An Open-Source Training Dataset for Foundation Models for Hyperparameter Optimization

Aaron Klein[1], Herilalaina Rakotoarison[2], Luca Thale-Bombien[3], David Salinas[1,4]

[1]ELLIS Institute Tübingen [2]University of Helsinki [3]Leipzig University, ScaDS.AI [4]Prior Labs

NeurIPS 2026 · Evaluations and Datasets Track

Motivation

- **In practice,** hyperparameter optimization methods require careful tuning of their own hyperparameters and no method performs best on all benchmarks.
- **Foundation models** showed great promise in related fields such as tabular prediction or time-series forecasting.
- **OptFormer** by Chen et. al. showed that an autoregressive model can learn the sampling distribution of different optimization methods.
- **Contributions:**
  - **Dataset:** We provide the first large-scale open-source dataset for hyperparameter optimization consisting of 557,100 optimization trajectories from 6 different optimizers on 3,095 HPO tasks and global optimization problems.
  - **Scaling:** We systematically analyze how well decoder-only transformers can imitate state-of-the-art hyperparameter optimization methods when parameter count and token budget are scaled.

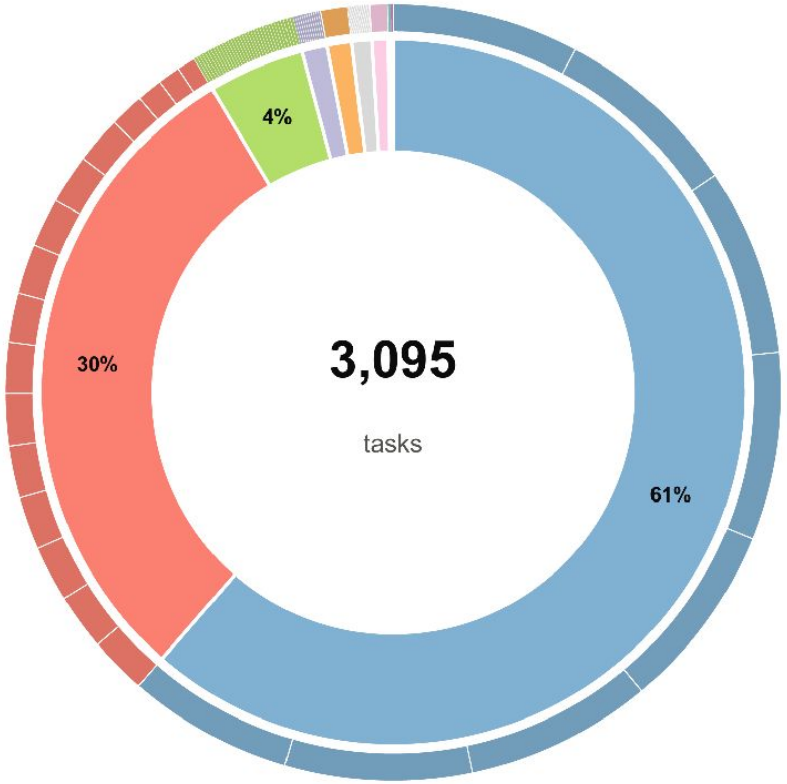
Foundation Models for Hyperparameter Optimization

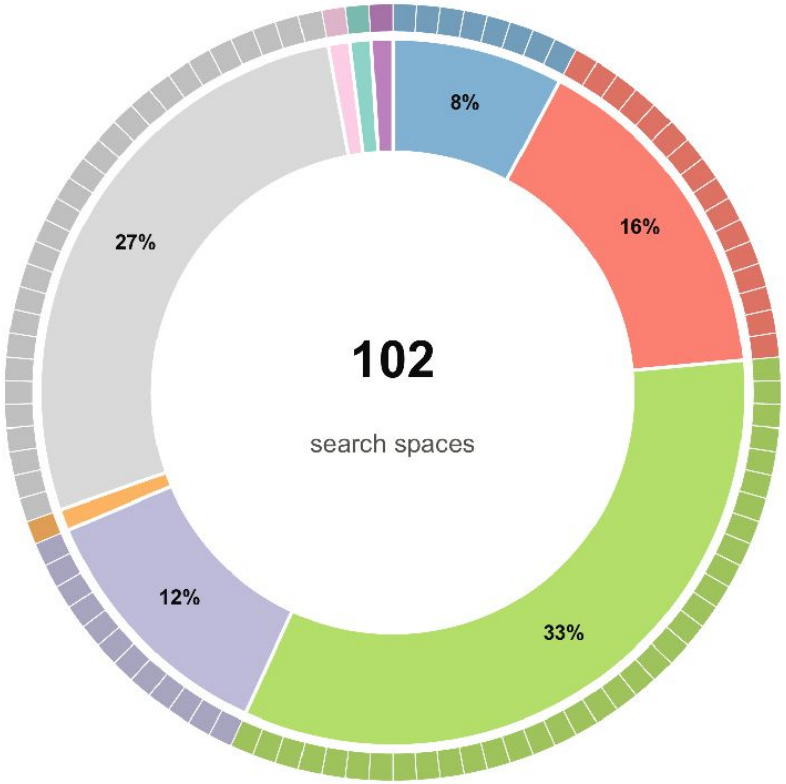
Optimization trajectory

$$\mathbf{x}_t \sim p\left(\mathbf{x} \middle| m, \mathbf{x}_0, y_0, \ldots, \mathbf{x}_{t-1}, y_{t-1}\right)$$

Meta-information

BBO-Pile: 557,100 optimization runs, fully open





	tasks	spaces
TabRepo	1,900	8
HPO-B	930	16
FC-Net (masked)	136	34
NAS-201 (masked)	36	12
LC-Bench	35	1
Global Opt.	28	28
PD1	23	1
FC-Net	4	1
NAS-201	3	1

557K trajectories **2.5B** tokens **6** optimizers × 30 seeds × 100 evals **50,595** CPU hours

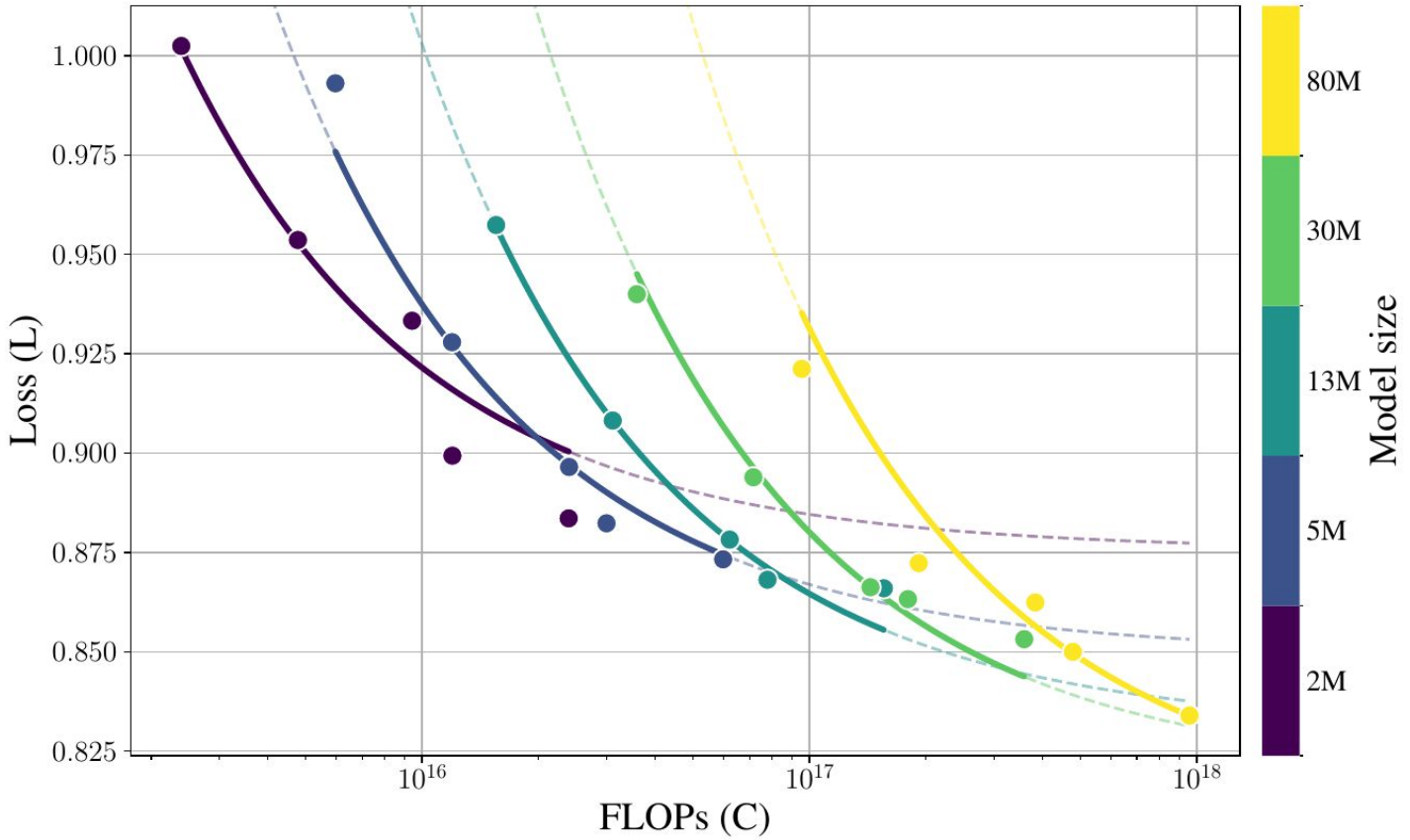
Optimization runs become token sequences

```
<algorithm>:RS
<type>:<UNI>,<min_value>:0.01,<max_value>:1.0,<log-scale>&
<type>:<UNI>,<min_value>:1,<max_value>:5,<linear-scale>&
<type>:<CATEGORICAL>,<categories>:[0, 1]
120,200,<1>*300|60,50,<0>*200|
```

Search Space: {"dropout": $Log\mathcal{U}(0.01, 1.0)$, "alpha": $\mathcal{U}(1.0, 5.0)$, "reg": {"l1", "l2"} }

$\boldsymbol{x}$ = {"dropout": 0.12, "alpha": 3.0, "reg": "l2"} $y = 0.3$ $\xrightarrow{\phi}$ $\xleftarrow{\phi^{-1}}$ s = "120,200,<1>*300|"

More tokens help more than more parameters



Chinchilla-style fit

$L = 0.79 + 21.4\, N^{-0.38} + 9.0{\cdot}10^{4}\, D^{-0.71}$

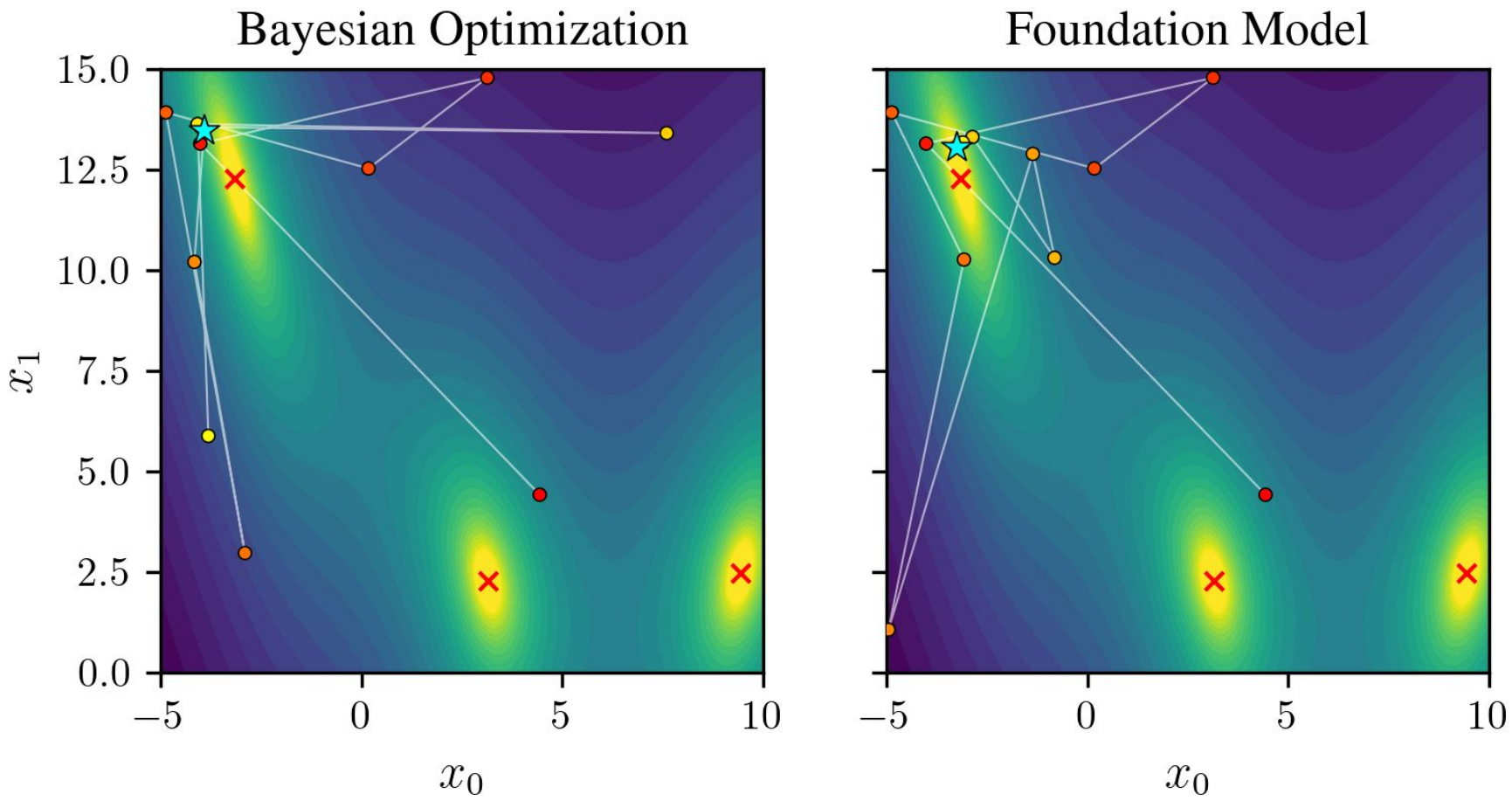
β > α: more data reduces the loss faster than more parameters

Caveat

Only ~2.6 orders of magnitude of compute and a shallow compute exponent of 0.0157

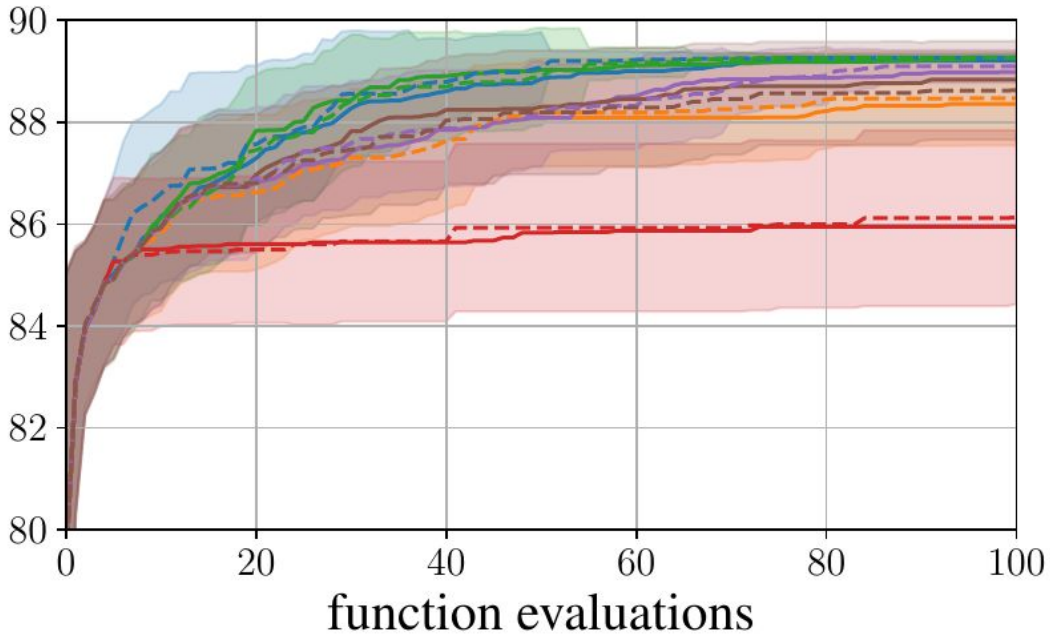
Dots: final validation loss of each (N, D) run. Lines: fitted L(N, D) per model size; dashed lines extrapolate.

Imitating Bayesian Optimization on a 2D function

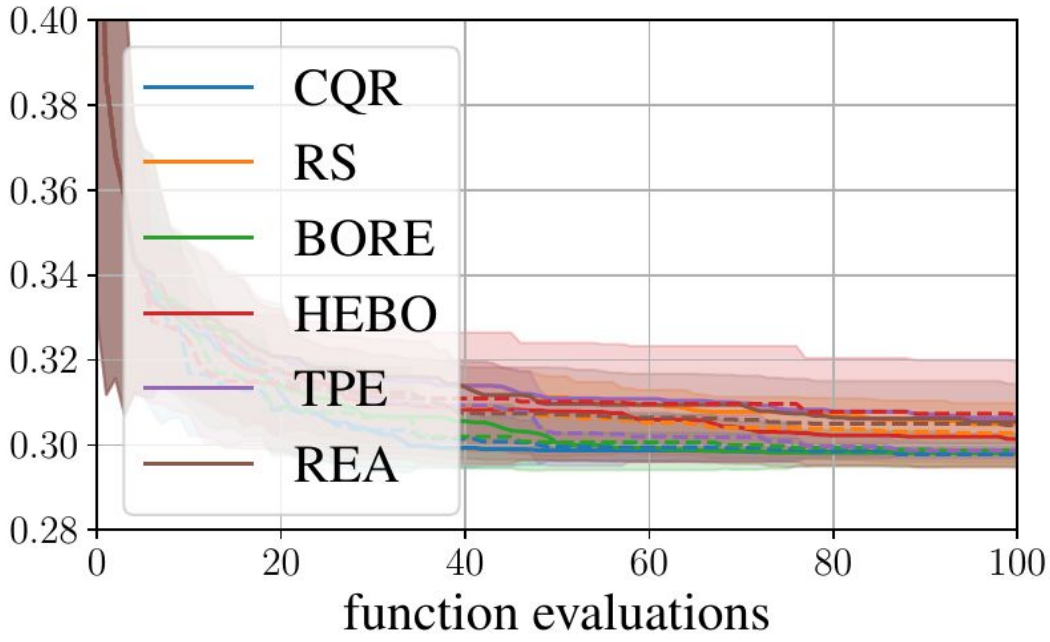


The model closely imitates its teachers

Seen search space: LC-Bench (accuracy ↑)



Unseen search space: TabRepo (error ↓)



Dashed: 80M model prompted to imitate each optimizer. Solid: original optimizer. Mean ± std over 30 seeds.

Takeaways

- **Open resource:** BBO-Pile, the first large-scale open dataset of optimization trajectories, released with training code and all model checkpoints
- **Imitation works:** transformers trained on BBO-Pile closely reproduce six optimizers, also on unseen search spaces and an unseen benchmark family
- **Scaling:** more training tokens help more than more parameters, but the gains saturate beyond ~1B tokens
- **Fast inference:** ~0.5 s per proposed configuration with vLLM
- **Limitations:** focus on hyperparameter optimization, and the models are bounded by their teachers
- **Next steps:** post-training and RL to go beyond the teachers, plus more domains and optimizers

Code: github.com/syne-tune/bbo-pile

Data and models: huggingface.co/datasets/synetune/bbo-pile