A Journey of Tabular Benchmarks: Lessons in Method Comparison and Curation

Oxford ML School

David Salinas. Aug 2025.

ellis INSTITUTE TÜBINGEN

OPEN EURO LLM

universität freiburg

MENU DU JOUR

"A Journey of Tabular Benchmarks: Lessons in Curation and Method Comparison"

ENTRÉES

- **TabRepo - A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications** **(25 min)**
  A carefully curated appetizer to stimulate your appetite for comprehensive tabular data evaluation

MENU PRINCIPAL

- **TabArena: A Living Benchmark for Machine Learning on Tabular Data** **(50 min)**
  Our signature dish - a robust and evolving benchmark that will satisfy your hunger for rigorous evaluation

DESSERT

- **A Delicious Case for Openness** **(5 min)**
  A sweet case promoting transparency and collaborative building of LLMs

❦

Questions can be asked throughout all the talk!

We will also keep ~10 minutes for discussion at the end.

❦

MENU DU JOUR

"A Journey of Tabular Benchmarks: Lessons in Curation and Method Comparison"

ENTRÉES

- **TabRepo - A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications** **(25 min)**
  A carefully curated appetizer to stimulate your appetite for comprehensive tabular data evaluation

MENU PRINCIPAL

- **TabArena: A Living Benchmark for Machine Learning on Tabular Data** **(50 min)**
  Our signature dish - a robust and evolving benchmark that will satisfy your hunger for rigorous evaluation

DESSERT

- **A Delicious Case for Openness** **(5 min)**
  A sweet case promoting transparency and collaborative building of LLMs

Questions can be asked throughout all the talk!

We will also keep ~10 minutes for discussion at the end.

Part I

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

Tabular prediction

Tabular prediction

- Tabular prediction: problem definition

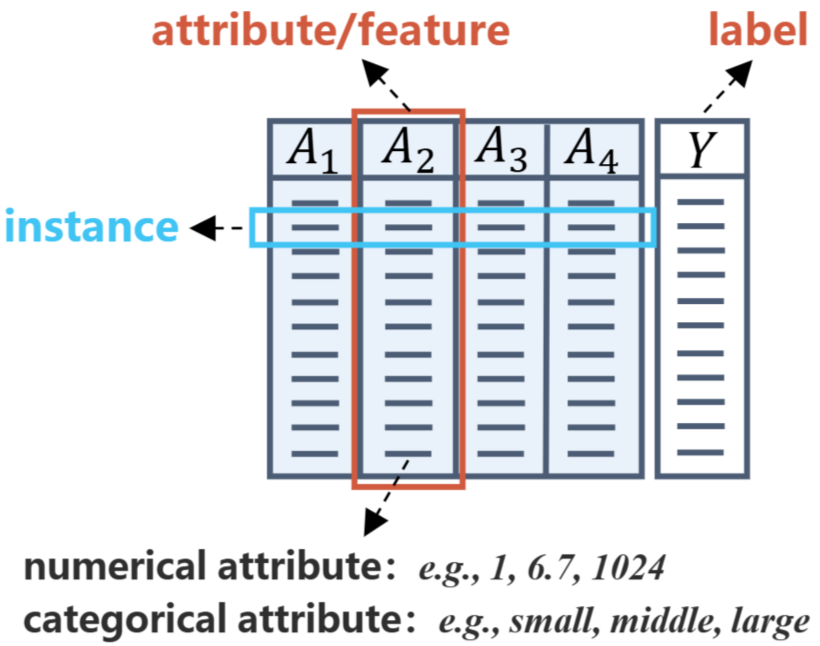
Tabular prediction

- Tabular prediction: problem definition
- A quick glance at the current SOTA tabular system: AutoGluon

Tabular prediction

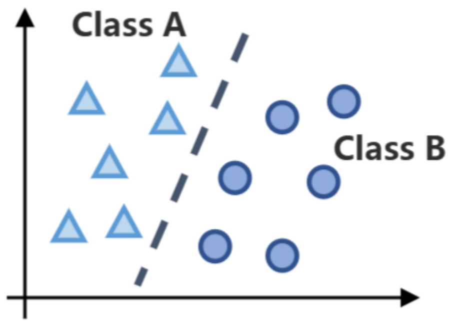
- Tabular prediction: problem definition
- A quick glance at the current SOTA tabular system: AutoGluon
- Improving AutoGluon with offline evaluations and portfolio (meta- learning

Tabular prediction



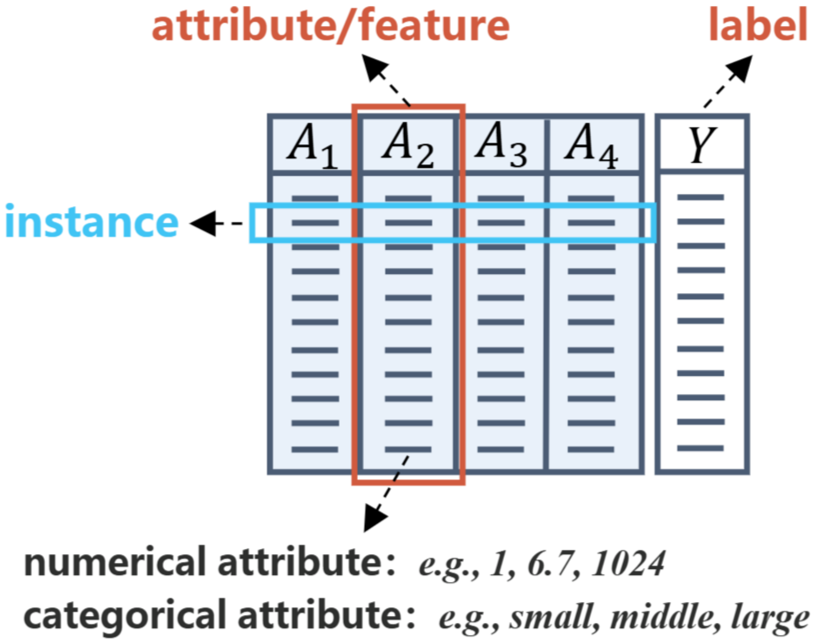

```python
import pandas as pd
from autogluon.tabular import TabularPredictor

df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('train.csv')

predictor = TabularPredictor(label='class').fit(df_train)
predictions = predictor.predict(df_test)
```

classification

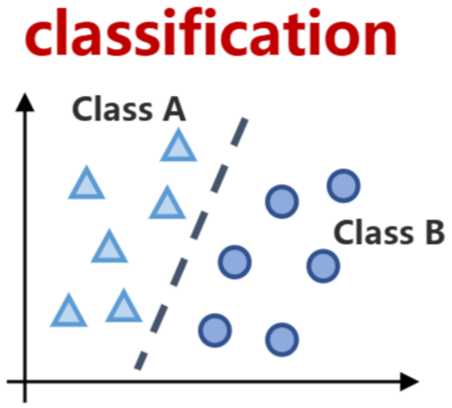


Tabular prediction

- Input: a training data frame, a target column and a training time budget

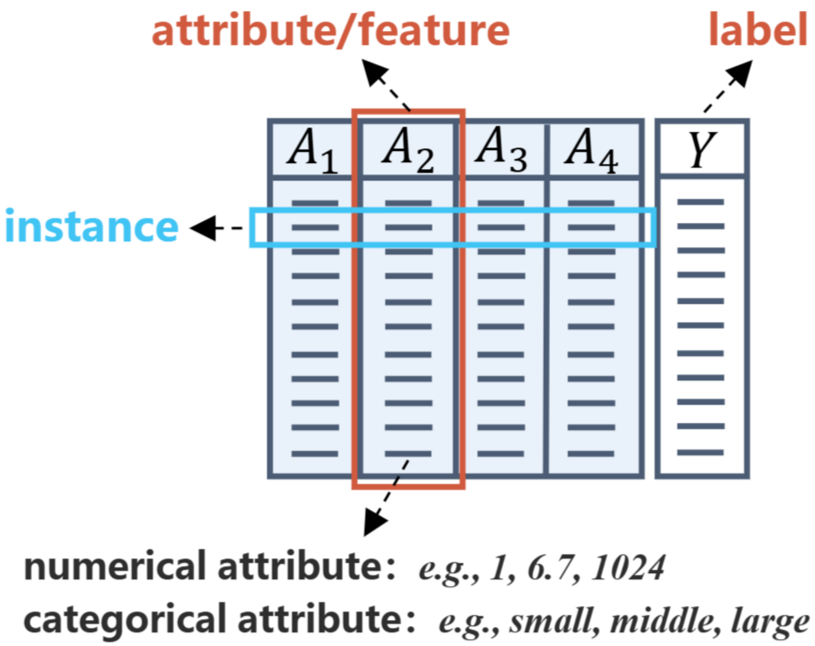



```python
import pandas as pd
from autogluon.tabular import TabularPredictor

df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('train.csv')

predictor = TabularPredictor(label='class').fit(df_train)
predictions = predictor.predict(df_test)
```

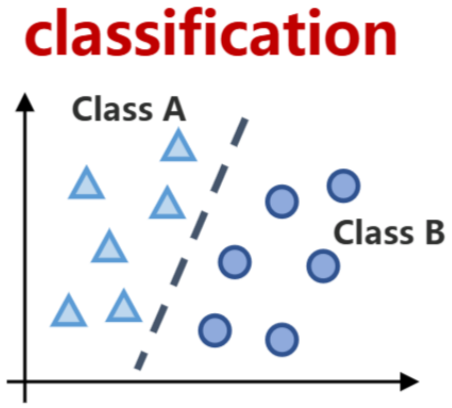


Tabular prediction

- Input: a training data frame, a target column and a training time budget
- Output: a predictor able to give predictions given a test dataframe



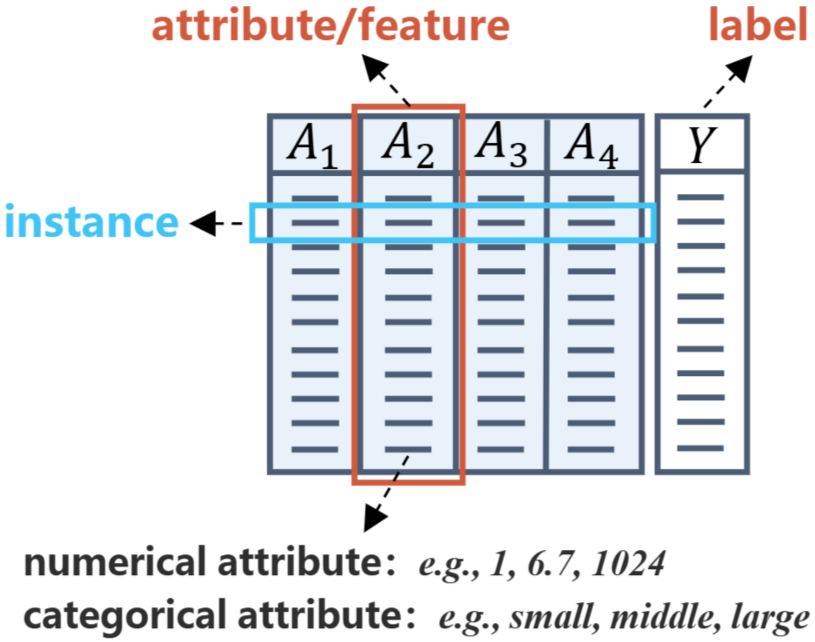

```python
import pandas as pd
from autogluon.tabular import TabularPredictor

df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('train.csv')

predictor = TabularPredictor(label='class').fit(df_train)
predictions = predictor.predict(df_test)
```



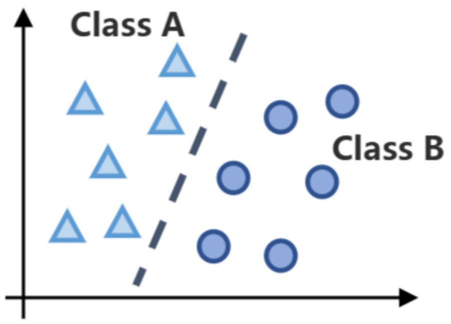
Tabular prediction

- Input: a training data frame, a target column and a training time budget
- Output: a predictor able to give predictions given a test dataframe
- Metrics:
    - RMSE (regression), log-prob (classification)
    - Prediction latency, memory, …

```
import pandas as pd
from autogluon.tabular import TabularPredictor

df_train = pd.read_csv('train.csv')
df_test = pd.read_csv('train.csv')

predictor = TabularPredictor(label='class').fit(df_train)
predictions = predictor.predict(df_test)
```



classification



What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

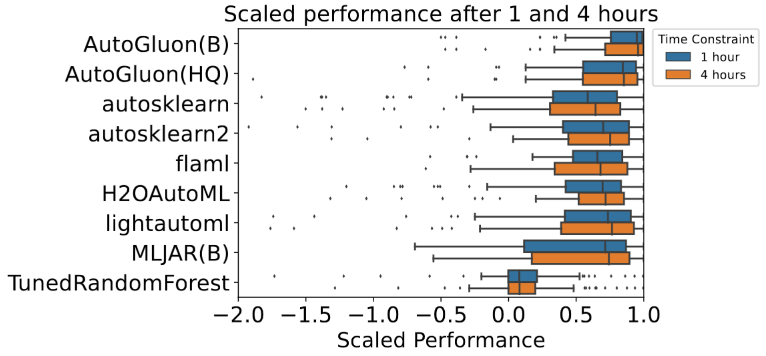


Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

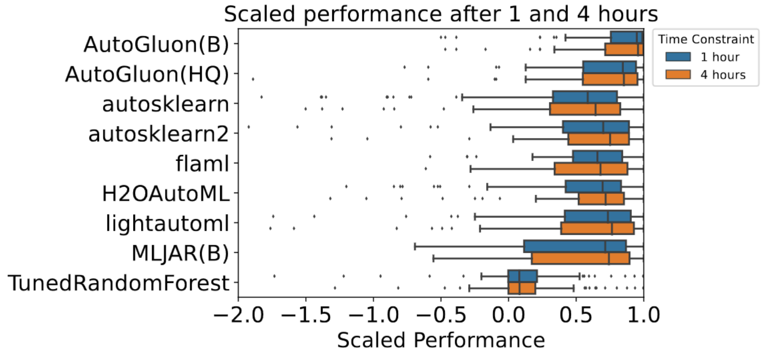


Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

- Considered 9 AutoML frameworks, evaluated on 1h and 4h fitting budget

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

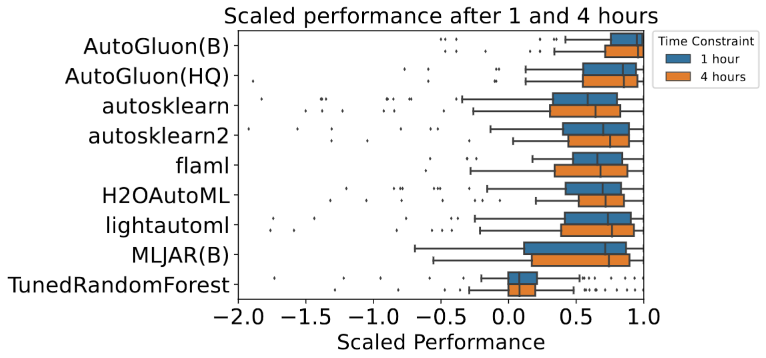


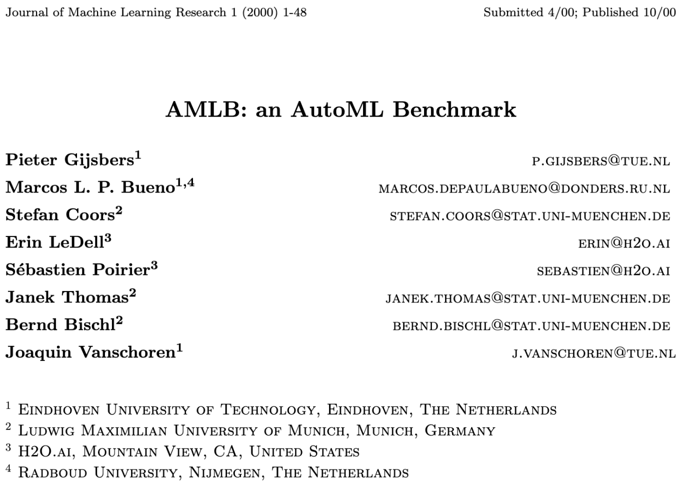
Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

- Considered 9 AutoML frameworks, evaluated on 1h and 4h fitting budget

- AutoGluon was then the best model by a large margin

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets


Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

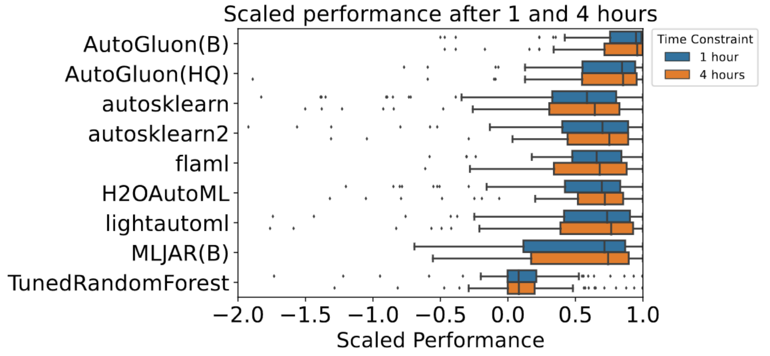


Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

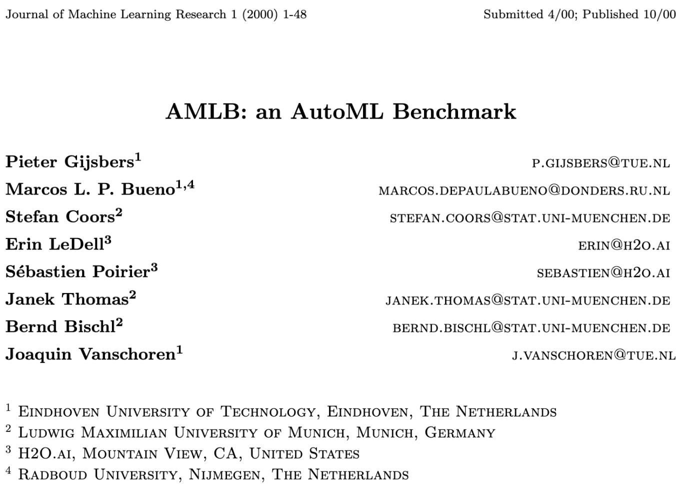
Evaluating a single method costs 40K CPU hours of compute!

- Considered 9 AutoML frameworks, evaluated on 1h and 4h fitting budget

- AutoGluon was then the best model by a large margin

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets



Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

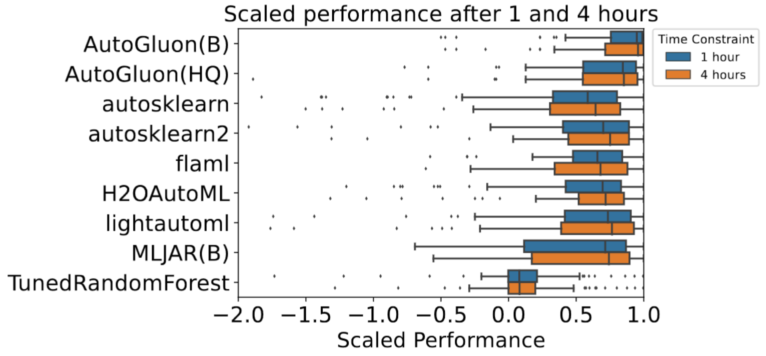


Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

Evaluating a single method costs 40K CPU hours of compute!

Can we limit this cost? 🤔

- Considered 9 AutoML frameworks, evaluated on 1h and 4h fitting budget

- AutoGluon was then the best model by a large margin

What is the best Tabular method?

- AutoML Benchmark [Ginsberg et al 2023] considered 71 classification and 33 regression datasets

Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

AMLB: an AutoML Benchmark

Pieter Gijsbers[1] P.GIJSBERS@TUE.NL
Marcos L. P. Bueno[1,4] MARCOS.DEPAULABUENO@DONDERS.RU.NL
Stefan Coors[2] STEFAN.COORS@STAT.UNI-MUENCHEN.DE
Erin LeDell[3] ERIN@H2O.AI
Sébastien Poirier[3] SEBASTIEN@H2O.AI
Janek Thomas[2] JANEK.THOMAS@STAT.UNI-MUENCHEN.DE
Bernd Bischl[2] BERND.BISCHL@STAT.UNI-MUENCHEN.DE
Joaquin Vanschoren[1] J.VANSCHOREN@TUE.NL

[1] Eindhoven University of Technology, Eindhoven, The Netherlands
[2] Ludwig Maximilian University of Munich, Munich, Germany
[3] H2O.ai, Mountain View, CA, United States
[4] Radboud University, Nijmegen, The Netherlands

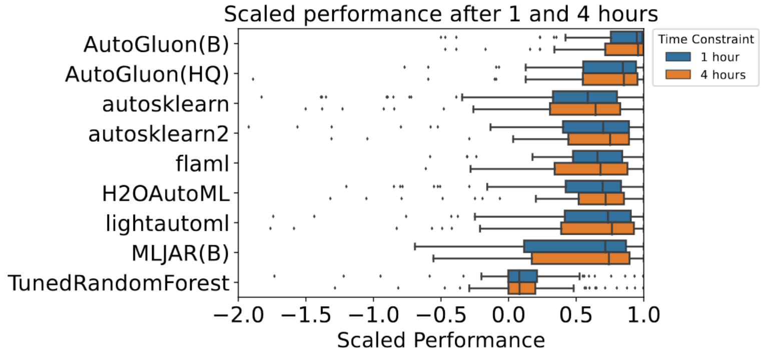


Figure 4: Scaled performance for each framework under different time constraints. Only frameworks which have evaluations on all tasks for both time constraints are shown. Performance generally does not improve much with more time.

Evaluating a single method costs 40K CPU hours of compute!

Can we limit this cost? 🤔

- Considered 9 AutoML frameworks, evaluated on 1h and 4h fitting budget

- AutoGluon How does this work? best model by a large margin

AutoGluon at a glance

AutoGluon at a glance

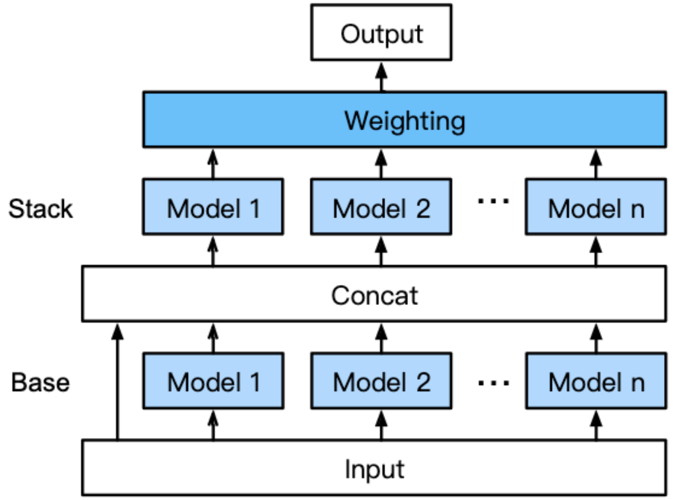


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

AutoGluon at a glance

- AutoGluon (1.1) recipe:

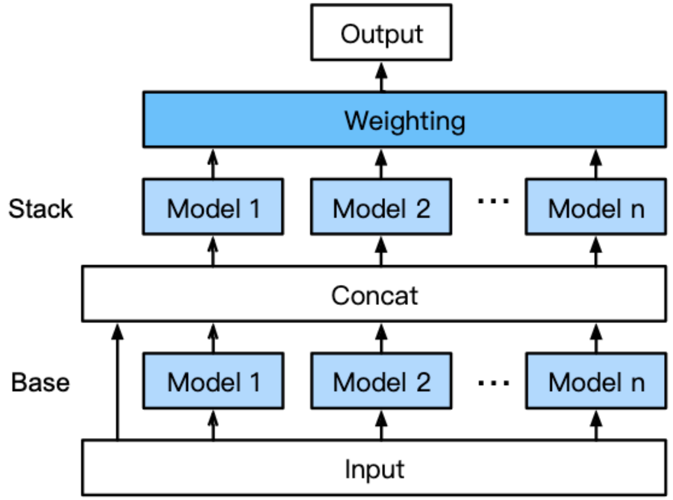


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

AutoGluon at a glance

- AutoGluon (1.1) recipe:
    - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*

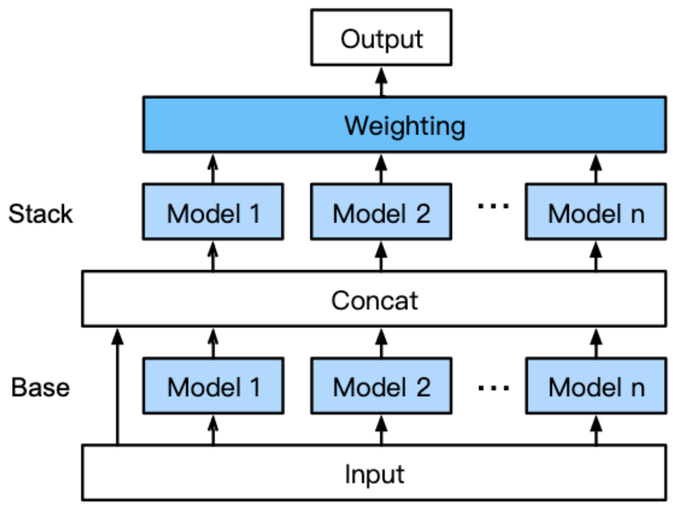


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

AutoGluon at a glance

- AutoGluon (1.1) recipe:
    - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
    - For each model, Autogluon performs **bagging with out of fold cross-validation**

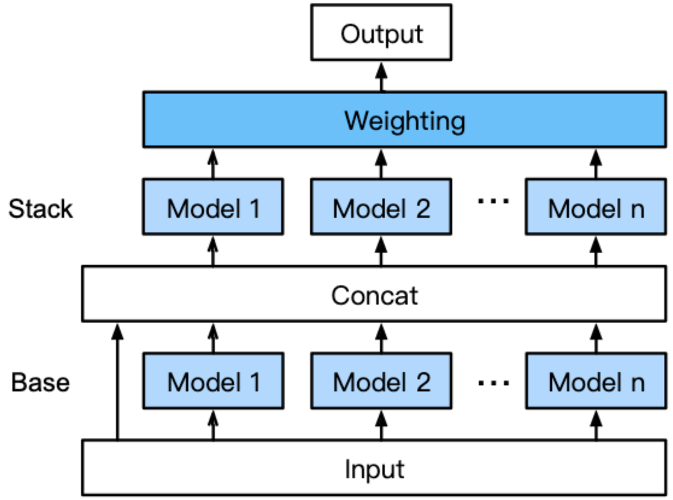


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

AutoGluon at a glance

- AutoGluon (1.1) recipe:
    - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
    - For each model, Autogluon performs **bagging with out of fold cross-validation**

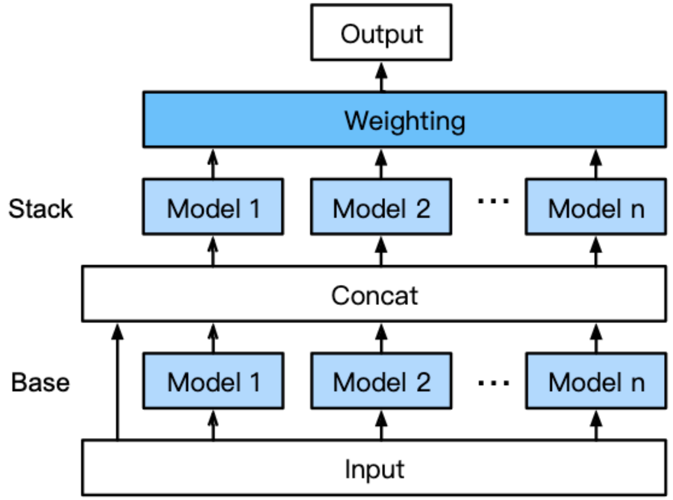
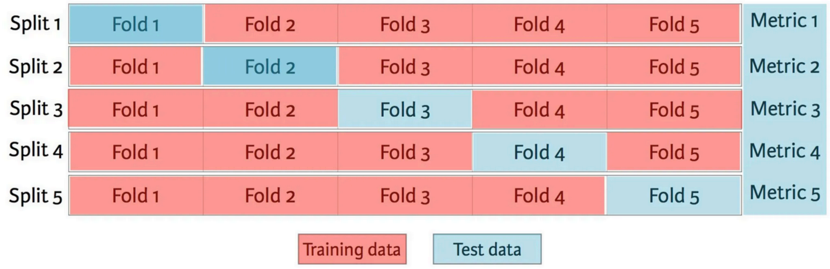


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.



Out of fold evaluation, image credit: data camp

AutoGluon at a glance

- AutoGluon (1.1) recipe:
  - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
  - For each model, Autogluon performs **bagging with out of fold cross-validation**
  - Each model is learned on 8 non-overlapping fold of the data and the predictions are averaged

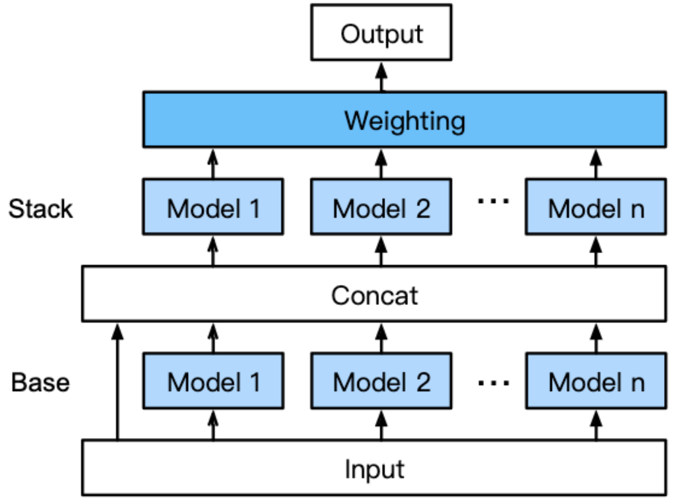
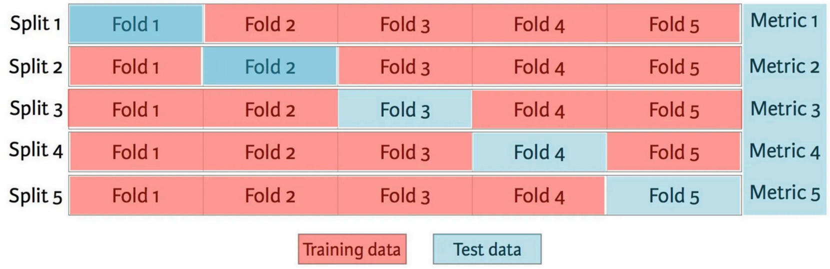


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.



Out of fold evaluation, image credit: data camp

AutoGluon at a glance

- AutoGluon (1.1) recipe:
  - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
  - For each model, Autogluon performs **bagging with out of fold cross-validation**
  - Each model is learned on 8 non-overlapping fold of the data and the predictions are averaged
  - Then perform ***stacking***: e.g. learn the models again while concatenating the predictions of the first *layer* with the original features

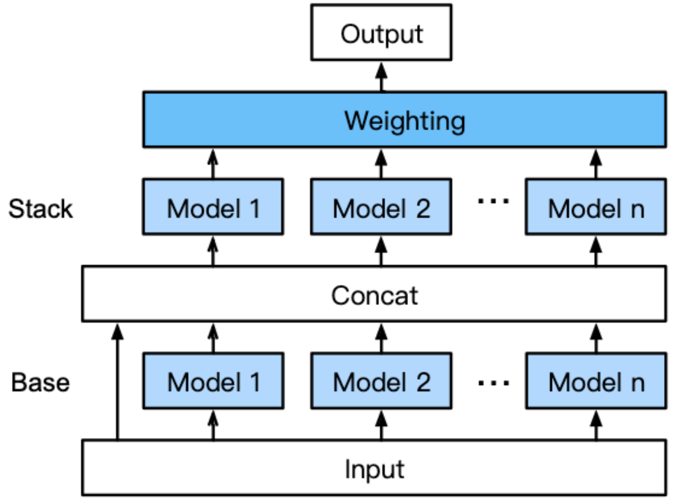
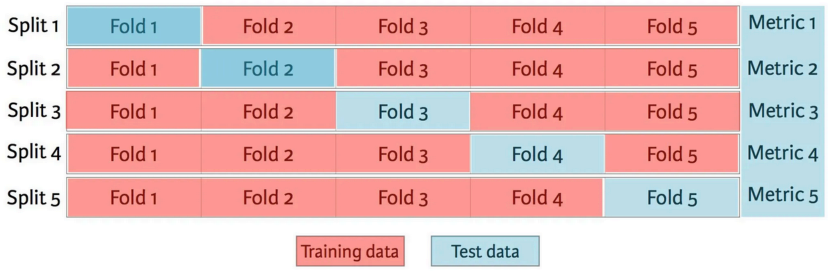


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.



Out of fold evaluation, image credit: data camp

AutoGluon at a glance

- AutoGluon (1.1) recipe:
  - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
  - For each model, Autogluon performs **bagging with out of fold cross-validation**
  - Each model is learned on 8 non-overlapping fold of the data and the predictions are averaged
  - Then perform ***stacking***: e.g. learn the models again while concatenating the predictions of the first *layer* with the original features
  - Then perform ***ensembling***: by estimating the weights on hold-out data (Caruana 2004) using validation scores

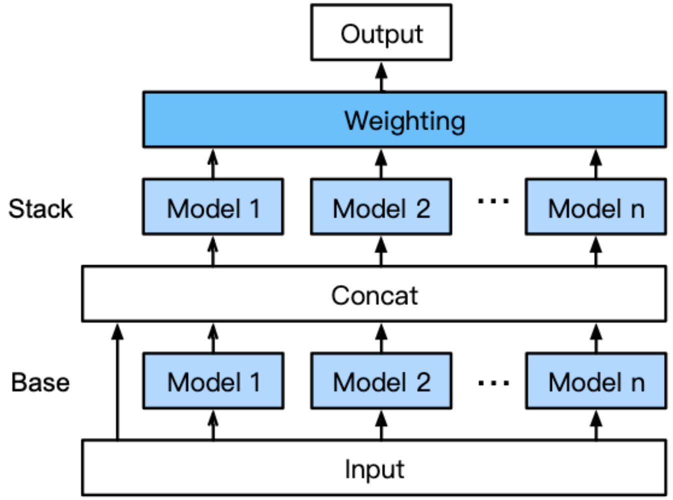
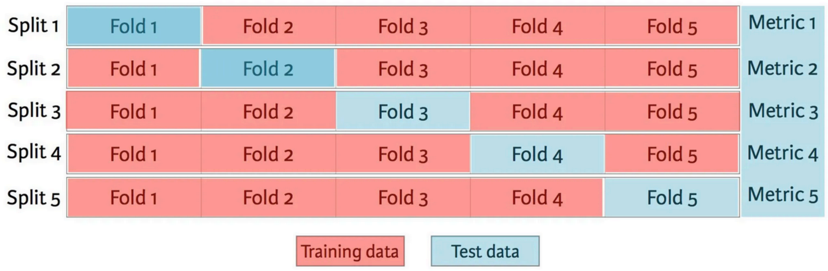


Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.



Out of fold evaluation, image credit: data camp

AutoGluon at a glance

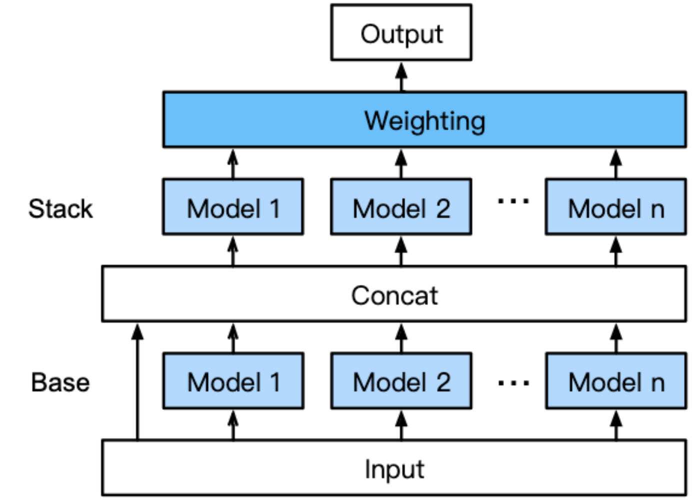
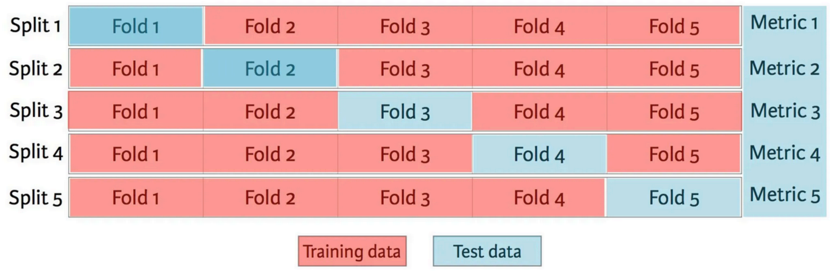


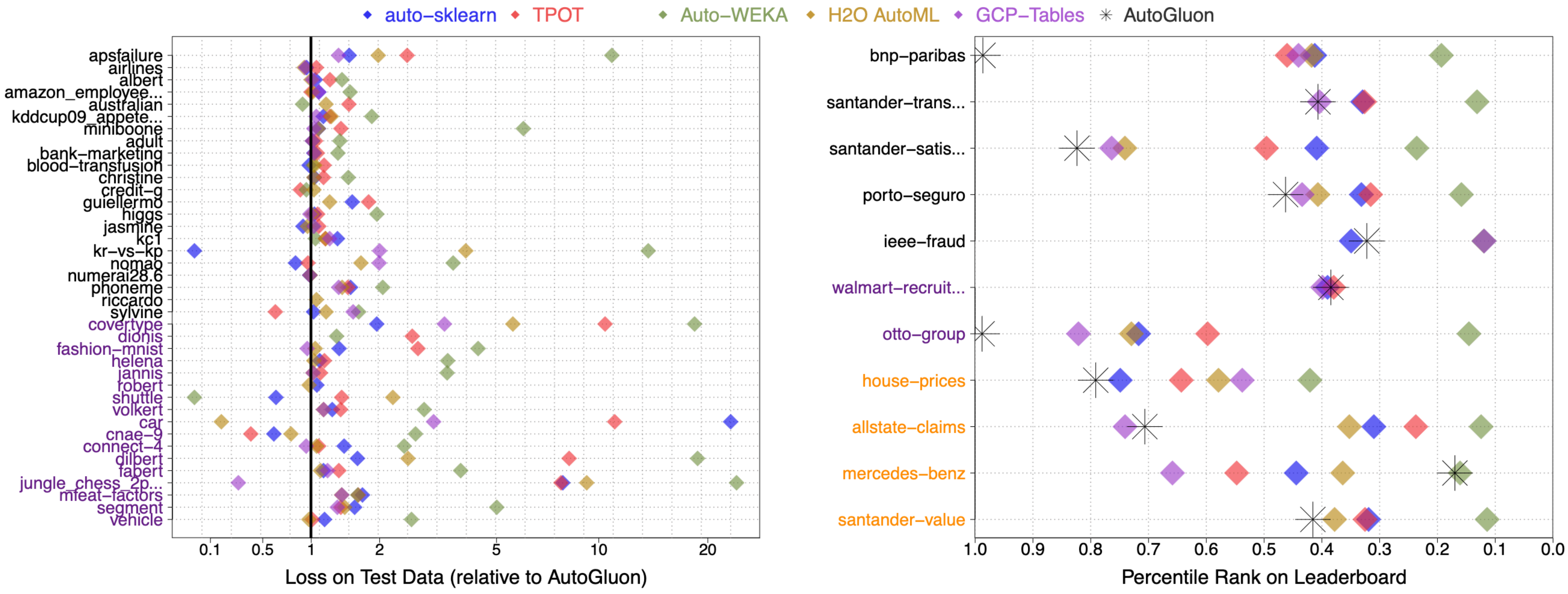
Figure 2. AutoGluon's multi-layer stacking strategy, shown here using two stacking layers and n types of base learners.

- AutoGluon (1.1) recipe:
  - Runs 13 models (KNN, linear, Catboost, LightGBM, MLPs, RandomForest, …) in a first *layer*
  - For each model, Autogluon performs **bagging with out of fold cross-validation**
  - Each model is learned on 8 non-overlapping fold of the data and the predictions are averaged
  - Then perform ***stacking***: e.g. learn the models again while concatenating the predictions of the first *layer* with the original features
  - Then perform ***ensembling***: by estimating the weights on hold-out data (Caruana 2004) using validation scores
- Let us take a look!



Out of fold evaluation, image credit: data camp

What is the best Tabular method?

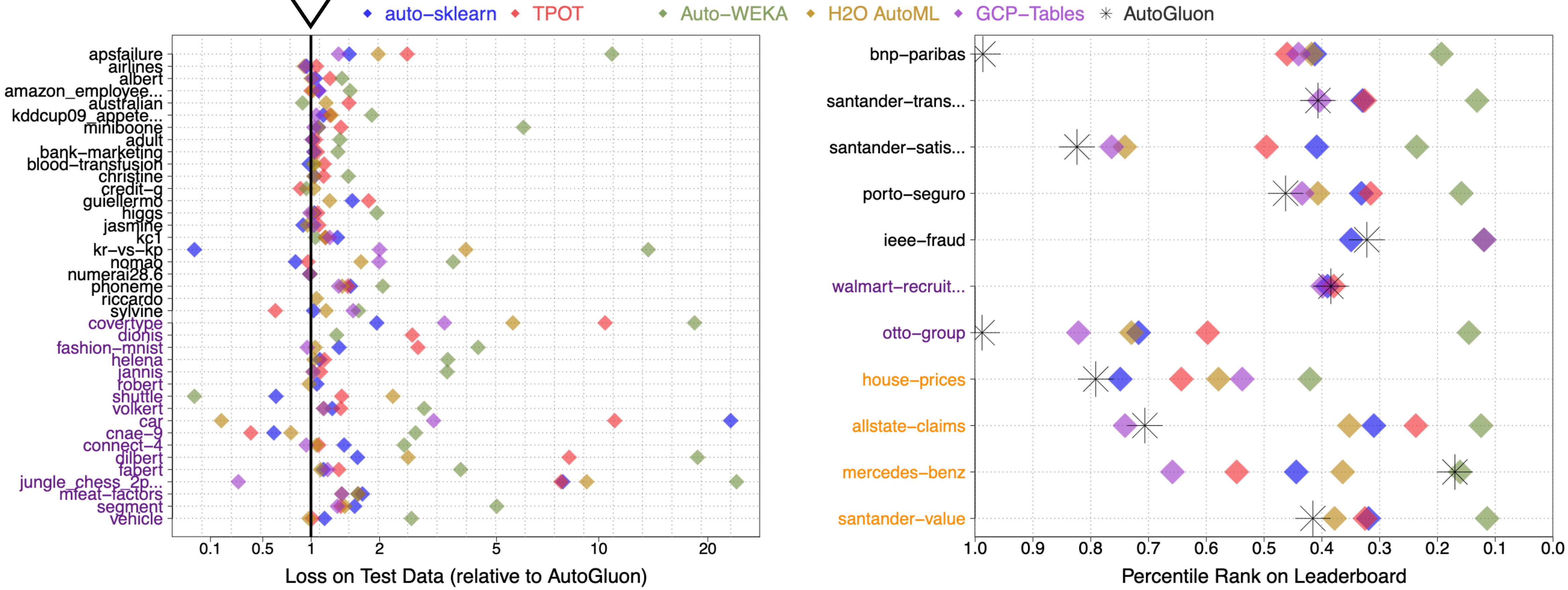


(A) AutoML Benchmark (1h) **(B)** Kaggle Benchmark (4h)

Erickson & Mueller et al 2020

What is the best Tabular method?

Better than all frameworks most of the time

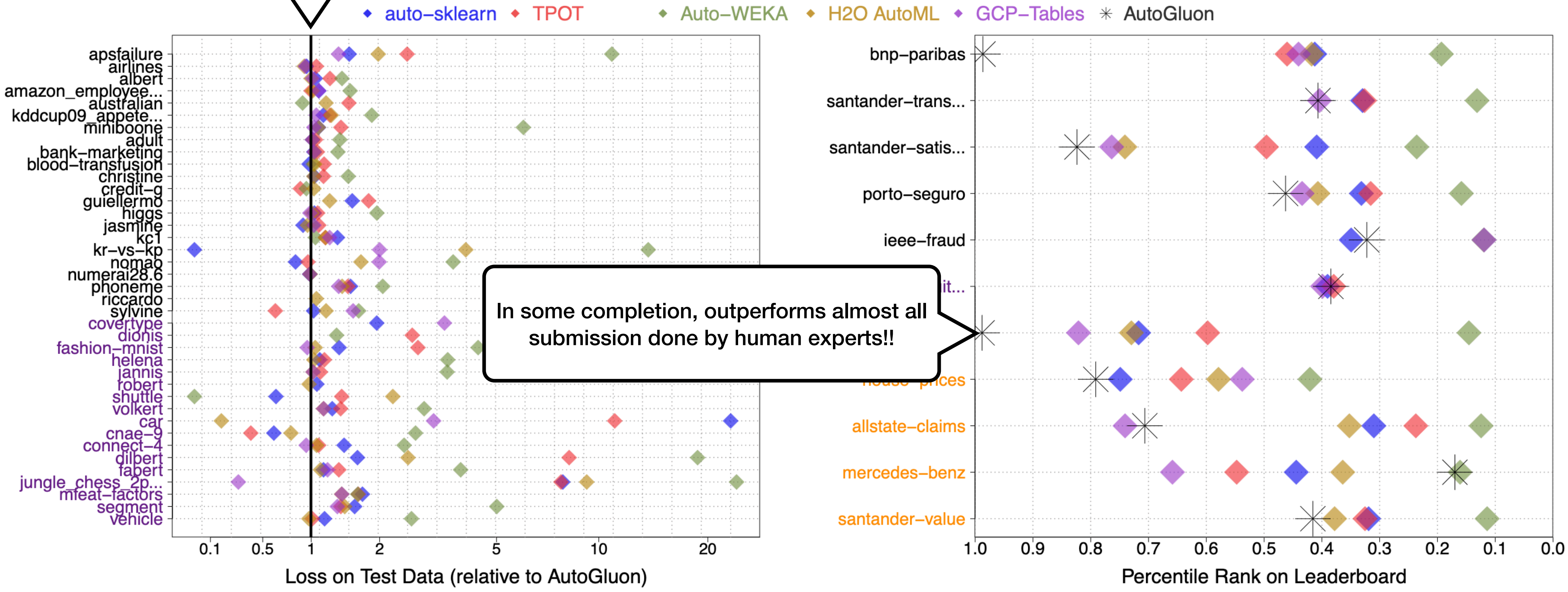


(A) AutoML Benchmark (1h) **(B)** Kaggle Benchmark (4h)

Erickson & Mueller et al 2020

What is the best Tabular method?

Better than all frameworks most of the time



(A) AutoML Benchmark (1h) **(B)** Kaggle Benchmark (4h)

Erickson & Mueller et al 2020

AutoGluon

Hyperparameter Optimization (HPO)

AutoGluon

Hyperparameter Optimization (HPO)

- Strikingly, AutoGluon achieved state-of-the-art results **without** HPO with its mix of bagging, stacking, ensembling and good heuristic featurizers

AutoGluon

Hyperparameter Optimization (HPO)

- Strikingly, AutoGluon achieved state-of-the-art results **without** HPO with its mix of bagging, stacking, ensembling and good heuristic featurizers
- It is not that HPO does not help, it does but compute is better spent evaluating a good set of default models (with more folds, more rounds, etc)

AutoGluon

Hyperparameter Optimization (HPO)

- Strikingly, AutoGluon achieved state-of-the-art results **without** HPO with its mix of bagging, stacking, ensembling and good heuristic featurizers
- It is not that HPO does not help, it does but compute is better spent evaluating a good set of default models (with more folds, more rounds, etc)
- AutoGluon default models: 13 default hyperparameters chosen manually by experts

AutoGluon

Hyperparameter Optimization (HPO)

- Strikingly, AutoGluon achieved state-of-the-art results **without** HPO with its mix of bagging, stacking, ensembling and good heuristic featurizers
- It is not that HPO does not help, it does but compute is better spent evaluating a good set of default models (with more folds, more rounds, etc)
- AutoGluon default models: 13 default hyperparameters chosen manually by experts
- Can we do better by automating this?

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

TabRepo

- Goals:

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
    - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
    - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
    - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
  - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)
  - 200 random configurations of models used in AutoGluon (CatBoost, MLP, LightGBM, RandomForest, …) on all datasets with 3 seeds

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
  - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)
  - 200 random configurations of models used in AutoGluon (CatBoost, MLP, LightGBM, RandomForest, …) on all datasets with 3 seeds
- Performance metrics (latency, accuracy, …) **and predictions** available for every dataset, model, seed

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
  - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)
  - 200 random configurations of models used in AutoGluon (CatBoost, MLP, LightGBM, RandomForest, …) on all datasets with 3 seeds
- Performance metrics (latency, accuracy, …) **and predictions** available for every dataset, model, seed
- ~100GB of data, ~200K CPU hours of compute

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
    - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
    - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
    - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)
    - 200 random configurations of models used in AutoGluon (CatBoost, MLP, LightGBM, RandomForest, …) on all datasets with 3 seeds
- Performance metrics (latency, accuracy, …) **and predictions** available for every dataset, model, seed
- ~100GB of data, ~200K CPU hours of compute

> 💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

TabRepo

TabRepo: A Large Scale Repository of Tabular Model Evaluations and its AutoML Applications

David Salinas[1,*] Nick Erickson[1,*]

- Goals:
  - 1) reduce cost of evaluation (40K CPU hours to evaluate a single method on AutoML Benchmark)
  - 2) improve over the manual selection of AutoGluon default models
- Precomputed evaluations and results on:
  - 200 datasets from regression, classification, multi-class (thanks OpenML 🥰)
  - 200 random configurations of models used in AutoGluon (CatBoost, MLP, LightGBM, RandomForest, …) on all datasets with 3 seeds
- Performance metrics (latency, accuracy, …) **and predictions** available for every dataset, model, seed
- ~100GB of data, ~200K CPU hours of compute

💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

🥳 The dataset combined with **portfolio learning** allows to outperform Autogluon!

TabRepo

Studying the effect of HPO and ensembling

TabRepo

Studying the effect of HPO and ensembling

💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

TabRepo

Studying the effect of HPO and ensembling

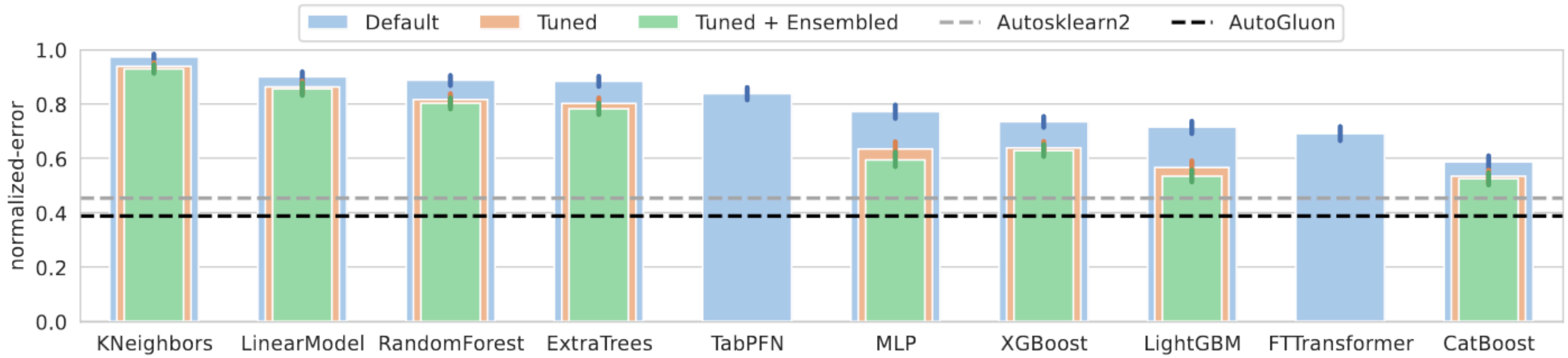


Figure 2: Normalized error for all model families when using default hyperparameters, tuned hyperparameters, and ensembling after tuning. All methods are run with a 4h budget.

💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

TabRepo

Studying the effect of HPO and ensembling

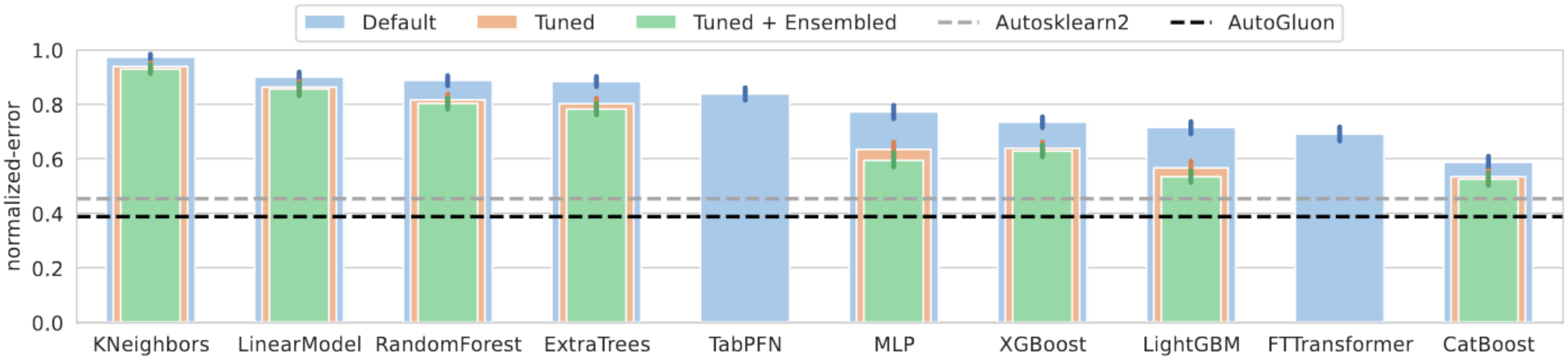


Figure 2: Normalized error for all model families when using default hyperparameters, tuned hyperparameters, and ensembling after tuning. All methods are run with a 4h budget.

💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

Doing this analysis just costs a few minutes on a laptop (as opposed to days on a cluster!)

TabRepo

Studying the effect of HPO and ensembling

Tuning hyperparameters helps a lot but it is not done in AutoGluon because of cost. Can transfer learning help?

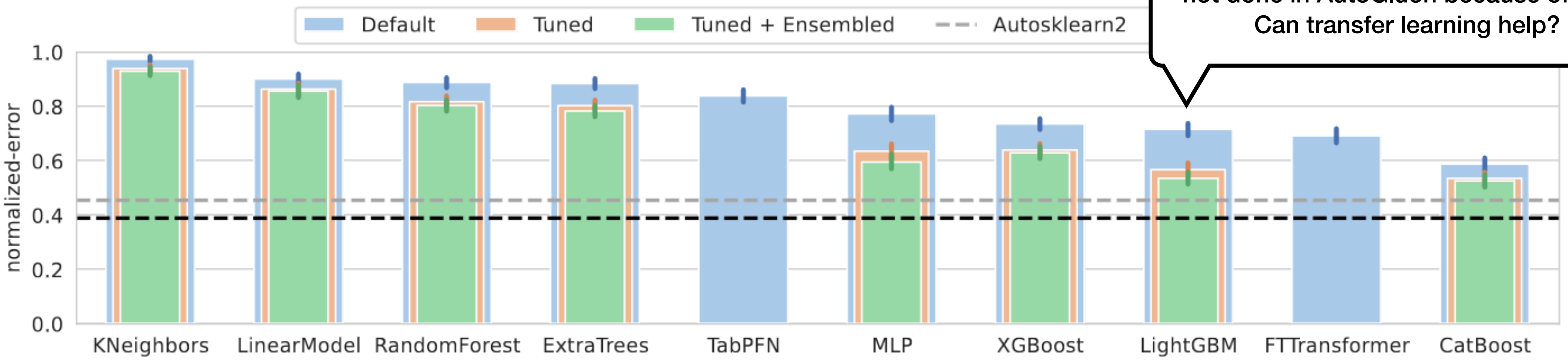


Figure 2: Normalized error for all model families when using default hyperparameters, tuned hyperparameters, and ensembling after tuning. All methods are run with a 4h budget.

💡 **Storing predictions** and target labels allows to obtain the performance of **any ensemble** on the fly!

Doing this analysis just costs a few minutes on a laptop (as opposed to days on a cluster!)

Portfolio learning

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1,\ldots,j_k)\in[m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

Select among all possible sets of k models

With best avg. performance across datasets …

… when using the best performing model on a given dataset

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

$$j_1 = \text{argmin}_{j_1 \in [m]} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij_1}, \qquad j_n = \text{argmin}_{j_n \in [m]} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_n})$$

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

$$j_1 = \text{argmin}_{j_1 \in [m]} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij_1}, \qquad j_n = \text{argmin}_{j_n \in [m]} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_n})$$

Pick the model performing best on average

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

$$j_1 = \text{argmin}_{j_1 \in [m]} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij_1}, \qquad j_n = \text{argmin}_{j_n \in [m]} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_n})$$

Pick the model performing best on average

Pick the model performing best on average when combined with the ones previously selected

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \ldots, j_k) = \text{argmin}_{(j_1, \ldots, j_k) \in [m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_k})$$

Benefits 👍:
- Approximation guarantees from the original (sub-modular) problem
- Tractable
- Works extremely well in practice

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

$$j_1 = \text{argmin}_{j_1 \in [m]} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij_1}, \qquad j_n = \text{argmin}_{j_n \in [m]} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \ldots, \varepsilon_{ij_n})$$

Pick the model performing best on average

Pick the model performing best on average when combined with the ones previously selected

Portfolio learning

- Assume we have access to error metrics of n datasets on m models, denoted as $\varepsilon \in \mathbb{R}^{n \times m}$
- How can we select the best set of k default models for an average dataset?
- Solve the optimization problem:

With best avg. performance across datasets …

… when using the best performing model on a given dataset

Select among all possible sets of k models

$$(j_1, \dots, j_k) = \text{argmin}_{(j_1,\dots,j_k)\in[m]^k} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \dots, \varepsilon_{ij_k})$$

- NP-hard [Feurer 2022], but admits an approximation
- Greedy algorithm:

$$j_1 = \text{argmin}_{j_1\in[m]} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{ij_1}, \qquad j_n = \text{argmin}_{j_n\in[m]} \frac{1}{n} \sum_{i=1}^{n} \min(\varepsilon_{ij_1}, \dots, \varepsilon_{ij_n})$$

Pick the model performing best on average

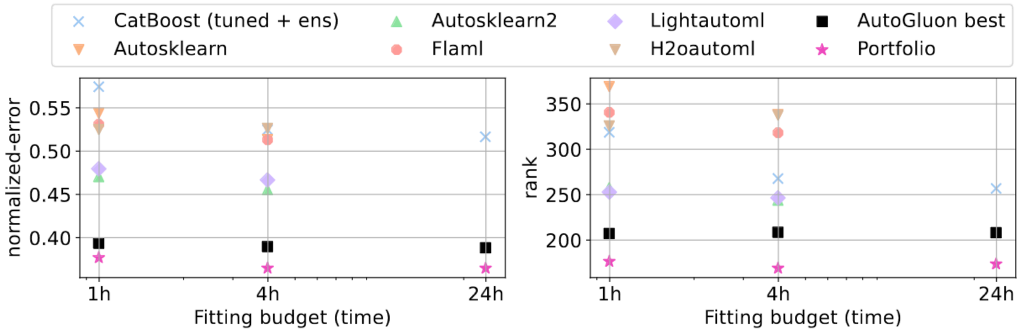
Pick the model performing best on average when combined with the ones previously selected

Benefits 👍:
- Approximation guarantees from the original (sub-modular) problem
- Tractable
- Works extremely well in practice

Disadvantage 👎: needs a grid or a surrogate

Results

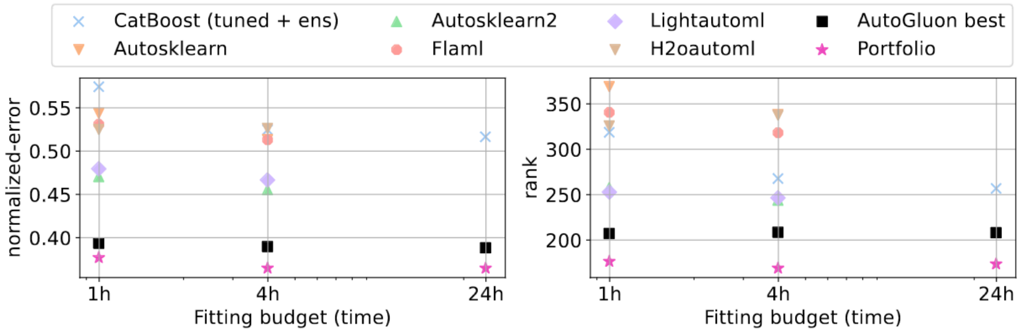
Results



CatBoost (tuned + ens)
Autosklearn
Autosklearn2
Flaml
Lightautoml
H2oautoml
AutoGluon best
Portfolio
normalized-error
0.55
0.50
0.45
0.40
1h
4h
24h
Fitting budget (time)
rank
350
300
250
200
1h
4h
24h
Fitting budget (time)

Results

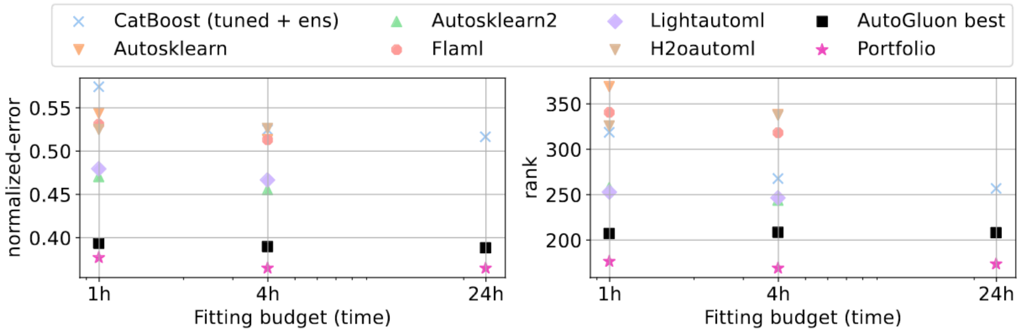
- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied



Results

- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied

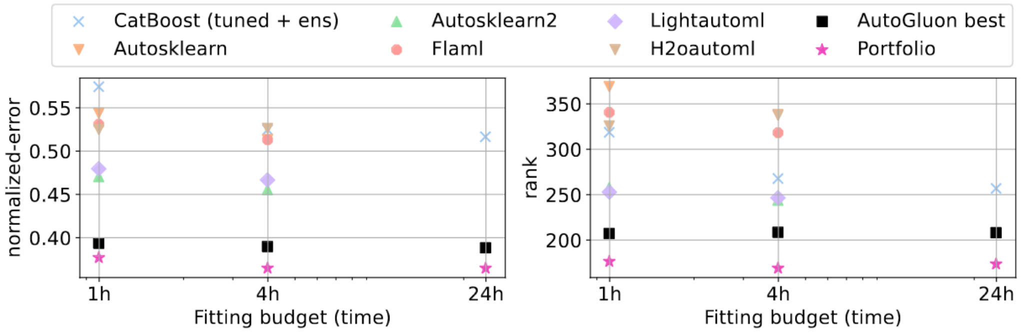
- We can analyse the performance of various components: #ensemble, #configurations, #datasets



Results

- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied

- We can analyse the performance of various components: #ensemble, #configurations, #datasets



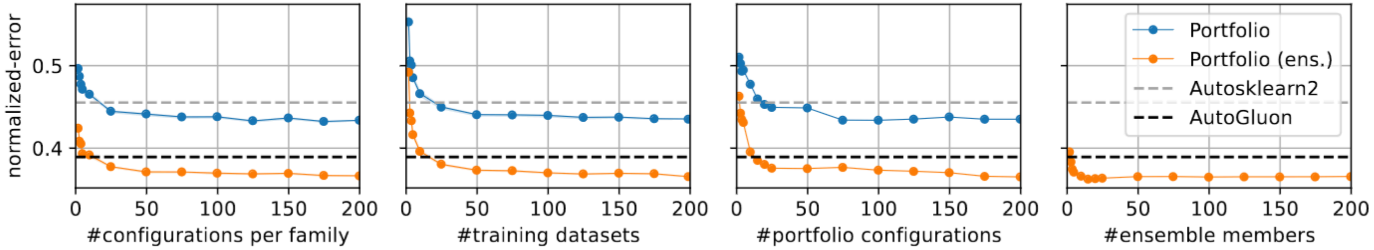


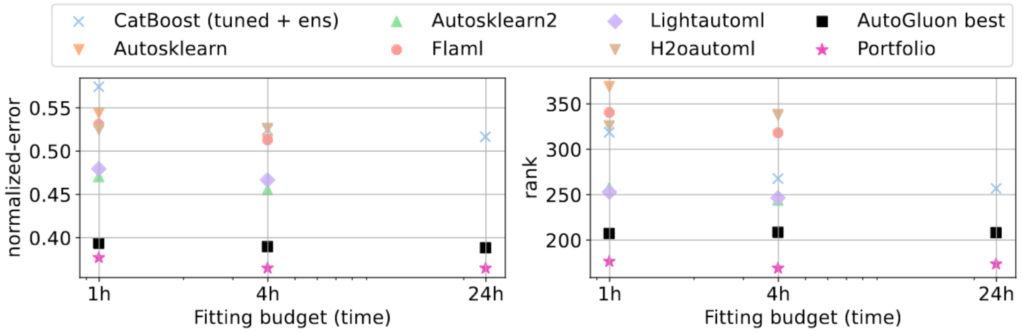
Figure 4: Impact on normalized error when varying the (a) number of configurations per family, (b) number of training datasets, (c) portfolio size and (d) number of ensemble members.

Results

- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied

- We can analyse the performance of various components: #ensemble, #configurations, #datasets

- Portfolio configurations has replaced the manually configured defaults and improved significantly AutoGluon



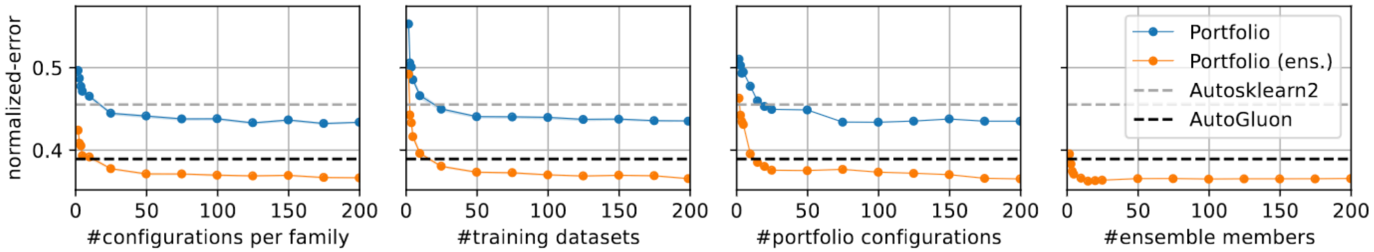


Figure 4: Impact on normalized error when varying the (a) number of configurations per family, (b) number of training datasets, (c) portfolio size and (d) number of ensemble members.

Results

- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied
- We can analyse the performance of various components: #ensemble, #configurations, #datasets
- Portfolio configurations has replaced the manually configured defaults and improved significantly AutoGluon

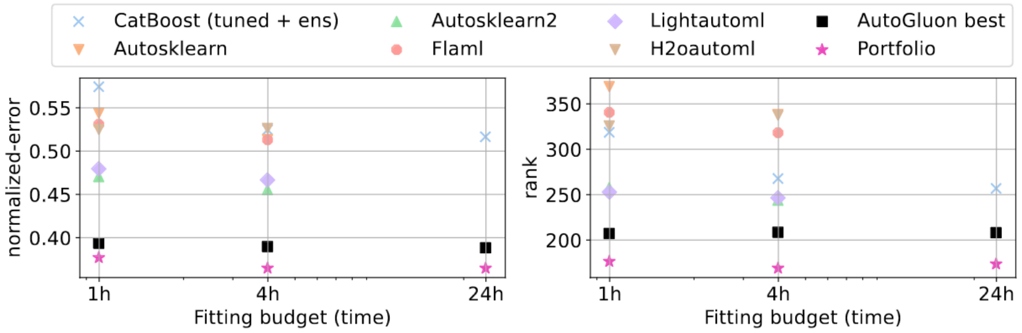


Table 2: Performance of AutoGluon combined with portfolios on AMLB.

method	win-rate	loss reduc.
AG + Portfolio (ours)	-	**0%**
AG	67%	2.8%
MLJAR	81%	22.5%
lightautoml	83%	11.7%
GAMA	86%	15.5%
FLAML	87%	16.3%
autosklearn	89%	11.8%
H2OAutoML	92%	10.3%
CatBoost	94%	18.1%
TunedRandomForest	94%	22.9%
RandomForest	97%	25.0%
XGBoost	98%	20.9%
LightGBM	98%	23.6%

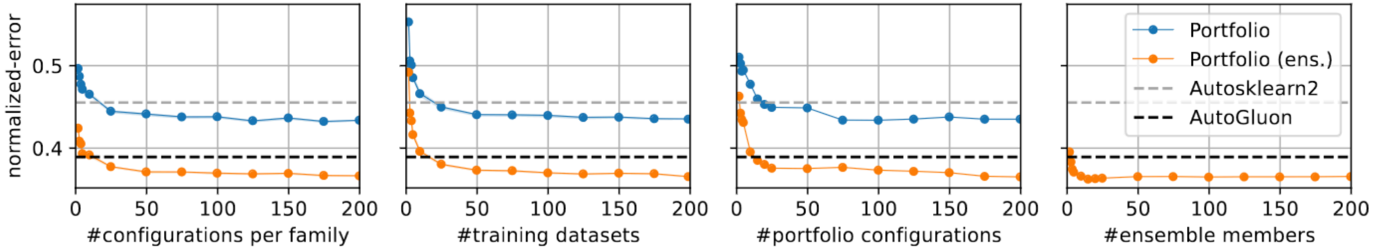
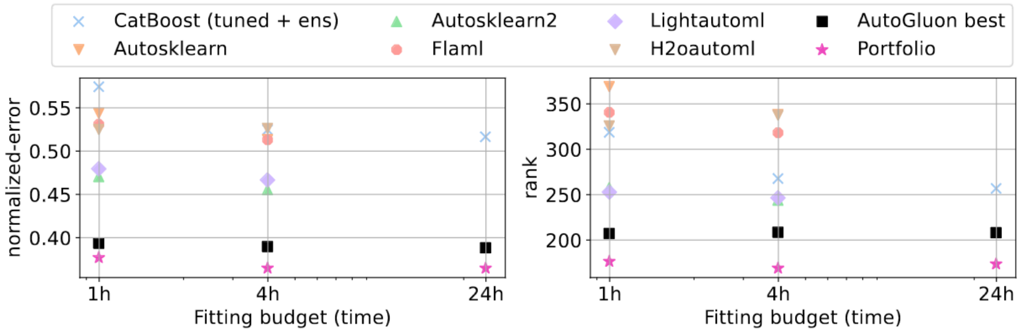


Figure 4: Impact on normalized error when varying the (a) number of configurations per family, (b) number of training datasets, (c) portfolio size and (d) number of ensemble members.

Results

- Just fitting portfolio configuration on evaluations of TabRepo outperforms all SOTA AutoML methods studied
- We can analyse the performance of various components: #ensemble, #configurations, #datasets
- Portfolio configurations has replaced the manually configured defaults and improved significantly AutoGluon



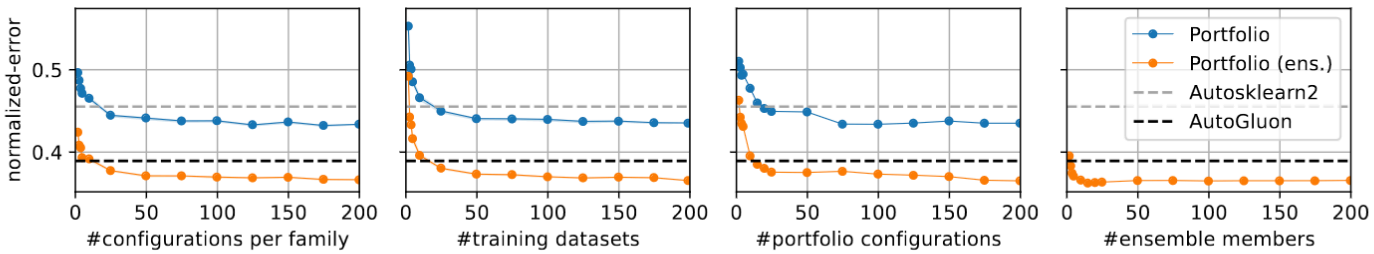


Figure 4: Impact on normalized error when varying the (a) number of configurations per family, (b) number of training datasets, (c) portfolio size and (d) number of ensemble members.

Table 2: Performance of AutoGluon combined with portfolios on AMLB.

method	win-rate	loss reduc.
AG + Portfolio (ours)	**-**	**0%**
AG	$\boxed{67\%}$	2.8%
MLJAR	81%	22.5%
lightautoml	83%	11.7%
GAMA	86%	15.5%
FLAML	87%	16.3%
autosklearn	89%	11.8%
H2OAutoML	92%	10.3%
CatBoost	94%	18.1%
TunedRandomForest	94%	22.9%
RandomForest	97%	25.0%
XGBoost	98%	20.9%
LightGBM	98%	23.6%

Results

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget
    - Apply different meta-heuristics to optimise the learned default portfolio list of configurations on a new dataset

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget
    - Apply different meta-heuristics to optimise the learned default portfolio list of configurations on a new dataset
    - Multiobjective optimization taking latency into account…

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget
    - Apply different meta-heuristics to optimise the learned default portfolio list of configurations on a new dataset
    - Multiobjective optimization taking latency into account…
    - All those experiments can be done… with your laptop!!

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget
    - Apply different meta-heuristics to optimise the learned default portfolio list of configurations on a new dataset
    - Multiobjective optimization taking latency into account…
    - All those experiments can be done… with your laptop!!
- 👨‍💻 https://github.com/autogluon/tabrepo

Results

- 🥳 All those experiments (fitting portfolio and evaluating) can be done using TabRepo for a very small cost (e.g. many table lookups)
- Possible research ideas:
    - Find best tabular configurations given time budget
    - Apply different meta-heuristics to optimise the learned default portfolio list of configurations on a new dataset
    - Multiobjective optimization taking latency into account…
    - All those experiments can be done… with your laptop!!
- 🧑‍💻 https://github.com/autogluon/tabrepo
- Quick demo

Limitations

Limitations

- Easy to rerun paper analysis but hard to compare your own method

Limitations

- Easy to rerun paper analysis but hard to compare your own method
- Large collections of datasets (216) but mostly grabbed everything we could

Limitations

- Easy to rerun paper analysis but hard to compare your own method
- Large collections of datasets (216) but mostly grabbed everything we could
- No good control on quality, duplication, domain

Limitations

- Easy to rerun paper analysis but hard to compare your own method
- Large collections of datasets (216) but mostly grabbed everything we could
- No good control on quality, duplication, domain
- Only TabPFN-v1 as In Context Learning (ICL) method

Any questions?

Paper: https://arxiv.org/pdf/2311.02971
Code: https://github.com/autogluon/tabrepo


David Salinas


Nick Erickson

Part II

TabArena: A Living Benchmark for Machine Learning on Tabular Data

Motivation 1: Unreliable Baselines

How to become SOTA on the highly used benchmark by McElfresh et al. (2023):

Model	Avg. Rank	Avg. norm. logloss	Avg. logloss
XGBoost	5.56	0.1	0.39
CatBoost	5.84	0.12	0.45
LightGBM	6.85	0.17	0.45
ResNet	8.12	0.22	0.49
SAINT	8.77	0.23	0.52
...			
MLP	10.79	0.39	0.96
...			
KNN	15.68	0.71	0.88

Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 1: Unreliable Baselines

How to become SOTA on the highly used benchmark by McElfresh et al. (2023):

Model	Avg. Rank	Avg. norm. logloss	Avg. logloss
XGBoost (ours, holdout)	4.13	0.06	0.36
XGBoost	5.56	0.1	0.39
CatBoost	5.84	0.12	0.45
MLP (ours, holdout)	6.09	0.15	0.4
LightGBM	6.85	0.17	0.45
ResNet	8.12	0.22	0.49
SAINT	8.77	0.23	0.52
...			
MLP	10.79	0.39	0.96
...			
KNN	15.68	0.71	0.88

Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 1: Unreliable Baselines

How to become SOTA on the highly used benchmark by McElfresh et al. (2023):

Model	Avg. Rank	Avg. norm. logloss	Avg. logloss
XGBoost (ours, holdout)	4.13	0.06	0.36
XGBoost	5.56	0.1	0.39
CatBoost	5.84	0.12	0.45
MLP (ours, holdout)	6.09	0.15	0.4
LightGBM	6.85	0.17	0.45
ResNet	8.12	0.22	0.49
SAINT	8.77	0.23	0.52
...			
MLP	10.79	0.39	0.96
...			
KNN	15.68	0.71	0.88

Accepted ICML and NeurIPS papers (that claim SOTA)

Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 1: Unreliable Baselines

How to become SOTA on the highly used benchmark by McElfresh et al. (2023):

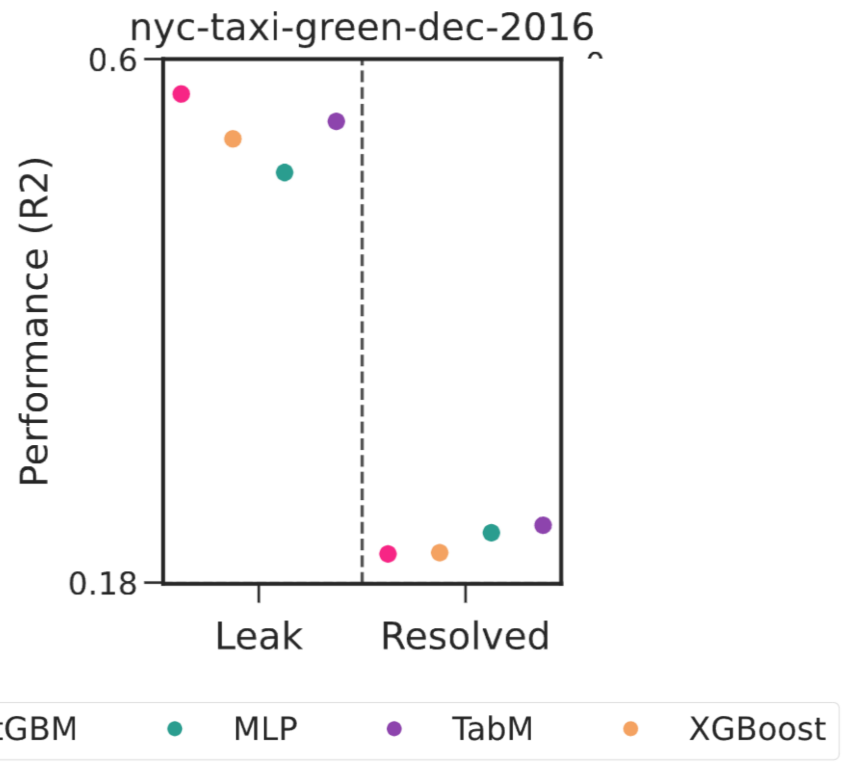
Model	Avg. Rank	Avg. norm. logloss	Avg. logloss
XGBoost (ours, 5CV)	1.77	0.03	0.34
MLP (ours, 5CV)	2.1	0.08	0.34
XGBoost (ours, holdout)	4.13	0.06	0.36
XGBoost	5.56	0.1	0.39
CatBoost	5.84	0.12	0.45
MLP (ours, holdout)	6.09	0.15	0.4
LightGBM	6.85	0.17	0.45
ResNet	8.12	0.22	0.49
SAINT	8.77	0.23	0.52
...			
MLP	10.79	0.39	0.96
...			
KNN	15.68	0.71	0.88

Accepted ICML and NeurIPS papers (that claim SOTA)

Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 2: Insufficient Dataset Curation

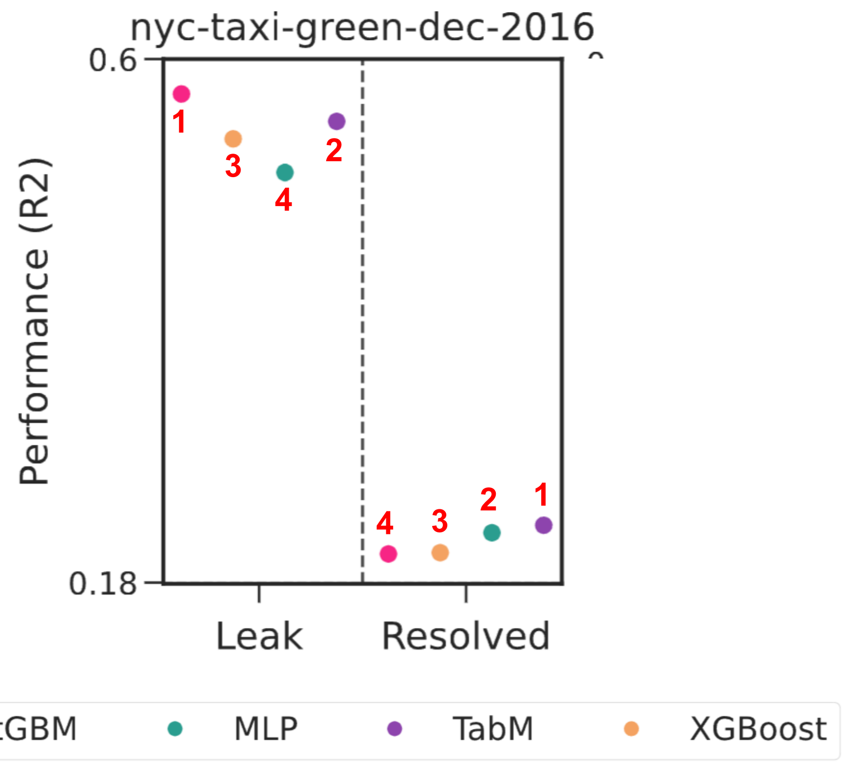
Faulty data influences the results:



Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 2: Insufficient Dataset Curation

Faulty data influences the results:



Tschalzev, Andrej, et al. "Unreflected use of tabular data repositories can undermine research quality." (2025).

Motivation 3: Inappropriate Evaluation Protocols

Splits must be appropriate for the data:

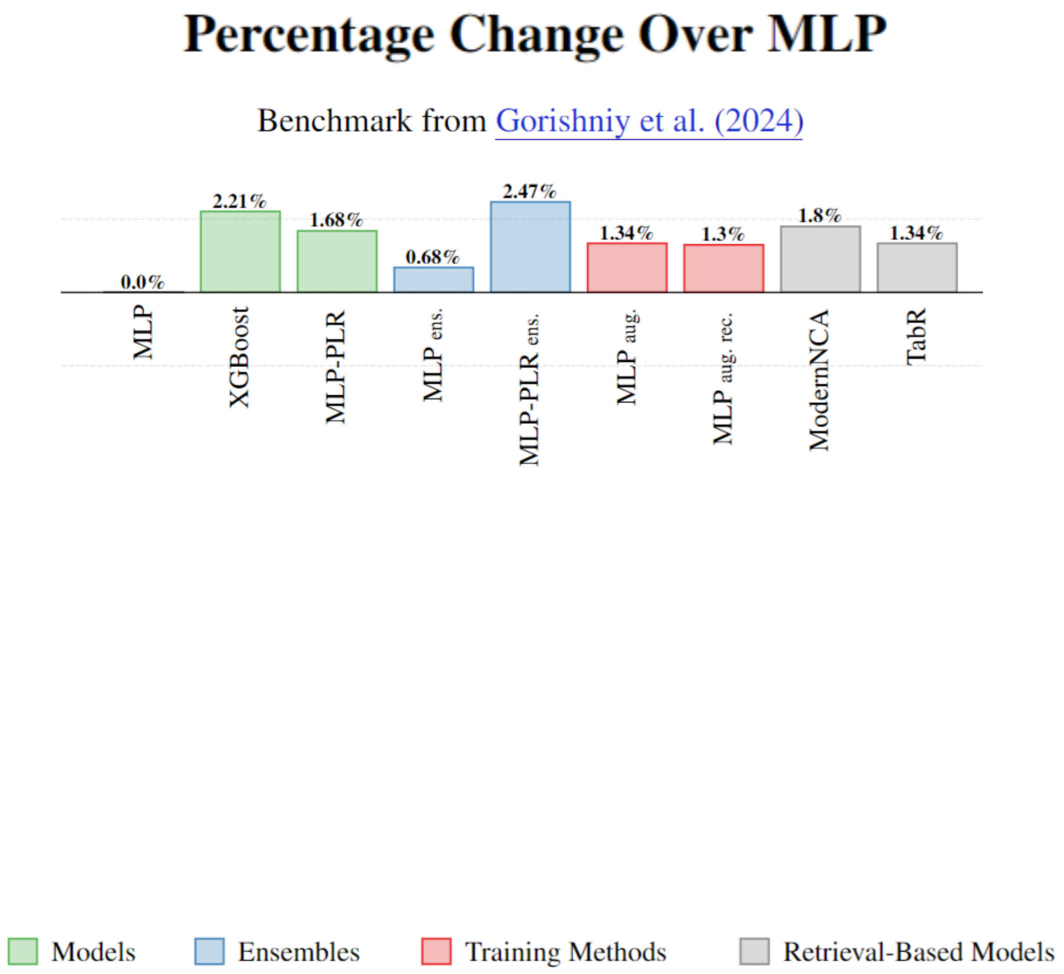
Benchmark	Time-split		
	Needed	Possible	Used
Grinsztajn et al. (2022)	22	5	
Tabzilla (McElfresh et al., 2023)	12	0	
WildTab (Kolesnikov, 2023)	1	1	✗
TableShift (Gardner et al., 2023)	15	8	
Gorishniy et al. (2024)	7	1	

Rubachev, Ivan, et al. "TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks." (2024)

Motivation 3: Inappropriate Evaluation Protocols

Splits must be appropriate for the data:

Benchmark	Time-split		
	Needed	Possible	Used
Grinsztajn et al. (2022)	22	5	
Tabzilla (McElfresh et al., 2023)	12	0	
WildTab (Kolesnikov, 2023)	1	1	✗
TableShift (Gardner et al., 2023)	15	8	
Gorishniy et al. (2024)	7	1	

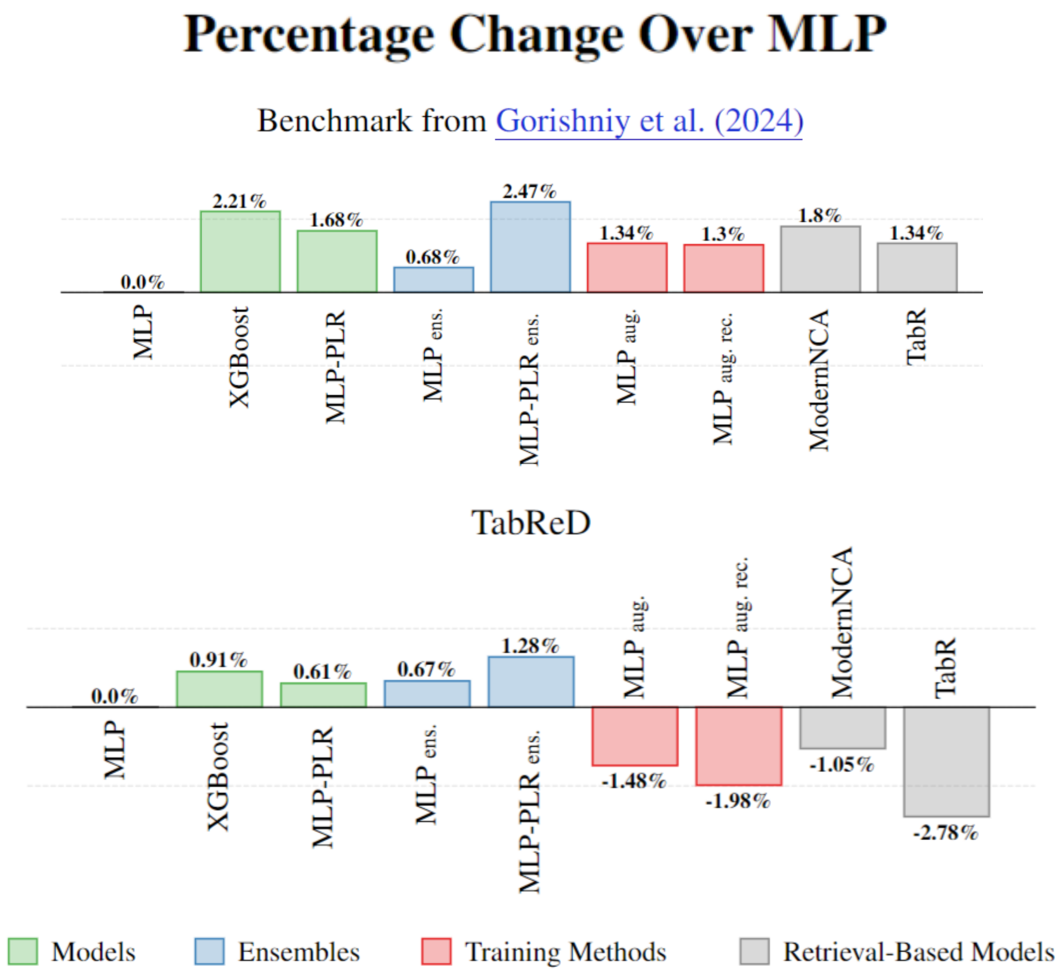


Rubachev, Ivan, et al. "TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks.“ (2024)

Motivation 3: Inappropriate Evaluation Protocols

Splits must be appropriate for the data:

Benchmark	Time-split		
	Needed	Possible	Used
Grinsztajn et al. (2022)	22	5	
Tabzilla (McElfresh et al., 2023)	12	0	
WildTab (Kolesnikov, 2023)	1	1	✗
TableShift (Gardner et al., 2023)	15	8	
Gorishniy et al. (2024)	7	1	



Rubachev, Ivan, et al. "TabReD: Analyzing Pitfalls and Filling the Gaps in Tabular Deep Learning Benchmarks.“ (2024)

Motivation Summary

(Partial) Overview of Tabular Benchmarks

Bischl et al. [28, 29]
Gorishniy et al. [30]
Shwartz-Ziv and Armon [31]
Grinsztajn et al. [32]
McElfresh et al. [33]
Fischer et al. [34]
Gijsbers et al. [35]
Kohli et al. [7]
Tschalzev et al. [8]
Holzmüller et al. [20]
Ye et al. [36]
Rubachev et al. [10]
Salinas and Erickson [37]

Erickson, Nick, et al. "TabArena: A Living Benchmark for Machine Learning on Tabular Data." (2025).

Motivation Summary

(Partial) Overview of Tabular Benchmarks

Bischl et al. [28, 29]
Gorishniy et al. [30]
Shwartz-Ziv and Armon [31]
Grinsztajn et al. [32]
McElfresh et al. [33]
Fischer et al. [34]
Gijsbers et al. [35]
Kohli et al. [7]
Tschalzev et al. [8]
Holzmüller et al. [20]
Ye et al. [36]
Rubachev et al. [10]
Salinas and Erickson [37]

One more benchmark should fix it!

Erickson, Nick, et al. "TabArena: A Living Benchmark for Machine Learning on Tabular Data." (2025).

Motivation Summary

(Partial) Overview of Tabular Benchmarks

Bischl et al. [28, 29]
Gorishniy et al. [30]
Shwartz-Ziv and Armon [31]
Grinsztajn et al. [32]
McElfresh et al. [33]
Fischer et al. [34]
Gijsbers et al. [35]
Kohli et al. [7]
Tschalzev et al. [8]
Holzmüller et al. [20]
Ye et al. [36]
Rubachev et al. [10]
Salinas and Erickson [37]

No!

One more benchmark should fix it!

Erickson, Nick, et al. "TabArena: A Living Benchmark for Machine Learning on Tabular Data." (2025).

Motivation Summary

(Partial) Overview of Tabular Benchmarks

Bischl et al. [28, 29]
Gorishniy et al. [30]
Shwartz-Ziv and Armon [31]
Grinsztajn et al. [32]
McElfresh et al. [33]
Fischer et al. [34]
Gijsbers et al. [35]
Kohli et al. [7]
Tschalzev et al. [8]
Holzmüller et al. [20]
Ye et al. [36]
Rubachev et al. [10]
Salinas and Erickson [37]

No!

~~One more benchmark should fix it!~~

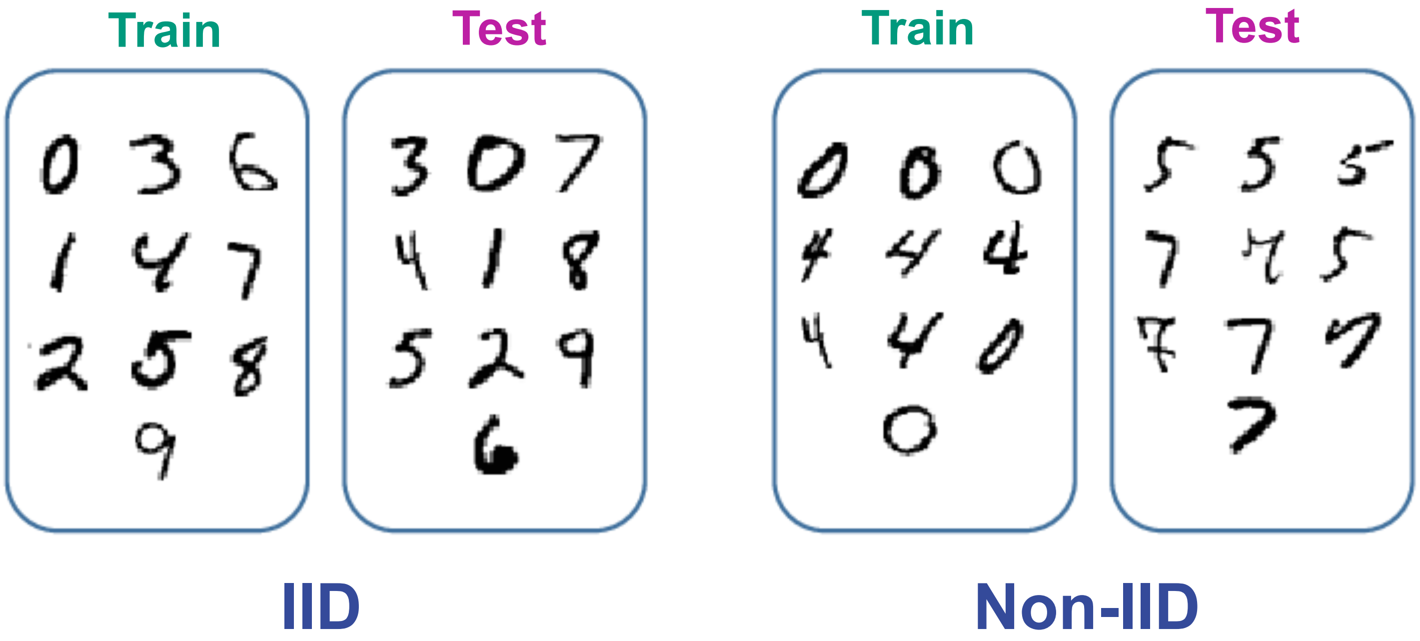
Benchmarks require <u>continuous</u> updates!

Erickson, Nick, et al. "TabArena: A Living Benchmark for Machine Learning on Tabular Data." (2025).

Background

Background

Independent and identically distributed (IID) Data



Figures from https://www.researchgate.net/profile/Henrik-Hellstroem/publication/344017123

Background

Relevance to AutoML: *many successful AutoML systems focus on IID tabular data*

Background

Relevance to AutoML: *many successful AutoML systems focus on IID tabular data*

As we show later, TabArena enables AutoML to:
- find the best models we should integrate into AutoML systems
- simulate complex ensemble pipelines
- meta-learn model portfolios (a.k.a. zero-shot HPO)
- transfer academic work/models into usable industry pipelines

Background

Relevance to AutoML: *many successful AutoML systems focus on IID tabular data*

As we show later, TabArena enables AutoML to:

- find the best models we should integrate into AutoML systems
- simulate complex ensemble pipelines
- meta-learn model portfolios (a.k.a. zero-shot HPO)
- transfer academic work/models into usable industry pipelines

TabArena, a research platform for AutoML ✨

TabArena-v0.1

Overview

Models

Datasets

Evaluations

Overview

For representative benchmarking, we need representative

Models

Datasets

Evaluations

and an explicit **Focus** to represent.

Overview

For representative benchmarking, we need representative

Models

Datasets

Evaluations

and an explicit **Focus** to represent.

Because of no free lunch theorem,
They *cannot* be a benchmark for "everything"

TabArena-v0.1

Focus

Focus Statement

Focus

We focus on:

- Tabular IID data spanning small to large data regime (500-250k samples)
- Predictive machine learning models for real-world classification and regression tasks
- Evaluating the peak performance of models

🎯The first **truly representative** benchmark for our focus **to guide researchers and practitioners**

Focus Statement

Focus

We focus on:

- Tabular IID data spanning small to large data regime (500-250k samples)
- Predictive machine learning models for real-world classification and regression tasks
- Evaluating the peak performance of models

🎯The first **truly representative** benchmark for our focus **to guide researchers and practitioners**

Not our focus / future work:

- Non-IID data (temporal dependencies or distribution shifts)
- Few-shot predictions, very small data (less than 500 training samples) or very large data
- Tabular data with text and/or semantic context information
- Other tasks such as clustering, subgroup discovery or survival analysis.
- Performance trade-offs

Clarifications

Focus

Why do we focus?

- Making the implicit assumptions explicit – "**I know that I know nothing**"
- **Clear communication** with practitioners and researchers
- Clearly **motivating the curation** of data and models

Clarifications

Focus

Why do we focus?

- Making the implicit assumptions explicit – "**I know that I know nothing**"
- **Clear communication** with practitioners and researchers
- Clearly **motivating the curation** of data and models

Why do we care about ML on tabular IID data?

- **Omnipresent traditional ML task** in industry and academia
- Playground for **model development** and a key task for **AutoML systems**
- **Stepping stone for exciting new avenues** such as context-aware or non-IID modelling

Clarifications

Focus

Why only small to large data (500-250k)?

- Among the **most common data**
- Smaller or larger necessitates **unique pipelines, models, and evaluation protocols**

Clarifications

Focus

Why only small to large data (500-250k)?

- Among the **most common data**
- Smaller or larger necessitates **unique pipelines, models, and evaluation protocols**

Why peak performance (and not trade-offs)?

- Most **models can be made much more efficient** if their performance is worth it
- **Trade-offs require user constraints** (per-dataset)
  - We already assume a limit of 1 hour!
- **Efficiency of the ensemble is relevant**, not the individual model
  - We can simulate and research this with TabArena!

Focus

TabArena-v0.1

Models

Why are models hard to get right?

Models

Search Space Problems:

CatBoost	
learning_rate	$\log \mathcal{U}(e^{-5}, 1)$
random_strength	$\mathcal{U}\{1, 2, \ldots, 20\}$
l2_leaf_reg	$\log \mathcal{U}(1, 10)$
bagging_temperature	$\mathcal{U}(0.0, 1.0)$
leaf_estimation_iterations	$\mathcal{U}\{1, 2, \ldots, 20\}$
iterations	$\mathcal{U}\{100, 101, \ldots, 4000\}$

Hollmann, Noah, et al. "Accurate predictions on small data with a tabular foundation model." (2025)

- Copied/summarized from prior work
- Disconnected from the pipeline and evaluation protocol

Models

Why are models hard to get right?

Search Space Problems:

CatBoost	
learning_rate	$\log \mathcal{U}(e^{-5}, 1)$
random_strength	$\mathcal{U}\{1, 2, \ldots, 20\}$
l2_leaf_reg	$\log \mathcal{U}(1, 10)$
bagging_temperature	$\mathcal{U}(0.0, 1.0)$
leaf_estimation_iterations	$\mathcal{U}\{1, 2, \ldots, 20\}$
iterations	$\mathcal{U}\{100, 101, \ldots, 4000\}$

Hollmann, Noah, et al. "Accurate predictions on small data with a tabular foundation model." (2025)

- Copied/summarized from prior work
- Disconnected from the pipeline and evaluation protocol

Implementation Problems:

- No pip package, undefined dependencies
- Untested research code
- Custom pipeline per model (with custom bugs)
- Insufficient data or know-how for model choices
- Ignorance of target metric or user constraints

Models, Hyperparameters, and Tuning

Models

1. **SOTA** tree-based, neural networks, and foundation **models**.
2. Implemented **with authors**
3. Good, **optimized** search spaces

Model	Short Name	Search Space	Type
Random Forests [12]	RandomForest	Prior Work + Us	
Extremely Randomized Trees [13]	ExtraTrees	Prior Work + Us	
XGBoost [14]	XGBoost	Prior Work + Us	
LightGBM [15]	LightGBM	Prior Work + Us	
CatBoost [16]	CatBoost	Prior Work + Us	
Explainable Boosting Machine [17, 18]	EBM	Authors	
FastAI MLP [19]	FastaiMLP	Authors	
Torch MLP [19]	TorchMLP	Authors	
RealMLP [20]	RealMLP	Authors	
$\text{TabM}^{\dagger}_{\text{mini}}$ [9]	TabM	Authors	
ModernNCA [21]	ModernNCA	Authors	
TabPFNv2 [5]	TabPFNv2	Authors	
TabICL [22]	TabICL	-	
TabDPT [23]	TabDPT	-	
Linear / Logistic Regression	Linear	TabRepo	
K-Nearest Neighbors	KNN	TabRepo	

tree-based ( ), neural network ( ), pretrained foundation models ( ), and baseline ( )

Models, Hyperparameters, and Tuning

Models

Benchmark	#splits inner
Bischl et al. [28, 29]	1
Gorishniy et al. [30]	1
Shwartz-Ziv and Armon [31]	1
Grinsztajn et al. [32]	1
McElfresh et al. [33]	1
Fischer et al. [34]	{1, 3, 10}
Gijsbers et al. [35]	-
Kohli et al. [7]	1
Tschalzev et al. [8]	10
Holzmüller et al. [20]	1
Ye et al. [36]	1
Rubachev et al. [10]	1
Salinas and Erickson [37]	8
TabArena (Ours)	8

Peak Performance by:

- Proper (inner) **cross-validation to avoid overfitting**

Models, Hyperparameters, and Tuning

Models

Benchmark	#splits inner	Ensembling
Bischl et al. [28, 29]	1	✗
Gorishniy et al. [30]	1	(✓)
Shwartz-Ziv and Armon [31]	1	(✓)
Grinsztajn et al. [32]	1	✗
McElfresh et al. [33]	1	✗
Fischer et al. [34]	{1, 3, 10}	✗
Gijsbers et al. [35]	-	(✓)
Kohli et al. [7]	1	✗
Tschalzev et al. [8]	10	(✓)
Holzmüller et al. [20]	1	(✓)
Ye et al. [36]	1	✗
Rubachev et al. [10]	1	(✓)
Salinas and Erickson [37]	8	✓
TabArena (Ours)	8	✓

Peak Performance by:

- Proper (inner) **cross-validation to avoid overfitting**
- Model-wise **post-hoc ensembling** (Caruana et al.)

Models, Hyperparameters, and Tuning

Models

Benchmark	#splits inner	Ensembling	HPO Limit #confs.	HPO Limit #hours
Bischl et al. [28, 29]	1	✗	1	-
Gorishniy et al. [30]	1	(✓)	100	6
Shwartz-Ziv and Armon [31]	1	(✓)	1000	-
Grinsztajn et al. [32]	1	✗	400	-
McElfresh et al. [33]	1	✗	30	10
Fischer et al. [34]	{1, 3, 10}	✗	{-, 500}	-
Gijsbers et al. [35]	-	(✓)	-	4
Kohli et al. [7]	1	✗	100	{3, -}
Tschalzev et al. [8]	10	(✓)	100	-
Holzmüller et al. [20]	1	(✓)	50	-
Ye et al. [36]	1	✗	100	-
Rubachev et al. [10]	1	(✓)	100	-
Salinas and Erickson [37]	8	✓	200	200
TabArena (Ours)	8	✓	200	200

Peak Performance by:

- Proper (inner) **cross-validation to avoid overfitting**
- Model-wise **post-hoc ensembling** (Caruana et al.)
- **Extensive HPO** (200 configs, 1 hour per config)

Focus

Models

TabArena-v0.1

Datasets

Datasets Curation

Datasets

1053 datasets with unique names from 13 tabular benchmarks

Deduplication
- 254 alternative version
- 167 same but other names
- 63 regex + sanity check
- 7 similar tasks

= 562

Other domain
- 66 image
- 39 forecasting
- 13 audio
- 12 text
- 5 control

= 427

Real predictive task
- 49 scientific discovery
- 44 deterministic
- 30 artificial or simulated

= 304

Other
- 142 tiny data
- 32 quality issues
- 9 License

= 121

IID
- 52 temporal
- 16 grouped

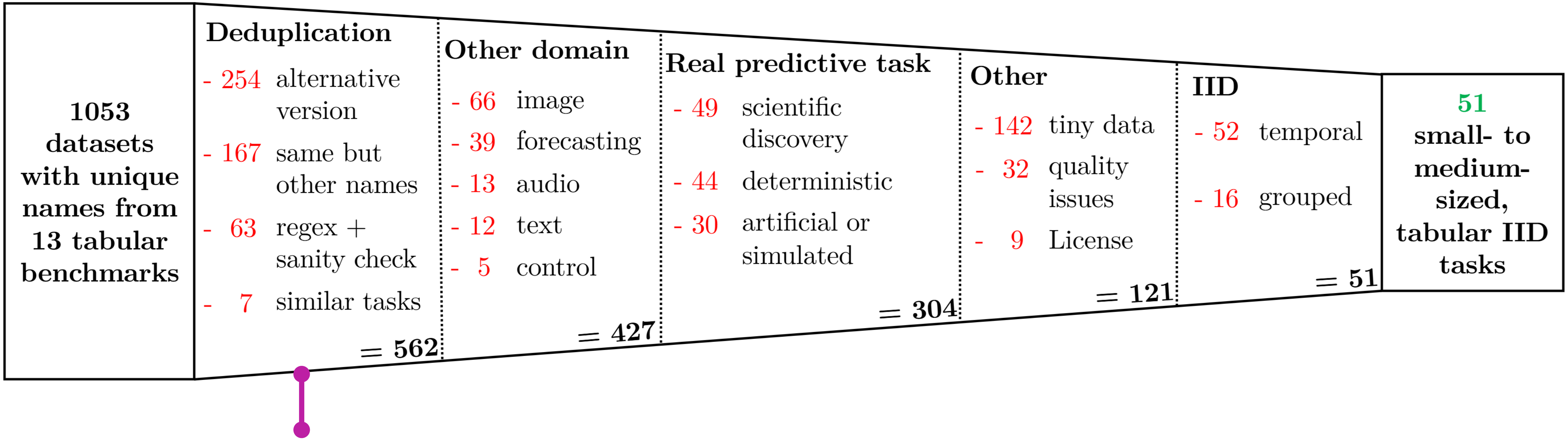
= 51

51 small- to medium-sized, tabular IID tasks

Results of our *manual* curation: *51 out of 1053*

Datasets Curation

Datasets

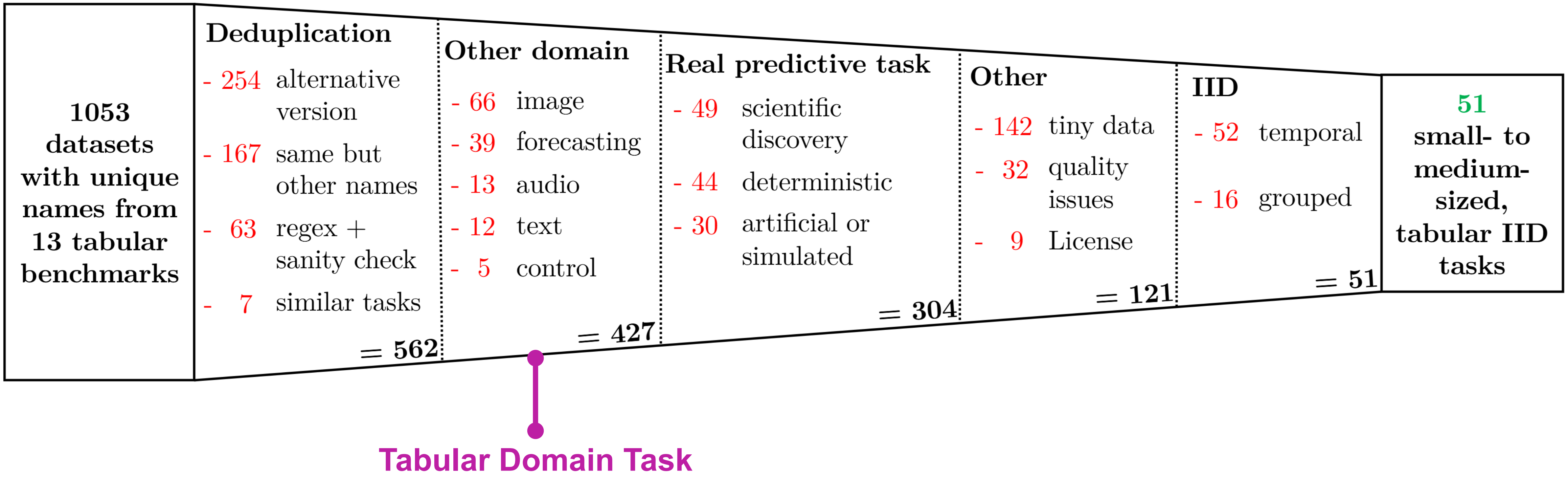


Unique datasets

- Many surprising duplicates (e.g., AutoML competition datasets)
- Very similar tasks (e.g., 5 datasets from one paper, same features different targets)

Datasets

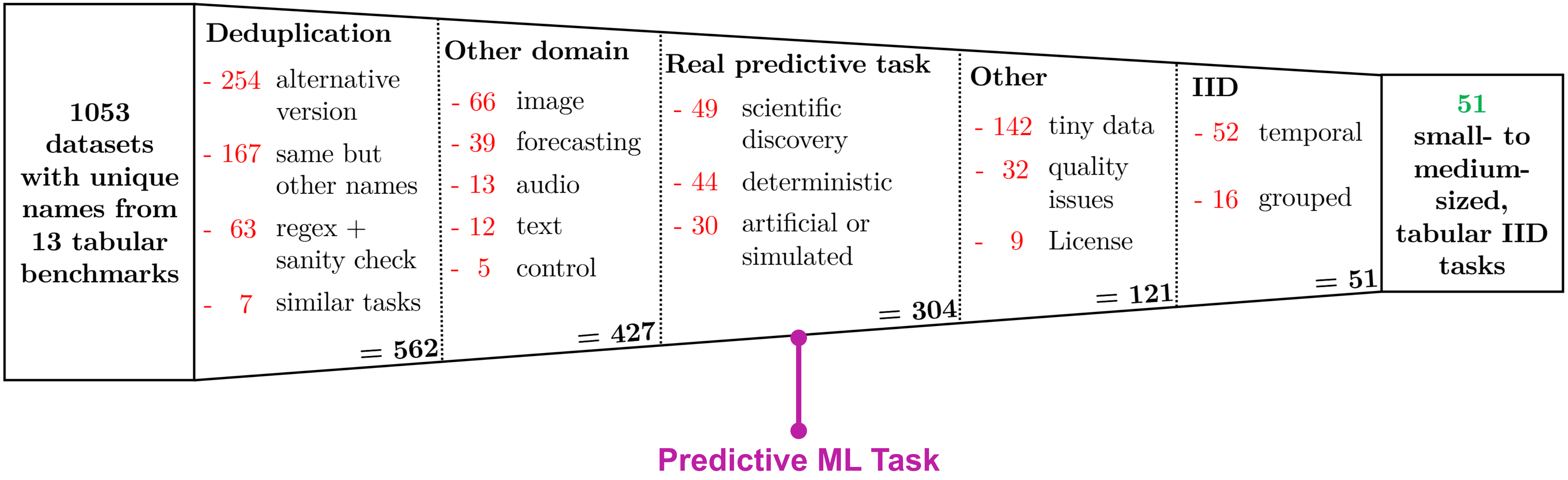
Datasets Curation



- Many datasets that treat images as tables (often very outdated)
- Often, only the original source described the data

Datasets Curation

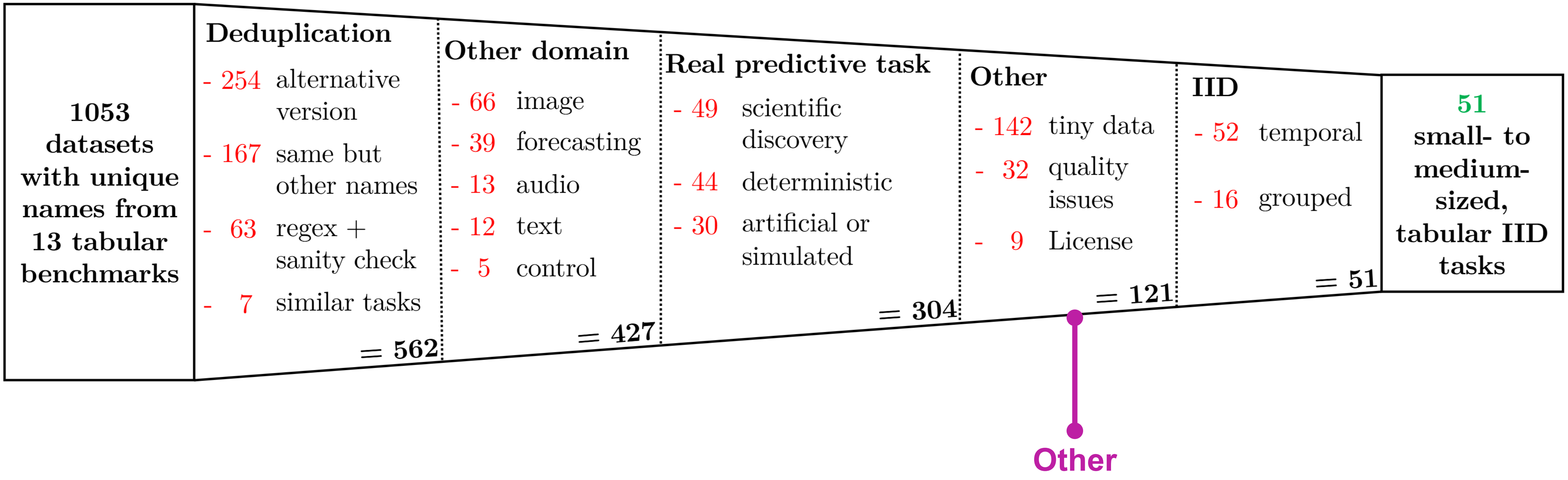
Datasets



- Scientific discovery (why/how questions) vs. predictive task
- Real-world data: not deterministic, not artificial, not simulated

Datasets Curation

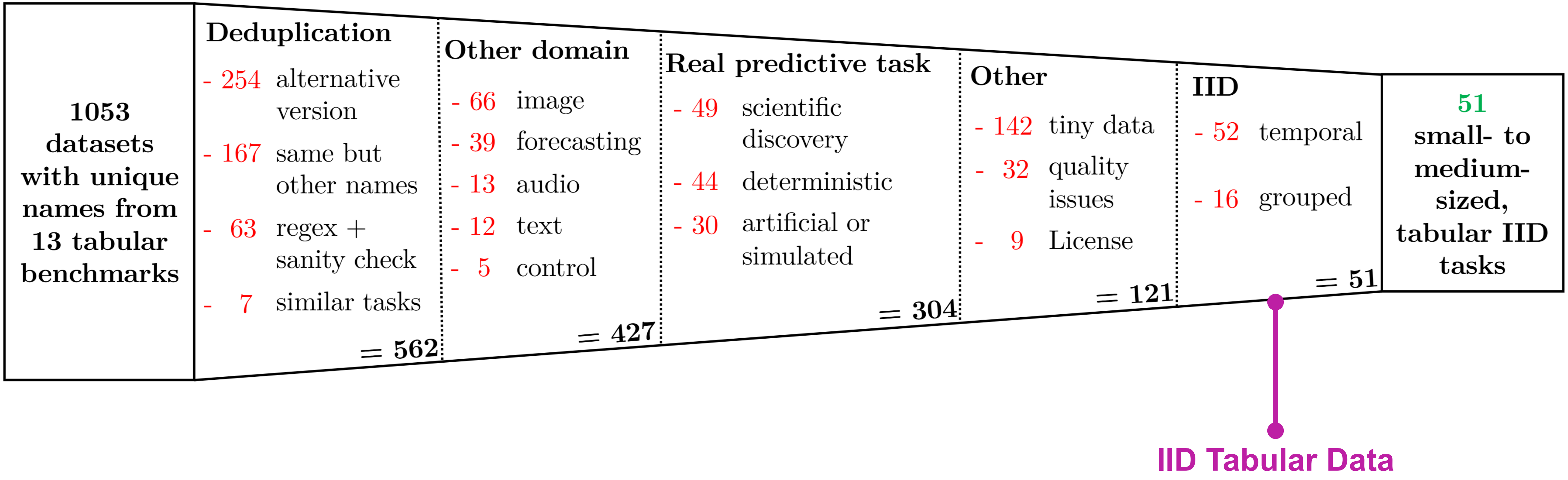
Datasets



- Many tiny (often old) datasets
- Datasets with preprocessing errors (PCA data leakage), missing source information, and target leakage

Datasets Curation

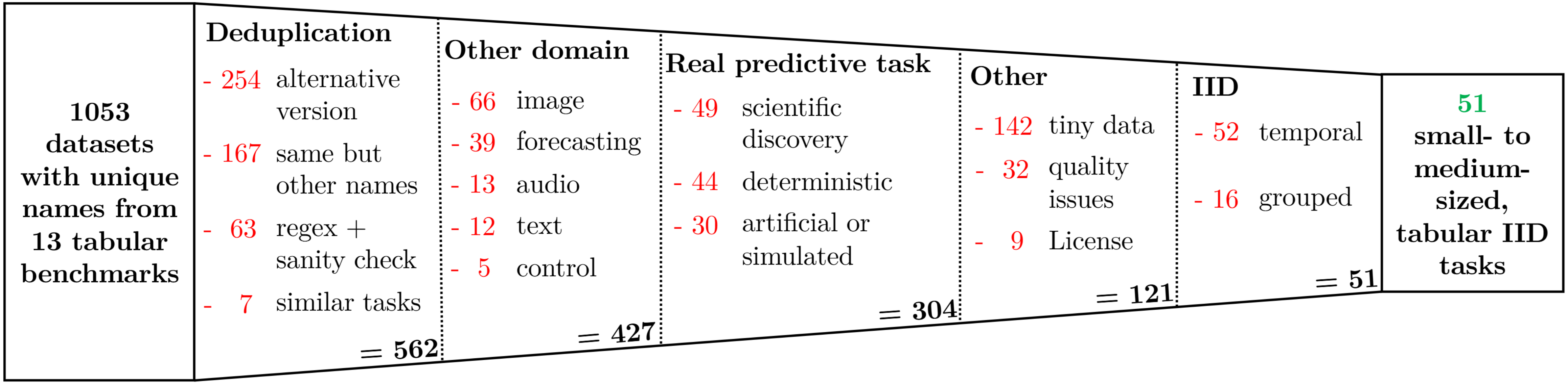
Datasets



- Tasks that require non-random splits
- Temporal-dependent features / grouped data (e.g., algorithm selection)
- Many borderline cases

Datasets Curation

Datasets



Check for yourself and verify our curation:
https://tabarena.ai/dataset-curation

Datasets Curation

Datasets

1053 datasets with unique names from 13 tabular benchmarks

Deduplication
- 254 alternative version
- 167 same but other names
- 63 regex + sanity check
- 7 similar tasks

= 562

Other domain
- 66 image
- 39 forecasting
- 13 audio
- 12 text
- 5 control

= 427

Real predictive task
- 49 scientific discovery
- 44 deterministic
- 30 artificial or simulated

= 304

Other
- 142 tiny data
- 32 quality issues
- 9 License

= 121

IID
- 52 temporal
- 16 grouped

= 51

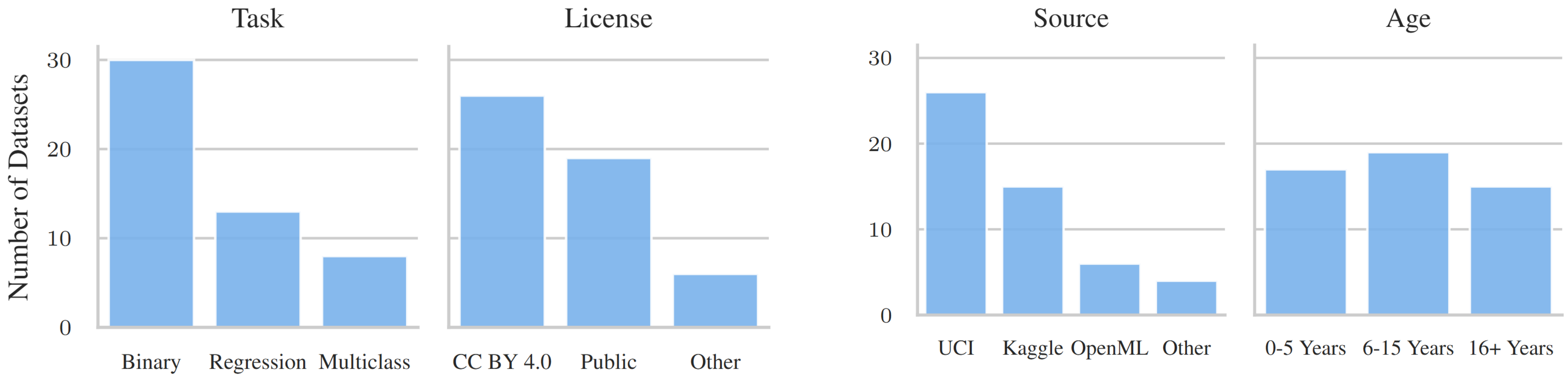
51 small- to medium-sized, tabular IID tasks

Check for yourself and verify our curation:
https://tabarena.ai/dataset-curation

Smaller is better!
Sometimes at least…

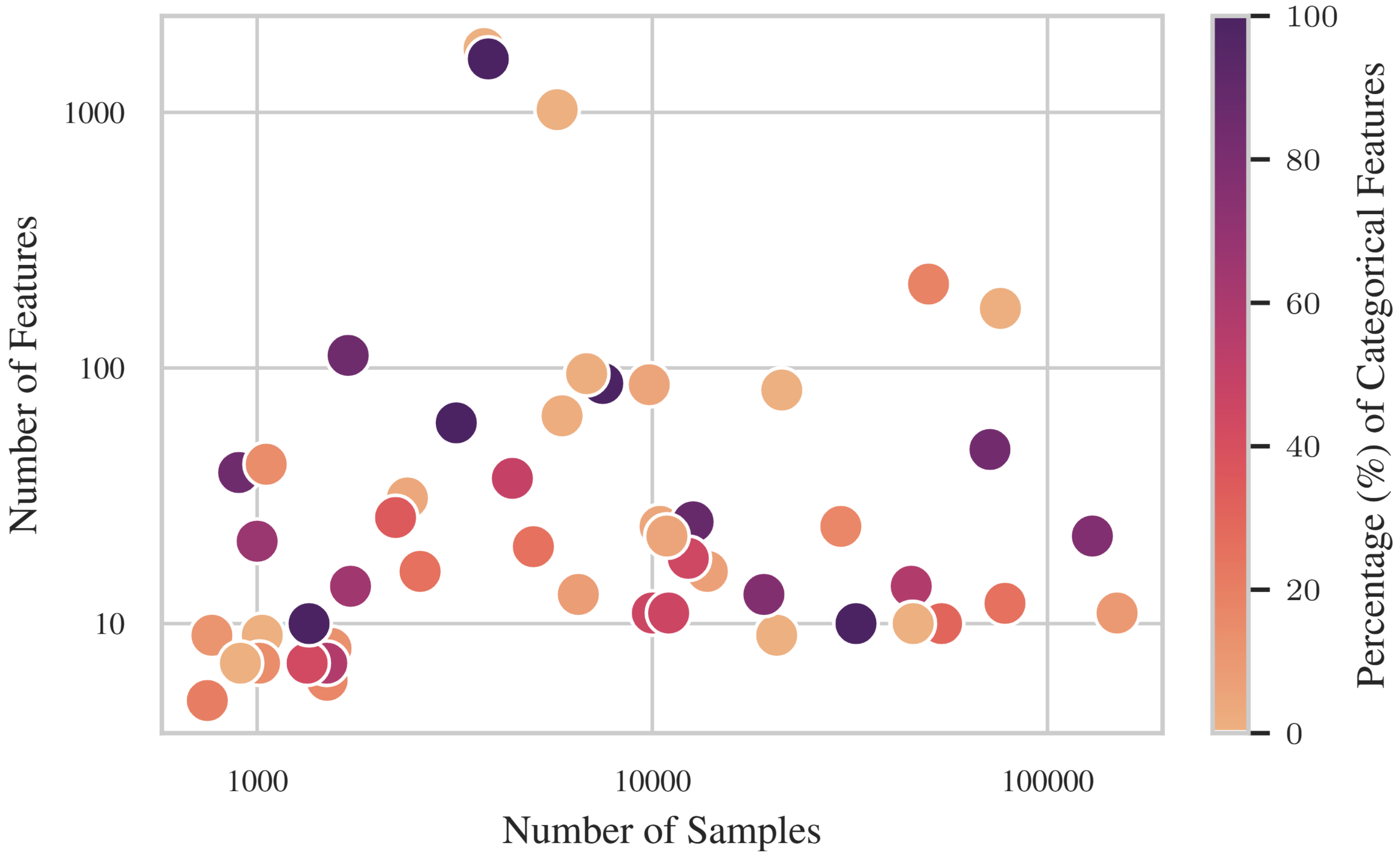
Datasets Overview

Datasets



Datasets Overview

Datasets



Datasets

Compared to Prior Benchmarks

Benchmark	Manual curation	#datasets remaining
Bischl et al. [28, 29]	✗	9/72
Gorishniy et al. [30]	✓	1/11
Shwartz-Ziv and Armon [31]	✗	1/11
Grinsztajn et al. [32]	✓	12/47
McElfresh et al. [33]	✗	13/196
Fischer et al. [34]	✓	8/35
Gijsbers et al. [35]	✓	15/104
Kohli et al. [7]	✓	17/187
Tschalzev et al. [8]	✓	1/10
Holzmüller et al. [20]	✓	10/118
Ye et al. [36]	✗	39/300
Rubachev et al. [10]	✓	0/8
Salinas and Erickson [37]	✗	19/200
TabArena (Ours)	✓	51/51

Focus

Models

Datasets

TabArena-v0.1

Evaluations

Evaluations

Evaluation Design

1. **Repeat experiments per dataset:**
   - 30 times for data with less than 2500 samples (10-repeated 3-fold cv)
   - 9 times for all other data (3-repeated 3-fold cv)
2. **Using the Elo rating system**
   - pairwise model comparison
   - 400-point Elo Gap corresponds to a 10 to 1 (91%) win rate
3. **Robust metrics appropriate for benchmarking**
   - Binary: ROC AUC
   - Multiclass: Log Loss
   - Regression: RMSE

Evaluations

Evaluation Design

1. **Repeat experiments per dataset:**
   - 30 times for data with less than 2500 samples (10-repeated 3-fold cv)
   - 9 times for all other data (3-repeated 3-fold cv)
2. **Using the Elo rating system**
   - pairwise model comparison
   - 400-point Elo Gap corresponds to a 10 to 1 (91%) win rate
3. **Robust metrics appropriate for benchmarking**
   - Binary: ROC AUC
   - Multiclass: Log Loss
   - Regression: RMSE
4. **Realistic reference pipeline for practitioners**
   - A pipeline practitioners can easily use
   - SOTA AutoML, AutoGluon trained for 4 hours

Evaluation Design

Evaluations

1. **Repeat experiments per dataset:**
   - 30 times for data with less than 2500 samples (10-repeated 3-fold cv)
   - 9 times for all other data (3-repeated 3-fold cv)
2. **Using the Elo rating system**
   - pairwise model comparison
   - 400-point Elo Gap corresponds to a 10 to 1 (91%) win rate
3. **Robust metrics appropriate for benchmarking**
   - Binary: ROC AUC
   - Multiclass: Log Loss
   - Regression: RMSE
4. **Realistic reference pipeline for practitioners**
   - A pipeline practitioners can easily use
   - SOTA AutoML, AutoGluon trained for 4 hours
5. **Store and share extensive metadata**

Evaluation Design

Evaluations

1. **Repeat experiments per dataset:**
   - 30 times for data with less than 2500 samples (10-repeated 3-fold cv)
   - 9 times for all other data (3-repeated 3-fold cv)
2. **Using the Elo rating system**
   - pairwise model comparison
   - 400-point Elo Gap corresponds to a 10 to 1 (91%) win rate
3. **Robust metrics appropriate for benchmarking**
   - Binary: ROC AUC
   - Multiclass: Log Loss
   - Regression: RMSE
4. **Realistic reference pipeline for practitioners**
   - A pipeline practitioners can easily use
   - SOTA AutoML, AutoGluon trained for 4 hours
5. **Store and share extensive metadata**
   - such as: validation predictions (per-fold), test predictions, training time, inference time, precomputed results on various metrics, hyperparameters – "**TabRepo 2.0**"

Evaluations

Evaluation Design

1. **Repeat experiments per dataset:**
   - 30 times for data with less than 2500 samples (10-repeated 3-fold cv)
   - 9 times for all other data (3-repeated 3-fold cv)
2. **Using the Elo rating system**
   - pairwise model comparison
   - 400-point Elo Gap corresponds to a 10 to 1 (91%) win rate
3. **Robust metrics appropriate for benchmarking**
   - Binary: ROC AUC
   - Multiclass: Log Loss
   - Regression: RMSE
4. **Realistic reference pipeline for practitioners**
   - A pipeline practitioners can easily use
   - SOTA AutoML, AutoGluon trained for 4 hours
5. **Store and share extensive metadata**
   - such as: validation predictions (per-fold), test predictions, training time, inference time, precomputed results on various metrics, hyperparameters – "**TabRepo 2.0**"

Evaluation Design

Evaluations

Benchmark	#splits inner	#splits outer	Results available
Bischl et al. [28, 29]	1	10	(✓)
Gorishniy et al. [30]	1	1	✗
Shwartz-Ziv and Armon [31]	1	{1, 3}	✗
Grinsztajn et al. [32]	1	{1, 2, 3, 5}	(✓)
McElfresh et al. [33]	1	10	(✓)
Fischer et al. [34]	{1, 3, 10}	{1, 10, 100}	(✓)
Gijsbers et al. [35]	-	10	(✓)
Kohli et al. [7]	1	1	✗
Tschalzev et al. [8]	10	1	✗
Holzmüller et al. [20]	1	10	✓
Ye et al. [36]	1	1	(✓)
Rubachev et al. [10]	1	1	(✓)
Salinas and Erickson [37]	8	3	✓
TabArena (Ours)	8	{9, 30}	✓

Focus

Models

Datasets

Evaluations

TabArena-v0.1

Results

The TabArena Team


Nick
Erickson

Lennart
Purucker


Andrej
Tschalzev


David
Holzmüller


Prateek
Mutalik Desai


David
Salinas


Frank
Hutter

aws

ufr

small data
MORE FROM LESS
CRC 1597

IN OMNIBUS VERITAS
UNIVERSITÄT MANNHEIM

Inria

ENS | PSL
UNIVERSITÉ PARIS

ellis
INSTITUTE
TÜBINGEN

PRIOR
LABS

The TabArena Team

Nick Erickson

Lennart Purucker

Andrej Tschalzev

David Holzmüller

Prateek Mutalik Desai

David Salinas

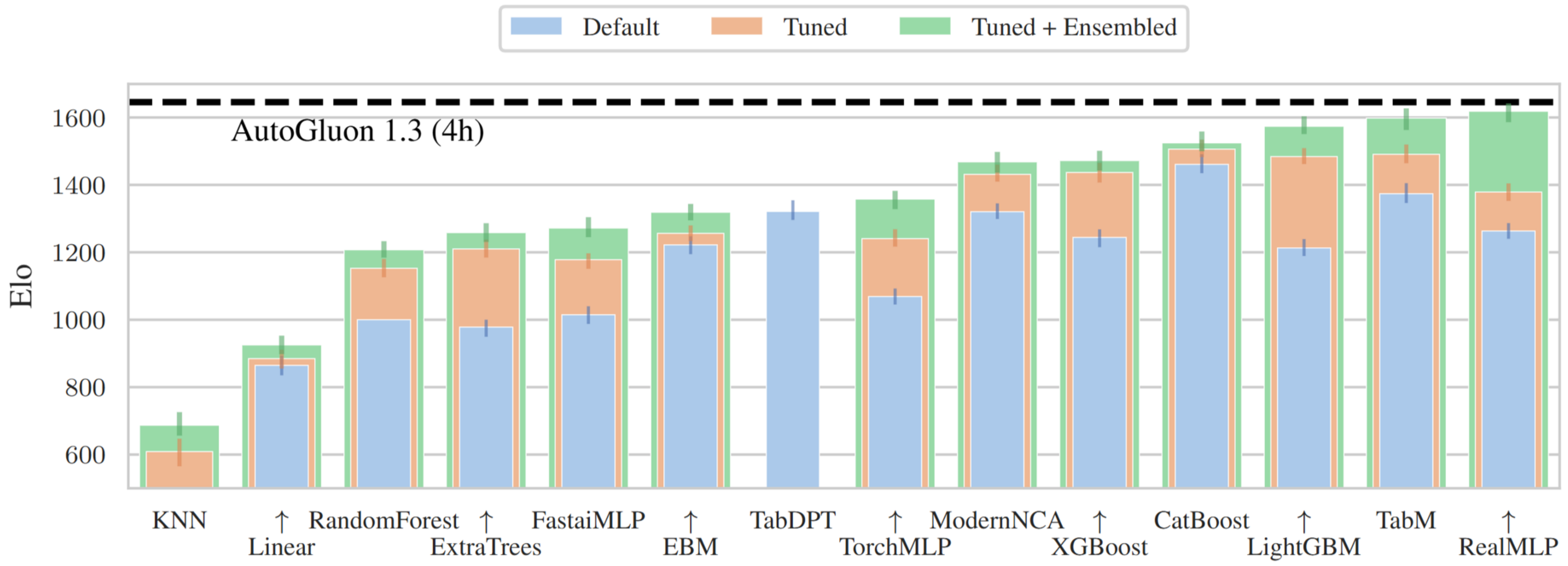
Frank Hutter

Competing interests

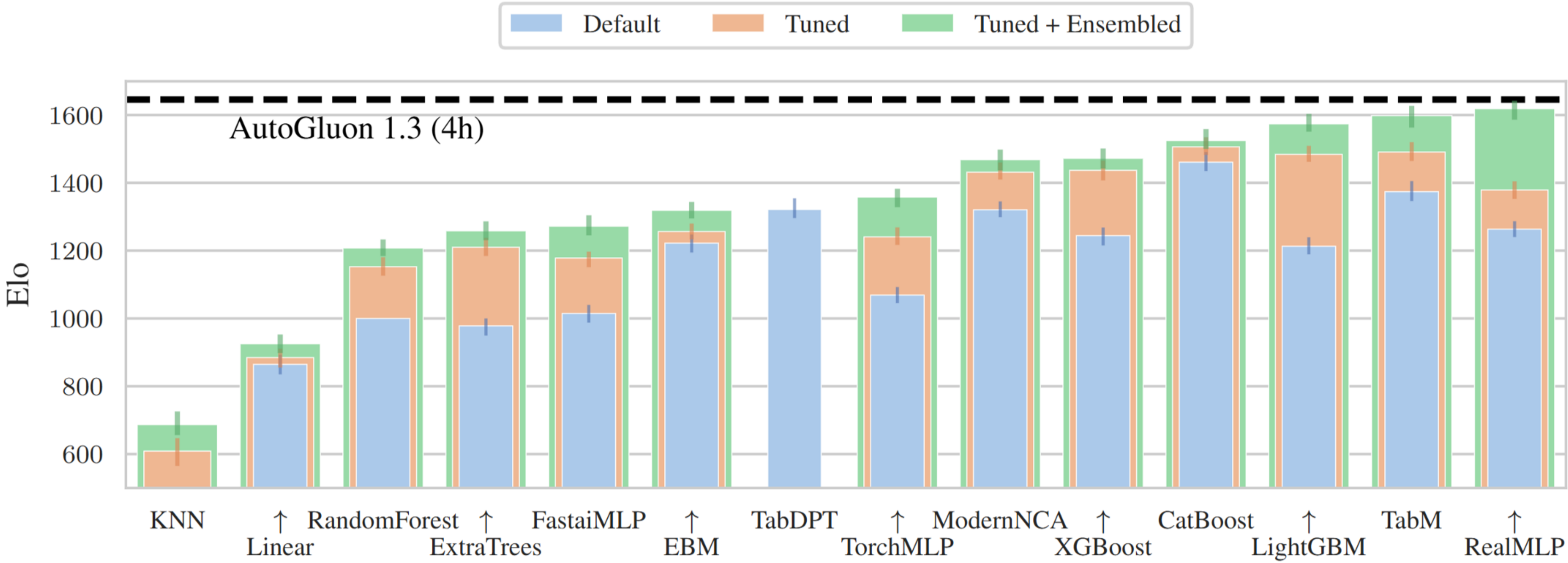
D.H. is one of the authors of RealMLP and one of the authors of TabICL.
D.S. and N.E. are the authors of TabRepo.
N.E., L.P., and P.M.D. are developers of AutoGluon, and in extension, the current maintainers of FastAI MLP and Torch MLP.
L.P. and F.H. are a subset of the authors of TabPFNv2.
L.P. is an OpenML core contributor.
F.H. is affiliated with PriorLabs, a company focused on developing tabular foundation models.
The authors declare no other competing interests.

INSTITUTE
TÜBINGEN

Main Results

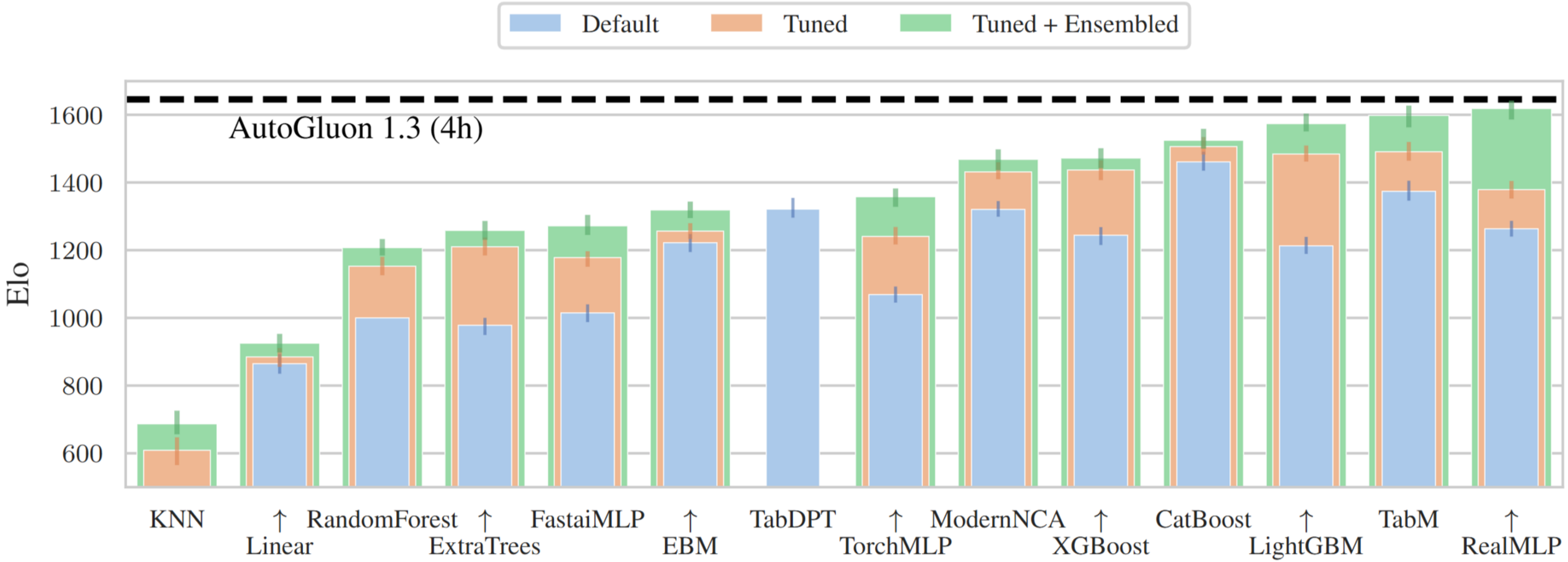


Main Results



CatBoost is best by default and with tuning.

Main Results



CatBoost is best by default and with tuning.

Deep learning models dominate with ensembling.

Main Results (cont.)

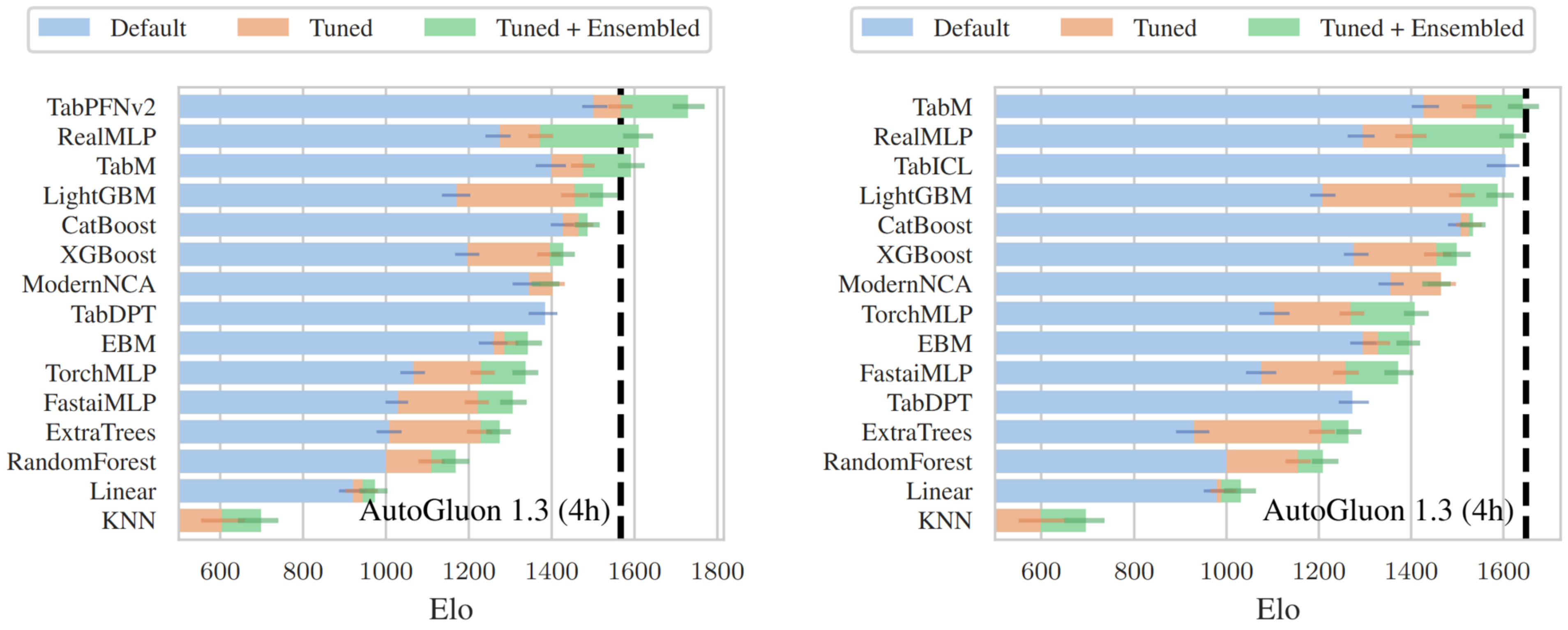


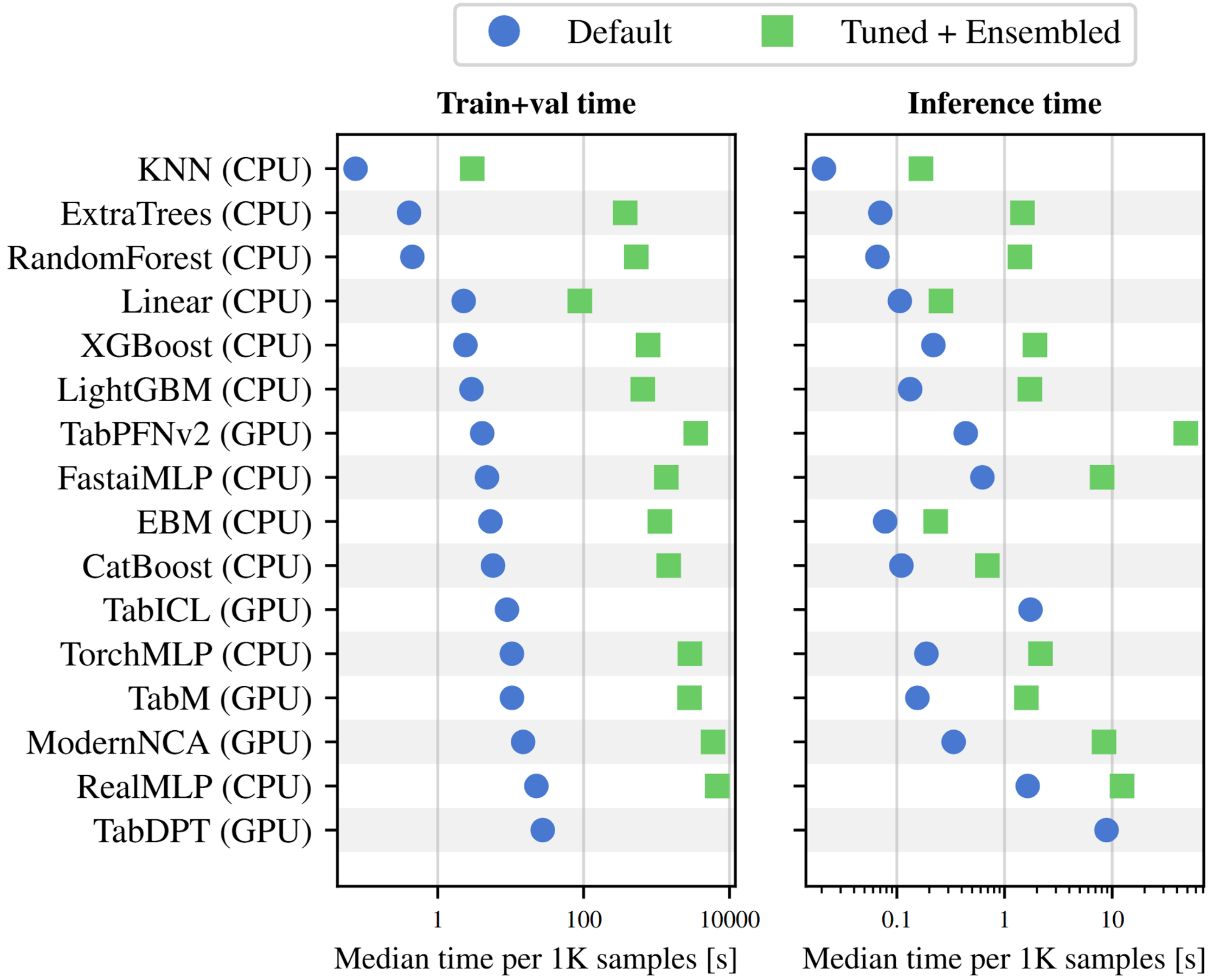
Figure 4: **Leaderboard for TabPFNv2-compatible (left) and TabICL-compatible (right) datasets.** For TabPFNv2, we obtain 33 datasets ($\leq$ 10K training samples, $\leq$ 500 features). For TabICL, we obtain 36 classification datasets ($\leq$ 100K, $\leq$ 500). Everything but the datasets is identical to Figure 1.

Foundation models dominate by default (and with tuning) within their constraints.

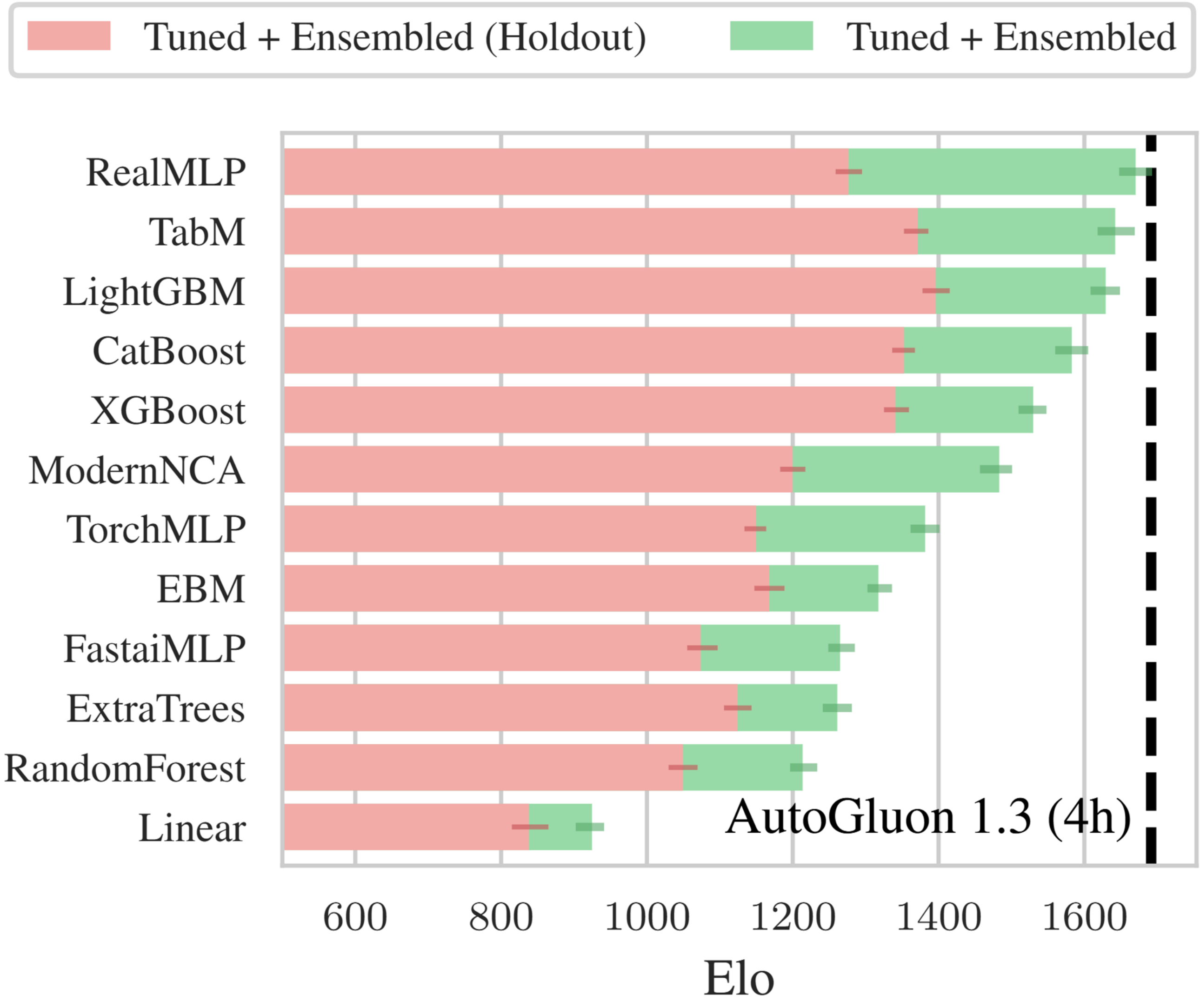
Additional Results: Time trade-off

Efficiency under peak performance:

- **Train+val time** is a must!
    - See TabDPT
- **Ensembling is expensive** but (often) worth it.
- **Deep learning models are more expensive** in general
- **Optimized implementations shine** (e.g. CatBoost)



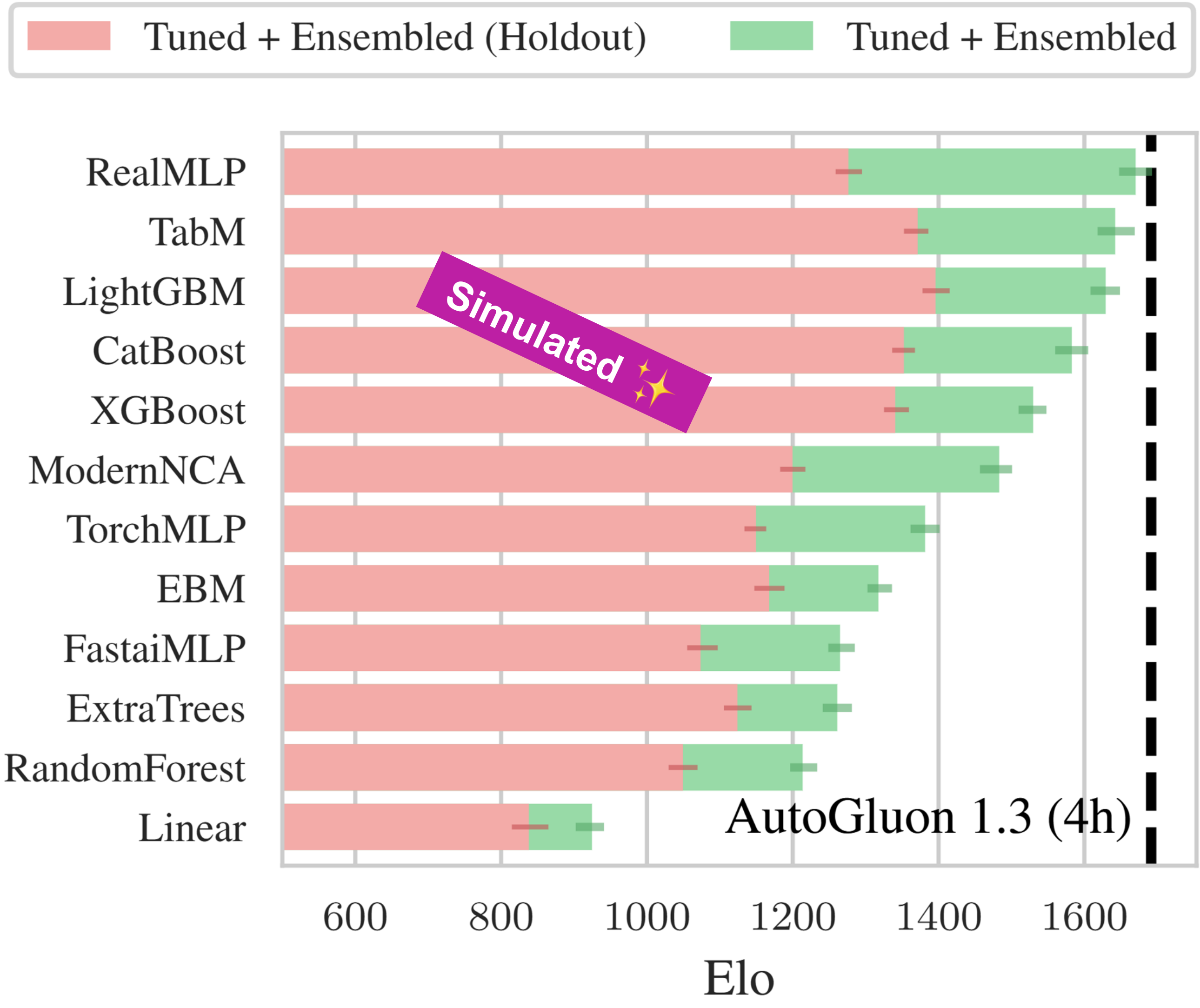
Additional Results: Hold Holdout!



Do not use holdout validation!

- **Worse peak performance** (after HPO + Ensembling)
- Relative **model ranking changes**
- **Unreliable for post-hoc analysis** (e.g., meta-feature analysis)

Additional Results: Hold Holdout!



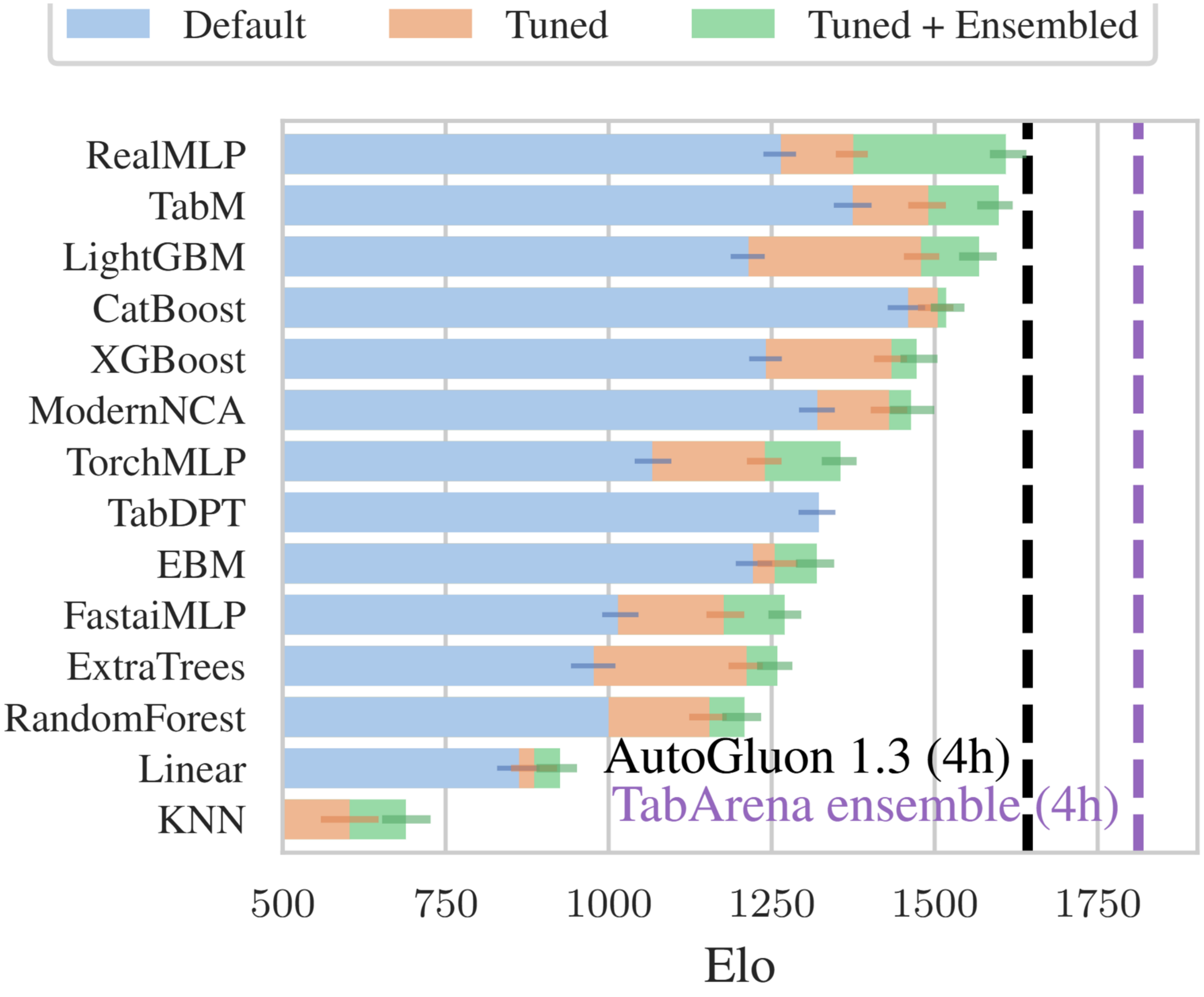
Do not use holdout validation!

- **Worse peak performance** (after HPO + Ensembling)
- Relative **model ranking changes**
- **Unreliable for post-hoc analysis** (e.g., meta-feature analysis)

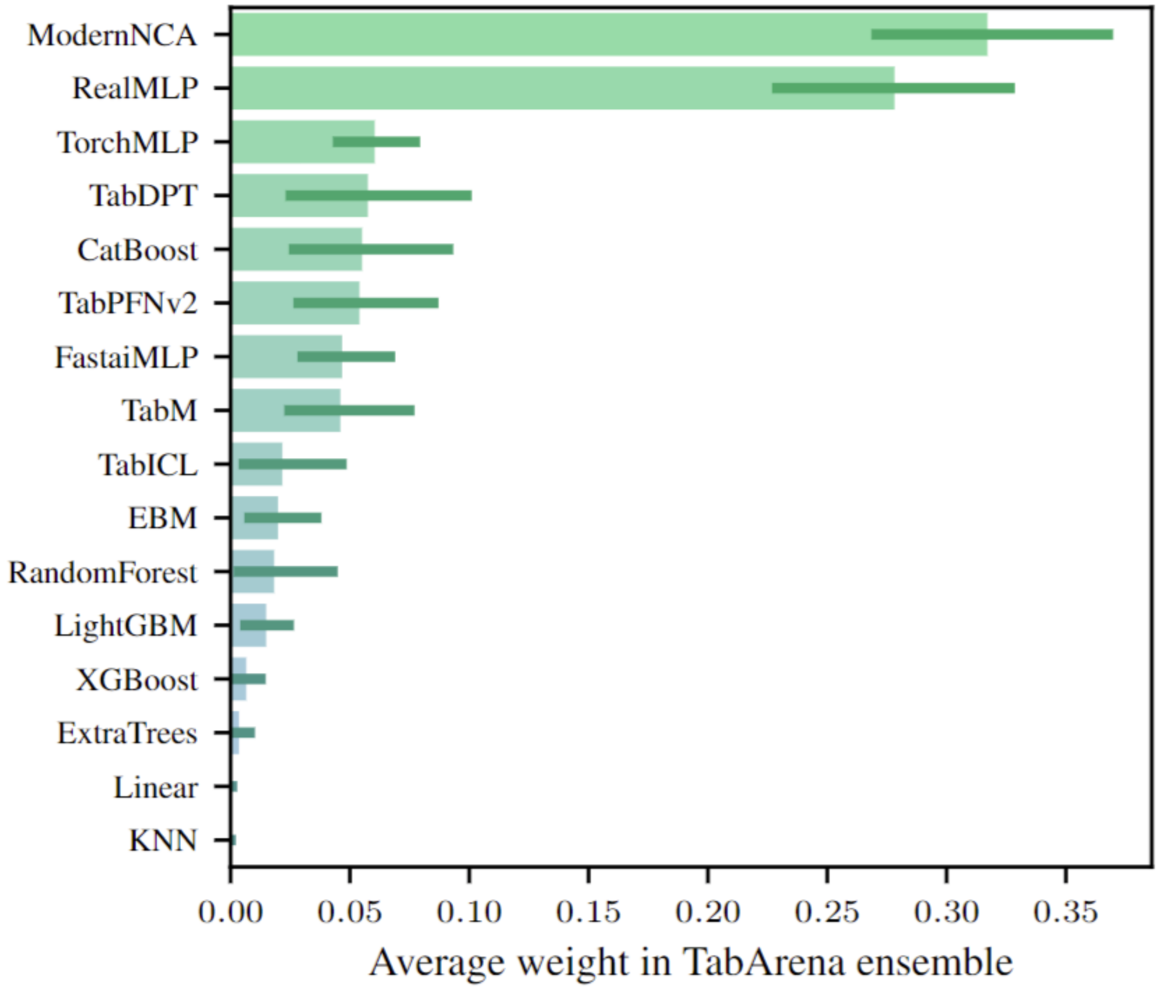
Additional Results: Ensembling

<u>**SOTA model-agnostic ensembles!**</u>

- Fully **simulated ✨ AutoML system** (AutoGluon-like)
- **Significantly better**, even with 4 hours instead of 200 configs
- **The real research goal**; GBDT vs. Deep learning is “just” framing



Additional Results: What are (maybe) important models?



Contributions to ensembles!

- **Contributing most to the ensemble must be important** (?)

Future work:

- Can we **deprecate unimportant models**?
- Approach **likely not representative due to overfitting**

TabArena Ecosystem

Hugging Face Leaderboard: https://tabarena.ai/

TabArena Leaderboard for Predictive Machine Learning on IID Tabular Data

TabArena is a living benchmark system for predictive machine learning on tabular data. The goal of TabArena and its leaderboard is to asses the peak performance of model-specific pipelines.

Datasets

Models

Metrics

Reference Pipeline

More Details

Citation

TabArena Overview

The ranking of all models (with imputation) across various leaderboards.

Search...

Type	Model	Main	Classification	Regression	TabICL-data	TabPFN-data	TabPFN/ICL-data	Lite
	RealMLP (tuned + ensemble)	1	2	1	2	2	4	1
	TabM (tuned + ensemble)	2	1	7	1	3	2	3
	LightGBM (tuned + ensemble)	3	3	5	4	5	7	2
	CatBoost (tuned + ensemble)	4	6	4	6	7	10	4
	CatBoost (tuned)	5	7	6	7	10	11	6
	TabM (tuned)	6	5	12	5	9	8	9
	LightGBM (tuned)	7	8	9	10	11	9	8
	XGBoost (tuned + ensemble)	8	11	8	11	12	15	7
	ModernNCA (tuned + ensemble)	9	14	2	14	17	19	5
	CatBoost (default)	10	10	13	9	13	13	10
	TabPFNv2 (tuned + ensemble)	11	9	15	8	1	1	13
	XGBoost (tuned)	12	13	10	13	16	17	11

Living Benchmark: First Steps

[WIP][New Model] TabFlex ✓
#171 opened 4 days ago by LennartPurucker · updated 4 days ago
new model

Mitra
#161 opened last month by xiyuanzh · updated last week

update to EBM hyperparameters
#158 opened on May 30 by paulbkoch • 1

[WIP][New Model] PerpetualBoosting ✓
#170 opened 4 days ago by LennartPurucker · updated 4 days ago
new model

[WIP][New Model] BETA-TabPFN ✓
#172 opened 4 days ago by LennartPurucker
new model

[WIP][New Model] Dynamic Programming Decision Trees
#176 opened 3 days ago by KohlerHECTOR · updated 3 days ago · 4 tasks
new model

Using all our models – or with the next version of AutoGluon :)

```python
from autogluon.core.data import LabelCleaner
from autogluon.features.generators import AutoMLPipelineFeatureGenerator
from sklearn.datasets import load_breast_cancer
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import train_test_split

# Import a TabArena model
from tabrepo.benchmark.models.ag.realmlp.realmlp_model import RealMLPModel

# Get Data
X, y = load_breast_cancer(return_X_y=True, as_frame=True)
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.5, random_state=42
)
# Preprocessing
feature_generator, label_cleaner = (
    AutoMLPipelineFeatureGenerator(),
    LabelCleaner.construct(problem_type="binary", y=y),
)
X_train, y_train = (
    feature_generator.fit_transform(X_train),
    label_cleaner.transform(y_train),
)
X_test, y_test = feature_generator.transform(X_test), label_cleaner.transform(y_test)

# Train TabArena Model
clf = RealMLPModel()
clf.fit(X=X_train, y=y_train)

# Predict and score
prediction_probabilities = clf.predict_proba(X=X_test)
print("ROC AUC:", roc_auc_score(y_test, prediction_probabilities))
```

https://tabarena.ai/code-examples

Public Dataset Curation: https://tabarena.ai/dataset-curation

tid	did	name	Comments	Year	License	Potential issue	Domain	Required split	Relevant task	Refe	Orig	Include (Andrej)	Explanation (Andrej)	Include (Lennart)	Explanation (Lennart)	Final Decision	Benchmark
2	2	anneal	Not much is known, might be legit; likely from steel production (annealing) as most attributes point to chemical components	1990		Outdated	Tabular	random	Maybe	https:	10.2	No	Not in TabRepo, so likely trivial	Maybe	As long as it is not trivial, this seems to be a legit dataset.	Yes	Tabular
6	6	letter	Numerical features extracted from images of letters; also includes data augmentation of the images	1991		Image domain	Image	-	No	P. W.	https	No	Image	No	Image	No	Image
11	11	balance-scale	generated data to model a pyschological experiment	1976		trivial, artificial, determinisitc	Artificial	-	No	Siegle	https	No	Artificial	No	Artificial	No	Deterministic
15	15	breast-w	Nowadays solved differently, domain features extracted from images	1995		Maybe Image domain, outdated	Image, tabu	random	No	This	https	No	Image	No	Image, Outdated	No	Image
24	24	mushroom	New knowledge about mushrooms likely is available nowadays; dataset from a book (I guess);	1981		trivial	Tabular	random	No	10.24	Aud	No	Trivial	No	Trivial	No	Scientific Discovery
26	26	nursery	Data was derived from a hierarchical decision model, likely trivial as samples cover all possible values; also originally a regression task; no ground truth that the	1989		Outdated, Simulated, ethical issues as reproduces biases	Simulated	-	Maybe	https:	https	No	Simulated	No	Simulated/ Ethical	No	Artificial/Simulated
28	28	optdigits	Yet another handwritten digits dataset…	1995		Image domain	Image	-	No	https:	https	No	Image	No	Image	No	Image
30	30	page-blocks	Grouped data, random splits may be inappropriate; meta-features extract from	1995		Image domain	Image	Grouped	No	https:	https	No	Image	No	Image	No	Image
32	32	pendigits	Yet another handwritten digits dataset…, Grouped data, random splits may be inappropriate, either image or weird	1998		Other domain	Image, Pixe	Grouped	No	https:	https	No	Image	No	Image, heavily preprocess	No	Image
37	37	diabetes	Rather interpretability than predictive performance task, nowadays done differently	1988		Outdated	Tabular	random	Maybe	Smith	Miss	Yes	Fits our criteria, but TabRepo resutls for this dataset are pretty random	Yes	No objection	Yes	Tabular
41	42	soybean	Some infrequent classes should not be used for prediction, may be outdated, maybe also rather an interpretability task, might require time split as date is available; categorical and nan values already preprocessed	1988		Preprocessing, Historic problems with classes (see e-mails from UCI download)	Tabular	random	Maybe	R.S.	https	Conditiona	Needs proper task definition and preprocessing.	Unclear	After some preprocessing, I can see this being added	no	Tiny data
43	44	spambase	Text formated as table, outdated task / solution, not meta-features but	1998		Text domain	Text	-	No	https:	https	No	Text	No	Text	No	Text
45	46	splice	Domain specific methods might exist; preprocssed DNA data	1991		-	Special tabu	random	Maybe	? http	https	Yes	Special domain and quite old, but no particular reason to exclude.	Yes	No objection	Yes	Tabular
49	50	tic-tac-toe	GBDTs & NNs perform perfectly	1991		trivial, artificial, determinisitc	Artificial	random	No	?	https	No	Artificial	No	Deterministic	No	Deterministic
58	60	waveform-500	19/40 features are pure noise, data describes waves and was simulated; data from a book	1984		Artificial, Deterministic with noise	Artificial	random	No	Breim	https	No	Artificial	No	Deterministic	No	Deterministic
219	151	electricity	leak if not temporal split; manually normalized but unclear how; day-wise	1996-1998		temporal split	tabular	temporal	Maybe	M. Ha	?	No	Temporal split	No	Temporal split	No	Temporal Tabular
223	155	pokerhand	game data, normalized version, solvable by a look-up table or deterministic algorithm	2002		artificial, deterministic	Artificial	random	No	https:	https	No	Artificial	No	Deterministic	No	Deterministic

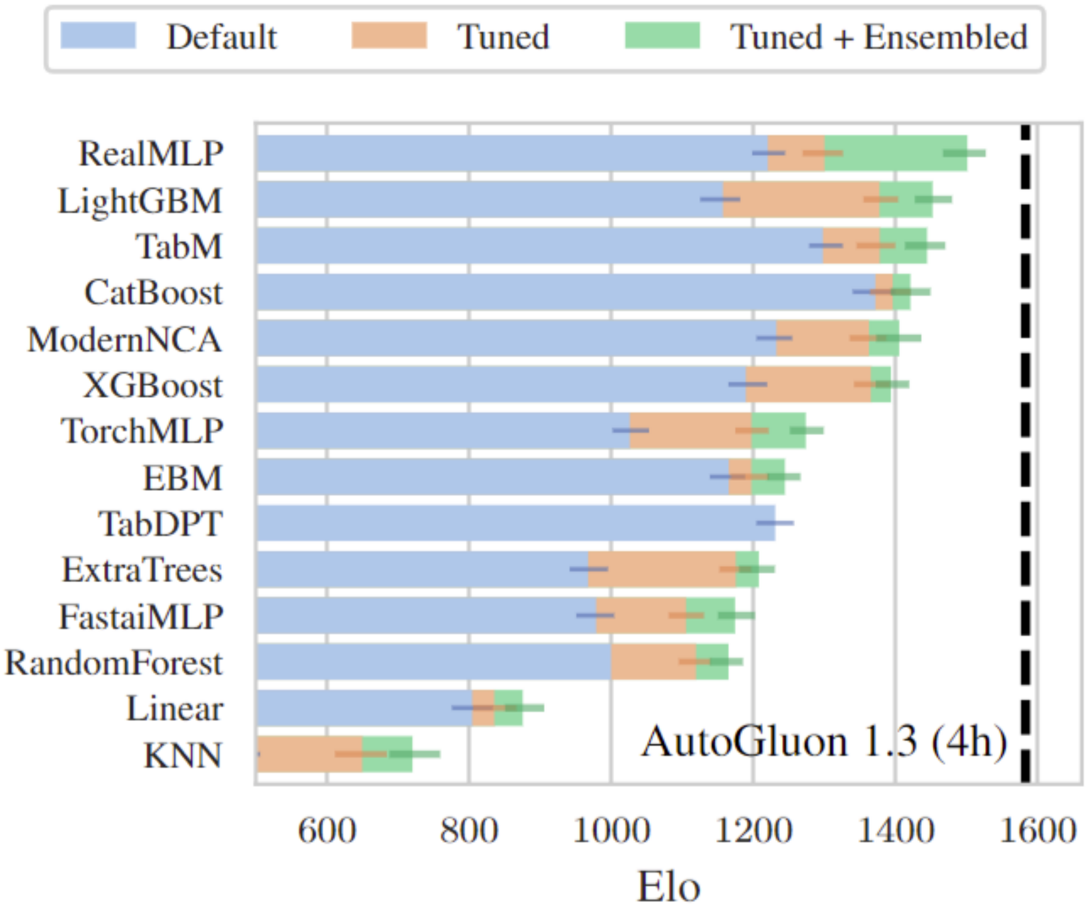
Public Dataset Curation: https://tabarena.ai/dataset-curation

tid	did	name	Comments	Year	License	Potential issue	Domain	Required split	Relevant task	Refe	Orig	Include (Andrej)	Explanation (Andrej)	Include (Lennart)	Explanation (Lennart)	Final Decision	Benchmark
2	2	anneal	Not much is known, might be legit; likely from steel production (annealing) as most attributes point to chemical components	1990		Outdated	Tabular	random	Maybe	https:	10.2	No	Not in TabRepo, so likely trivial	Maybe	As long as it is not trivial, this seems to be a legit dataset.	Yes	Tabular
6	6	letter	Numerical features extracted from images of letters; also includes data augmentation of the images	1991		Image domain	Image	-	No	P. W.	http:	No	Image	No	Image	No	Image
11	11	balance-scale	generated data to model a pyschological experiment	1976		trivial, artificial, determinisitc	Artificial	-	No	Siegle	http:	No	Artificial	No	Artificial	No	Determini stic
15	15	breast-w	Nowadays solved differently, domain features extracted from images	1995		Maybe Image domain, outdated	Image, tabu	random	No	This	http:	No	Image	No	Image, Outdated	No	Image
24	24	mushroom	New knowledge about mushrooms likely is available nowadays; dataset from a book (I guess);	1981		trivial	Tabular	random	No	10.24	Aud	No	Trivial	No	Trivial	No	Scientific Discovery
26	26	nursery	Data was derived from a hierarchical decision model, likely trivial as samples cover all possible values; also originally a regression task; no ground truth that the	1989		Outdated, Simulated, ethical issues as reproduces biases	Simulated	-	Maybe	https:	http:	No	Simulated	No	Simulated/ Ethical	No	Artificial/S imulated
28	28	optdigits	Yet another handwritten digits dataset...	1995		Image domain	Image	-	No	https:	http:	No	Image	No	Image	No	Image
30	30	page-blocks	Grouped data, random splits may be inappropriate; meta-features extract from	1995		Image domain	Image	Grouped	No	https:	http:	No	Image	No	Image	No	Image
32	32	pendigits	Yet another handwritten digits dataset..., Grouped data, random splits may be inappropriate, either image or weird	1998		Other domain	Image, Pixe	Grouped	No	https:	http:	No	Image	No	Image, heavily preprocess	No	Image
37	37	diabetes	Rather interpretability than predictive performance task, nowadays done differently	1988		Outdated	Tabular	random	Maybe	Smith	Miss	Yes	Fits our criteria, but TabRepo resutls for this dataset are pretty random	Yes	No objection	Yes	Tabular
41	42	soybean	Some infrequent classes should not be used for prediction, may be outdated, maybe also rather an interpretability task, might require time split as date is available; categorical and nan values already preprocessed	1988		Preprocessing, Historic problems with classes (see e-mails from UCI download)	Tabular	random	Maybe	R.S.	http:	Conditiona	Needs proper task definition and preprocessing.	Unclear	After some preprocessi ng, I can see this being added	no	Tiny data
43	44	spambase	Text formated as table, outdated task / solution, not meta-features but	1998		Text domain	Text	-	No	https:	http:	No	Text	No	Text	No	Text
45	46	splice	Domain specific methods might exist; preprocssed DNA data	1991		-	Special tabu	random	Maybe	? http	http:	Yes	Special domain and quite old, but no particular reason to exclude.	Yes	No objection	Yes	Tabular
49	50	tic-tac-toe	GBDTs & NNs perform perfectly	1991		trivial, artificial, determinisitc	Artificial	random	No	?	http:	No	Artificial	No	Determinist ic	No	Determini stic
58	60	waveform-500	19/40 features are pure noise, data describes waves and was simulated; data from a book	1984		Artificial, Deterministic with noise	Artificial	random	No	Breim	http:	No	Artificial	No	Determinist ic	No	Determini stic
219	151	electricity	leak if not temporal split; manually normalized but unclear how; day-wise	1996-1998		temporal split	tabular	temporal	Maybe	M. Ha	?	No	Temporal split	No	Temporal split	No	Temporal Tabular
223	155	pokerhand	game data, normalized version, solvable by a look-up table or deterministic algorithm	2002		artificial, deterministic	Artificial	random	No	https:	http:	No	Artificial	No	Determinist ic	No	Determini stic

Public Dataset Curation: https://tabarena.ai/dataset-curation

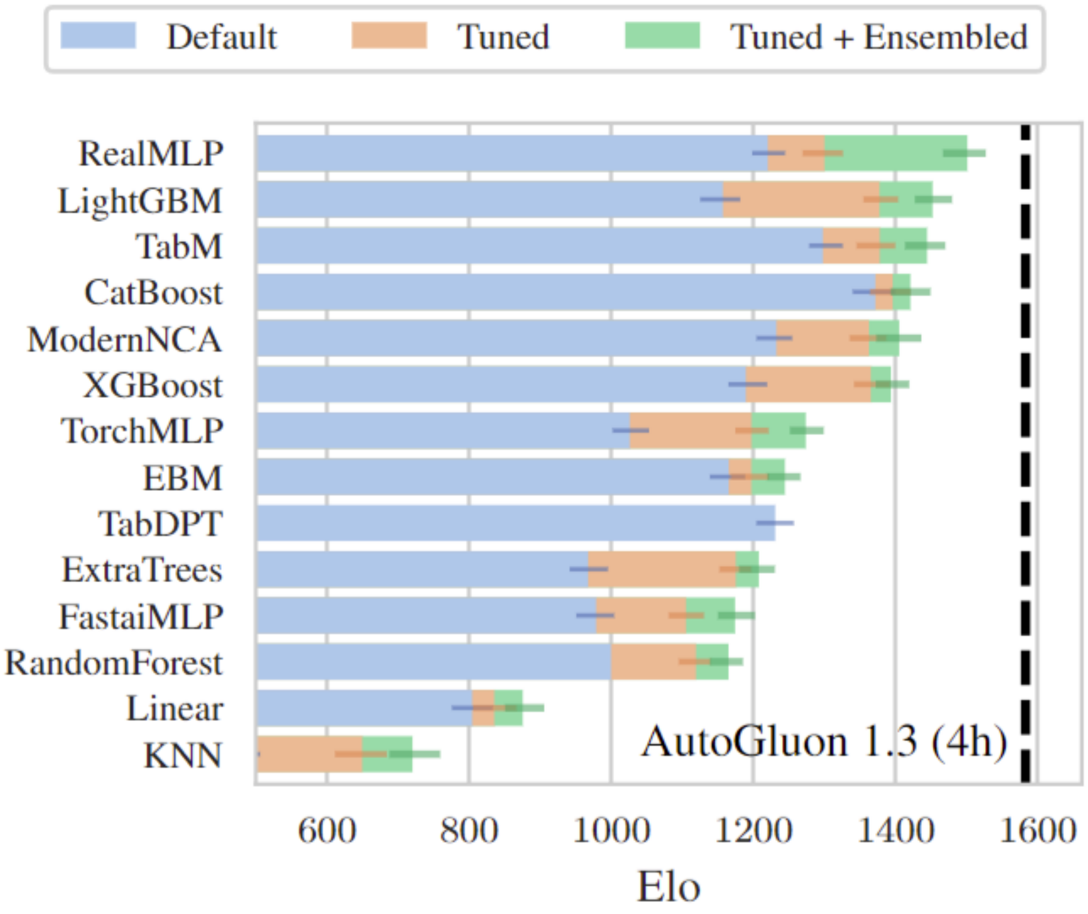
tid	did	name	Comments	Year	License	Potential issue	Domain	Required split	Relevant task	Refe	Orig	Include (Andrej)	Explanation (Andrej)	Include (Lennart)	Explanation (Lennart)	Final Decision	Benchmark
2	2	anneal	Not much is known, might be legit; likely from steel production (annealing) as most attributes point to chemical components	1990		Outdated	Tabular	random	Maybe	https:	10.2	No	Not in TabRepo, so likely trivial	Maybe	As long as it is not trivial, this seems to be a legit dataset.	Yes	Tabular
6	6	letter	Numerical features extracted from images of letters; also includes data augmentation of the images	1991		Image domain	Image	-	No	P. W.	http	No	Image	No	Image	No	Image
11	11	balance-scale	generated data to model a pyschological experiment	1976		trivial, artificial, determinisitc	Artificial	-	No	Siegle	http	No	Artificial	No	Artificial	No	Determini stic
15	15	breast-w	Nowadays solved differently, domain features extracted from images	1995		Maybe Image domain, outdated	Image, tabu	random	No	This	http	No	Image	No	Image, Outdated	No	Image
24	24	mushroom	New knowledge about mushrooms likely is available nowadays; dataset from a book (I guess);	1981		trivial	Tabular	random	No	10.24	Aud	No	Trivial	No	Trivial	No	Scientific Discovery
26	26	nursery	Data was derived from a hierarchical decision model, likely trivial as samples cover all possible values; also originally a regression task; no ground truth that the	1989		Outdated, Simulated, ethical issues as reproduces biases	Simulated	-	Maybe	https:	http	No	Simulated	No	Simulated/ Ethical	No	Artificial/S imulated
28	28	optdigits	Yet another handwritten digits dataset…	1995		Image domain	Image	-	No	https:	http	No	Image	No	Image	No	Image
30	30	page-blocks	Grouped data, random splits may be inappropriate; meta-features extract from	1995		Image domain	Image	Grouped	No	https:	http	No	Image	No	Image	No	Image
32	32	pendigits	Yet another handwritten digits dataset…, Grouped data, random splits may be inappropriate, either image or weird	1998		Other domain	Image, Pixe	Grouped	No	https:	http	No	Image	No	Image, heavily preprocess	No	Image
37	37	diabetes	Rather interpretability than predictive performance task, nowadays done differently	1988		Outdated	Tabular	random	Maybe	Smith	Miss	Yes	Fits our criteria, but TabRepo resutls for this dataset are pretty random	Yes	No objection	Yes	Tabular
41	42	soybean	Some infrequent classes should not be used for prediction, may be outdated, maybe also rather an interpretability task, might require time split as date is available; categorical and nan values already preprocessed	1988		Preprocessing, Historic problems with classes (see e-mails from UCI download)	Tabular	random	Maybe	R.S.	http	Conditiona	Needs proper task definition and preprocessing.	Unclear	After some preprocessing, I can see this being added	no	Tiny data
43	44	spambase	Text formated as table, outdated task / solution, not meta-features but	1998		Text domain	Text	-	No	https:	http	No	Text	No	Text	No	Text
45	46	splice	Domain specific methods might exist; preprocssed DNA data	1991		-	Special tabu	random	Maybe	? http	http	Yes	Special domain and quite old, but no particular reason to exclude.	Yes	No objection	Yes	Tabular
49	50	tic-tac-toe	GBDTs & NNs perform perfectly	1991		trivial, artificial, determinisitc	Artificial	random	No	?	http	No	Artificial	No	Deterministic	No	Determini stic
58	60	waveform-500	19/40 features are pure noise, data describes waves and was simulated; data from a book	1984		Artificial, Deterministic with noise	Artificial	random	No	Breim	http	No	Artificial	No	Deterministic	No	Determini stic
219	151	electricity	leak if not temporal split; manually normalized but unclear how; day-wise	1996-1998		temporal split	tabular	temporal	Maybe	M. Ha	?	No	Temporal split	No	Temporal split	No	Temporal Tabular
223	155	pokerhand	game data, normalized version, solvable by a look-up table or deterministic algorithm	2002		artificial, deterministic	Artificial	random	No	https:	http	No	Artificial	No	Deterministic	No	Determini stic
															Likely too		

Cheaper Evaluation For Papers: TabArena Lite



Only one repeat: 816× fewer jobs

Cheaper Evaluation For Papers: TabArena Lite



Benchmarking TabFlex with TabArena Lite takes about 20 minutes

Only one repeat: 816× fewer jobs

TabArena-v1.0?

Open Problems and Future Work

Datasets

- **More data diversity**: domains, tiny, large, non-IID, with text, with images, …
- Evaluation with (expert) **preprocessing and feature engineering**

Benchmarking

- **Overfitting** the benchmark (?)
- **Bias from data contamination** due to pretraining foundation models or LLMs
- More **realistic user constraints and metrics**

Takeaways

Benchmarks ❤️	TabArena is a truly representative benchmark for machine learning on small- to medium sized IID tabular data.
SOTA with Ensembling 📈	CatBoost shines. Deep learning with ensembling dominates. Promising future for foundation models!
Living benchmark baby!	TabArena will be updated and support more (non-IID) data, models, and tasks.

Thank you, any questions?

Leaderboard: https://tabarena.ai
Paper: https://arxiv.org/abs/2506.16791
Code: https://tabarena.ai/code

Nick
Erickson

Lennart
Purucker

Andrej
Tschalzev

David
Holzmüller

Prateek
Mutalik Desai

**David
Salinas**

Frank
Hutter

Part III

A Case for Openness

The Case of LLMs

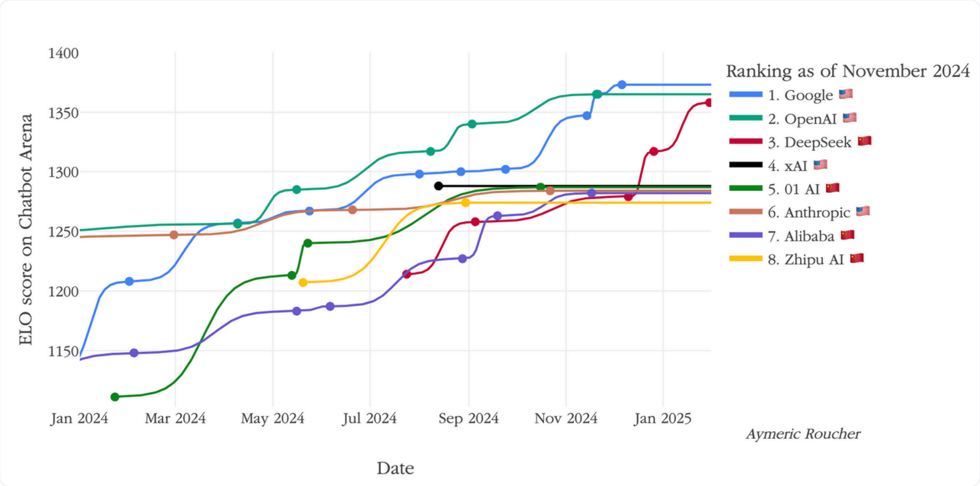
The Case of LLMs

- Currently an arm race

The Case of LLMs

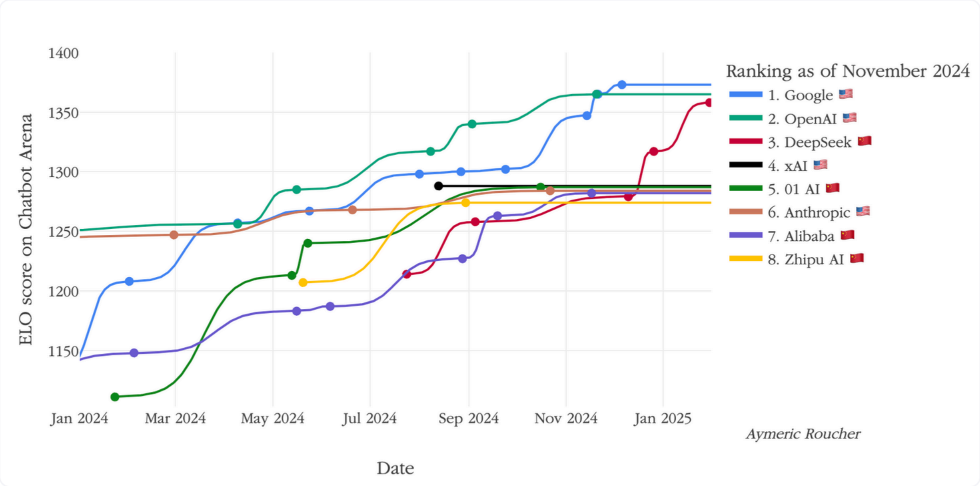
- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time

The Case of LLMs



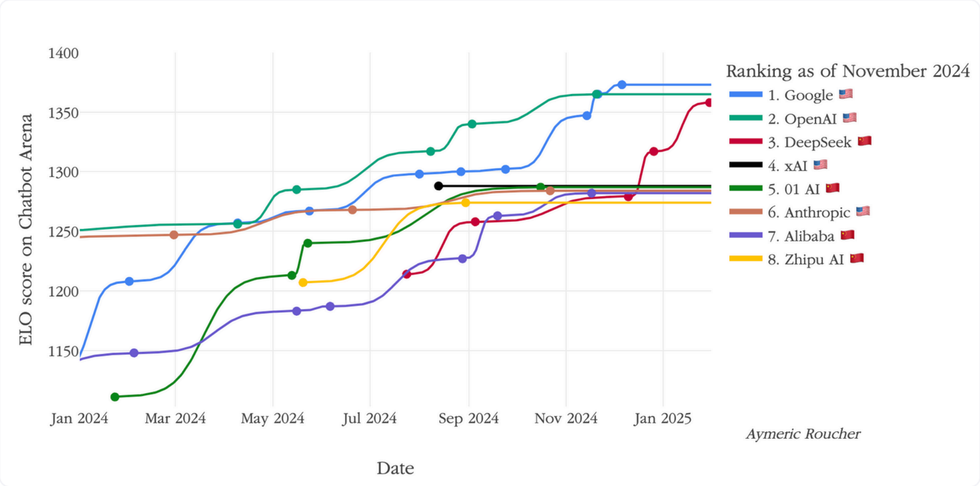
- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time

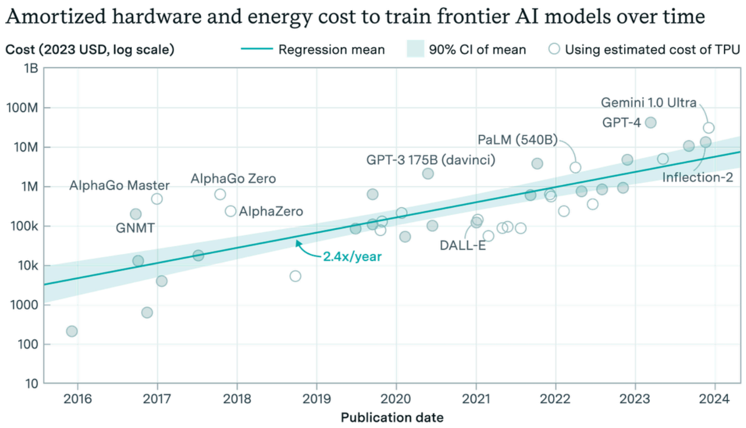
The Case of LLMs



- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time
    - Scaling compute efficiency (the bitter lesson from Sutter)

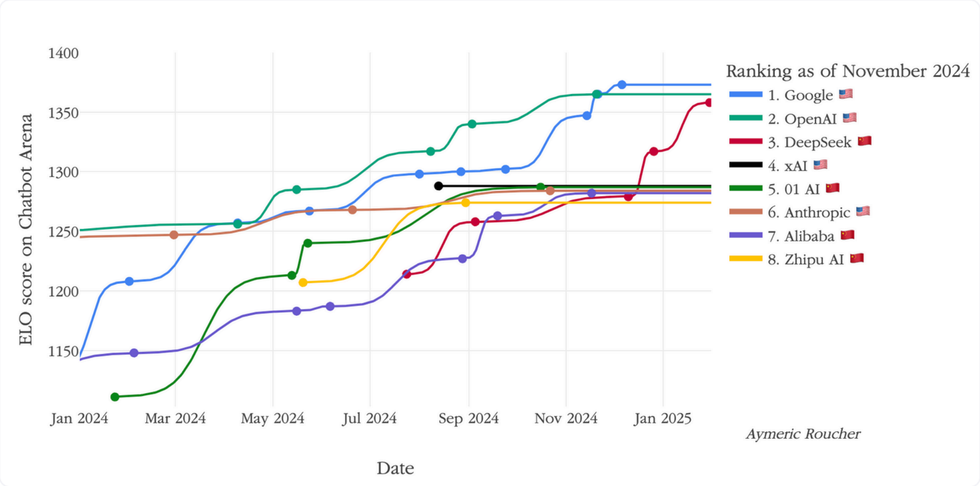
The Case of LLMs

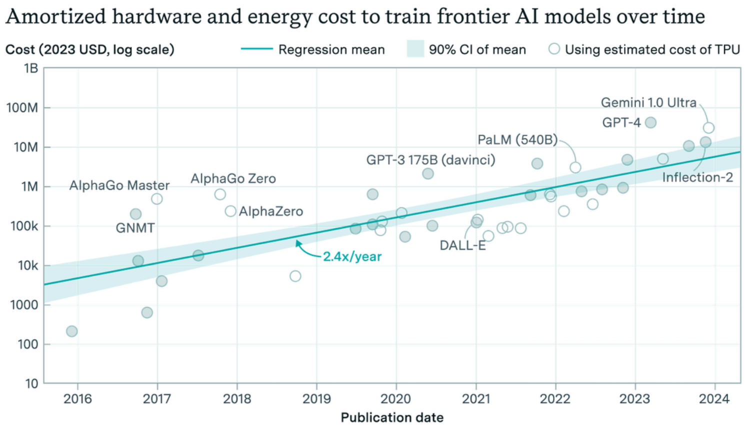




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)

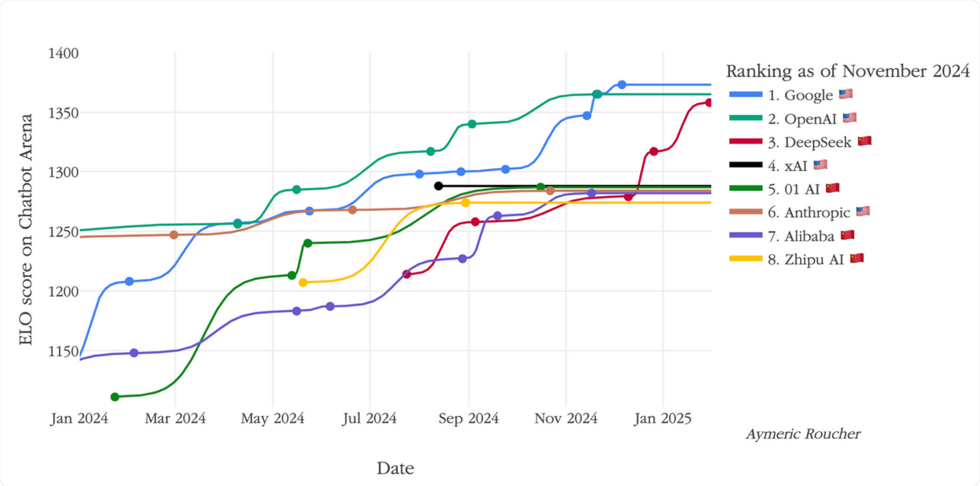
The Case of LLMs

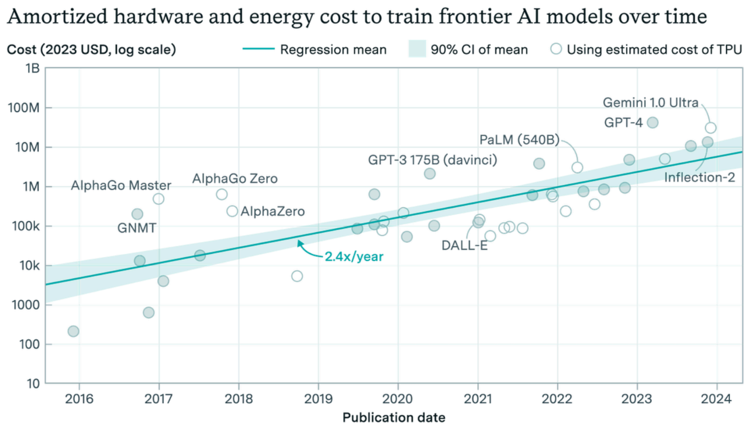




- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time
    - Scaling compute efficiency (the bitter lesson from Sutter)
    - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls

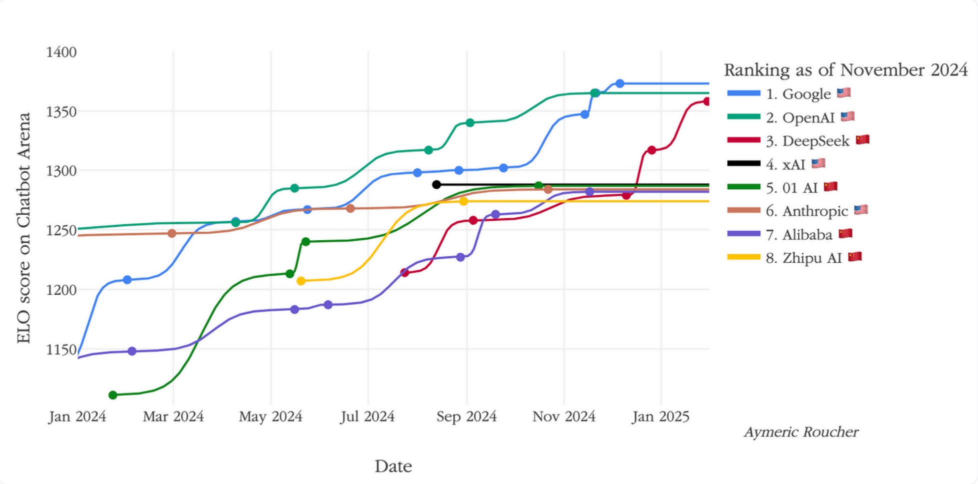
The Case of LLMs

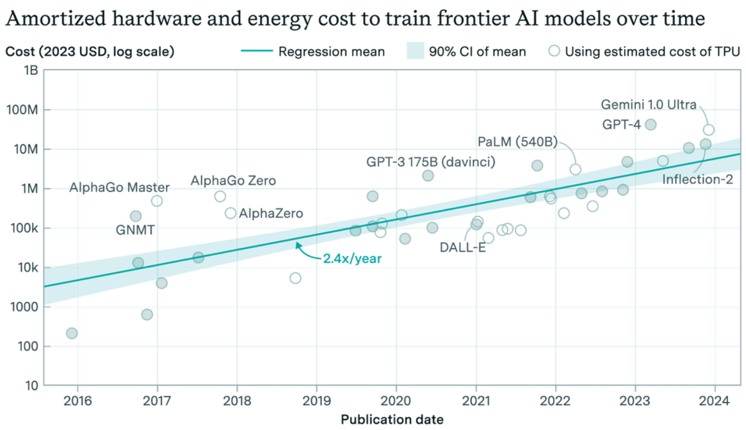




- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
  - Large ecological cost and human cost (safety annotations done by South developing countries)

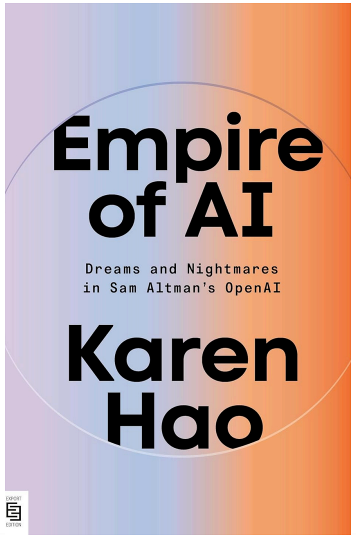
The Case of LLMs



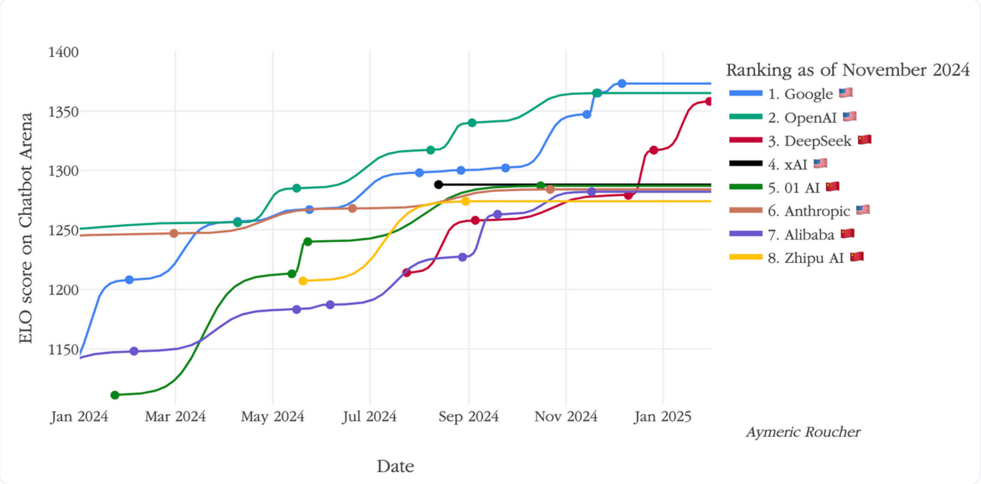


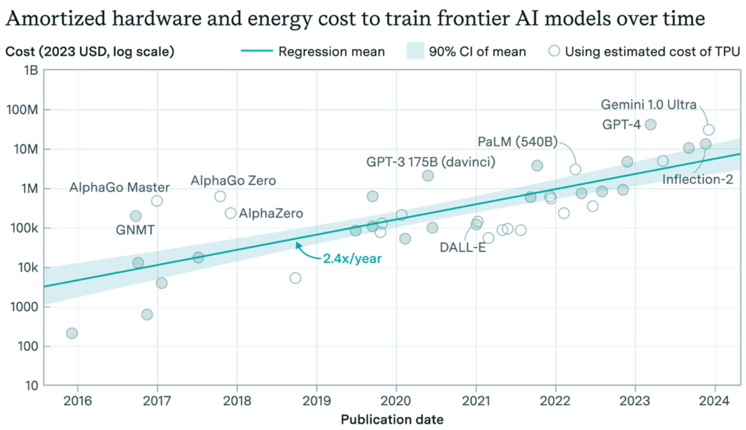
- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
  - Large ecological cost and human cost (safety annotations done by South developing countries)

Recommended reading 📚



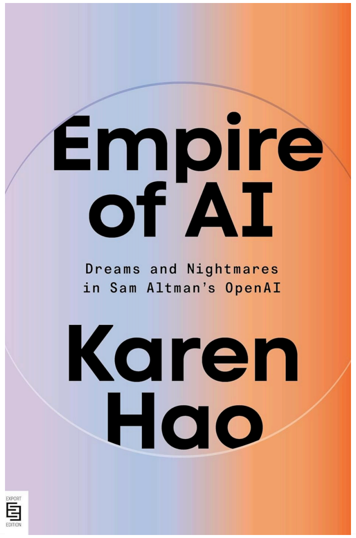
The Case of LLMs



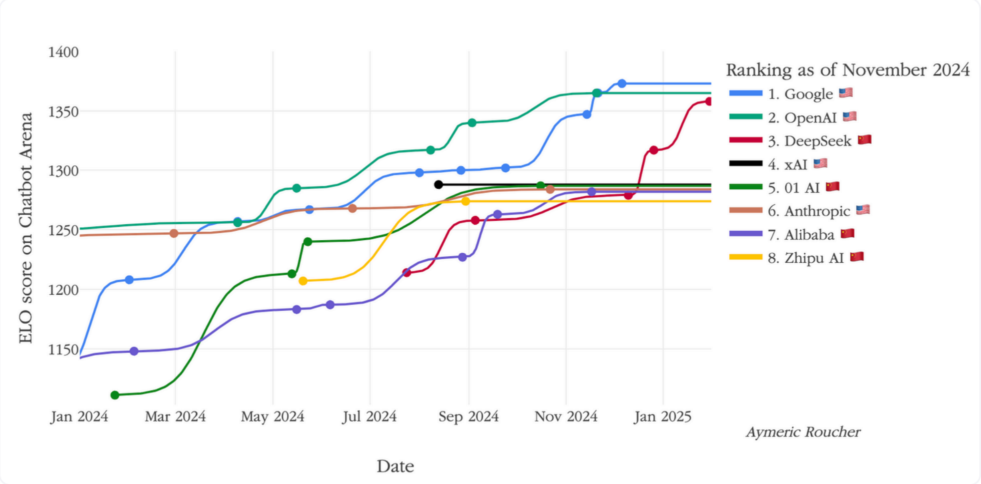


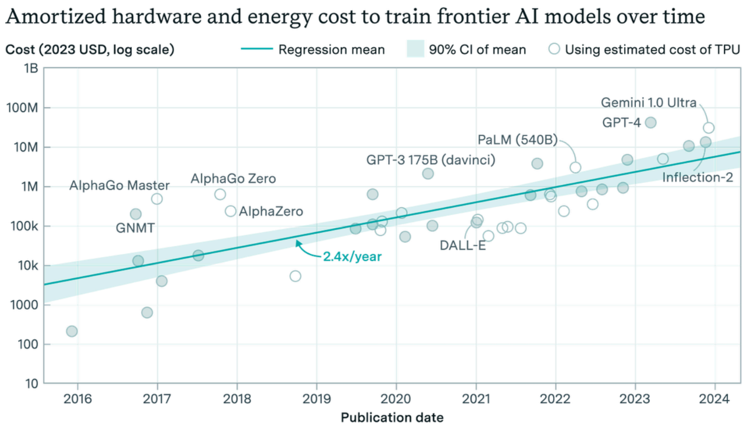
- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time
    - Scaling compute efficiency (the bitter lesson from Sutter)
    - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
    - Large ecological cost and human cost (safety annotations done by South developing countries)
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models

Recommended reading 📚



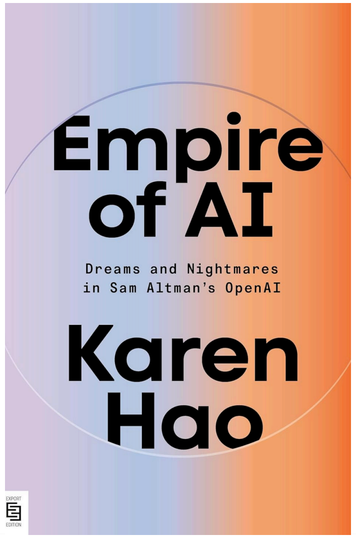
The Case of LLMs



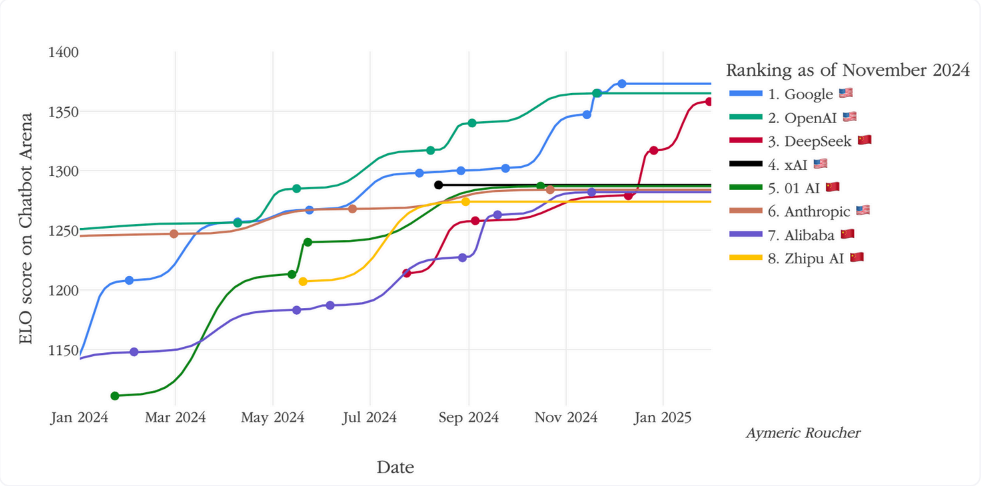


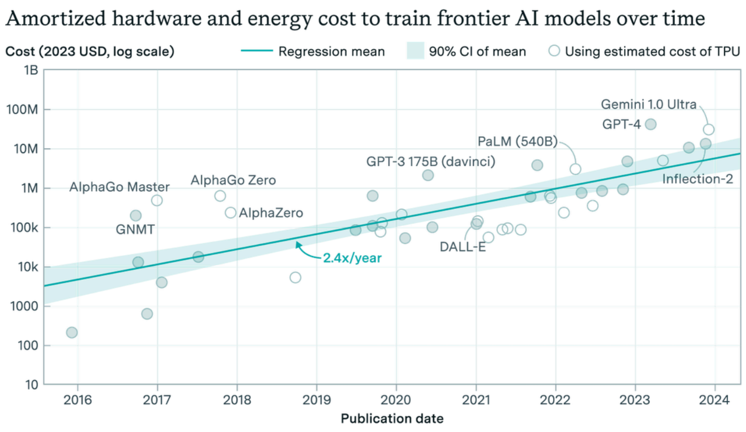
- Currently an arm race
    - One world with N actors developing N models and sharing less and less over time
    - Scaling compute efficiency (the bitter lesson from Sutter)
    - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
    - Large ecological cost and human cost (safety annotations done by South developing countries)
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models
    - Open-weights: Meta, Google, Mistral, …

Recommended reading 📚



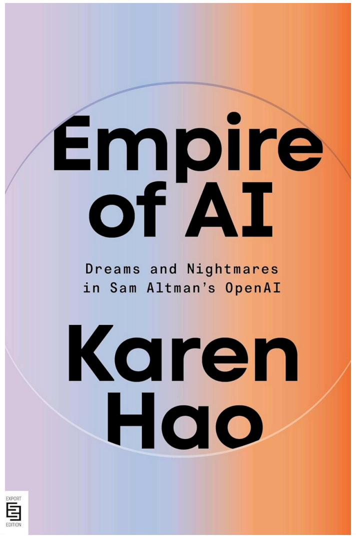
The Case of LLMs





- Currently an arm race
  - One world with N actors developing N models and sharing less and less over time
  - Scaling compute efficiency (the bitter lesson from Sutter)
  - Algorithmic progress: ~4x/year? https://www.darioamodei.com/post/on-deepseek-and-export-controls
  - Large ecological cost and human cost (safety annotations done by South developing countries)
- Alternate model: companies & universities sharing open-weight models and sometimes fully open models
  - Open-weights: Meta, Google, Mistral, …
  - Fully open models: Stanford, AllenAI institute, Apple …

Recommended reading 📚

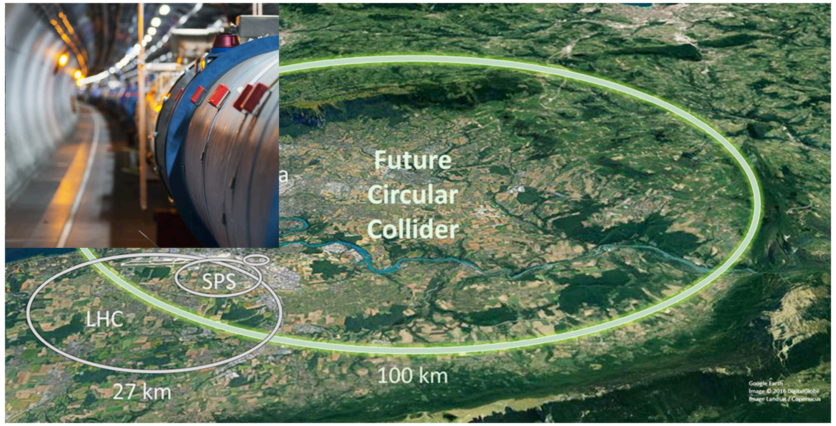


A Case of Openness

Some of humanity largest projects

A Case of Openness

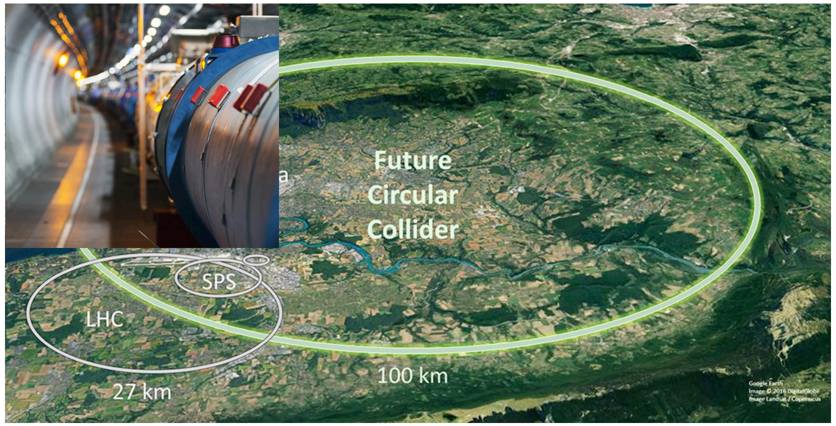
Some of humanity largest projects

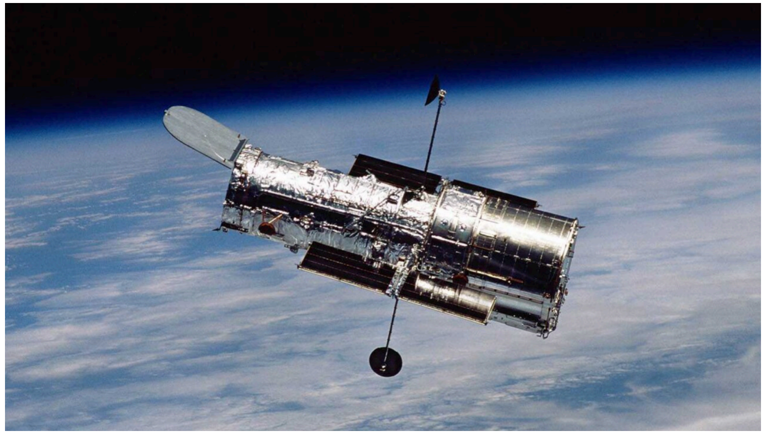


LHC: $5 Billion, 23 countries

A Case of Openness

Some of humanity largest projects

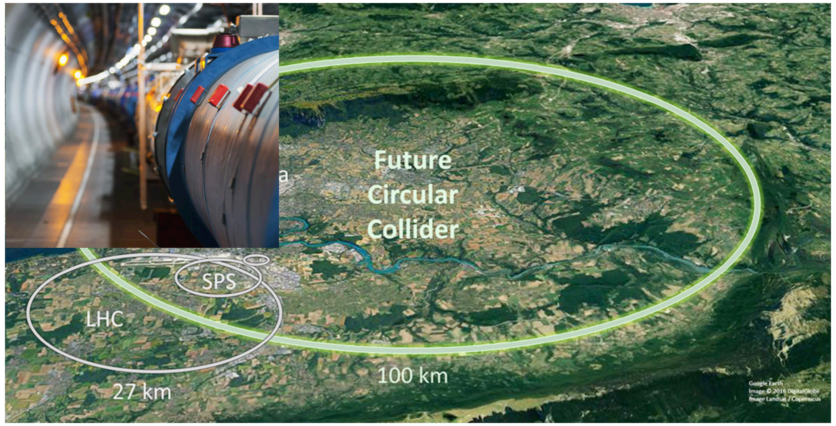




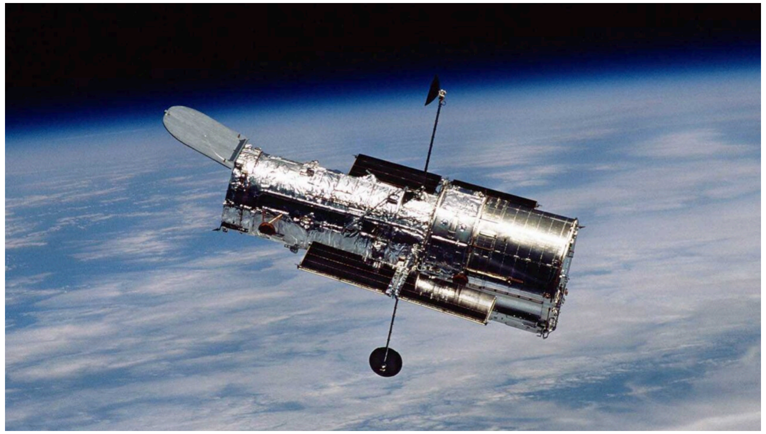
LHC: $5 Billion, 23 countries

A Case of Openness

Some of humanity largest projects



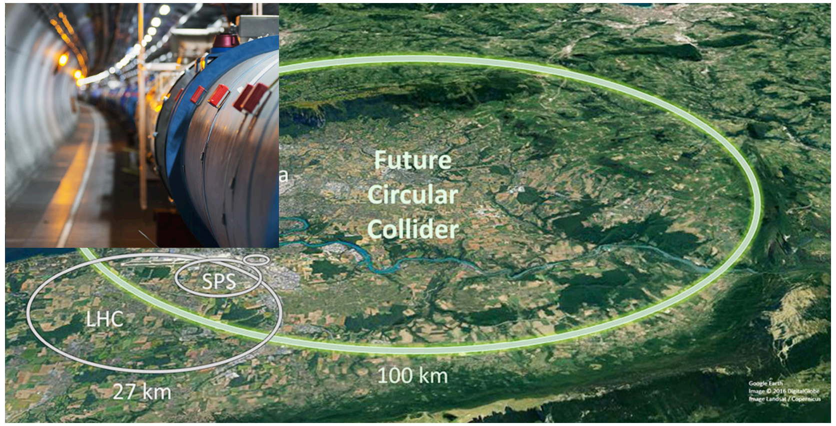
LHC: $5 Billion, 23 countries



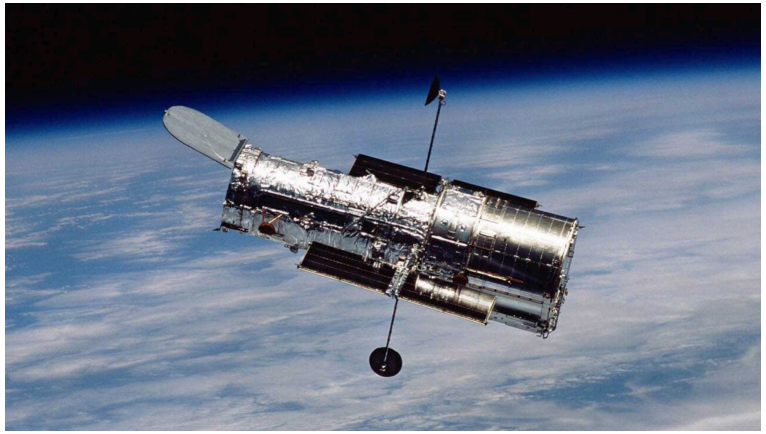
Hubble $16 billion, 11 countries

A Case of Openness

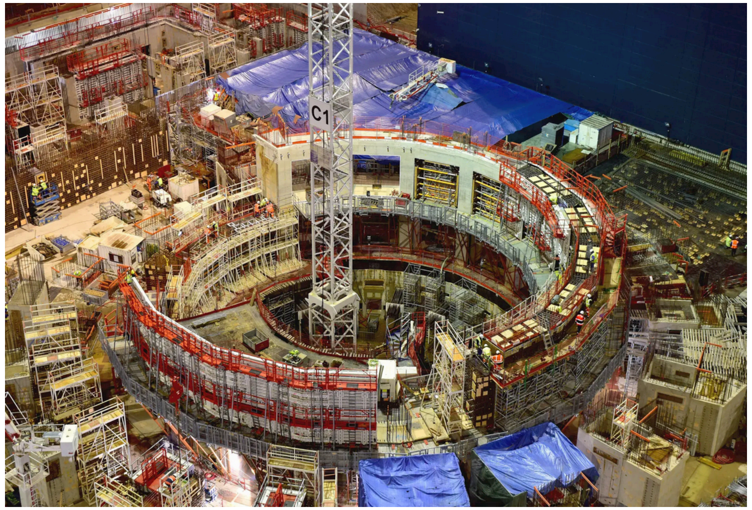
Some of humanity largest projects



LHC: $5 Billion, 23 countries

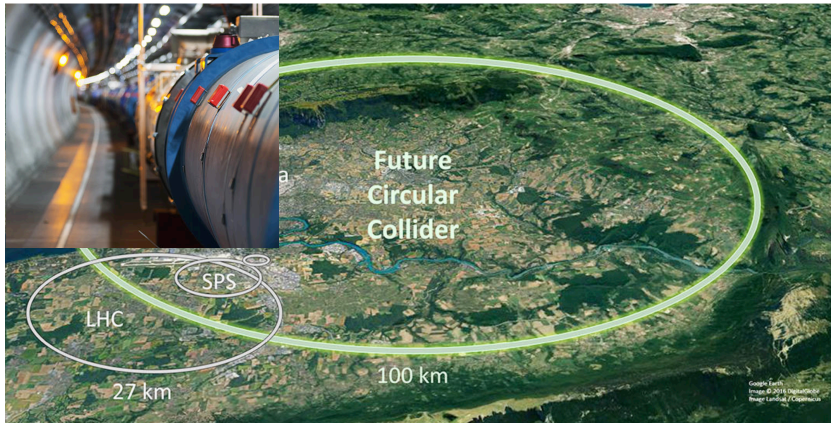


Hubble $16 billion, 11 countries

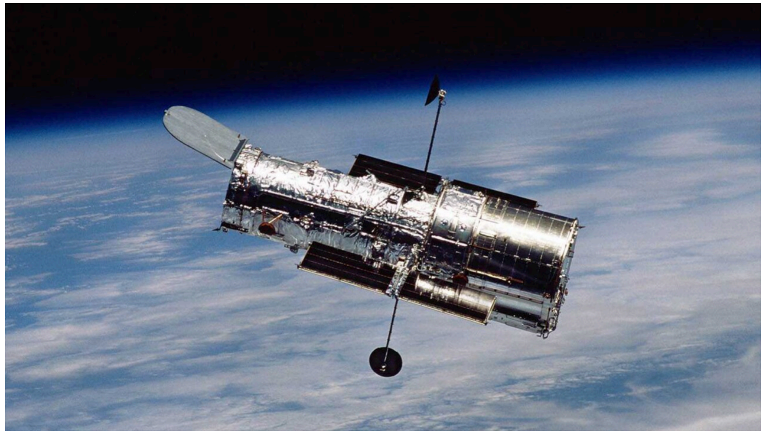


A Case of Openness

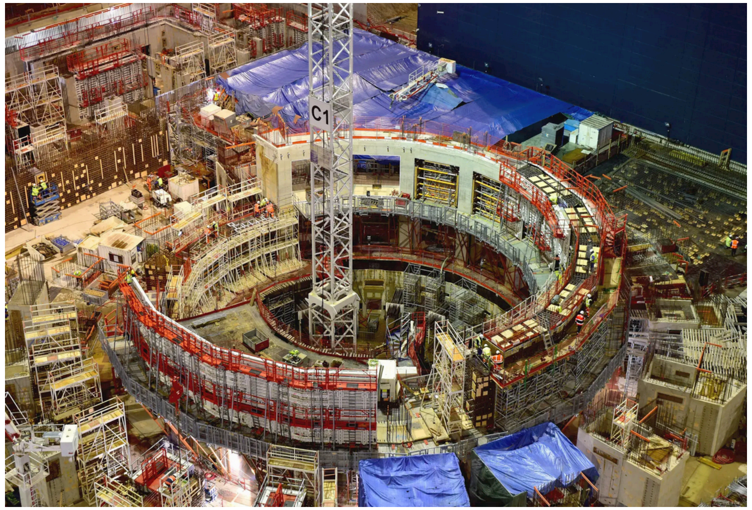
Some of humanity largest projects



LHC: $5 Billion, 23 countries



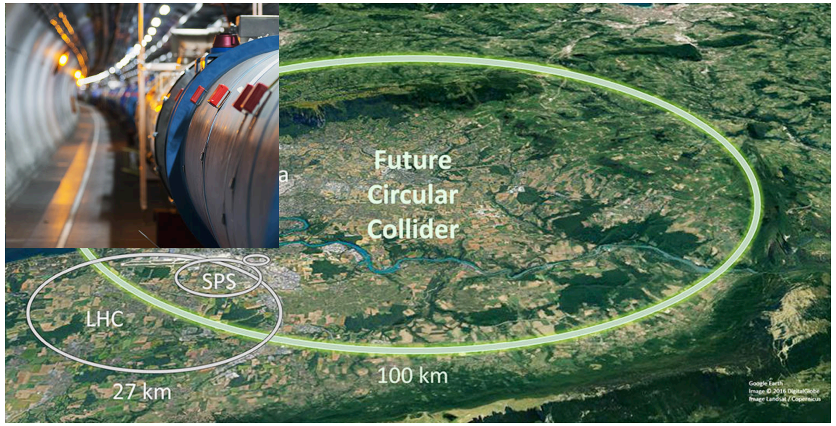
Hubble $16 billion, 11 countries



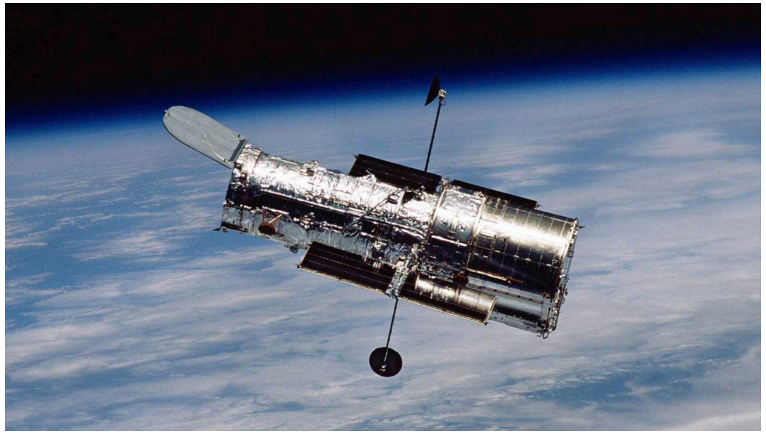
ITER: $45 Billion, 35 countries

A Case of Openness

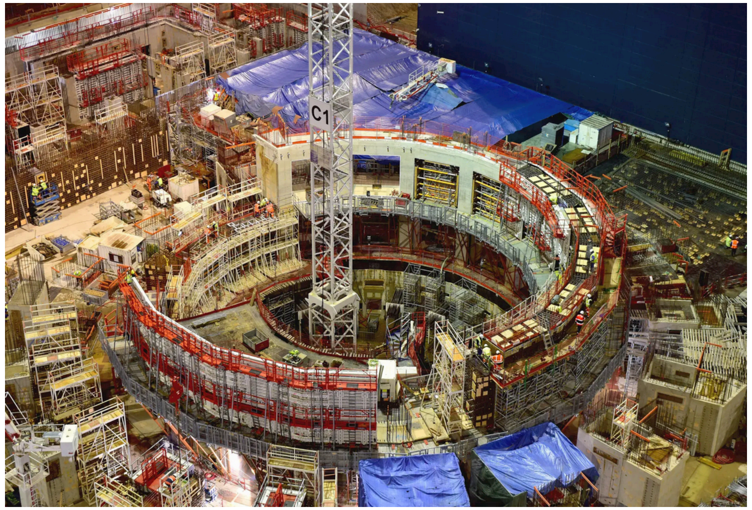
Some of humanity largest projects



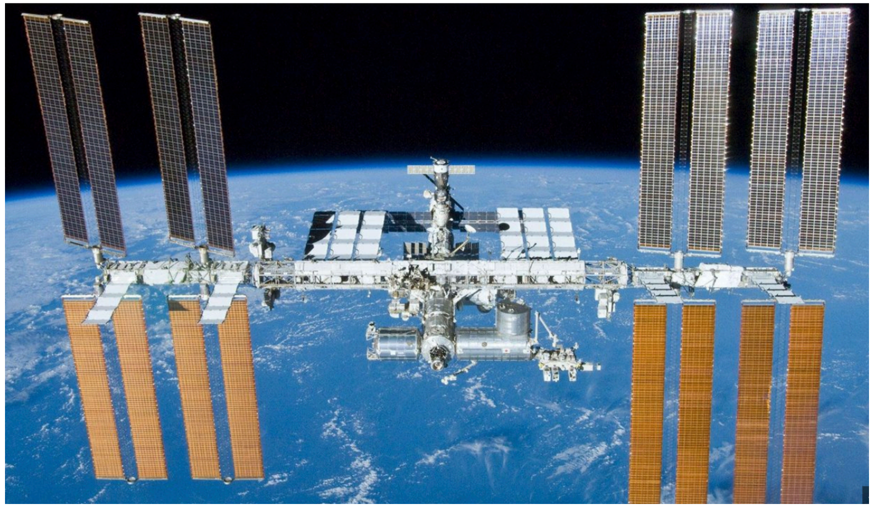
LHC: $5 Billion, 23 countries



Hubble $16 billion, 11 countries

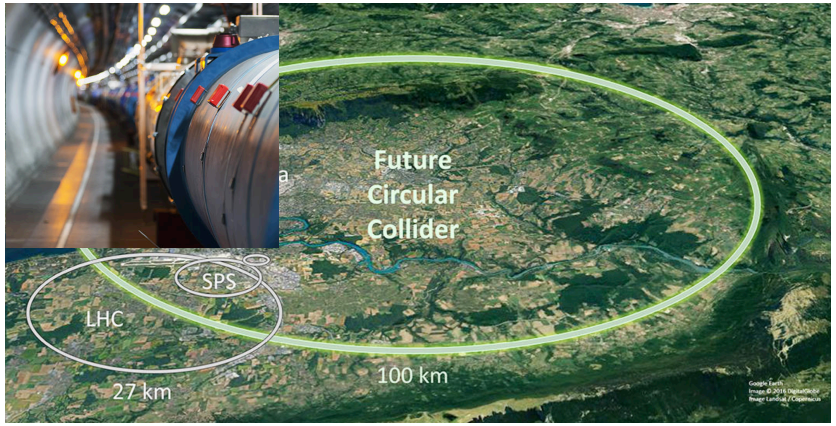


ITER: $45 Billion, 35 countries

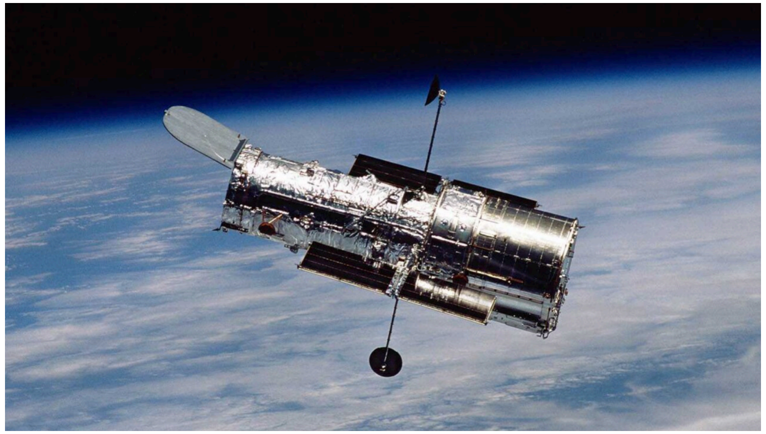


A Case of Openness

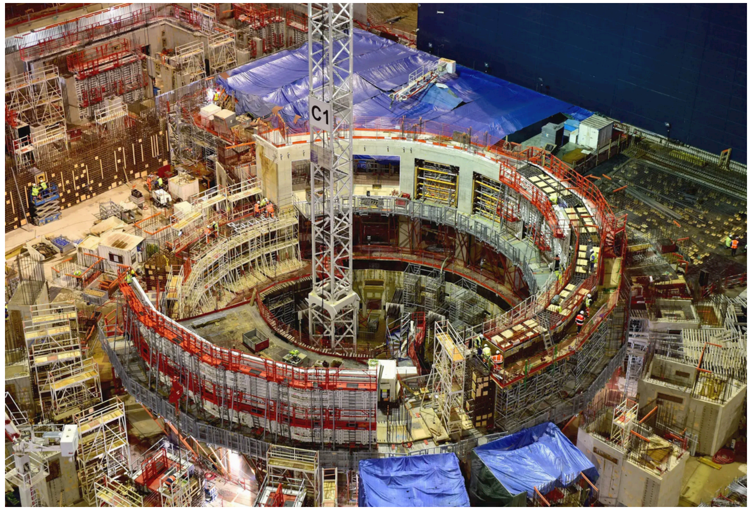
Some of humanity largest projects



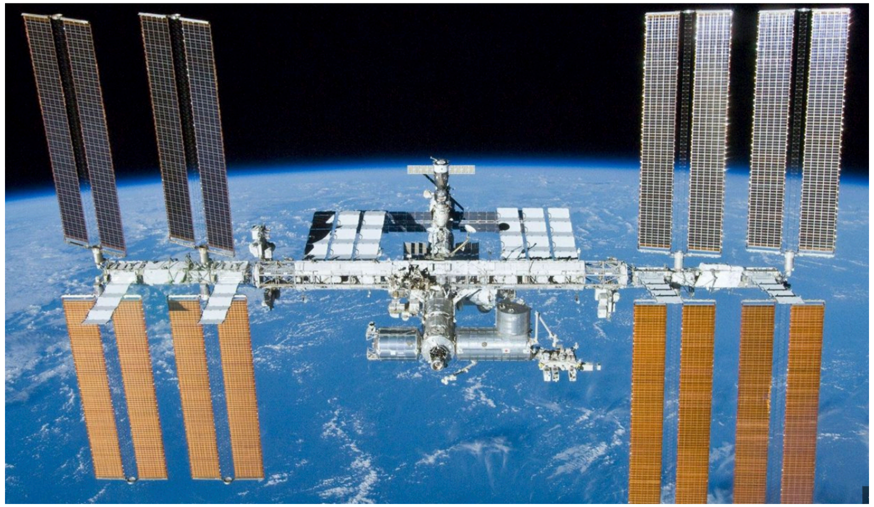
LHC: $5 Billion, 23 countries



Hubble $16 billion, 11 countries



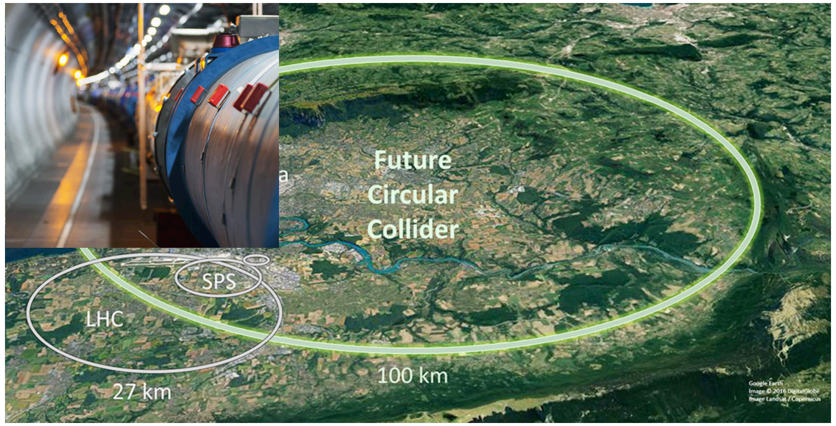
ITER: $45 Billion, 35 countries



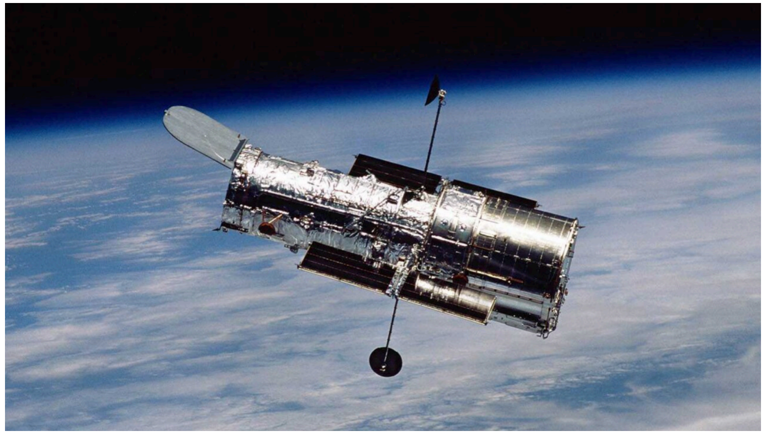
ISS: $100 Billion, 16 countries

A Case of Openness

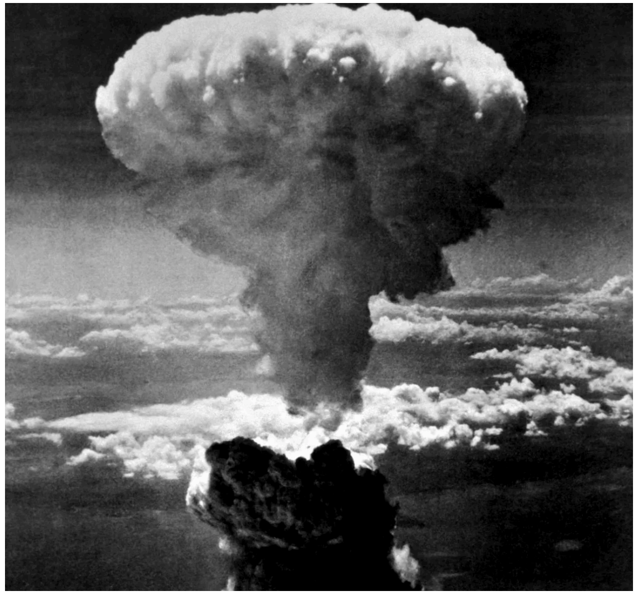
Some of humanity largest projects

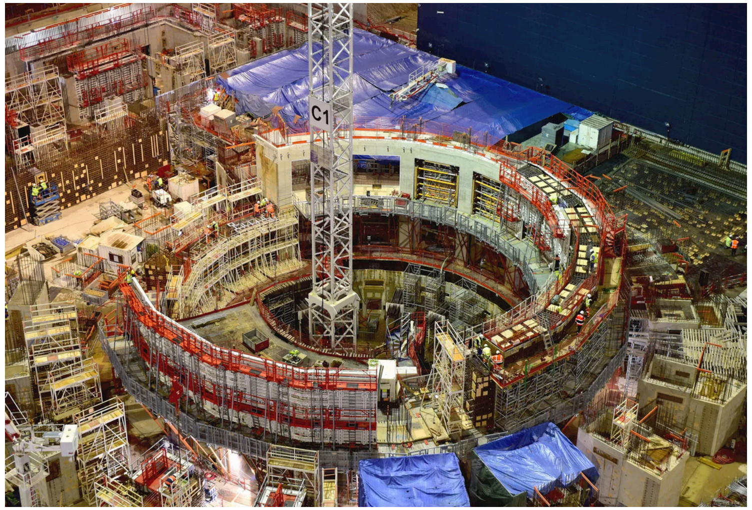


LHC: $5 Billion, 23 countries

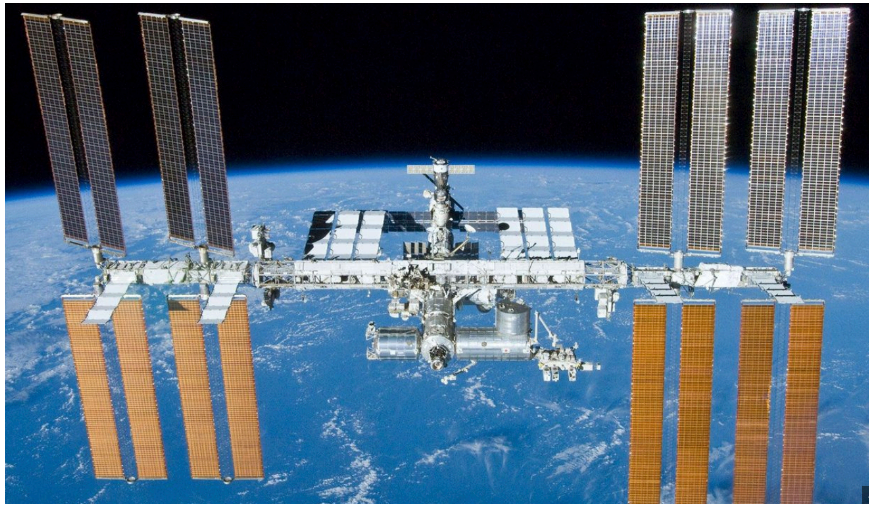


Hubble $16 billion, 11 countries





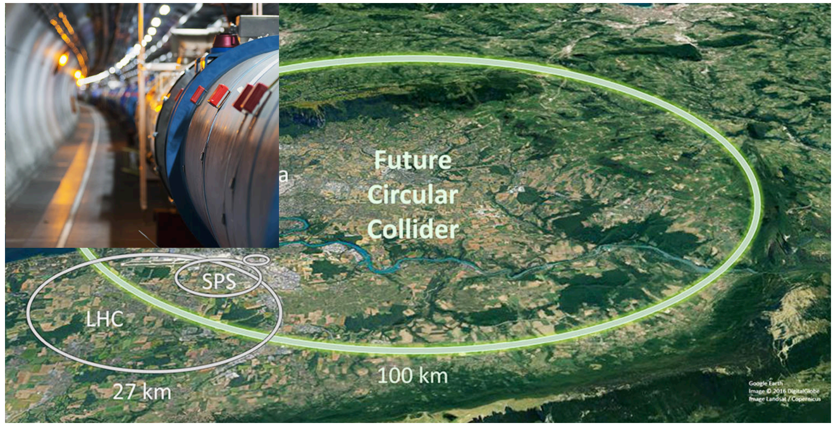
ITER: $45 Billion, 35 countries



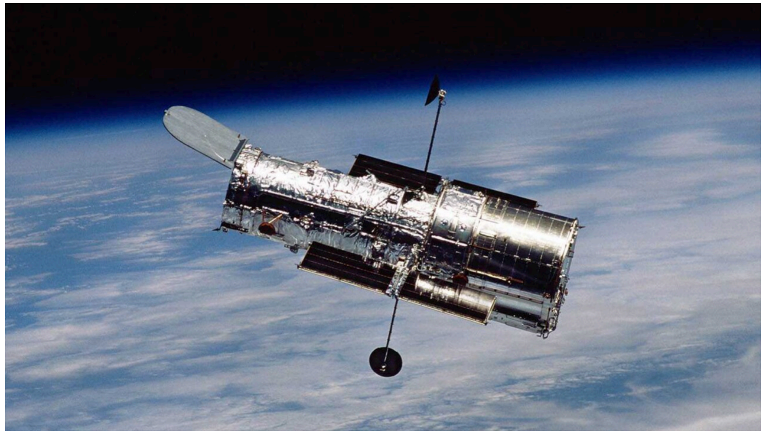
ISS: $100 Billion, 16 countries

A Case of Openness

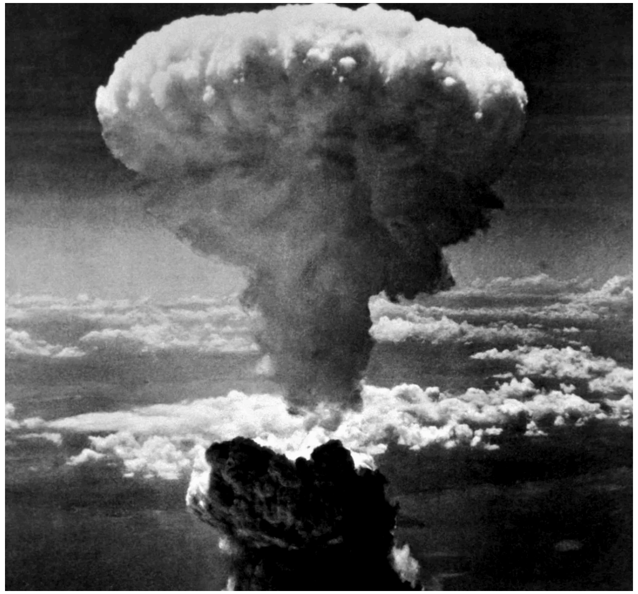
Some of humanity largest projects



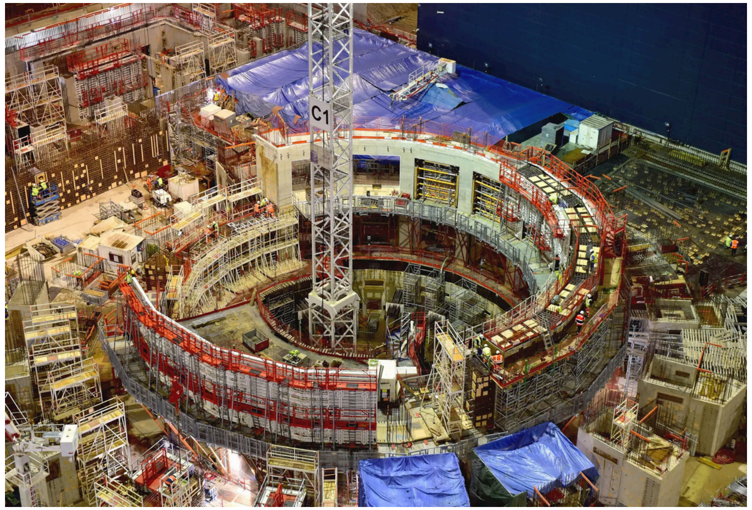
LHC: $5 Billion, 23 countries



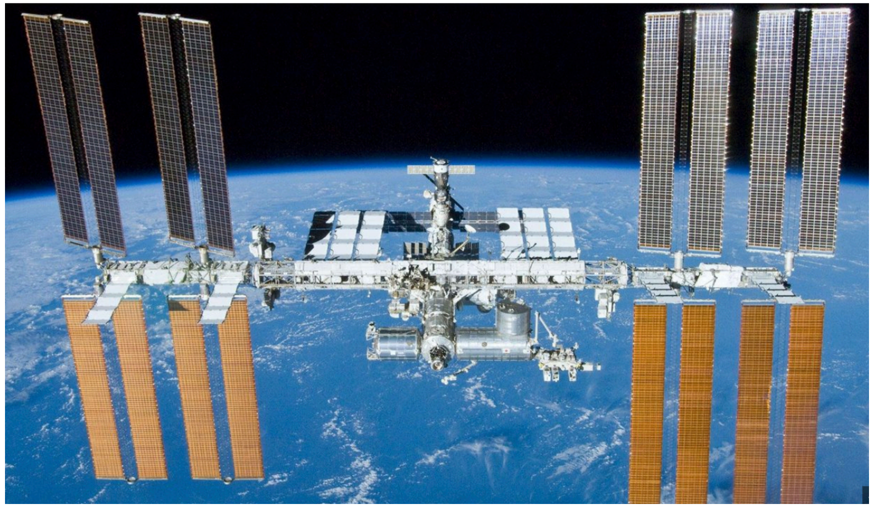
Hubble $16 billion, 11 countries



Manhattan project $30 billion, 3 countries



ITER: $45 Billion, 35 countries



ISS: $100 Billion, 16 countries

OpenEuroLLM

Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC

Universities and Research Organizations

Charles University
AI Sweden
alt-edic
Eberhard Karls Universität Tübingen
ellis Institute Tübingen
Fraunhofer IAIS
Barcelona Supercomputing Center
Jülich Forschungszentrum
TU/e Eindhoven University of Technology
University of Helsinki
University of Turku

Companies

Aleph Alpha
AMD Silo AI
ellamind
LightOn
prompsit

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
    - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
    - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models

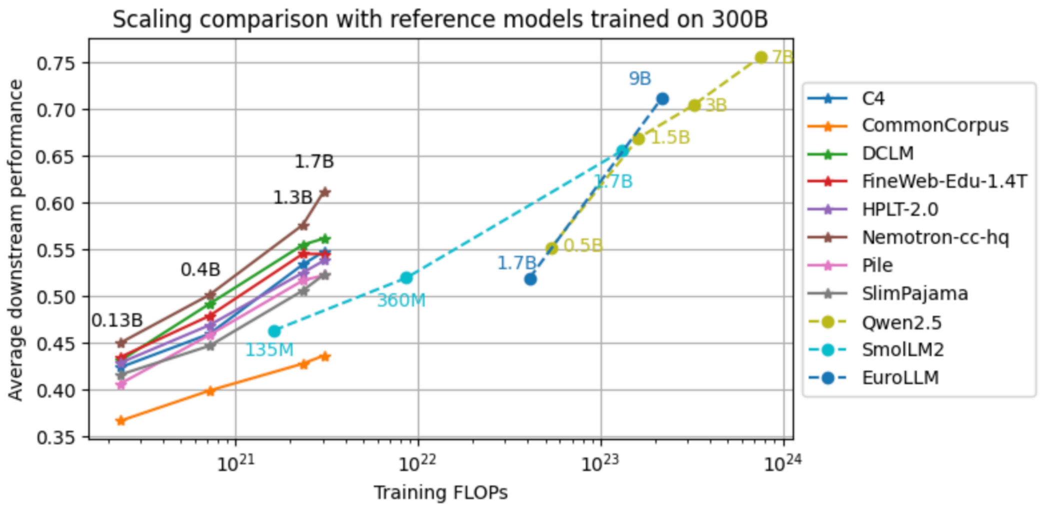
Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
  - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
  - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Charles University
AI Sweden
alt-edic
Eberhard Karls Universität Tübingen
ellis Institute Tübingen
Fraunhofer IAIS
Barcelona Supercomputing Center Centro Nacional de Supercomputación
Jülich Forschungszentrum
TU/e Eindhoven University of Technology
University of Helsinki
University of Turku

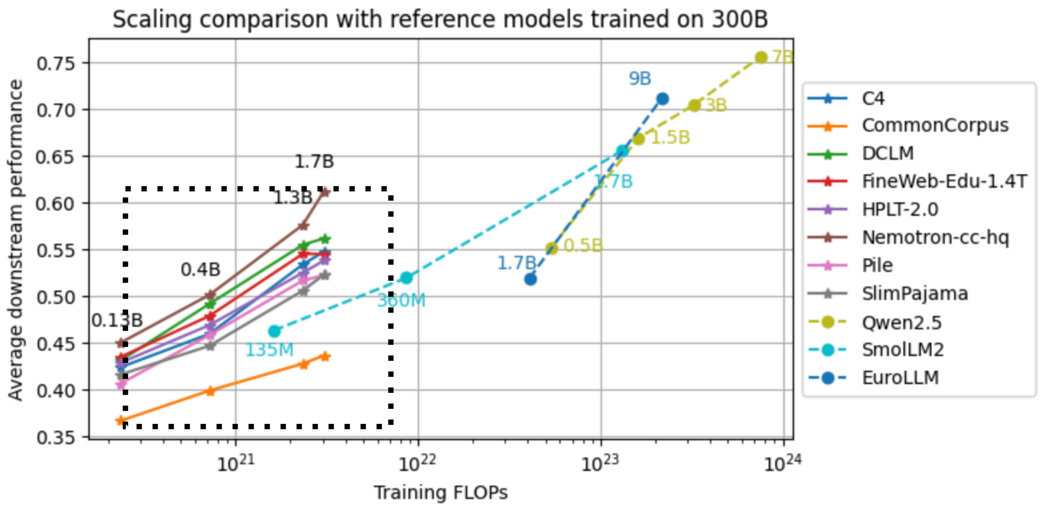
Companies

Aleph Alpha
AMD Silo AI
ellamind
LightOn
prompsit

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
    - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
    - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Charles University · AI Sweden · alt-edic · Eberhard Karls Universität Tübingen · ellis Institute Tübingen · Fraunhofer IAIS · BSC Barcelona Supercomputing Center · Jülich Forschungszentrum · TU/e Eindhoven University of Technology · University of Helsinki · University of Oslo · University of Turku

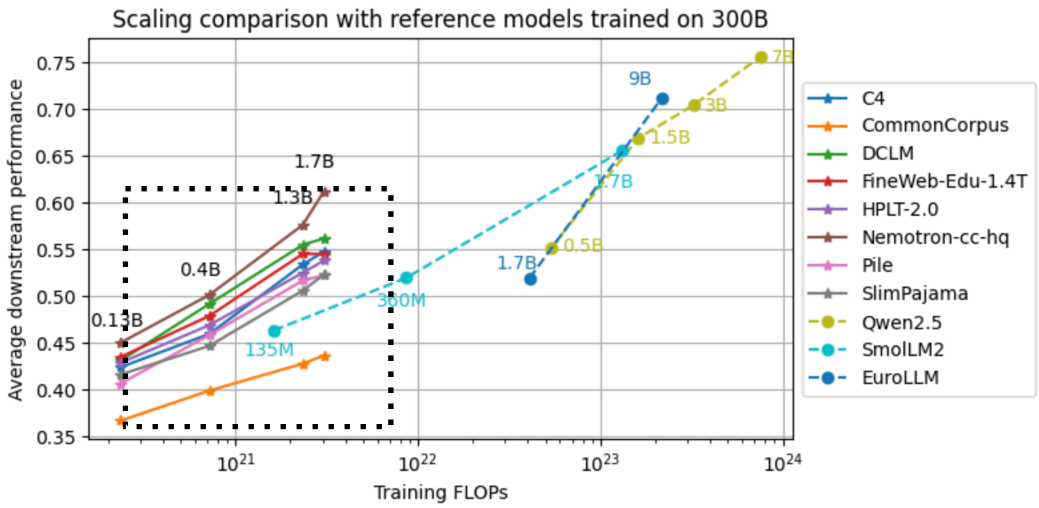
Companies

Aleph Alpha · AMD Silo AI · ellamind · LightOn · prompsit

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
    - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
    - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋



Reference analysis training 1.7B models from scratch for different datasets

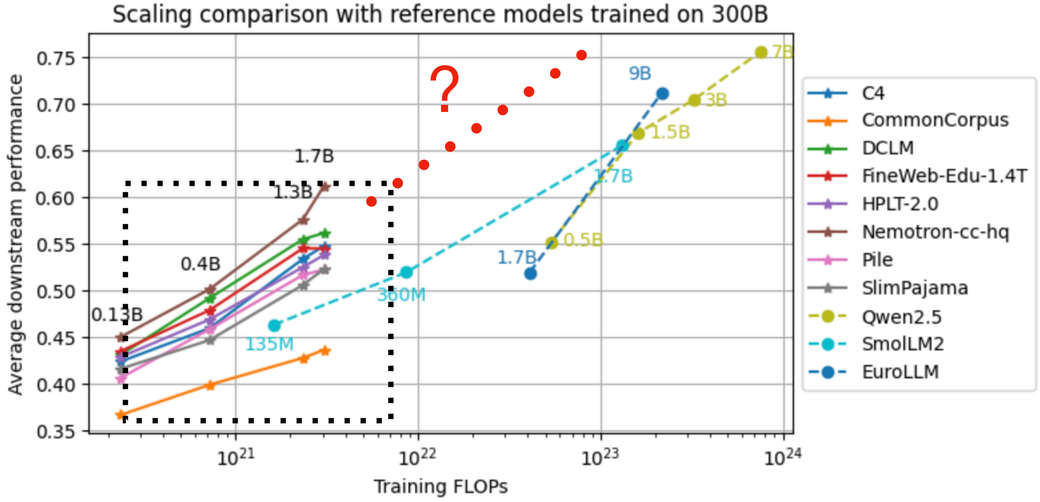
Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
    - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
    - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋



Reference analysis training 1.7B models from scratch for different datasets

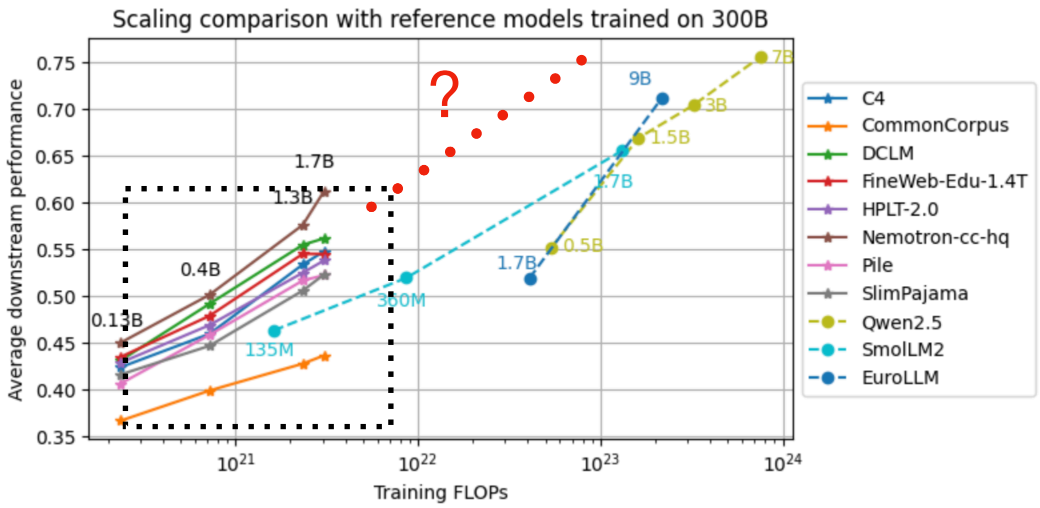
Universities and Research Organizations

Companies

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
    - Started in February 2025
    - Fully open: weights & code & data
    - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
    - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
    - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋
- Ping me if interested 🤗



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Charles University
AI Sweden
alt-edic
Eberhard Karls Universität Tübingen
ellis Institute Tübingen
Fraunhofer IAIS
Barcelona Supercomputing Center
Jülich Forschungszentrum
TU/e Eindhoven University of Technology
University of Helsinki
University of Turku

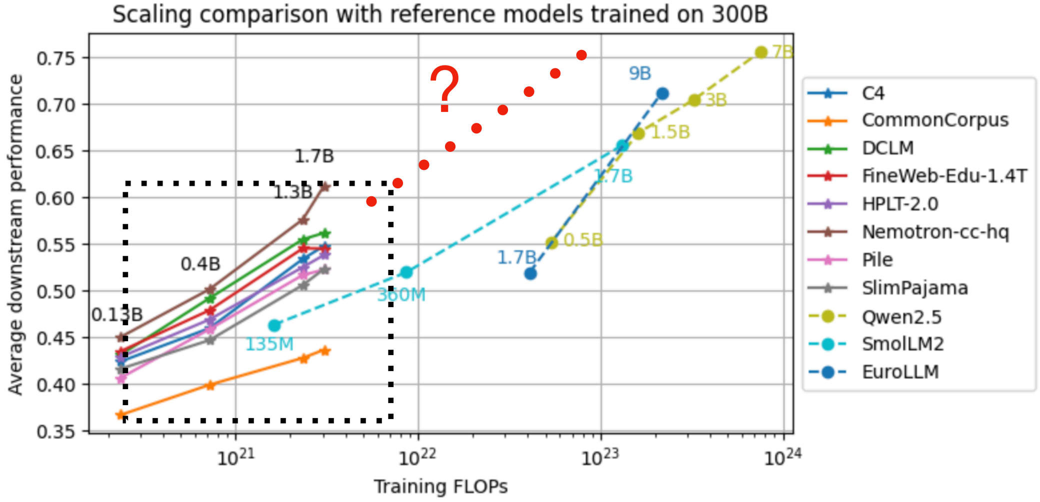
Companies

Aleph Alpha
AMD Silo AI
ellamind
LightOn
prompsit

Co-funded by the European Union

OpenEuroLLM

- An effort to build multilingual LLMs from scratch by 2028
  - Started in February 2025
  - Fully open: weights & code & data
  - €37.4 million funding. In addition many millions of GPU hours allocated in EuroHPC
- Just released:
  - Reference 2B models with SOTA performance among fully open models https://huggingface.co/collections/open-sci/open-sci-ref-001-685905e598be658fbcebff4f
  - 38 Monolingual 2B LLMs https://openeurollm.eu/blog/hplt-oellm-38-reference-models
- Currently hiring 9 ML researchers / engineers at ELLIS! Also internships 👋
- Ping me if interested 🤗
- Lots of areas for AutoML in pre-training, post-training, evaluation 🎉



Reference analysis training 1.7B models from scratch for different datasets

Universities and Research Organizations

Charles University · AI Sweden · alt-edic · Eberhard Karls Universität Tübingen · ellis Institute Tübingen · Fraunhofer IAIS · Barcelona Supercomputing Center · Jülich Forschungszentrum · TU/e Eindhoven University of Technology · University of Helsinki · University of Oslo · University of Turku

Companies

Aleph Alpha · AMD Silo AI · ellamind · LightOn · prompsit

Co-funded by the European Union

Any questions or discussion point?