Introducing Aspen Open Jets: a real-world ML-ready dataset for jet physics

<u>Ian Pang</u>, Oz Amram, Luca Anzalone, Joschka Birk, Darius Faroughy, Anna Hallin, Gregor Kasieczka, Michael Krämer, Alexander Mück, Humberto Reyes-Gonzalez, David Shih

Nov 4, 2024
ML4Jets, Paris

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
Fermilab
RUTGERS
RWTH AACHEN UNIVERSITY
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

ian.pang@physics.rutgers.edu

https://www.travelandleisure.com/travel-guide/aspen

New dataset named after Aspen, Colorado where the initial ideas were discussed!

Outline

1. Aspen Open Jets (AOJ)
    - Dataset overview
    - First ML-ready dataset with real jets
    - Jet and constituent features
2. AOJ for ML
    - Using AOJ
    - Unsupervised pre-training
    - Results

Aspen Open Jets

Dataset overview

- **CMS** released 16.4 fb^{-1} of data from their 2016 run (CMS open data ‘JetHT’)
- Data provided in MINIAOD format

- We then processed the data to PFNANO format
- Select **AK8 jets** of interests:
  - One or more triggers related to jet momenta or total event hadronic activity
  - **Jet $p_T >$ 300 GeV , jet $|\eta| <$ 2.5**
  - Other data quality filters
- **Total of ~180M jets in ML-ready format!**

Aspen Open Jets

Dataset overview

Not easily useable for ML tasks

- **CMS** released 16.4 fb^{-1} of data from their 2016 run (CMS open data 'JetHT')
- Data provided in MINIAOD format

- We then processed the data to PFNANO format
- Select **AK8 jets** of interests:
  - One or more triggers related to jet momenta or total event hadronic activity
  - **Jet $p_T >$ 300 GeV , jet $|\eta| <$ 2.5**
  - Other data quality filters
- **Total of ~180M jets in ML-ready format!**

Aspen Open Jets

Dataset overview

- **CMS** released 16.4 fb^{-1} of data from their 2016 run (CMS open data 'JetHT')
- Data provided in MINIAOD format

- We then processed the data to PFNANO format
- Select **AK8 jets** of interests:
    - One or more triggers related to jet momenta or total event hadronic activity
    - **Jet $p_T >$ 300 GeV , jet $|\eta| <$ 2.5**
    - Other data quality filters
- **Total of ~180M jets in ML-ready format!**

Aspen Open Jets

First ML-ready dataset with real jets

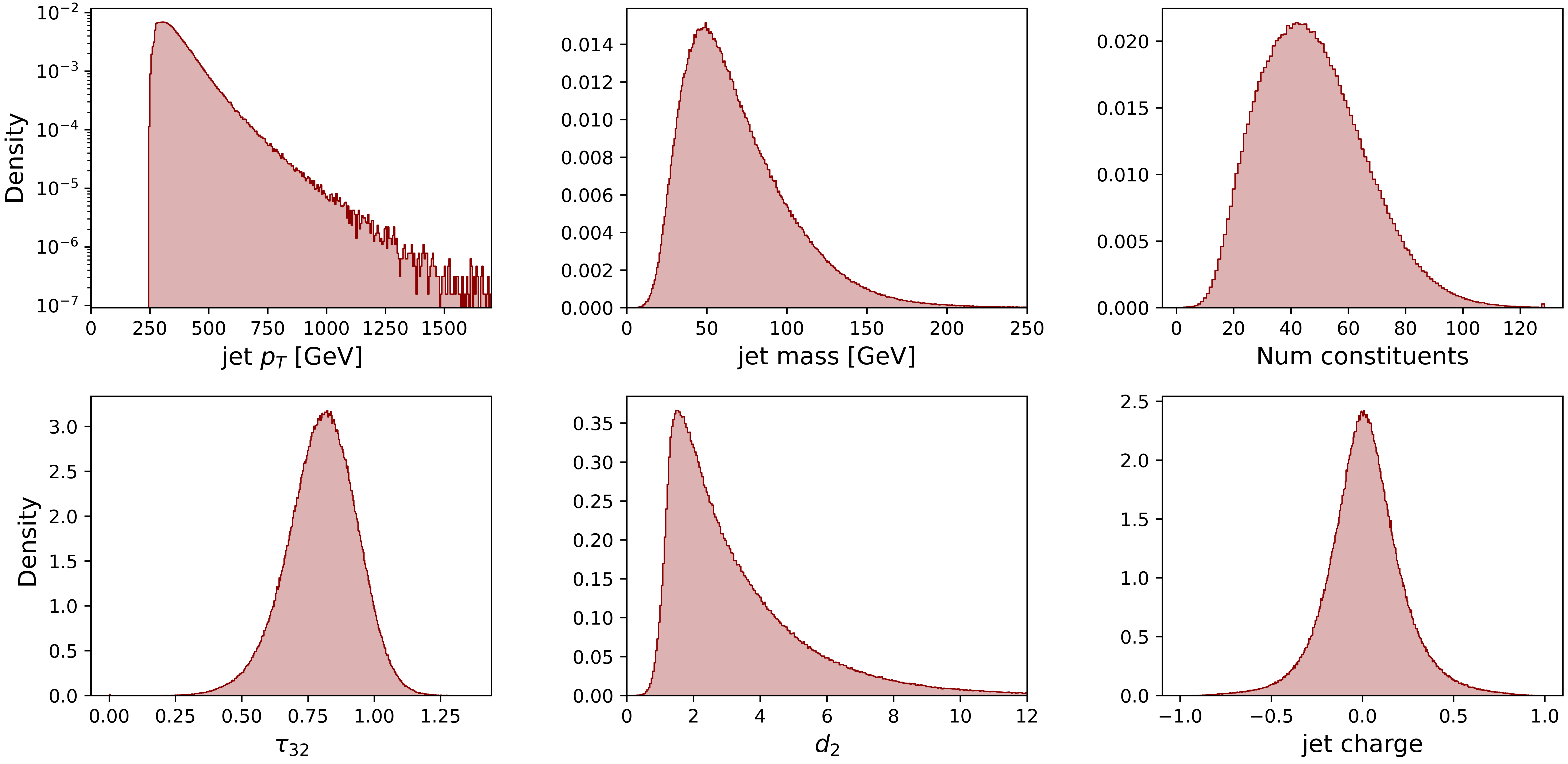
- Total of ~180M jets!
- Mostly QCD jets
- For **each jet**:
  - Jet p_T, η, ϕ
  - Soft drop mass
  - N-subjettiness: $\tau_1, \tau_2, \tau_3, \tau_4$
  - Number of constituents

- Up to 150 constituents per jet
- For **each constituent**:
  - 4-momenta (p_x, p_y, p_z, E)
  - Trajectory displacements d_0 , d_z and their uncertainties σ_{d_0} , σ_{d_z}
  - Particle charge and PID
  - PUPPI weights

Aspen Open Jets

Jet and constituent features

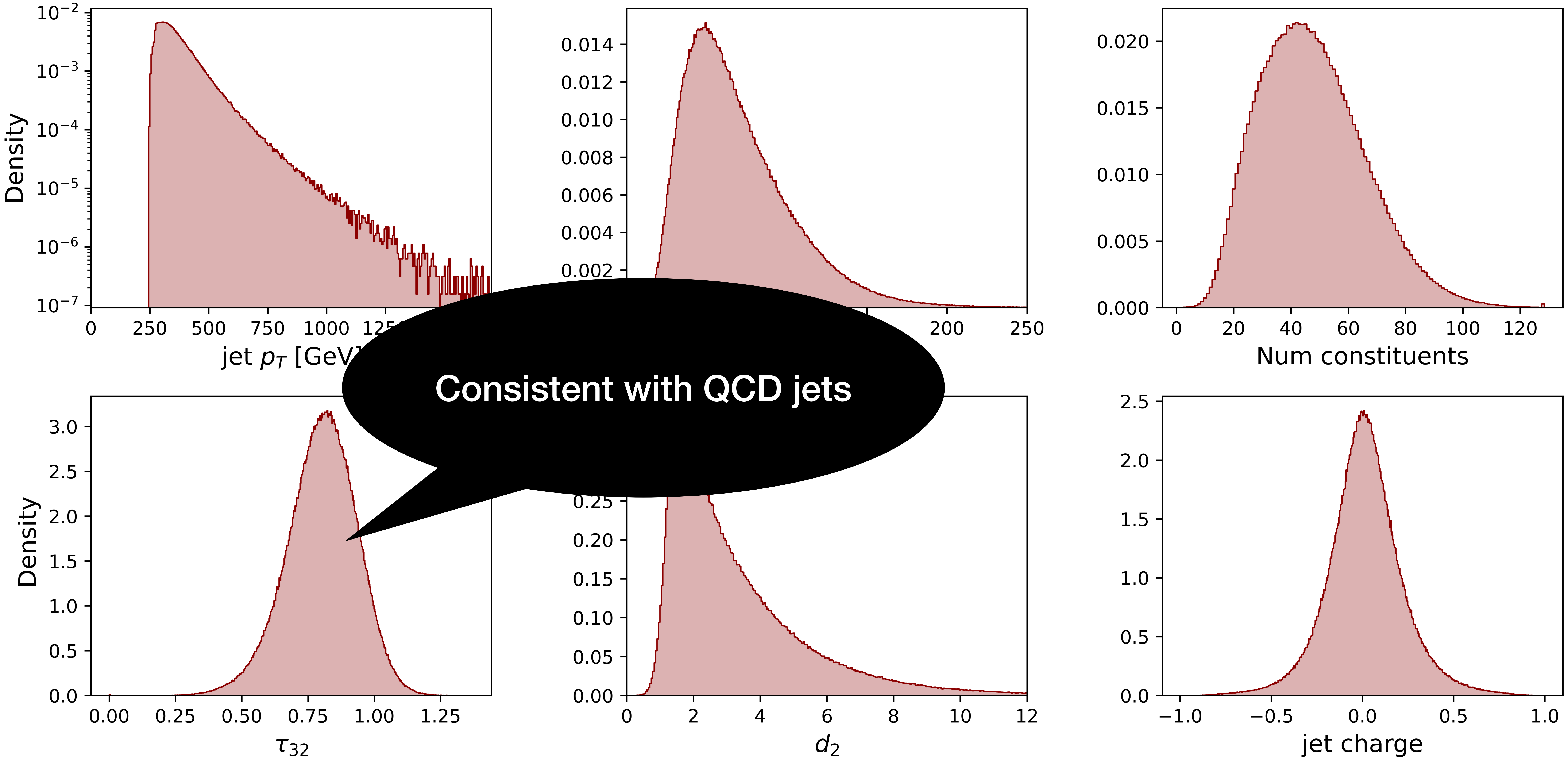
Features in plots are computed from jet constituents



Aspen Open Jets

Jet and constituent features

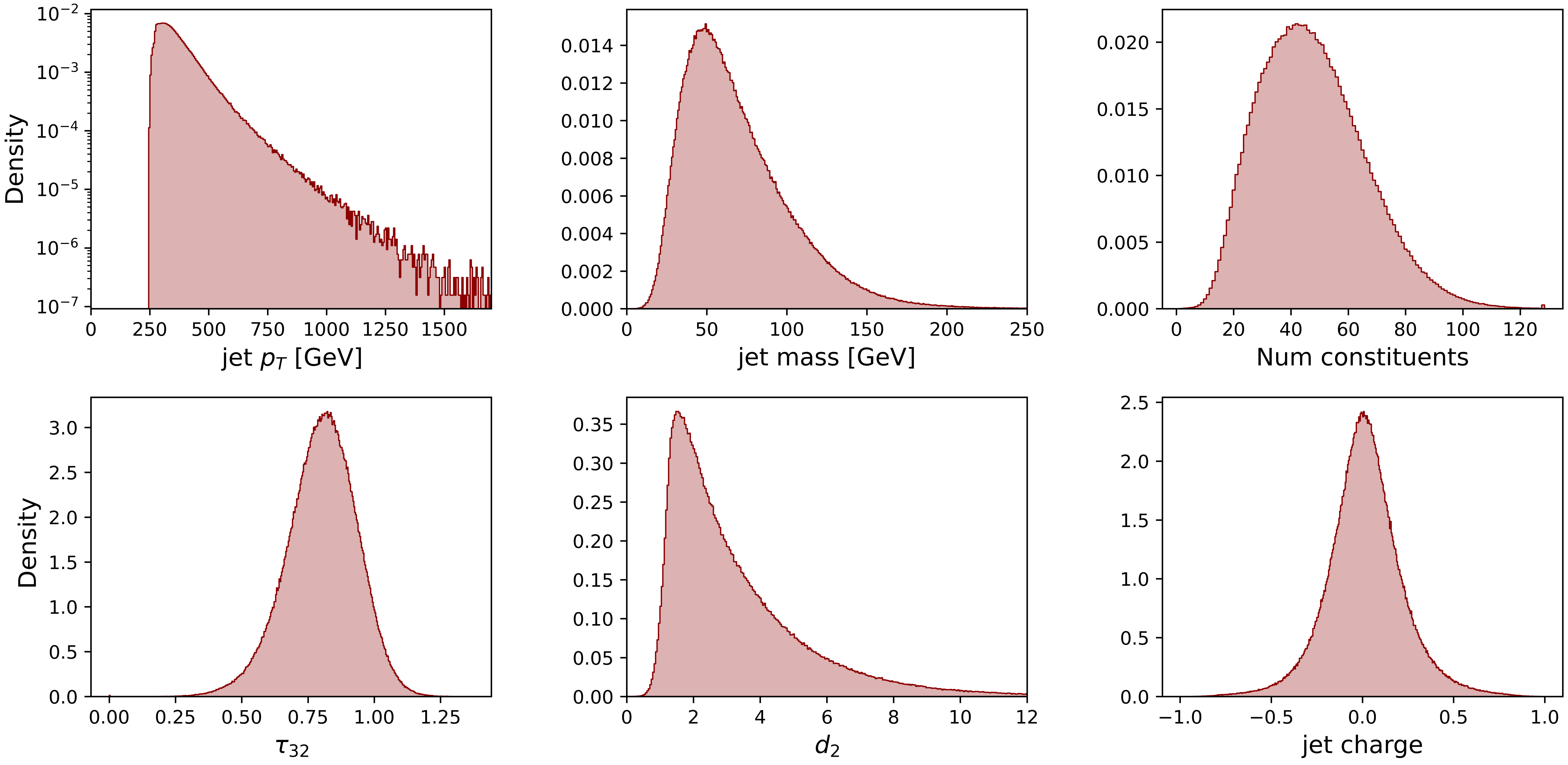
Features in plots are computed from jet constituents



Aspen Open Jets

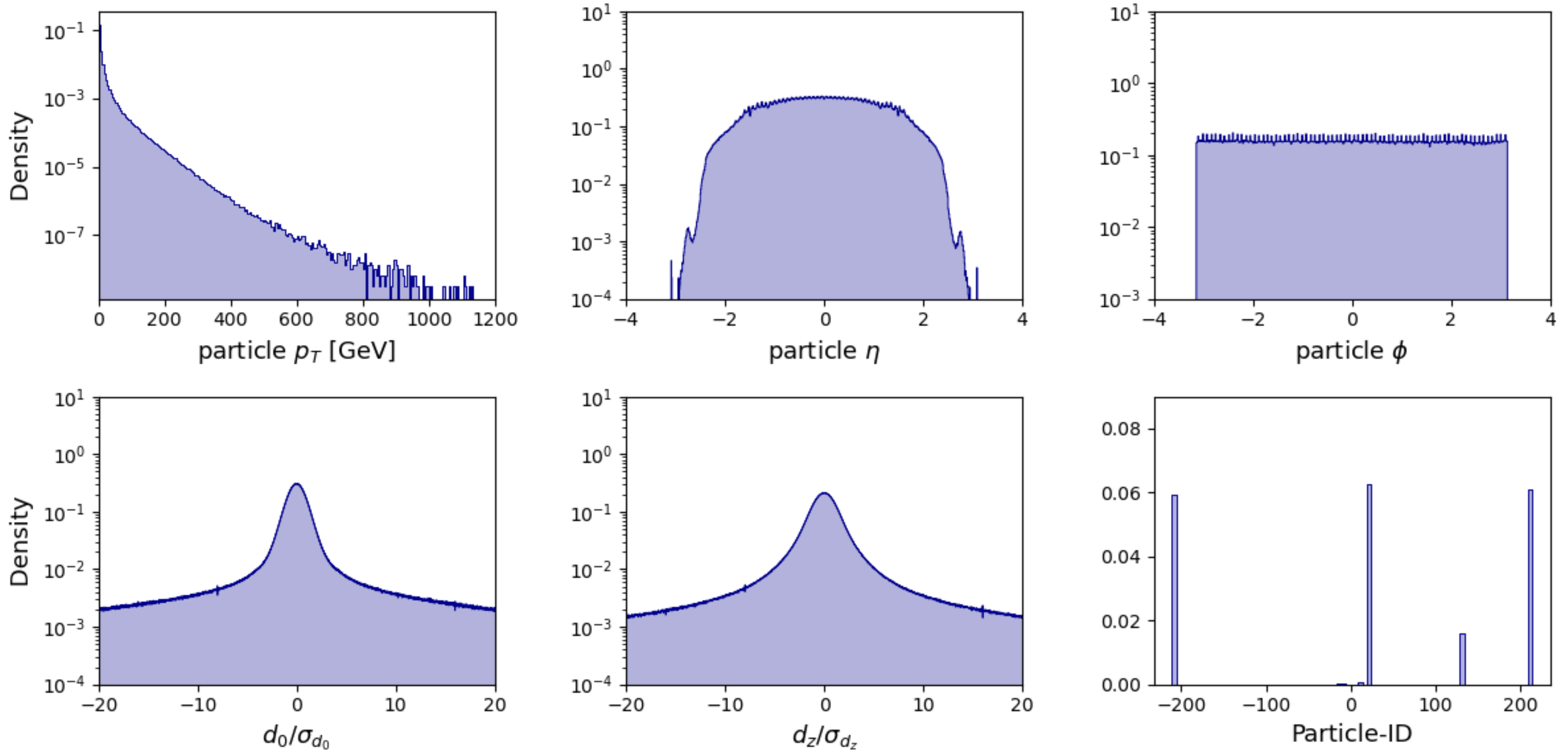
Jet and constituent features

Features in plots are computed from jet constituents



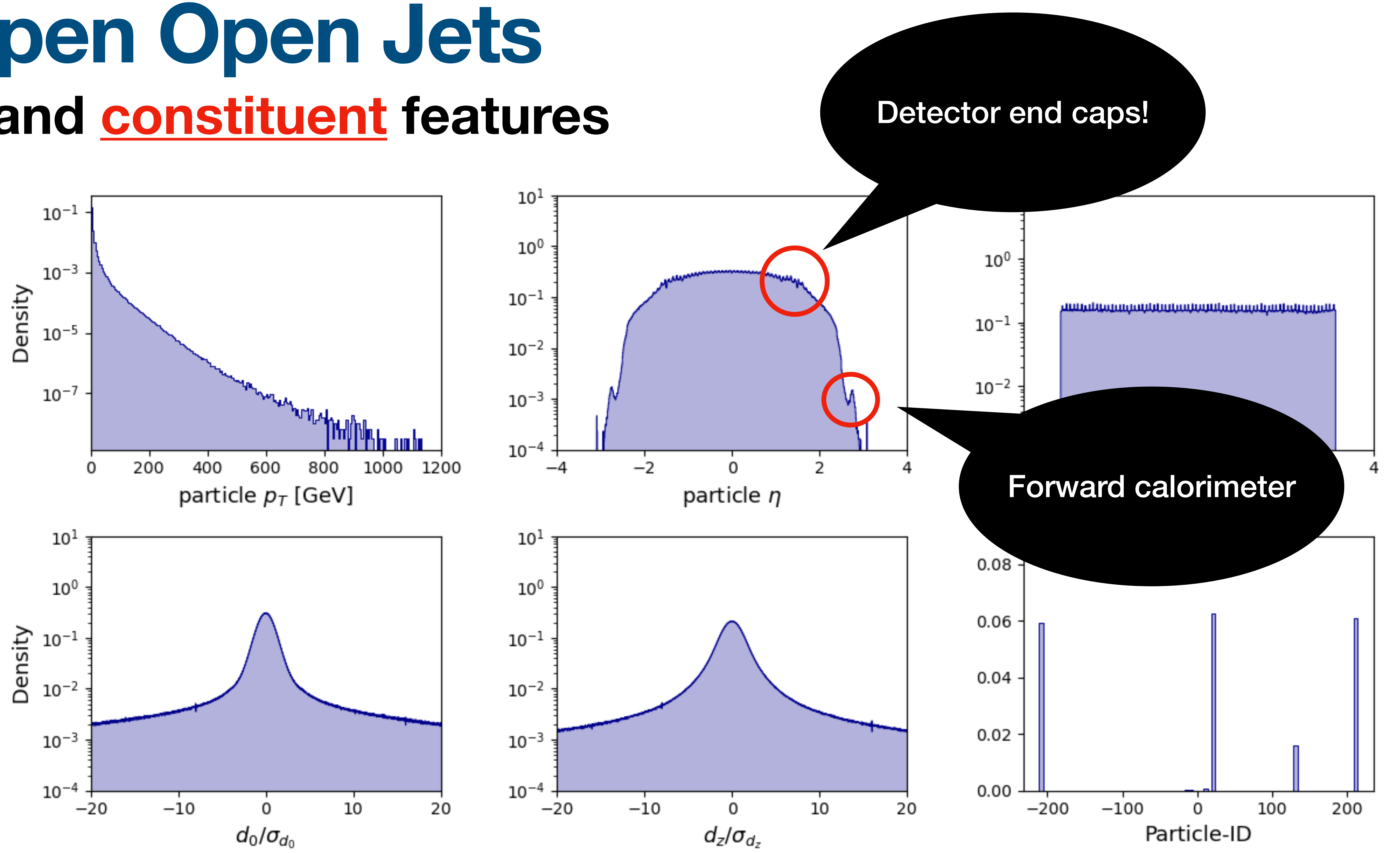
Aspen Open Jets

Jet and constituent features



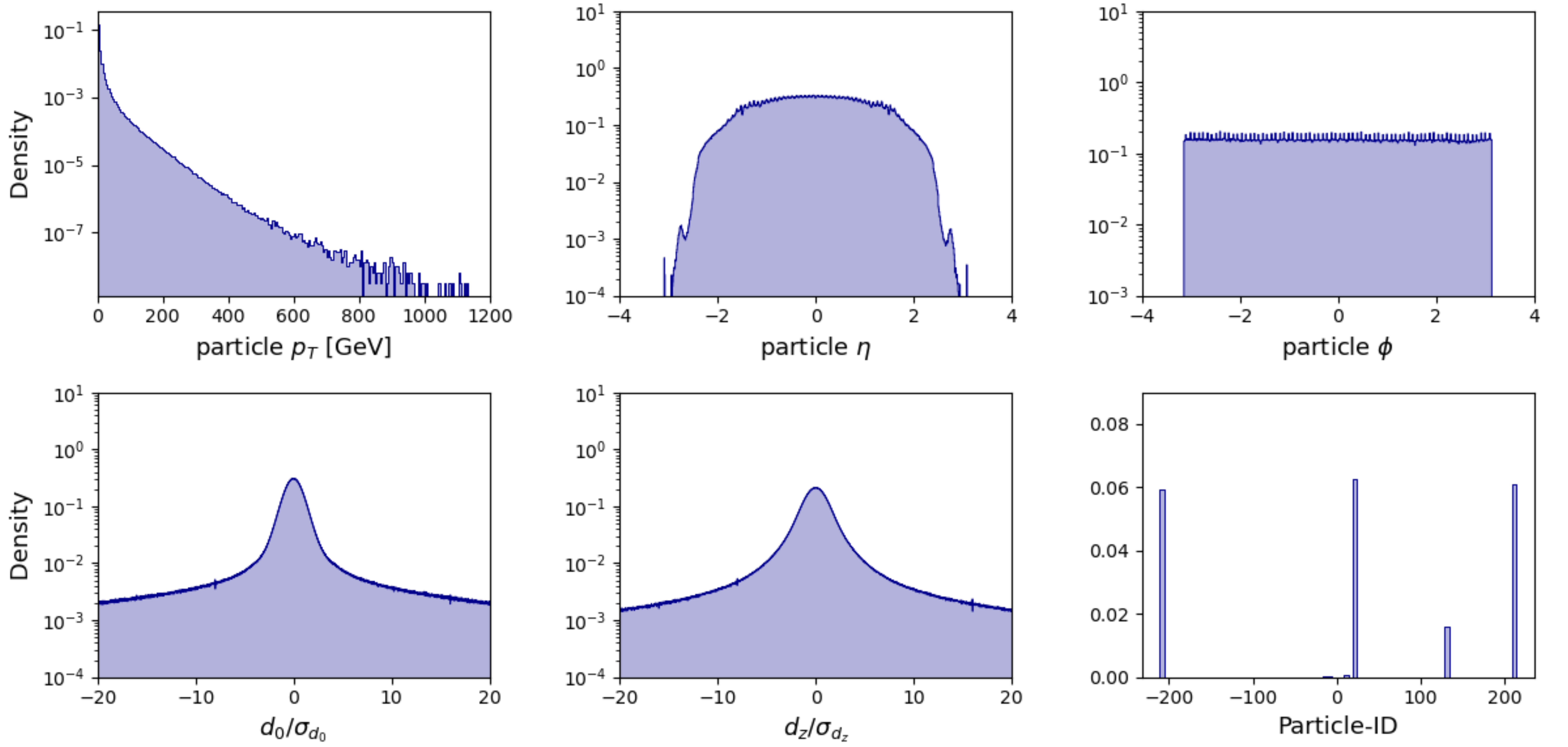
Aspen Open Jets

Jet and constituent features



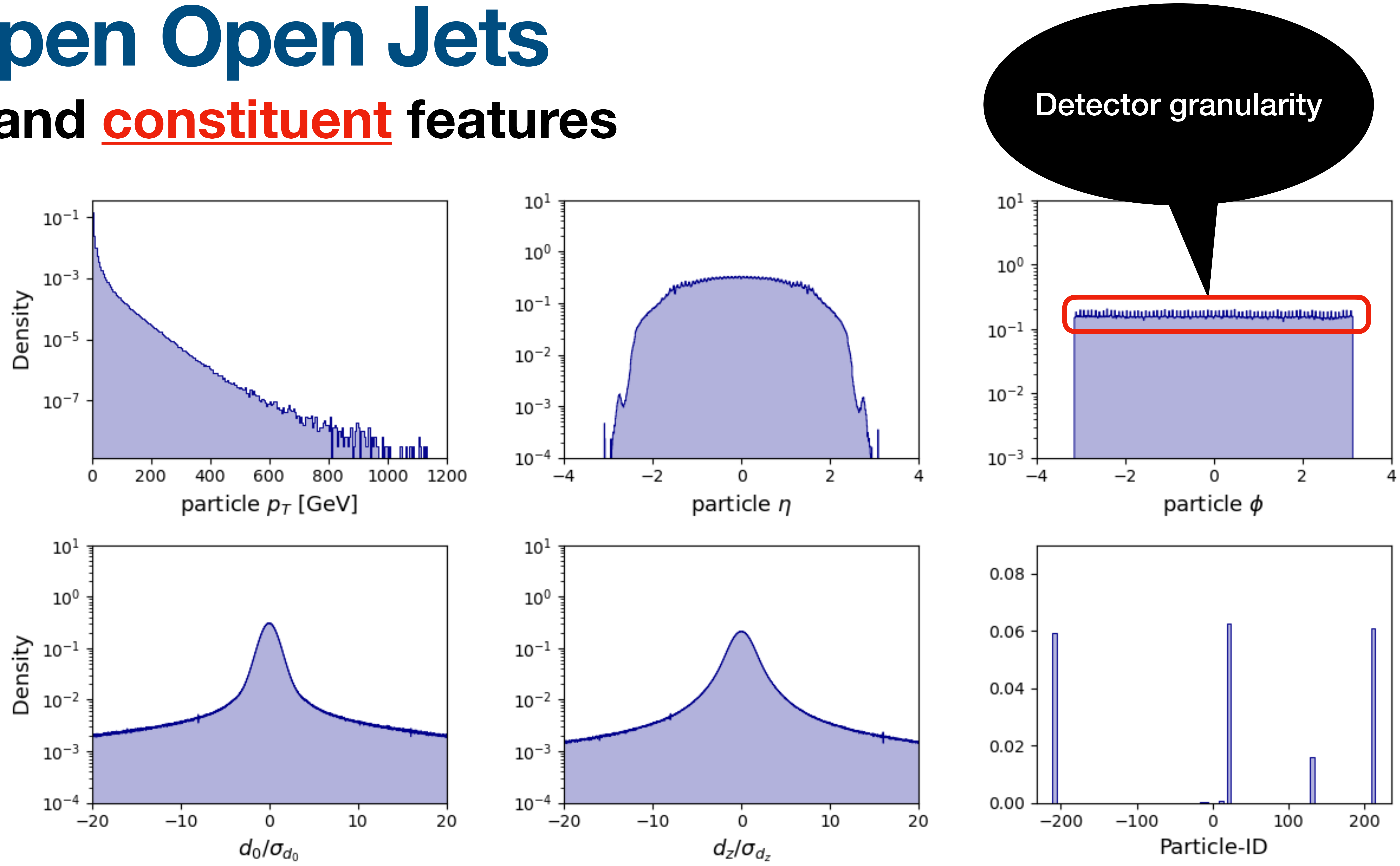
Aspen Open Jets

Jet and constituent features



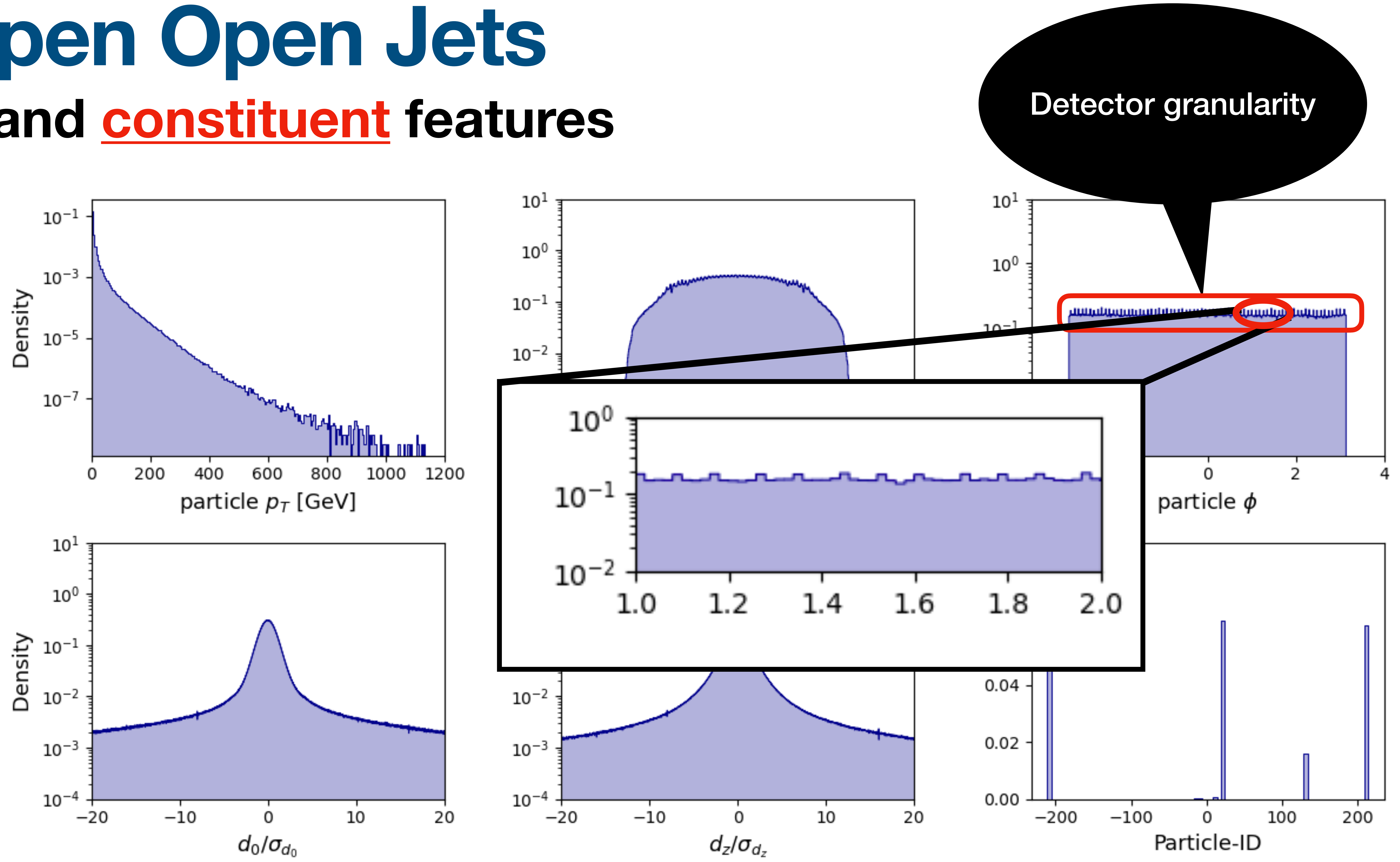
Aspen Open Jets

Jet and constituent features



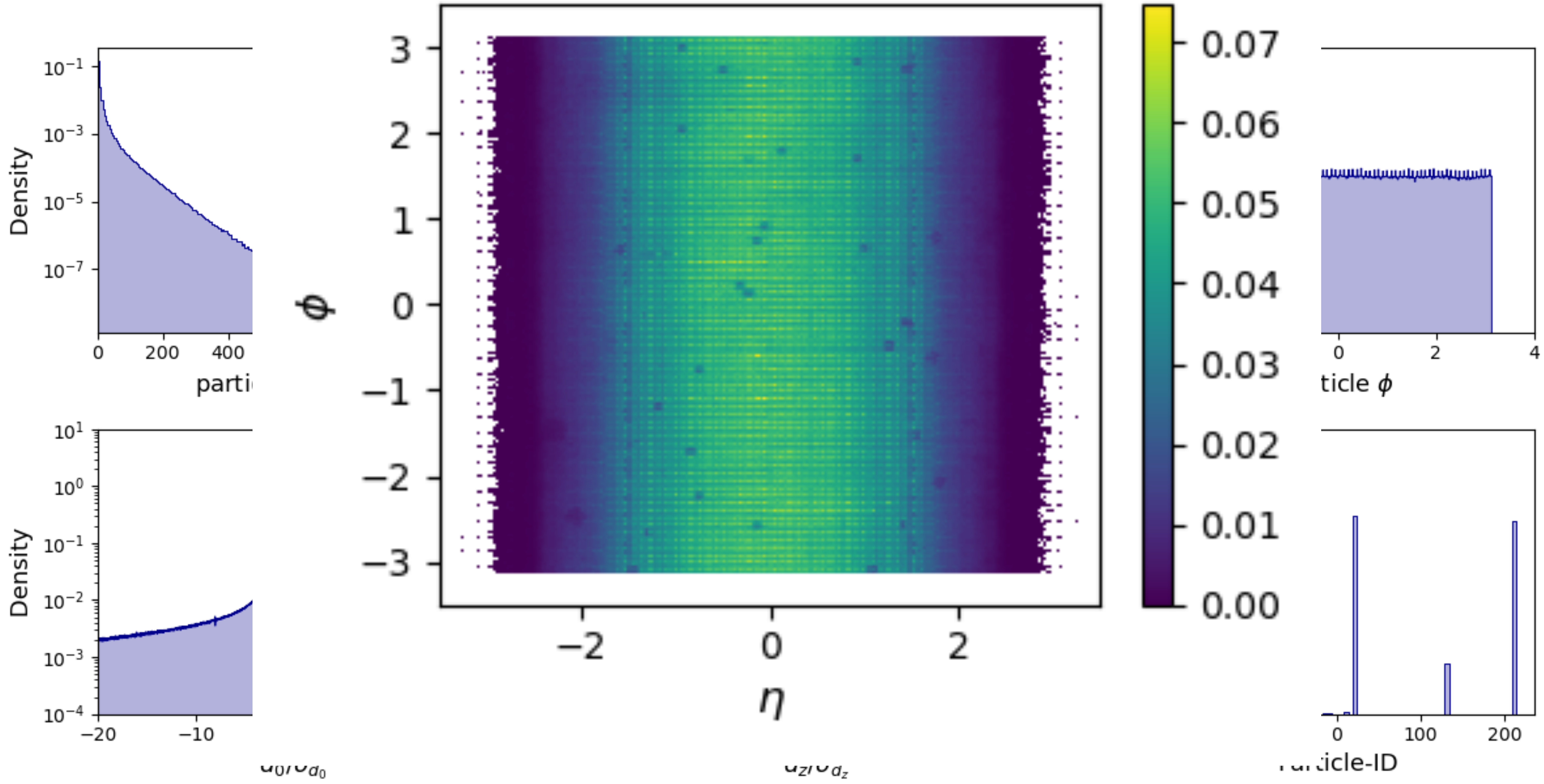
Aspen Open Jets

Jet and constituent features



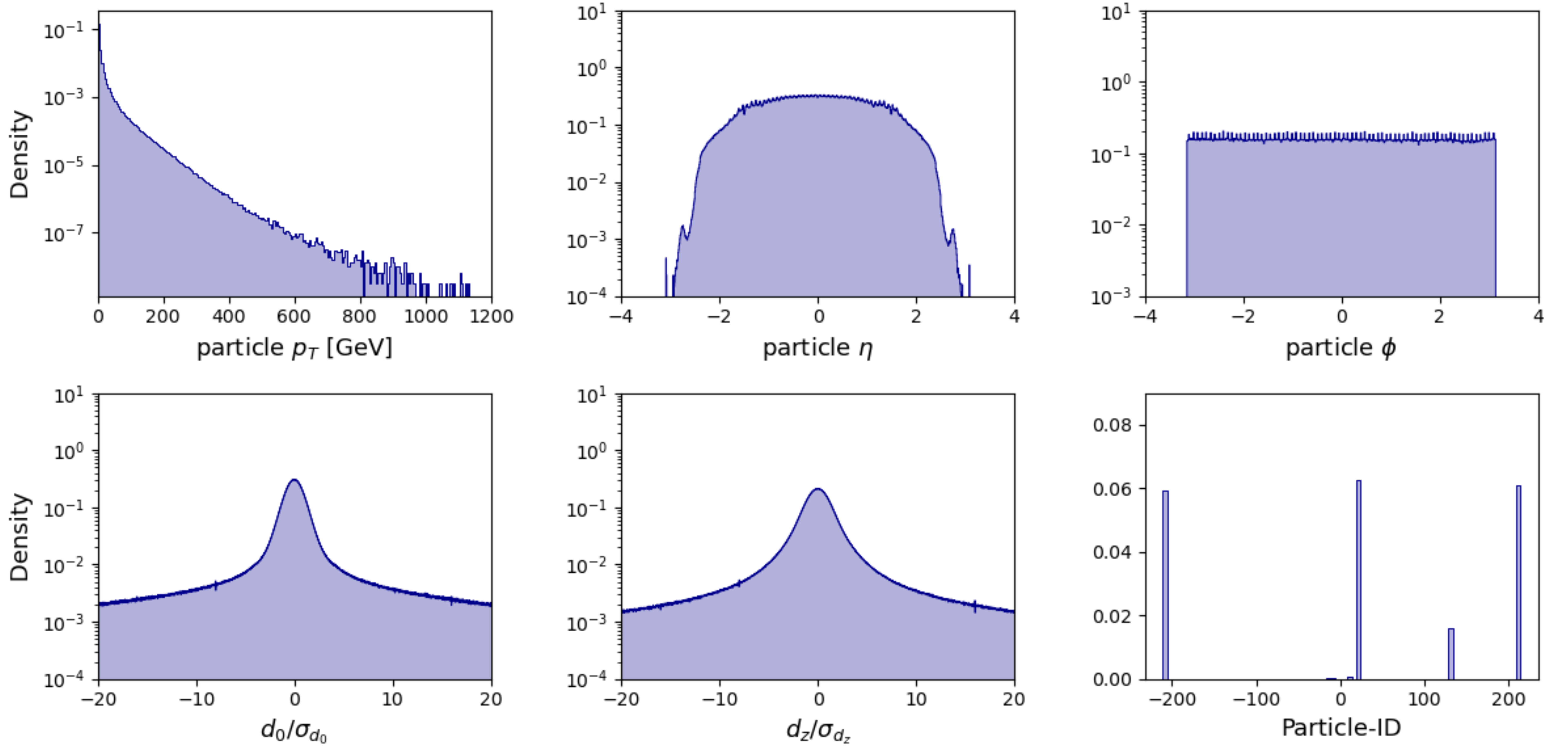
Aspen Open Jets

Jet and constituent features



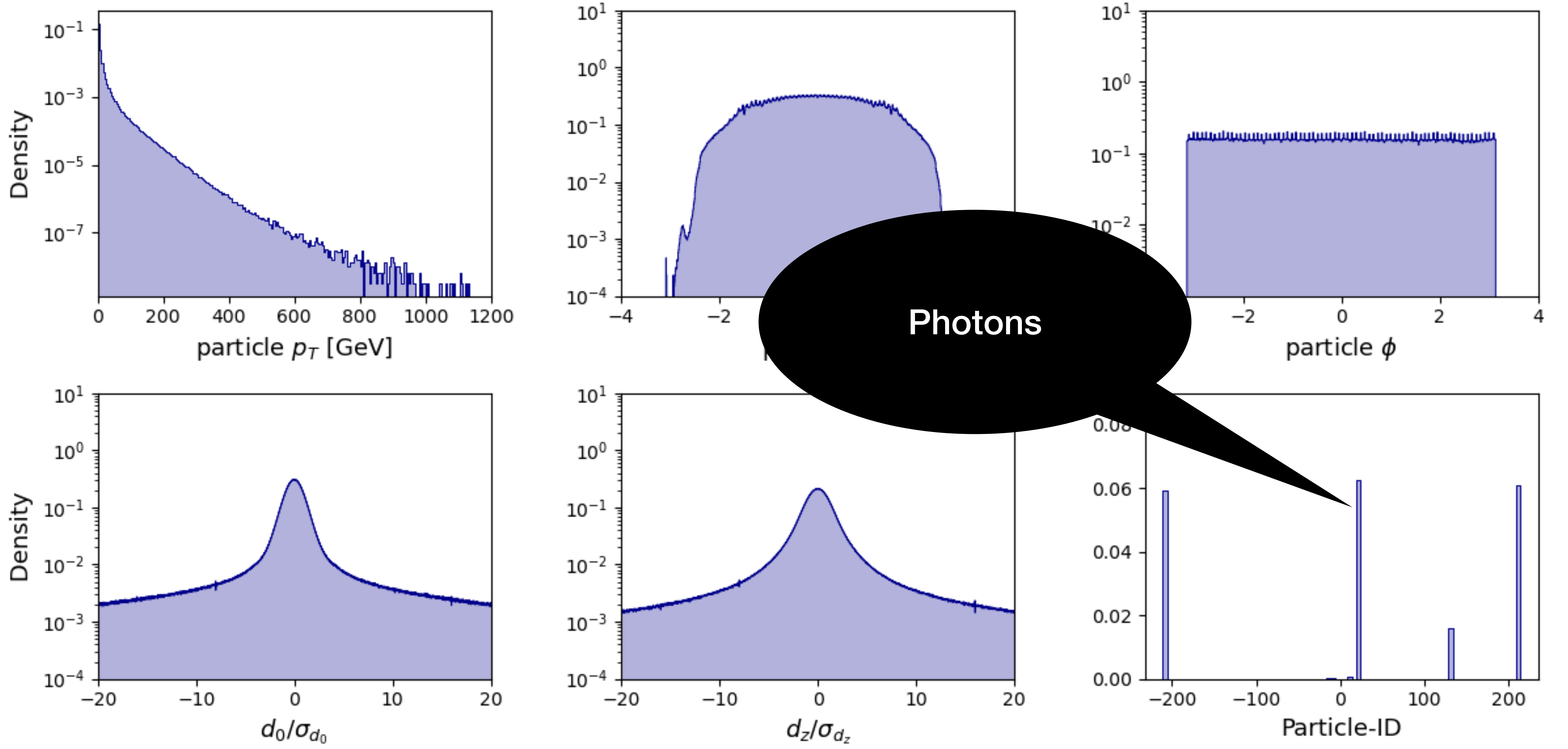
Aspen Open Jets

Jet and constituent features



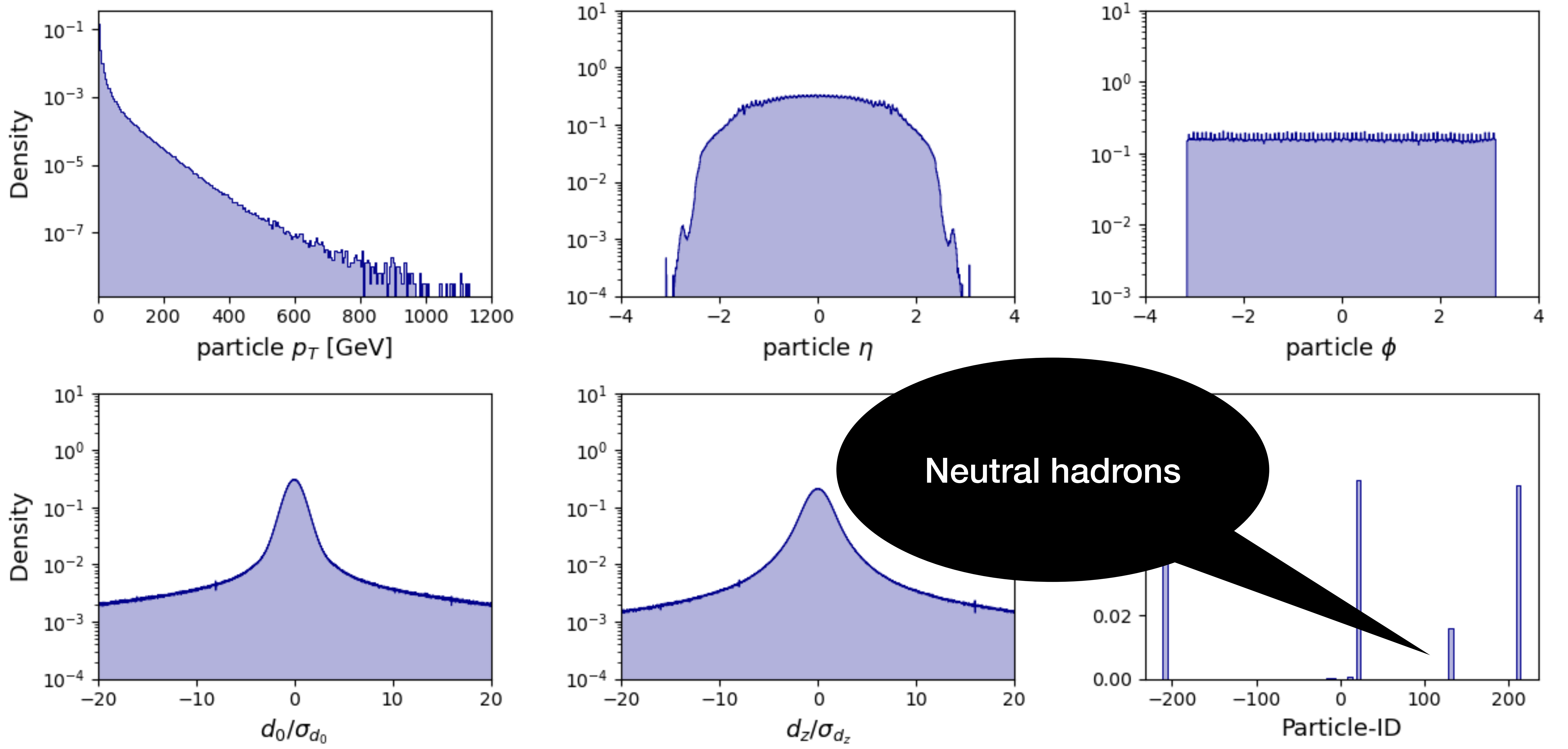
Aspen Open Jets

Jet and constituent features



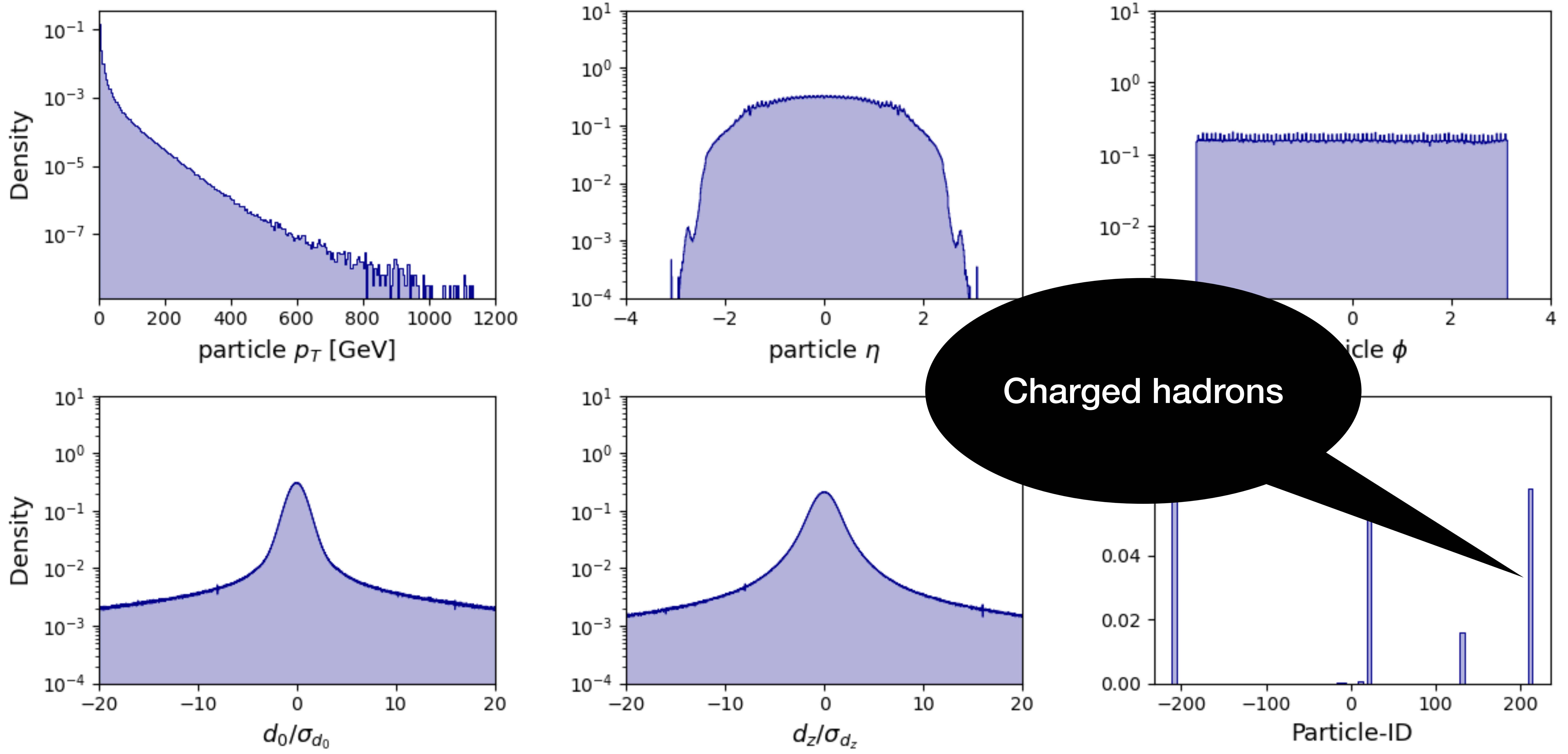
Aspen Open Jets

Jet and constituent features



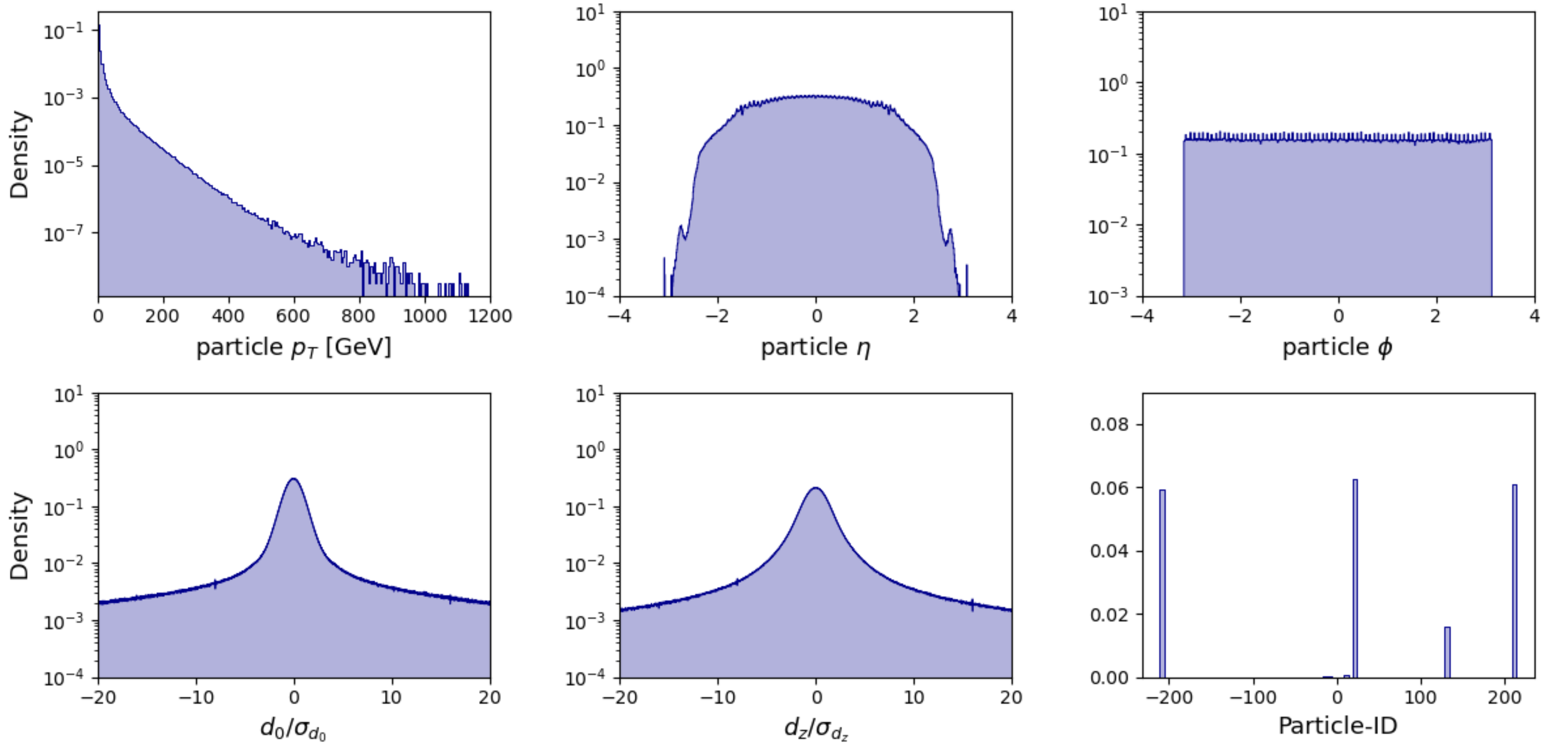
Aspen Open Jets

Jet and constituent features



Aspen Open Jets

Jet and constituent features



AOJ for ML

Using AOJ

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data

AOJ for ML

Using AOJ

No explicit class info about type of jet (e.g. top jet)

- AOJ is a large dataset (~180M jets) of <u>unlabeled data</u>

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data

Different from JetClass [1] dataset which has 125M jets with a total 10 jet types

[1] https://zenodo.org/records/6619768

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of <u>unlabeled data</u>
- Expected to be <u>mostly QCD jets</u> (~125k top jets)

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data
- Expected to be mostly QCD jets (~125k top jets)
- Can we train pre-train a large model in an unsupervised way on AOJ?

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data
- Expected to be mostly QCD jets (~125k top jets)
- Can we train pre-train a large model in an unsupervised way on AOJ?
- Does finetuning the pre-trained model on downstream tasks provide performance gain?

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data
- Expected to be mostly QCD jets (~125k top jets)
- Can we train pre-train a large model in an unsupervised way on AOJ?
- Does finetuning the pre-trained model on downstream tasks provide performance gain?
- **Example:**

AOJ for ML

Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data
- Expected to be mostly QCD jets (~125k top jets)
- Can we train pre-train a large model in an unsupervised way on AOJ?
- Does finetuning the pre-trained model on downstream tasks provide performance gain?
- **Example:**
    1. **Pre-train generative model on AOJ (~180M jets)**

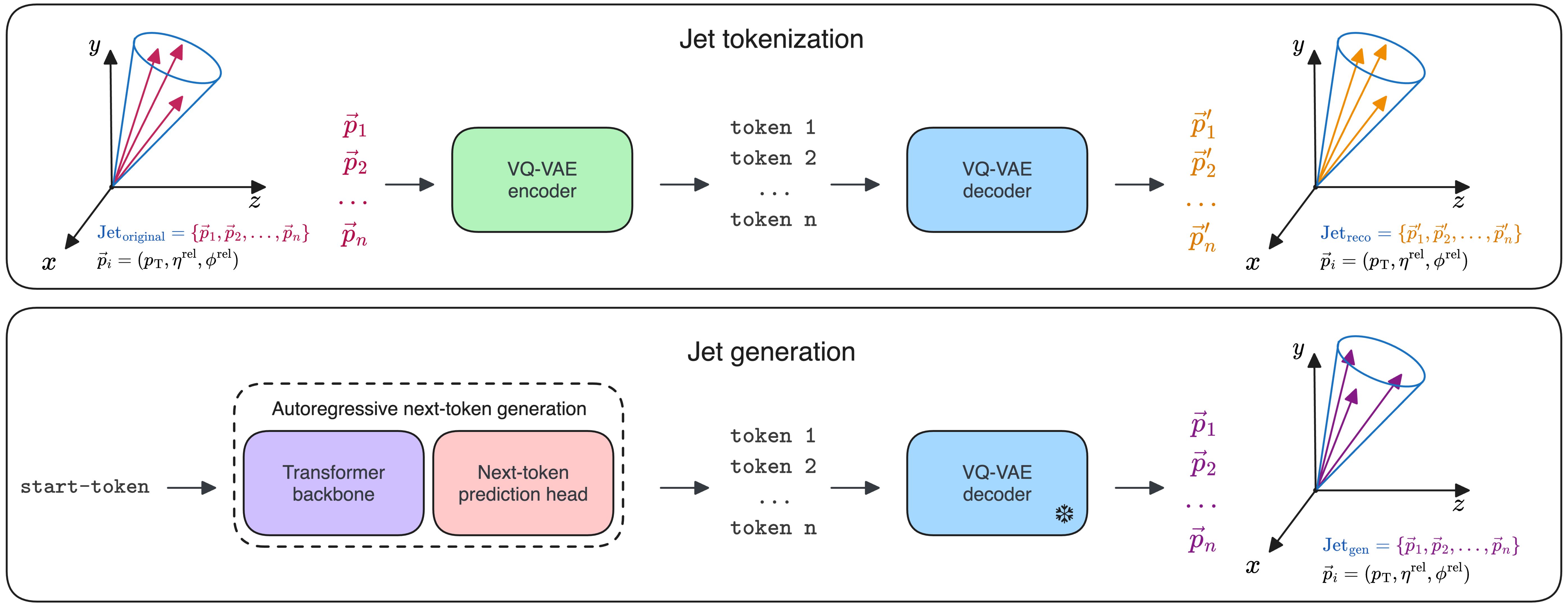
AOJ for ML

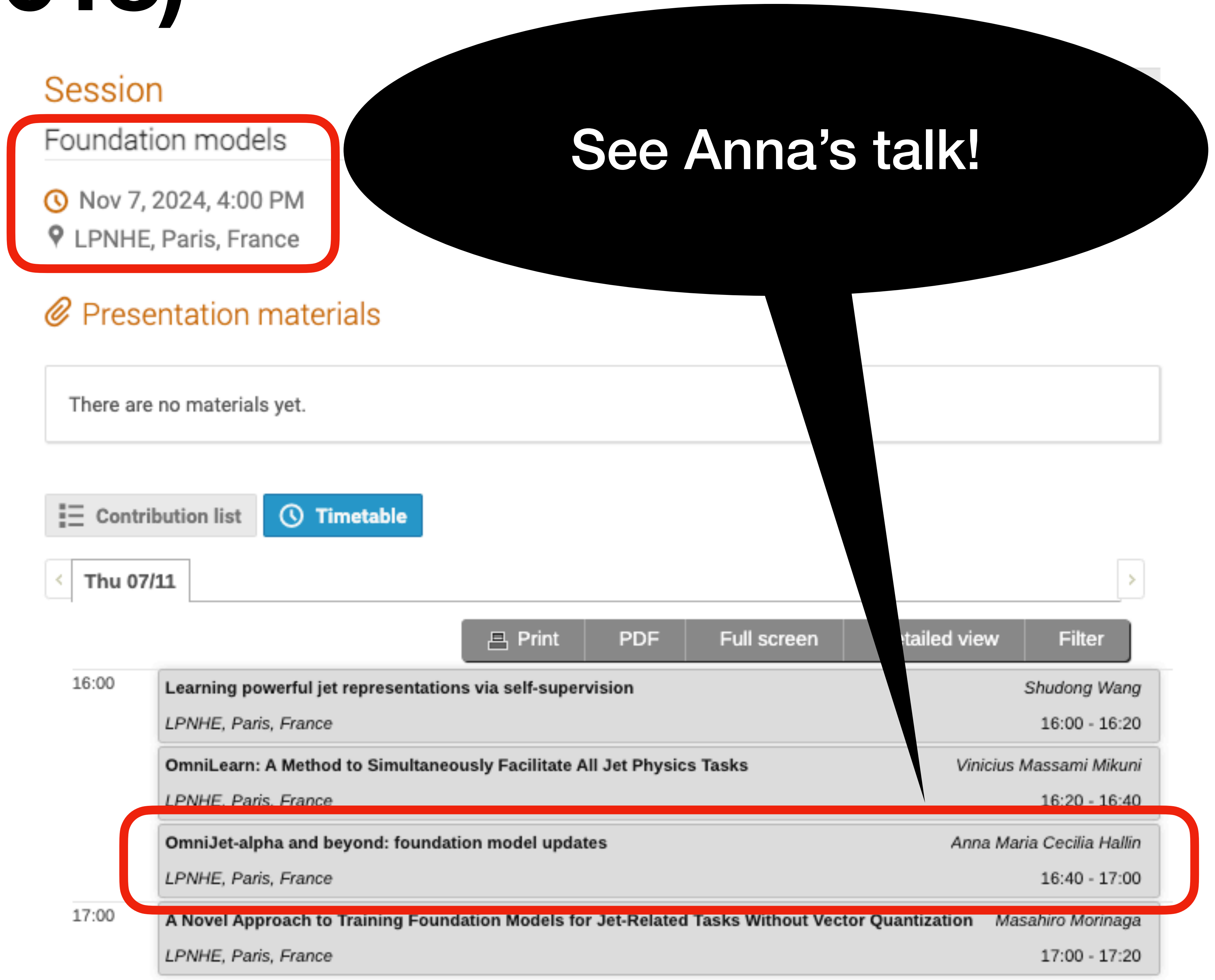
Using AOJ

- AOJ is a large dataset (~180M jets) of unlabeled data
- Expected to be mostly QCD jets (~125k top jets)
- Can we train pre-train a large model in an unsupervised way on AOJ?
- Does finetuning the pre-trained model on downstream tasks provide performance gain?
- **Example:**
    1. **Pre-train generative model on AOJ (~180M jets)**
    2. **Fine-tune on generating JetClass top jets**

Unsupervised pre-training

Based on Omnijet-α architecture (2403.05618)

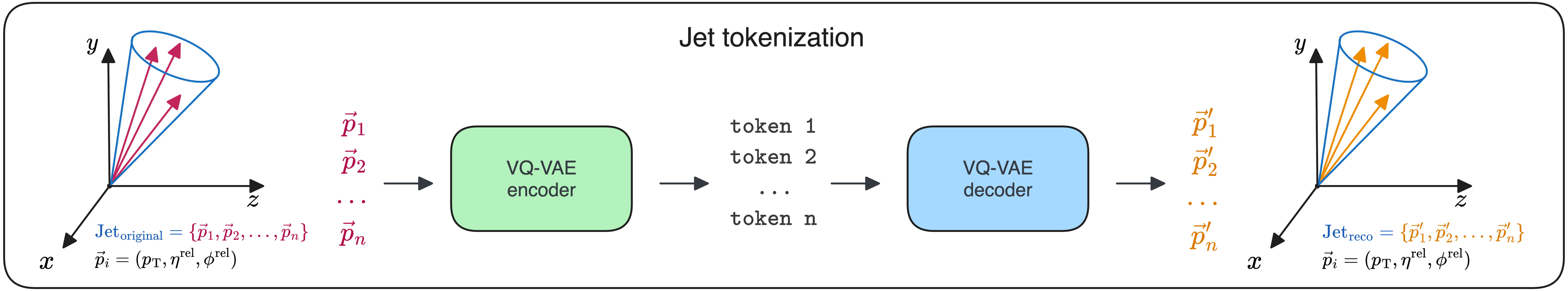




- Tokenized jet constituents $\left(p_T, \eta^{\mathrm{rel}}, \phi^{\mathrm{rel}}\right)$
- GPT-style generation: Next-token prediction

Unsupervised pre-training

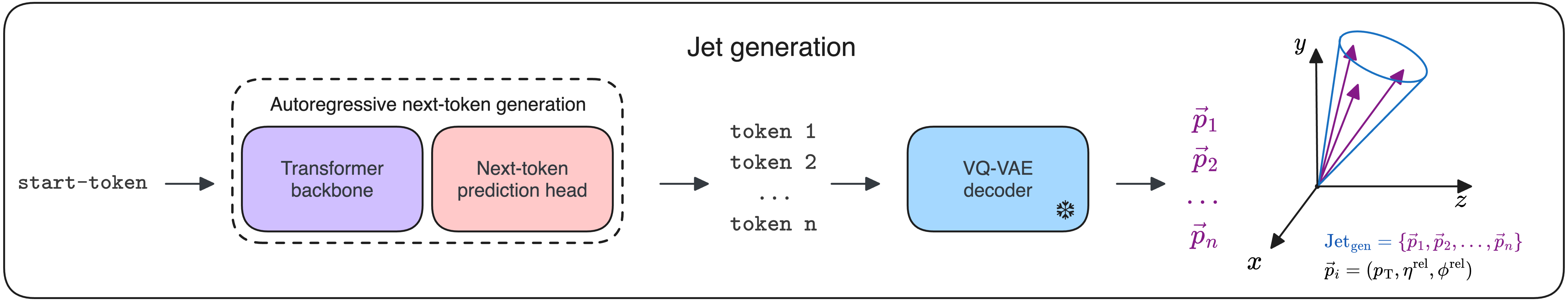
Based on Omnijet-α architecture (2403.05618)



- Tokenized jet constituents $\left(p_T, \eta^{\text{rel}}, \phi^{\text{rel}}\right)$
- GPT-style generation: Next-token prediction

Unsupervised pre-training

Based on Omnijet-α architecture (2403.05618)



- Tokenized jet constituents $\left(p_T, \eta^{\mathrm{rel}}, \phi^{\mathrm{rel}}\right)$
- GPT-style generation: Next-token prediction

Results

Does fine-tuning provide performance gain?

- **Fine-tuned:**

  Tokenizer: Trained on all AOJ jets

  Generative model: Pre-trained on all AOJ jets

- **From scratch:**

  Tokenizer: Trained on all JetClass [1] jets

  Generative model: No pre-training

[1] https://zenodo.org/records/6619768

Results

Does fine-tuning provide performance gain?

- **Fine-tuned:**

  Tokenizer: Trained on all AOJ jets

  Generative model: Pre-trained on all AOJ jets

- **From scratch:**

  Tokenizer: Trained on all JetClass [1] jets

  Generative model: No pre-training

Next: Trained to generate jets from JetClass [1] dataset

[1] https://zenodo.org/records/6619768

Results

Downstream task: Generating TOP jets from JetClass [1]

[1] https://zenodo.org/records/6619768

Results

Metrics for comparing HLF histograms

- Kullback-Leibler divergence (KLD)

$$KL(P\,||\,Q) = \sum_x p(x)\ \log\left(\frac{p(x)}{q(x)}\right)$$

- Wasserstein-1 distance

$$W_1(P,Q) = min_{\gamma\in\Pi}\sum_{x,y}|\,x-y\,|\ \gamma(x,y)$$

Results

Metrics for comparing HLF histograms

- Kullback-Leibler divergence (KLD)

$$KL(P||Q) = \sum_{x} p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

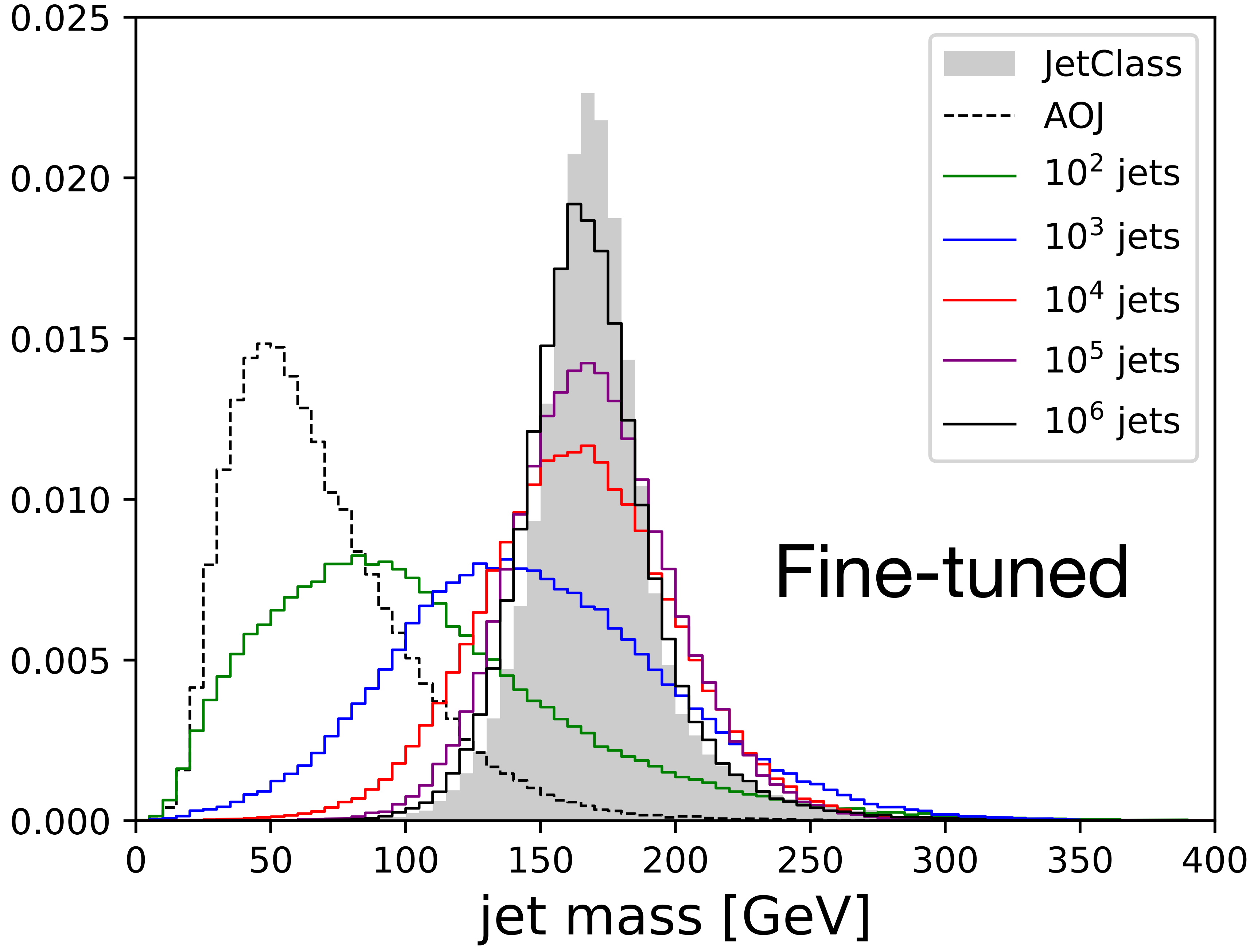
- Wasserstein-1 distance

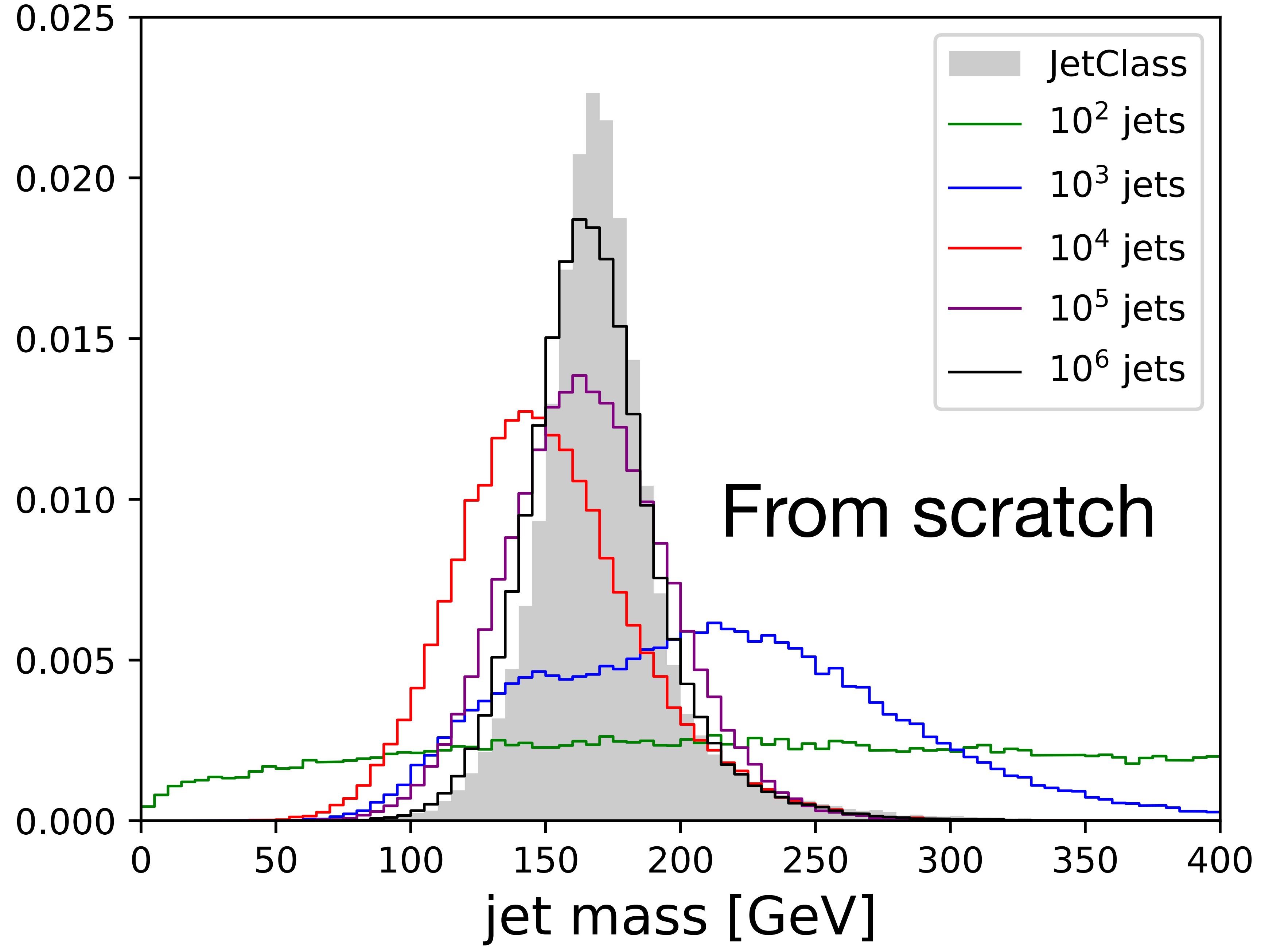
$$W_1(P,Q) = min_{\gamma \in \Pi} \sum_{x,y} |x - y| \, \gamma(x,y)$$

Two different metrics for computing distance between histograms

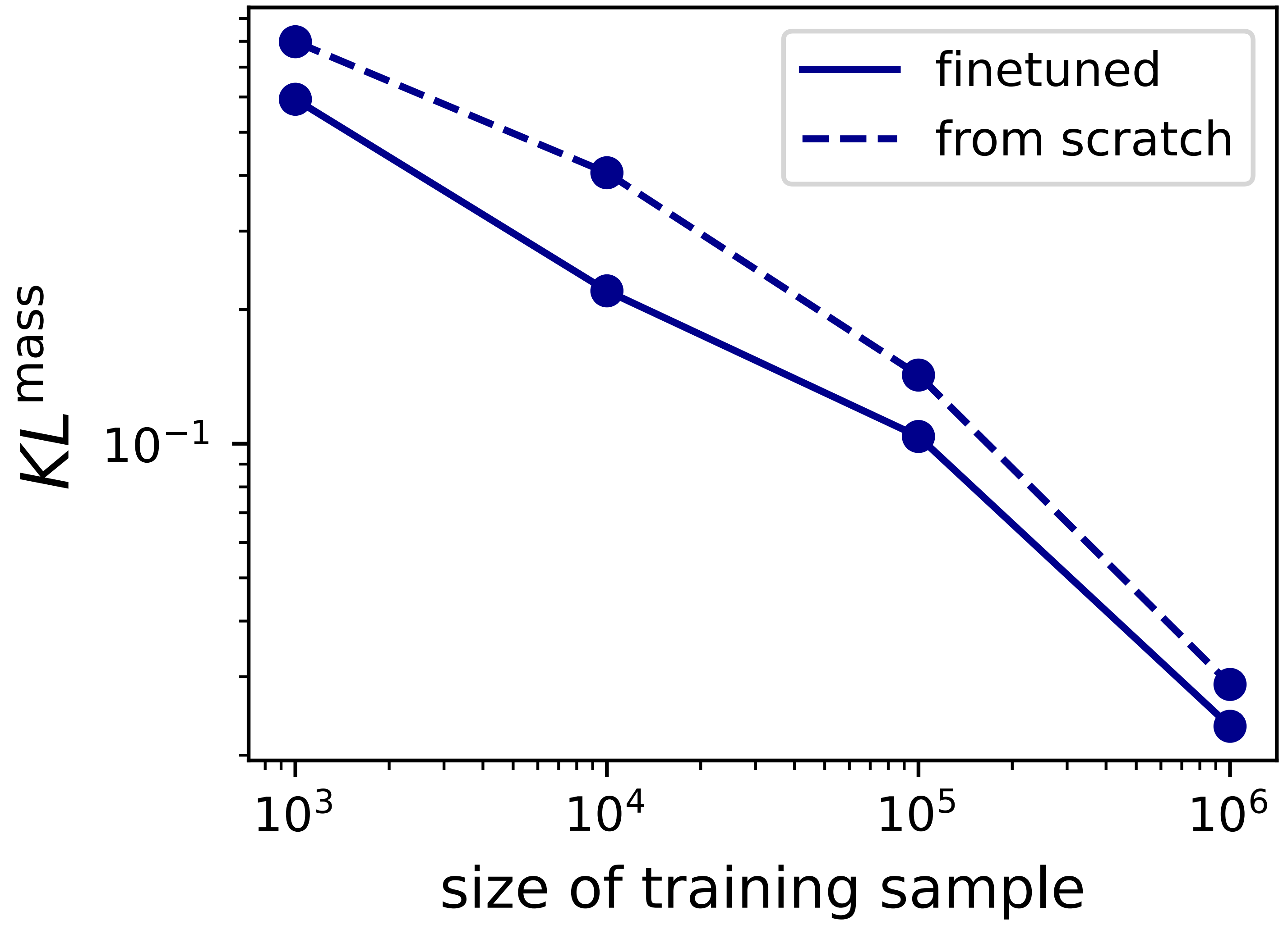
Results

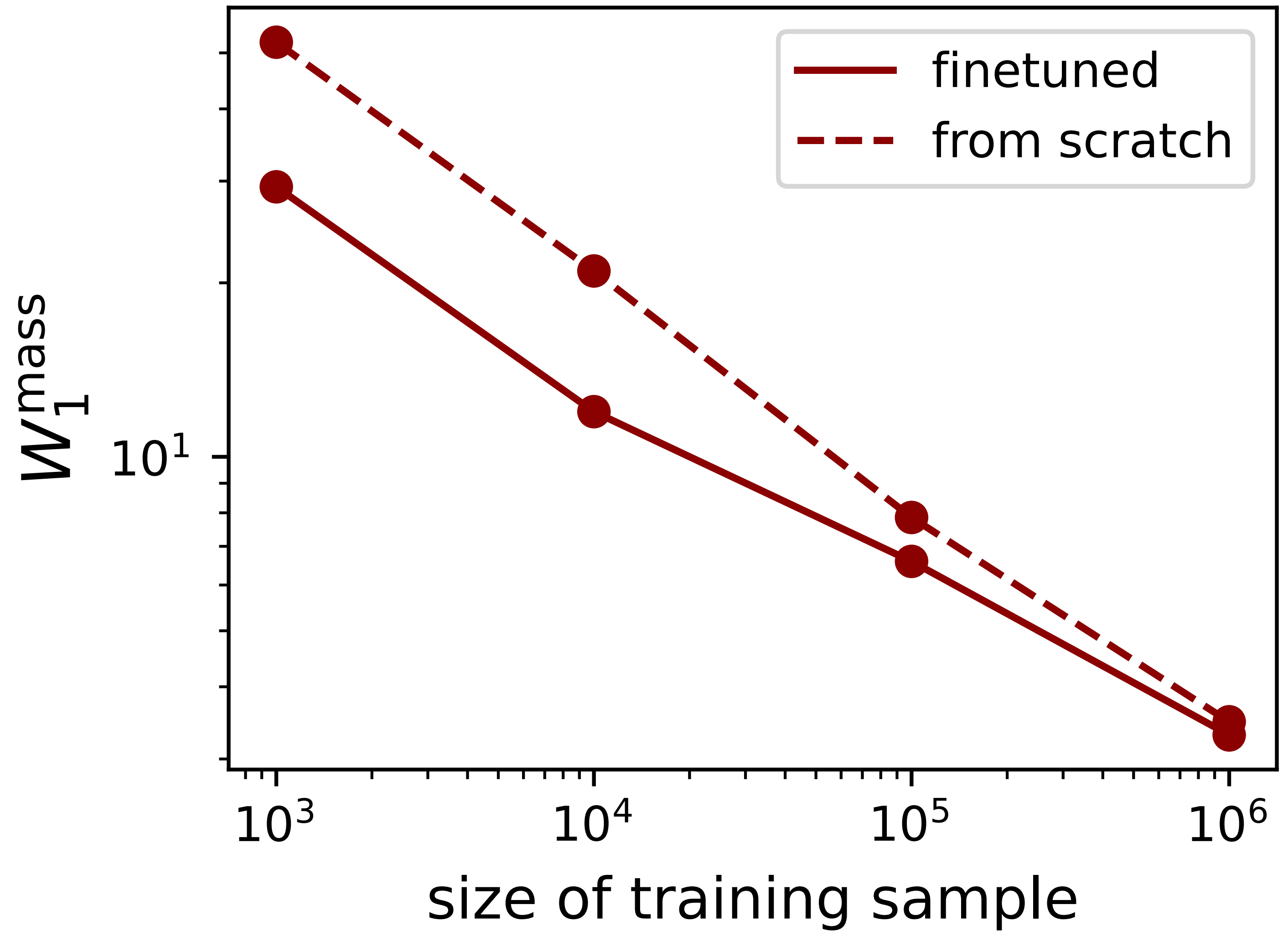
Jet kinematics





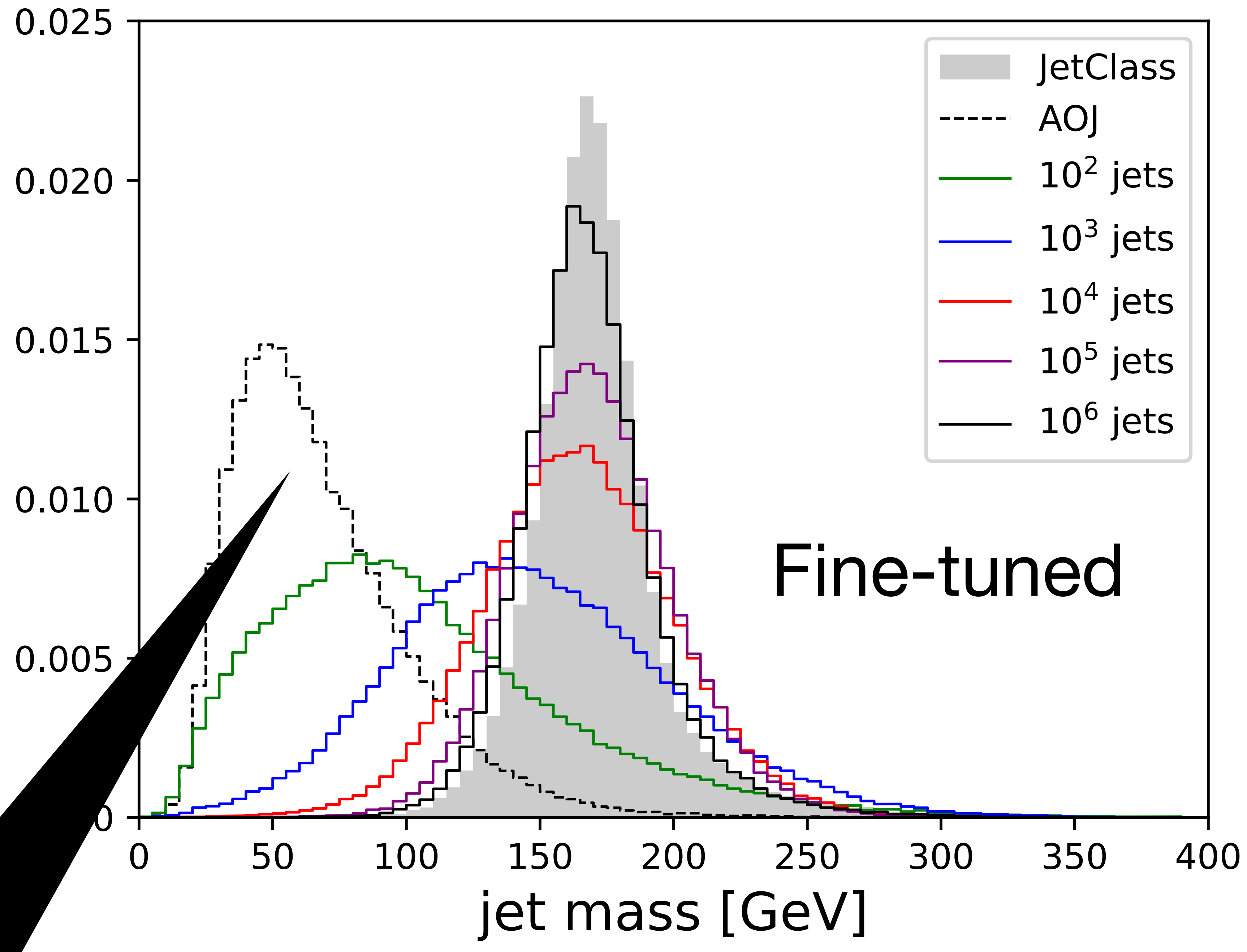
Better

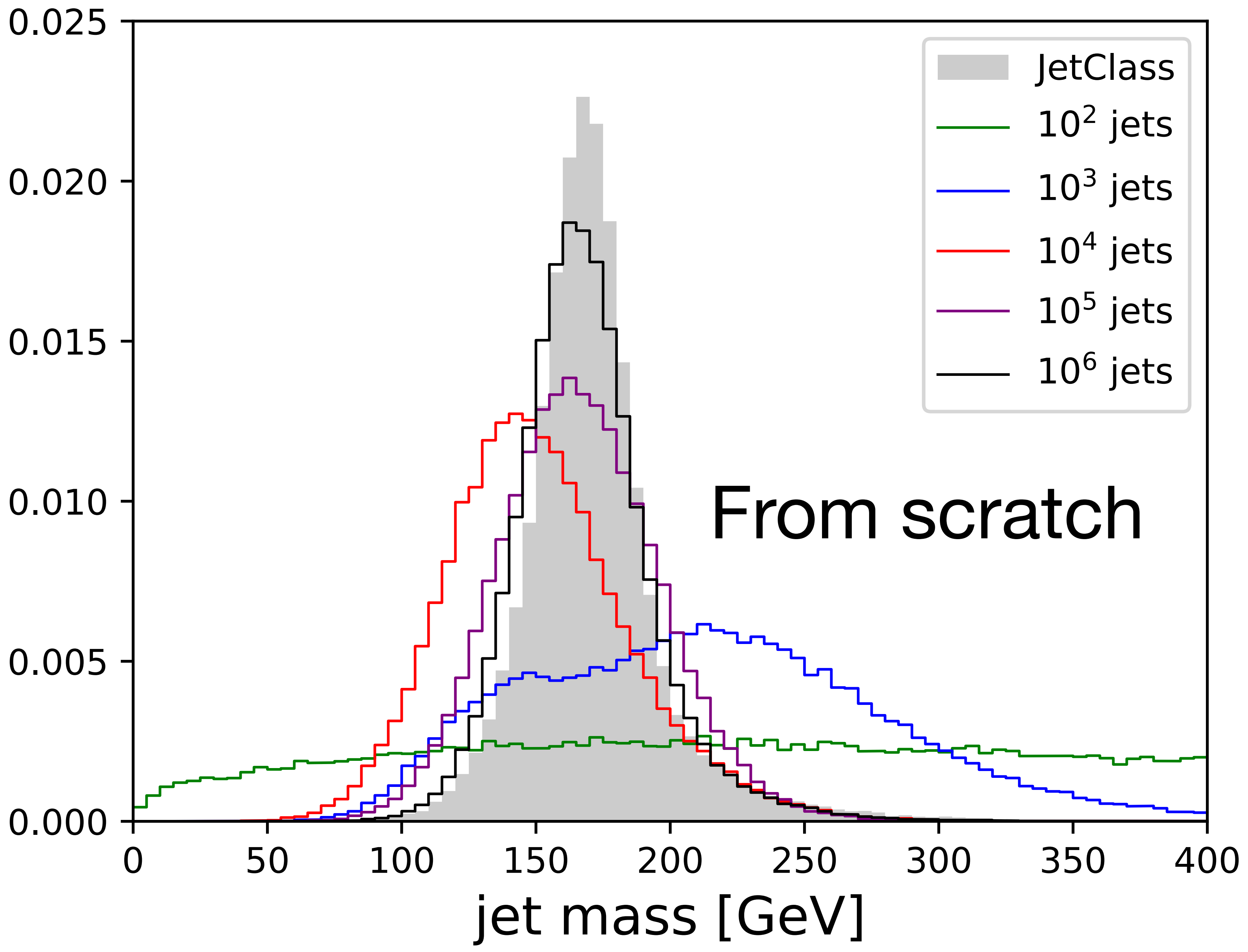




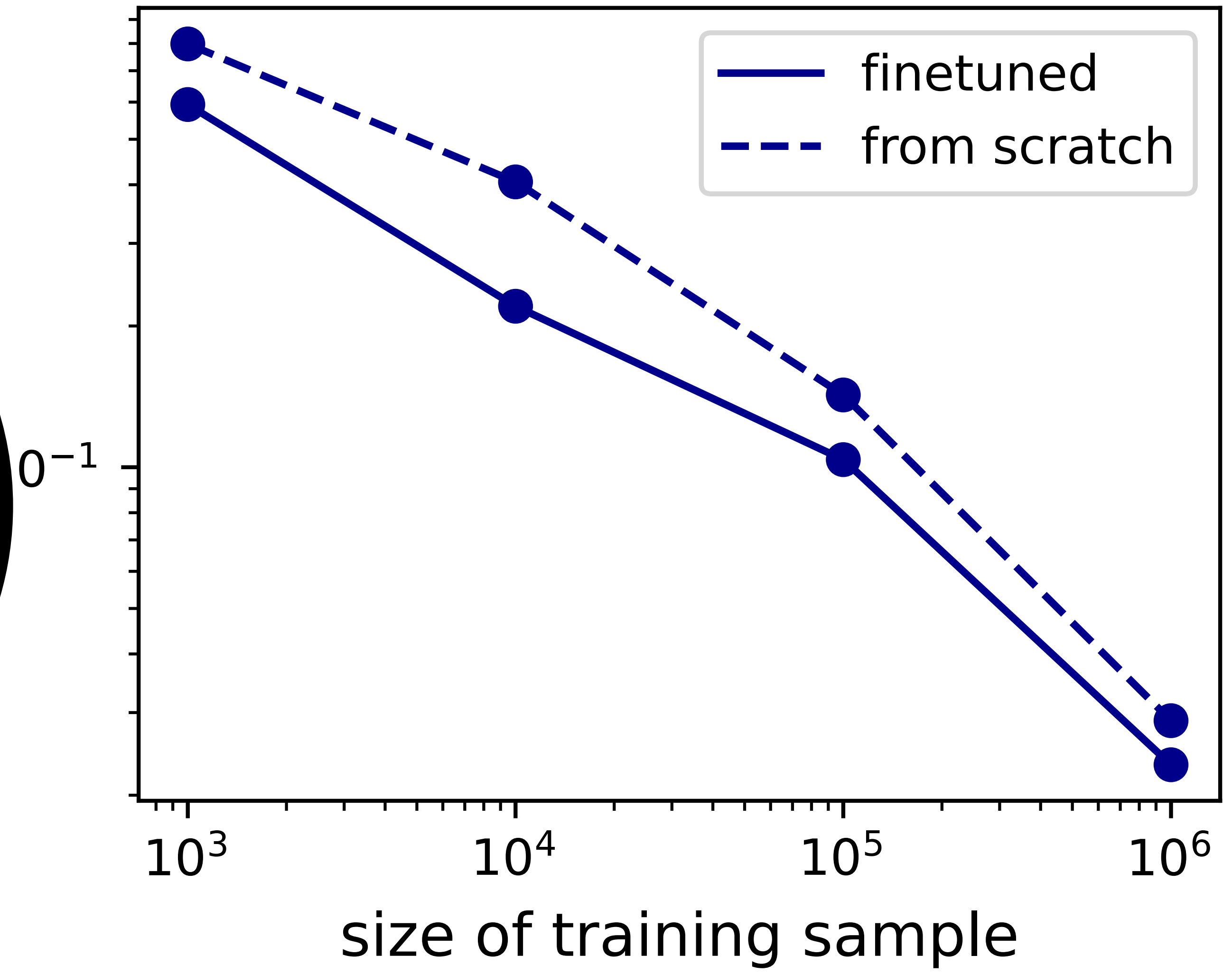
Results

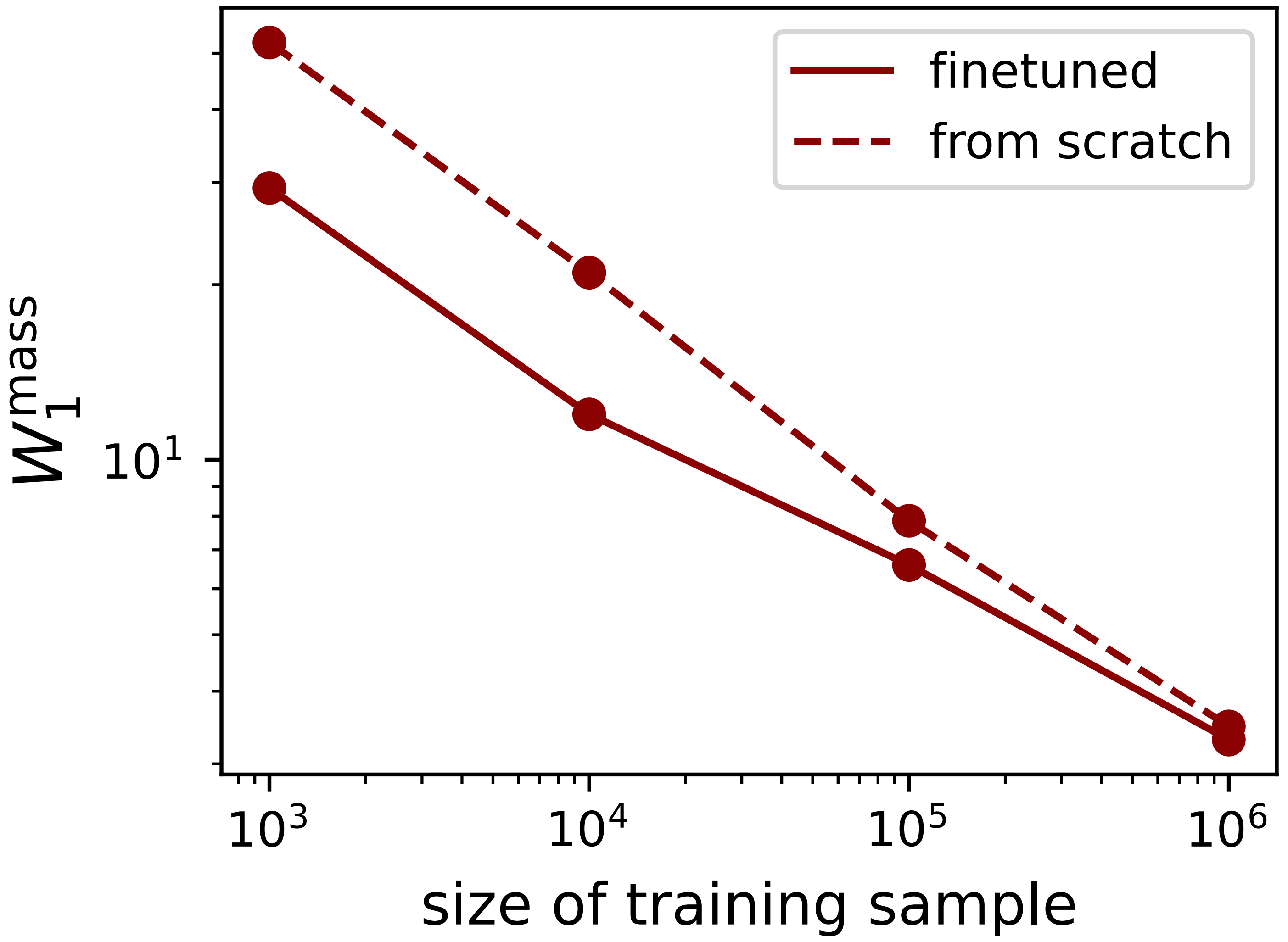
Jet kinematics





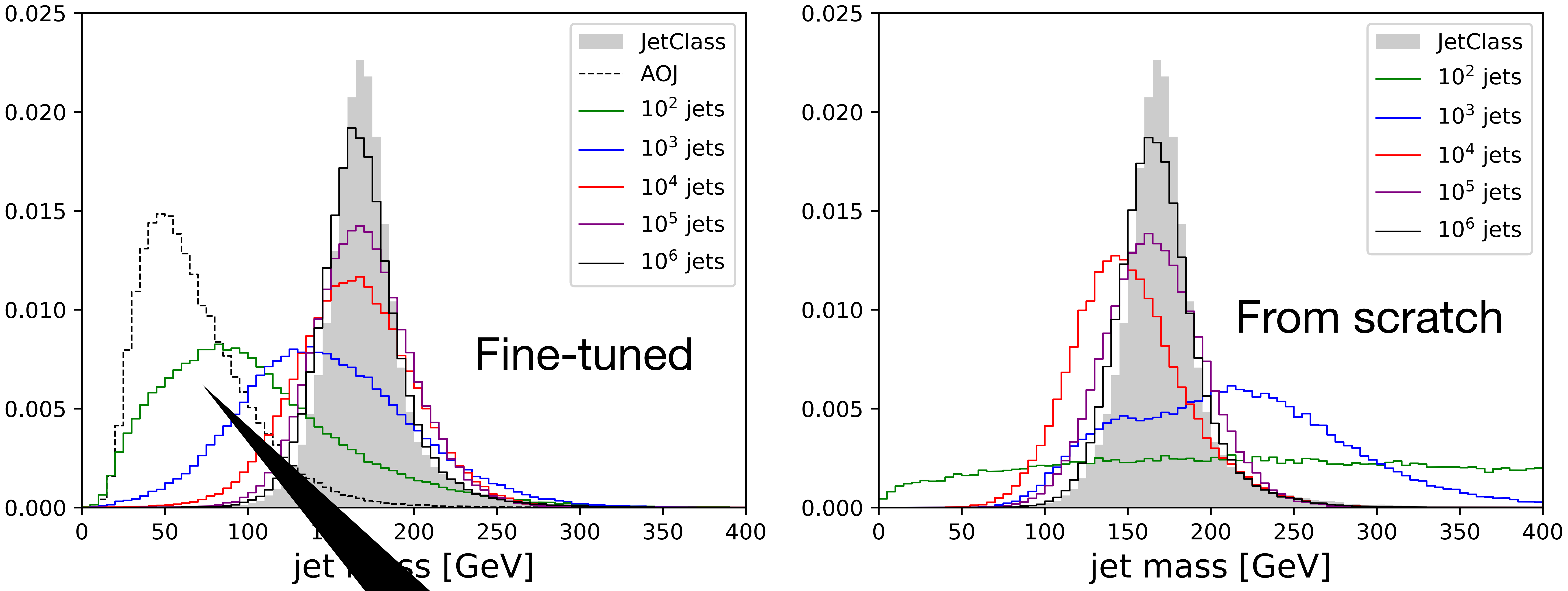
AOJ distribution shown with dotted line



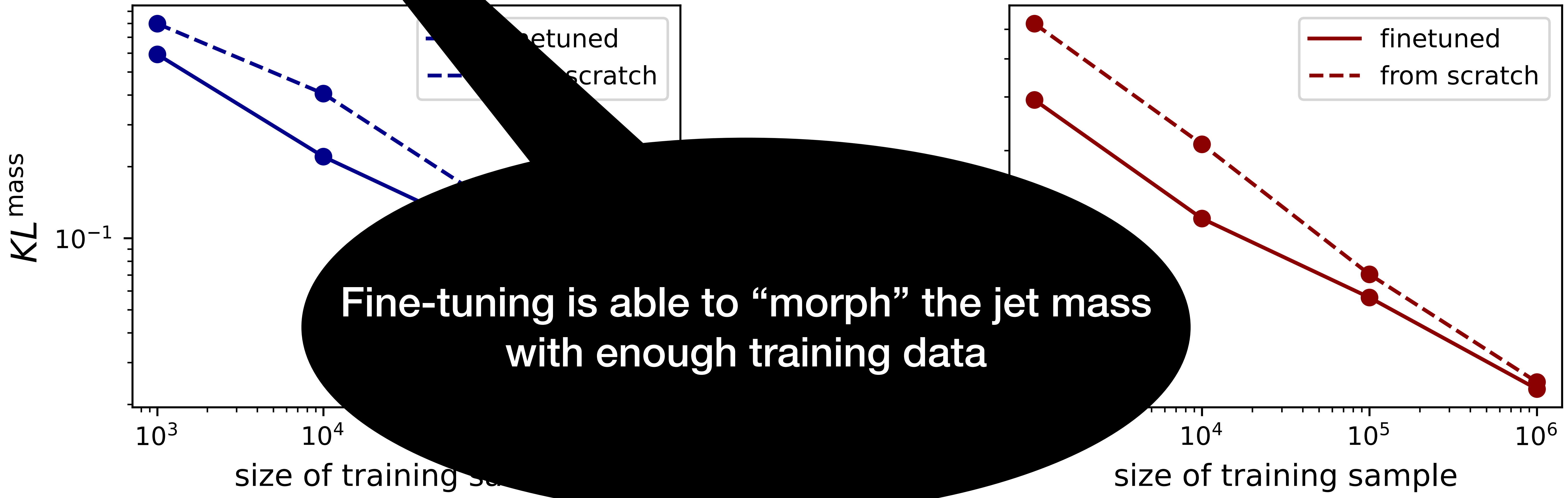


Results

Jet kinematics



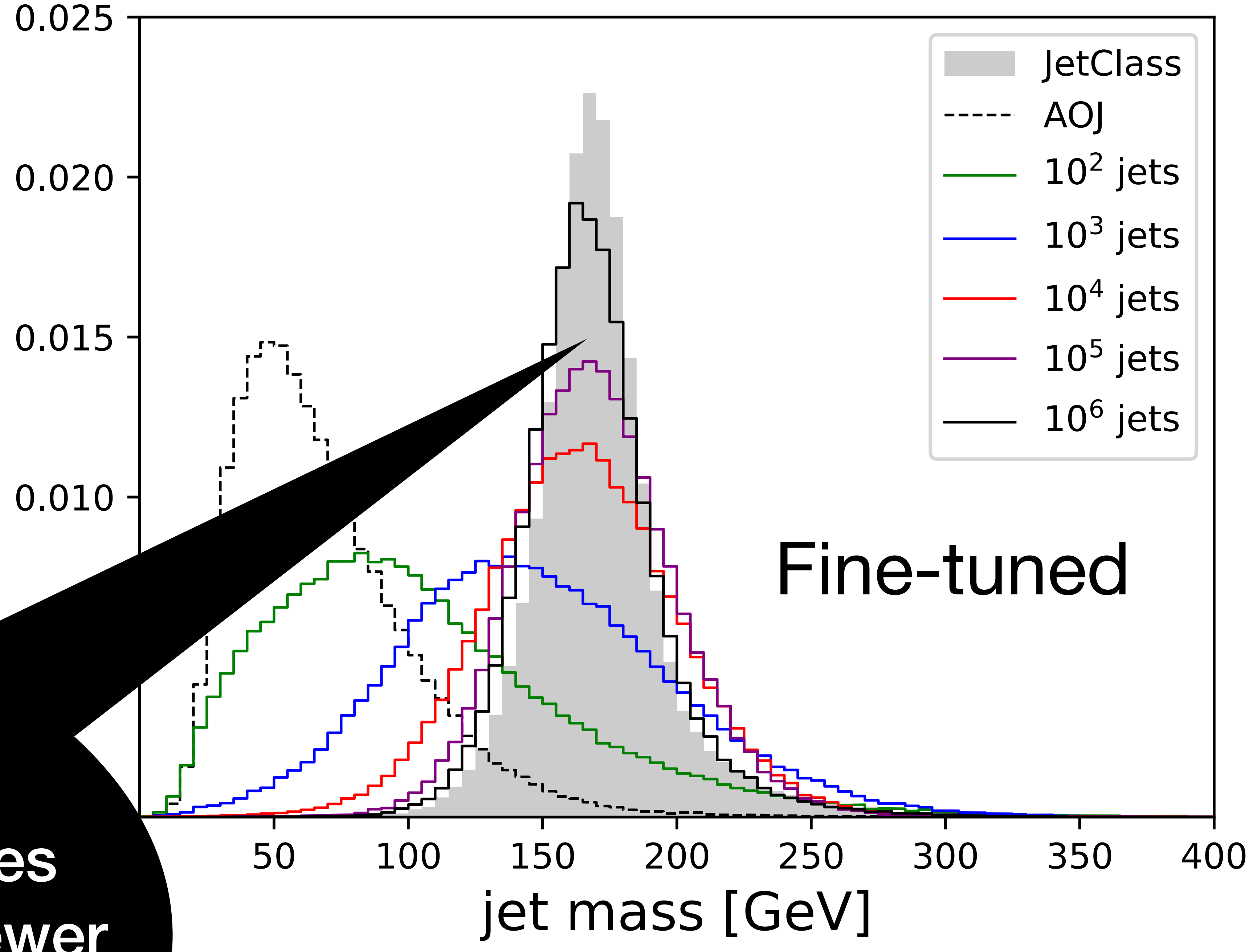
Better

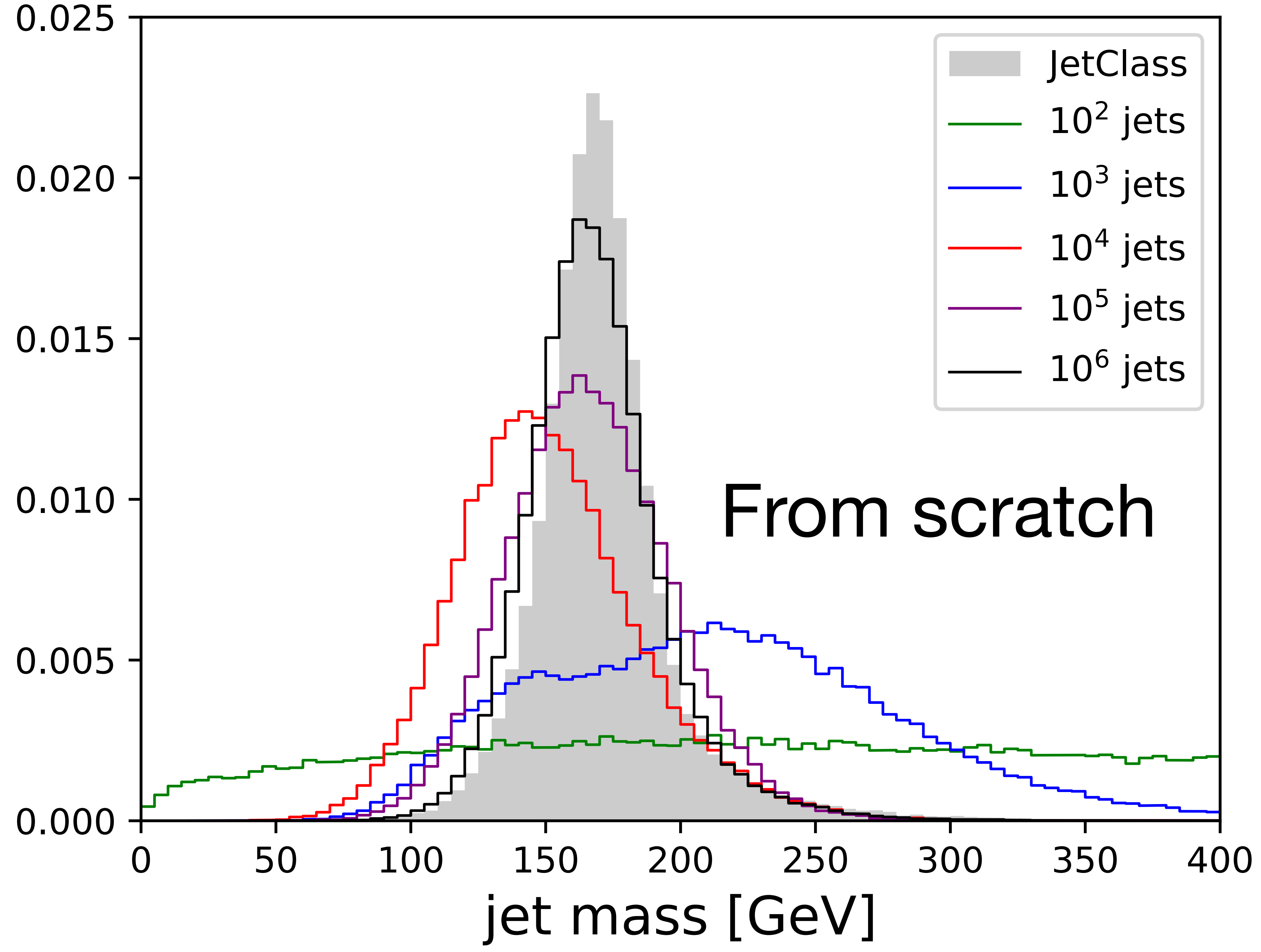


Fine-tuning is able to “morph” the jet mass with enough training data

Results

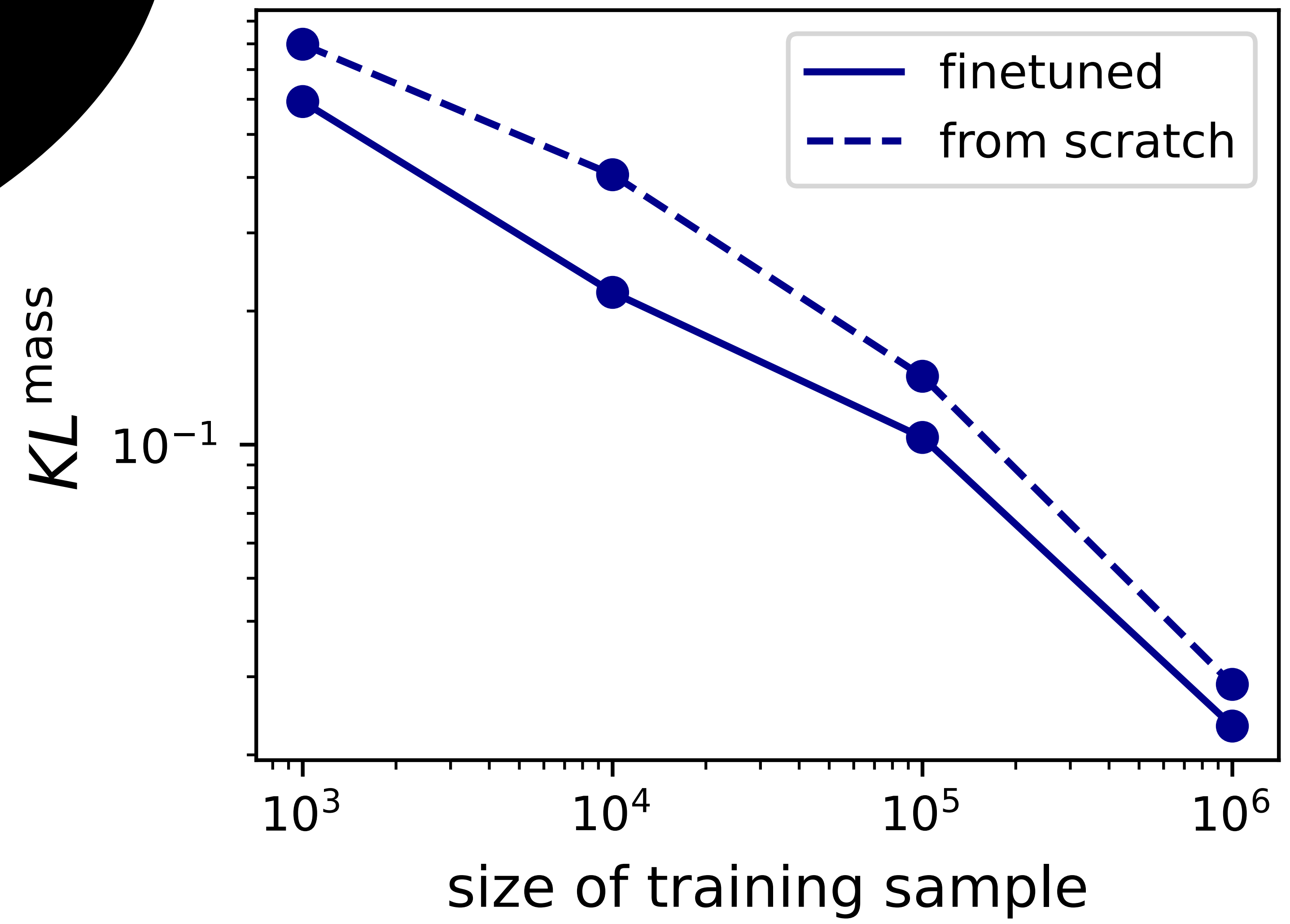
Jet kinematics

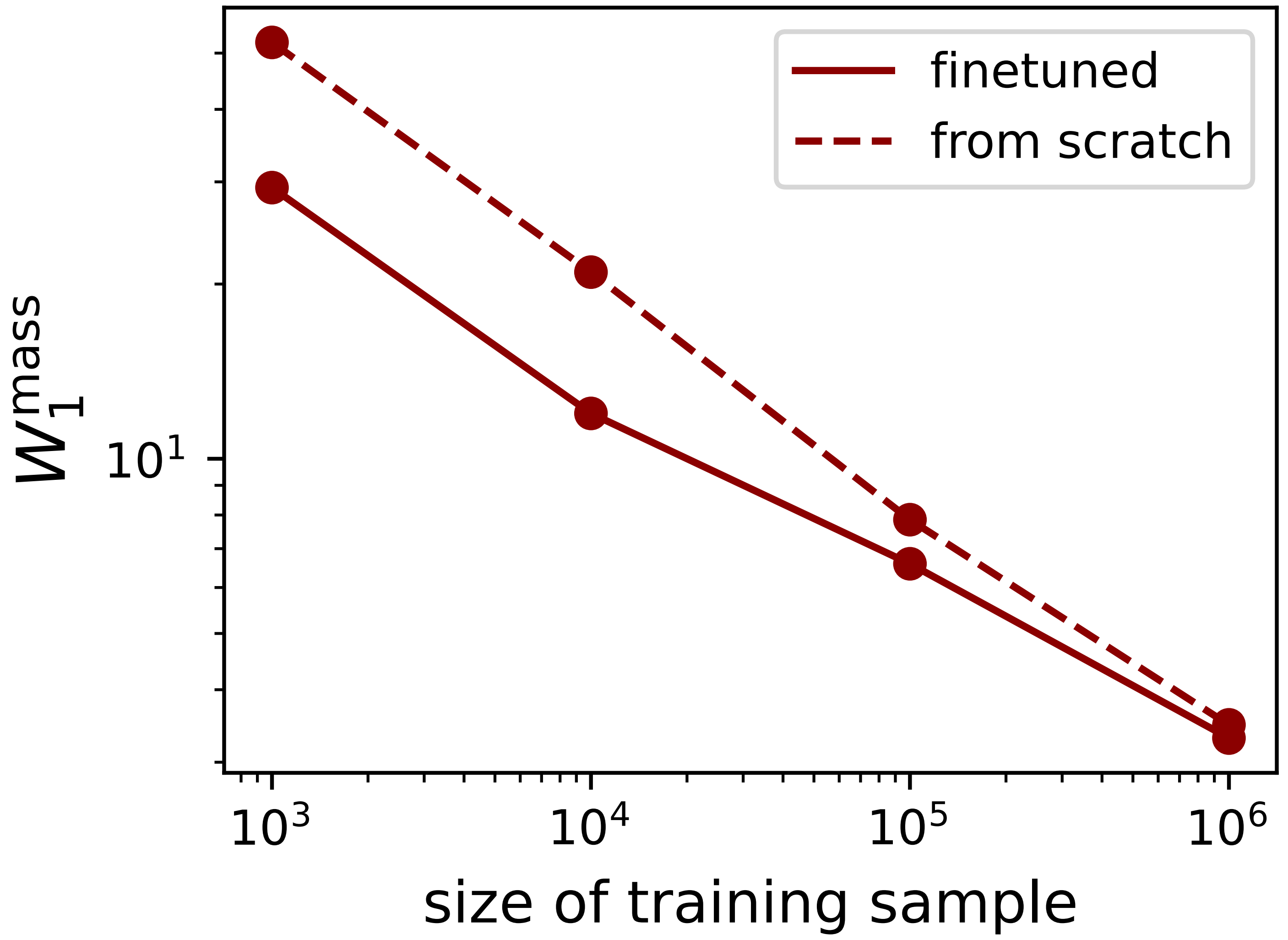




Fine-tuning generally achieves better generation quality for fewer number of training samples

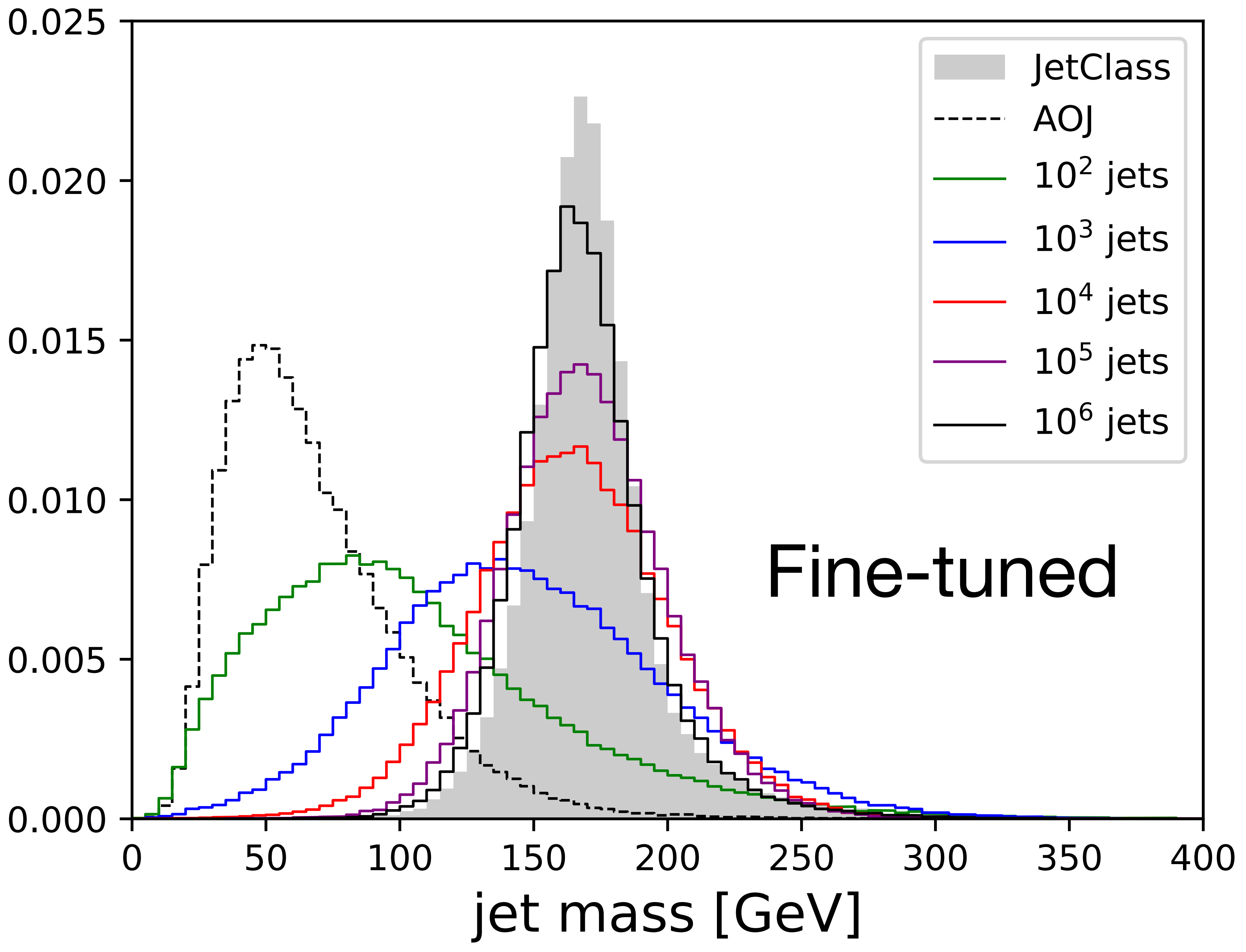
Better

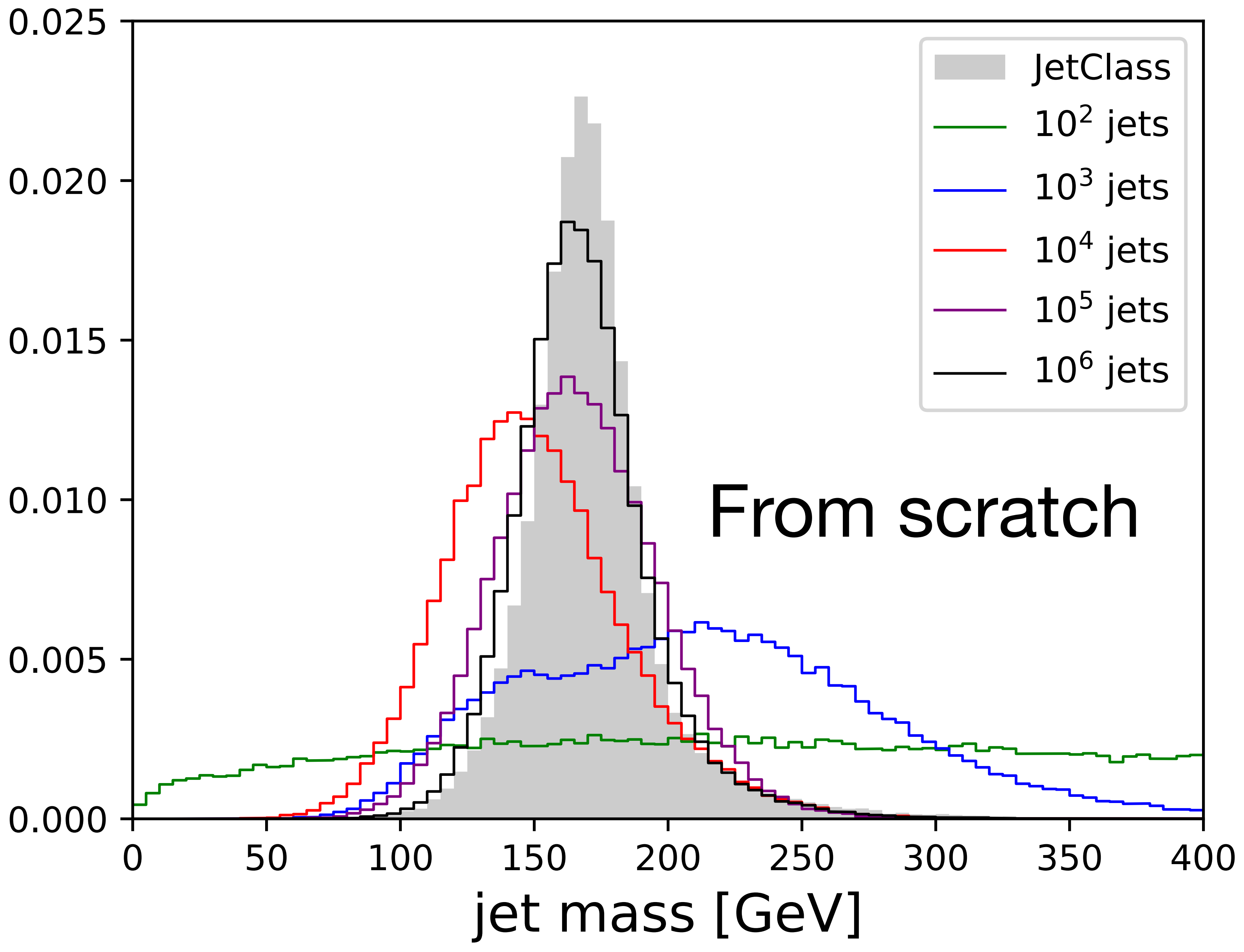




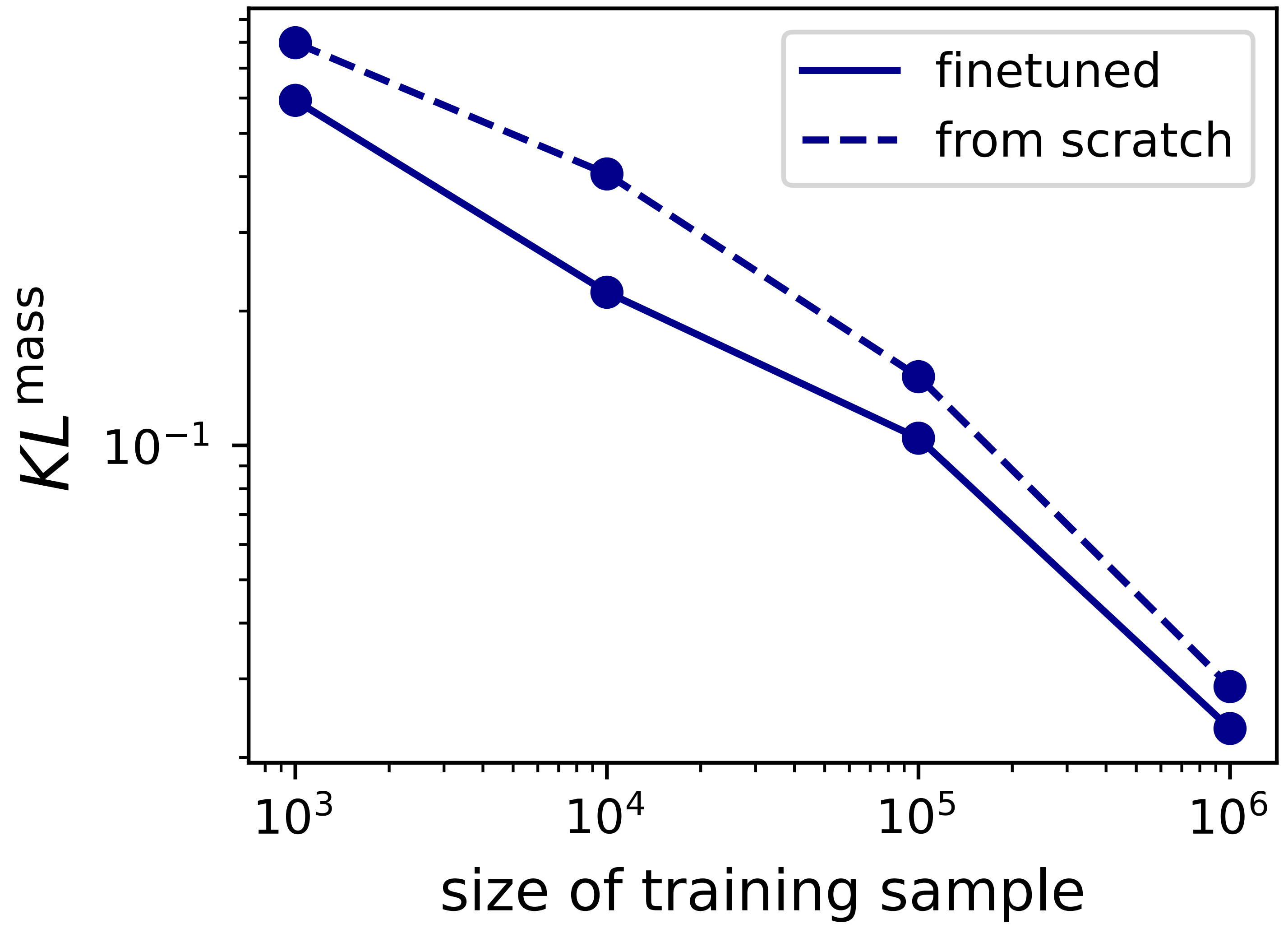
Results

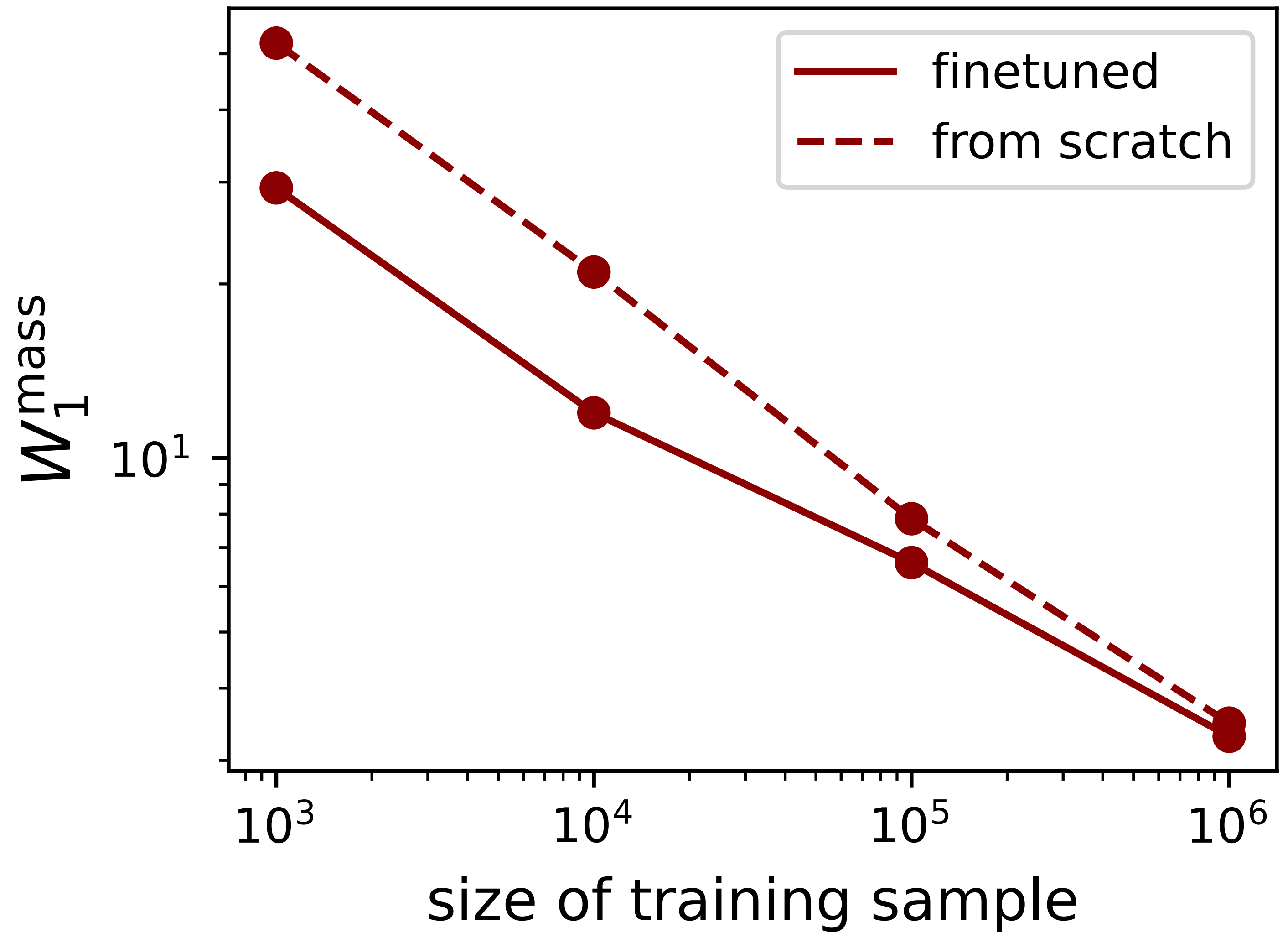
Jet kinematics





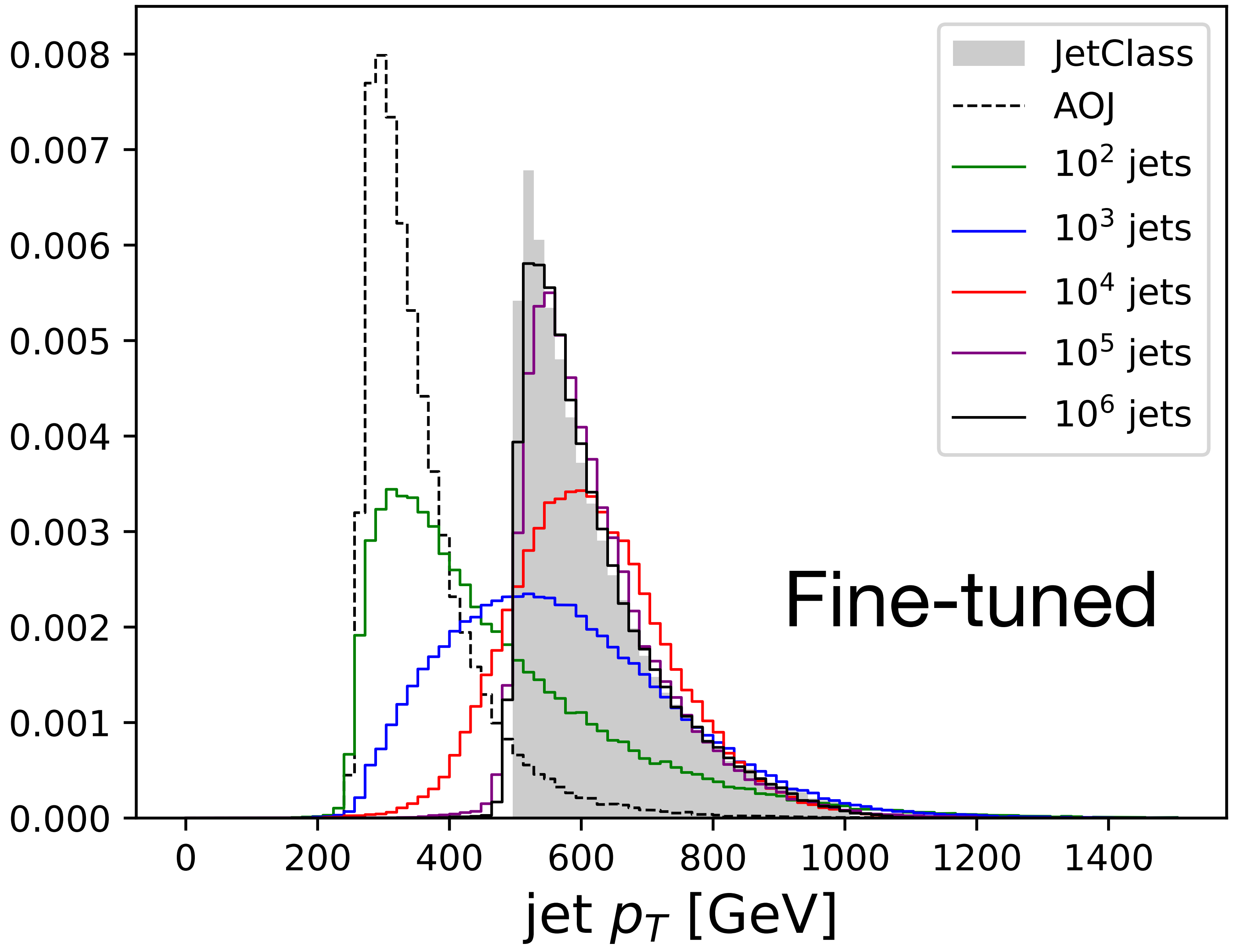
Better

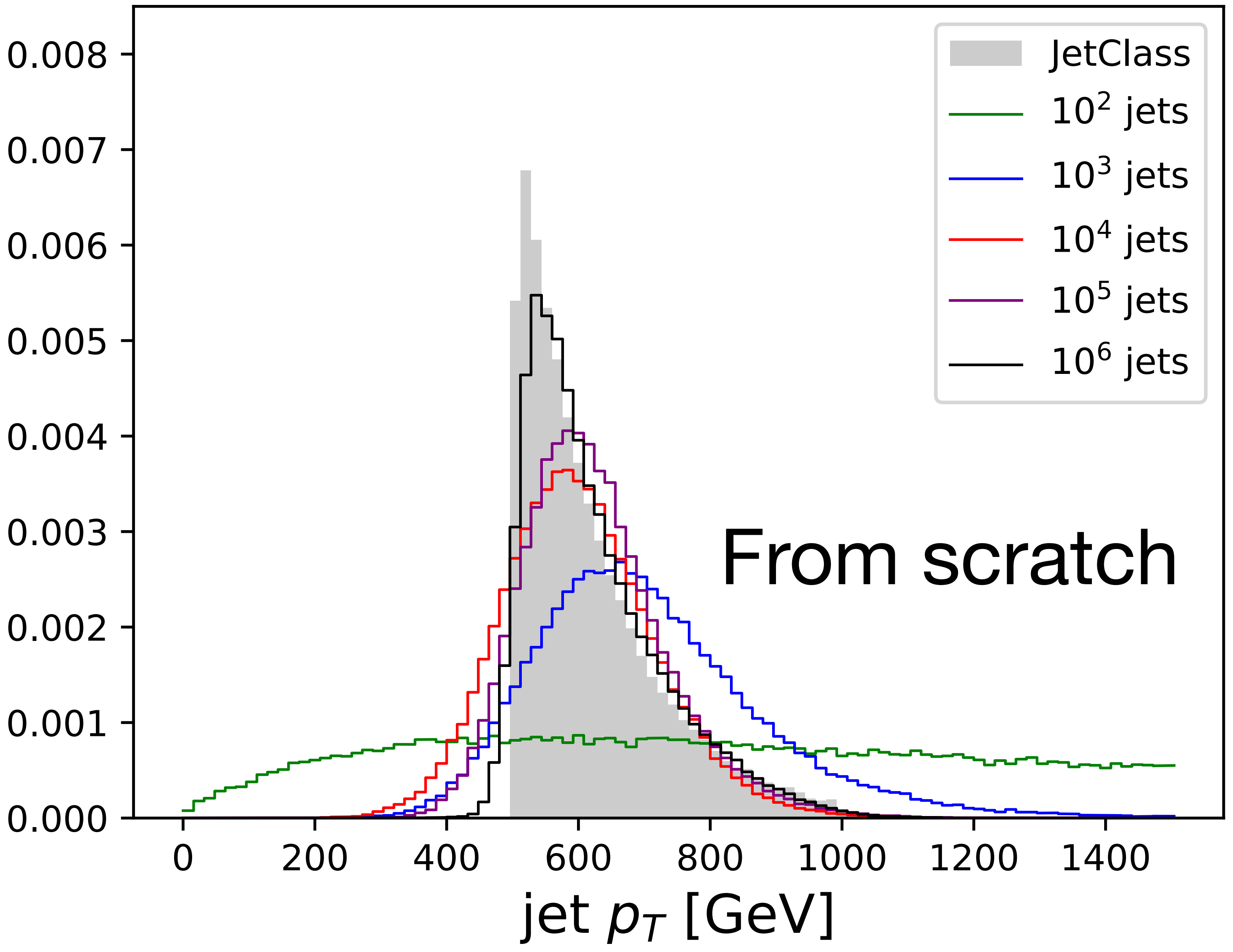




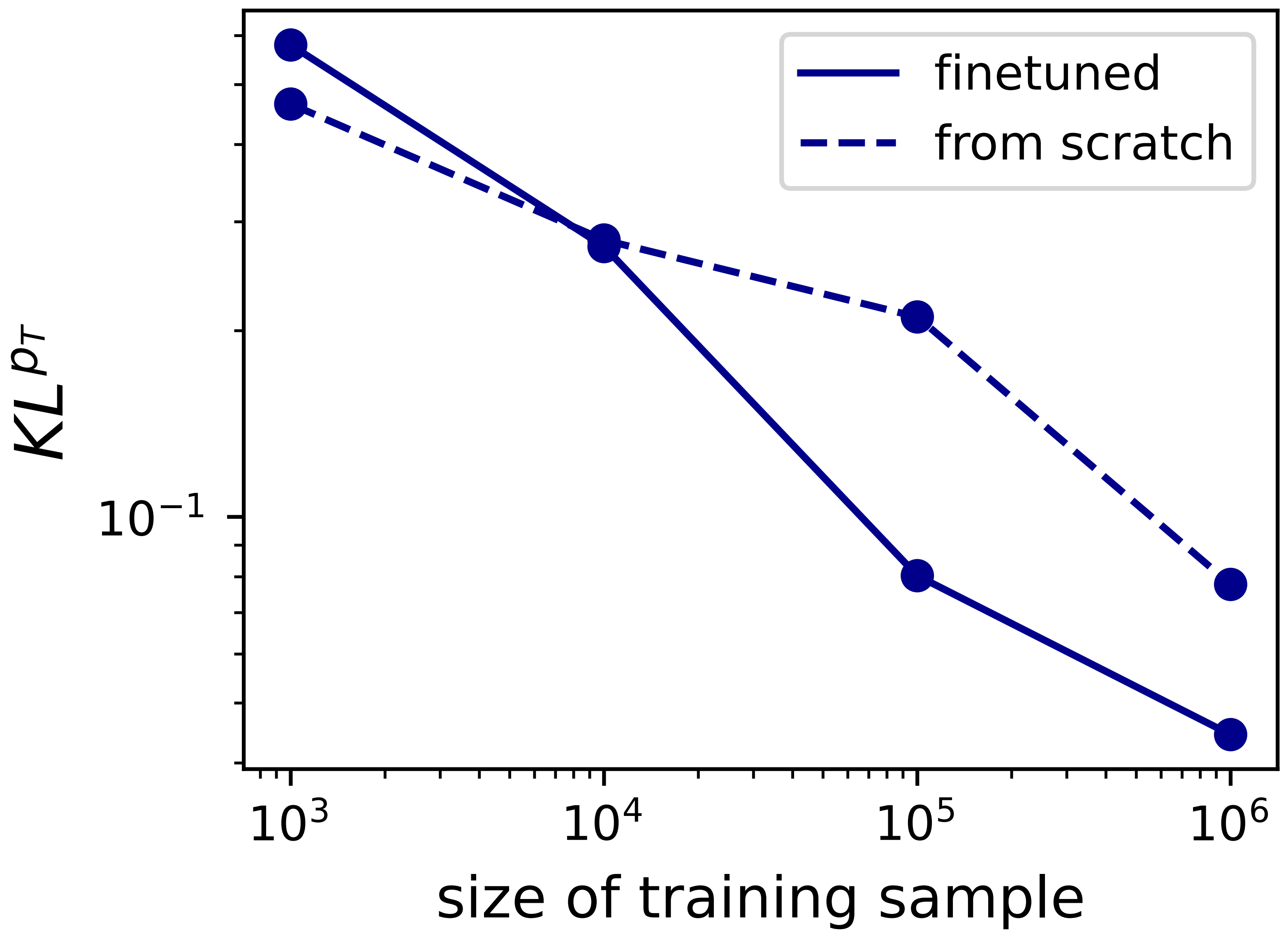
Results

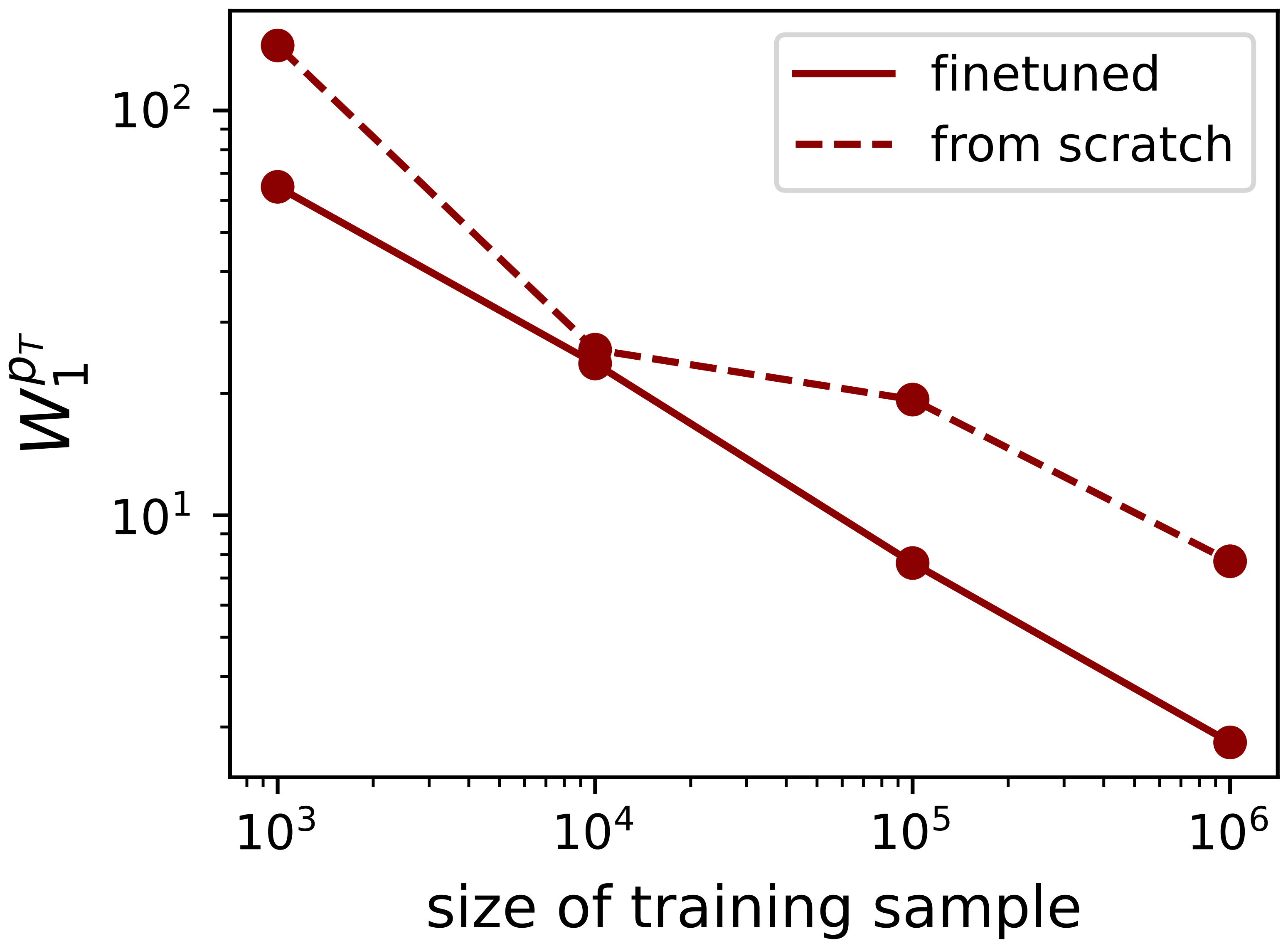
Jet kinematics





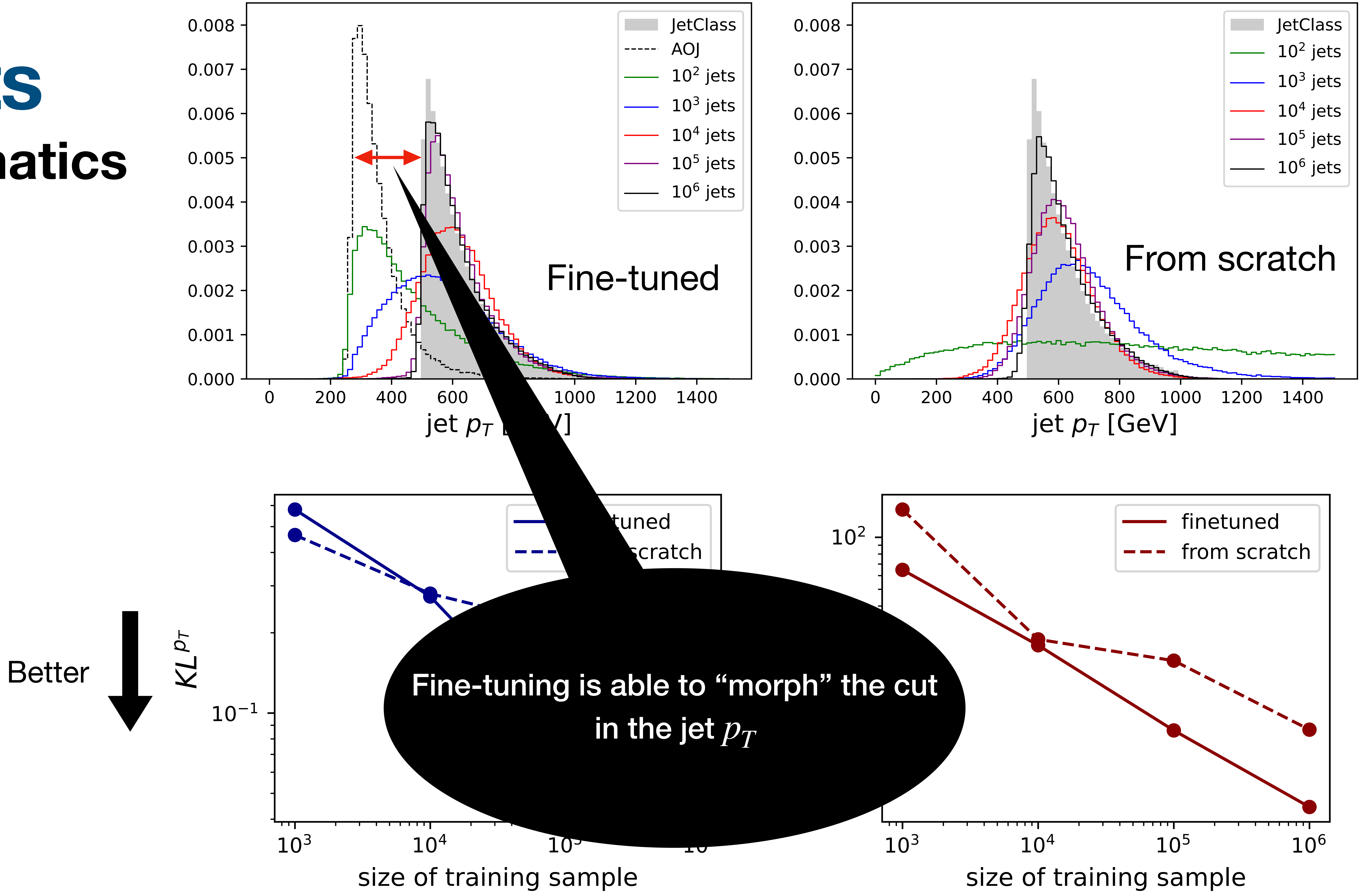
Better





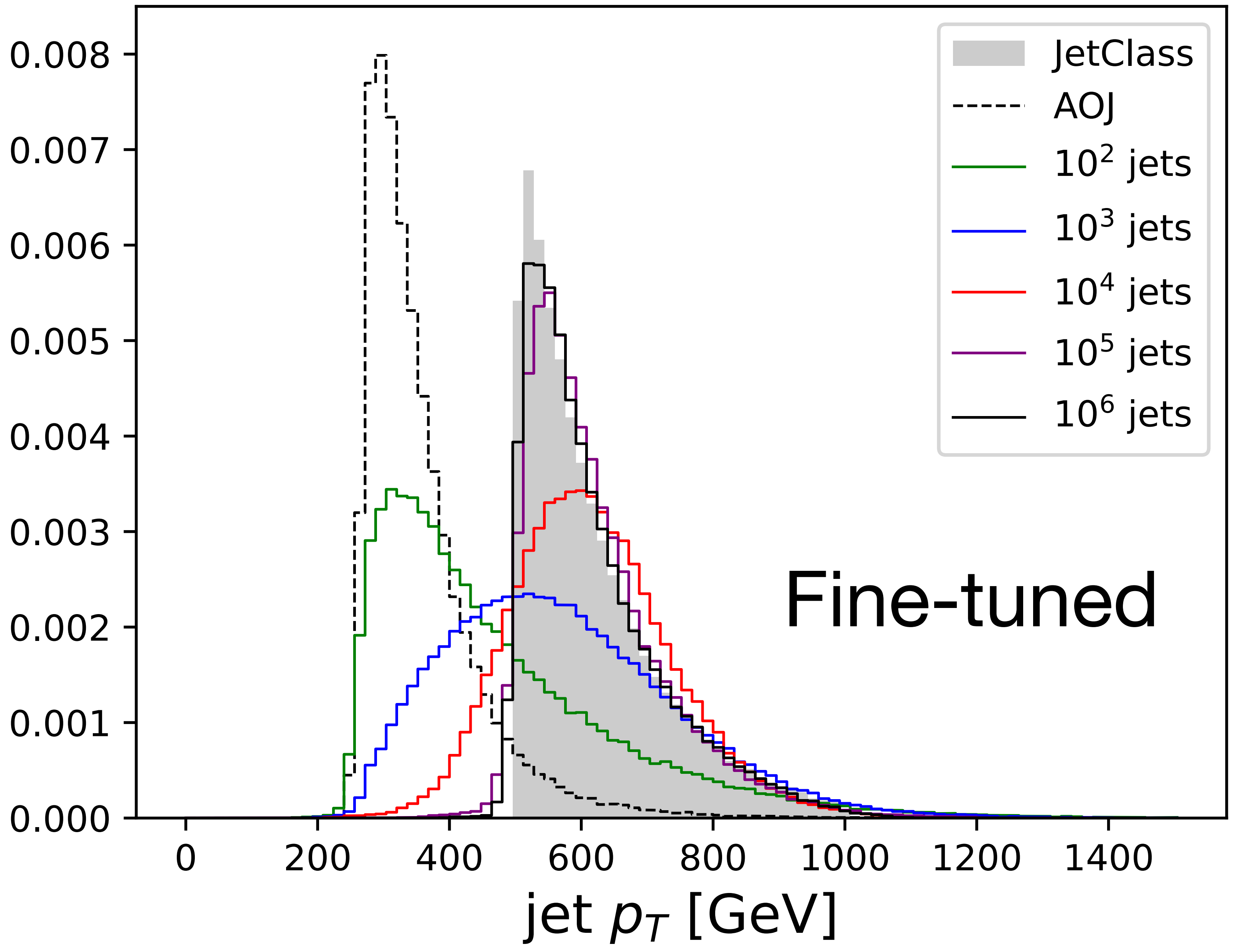
Results

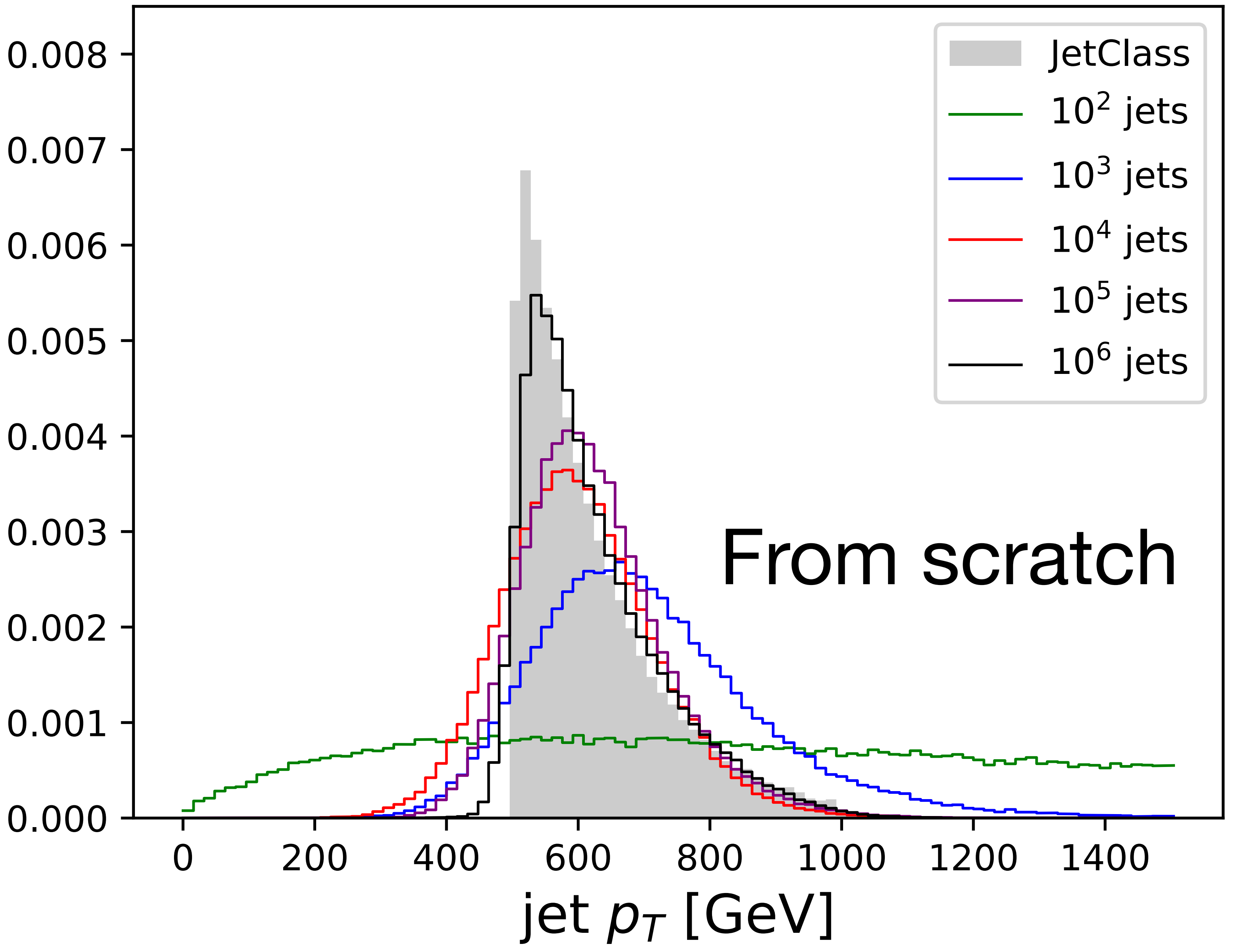
Jet kinematics



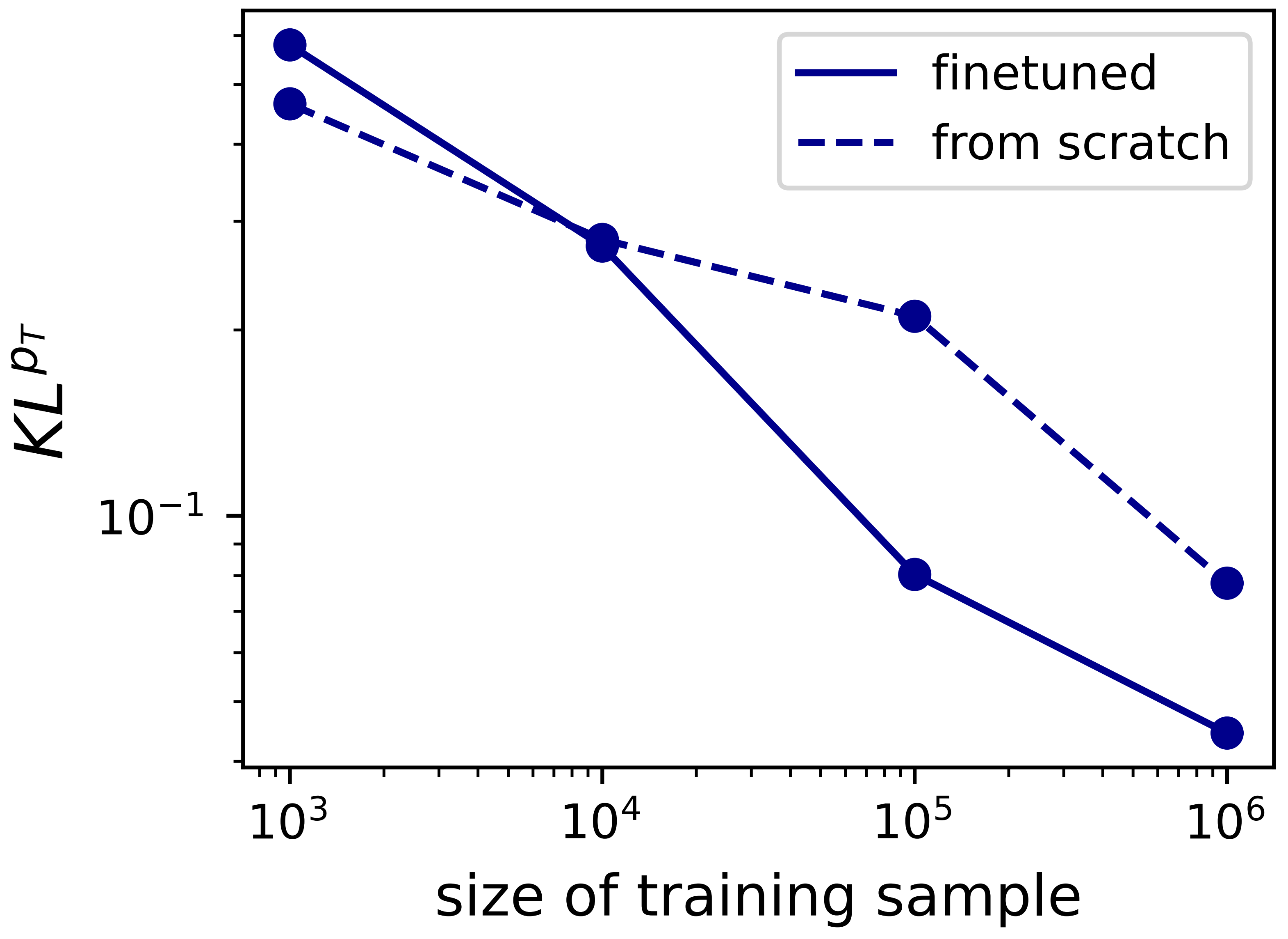
Results

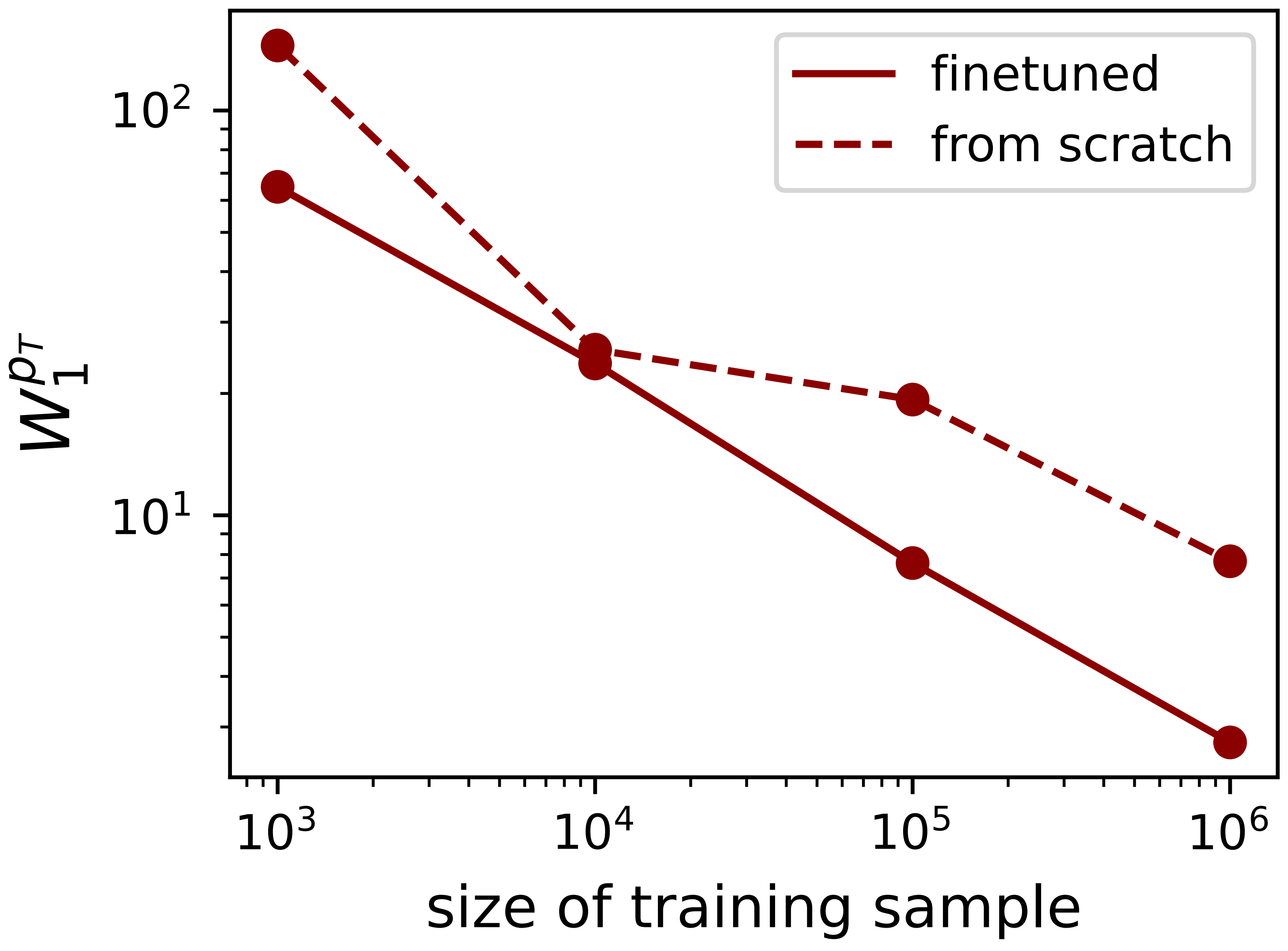
Jet kinematics





Better

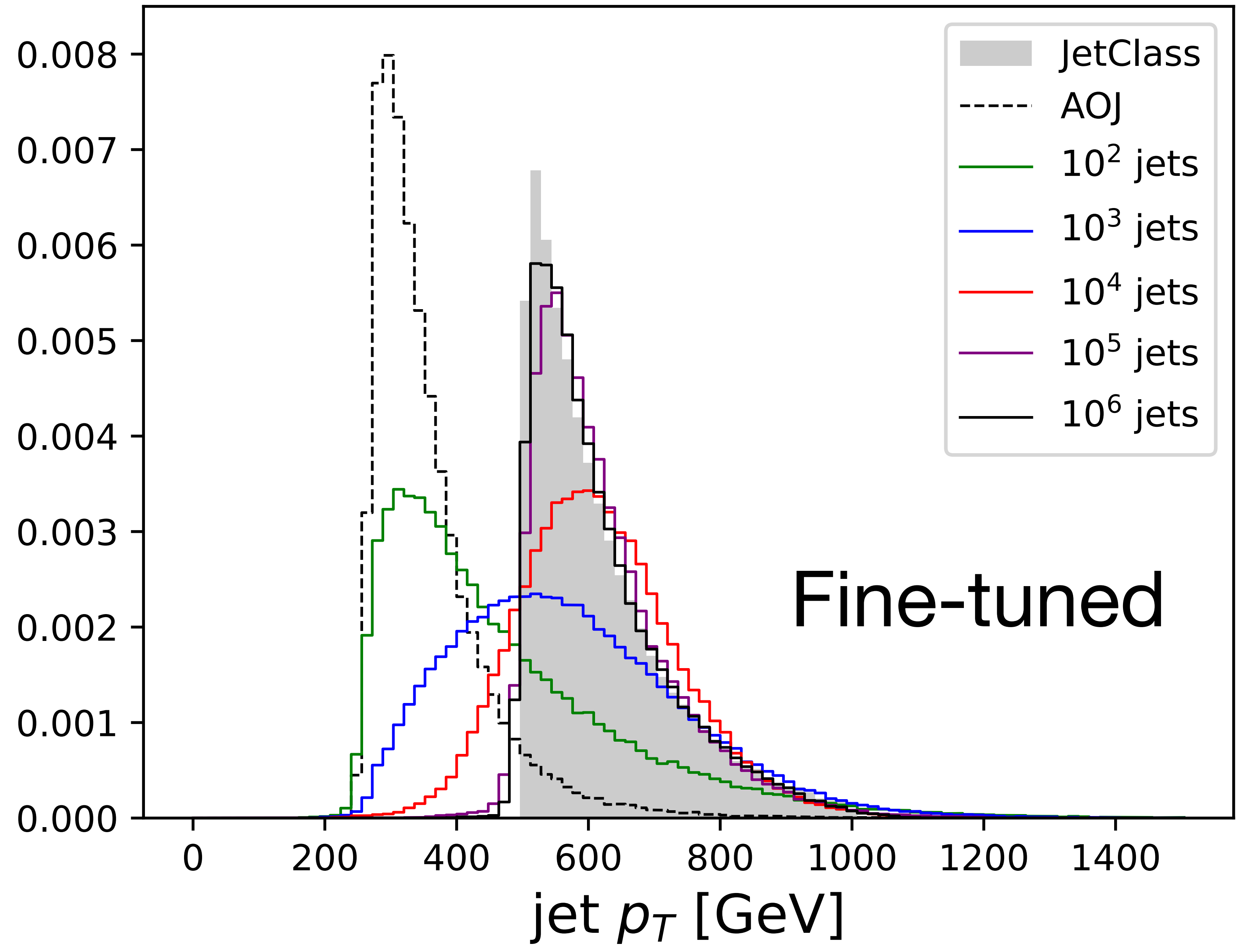


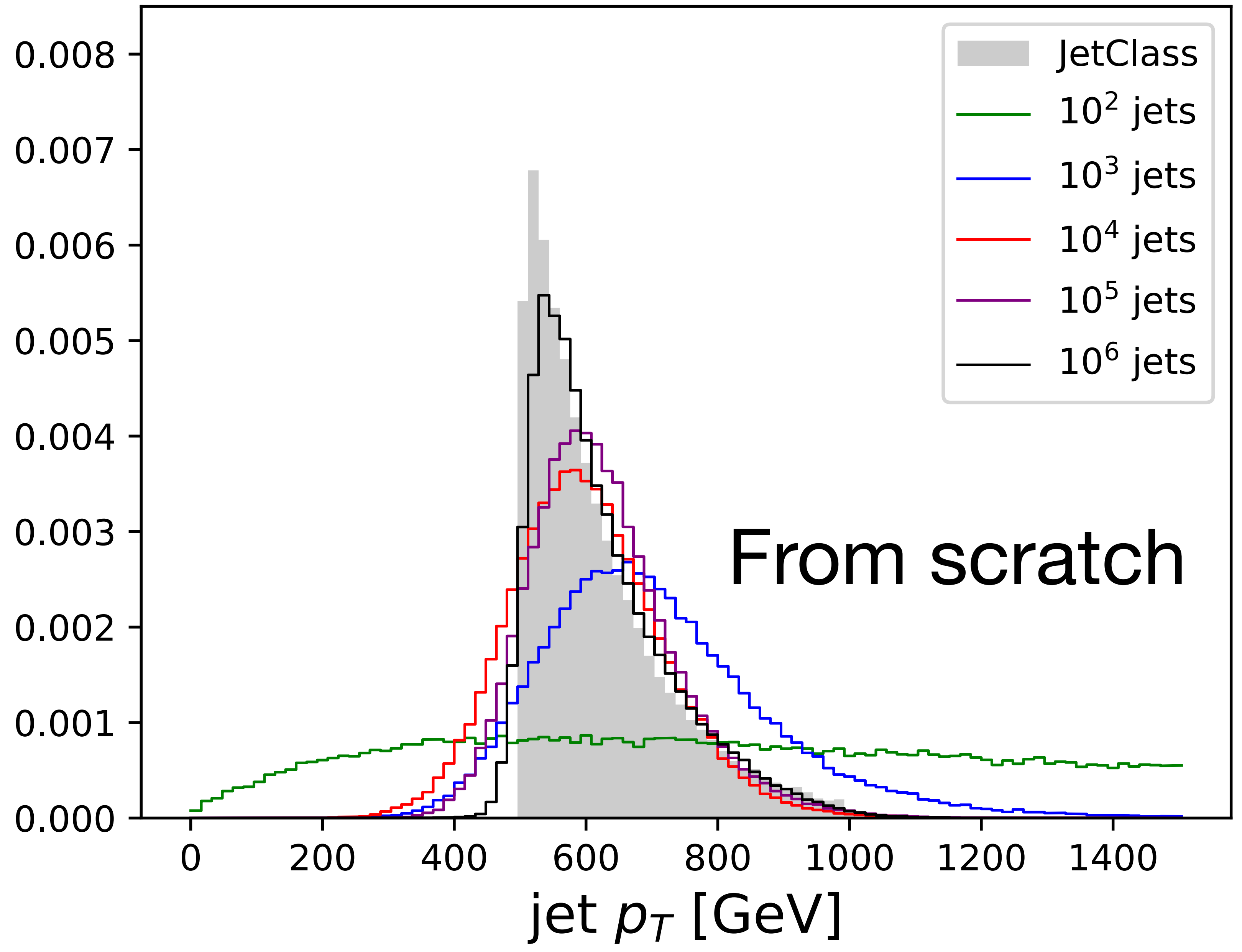


Results

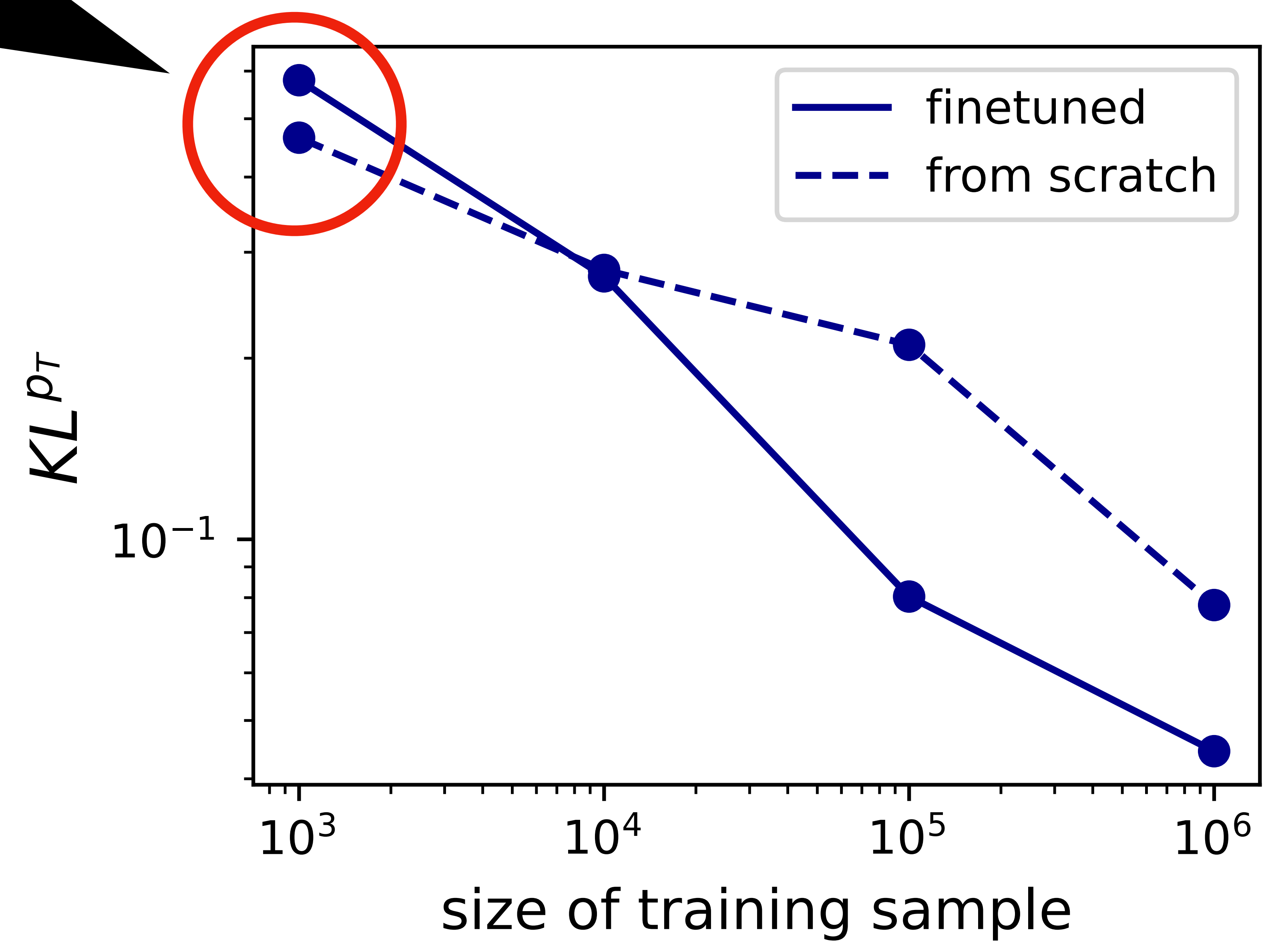
Jet kinematics

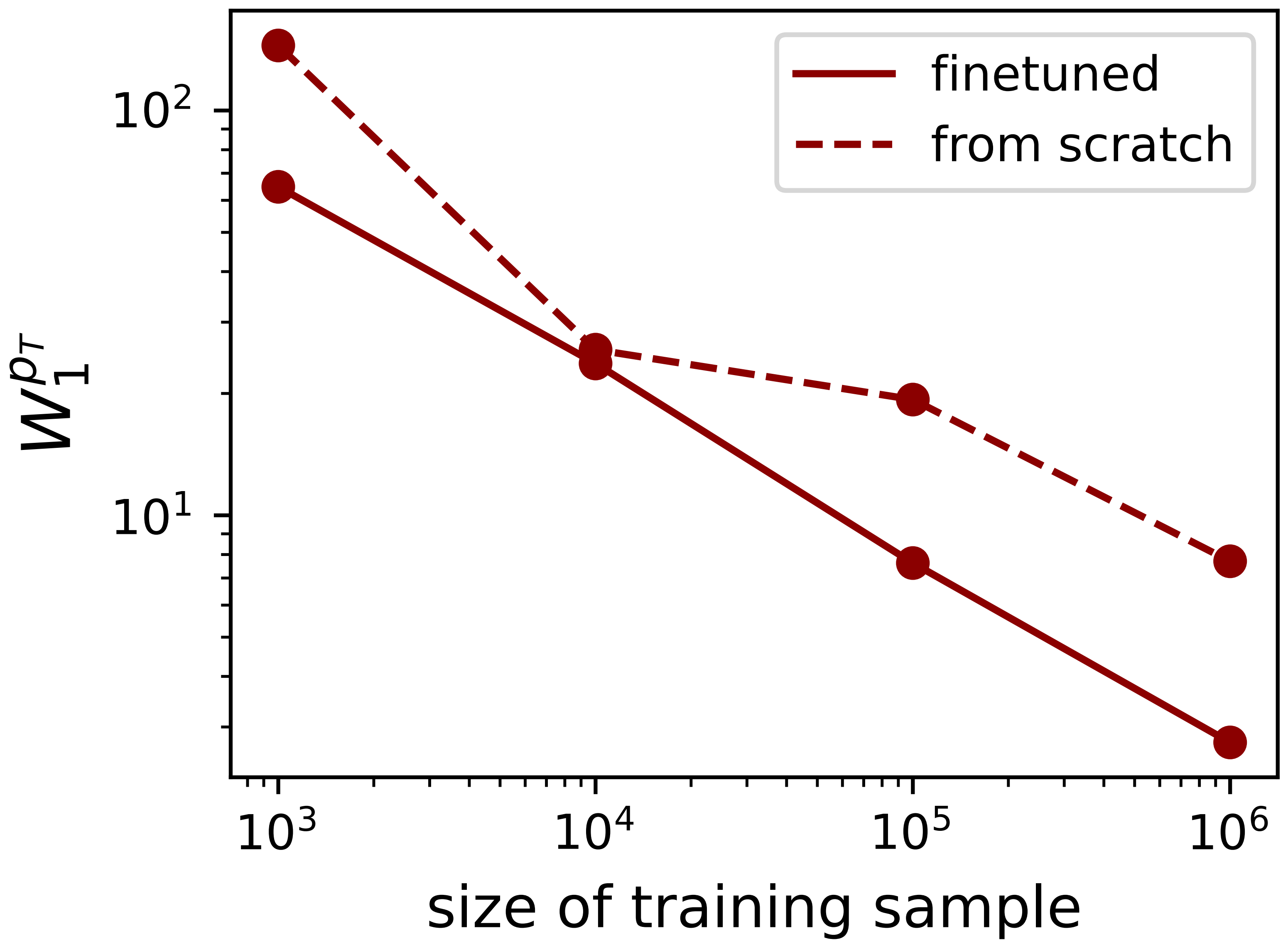
Sometimes “from scratch” does better based on one metric when trained on few jets





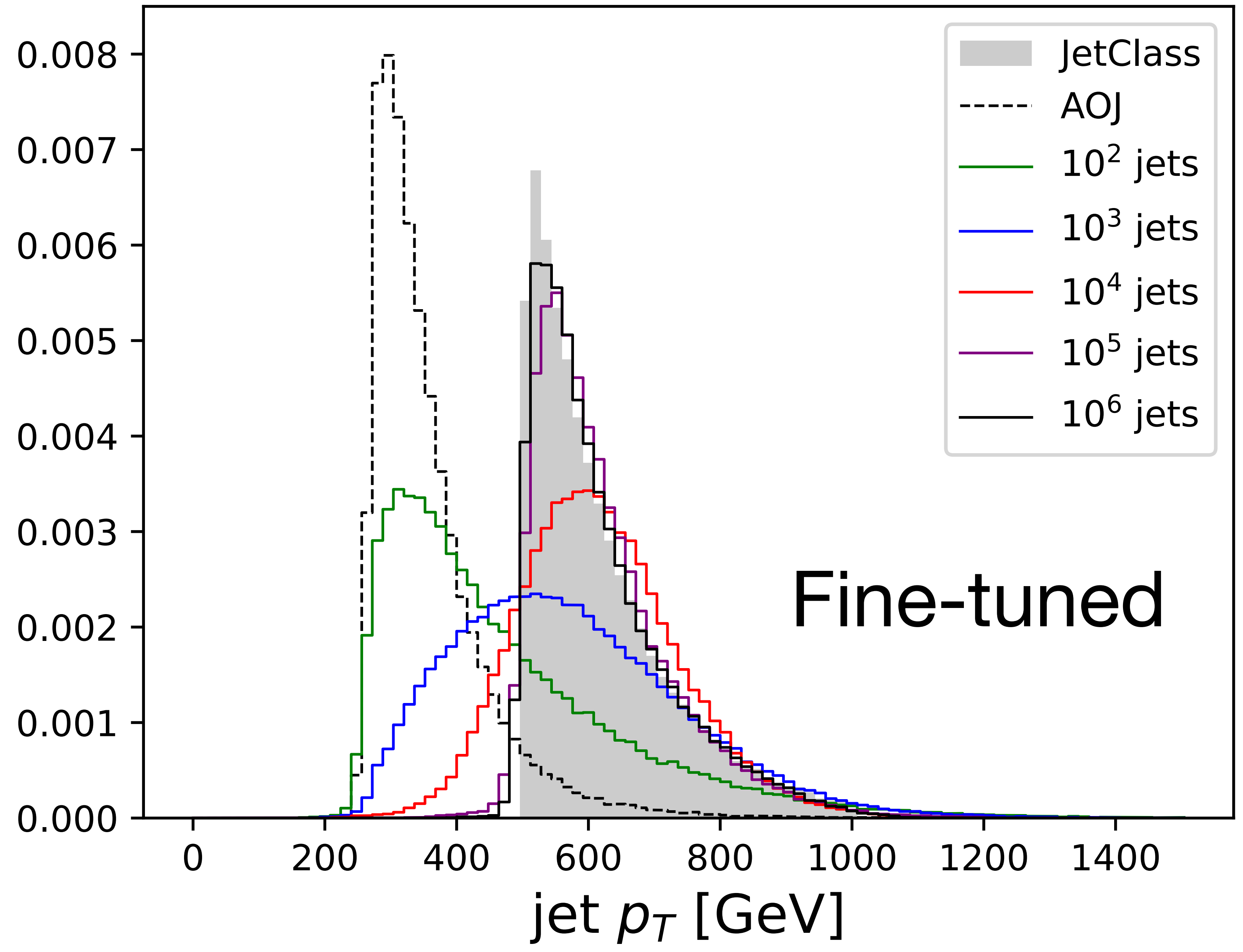
Better

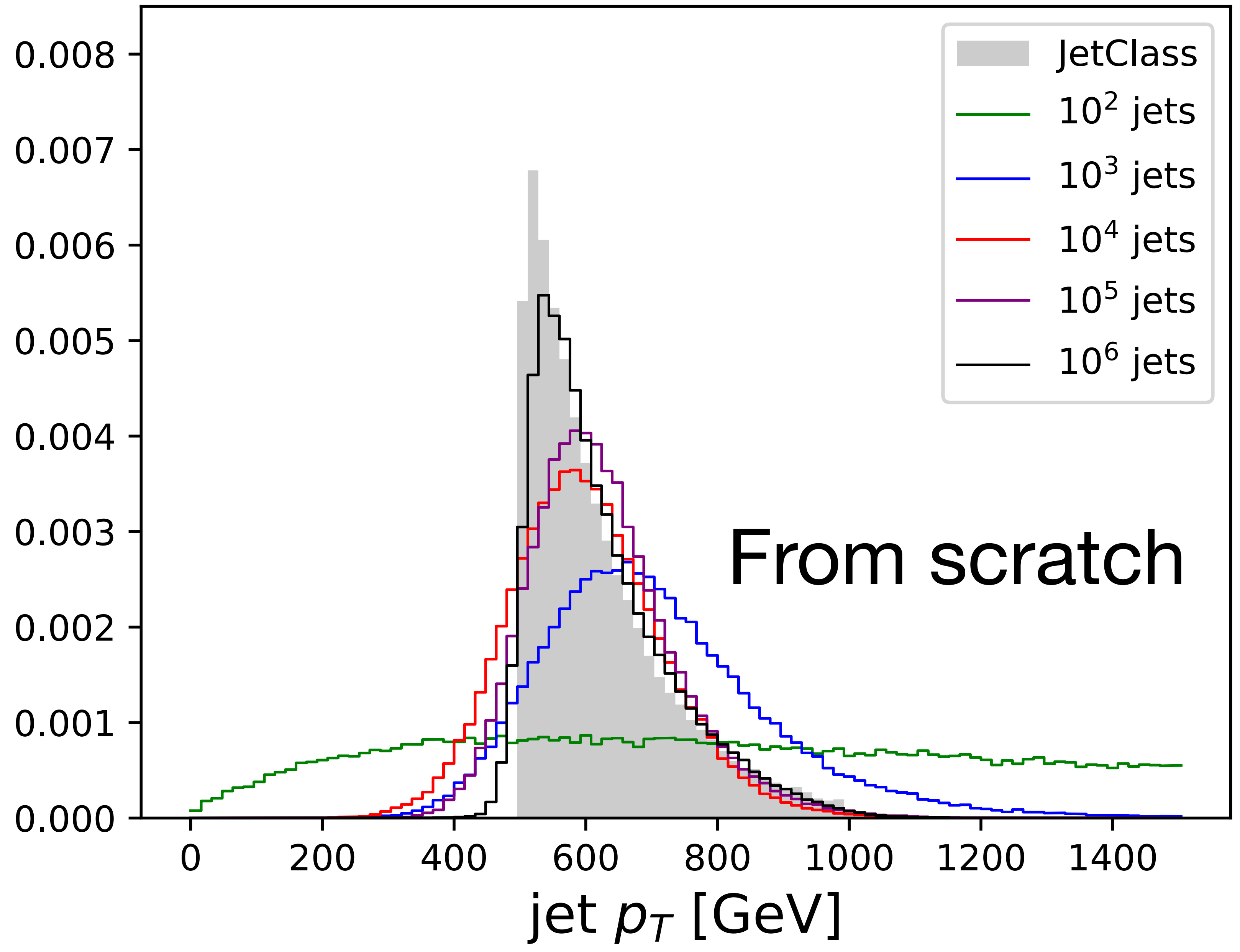




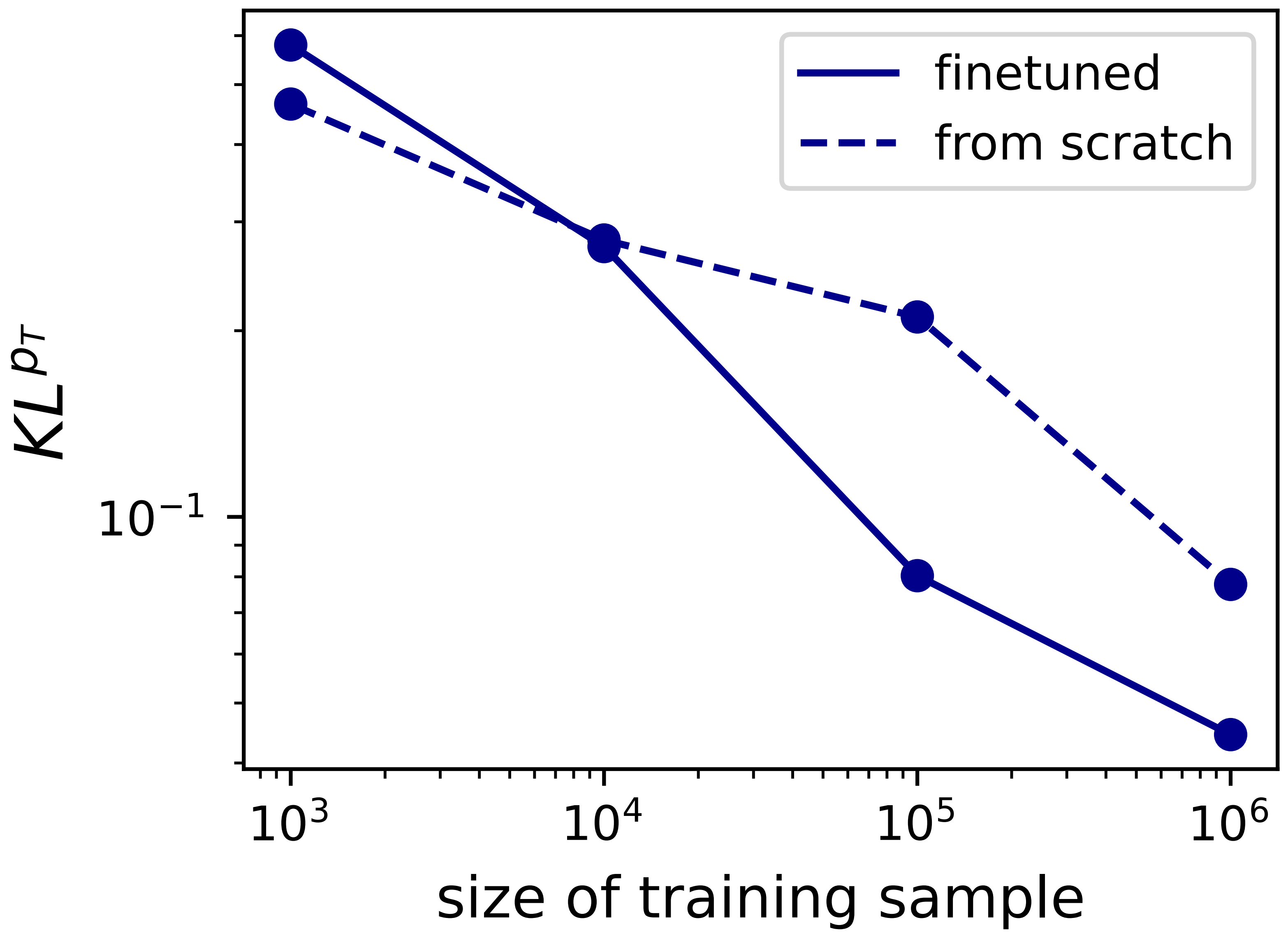
Results

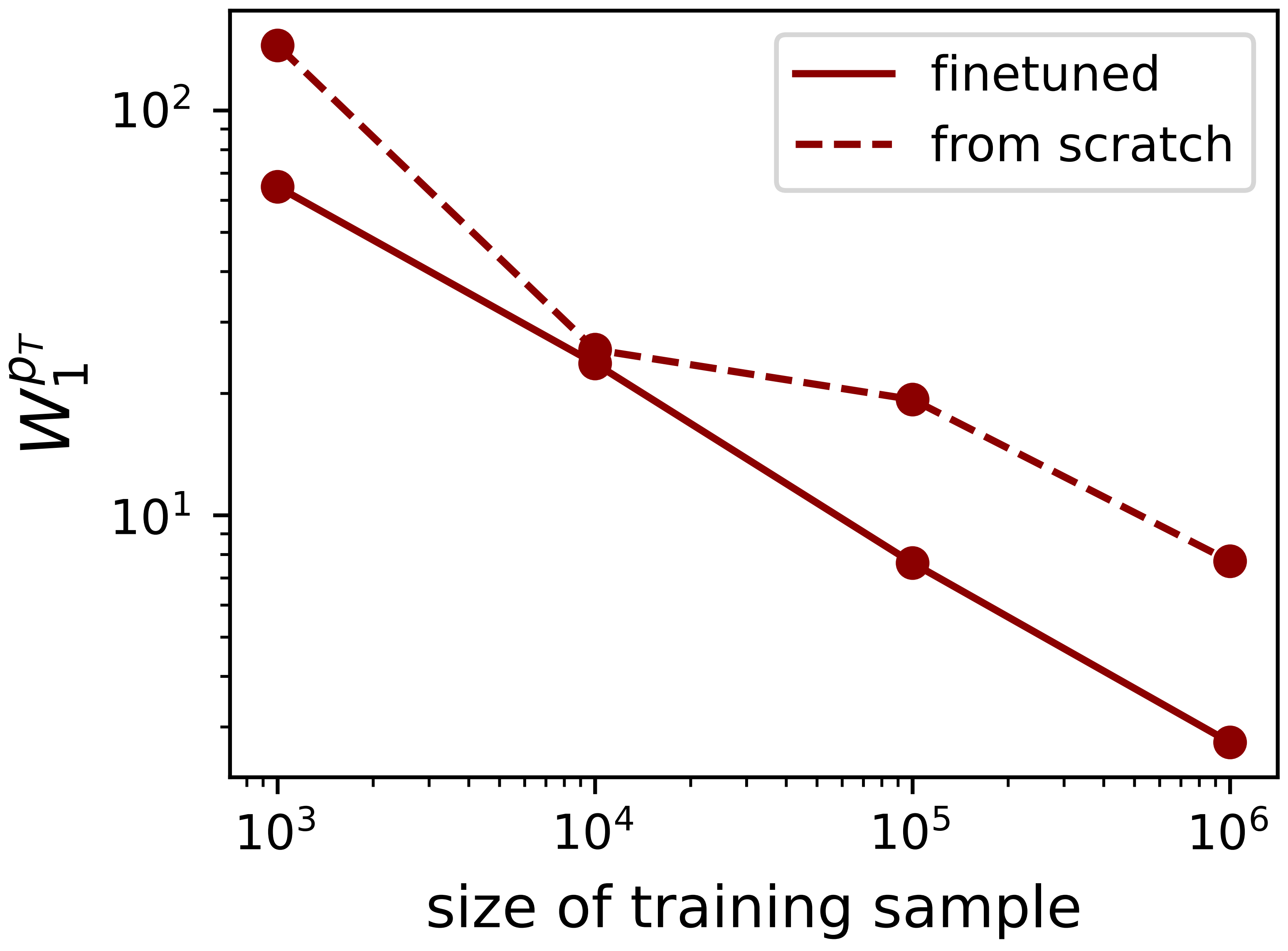
Jet kinematics





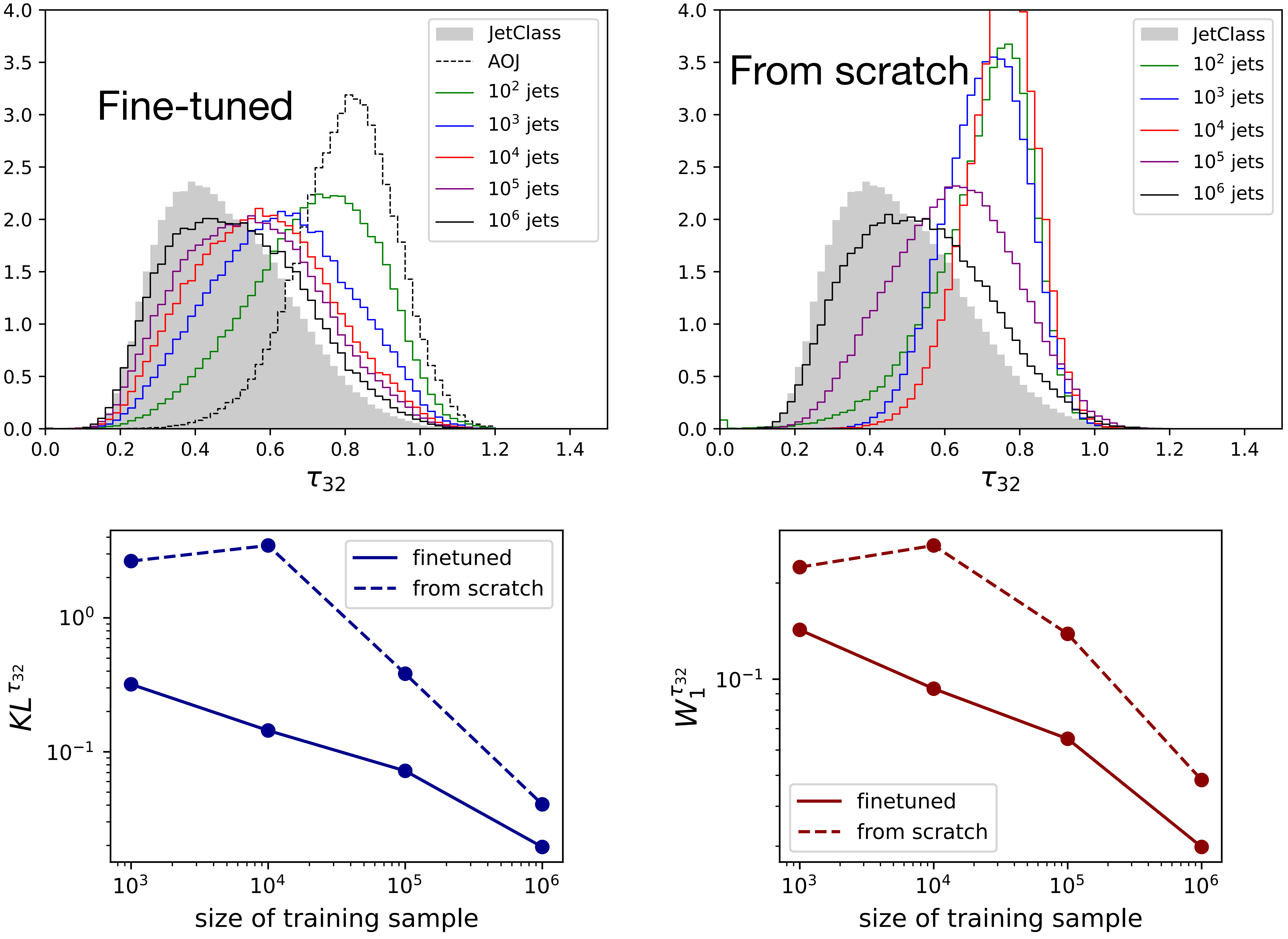
Better





Results

Substructure

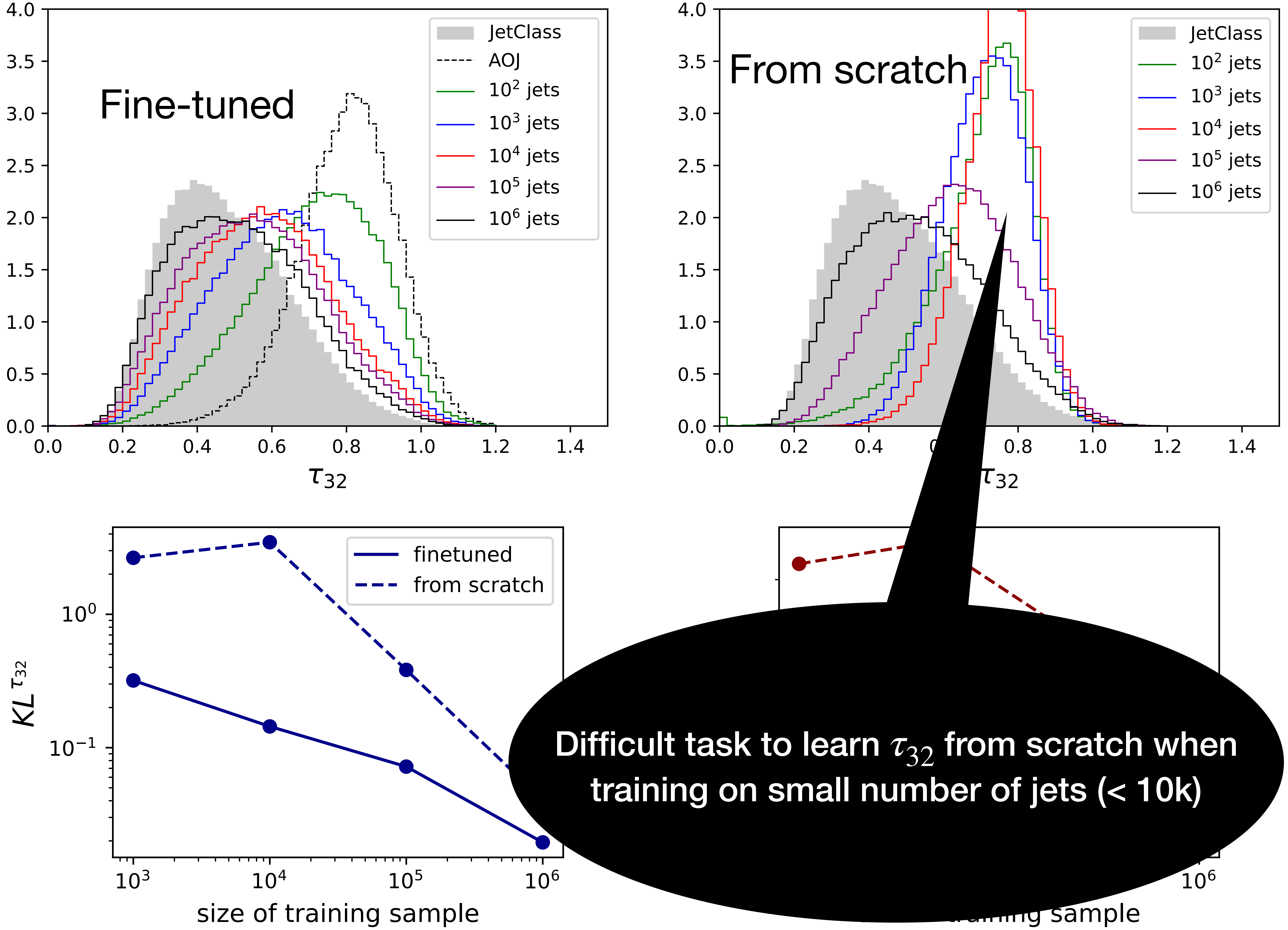


Better

Results

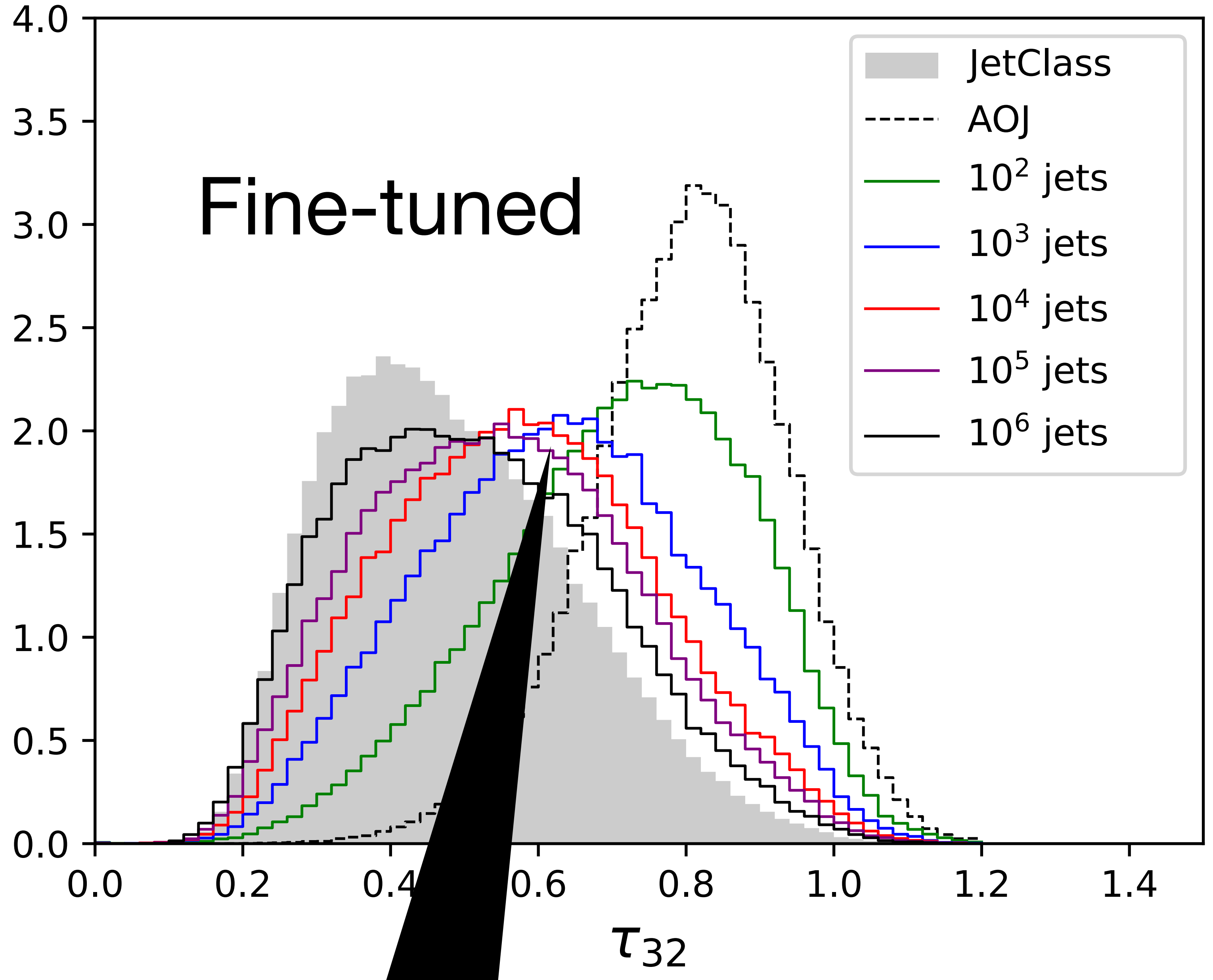
Substructure

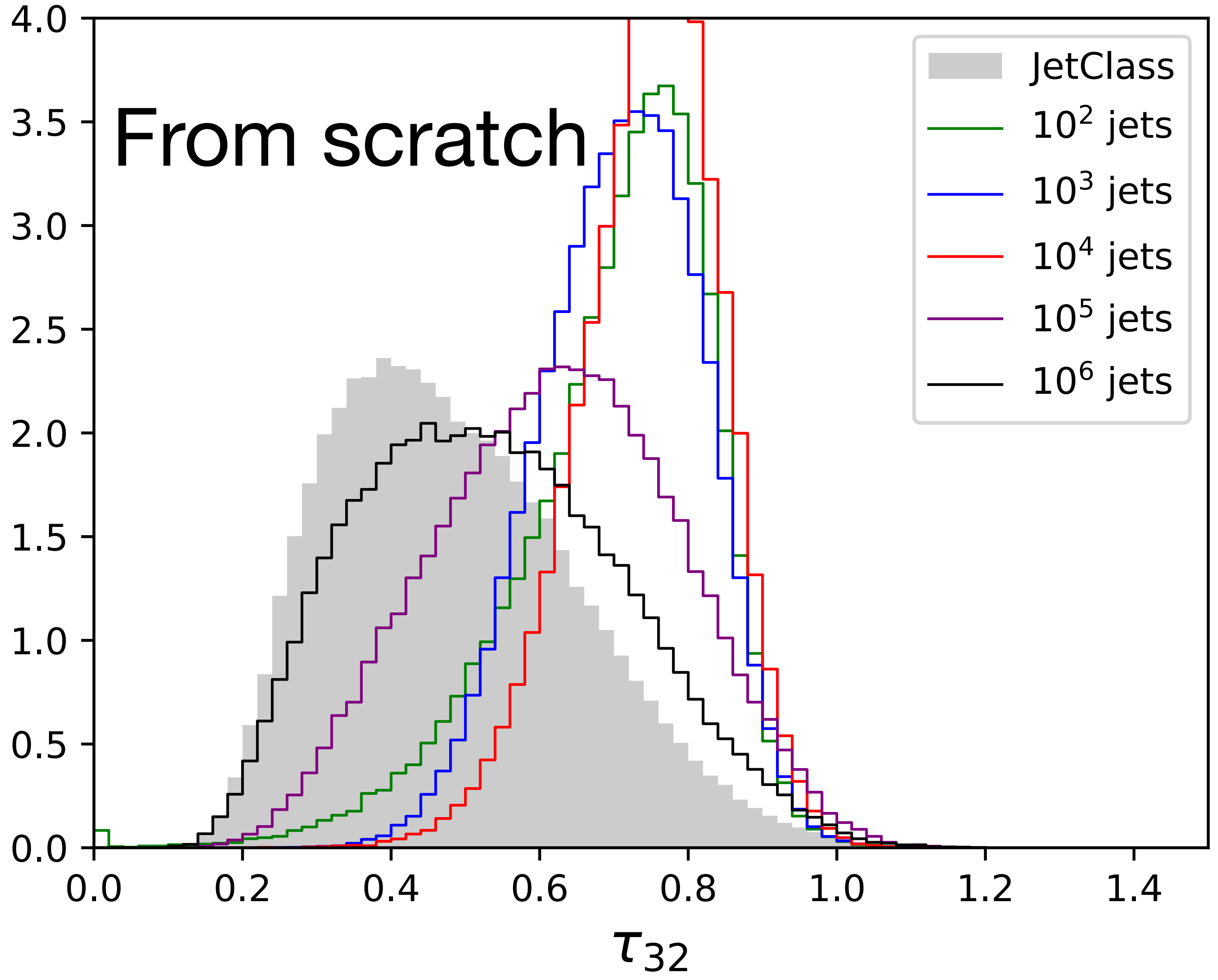
Better

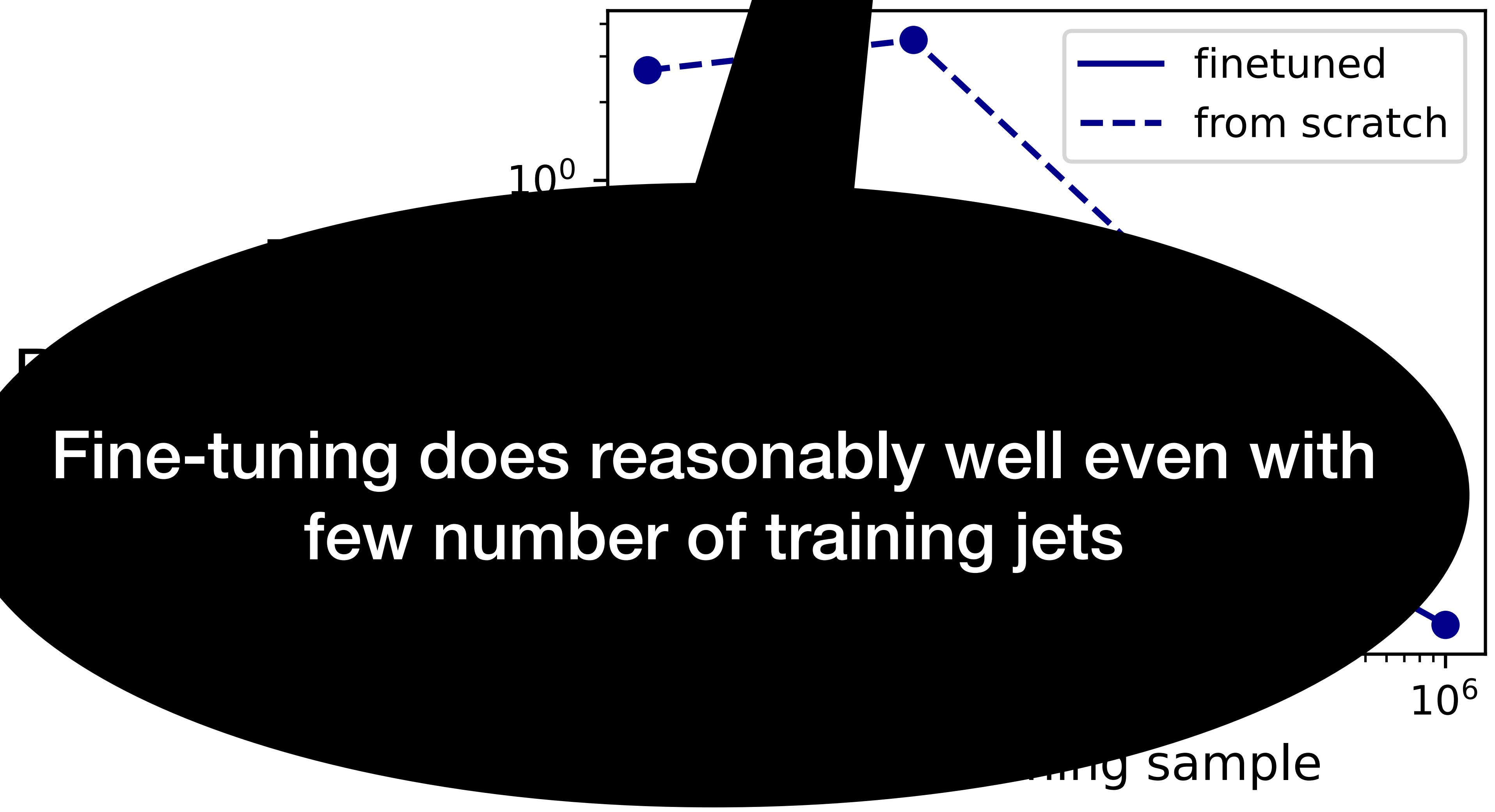


Results

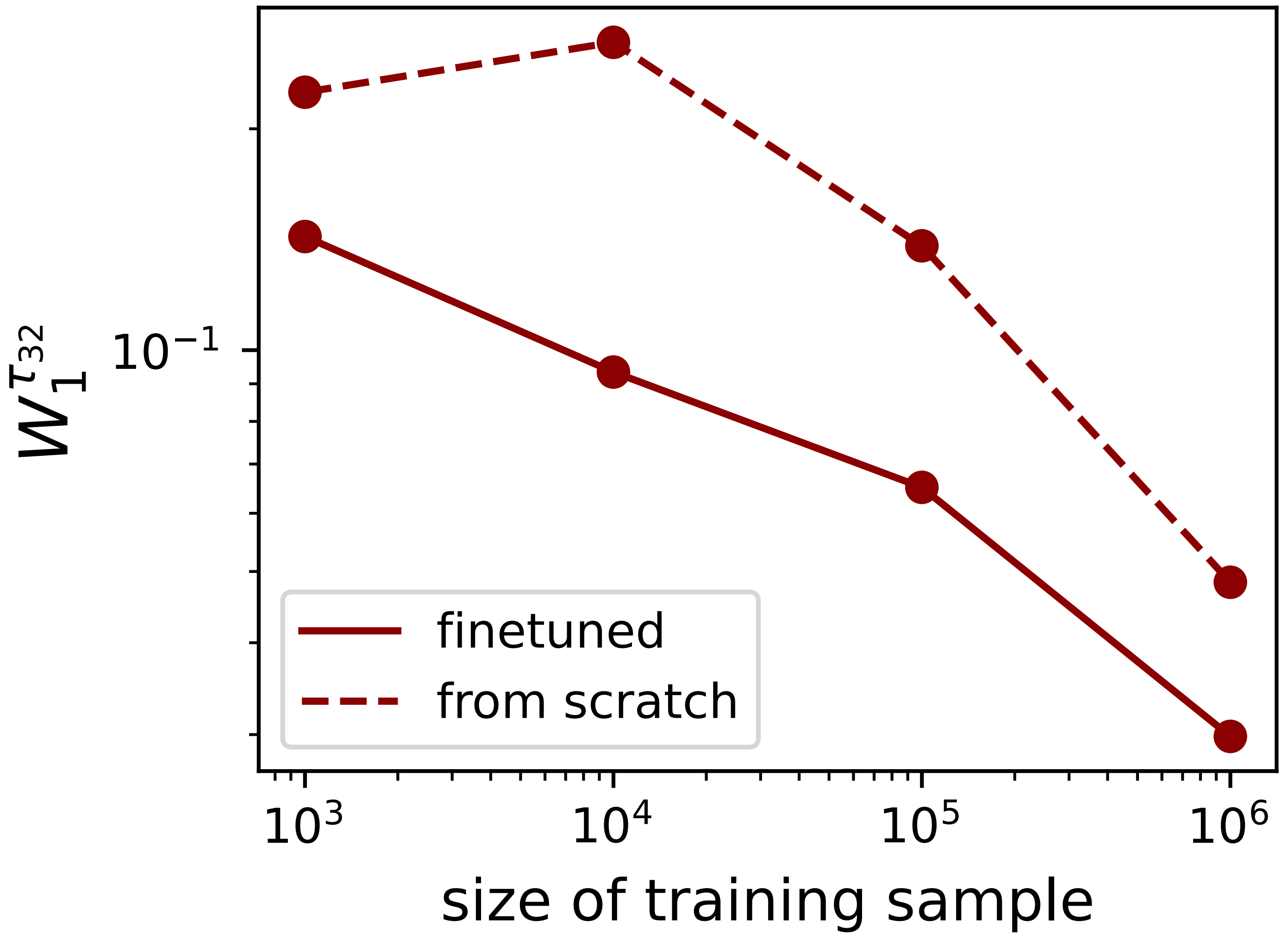
Substructure





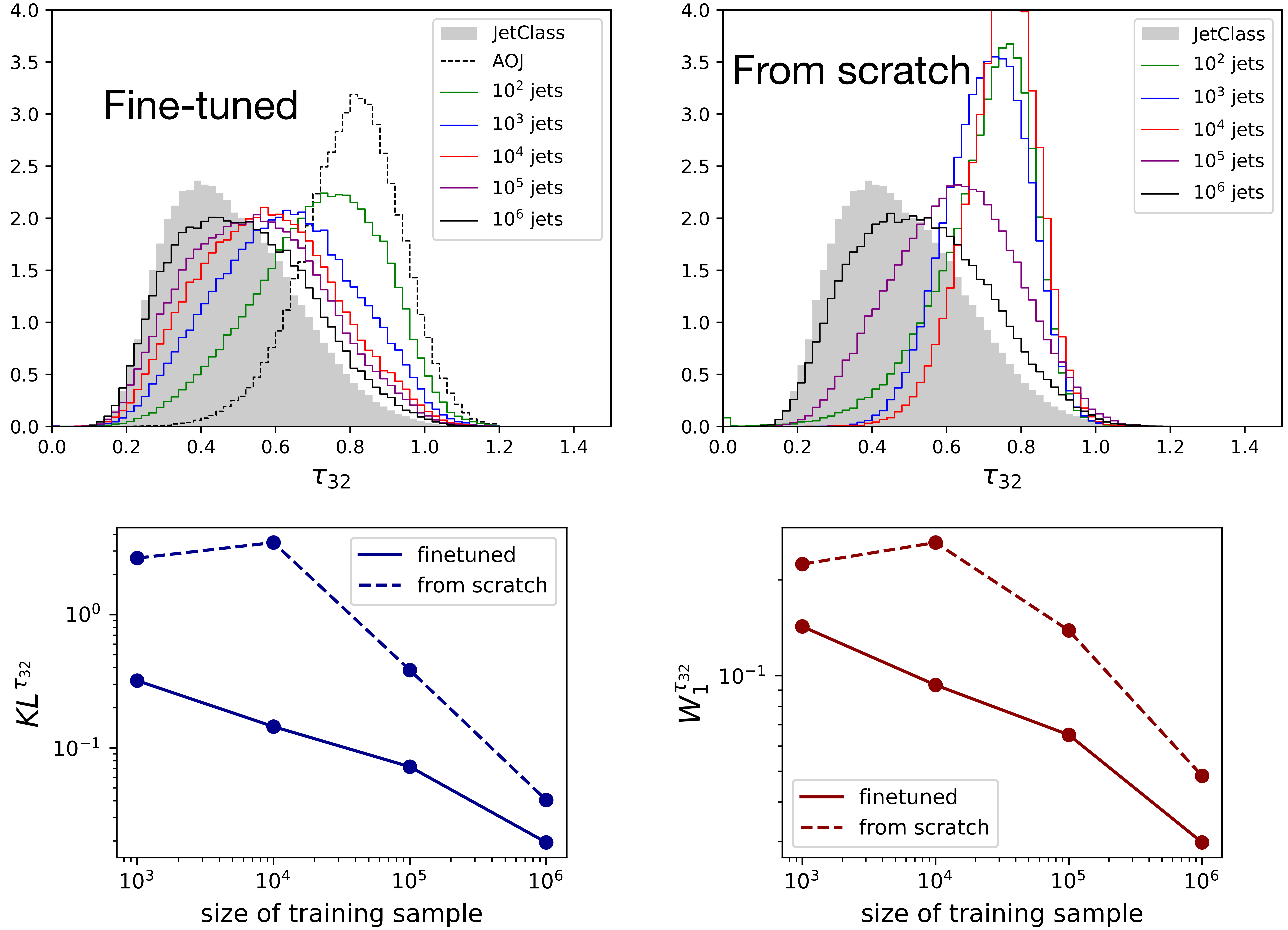


Fine-tuning does reasonably well even with few number of training jets



Results

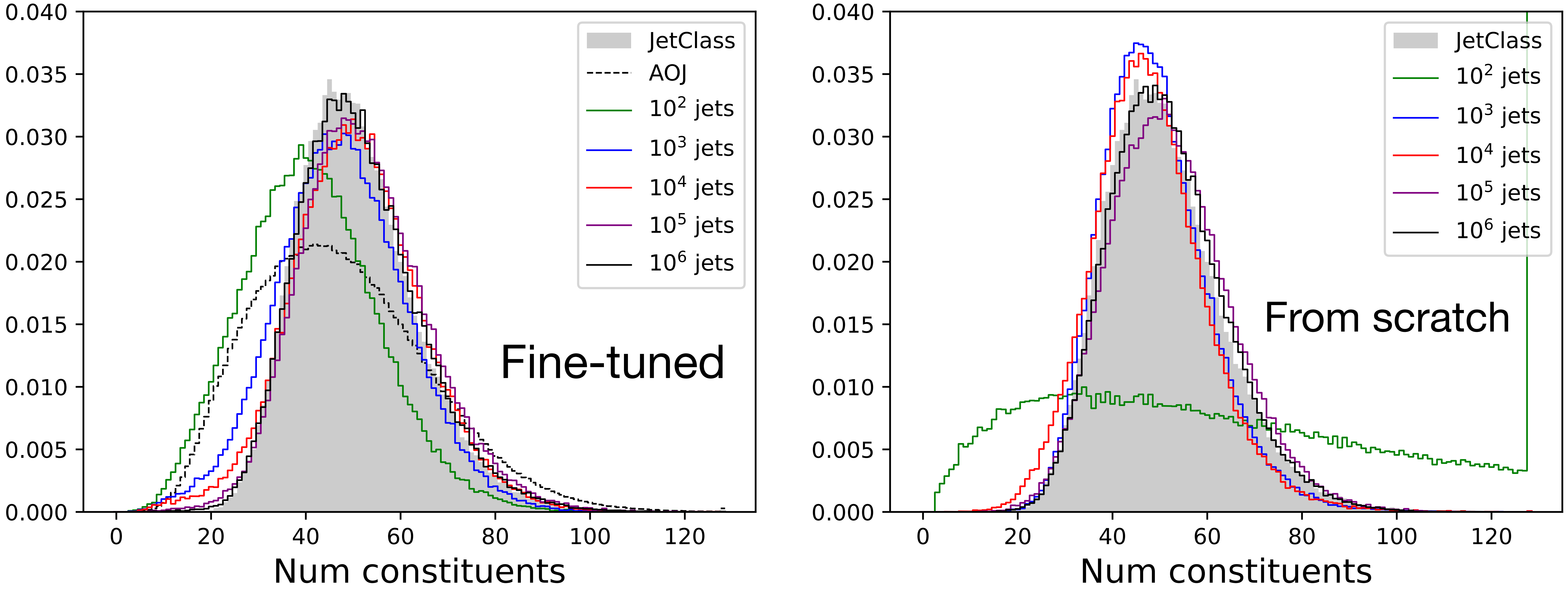
Substructure



Better

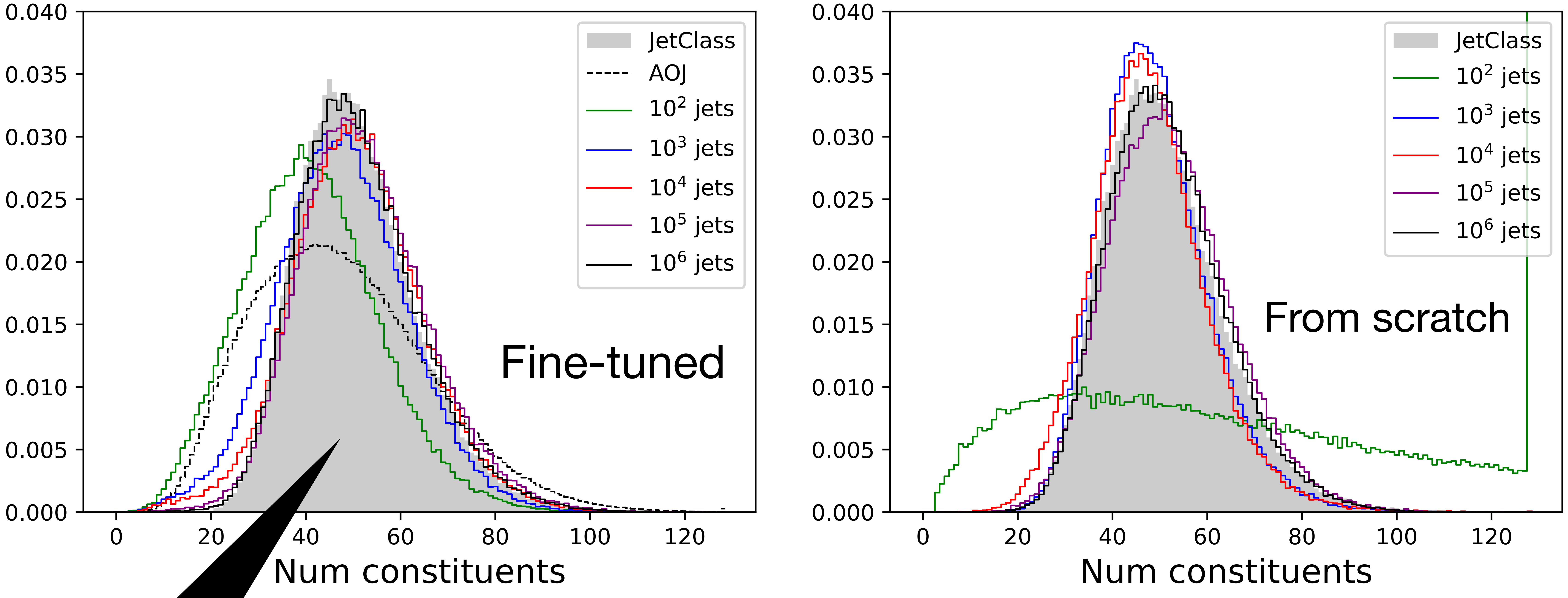
Results

Number of constituents



Results

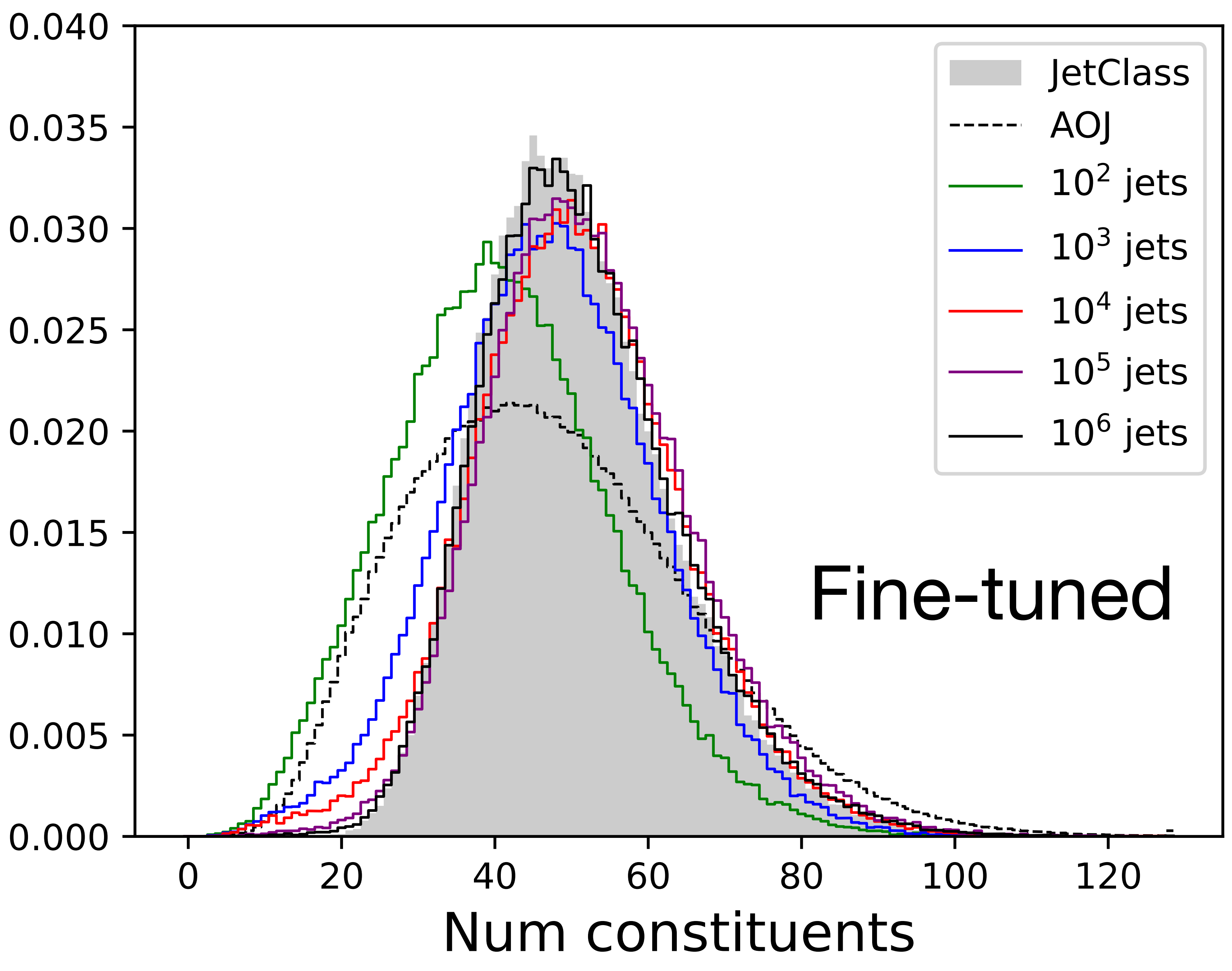
Number of constituents

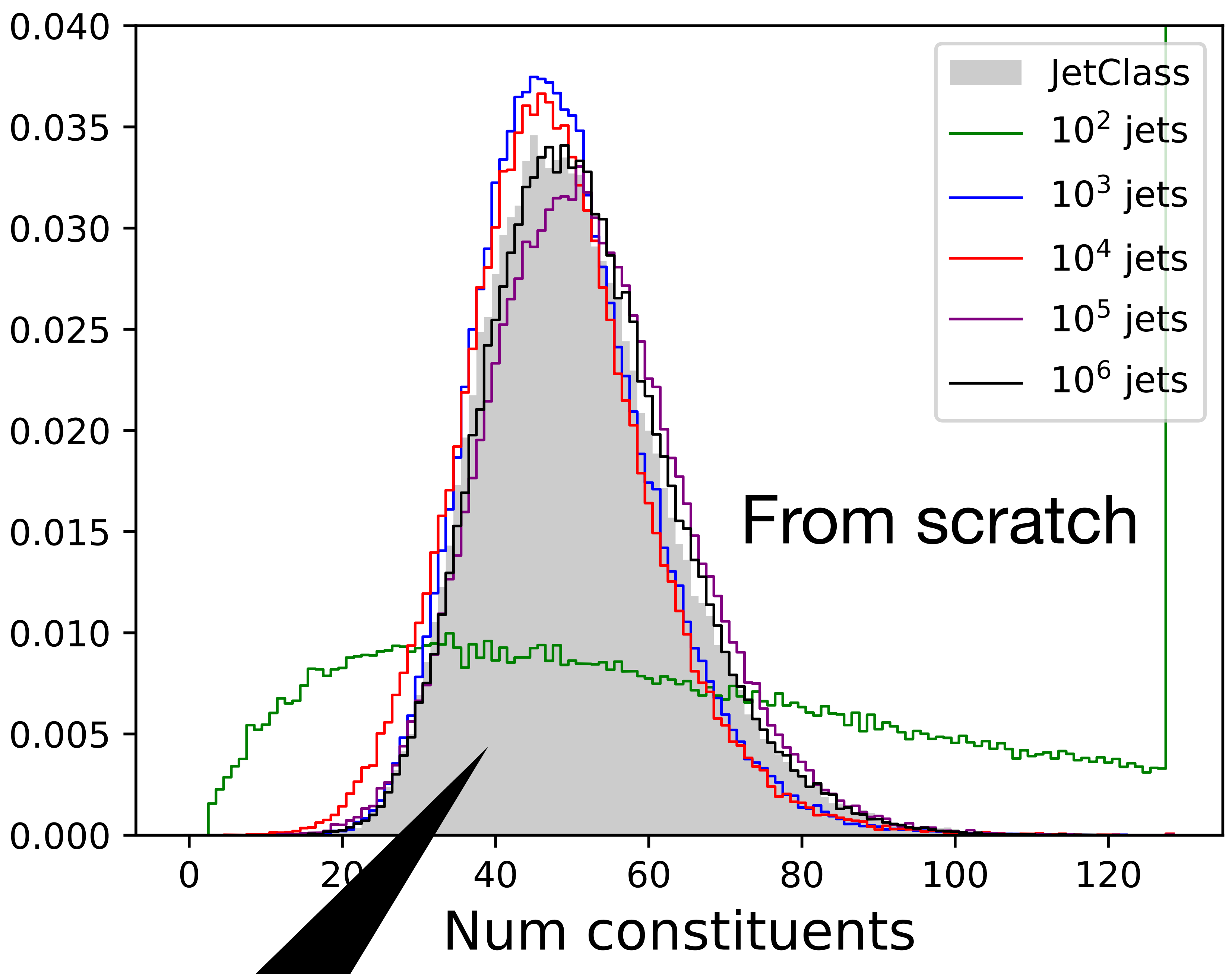


By construction, next-token prediction models can predict the number of jet constituents (i.e. predicting position of stop token)

Results

Number of constituents

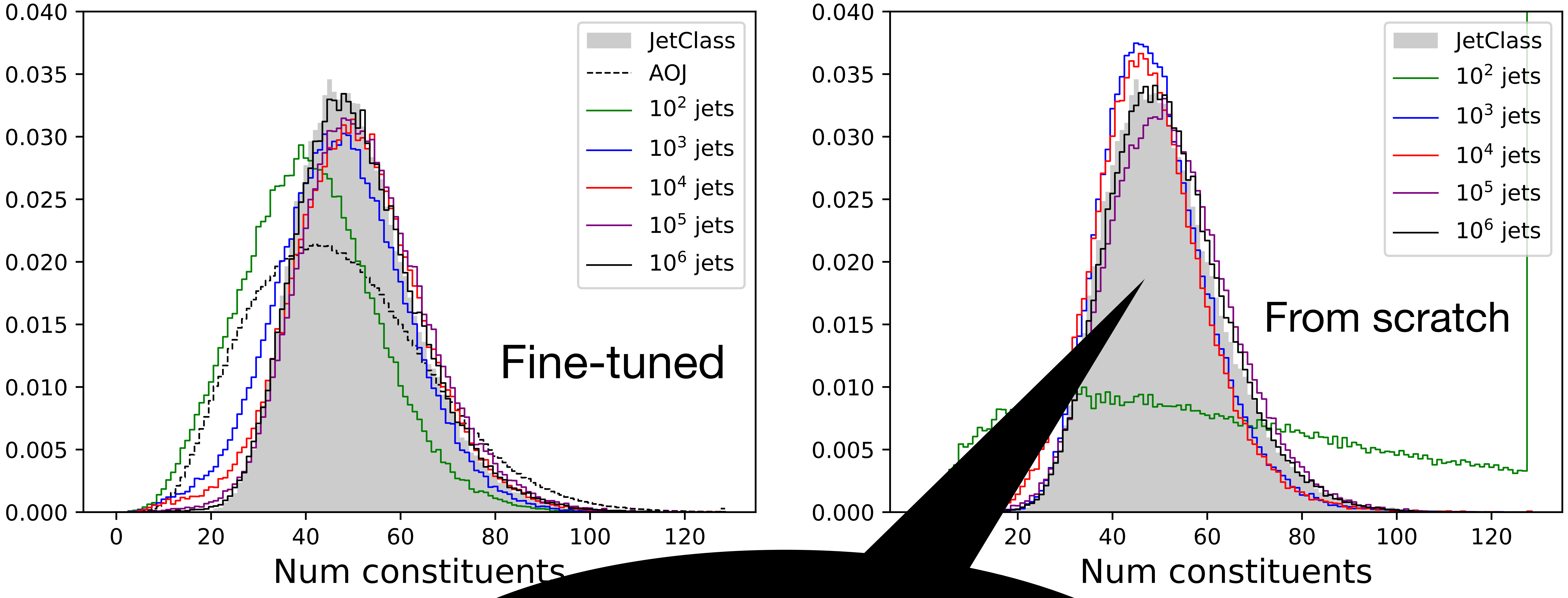




Number of constituents is not learned when training on only a 100 jets

Results

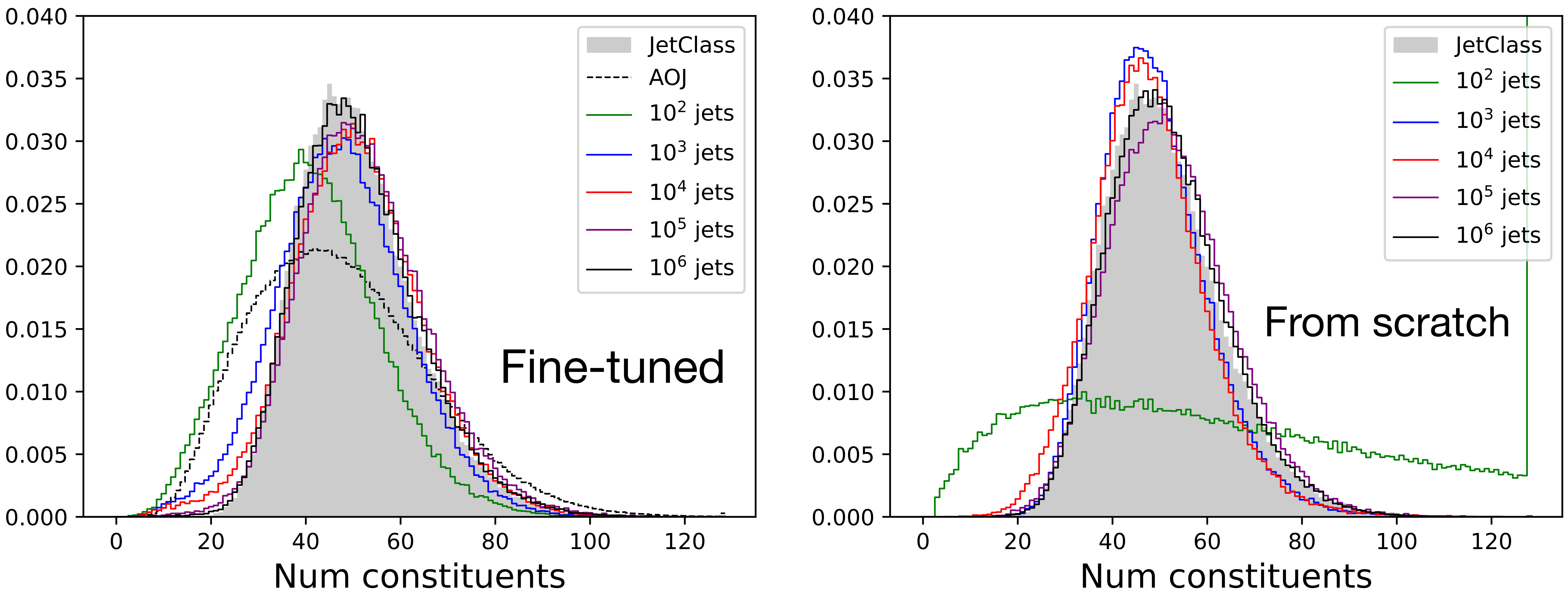
Number of constituents



The number of constituents is learned when training on more jets

Results

Number of constituents



Conclusions

Conclusions

- Aspen Open Jets is a new, large dataset with real/actual CMS jets

Conclusions

- Aspen Open Jets is a new, large dataset with real/actual CMS jets
- As a proof-of-concept, we show how pre-training on generating AOJ jets results in performance gain in a downstream generation task

Conclusions

- Aspen Open Jets is a new, large dataset with real/actual CMS jets
- As a proof-of-concept, we show how pre-training on generating AOJ jets results in performance gain in a downstream generation task
- Stay tuned! Plan to release the Aspen Open Jets dataset on Zenodo at the same time as our paper arXiv 2411.XXXXX

Conclusions

- Aspen Open Jets is a new, large dataset with real/actual CMS jets
- As a proof-of-concept, we show how pre-training on generating AOJ jets results in performance gain in a downstream generation task
- Stay tuned! Plan to release the Aspen Open Jets dataset on Zenodo at the same time as our paper arXiv 2411.XXXXX

Thank you!

Backup

Previous ML works with real CMS data

- 1704.05066
- 1704.05842
- 1908.08542

Jet datasets with fewer jets than AOJ

- 2312.06909 - single-lepton datasets

Tokenized features

- Total of 8192 tokens
- Found that increasing number of tokens did not significantly increase reconstruction quality