End to end optimization for data science in the wild

A. Kontaxakis, D. Sacharidis, A. Abelló, S. Nadal, A. Simitsis

EDBT 2025 PhD Workshop

Summary. Our work focuses on minimizing the cost of developing machine learning solutions (Exploratory (**EML**) or Automated (**AutoML**) Machine Learning). We can minimize cost by 1) optimizing the execution plan 2) selecting which solutions are the most promising. Our results show that by leveraging reuse, equivalence, and pipeline selection techniques we can achieve up to 10× more cost-effective ML development

Minimizing Cost

We can formulate the previous goals as optimization problems:

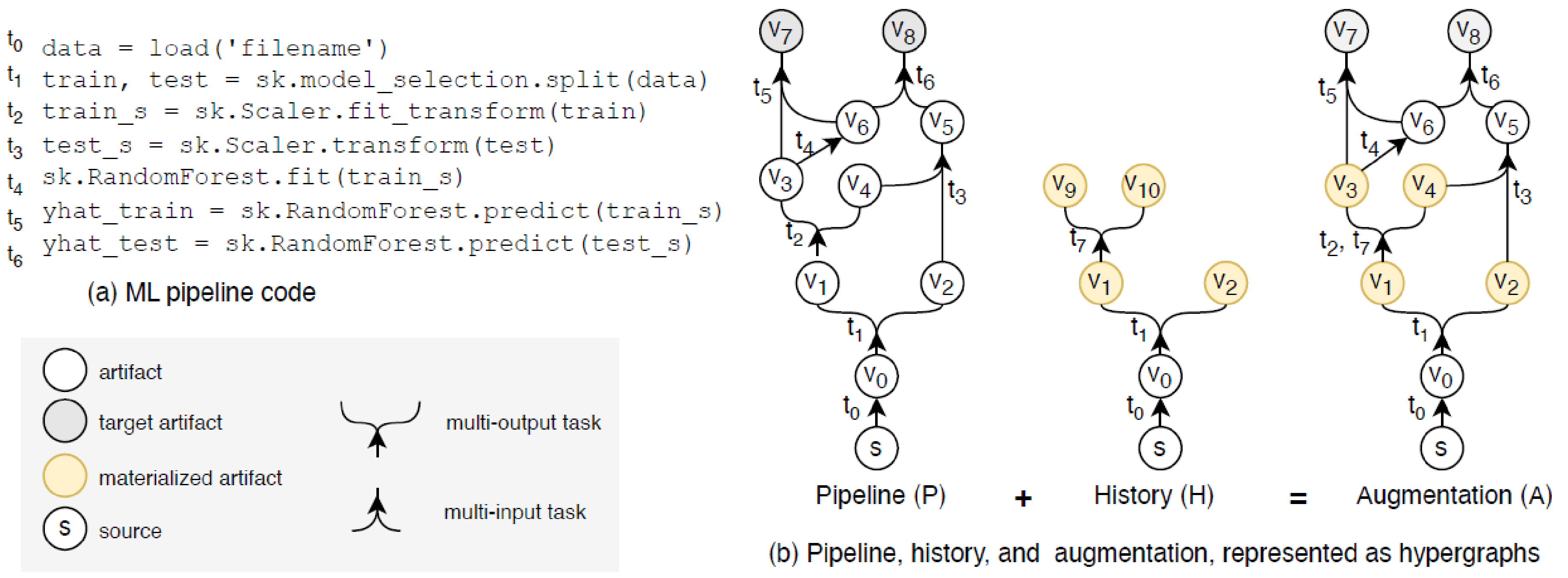
1. (**optimal plan**) Given a set of pipelines minimize their execution cost by exploiting reuse and equivalence

Sub-problem: (**Materialization**) after each execution of a pipeline select which results to store

2. (**Pipeline Selection**) Given a set of pipelines find the set of pipelines that maximize quality and minimize cost

Sub-problem: (**Quality-Cost Estimation**) given a set of pipeline use Historical knowledge to estimate their expected quality and cost

A Novel Pipeline Representation



```
t0 data = load('filename')
t1 train, test = sk.model_selection.split(data)
t2 train_s = sk.Scaler.fit_transform(train)
t3 test_s = sk.Scaler.transform(test)
t4 sk.RandomForest.fit(train_s)
t5 yhat_train = sk.RandomForest.predict(train_s)
t6 yhat_test = sk.RandomForest.predict(test_s)
```

(a) ML pipeline code

(b) Pipeline, history, and augmentation, represented as hypergraphs

Our Approach

Given a set of pipelines, CAPS:

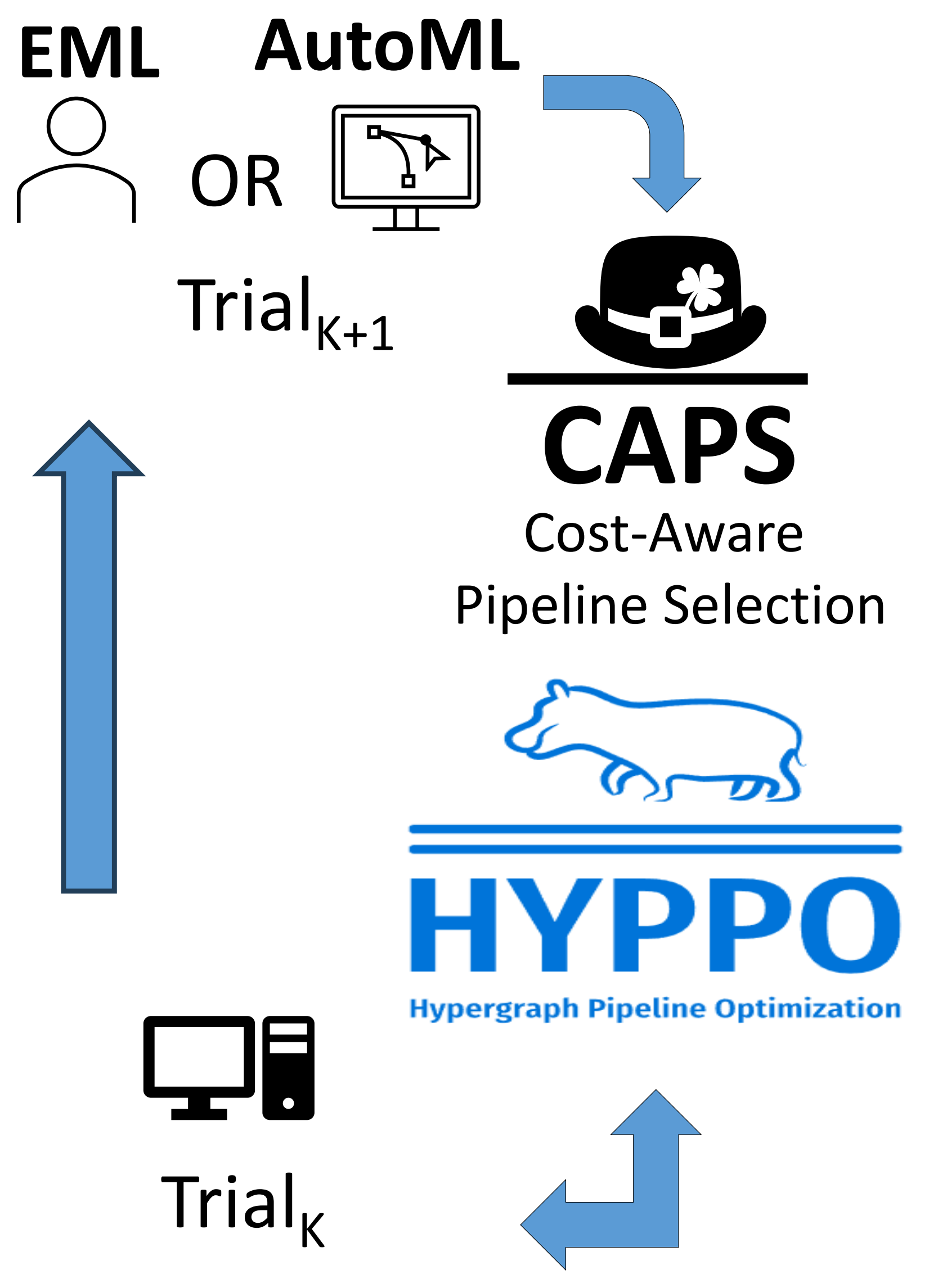
(a) **estimates** the **cost** and **performance** of unseen pipelines

(b) selects a set of pipelines with the best **trade-off** between quality and cost

Given a pipeline code, HYPPO:

(a) searches for an **optimized execution plan**

(b) decides what artifacts to **materialize**



HYPPO: https://github.com/akontaxakis/HYPPO

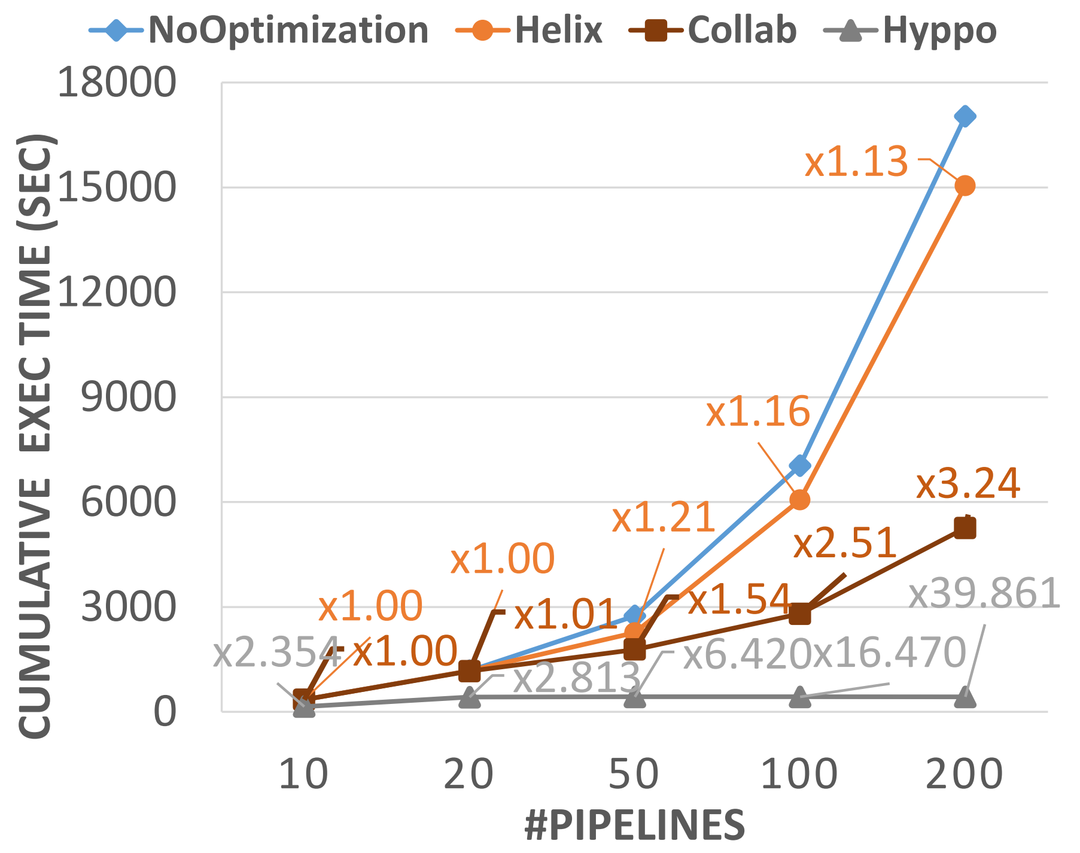
antonios.kontaxakis@ulb.be

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955895.

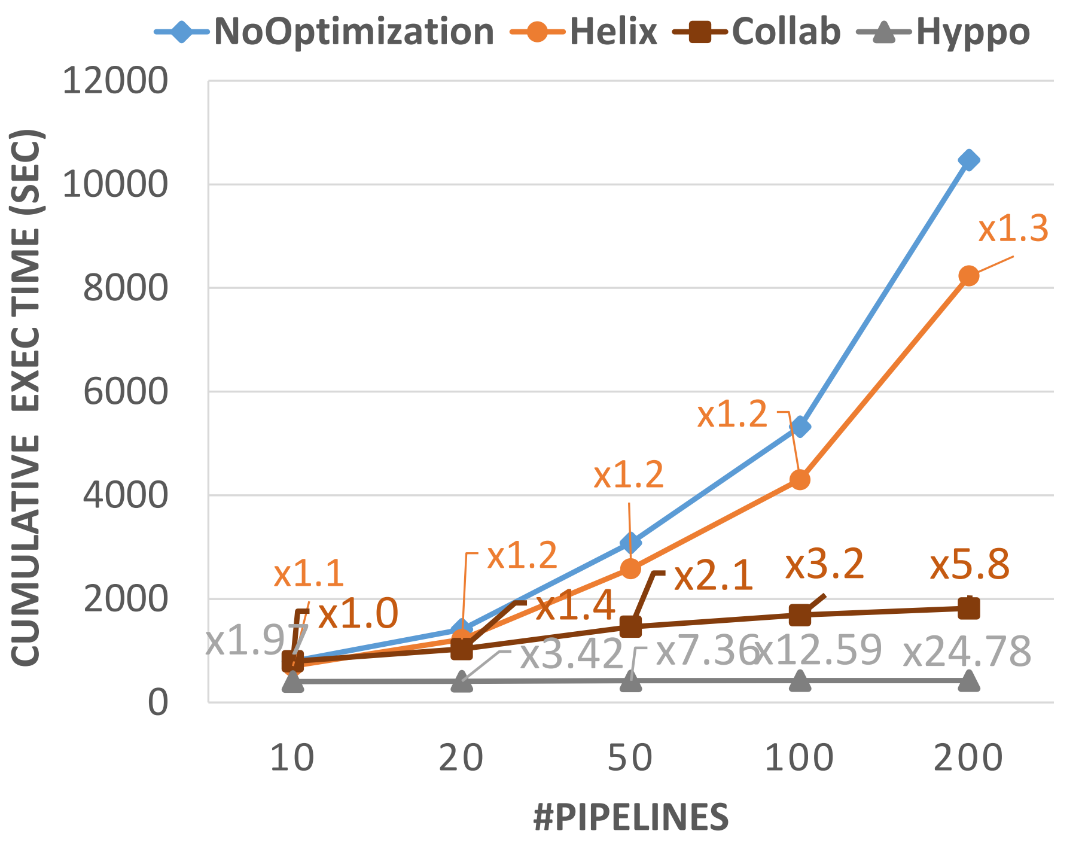
Evaluation (Optimal Plan)

Method	None	Sharing	Reuse	Materialization	Equivalence
NoOptimization	○				
Sharing		○			
Helix [VLDB'18]		○	○	◖	
Collab [SIGMOD'20]		○	◖	○	
HYPPO[ICDE'24]		○	○	○	○

HIGGS



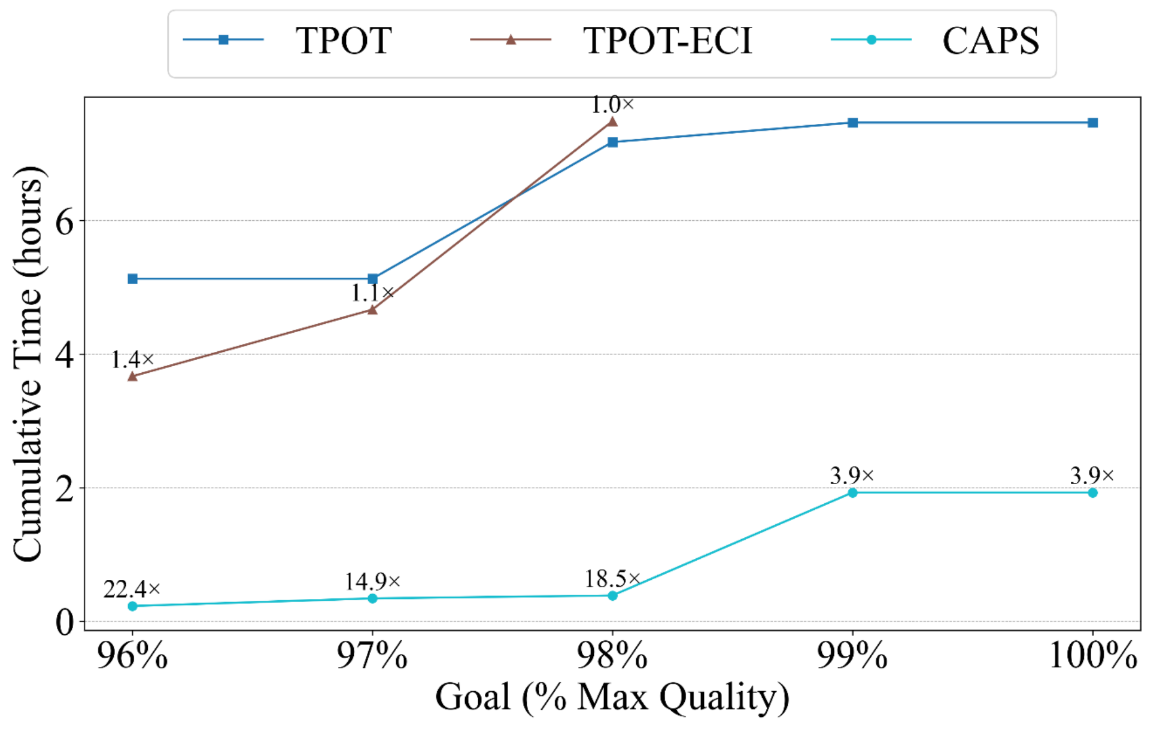
TAXI



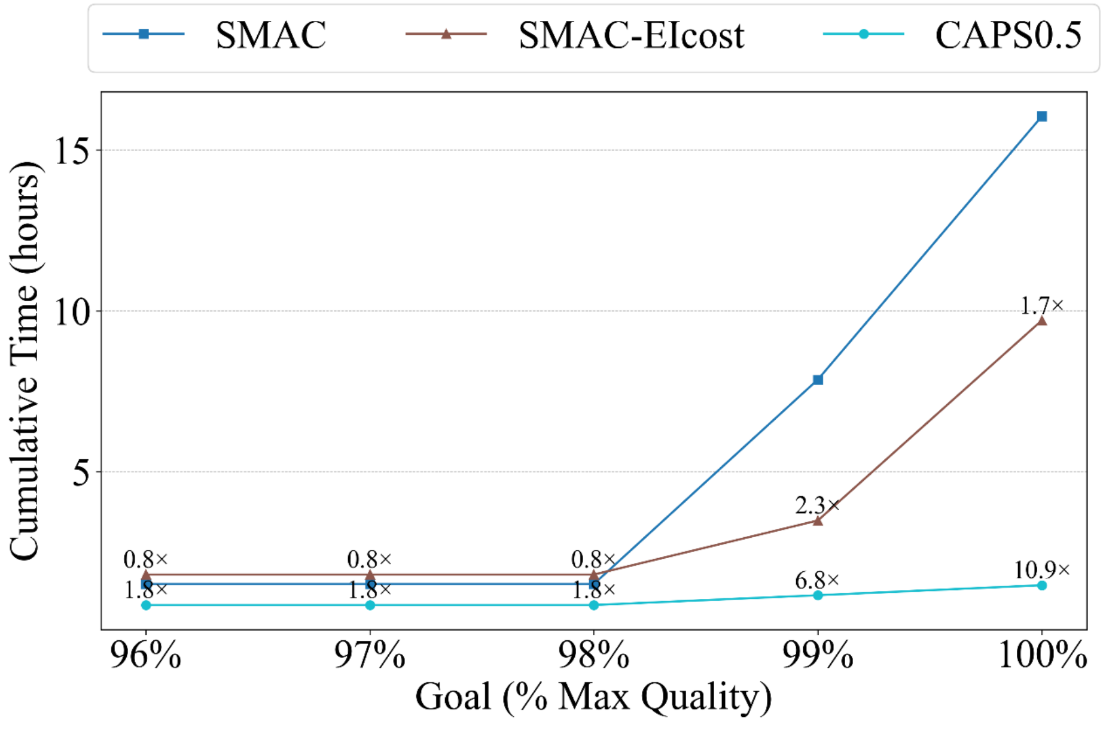
Evaluation (Pipeline Selection)

Method	Search Strategy	Quality Estimation	Cost Estimation
TPOT	Genetic	None	None
SMAC	Bayesian	EI (Model)	None
TPOT_ECI [MLsys'21]	Genetic	History	History
SMAC_EIperSec [NIPS'12]	Bayesian	EI (Model)	Model
Our Approach (CAPS)	Genetic or Bayesian	Model or History	History or Meta learning

Genetic search



Bayesian Search



Contributions

[ICDE'24] HYPPO: Using Equivalences to Optimize Pipelines in Exploratory Machine Learning

A novel representation for ML pipelines

An optimization and materialization algorithm

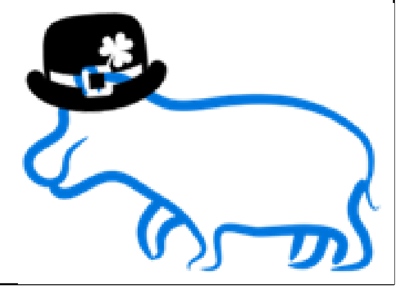
[EDBT'25] HYPPO: Efficient Discovery and Execution of Data Science Pipelines in Collaborative Environments

An API for retrieving and optimizing pipelines

[TBS] CAPS: Cost-Awareness ML Pipeline Selection

A pipeline selection algorithm

A novel cost estimator for ML pipelines



References

[VLDB 18] D. Xin, S. Macke, L. Ma, J. Liu, S. Song, and A. Parameswaran, "HELIX: Holistic optimization for accelerating iterative machine learning"

[SIGMOD 20] B. Derakhshan, A. Rezaei Mahdiraji, Z. Abedjan, T. Rabl, and V. Markl, "Optimizing machine learning workloads in collaborative environments"

[ICDE'24] Antonios I. Kontaxakis, Dimitris Sacharidis, Alkis Simitsis, Alberto Abelló, and Sergi Nadal. "HYPPO: Using Equivalences to Optimize Pipelines in Exploratory Machine Learning"

[MLsys'21] Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. 2021. "FLAML: A Fast and Lightweight AutoML Library"

[NIPS'12]_Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. "Practical Bayesian Optimization of Machine Learning Algorithms"