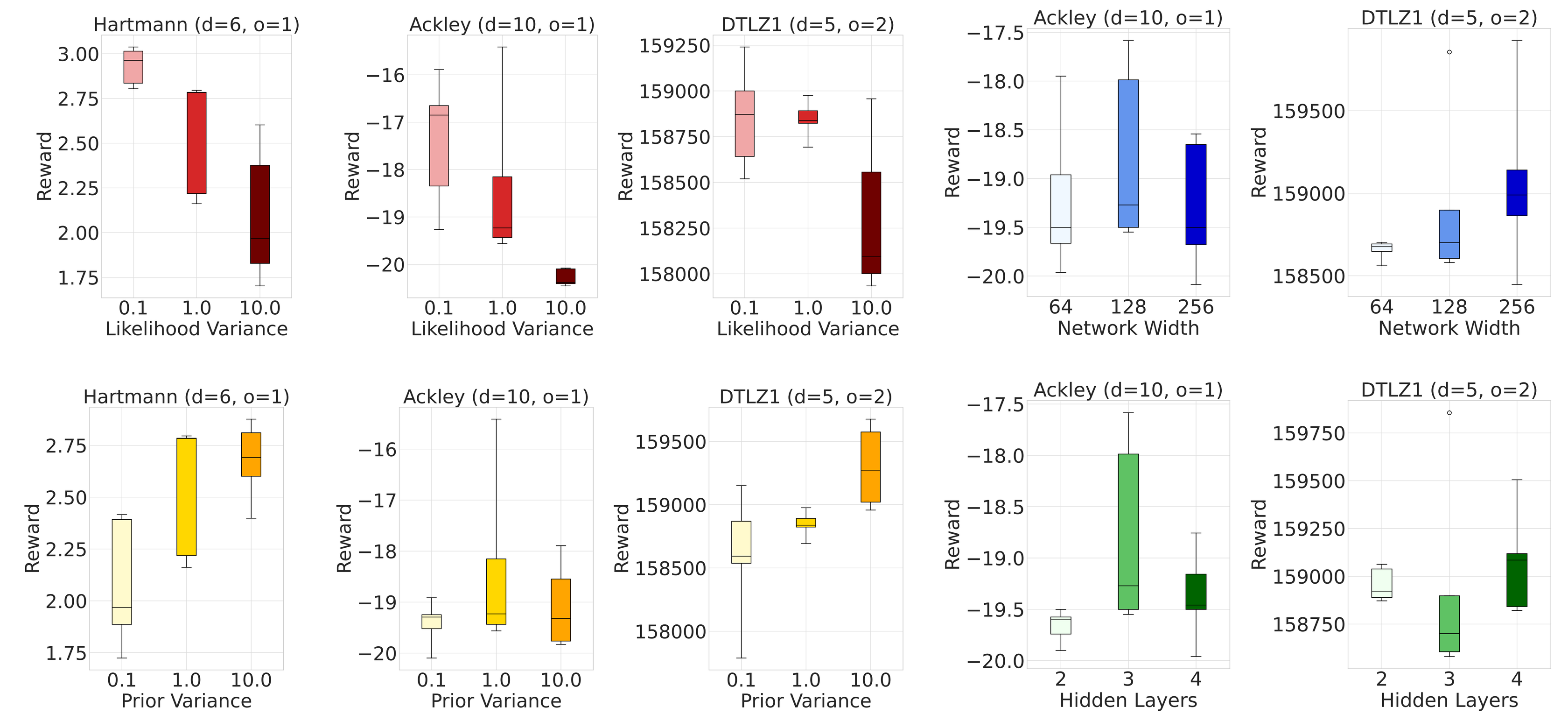
- Gaussian Process with Matérn Kernel (GP)
- Fully Stochastic Neural Networks
  - Hamiltonian Monte Carlo (HMC)
  - Stochastic Gradient HMC (SGHMC)
  - Deep Ensembles
- Infinite-width Bayesian Neural Networks (I-BNN)
- Deep Kernel Learning (DKL)
- Linearized Laplace Approximation (LLA)

Impact of BNN Architecture



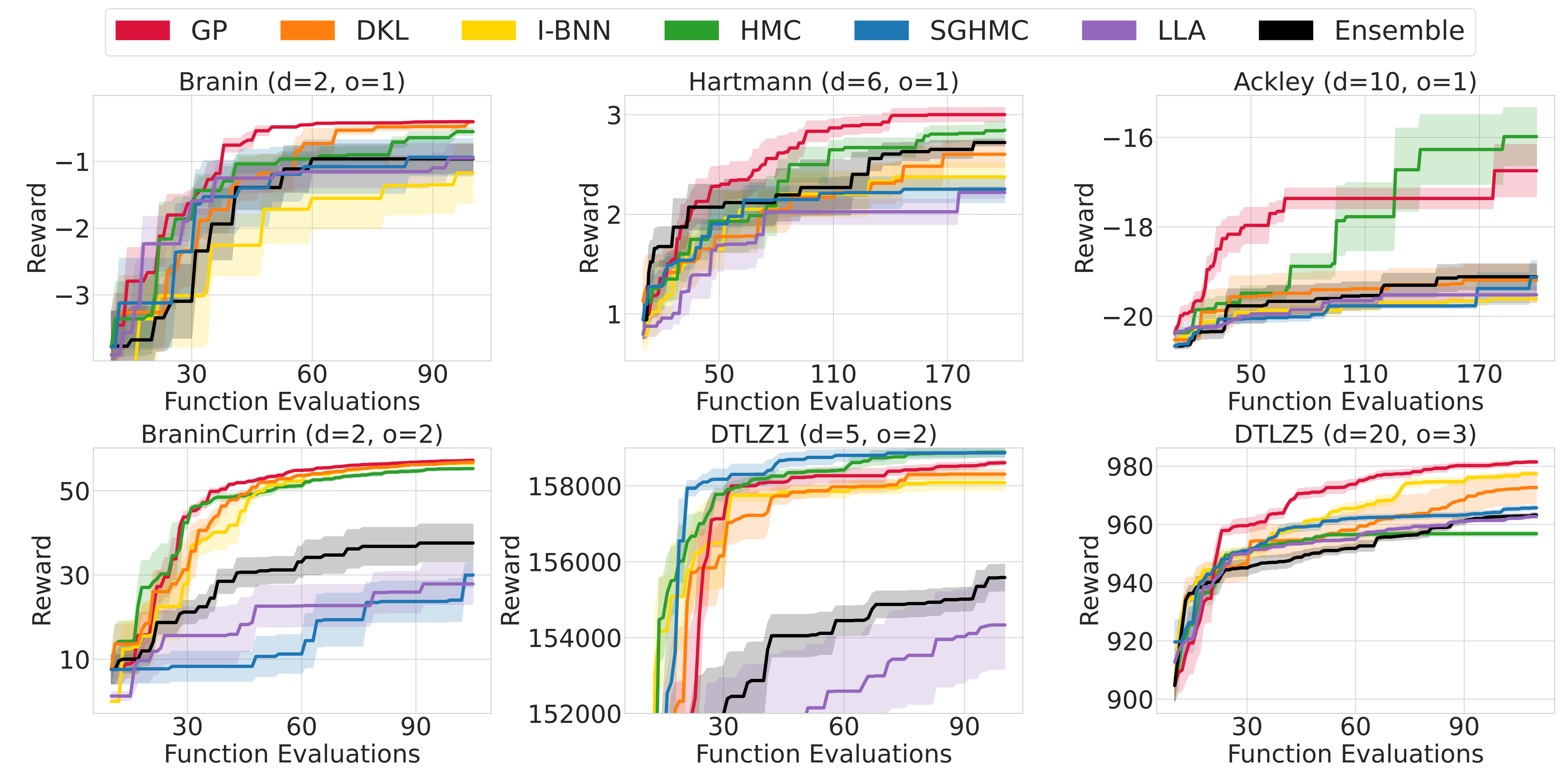
Empirical Results

Different Architectures



→ The best BNN architecture is problem-dependent

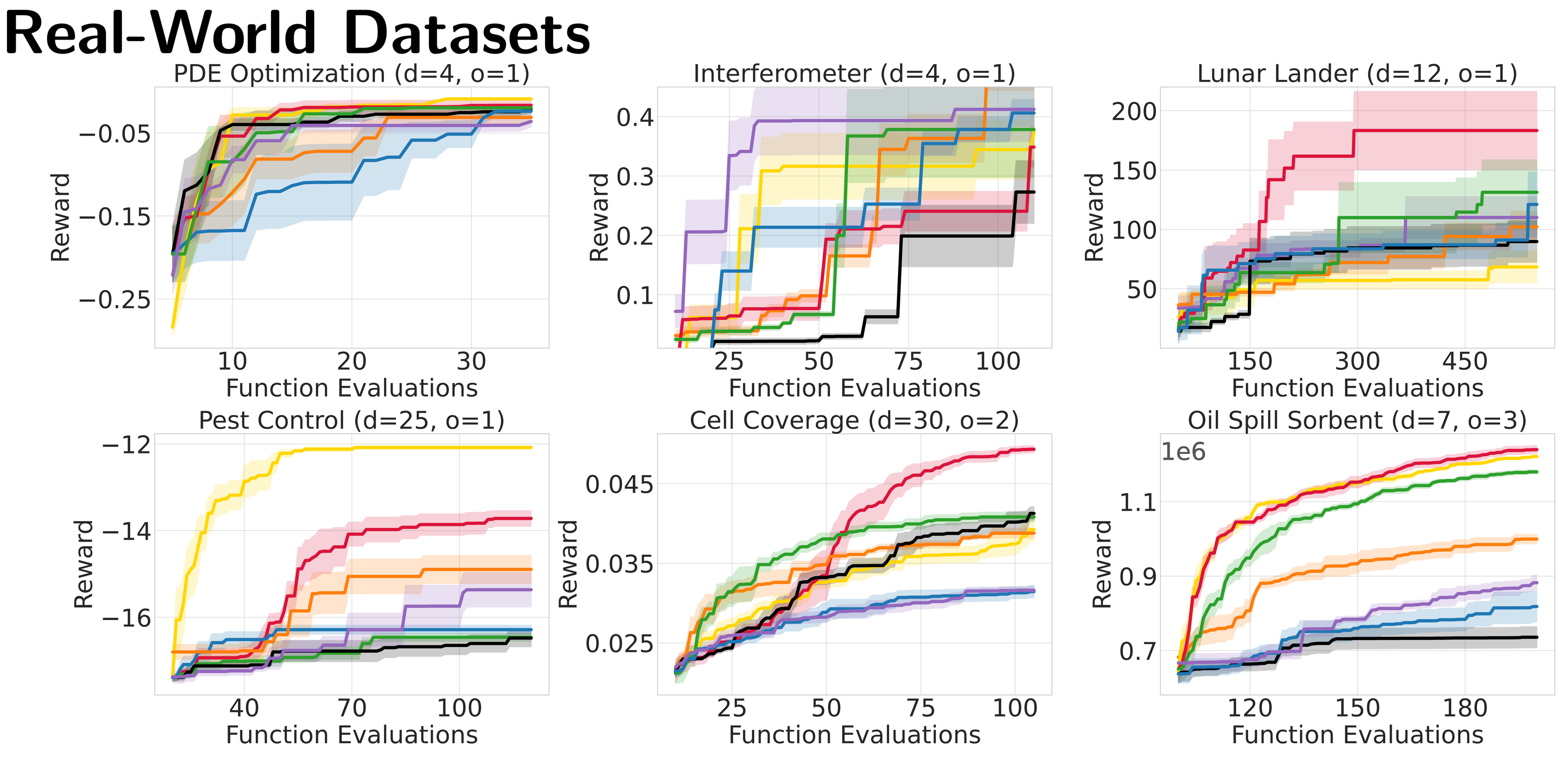
Synthetic Datasets



→ BNNs are often comparable to GPs on synthetic datasets

→ The different types of BNNs have distinct behaviors

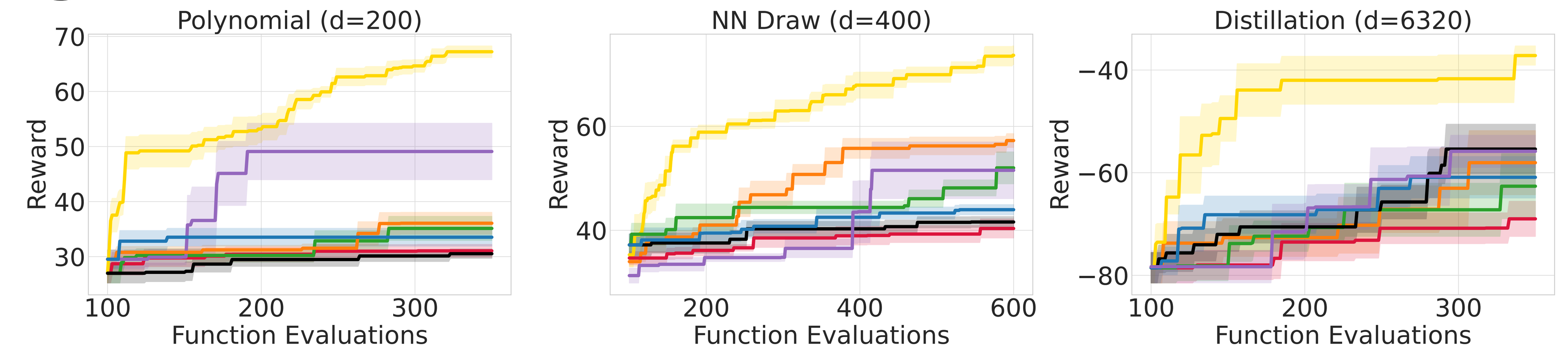
Real-World Datasets



→ Real-world datasets show mixed results

→ Deep ensembles and SGHMC often underperform

High-dimensional Datasets



→ I-BNNs outperform other surrogates across many settings

Further Empirical Results

Read our paper for additional results such as:

- Neural architecture search can improve the performance of BNNs
- Deep ensembles may underperform due to low model diversity
- I-BNNs consistently outperform GPs in high dimensions
- The optimal combination of hyperparameter selection method and kernel choice for GPs is problem-dependent