Finetuning Foundation Models for joint Analysis Optimization

arXiv:2401.13536

Lukas Heinrich, Nicole Hartman, Matthias Vigl

ACAT 2024, Stony Brook University, Stony Brook, Long Island NY, USA - 12 Mar 2024

Technische Universität München

ORIGINS Excellence Cluster

ATLAS EXPERIMENT

MAX-PLANCK-INSTITUT FÜR PHYSIK

Analysis pipeline at the LHC

$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \not{D} \psi + h.c. + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$

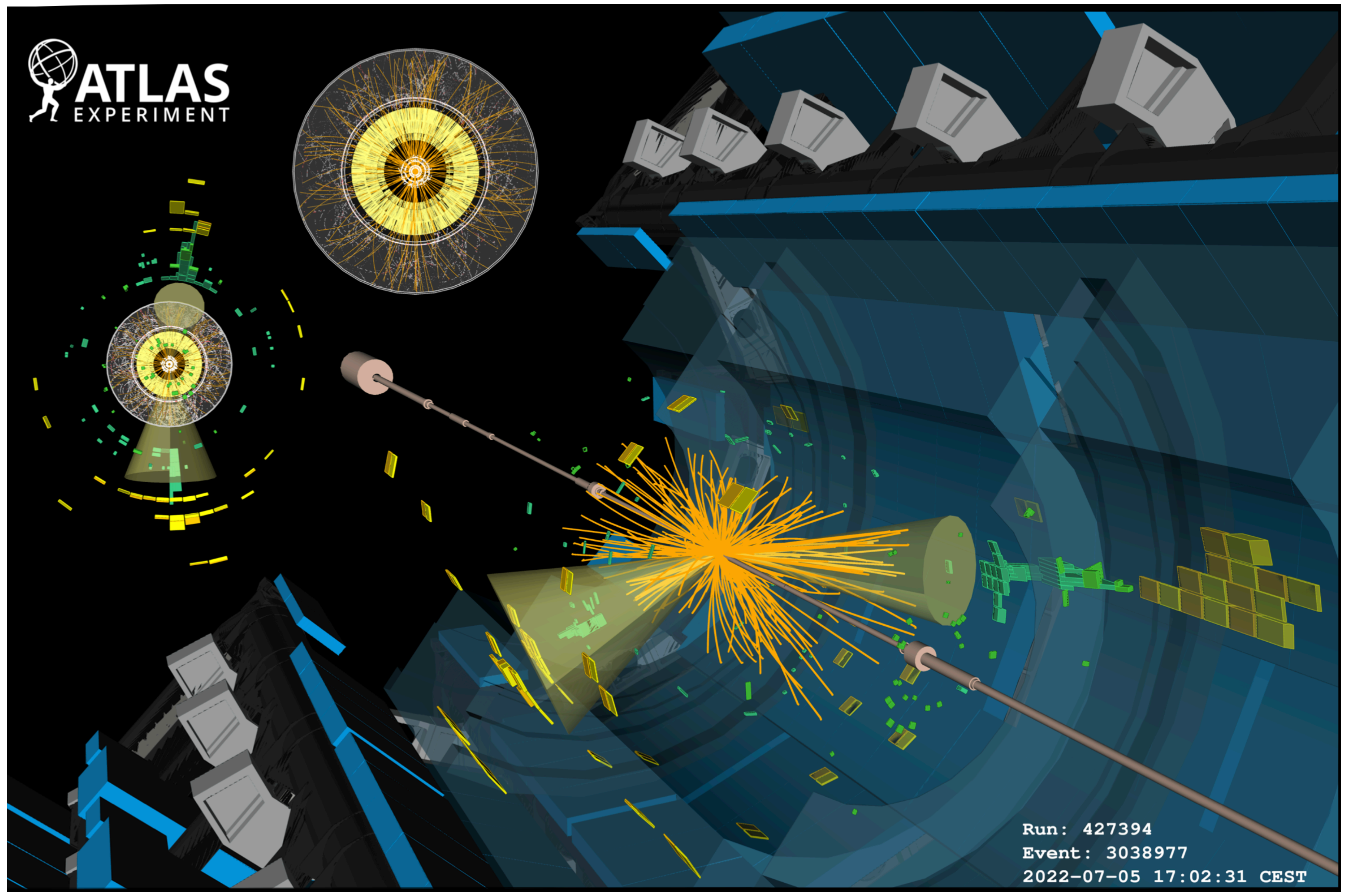
Theory

Few parameters

- ***reconstruction*** both reduces dimensionality of the data and gives physics interpretable representation (particles)

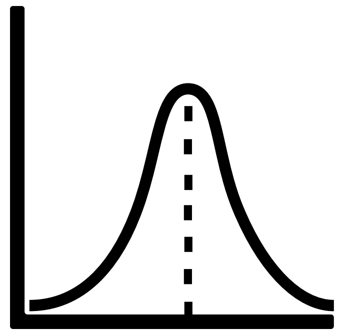
Raw Data

O(100M) channels!



Particles — Analysis — Result

summary statistics

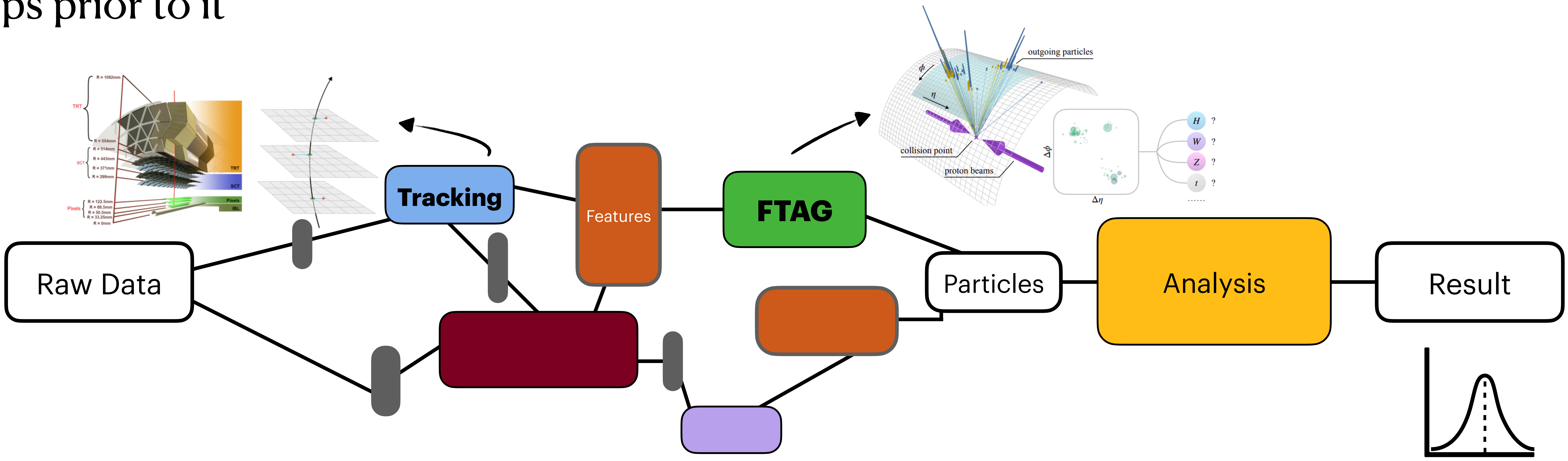


Lukas Heinrich, Nicole Hartman, Matthias Vigl

Reconstruction

Lots of (also ML) components - e.g. ***tracking***, ***jet tagging*** *(ftag)*...

But each **optimised separately** and downstream components are optimised based on the steps prior to it



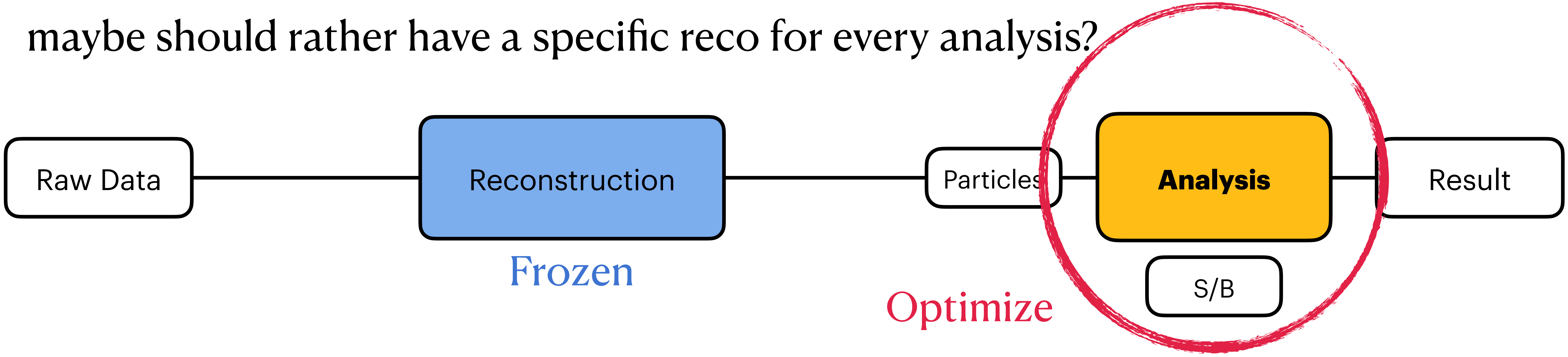
Lukas Heinrich, Nicole Hartman, Matthias Vigl

Analysis optimisation

The optimisation of the sensitivity is primarily the job of the **analysis**, given a fixed **reconstruction** - mostly common for all analysis

Is this the best way to do it?

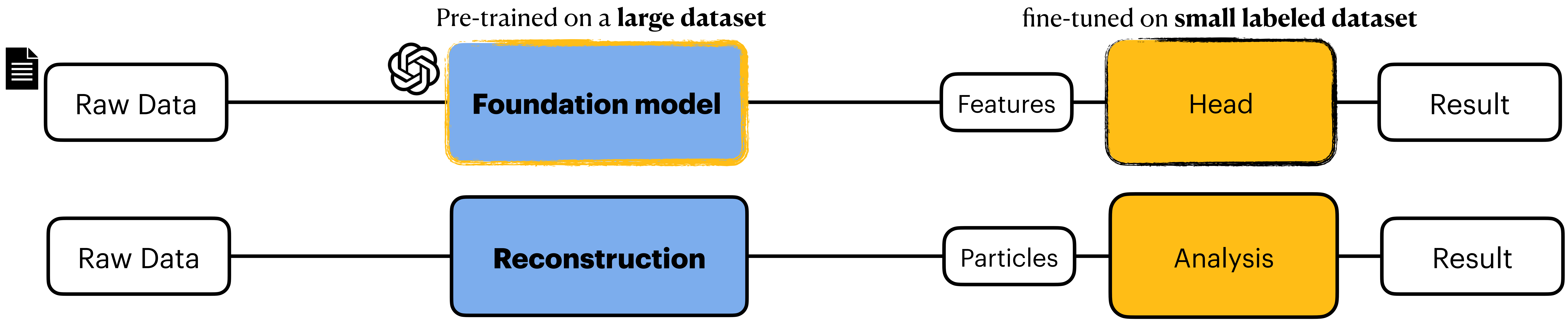
- a fixed reco loses some information irrevocably
- maybe should rather have a specific reco for every analysis?



Lukas Heinrich, Nicole Hartman, Matthias Vigl

Reconstruction = Foundation model

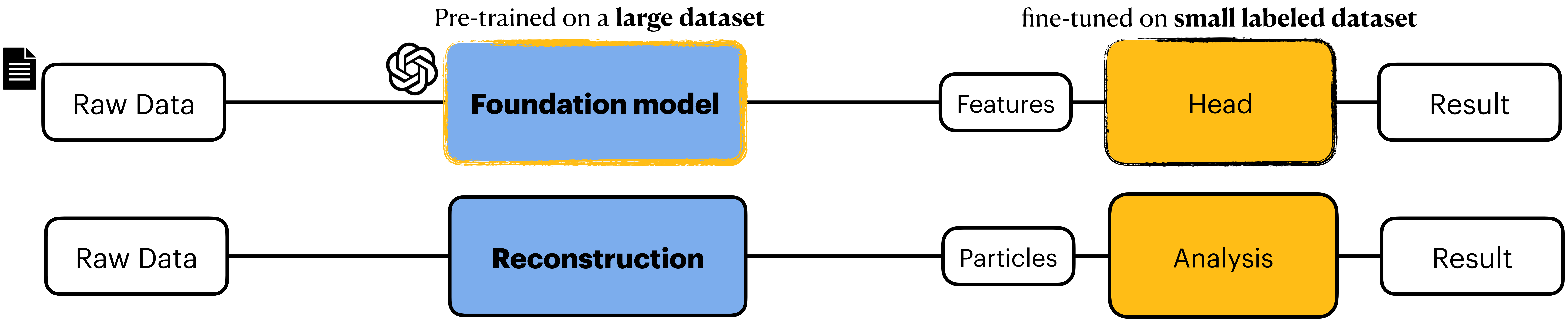
- ML and HEP setups are fortunately very aligned
- But everything is differentiable so can be fine-tuned w/ gradient descent



Lukas Heinrich, Nicole Hartman, Matthias Vigl

Reconstruction = Foundation model

- ML and HEP setups are fortunately very aligned
- But everything is differentiable so can be fine-tuned w/ gradient descent



Key difference: reconstruction is mostly common and Frozen for each downstream task (analysis)

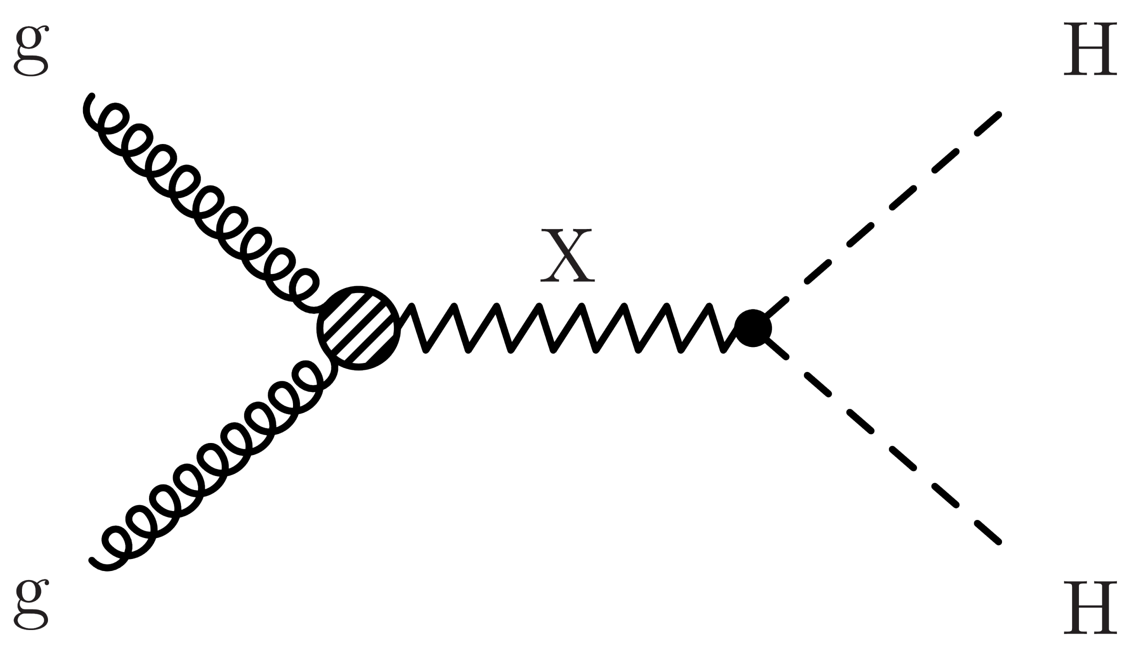
Q: Could this Finetuning workflow also work in HEP?

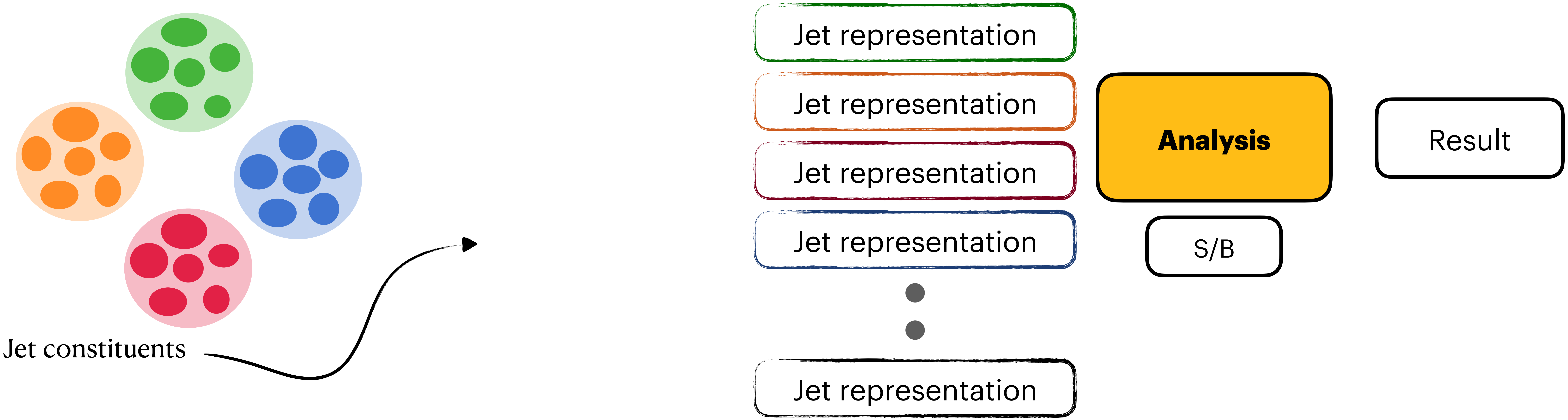
Lukas Heinrich, Nicole Hartman, Matthias Vigl

A toy end-to-end Analysis

$X \rightarrow HH \rightarrow b\bar{b}b\bar{b}$[1]

Final state with Higgs/ QCD Jets



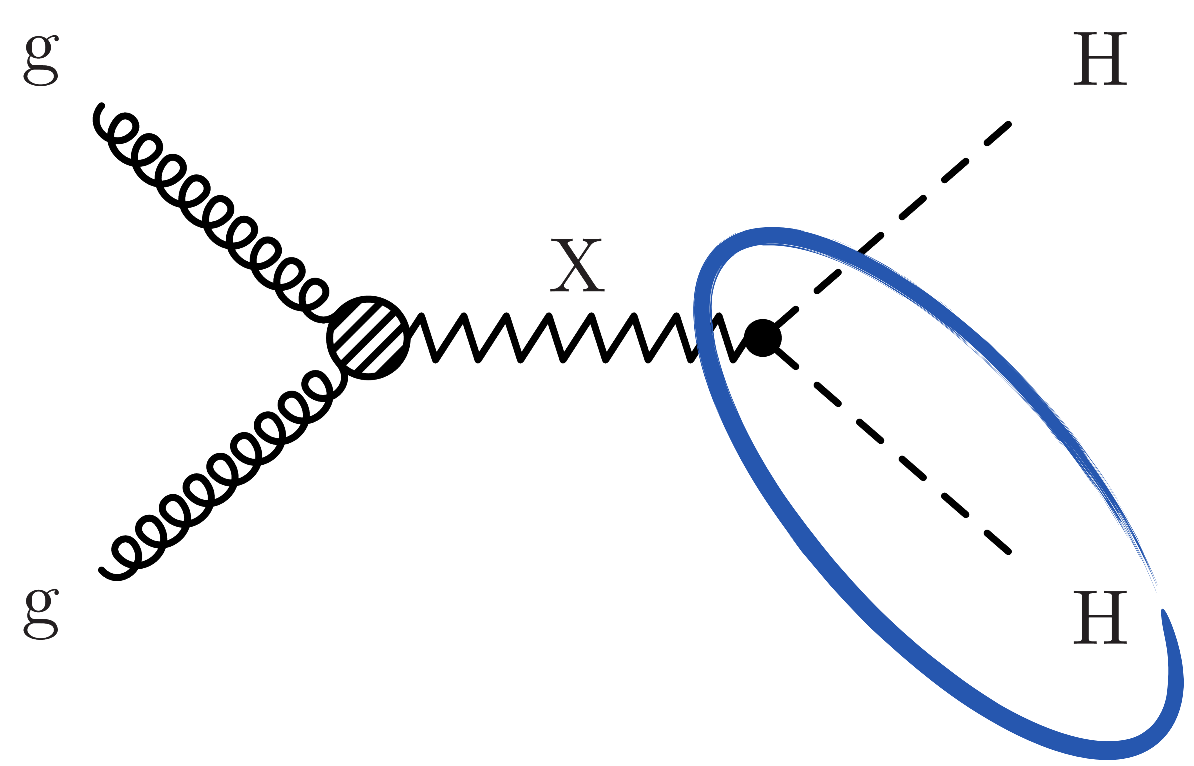


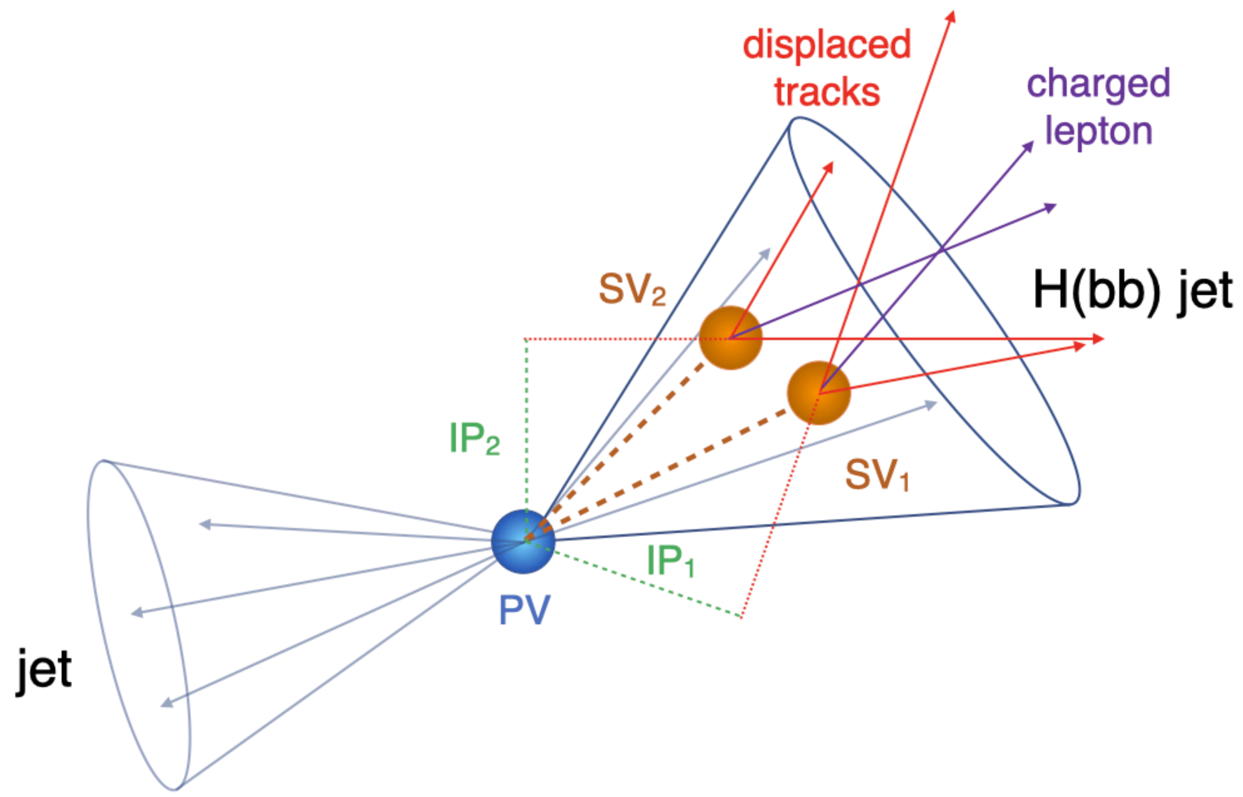
[1]: Duarte Javier, CMS open data [http://opendata.cern.ch/record/12102]

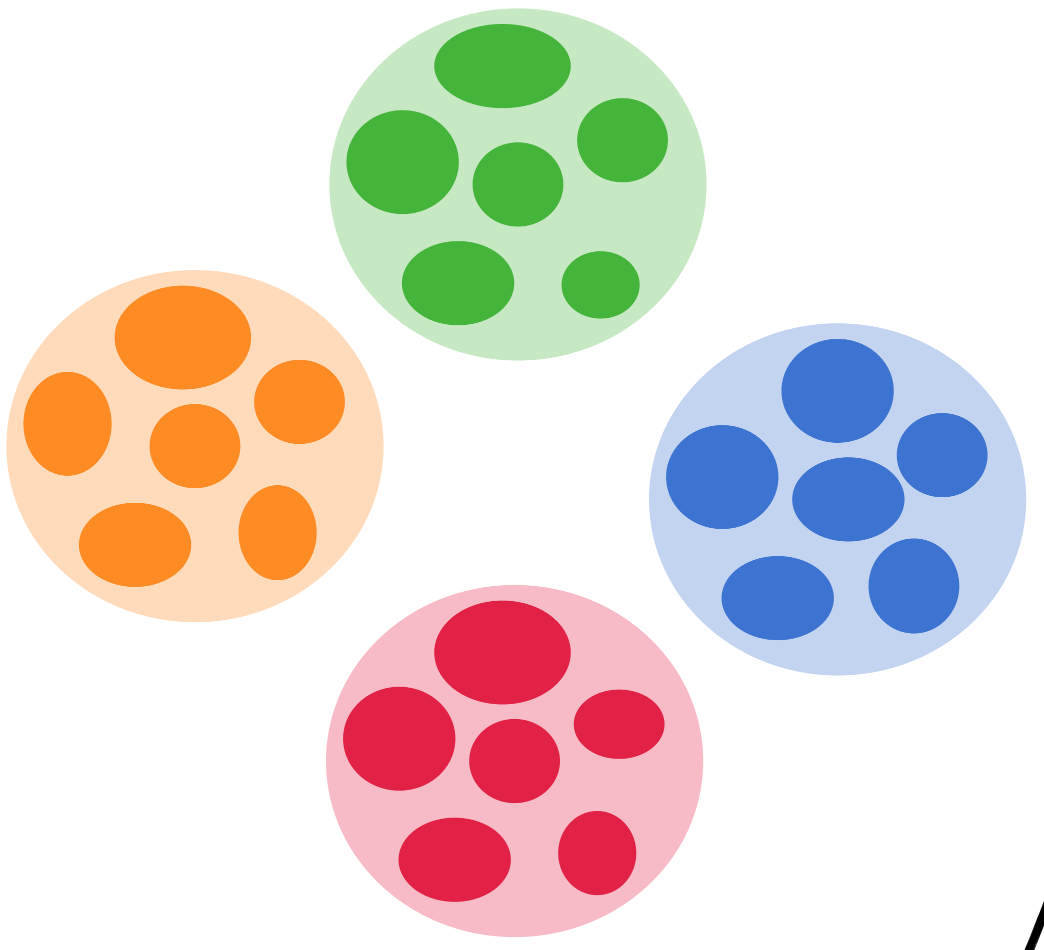
A toy end-to-end Analysis

$X \rightarrow HH \rightarrow b\bar{b}b\bar{b}$[1]

Final state with Higgs/ QCD Jets







Jet constituents

Backbone FM

Transformer (ParT)[2]

Xbb

Xbb flavour tagging (supervised)

Jet representation

Jet representation

Jet representation

Jet representation

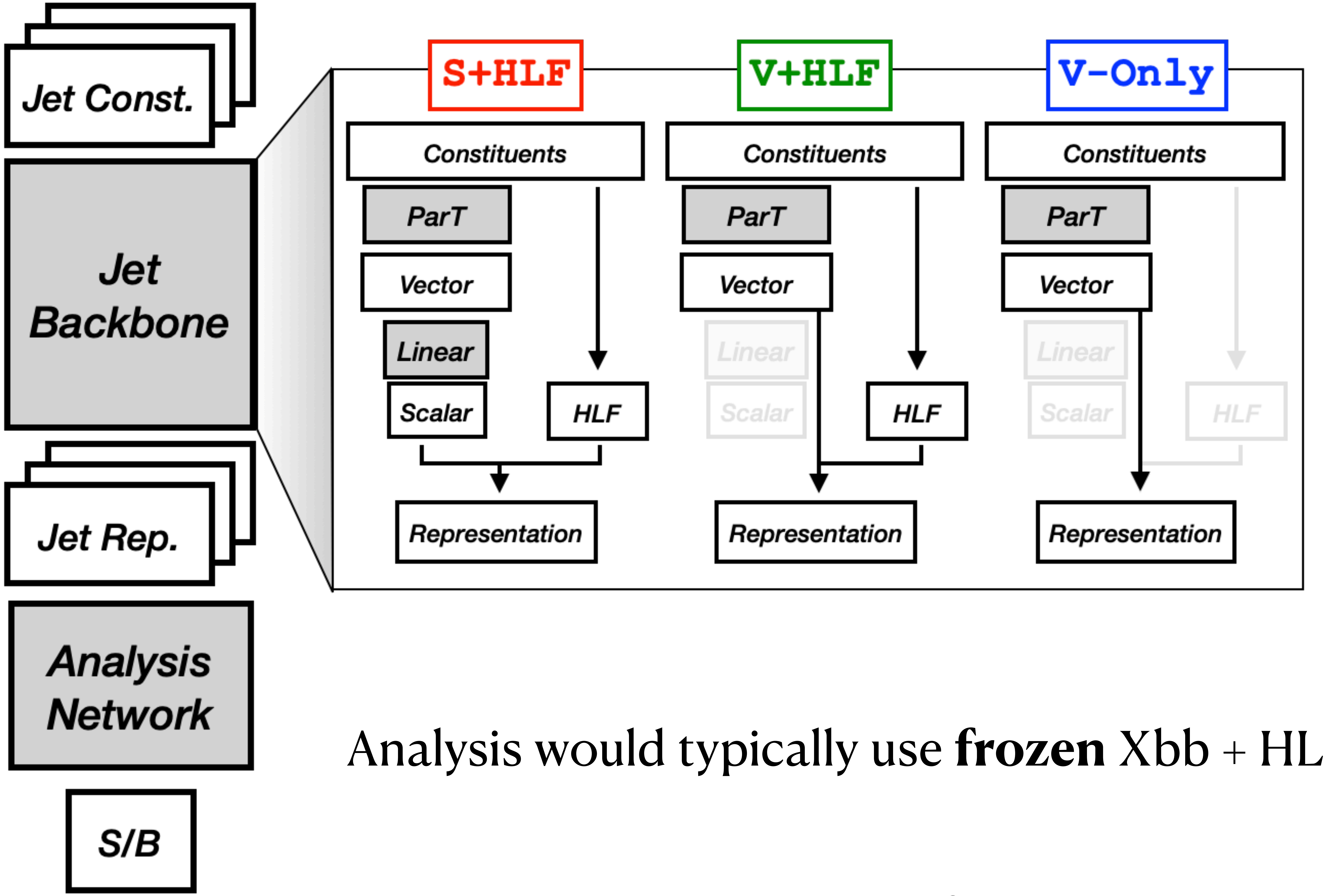
Jet representation

[2]: Huilin Qu, Congqiao Li, and Sitian Qian, "Particle Transformer for Jet Tagging," (2022), arXiv:2202.03772

[1]: Duarte Javier, CMS open data [http://opendata.cern.ch/record/12102]

Lukas Heinrich, Nicole Hartman, Matthias Vigl

Backbone Jet representation



Q: Do high-dim embeddings hold more (useful) info than Xbb+HL Features?

Analysis would typically use **frozen** Xbb + HL Features (jet 4-momenta)

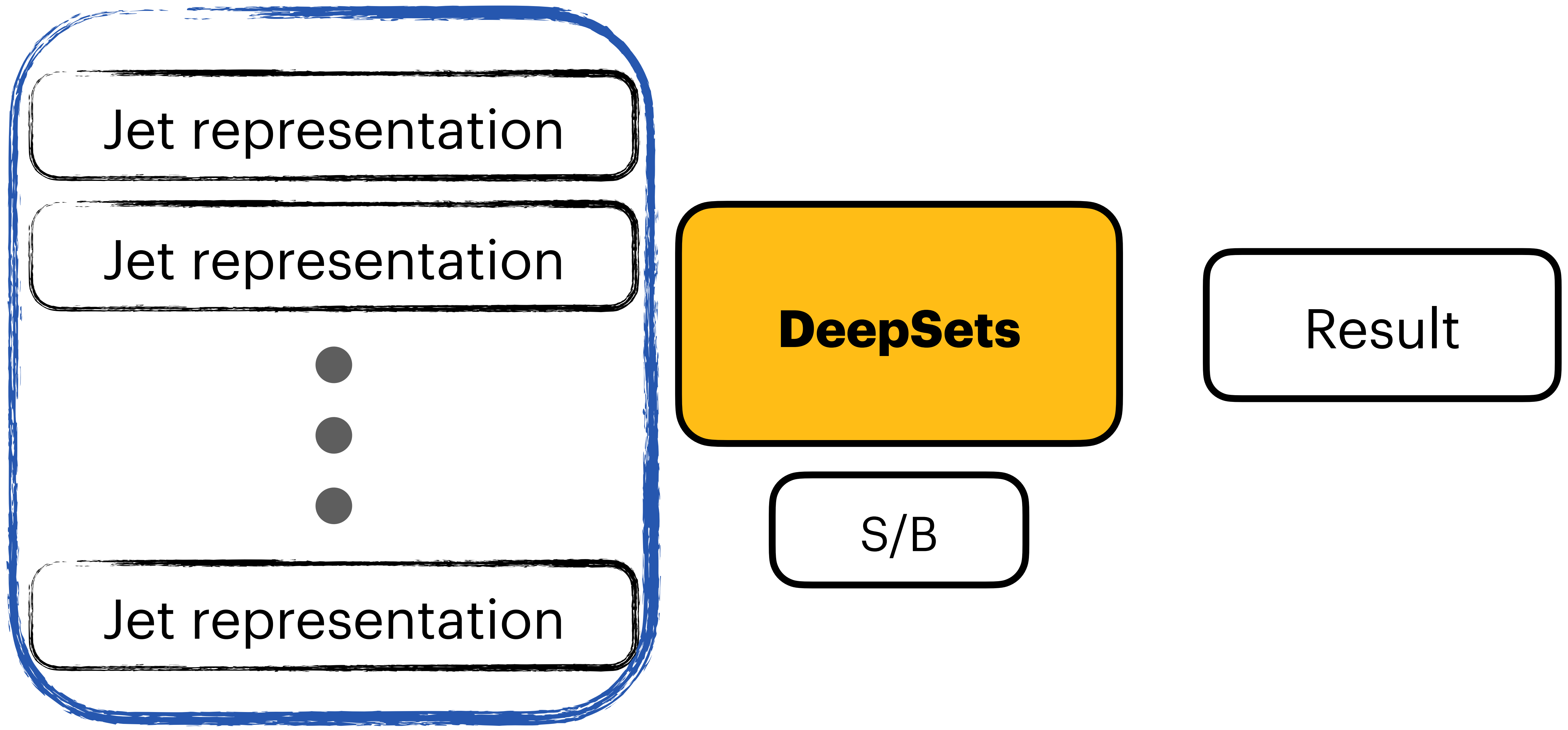
Lukas Heinrich, Nicole Hartman, Matthias Vigl

Analysis head

The head is trained for S/B discrimination with Jet representations from backbone as inputs

Variable number of jets per event + Permutation Invariance -> DeepSets

Q: Does fine-tuning the jet representation help?

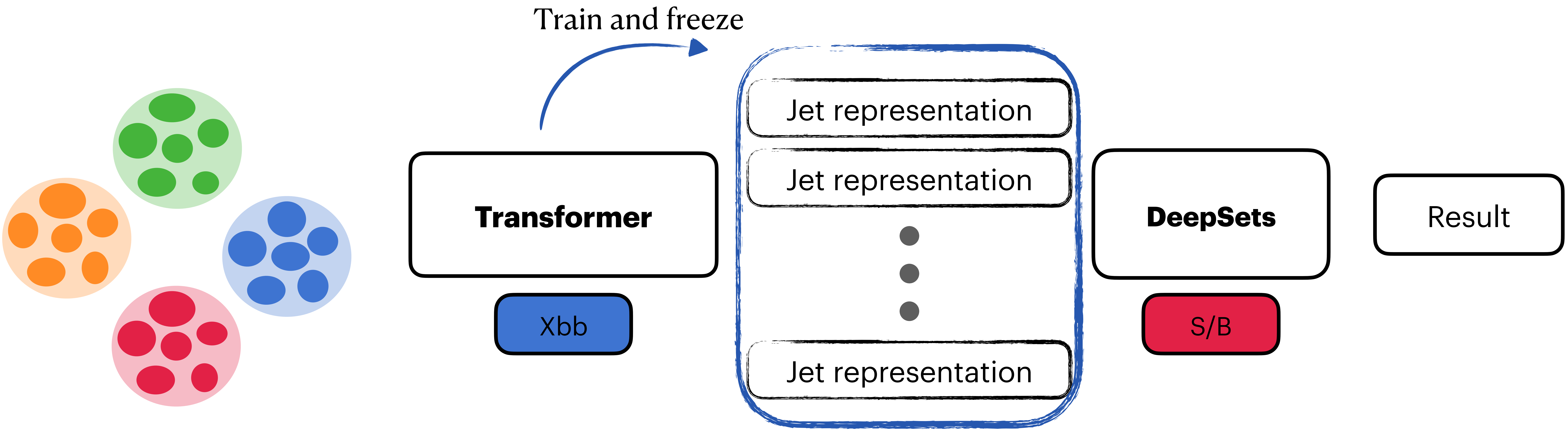


Lukas Heinrich, Nicole Hartman, Matthias Vigl

Frozen workflow

Backbone trained on **Xbb** task and then frozen

DeepSets + binary classification trained on **S/B**

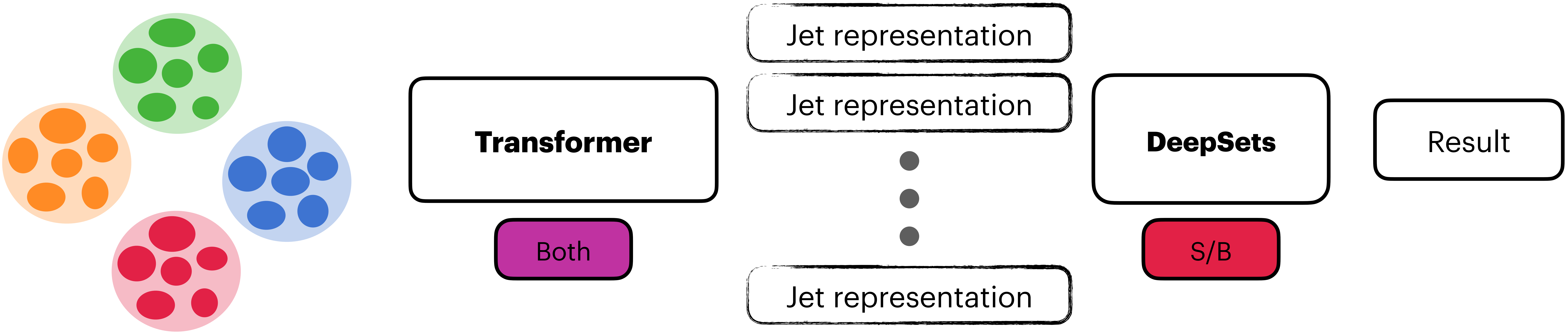


Lukas Heinrich, Nicole Hartman, Matthias Vigl

Fine-tuned workflow

Backbone pre-trained on **Xbb** task

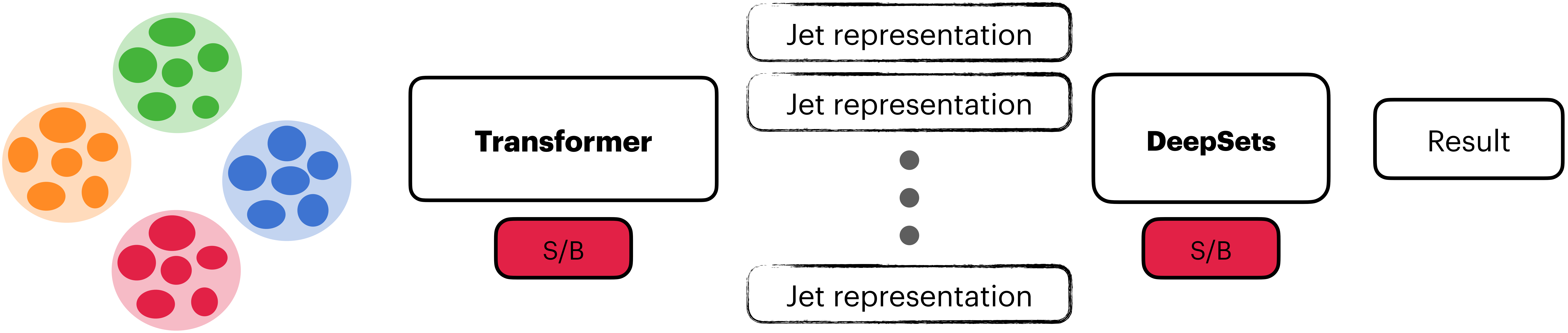
Then **fine-tuned** (end-to-end) on **S/B**



From scratch training

No backbone pre-training

Backbone + head trained from scratch on **S/B**



Lukas Heinrich, Nicole Hartman, Matthias Vigl

Architecture autonomy

Structural Autonomy →

Representational Autonomy ↓

Training \ Architecture	Scalar + HL	Vector + HL	Vector
Frozen	Standard HEP		Hope for sufficient stat
Fine-tuned		ML-assisted HEP	
From scratch	Inductive Bias is all you need		‘Hits to Higgs’

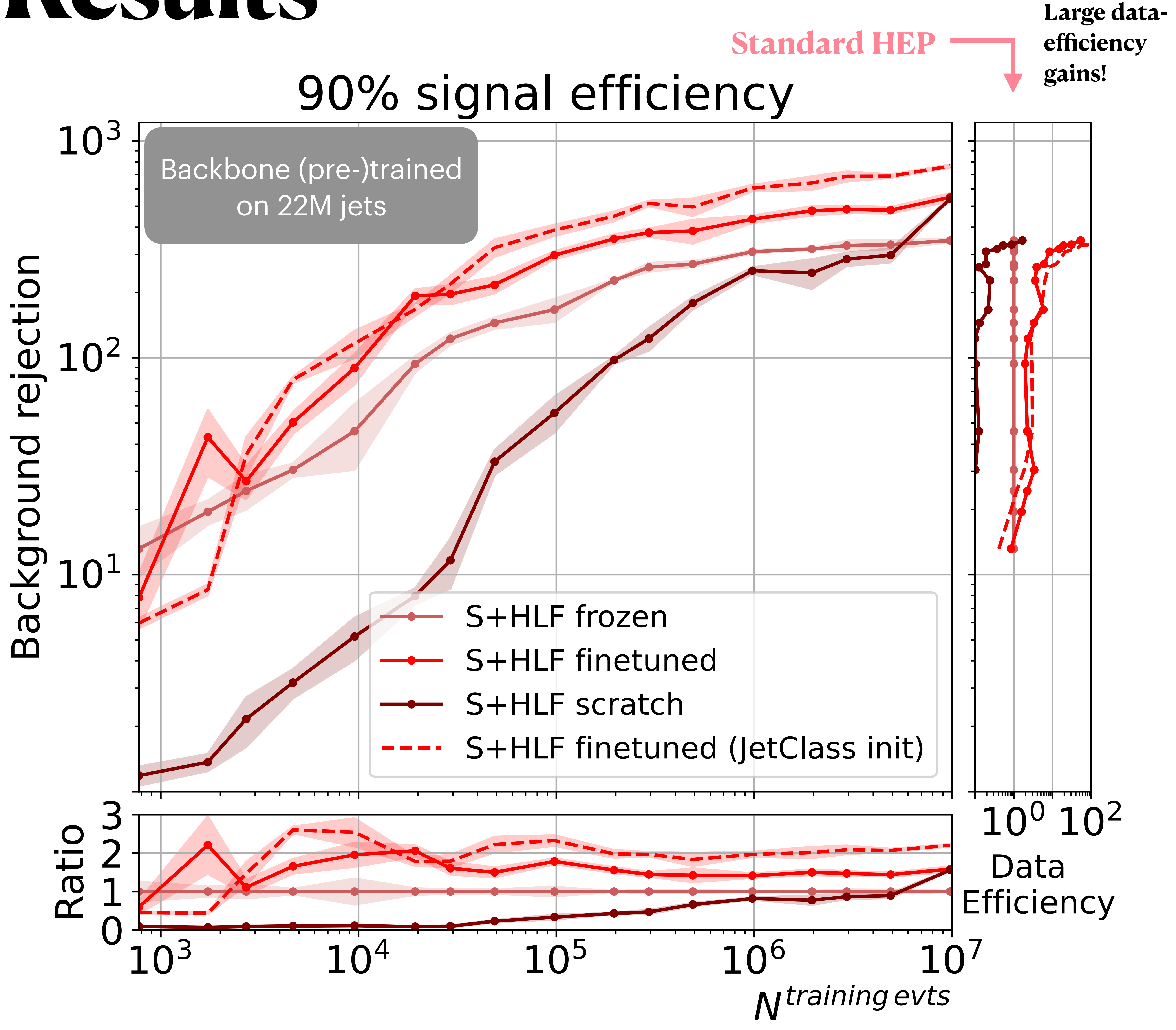
Lukas Heinrich, Nicole Hartman, Matthias Vigl

Results

Well-known patterns from ML seem to hold also in HEP

- Fine-tuning workflow improves both **performance** & **data efficiency** (10-100x wrt standard hep)

- **Domain adaptation**: Pre-training on a different dataset (JetClass[3]) helps



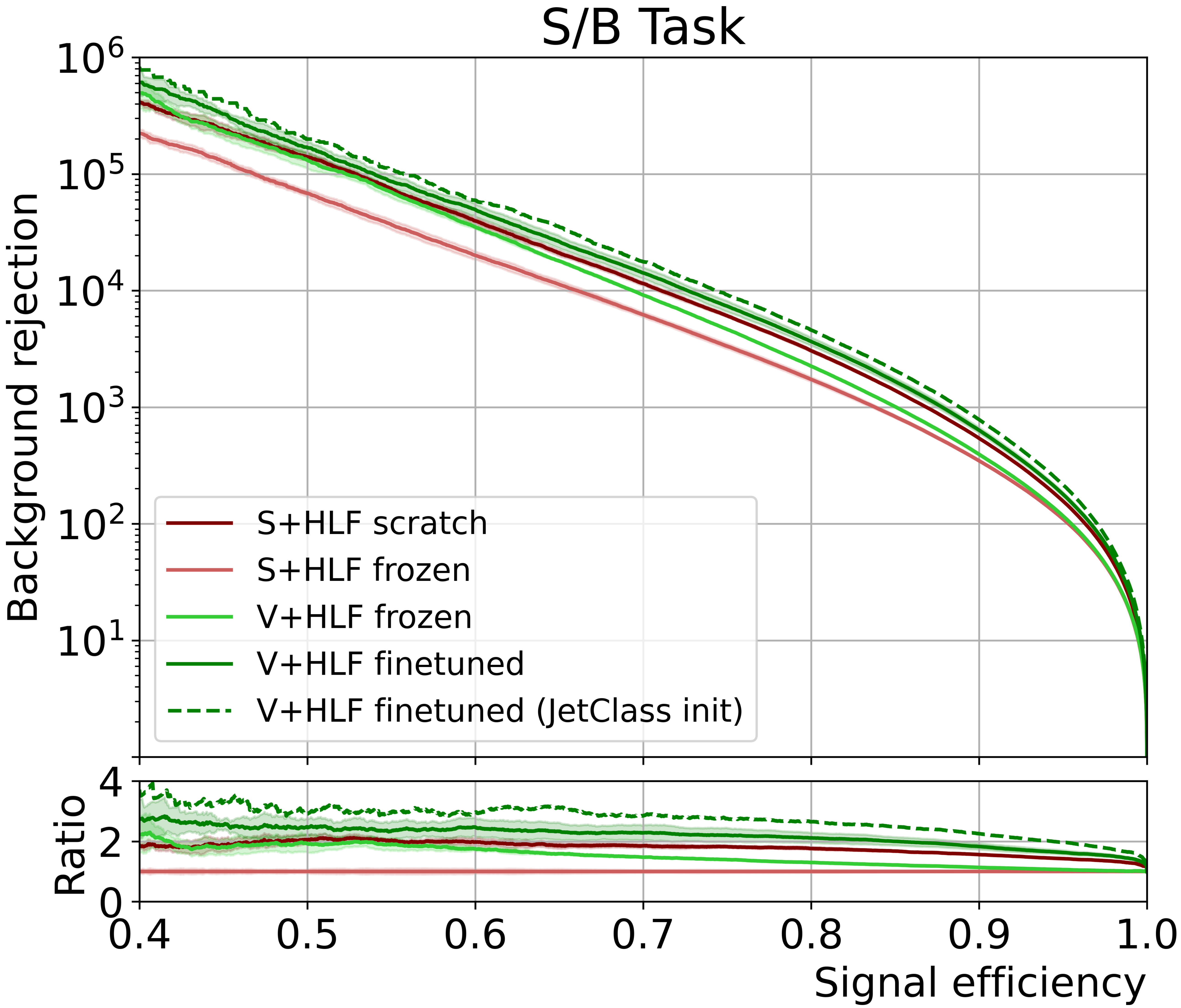
[3]: Huilin Qu, Congqiao Li, and Sitian Qian, "JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics," (2022).

Lukas Heinrich, Nicole Hartman, Matthias Vigl

Results

Well-known patterns from ML seem to hold also in HEP

- Fine-tuning workflow improves both **performance** & **data efficiency** (10-100x wrt standard hep)
- **High-dim embeddings** also seem to be useful in the frozen case
- **Domain adaptation**: Pre-training on a different dataset (JetClass[3]) helps



[3]: Huilin Qu, Congqiao Li, and Sitian Qian, "JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics," (2022).

Conclusions

1) Fine-tuning workflow for end to end analysis works and is useful even for simple examples

Compared to standard HEP approach:

- 2x in **background rejection,** 10-100x **in data efficiency**
- There might be more to gain in more complex topologies

2) Key question now: what's the best pre-training (e.g. supervised or self-supervised)?

SSL approaches are also being explored:

- e.g. "**Masked Particle Modeling**", **yesterday**, **today**, and **tomorrow** talks
- self-supervised training doesn't need labels: can pre-train on real data!
  - Huge amount of pre-training possible

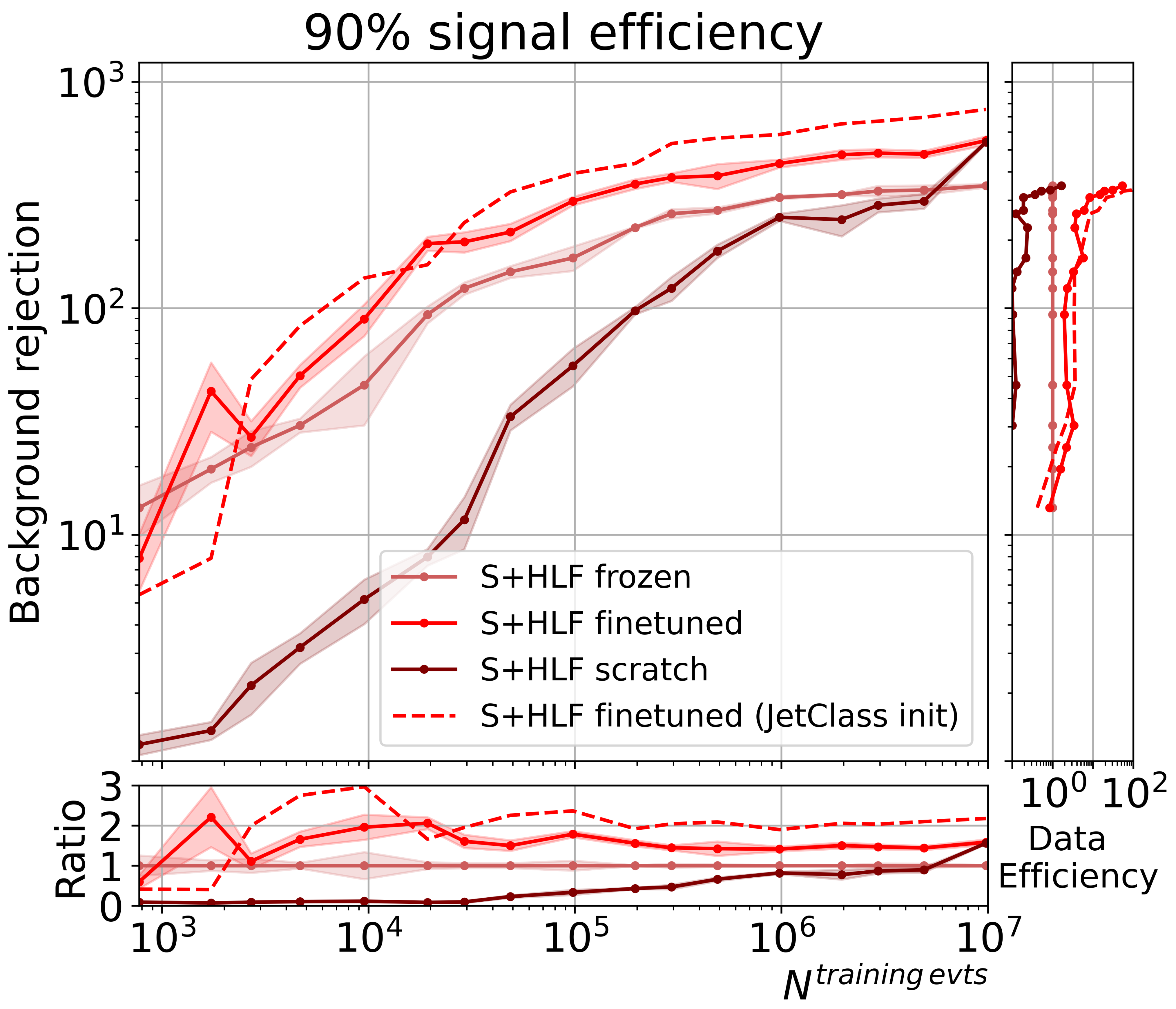
3) ...and calibration

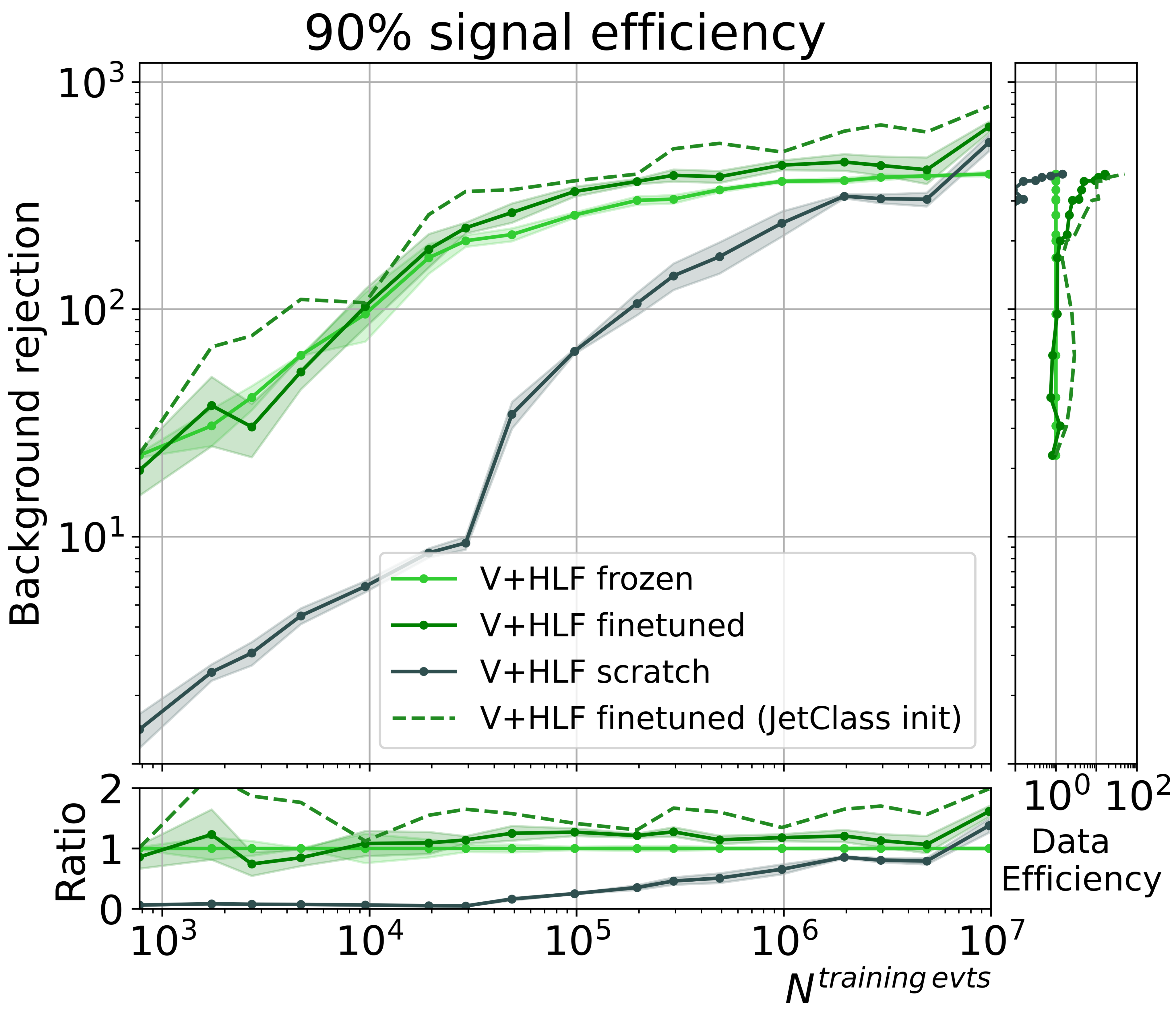
Thank You!

Link to the paper: **arXiv:2401.13536**

Backup

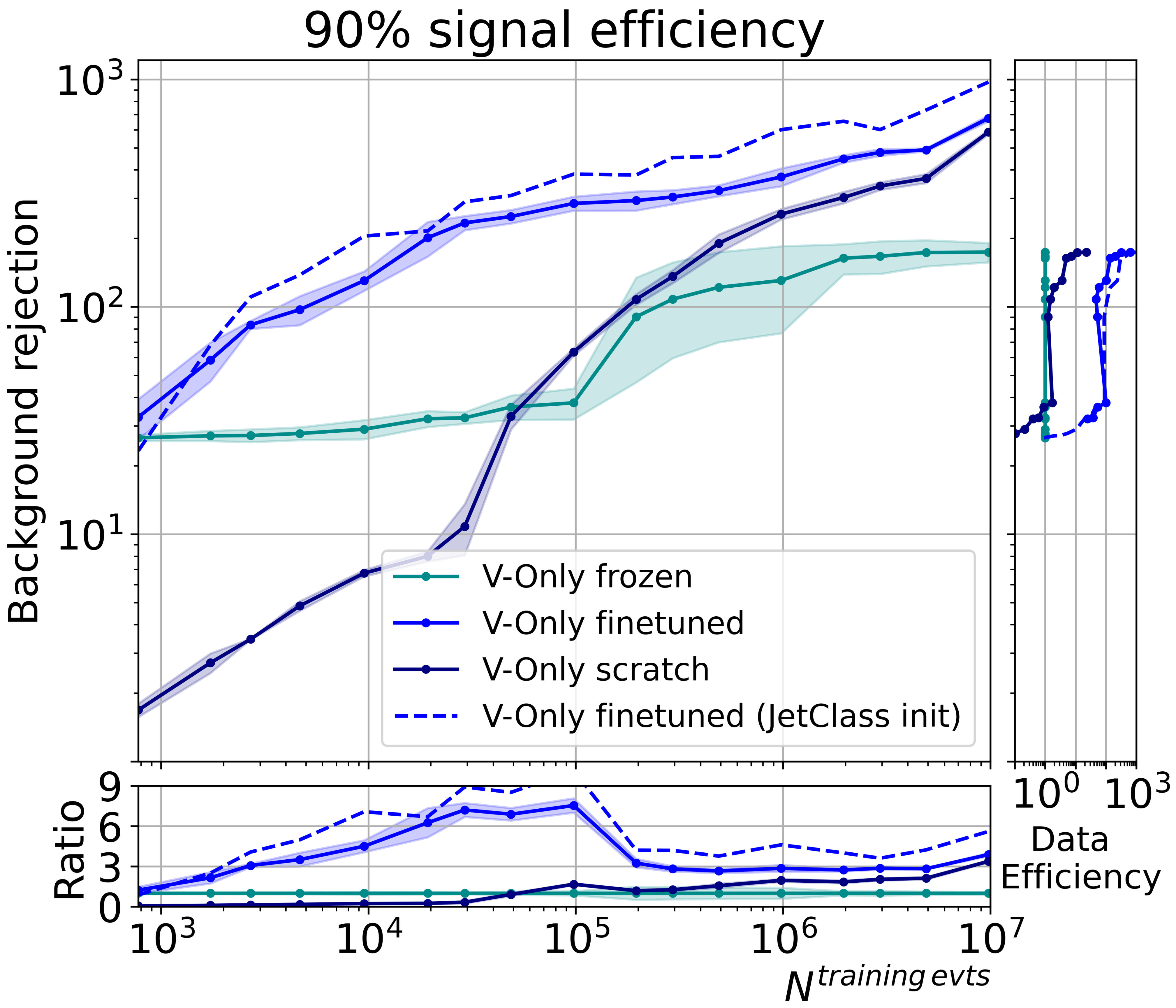
From scratch training eventually surpasses frozen models, it's just slow



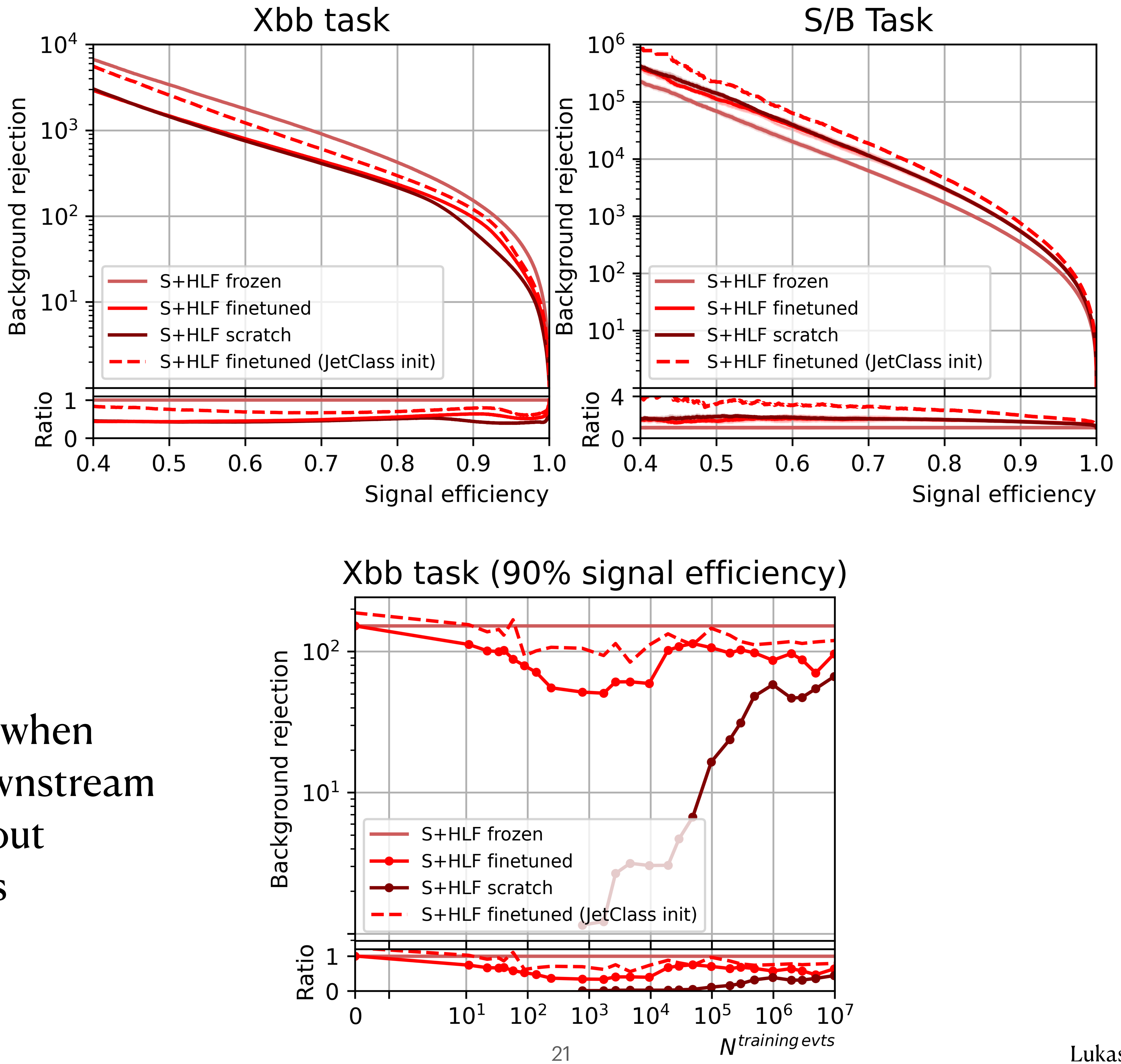


Lukas Heinrich, Nicole Hartman, Matthias Vigl

From scratch training eventually surpasses frozen models, it's just slow



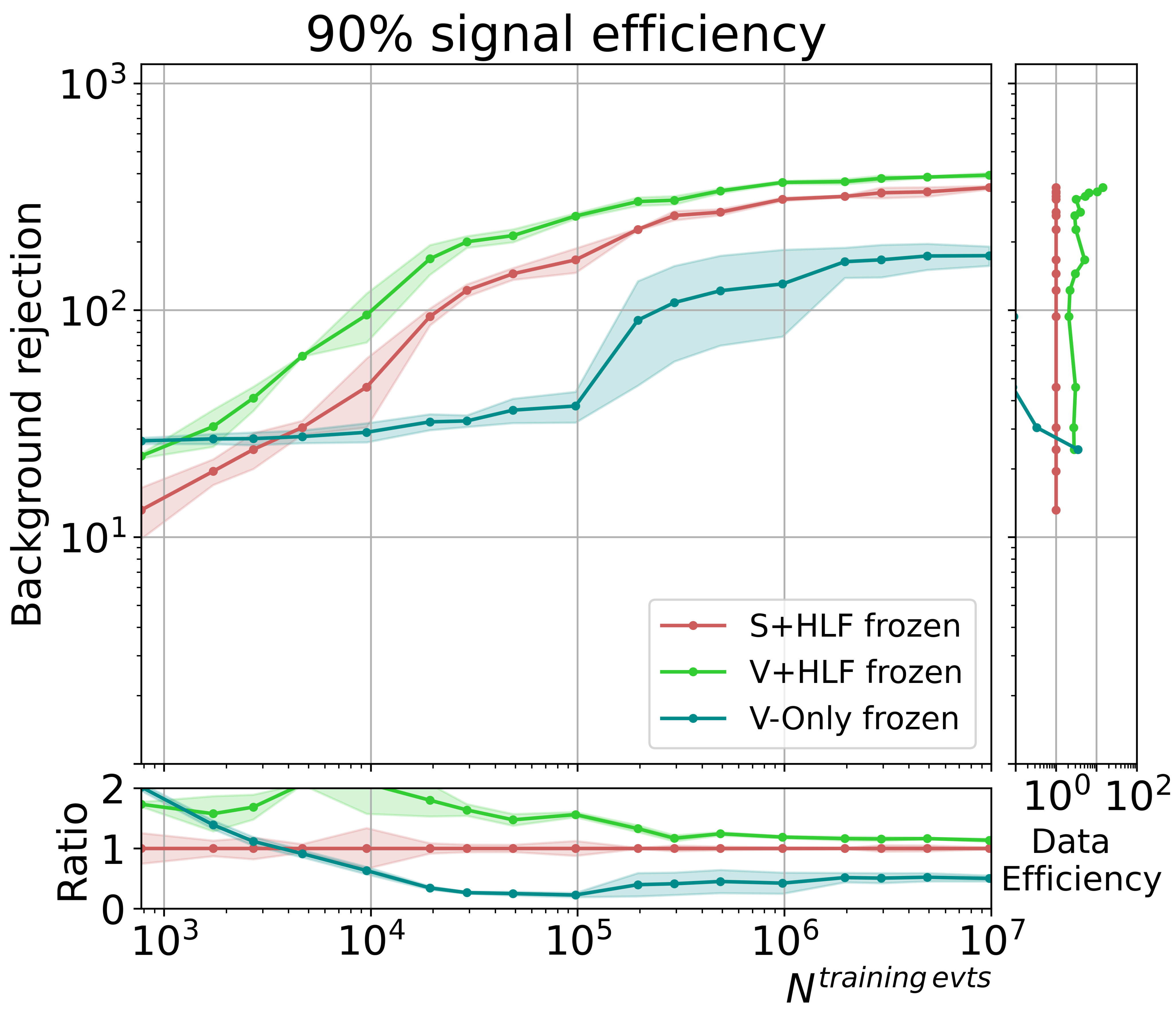
Lukas Heinrich, Nicole Hartman, Matthias Vigl



Xbb is learned when solving the downstream task even without actual jet labels

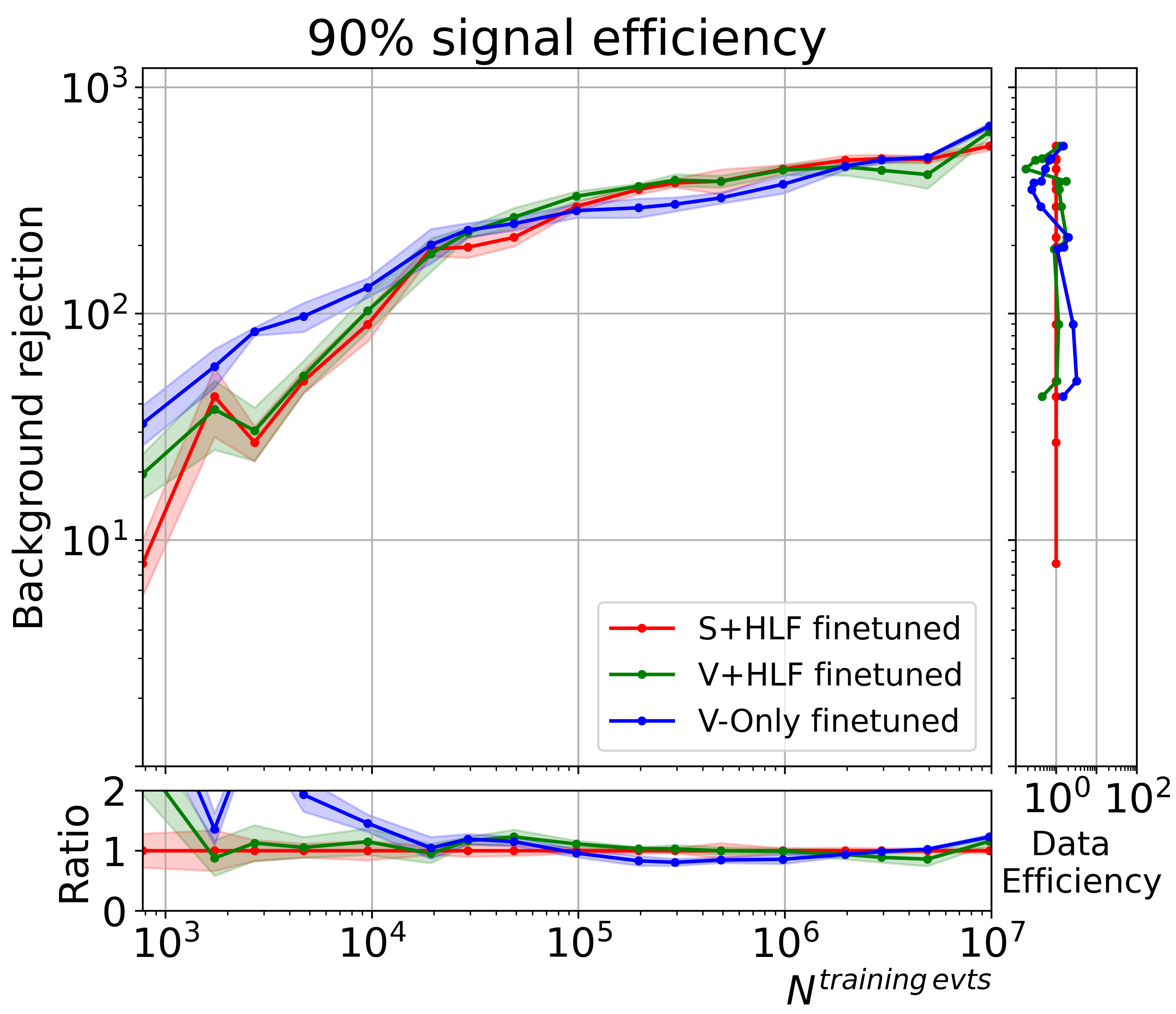
Lukas Heinrich, Nicole Hartman, Matthias Vigl

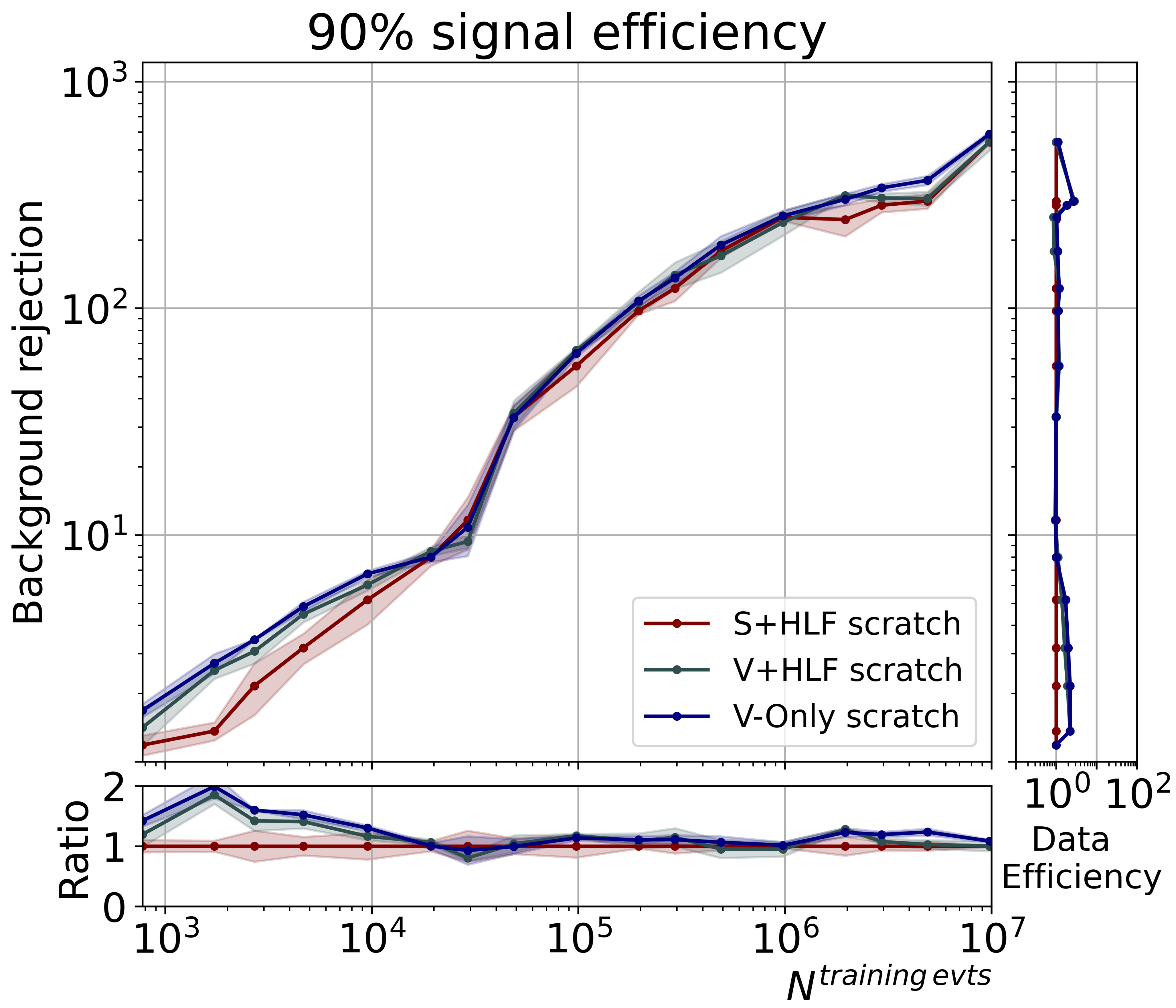
High dim embeddings help for frozen jet representations



Lukas Heinrich, Nicole Hartman, Matthias Vigl

Dimensionality becomes less important when training end-to-end





Lukas Heinrich, Nicole Hartman, Matthias Vigl

Setup: CMS open data and ParT

CMS open data: Duarte Javier, [http://opendata.cern.ch/record/12102]

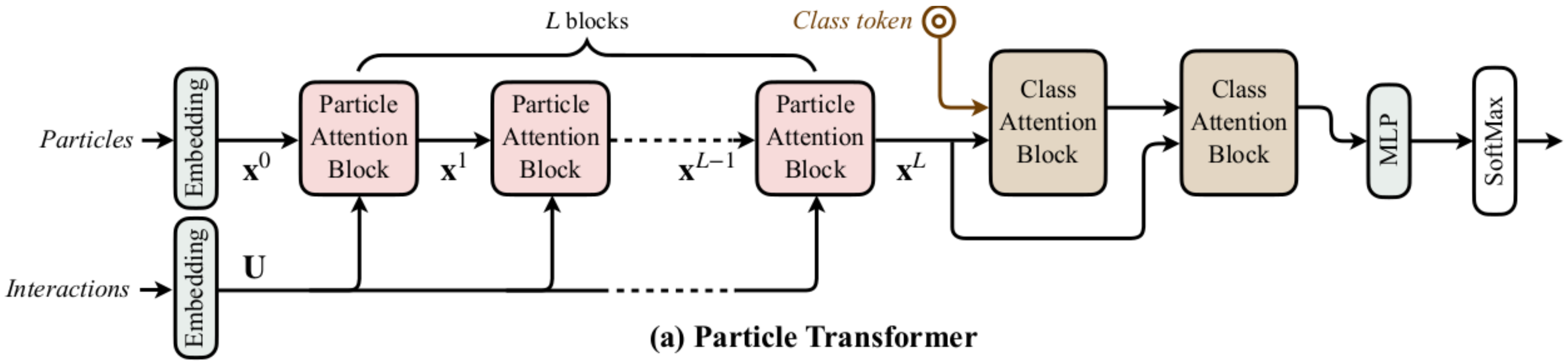
Jets are clustered using the anti-kT algorithm with R=0.8 from particle flow (PF) candidates

Constituents features:

- up to 100 PF per jet
- 17 features per PF

High-level features:

- Jet 4-momenta
- Xbb scores from ParT



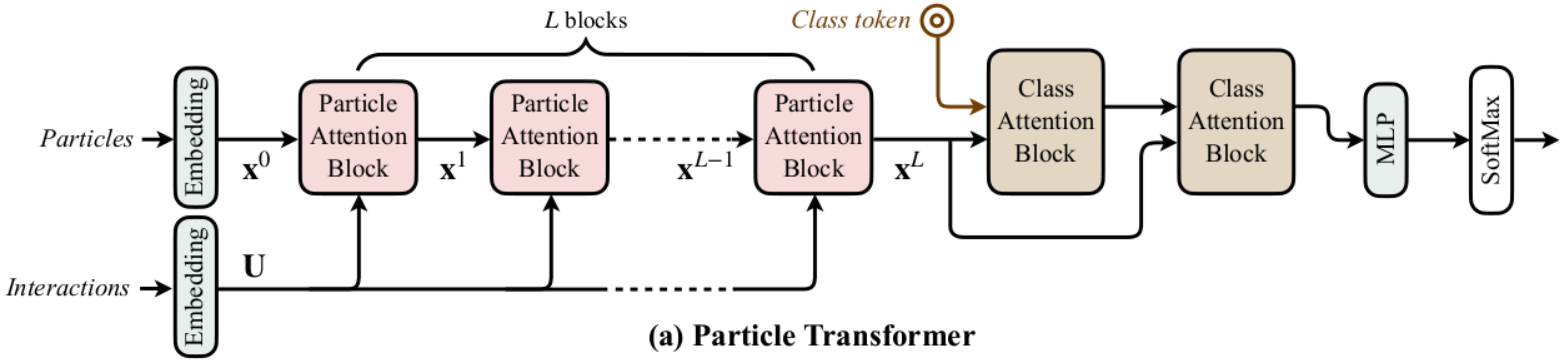
(a) Particle Transformer

Particle transformer for FTAG [arXiv:2202.03772]

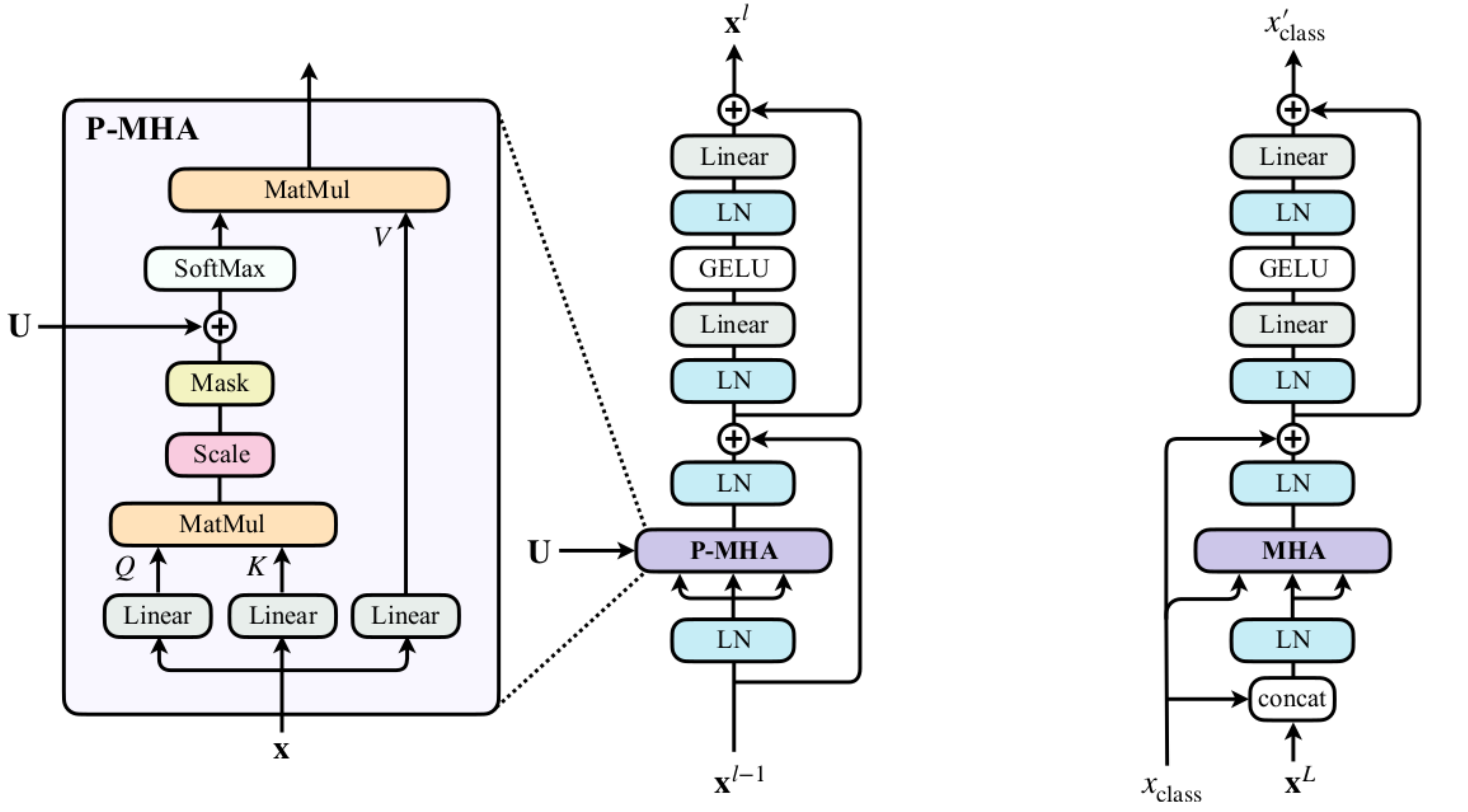
Training: QCD vs Higgs jets

10M events / 22M jets

ParT



(a) Particle Transformer



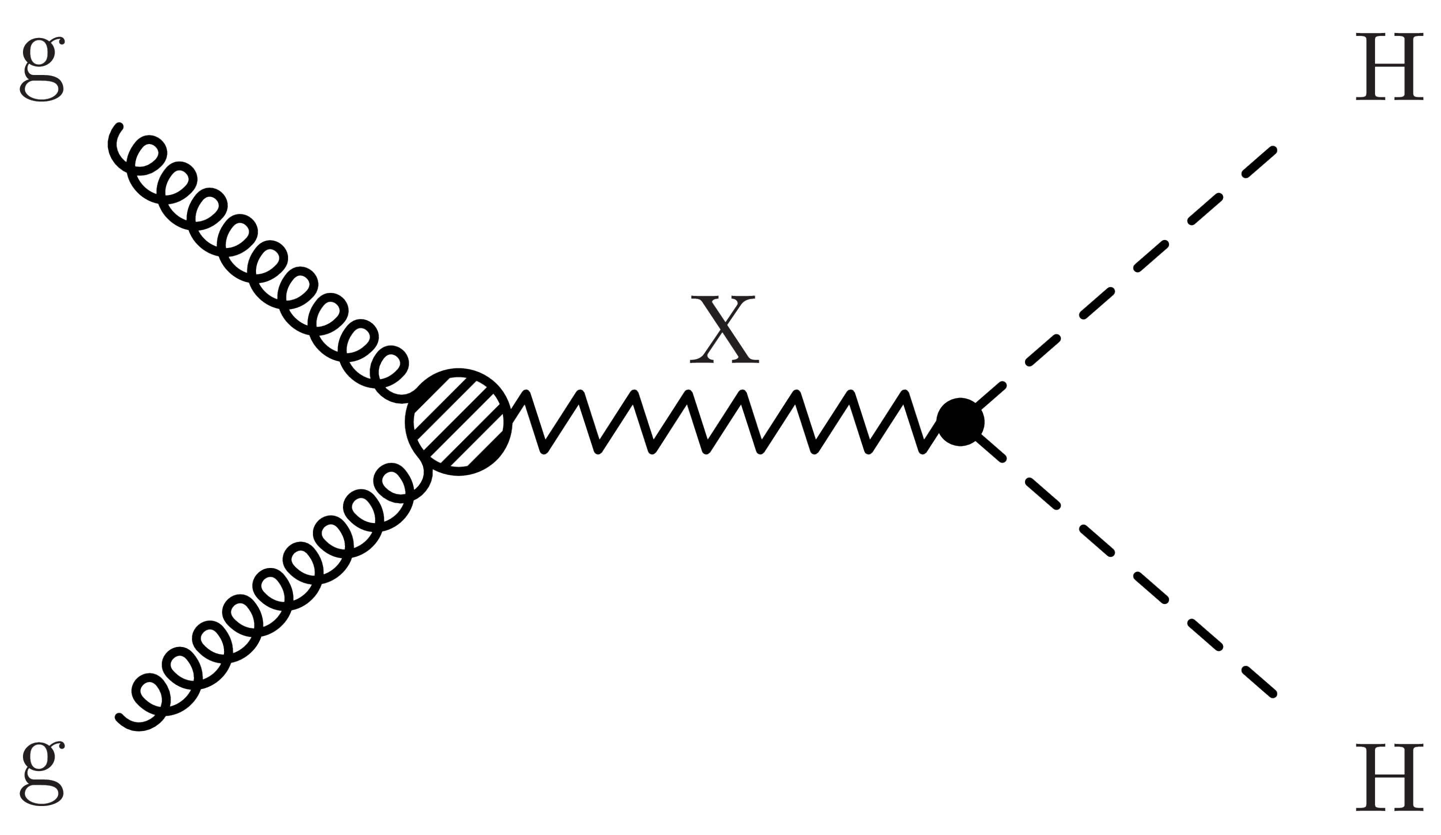
(b) Particle Attention Block

(c) Class Attention Block

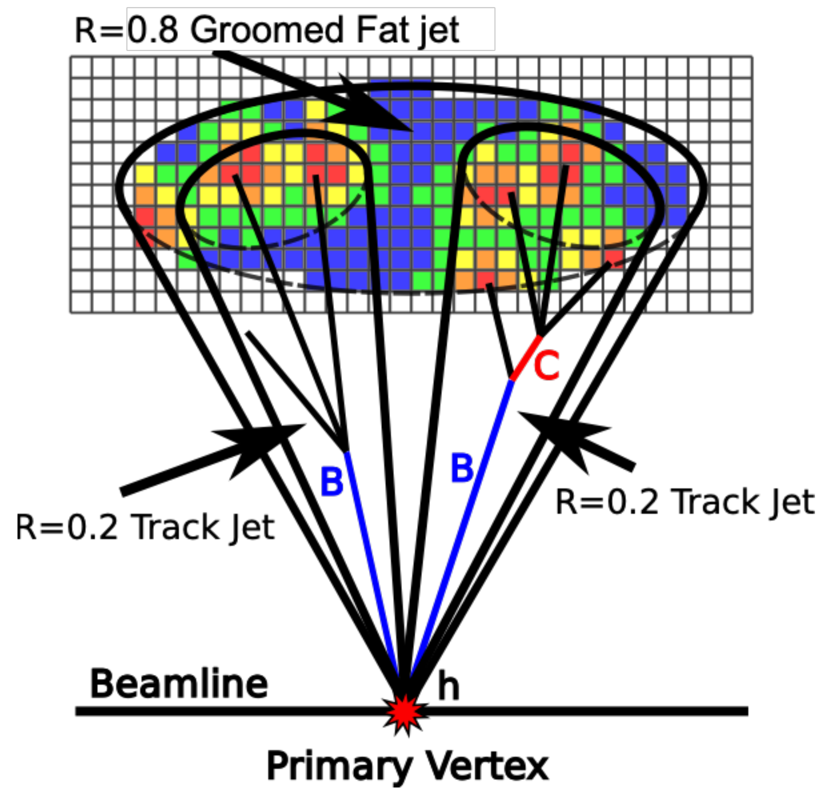
[arXiv:2202.03772]

CMS open data

- CMS simulated dataset:
- Sample with jet, track and secondary vertex properties for H(bb) tagging (http://opendata.cern.ch/record/12102)
- meant for jet tagging, up to 100 pf cand per jet - 17 feats each
- signal samples: 11 mass points - M_x from 600 GeV to 4500 GeV, bkg: QCD multijet
- 'fat jets' (fj) 4-momenta and (old) Xbb score



10M events / 22M jets




```
      - ['pfcand_pt_log', null]
      - ['pfcand_e_log', null]
      - ['pfcand_logptrel', null]
      - ['pfcand_logerel', null]
      - ['pfcand_deltaR', null]
      - ['pfcand_charge', null]
      - ['pfcand_isChargedHad', null]
      - ['pfcand_isNeutralHad', null]
      - ['pfcand_isGamma', null]
      - ['pfcand_isEl', null]
      - ['pfcand_isMu', null]
      - ['pfcand_dz', null]
      - ['pfcand_dzerr', null]
      - ['pfcand_dz', null]
      - ['pfcand_dzerr', null]
      - ['pfcand_deta', null]
      - ['pfcand_dphi', null]
pf_vectors:
   length: 110
   pad_mode: wrap
   vars:
      - [pfcand_px, null]
      - [pfcand_py, null]
      - [pfcand_pz, null]
      - [pfcand_energy, null]
```

[http://cms-results.web.cern.ch/cms-results/public-results/publications/BTV-16-002/]