Flow Matching Beyond Kinematics: Generating Jets with Particle-ID and Trajectory Displacement Information

Joschka Birk, Erik Buhmann, Cedric Ewen, Gregor Kasieczka, David Shih
joschka.birk@uni-hamburg.de

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

DESY Round Table on Deep Learning
November 14, 2023

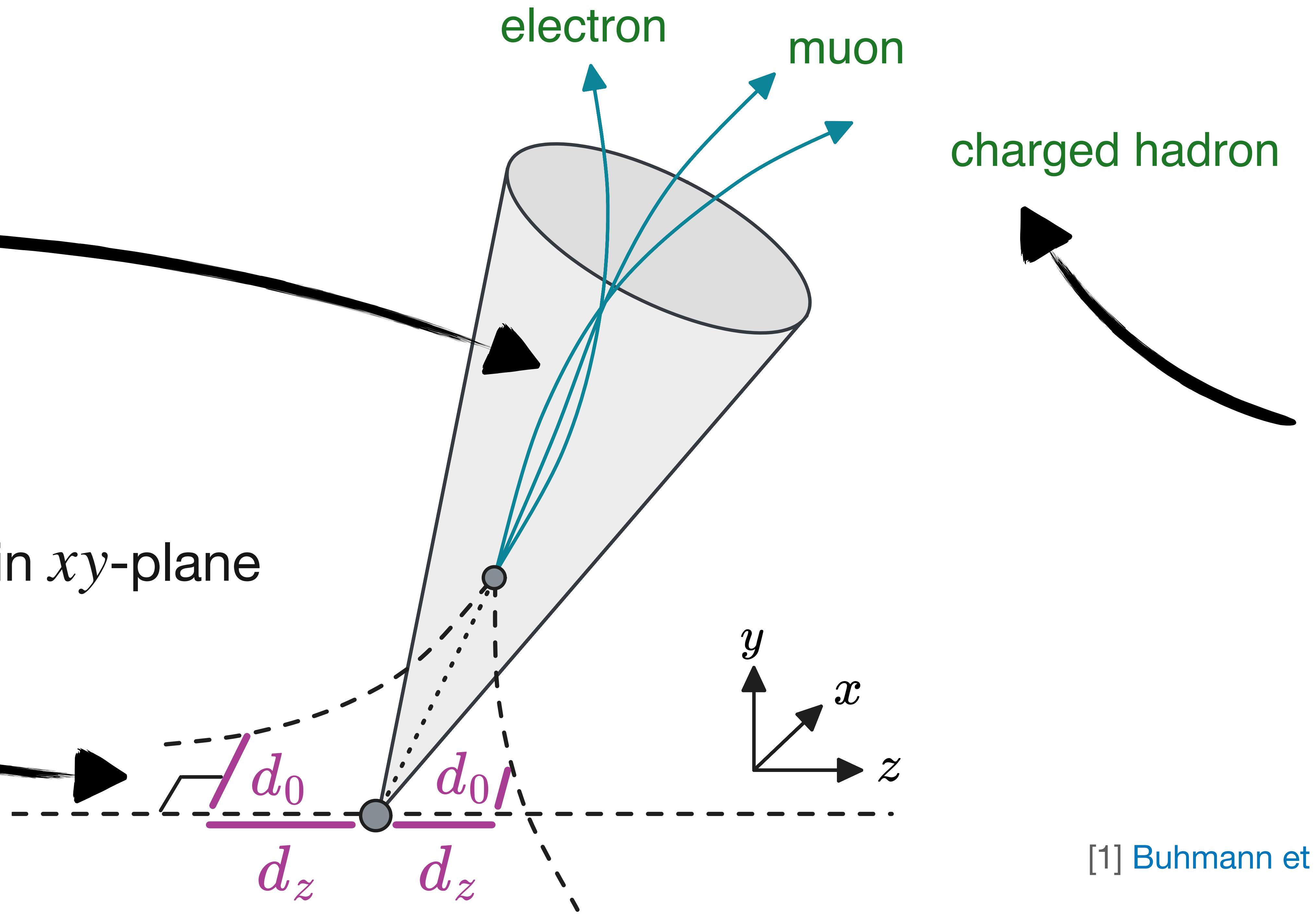
Motivation

- **Generative ML models** in HEP are intensively researched (detector simulation, event generation, **jet generation**, …)
- Jet generation can e.g. **help in anomaly detection** [1]
- Jets = **point cloud of jet constituents**
- **So far**, most studies in jet generation are **limited to kinematic features** of constituents

→ Can we also generate jets with **Particle-ID** and **trajectory displacement information**?

Kinematics :
p_T, η, ϕ

Trajectory displacement :
d_0 : closest approach to interaction point in xy-plane
d_z : z position where d_0 is evaluated



Particle-ID and charge :
`isElectron, isPhoton, …`

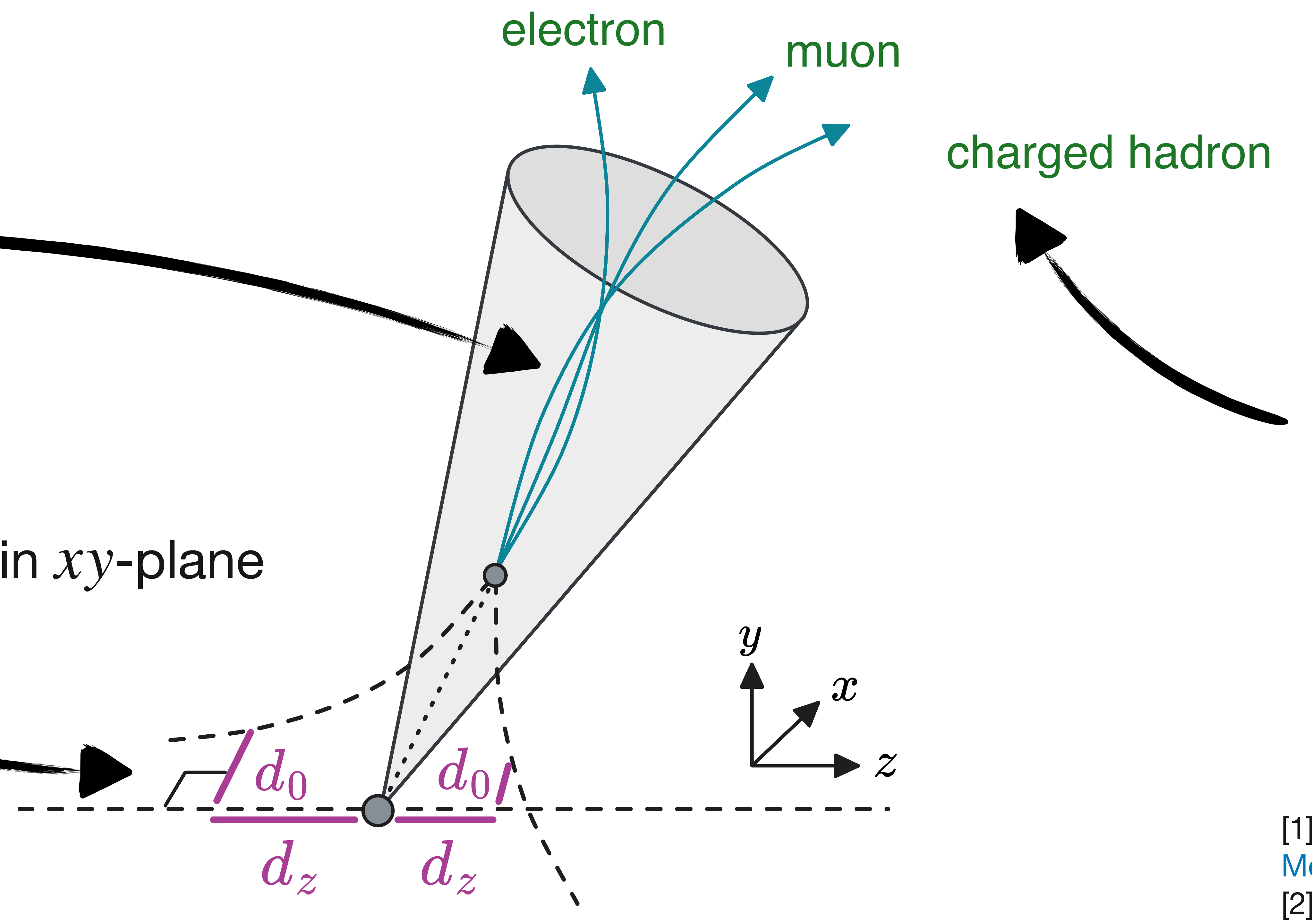
[1] Buhmann et al. "Full Phase Space Resonant Anomaly Detection"

The JetClass dataset

	JetNet [1]	**JetClass [2]** (used here)
Jet types	5 types	10 types (several decay channels for top and H jets)
Dataset size	180 thousand jets per class	12.5 million jets per class (70x more than JetNet)
Features	Kinematics	Kinematics, Particle-ID and charge, trajectory displacement

Kinematics :
p_T, η, ϕ

Trajectory displacement :
d_0 : closest approach to interaction point in xy-plane
d_z : z position where d_0 is evaluated



Particle-ID and charge :
`isElectron, isPhoton, …`

[1] Kansal et al. (2021) “Particle Cloud Generation with Message Passing Generative Adversarial Networks”
[2] Qu et al. (2022) "JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics"

Flow matching [1]

- Main idea:
  learn a CNF which transforms the data distribution p_{data} into a simple base distribution p_{base}

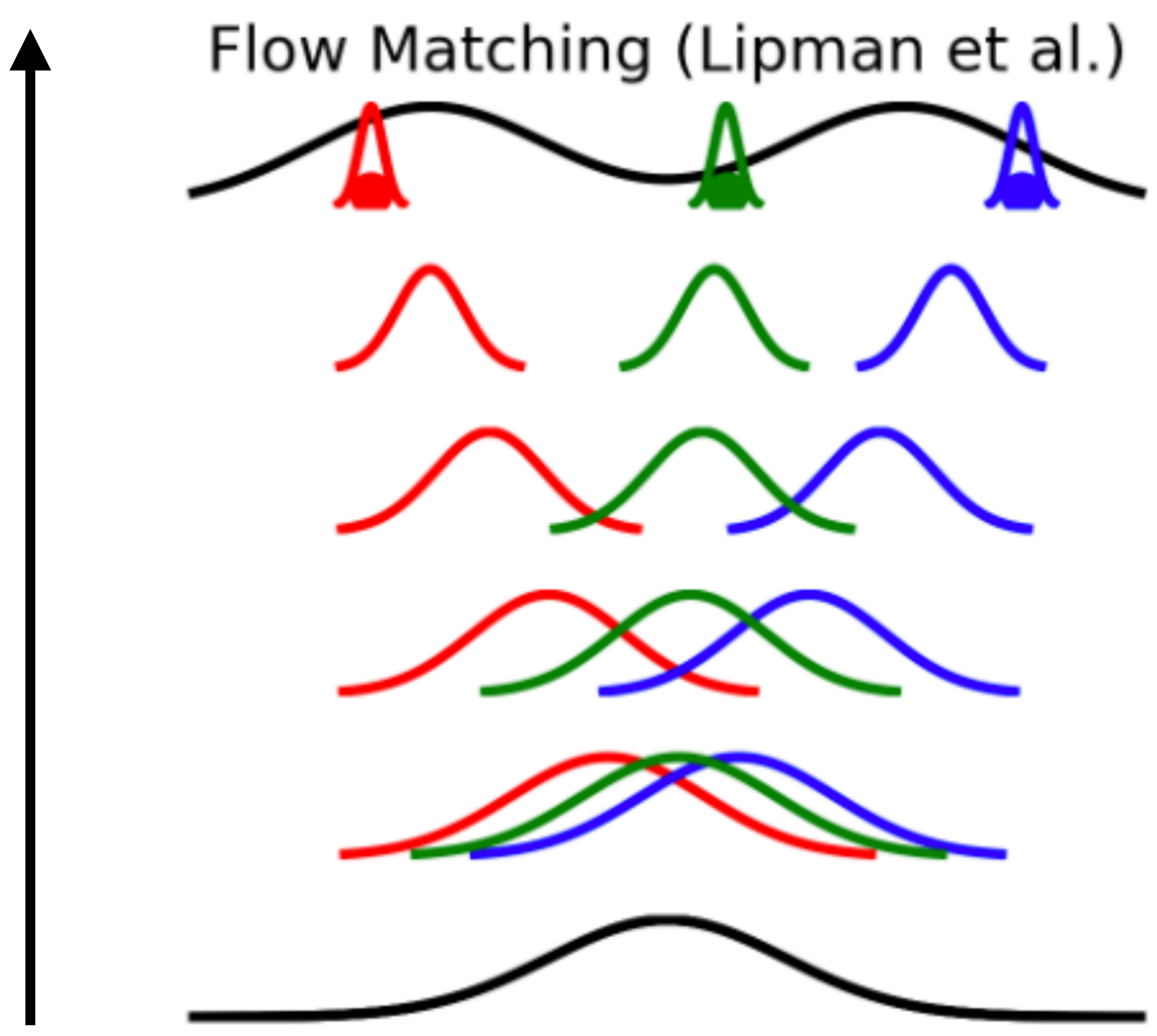
- Model $v_\theta(x_t, t)$ approximates the vector field
  $$u_t(x_t \mid x_0) = \frac{dx_t}{dt}$$

- Construct the target vector field $u_t(x_t \mid x_0)$ corresponding to trajectories $x_t = \gamma_t x_0 + \sigma_t \epsilon$

- Train by minimizing the MSE loss
  $\mathscr{L}_{\text{MSE}} = |v_\theta(x_t, t) - u_t(x_t \mid x_0)|^2$

- **Inference / jet generation:**
  sample from $p_{\text{base}}(x)$ and **integrate ODE**

$x \propto p_{\text{data}}$ at $t = 0$



$x \propto p_{\text{base}}$ at $t = 1$

Image from arXiv:2302.00482

[1] Lipman et al. (2023) "Flow Matching for Generative Modeling"

Model overview

- Training on the **flow matching objective** [1]
- Model architecture based on EPiC layers [2] and similar to EPiC-FM model [3]
- Model is **conditioned on jet type**, jet η and jet p_{T}
- **Model generates 13 features per constituent**:
  - 3 kinematic features: $p_{\mathrm{T}}^{\mathrm{rel}}, \eta^{\mathrm{rel}}, \phi^{\mathrm{rel}}$
  - 4 trajectory displacement features: $d_0, d_z, \sigma_{d_0}, \sigma_{d_z}$
  - 6 discrete features: **`isElectron, isMuon, isChargedHadron, isNeutralHadron, isPhoton, charge`**

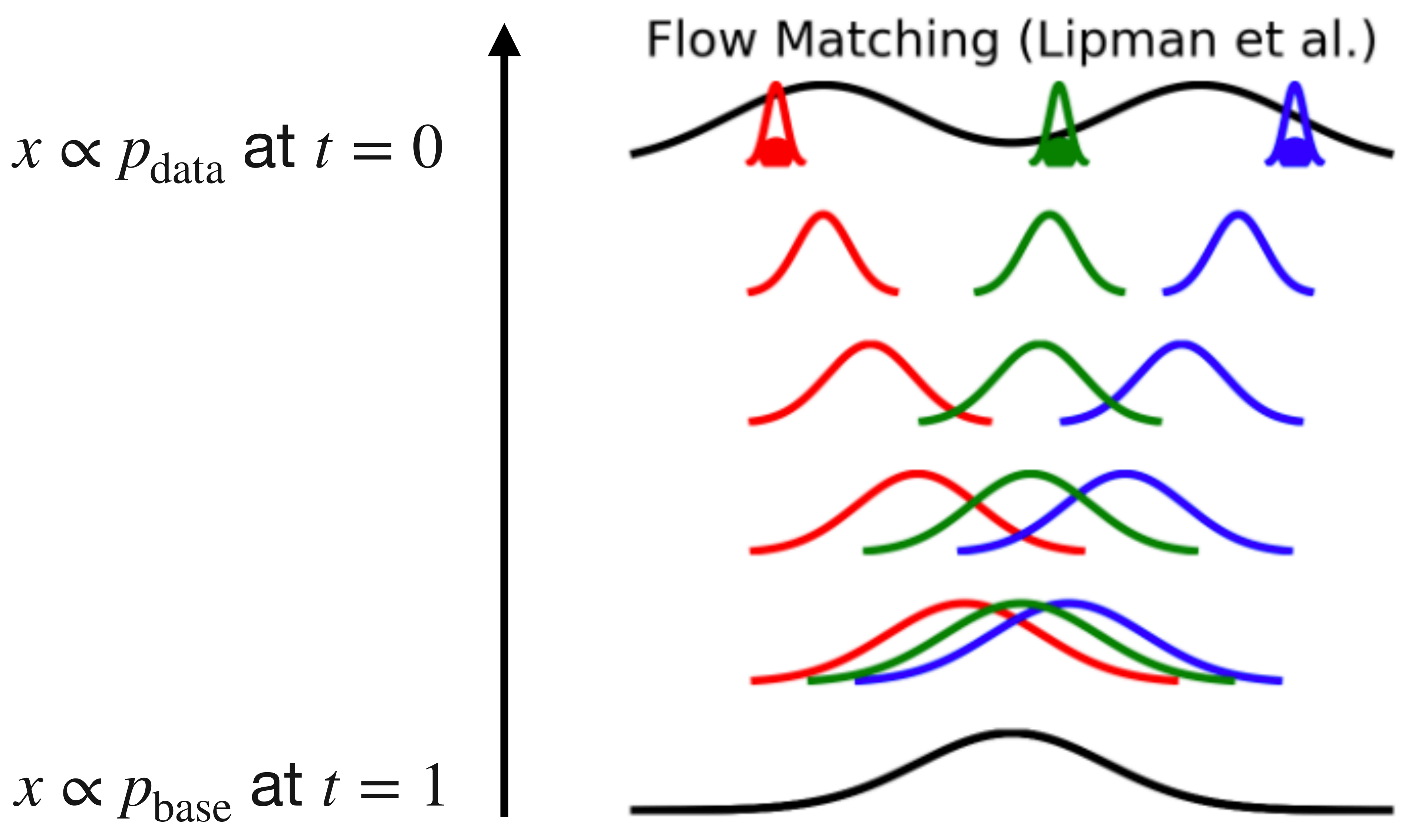


Image from arXiv:2302.00482

[1] Lipman et al. (2023) "Flow Matching for Generative Modeling"
[2] Buhmann et al. (2023) "EPiC-GAN: Equivariant point cloud generation for particle jets"
[3] Buhmann et al. (2023) “EPiC-ly Fast Particle Cloud Generation with Flow-Matching and Diffusion”

Model architecture sketch

x20 for the model presented here

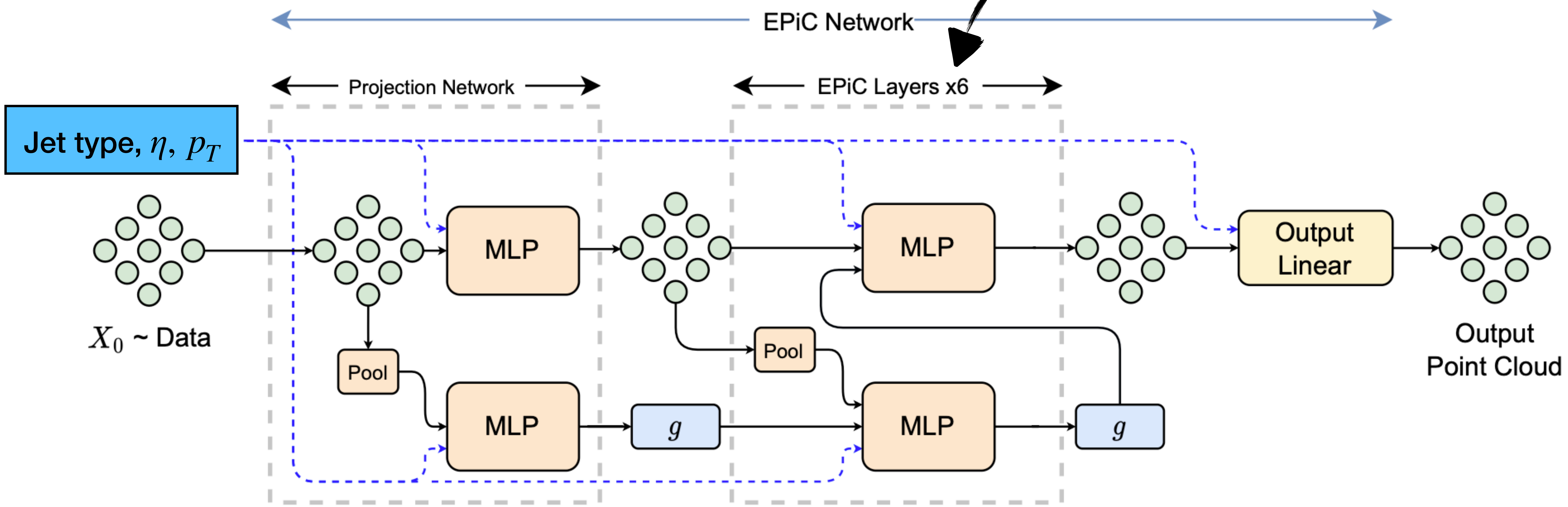
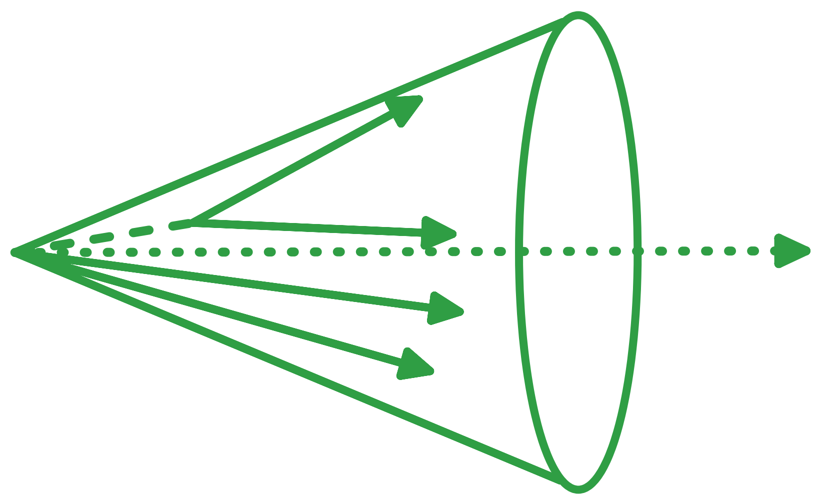


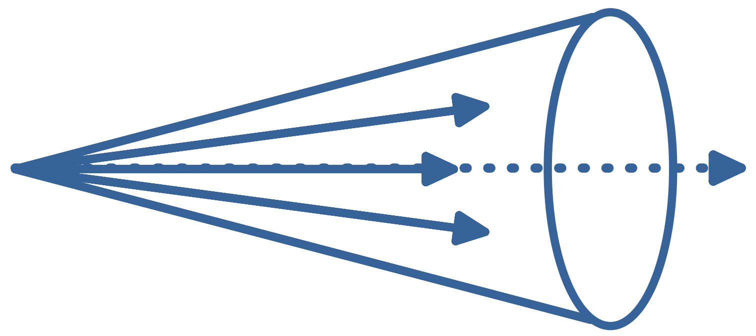
Fig. 2: Model schema of the *EPiC Network* generator architecture used in both EPiC-JeDi and EPiC-FM. Each multi-layer perceptron (MLP) is a two-layer neural network with LeakyReLU activation. The pooling operation is a concatenation of both average and summation pooling.

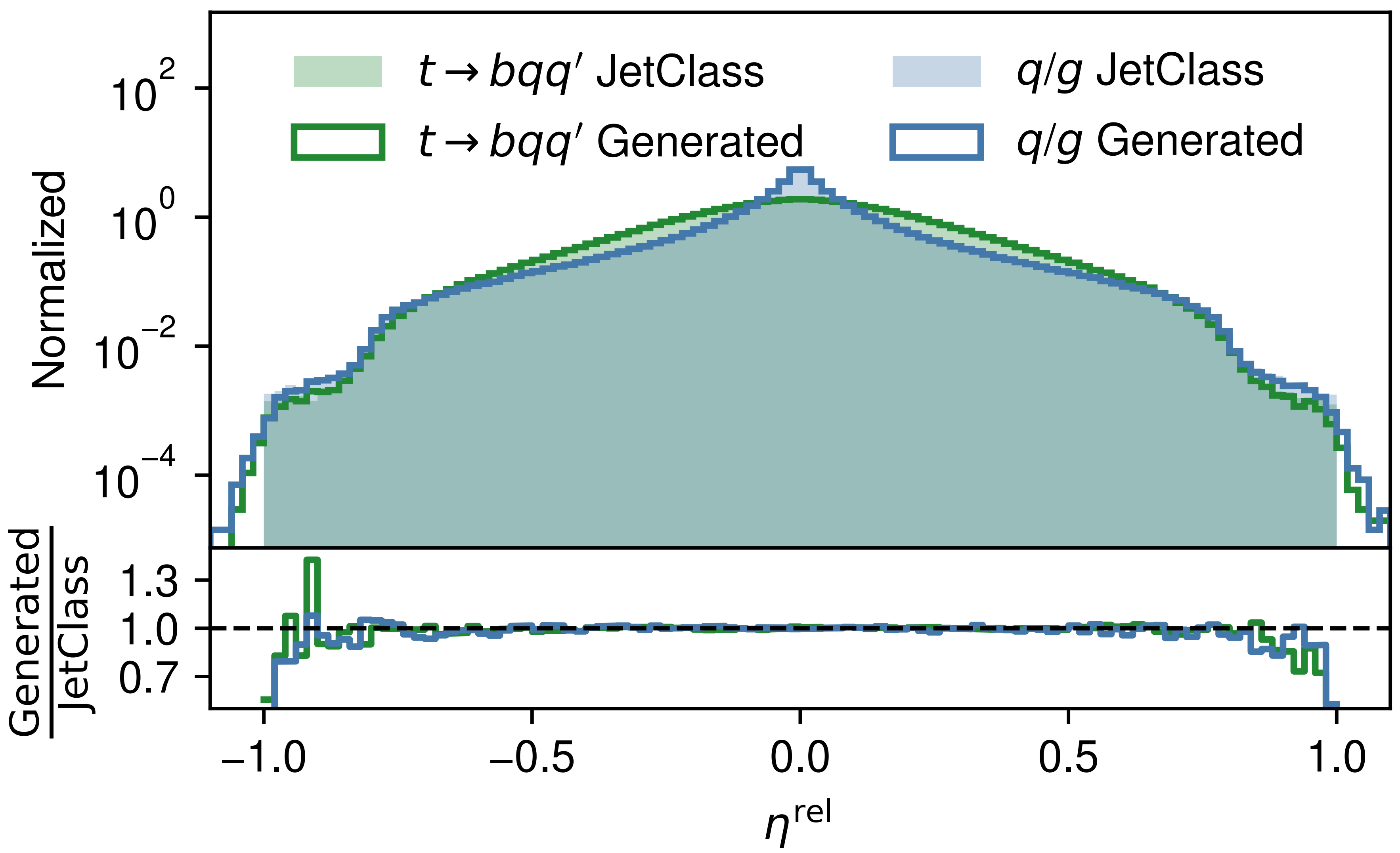
Image adapted from Buhmann et al. (2023) “EPiC-ly Fast Particle Cloud Generation with Flow-Matching and Diffusion”

Kinematic particle features

- Looking at two very different types of jets here
  - $t \to bqq'$ jets:
    - Wider: larger η^{rel}
    - More constituents: smaller $p_{\mathrm{T}}^{\mathrm{rel}}$
  - q/g jets:
    - Narrow: smaller η^{rel}
    - Fewer constituents: larger $p_{\mathrm{T}}^{\mathrm{rel}}$

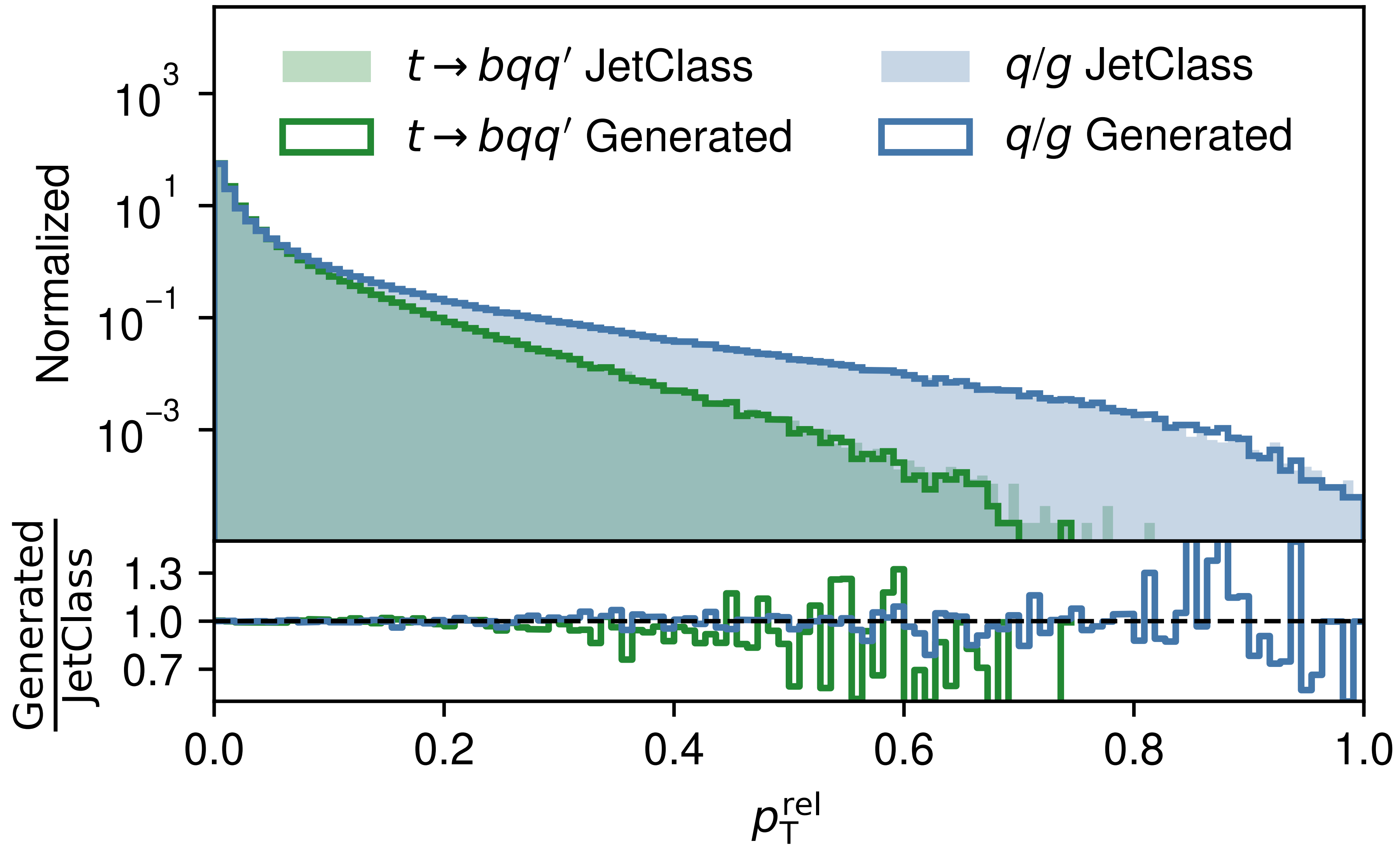






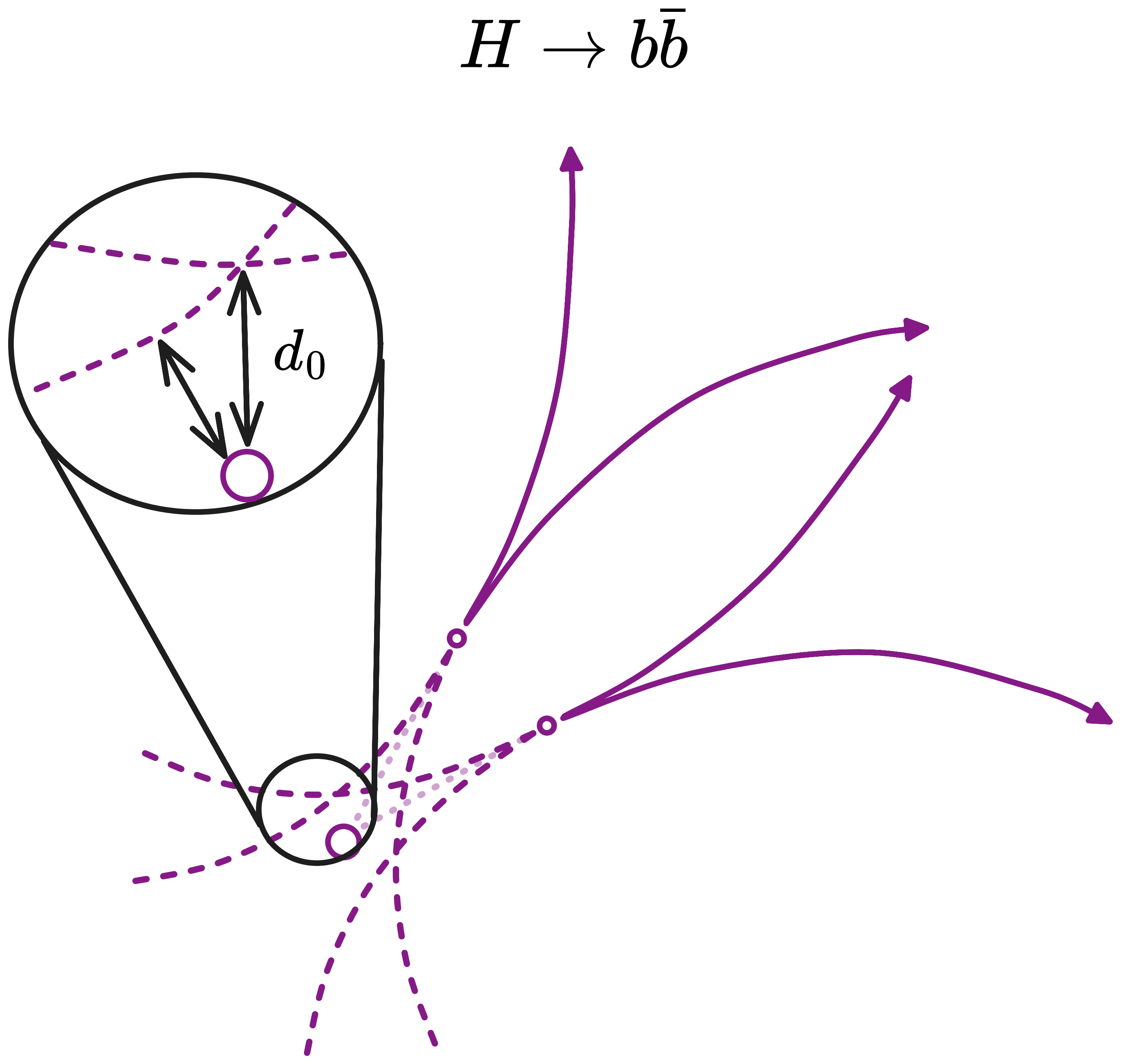
- Very good agreement in KL-divergence:

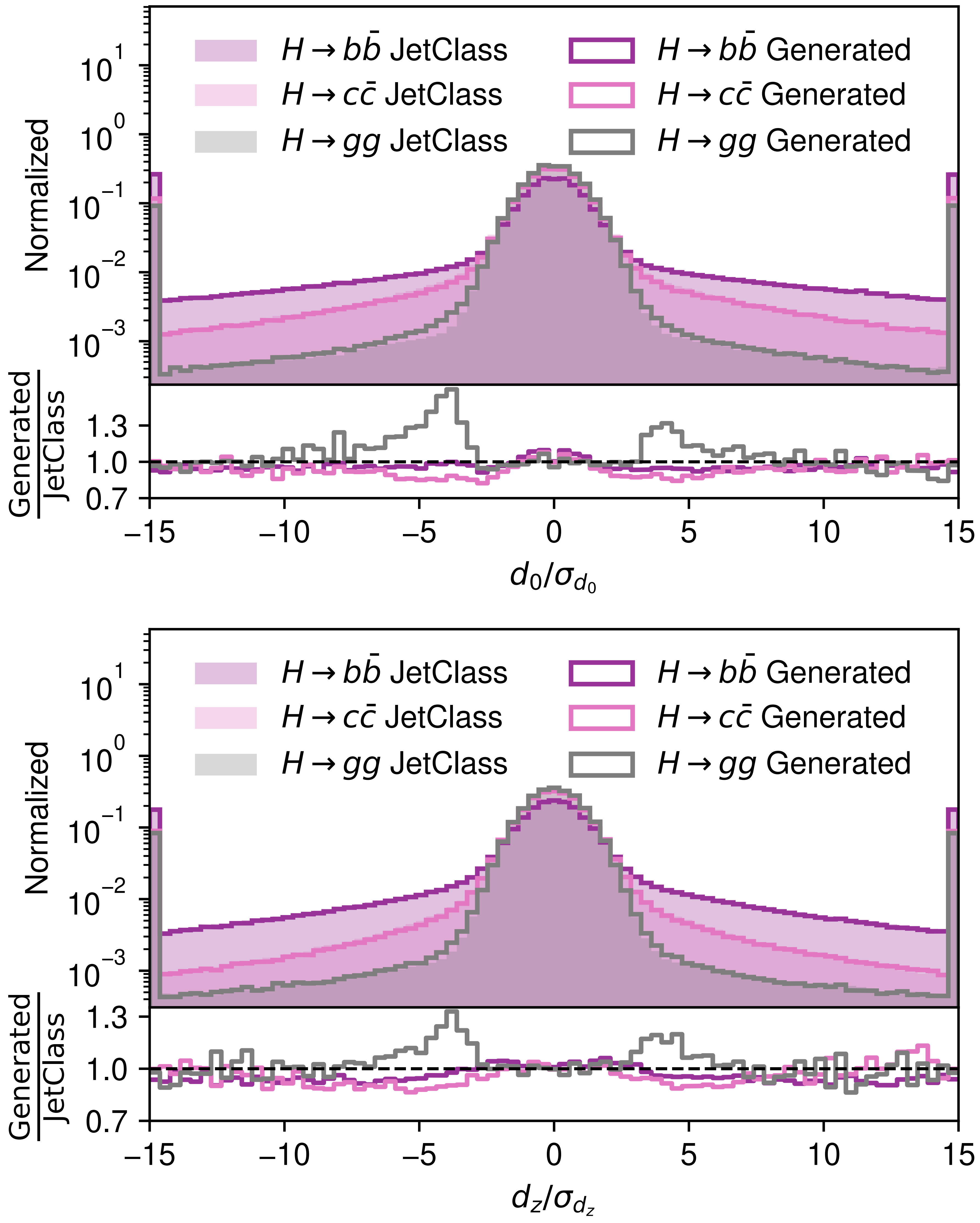
	$\mathrm{KL}^{p_{\mathrm{T}}^{\mathrm{rel}}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\eta^{\mathrm{rel}}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\phi^{\mathrm{rel}}}$ $(\times 10^{-3})$
Truth (q/g)	0.06 ± 0.01	0.08 ± 0.01	0.07 ± 0.01
EPiC-FM (q/g)	0.073 ± 0.007	0.13 ± 0.02	0.12 ± 0.01
Truth $(t \to bqq')$	0.043 ± 0.007	0.060 ± 0.005	0.057 ± 0.007
EPiC-FM $(t \to bqq')$	0.08 ± 0.01	0.09 ± 0.01	0.10 ± 0.01



Trajectory displacement modeling

- Impact parameter (IP) d_0 and d_z:
  closest approach to the interaction point
- IP of neutral particles is set to 0
- Depending on jet type, we expect wide / narrow IP distribution
  - $H \rightarrow b\bar{b}$ and $H \rightarrow c\bar{c}$: hadrons with long lifetime → large IP
  - $H \rightarrow gg$: tracks start very close to the interaction point → small IP



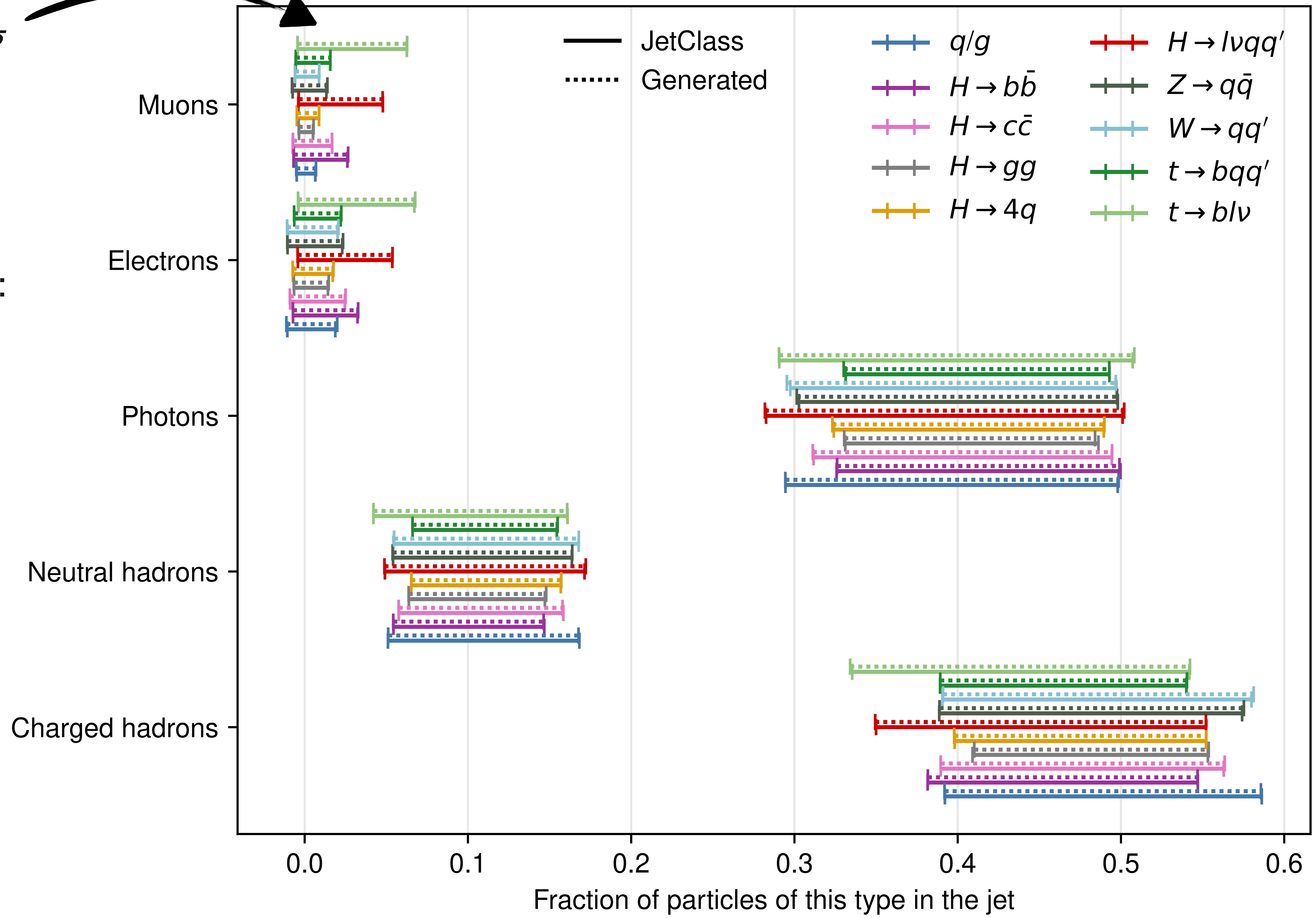


Particle-ID features

- **Post-processing** for discrete features:
  - Particle-ID: `argmax`
  - Charge: rounded

- Particle-ID is **very well modeled**

Mean values $\pm 1\sigma$



Jet mass and jet substructure

Mass distributions:

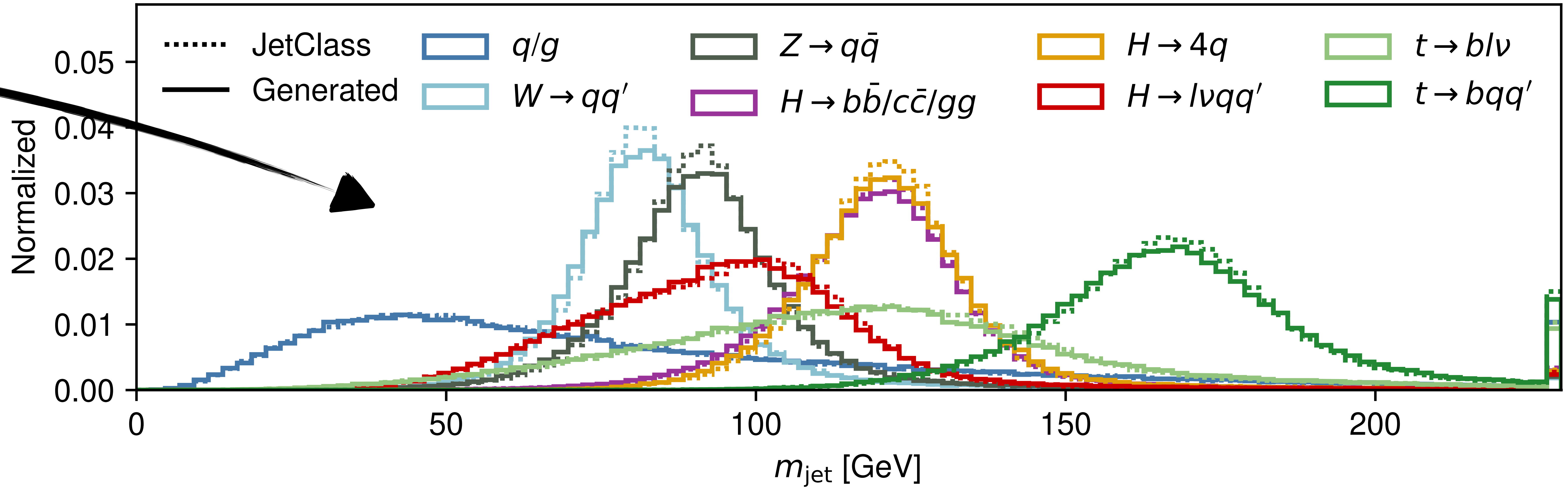
- Very good agreement!

We know:

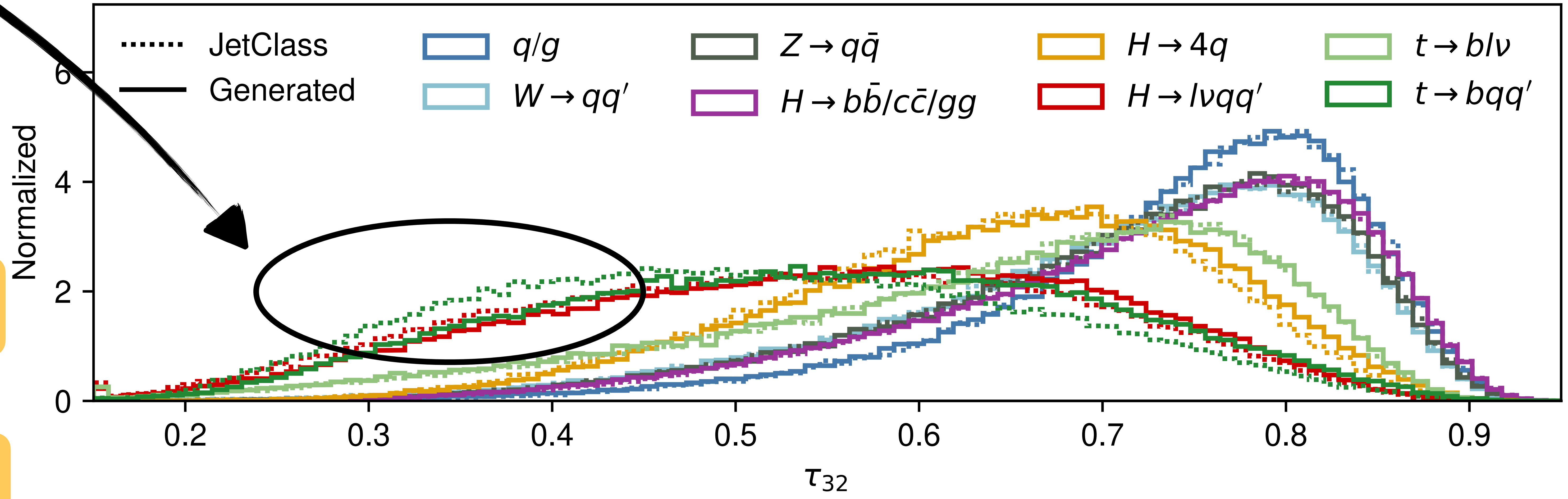
subjettiness ratio τ_{32} hard to model

- Good agreement for most jet types
- Largest disagreement for $t \rightarrow bqq'$ and $H \rightarrow \ell\nu qq'$ jets

→ model struggles a bit with 3-prong jets

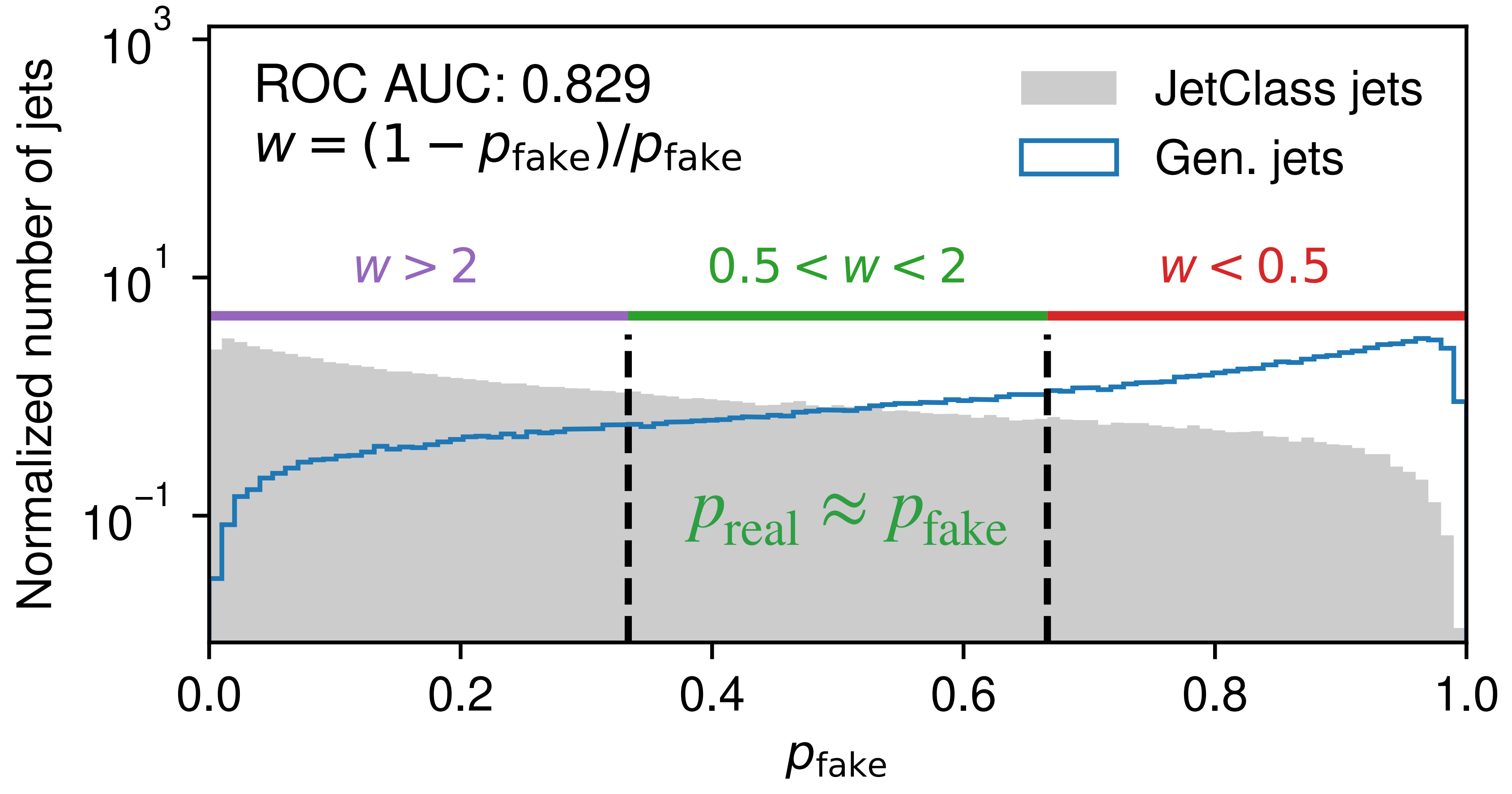


	$\mathrm{KL}^{m_{\mathrm{jet}}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\tau_{21}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\tau_{32}}$ $(\times 10^{-3})$
Truth (q/g)	2.0 ± 0.2	1.9 ± 0.3	2.0 ± 0.3
EPiC-FM (q/g)	2.0 ± 0.3	4.3 ± 0.5	2.3 ± 0.2
Truth $(H \to b\bar{b})$	1.9 ± 0.2	2.0 ± 0.2	2.1 ± 0.2
EPiC-FM $(H \to b\bar{b})$	3.3 ± 0.5	3.6 ± 0.6	2.7 ± 0.4
Truth $(H \to \ell\nu qq')$	1.9 ± 0.3	1.9 ± 0.3	2.0 ± 0.2
EPiC-FM $(H \to \ell\nu qq')$	3.5 ± 0.5	3.5 ± 0.5	9.1 ± 0.9
Truth $(Z \to q\bar{q})$	2.1 ± 0.2	2.0 ± 0.3	1.9 ± 0.2
EPiC-FM $(Z \to q\bar{q})$	4.8 ± 0.5	3.5 ± 0.5	2.0 ± 0.2
Truth $(t \to bqq')$	1.9 ± 0.2	2.0 ± 0.3	2.1 ± 0.2
EPiC-FM $(t \to bqq')$	4.7 ± 0.6	7.9 ± 0.9	30 ± 1



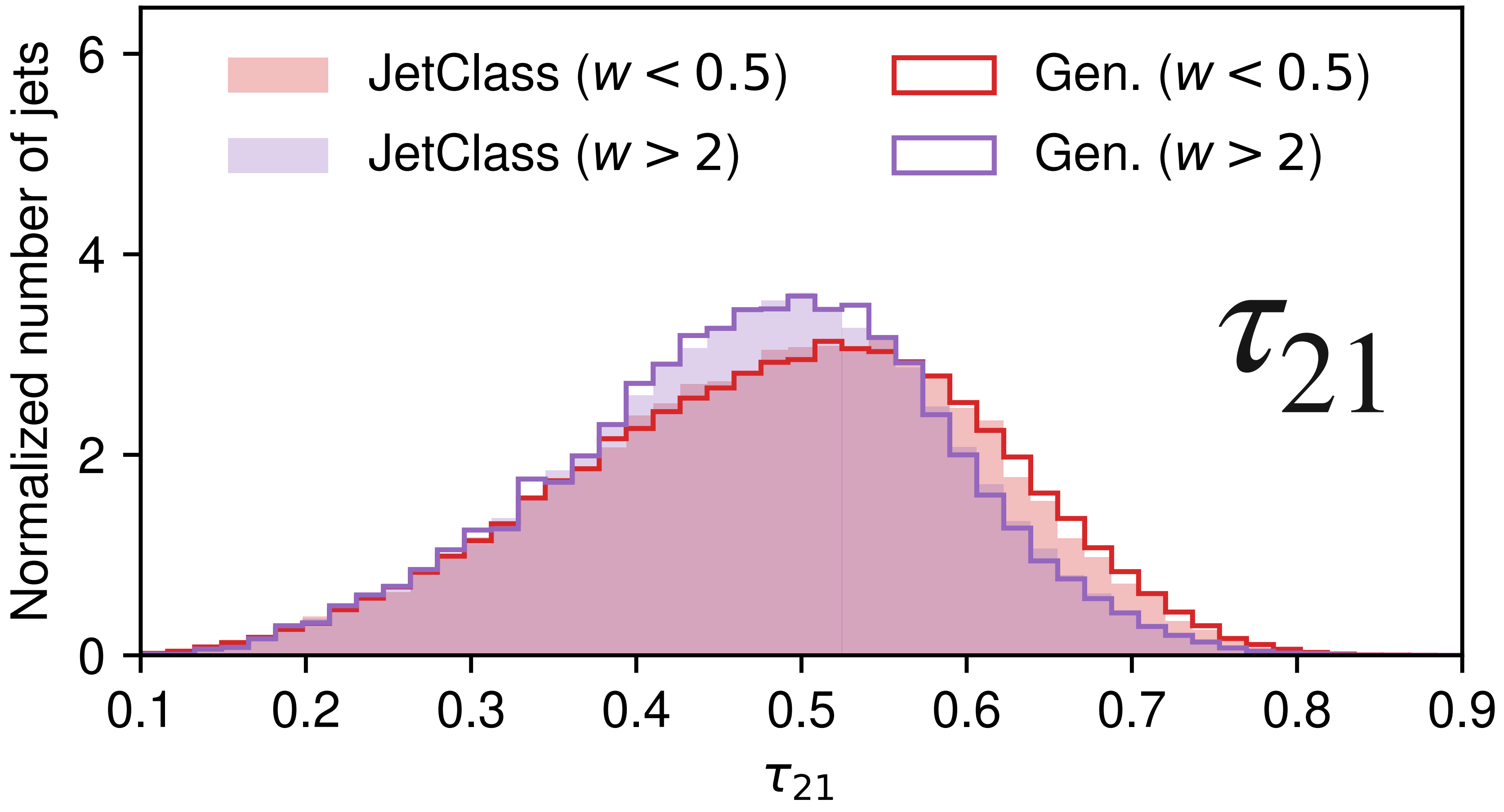
Investigate $t \rightarrow bqq'$ jets with a classifier test

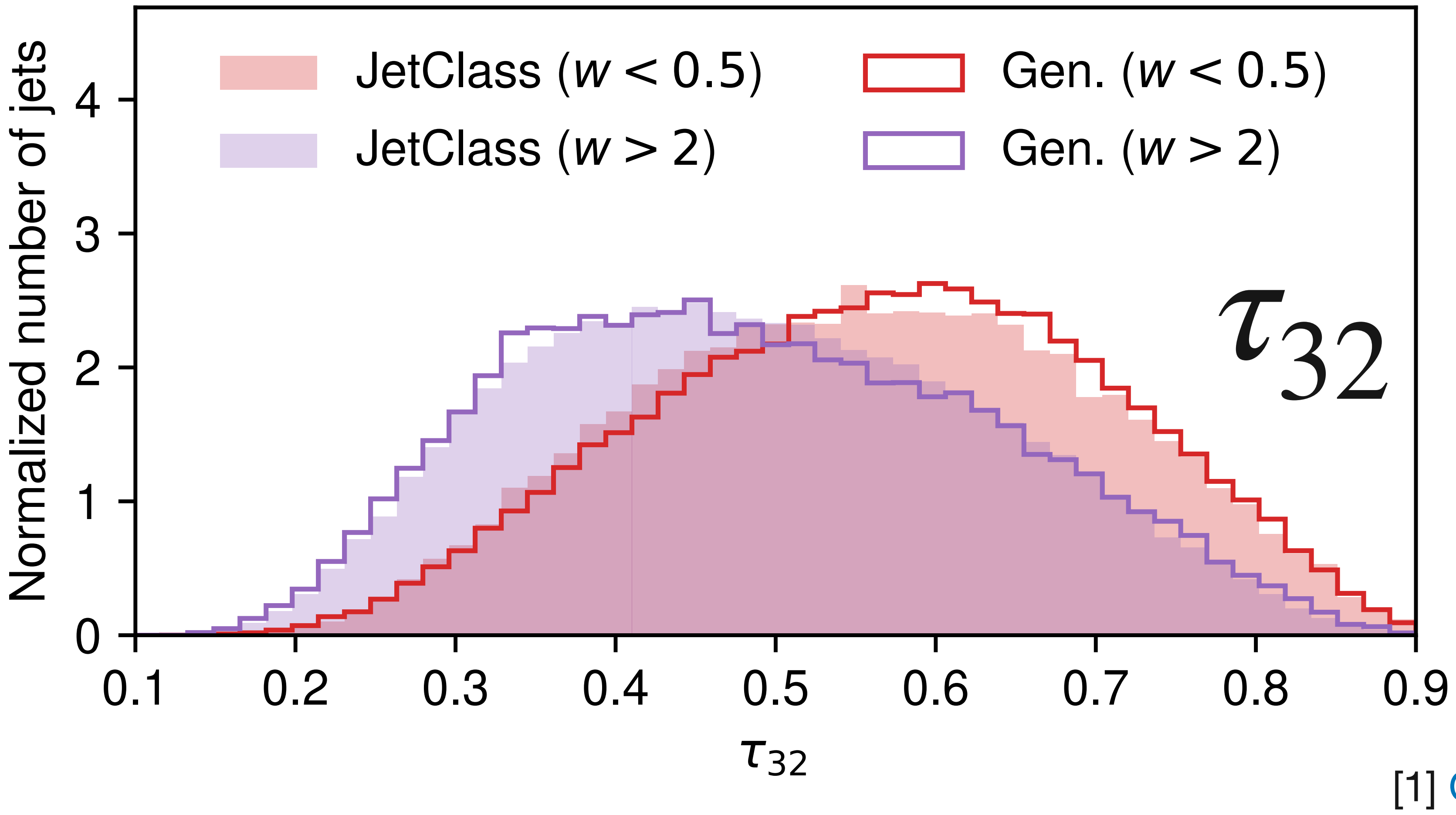
- Train **ParT-kin** [1] to distinguish between real and fake jets ($t \rightarrow bqq'$ only)
- Split into three regions: $p_{\mathrm{real}} \gg p_{\mathrm{fake}}$, $p_{\mathrm{real}} \approx p_{\mathrm{fake}}$, $p_{\mathrm{real}} \ll p_{\mathrm{fake}}$



- Substructure for fake jets from different regions:
  - Jets of which classifier thinks they are fake: larger τ_{32}
  - Jets of which classifier thinks they are real: smaller τ_{32}

→ Model seems to struggle with getting the 3-prong like structure right





[1] Qu et al. (2022) "Particle Transformer for Jet Tagging"

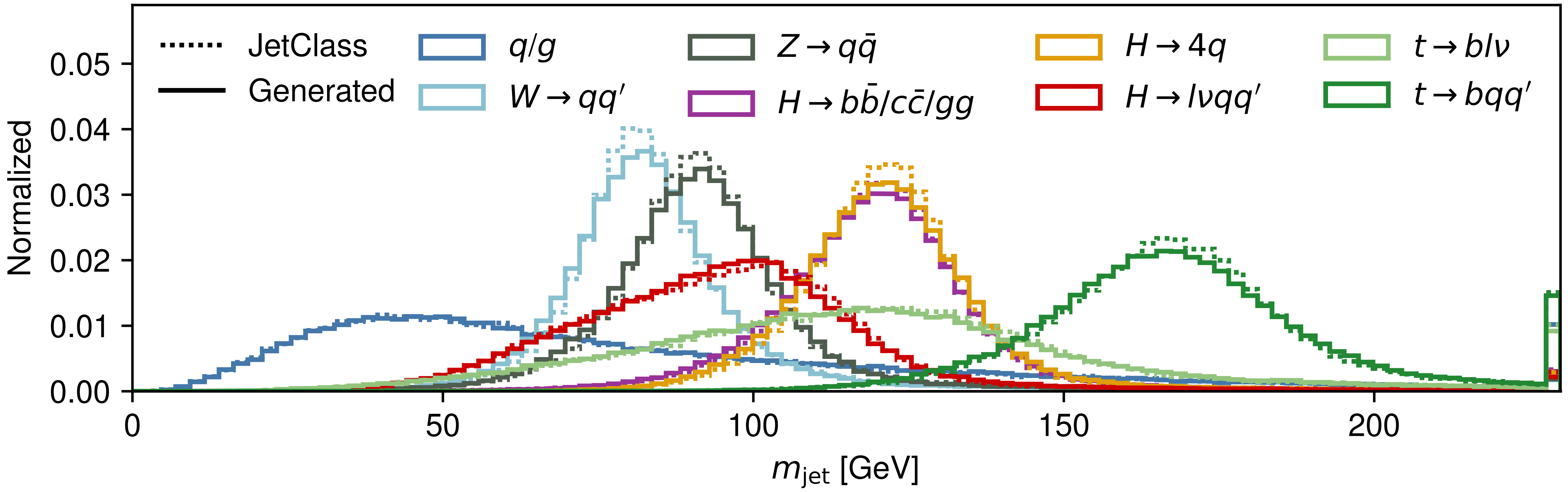
Summary

- **JetClass dataset** [1] **opens new possibilities** for generative modeling prototyping

- **Our flow-matching-based model accurately generates jets from the JetClass dataset**
    - One model for all 10 jet types
    - Generated jet constituents include Particle-ID and trajectory displacement information
    - pre-print available at arXiv:2312.00123

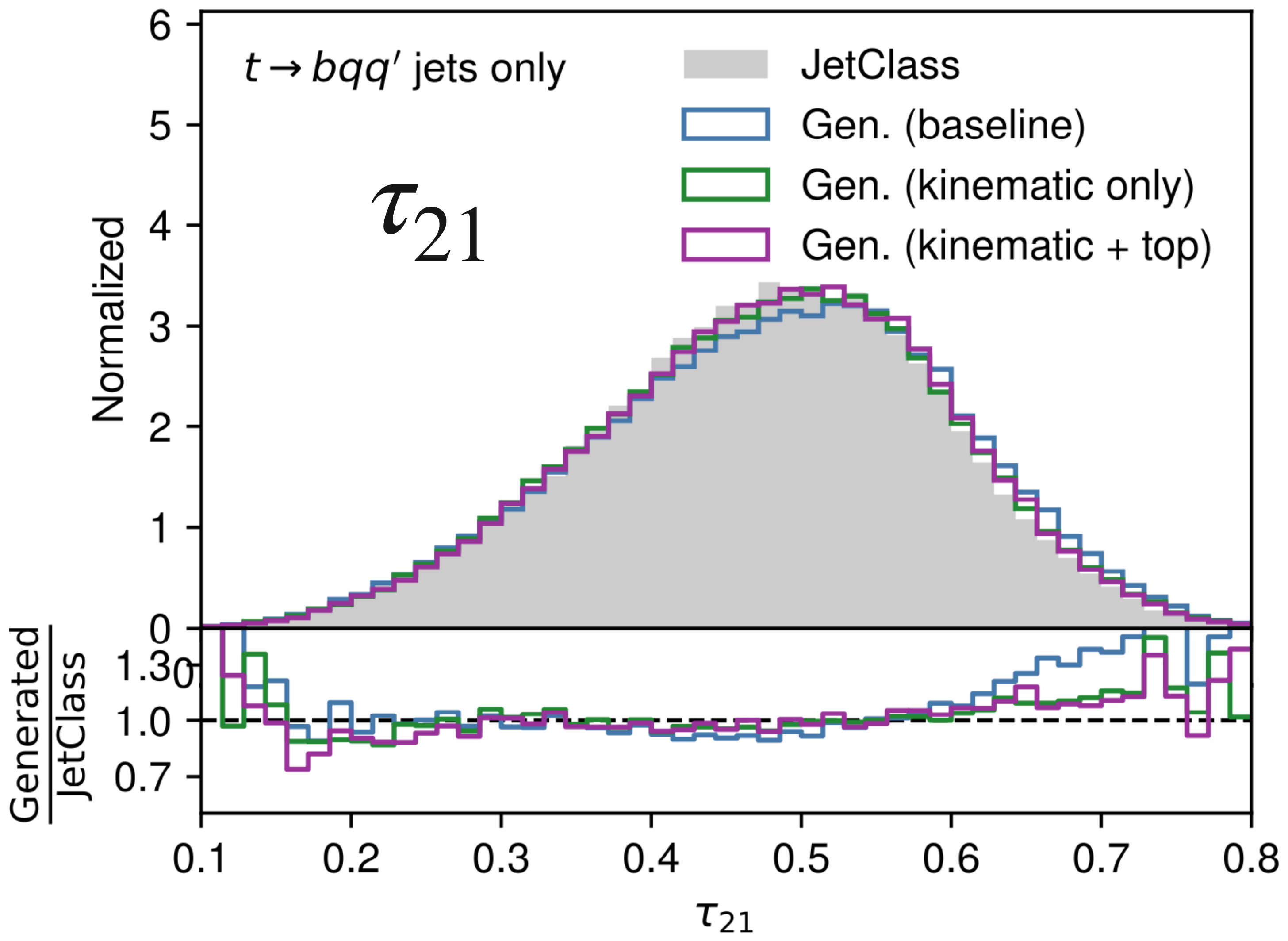
- Using **low-level features in anomaly detection** can drastically **increase significance improvement** [2]

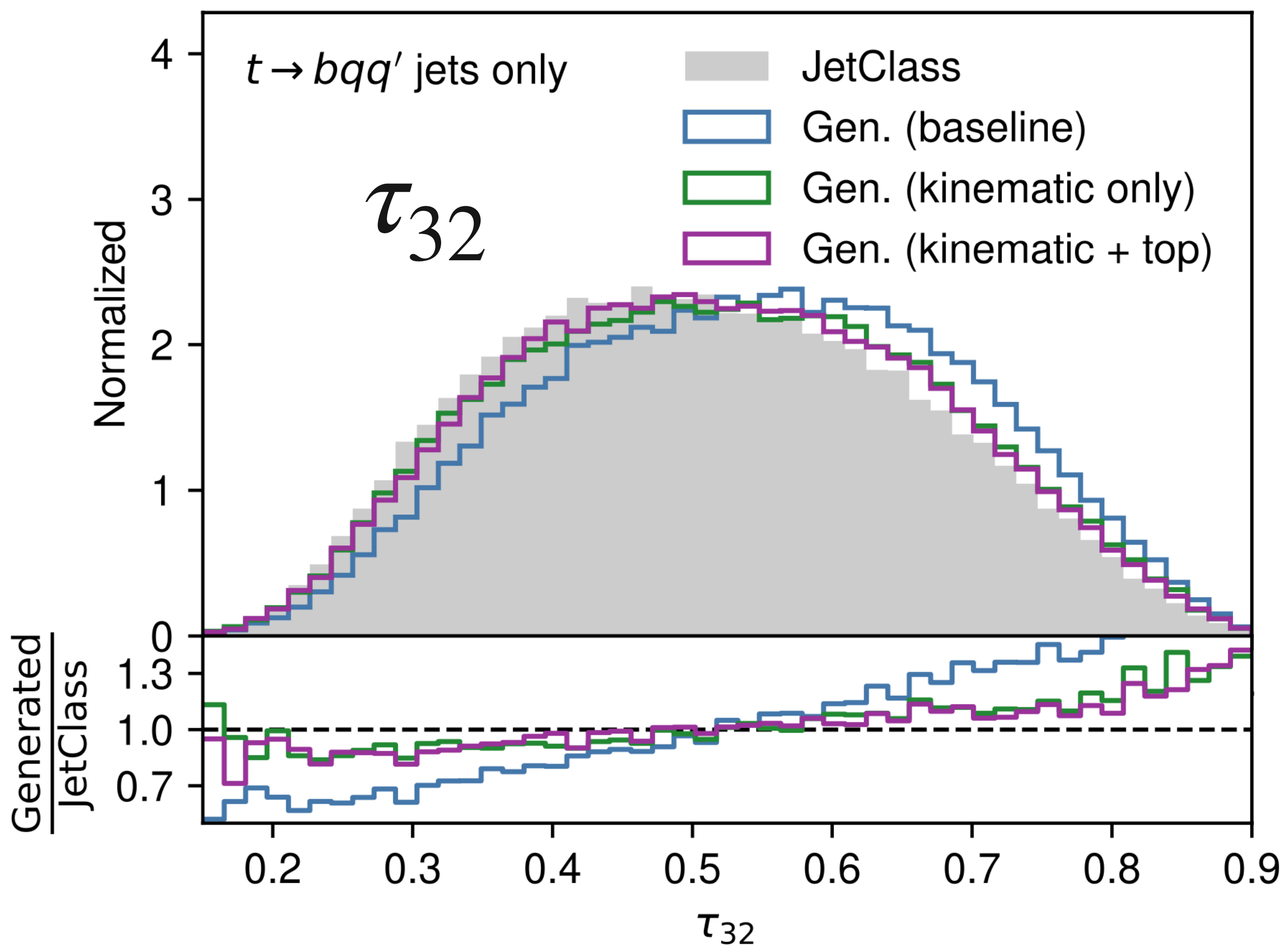
- Having **PID and trajectory displacement in the generated jets:**
  → sensitive to **final states with b-flavor and potentially long-lived final states** [3]



[1] Qu et al. (2022) "JetClass: A Large-Scale Dataset for Deep Learning in Jet Physics"
[2] Buhmann et al. (2023) “Full Phase Space Resonant Anomaly Detection”
[3] Alimena et al. (2020) “Searching for long-lived particles beyond the Standard Model at the Large Hadron Collider”

Improved models for $t \to bqq'$ jets





	$W_1^{\tau_{21}}$ $(\cdot 10^{-3})$	$W_1^{\tau_{32}}$ $(\cdot 10^{-3})$
Truth	1.7 ± 0.4	1.7 ± 0.5
Gen. (all features)	7.9 ± 0.5	35.4 ± 0.8
Gen. (kin. features)	3.6 ± 0.7	13.0 ± 0.7
Gen. (kin. features + top only)	4.8 ± 0.7	11.1 ± 0.9

- Trained two modified versions of our model:
  1. Generating kinematic features only (η^{rel}, ϕ^{rel}, $p_{\mathrm{T}}^{\mathrm{rel}}$)
  2. Generating kinematic features only (η^{rel}, ϕ^{rel}, $p_{\mathrm{T}}^{\mathrm{rel}}$) + trained on $t \to bqq'$ only
- ***Kinematic features only improves modeling of jet substructure***
- Training on ***top jets only increases agreement in τ_{32} but decreases agreement in τ_{21}***

  → training on different jet types simultaneously seems to help

Training and inference

This represents $\hat{x} = \frac{dx}{dt}$

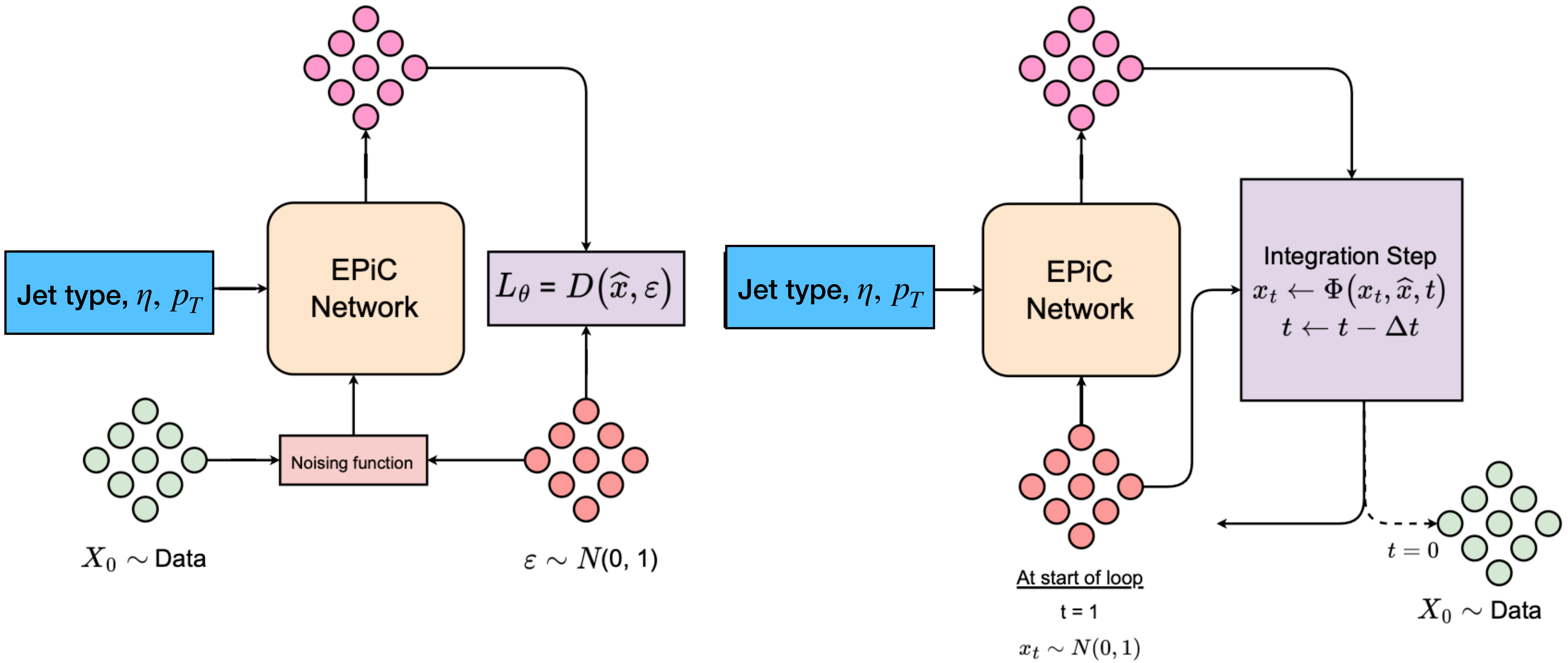


Fig. 1: Schematic overview of the EPiC-JeDi and EPiC-FM training (left) and generation (right) pipeline.

Image adapted from Buhmann et al. (2023) “EPiC-ly Fast Particle Cloud Generation with Flow-Matching and Diffusion”

N-subjettiness calculation and visualization

- Cluster into N exclusive subjets
- Calculate p_T -weighted sum of distances to closest subjet axis

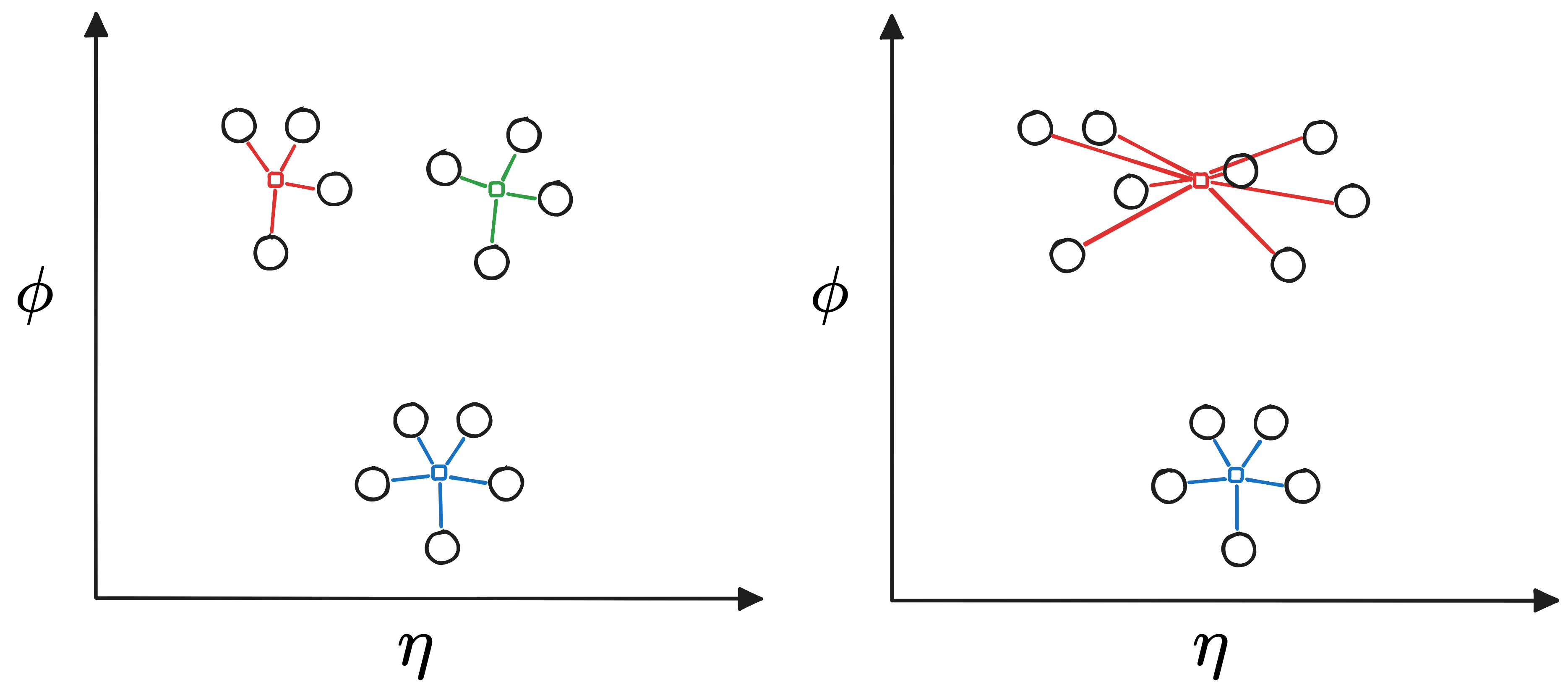
$$\tau_2 = \frac{1}{d_0} \sum_i p_{T,i} \min\{\Delta R_{1,i}, \Delta R_{2,i}\}$$

$$\tau_3 = \frac{1}{d_0} \sum_i p_{T,i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, \Delta R_{3,i}\}$$

- We then look at ratio $\tau_{32} = \tau_3/\tau_2$
- **3-prong jets:** expect τ_{32} to **peak at small value**
- **2-prong jets:** expect τ_{32} to **peak at larger value**

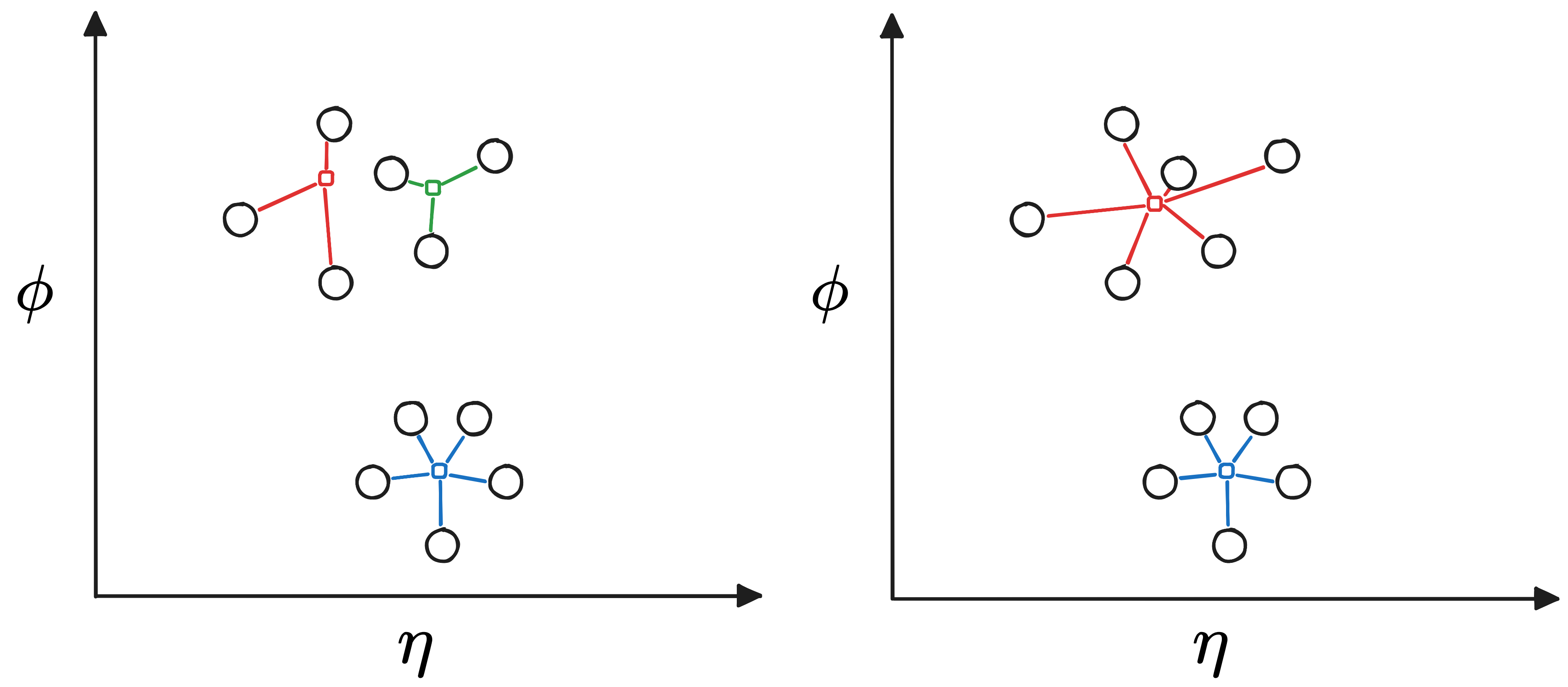
$t \rightarrow bqq'$

Here, τ_3 will be much smaller than τ_2

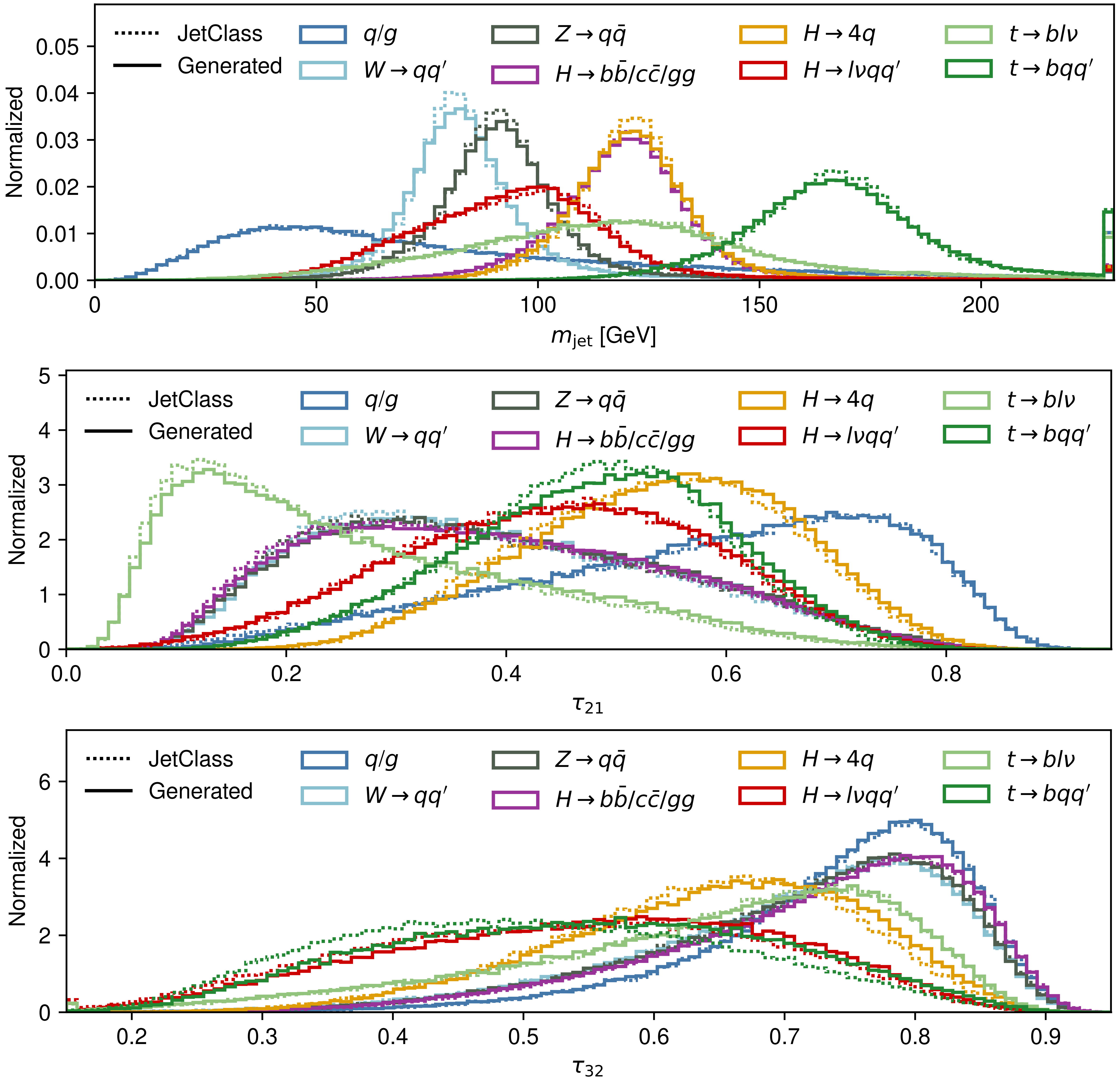


$H \rightarrow b\bar{b}$

Here, τ_3 will be only slightly smaller than τ_2



Jet substructure



Feature definitions

TABLE I: Features used for the studies in this paper. All features are either taken directly from the JetClass dataset or calculated from existing entries in the JetClass dataset. The model is conditioned on the jet features (first block) during generation. The remaining features are the constituent features our model is trained to generate.

Category	Variable	Definition
Jet features	Jet type	Indicating the type of the jet (i.e. $t \to bqq'$, q/g, ...)
	n_{const}	Number of constituents in the jet
	$p_{\mathrm{T}}^{\text{jet}}$	Transverse momentum of the jet
	η^{jet}	Pseudorapidity of the jet
Kinematics	η^{rel}	Difference in pseudorapidity η between the constituent and the jet axis
	ϕ^{rel}	Difference in azimuthal angle ϕ between the constituent and the jet axis
	$p_{\mathrm{T}}^{\text{rel}}$	Fraction of the constituent p_{T} and the jet p_{T}
Trajectory displacement	d_0	Transverse impact parameter value
	d_z	Longitudinal impact parameter value
	σ_{d_0}	Error of measured transverse impact parameter
	σ_{d_z}	Error of measured longitudinal impact parameter
Particle identification	charge	Electric charge of the particle
	isChargedHadron	Flag if the particle is a charged hadron (`|pid|==211 or 321 or 2212`)
	isNeutralHadron	Flag if the particle is a neutral hadron (`|pid|==130 or 2112 or 0`)
	isPhoton	Flag if the particle is a photon (`pid==22`)
	isElectron	Flag if the particle is an electron (`|pid|==11`)
	isMuon	Flag if the particle is a muon (`|pid|==13`)

Metrics for jet-level features

TABLE III: Kullback-Leibler divergence of jet-level observables with respect to the JetClass dataset.

	$\mathrm{KL}^{m_{\mathrm{jet}}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\tau_{21}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\tau_{32}}$ $(\times 10^{-3})$	KL^{D_2} $(\times 10^{-3})$
Truth (q/g)	2.0 ± 0.2	1.9 ± 0.3	2.0 ± 0.3	1.9 ± 0.4
EPiC-FM (q/g)	2.0 ± 0.3	4.3 ± 0.5	2.3 ± 0.2	4.7 ± 0.4
Truth $(H \to b\bar{b})$	1.9 ± 0.2	2.0 ± 0.2	2.1 ± 0.2	2.0 ± 0.2
EPiC-FM $(H \to b\bar{b})$	3.3 ± 0.5	3.6 ± 0.6	2.7 ± 0.4	3.3 ± 0.5
Truth $(H \to c\bar{c})$	2.0 ± 0.3	2.0 ± 0.2	2.2 ± 0.3	1.9 ± 0.2
EPiC-FM $(H \to c\bar{c})$	4.8 ± 0.5	6.4 ± 0.6	3.1 ± 0.5	3.7 ± 0.5
Truth $(H \to gg)$	2.0 ± 0.2	1.8 ± 0.3	1.9 ± 0.2	2.0 ± 0.2
EPiC-FM $(H \to gg)$	3.8 ± 0.4	3.8 ± 0.6	2.7 ± 0.4	3.2 ± 0.5
Truth $(H \to 4q)$	2.2 ± 0.4	2.1 ± 0.3	2.0 ± 0.3	1.9 ± 0.3
EPiC-FM $(H \to 4q)$	5.2 ± 0.5	5.1 ± 0.5	7.8 ± 0.7	3.2 ± 0.5
Truth $(H \to \ell\nu qq')$	1.9 ± 0.3	1.9 ± 0.3	2.0 ± 0.2	2.0 ± 0.2
EPiC-FM $(H \to \ell\nu qq')$	3.5 ± 0.5	3.5 ± 0.5	9.1 ± 0.9	2.3 ± 0.3
Truth $(Z \to q\bar{q})$	2.1 ± 0.2	2.0 ± 0.3	1.9 ± 0.2	2.0 ± 0.2
EPiC-FM $(Z \to q\bar{q})$	4.8 ± 0.5	3.5 ± 0.5	2.0 ± 0.2	2.4 ± 0.2
Truth $(W \to qq')$	2.1 ± 0.4	2.0 ± 0.2	2.0 ± 0.2	2.0 ± 0.2
EPiC-FM $(W \to qq')$	5.8 ± 0.5	4.2 ± 0.6	2.6 ± 0.4	3.1 ± 0.5
Truth $(t \to bqq')$	1.9 ± 0.2	2.0 ± 0.3	2.1 ± 0.2	1.9 ± 0.3
EPiC-FM $(t \to bqq')$	4.7 ± 0.6	7.9 ± 0.9	30 ± 1	3.0 ± 0.4
Truth $(t \to b\ell\nu)$	2.0 ± 0.1	1.9 ± 0.2	2.0 ± 0.2	2.0 ± 0.3
EPiC-FM $(t \to b\ell\nu)$	3.4 ± 0.2	5.8 ± 0.5	2.6 ± 0.4	3.8 ± 0.5

Metrics for constituent-level features

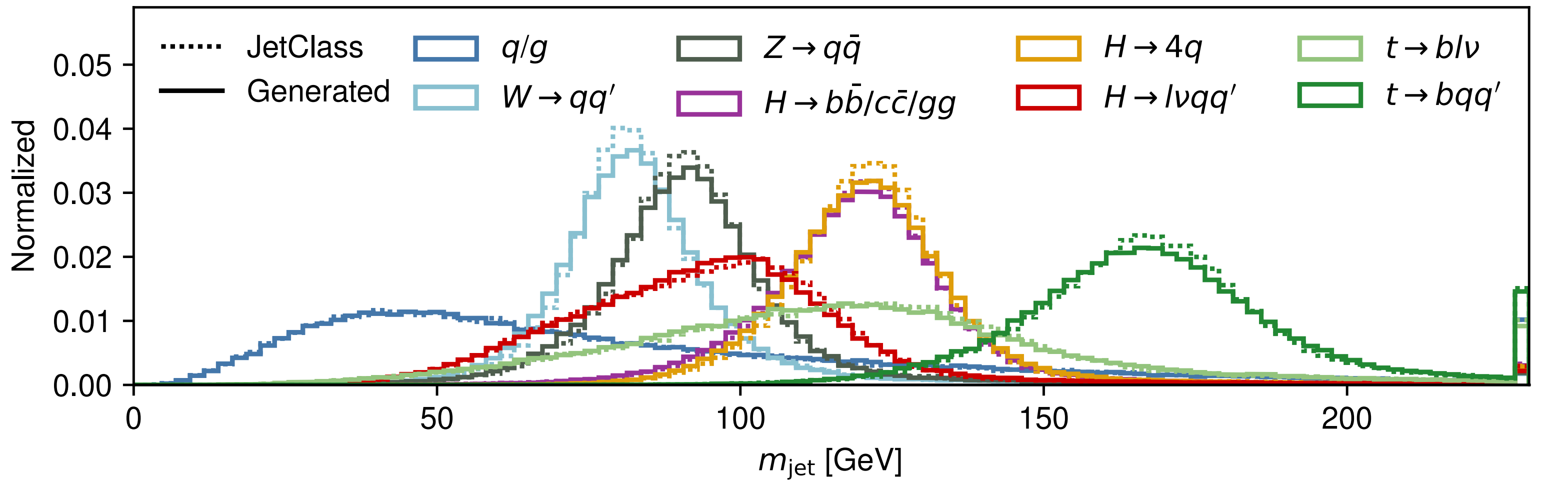
TABLE II: Kullback-Leibler divergence of constituent-level features with respect to the JetClass dataset.

	$\mathrm{KL}^{p_\mathrm{T}^\mathrm{rel}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\eta^\mathrm{rel}}$ $(\times 10^{-3})$	$\mathrm{KL}^{\phi^\mathrm{rel}}$ $(\times 10^{-3})$	$\mathrm{KL}^{d_0/\sigma_{d_0}}$ $(\times 10^{-3})$	$\mathrm{KL}^{d_z/\sigma_{d_z}}$ $(\times 10^{-3})$
Truth (q/g)	0.06 ± 0.01	0.08 ± 0.01	0.07 ± 0.01	0.13 ± 0.01	0.11 ± 0.02
EPiC-FM (q/g)	0.073 ± 0.007	0.13 ± 0.02	0.12 ± 0.01	2.20 ± 0.06	3.72 ± 0.09
Truth $(H \to b\bar{b})$	0.048 ± 0.004	0.07 ± 0.01	0.060 ± 0.008	0.10 ± 0.01	0.09 ± 0.01
EPiC-FM $(H \to b\bar{b})$	0.071 ± 0.008	0.12 ± 0.02	0.12 ± 0.02	2.64 ± 0.09	1.47 ± 0.07
Truth $(H \to c\bar{c})$	0.054 ± 0.007	0.07 ± 0.01	0.08 ± 0.02	0.11 ± 0.02	0.11 ± 0.02
EPiC-FM $(H \to c\bar{c})$	0.078 ± 0.007	0.11 ± 0.02	0.134 ± 0.009	2.8 ± 0.1	1.96 ± 0.08
Truth $(H \to gg)$	0.034 ± 0.002	0.06 ± 0.02	0.053 ± 0.008	0.07 ± 0.01	0.07 ± 0.01
EPiC-FM $(H \to gg)$	0.057 ± 0.009	0.10 ± 0.01	0.10 ± 0.01	2.3 ± 0.1	3.5 ± 0.1
Truth $(H \to 4q)$	0.039 ± 0.005	0.057 ± 0.007	0.05 ± 0.01	0.08 ± 0.01	0.08 ± 0.02
EPiC-FM $(H \to 4q)$	0.057 ± 0.006	0.12 ± 0.02	0.11 ± 0.02	2.26 ± 0.07	3.2 ± 0.1
Truth $(H \to \ell\nu qq')$	0.07 ± 0.01	0.10 ± 0.02	0.10 ± 0.01	0.14 ± 0.01	0.13 ± 0.02
EPiC-FM $(H \to \ell\nu qq')$	0.10 ± 0.01	0.15 ± 0.02	0.137 ± 0.008	2.11 ± 0.08	2.61 ± 0.09
Truth $(Z \to q\bar{q})$	0.061 ± 0.008	0.08 ± 0.01	0.08 ± 0.01	0.14 ± 0.03	0.13 ± 0.01
EPiC-FM $(Z \to q\bar{q})$	0.07 ± 0.01	0.15 ± 0.02	0.13 ± 0.02	3.3 ± 0.2	2.1 ± 0.1
Truth $(W \to qq')$	0.06 ± 0.01	0.088 ± 0.009	0.086 ± 0.008	0.12 ± 0.01	0.13 ± 0.02
EPiC-FM $(W \to qq')$	0.082 ± 0.008	0.14 ± 0.02	0.13 ± 0.02	2.18 ± 0.05	2.6 ± 0.1
Truth $(t \to bqq')$	0.043 ± 0.007	0.060 ± 0.005	0.057 ± 0.007	0.08 ± 0.01	0.08 ± 0.01
EPiC-FM $(t \to bqq')$	0.08 ± 0.01	0.09 ± 0.01	0.10 ± 0.01	1.79 ± 0.07	1.8 ± 0.1
Truth $(t \to b\ell\nu)$	0.07 ± 0.01	0.13 ± 0.02	0.15 ± 0.03	0.15 ± 0.02	0.15 ± 0.02
EPiC-FM $(t \to b\ell\nu)$	0.11 ± 0.02	0.20 ± 0.03	0.19 ± 0.06	1.76 ± 0.09	1.46 ± 0.06

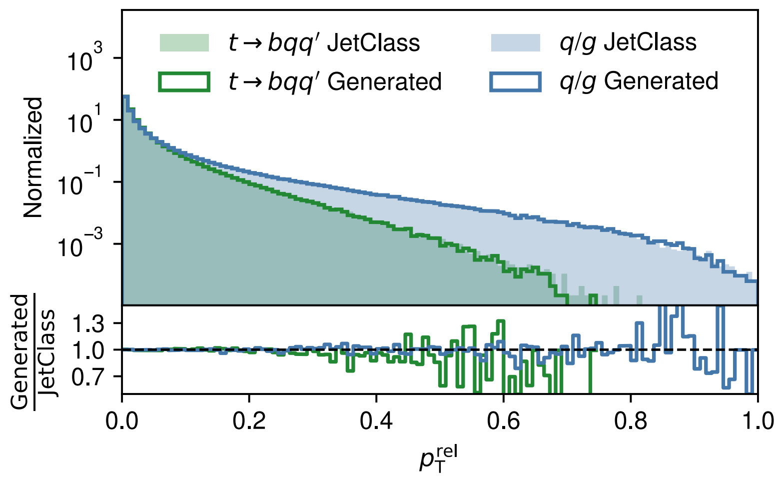
Summary

JetClass dataset [1] opens new possibilities for generative modeling prototyping

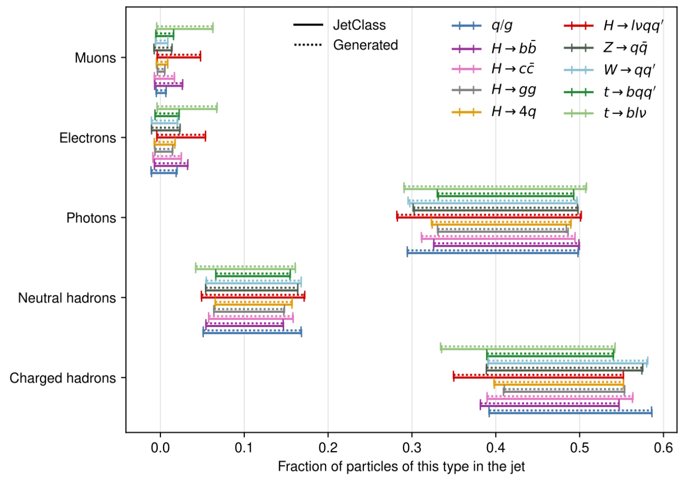
All 10 jet types in one model ✔



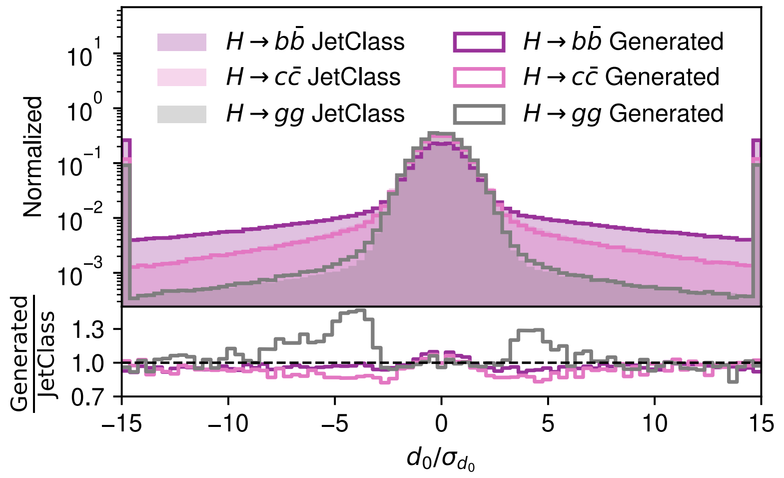
Kinematic features of jet constituents ✔



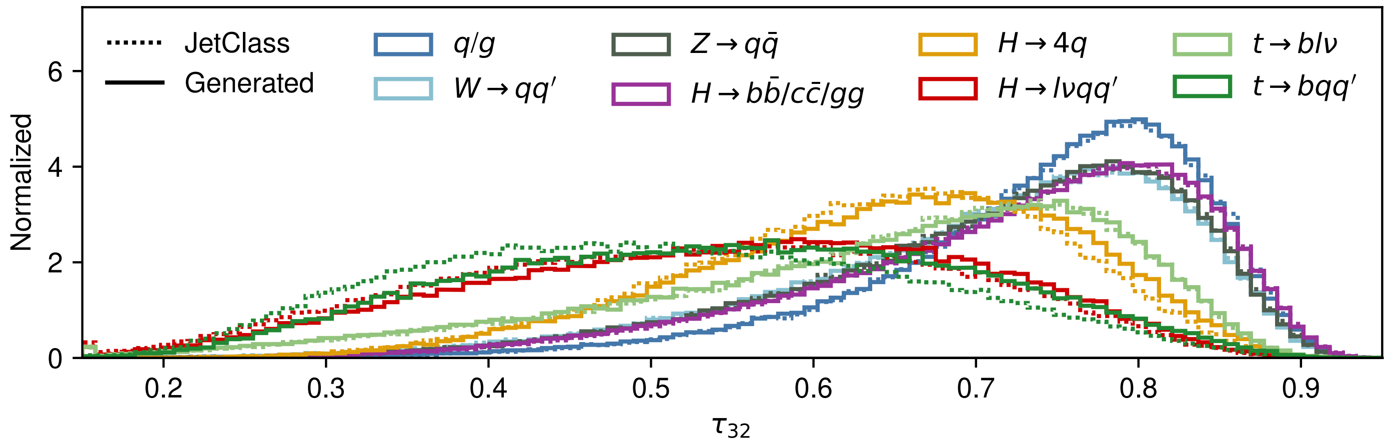
Particle-ID features accurately modeled ✔



Trajectory displacement well modeled ✔



Jet substructure is hard to model perfectly



Used classifier test to investigate results