Generative Models for Particle Physics Data

Erik Buhmann, Thorsten Buss, Sascha Diefenbacher, Engin Eren, Cedric Ewen, Frank Gaede, Gregor Kasieczka, William Korcari*, Anatolii Korol, Katja Krüge, Peter McKeown, Martina Mozzanica, Lennard Rustige, Lorenzo Valente

*william.korcari@uni-hamburg.de

Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE QUANTUM UNIVERSE

DESY.

CMS Compact Muon Solenoid

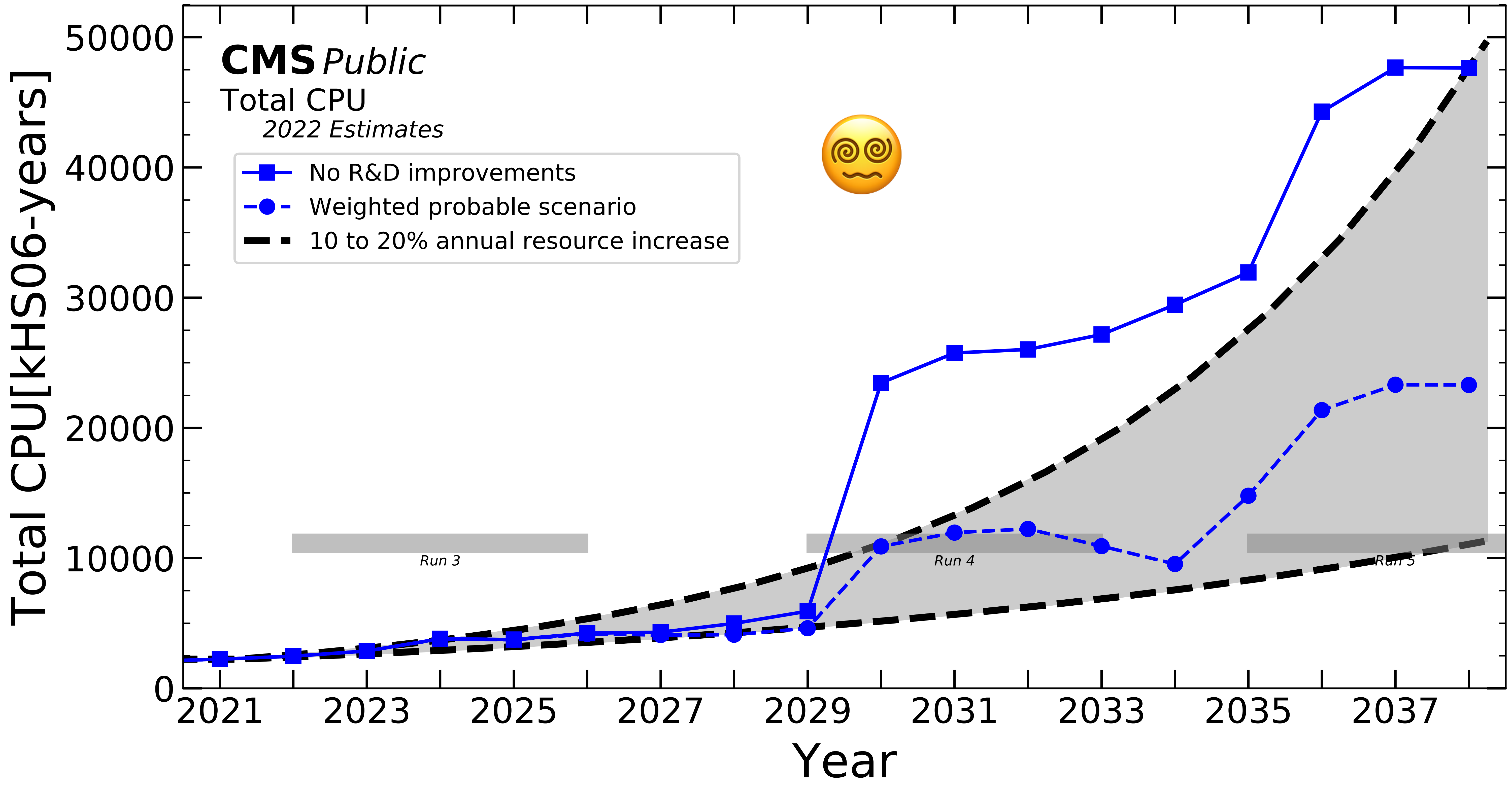
KISS

ErUM-Data IDT

SPONSORED BY THE Federal Ministry of Education and Research

Monte Carlo Simulation in HEP

- Simulation fundamental to compare theory to measurements
- Detector simulation most expensive block of the chain
- Without R&D, increase in computing time exceeds the one in resources



CMS Offline and Computing Public Results

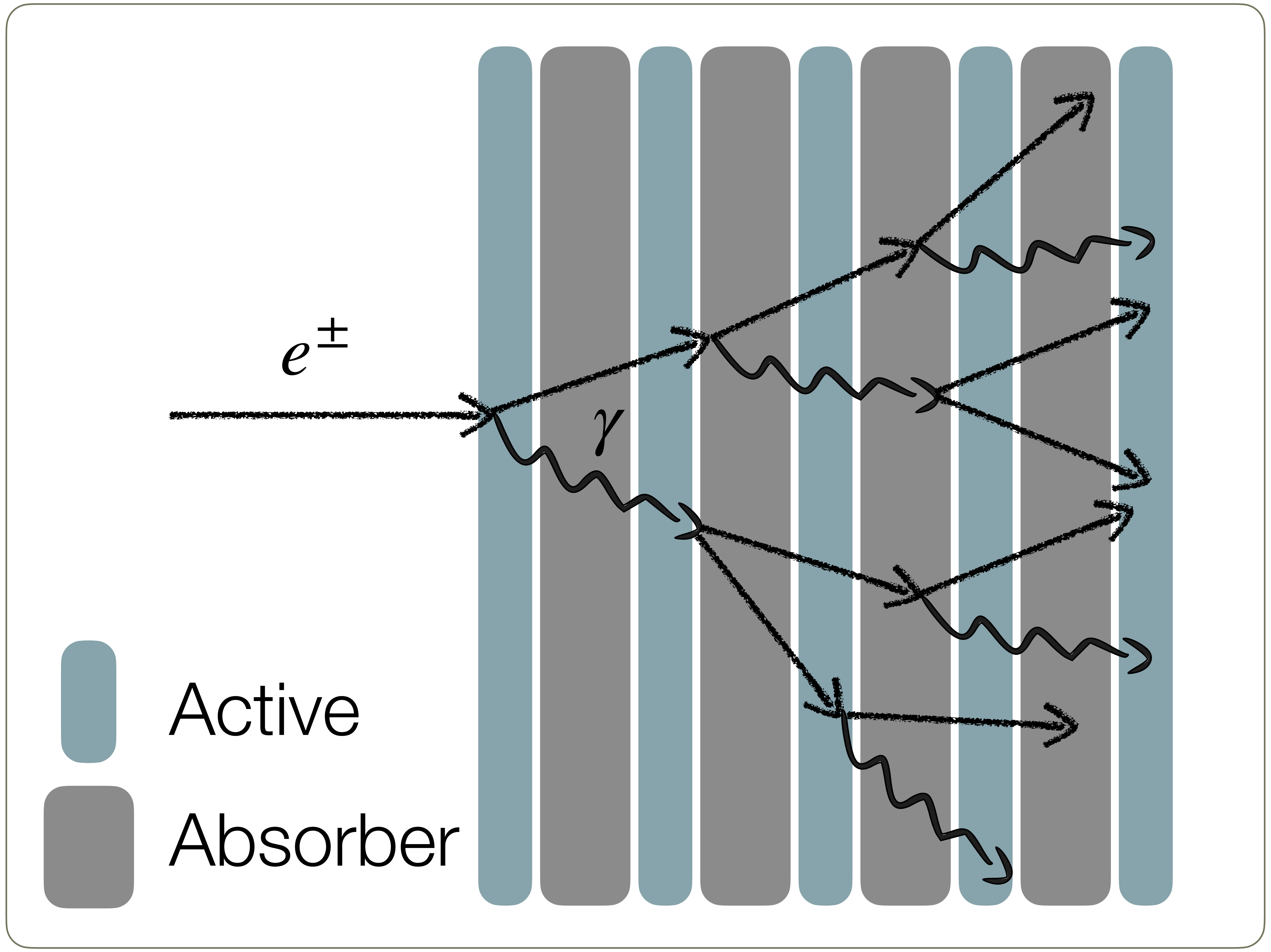
Calorimeters

- A calorimeter is a detector which fully absorbs the particles. The signals produced are a measure for the energy of the particle

- Incoming particle initiate the shower. Each secondary shower deposits energy and produces further particles

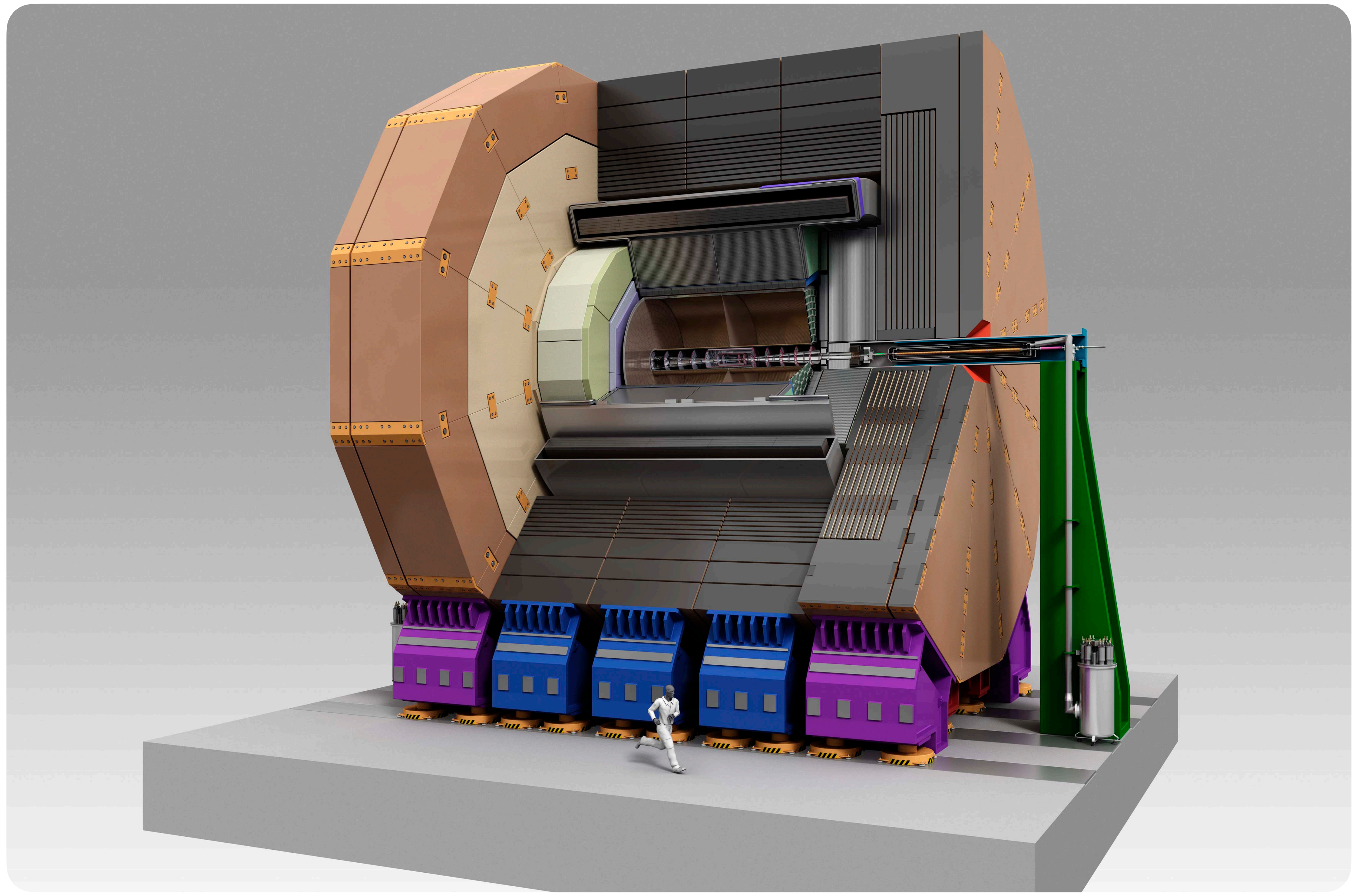
- Sampling calorimeter consists of alternating layers of passive absorbers and active detectors

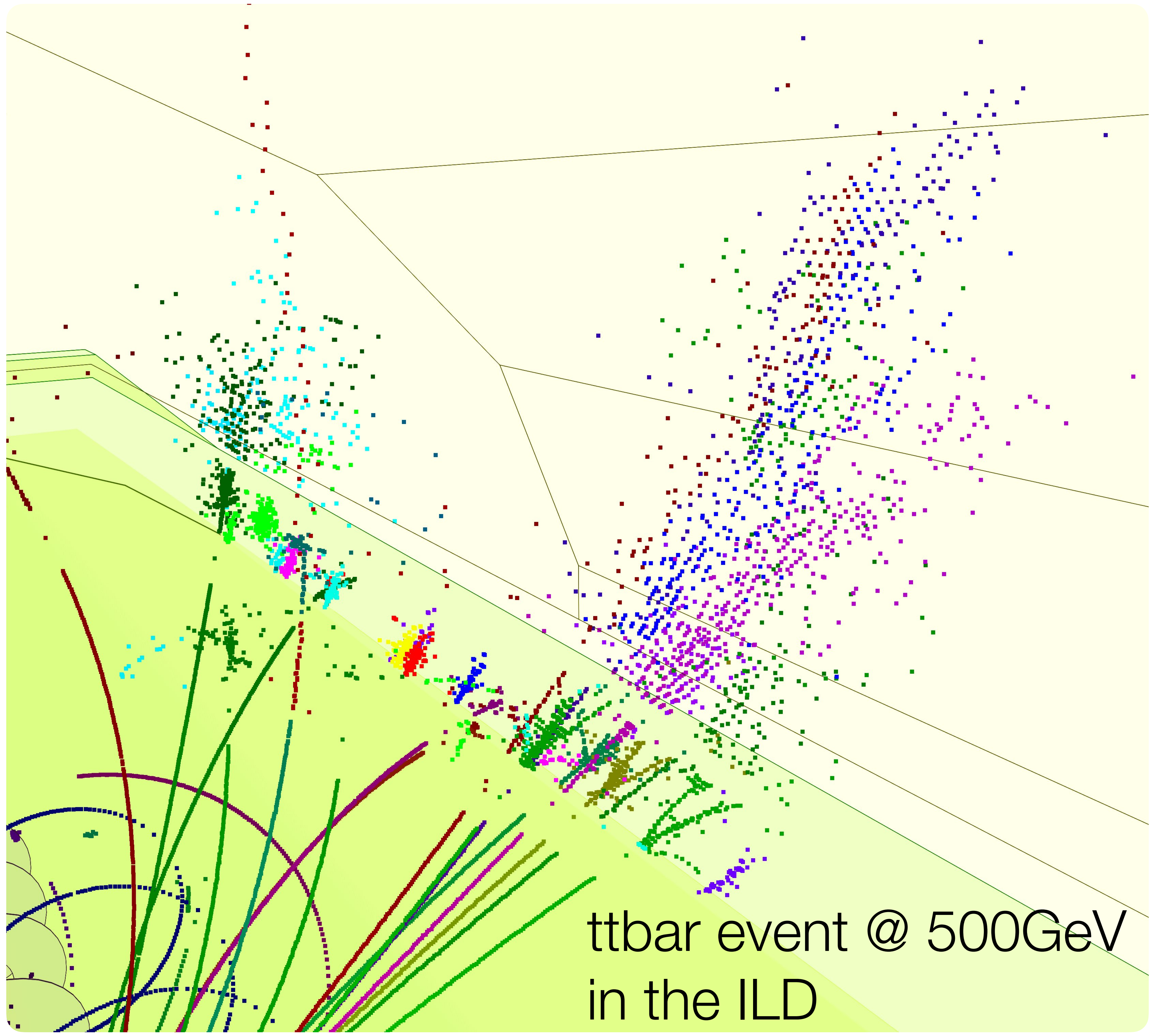
- Electromagnetic and hadronic calorimeters



Schematic view of EM shower in a eCal

Case Study: ILD Calorimeter





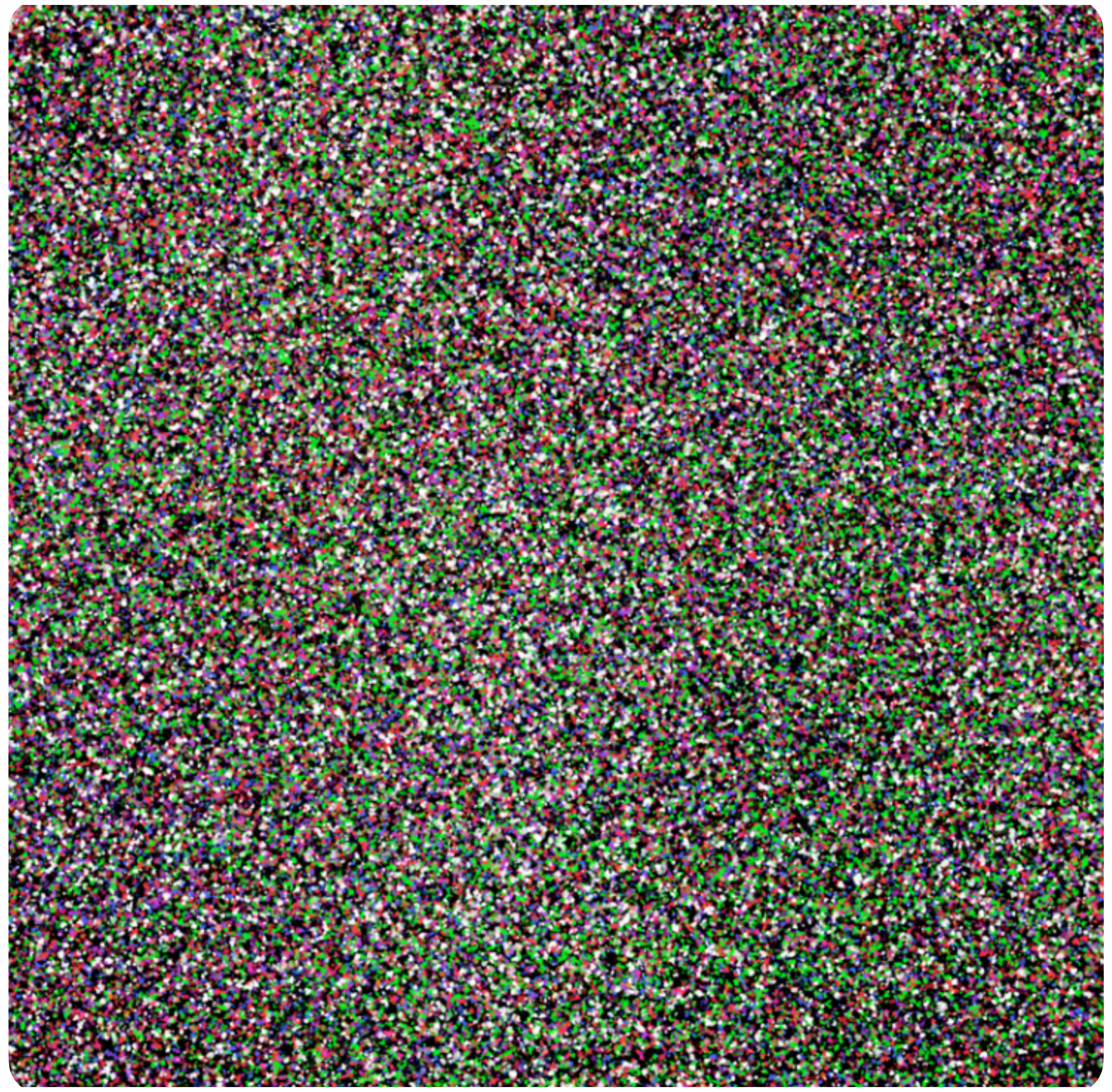
- International Large Detector (ILD) concept for the International Linear Collider (ILC)

- High-Granularity calorimeters:
  - ECal: Si-W - 5mm x 5mm - 30 layers
  - HCal: Sci-Fe - 30mm x 30mm - 49 layers

High granularity ⟶ Need for high fidelity simulation

Detector Simulation in HEP

- Generative models: map noise to structured data
- In our cases: to calorimeter showers



Numerous generative models:

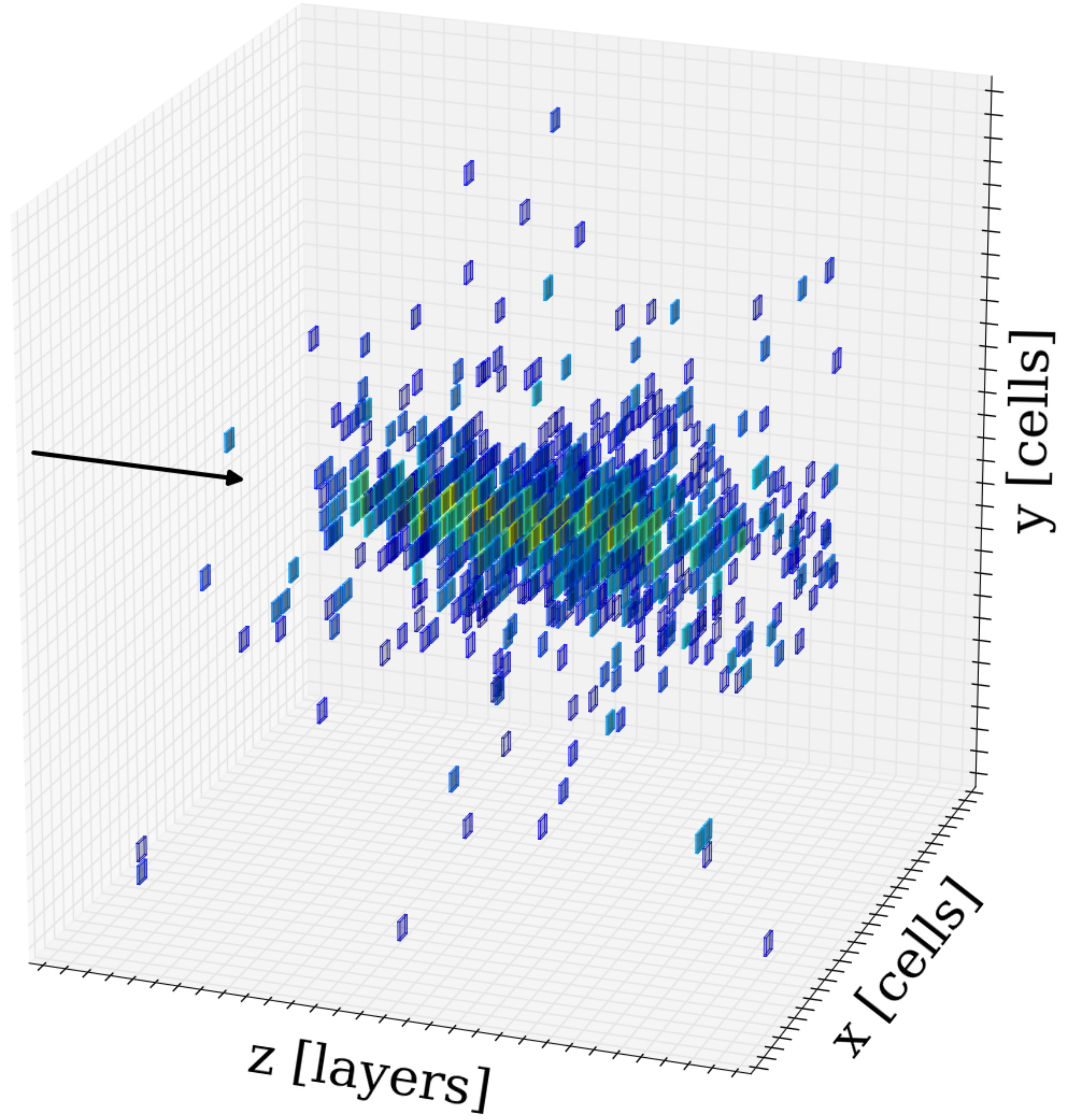
- GANs
- Autoencoders, e.g. BiB-AE
- Flow-based models
- Diffusion models

Data Representation of Showers

- Generative models for calorimeter showers applied to fixed geometries, i.e. CNNs for 3D **images**
- Calorimeter showers are very sparse (only ~ 4% filled pixels)
- **Point clouds** more memory efficient (variable-length, permutation-invariant sets)
- Generation of only non-zero points
- Can use clustered Geant4 steps with higher granularity than sensor size
- Allows for cell-geometry independent model

EM shower as 3D image

27,000 pixels (~ 1,000 non-zero)



EM shower as point cloud

~ 40,000 points

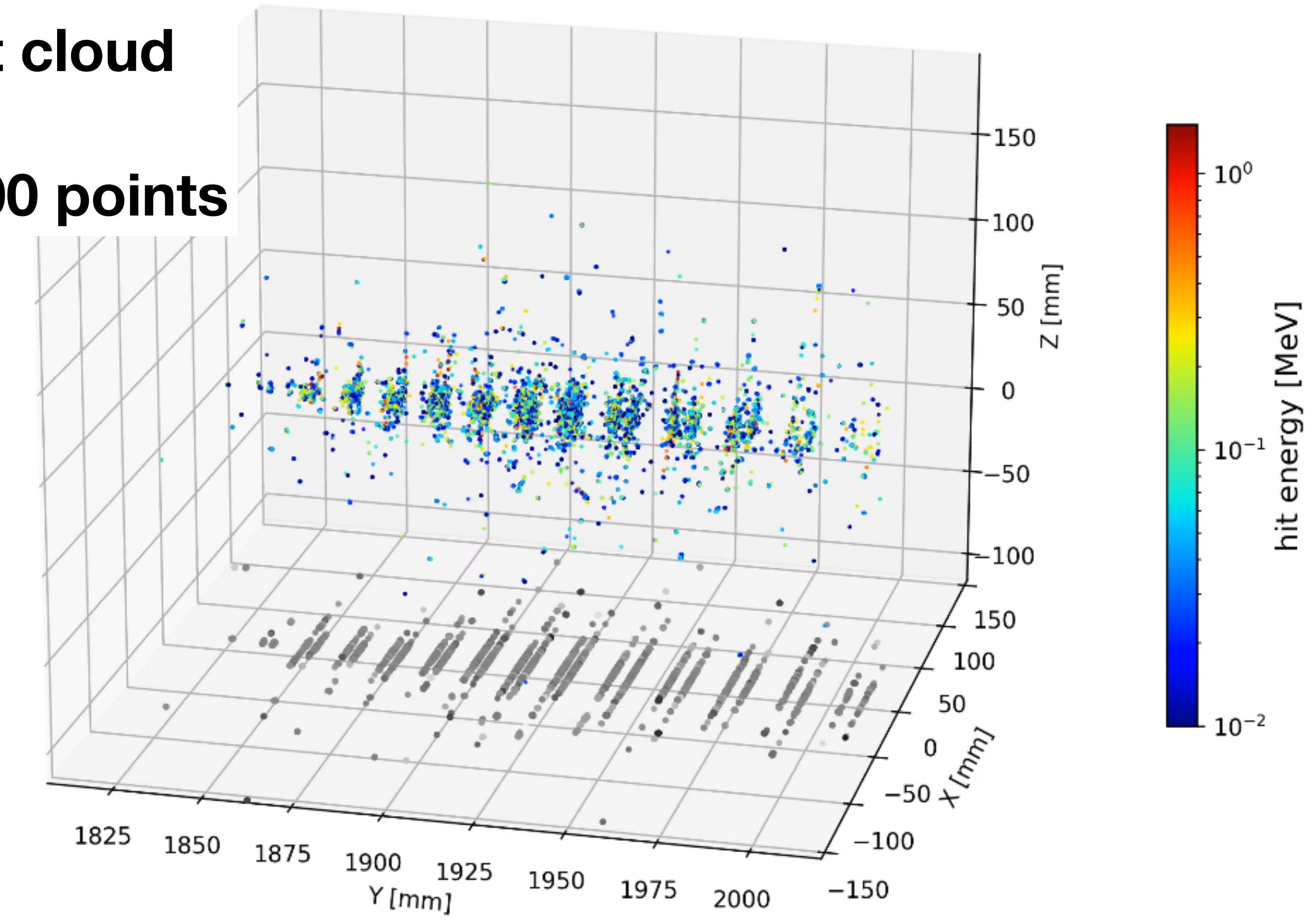
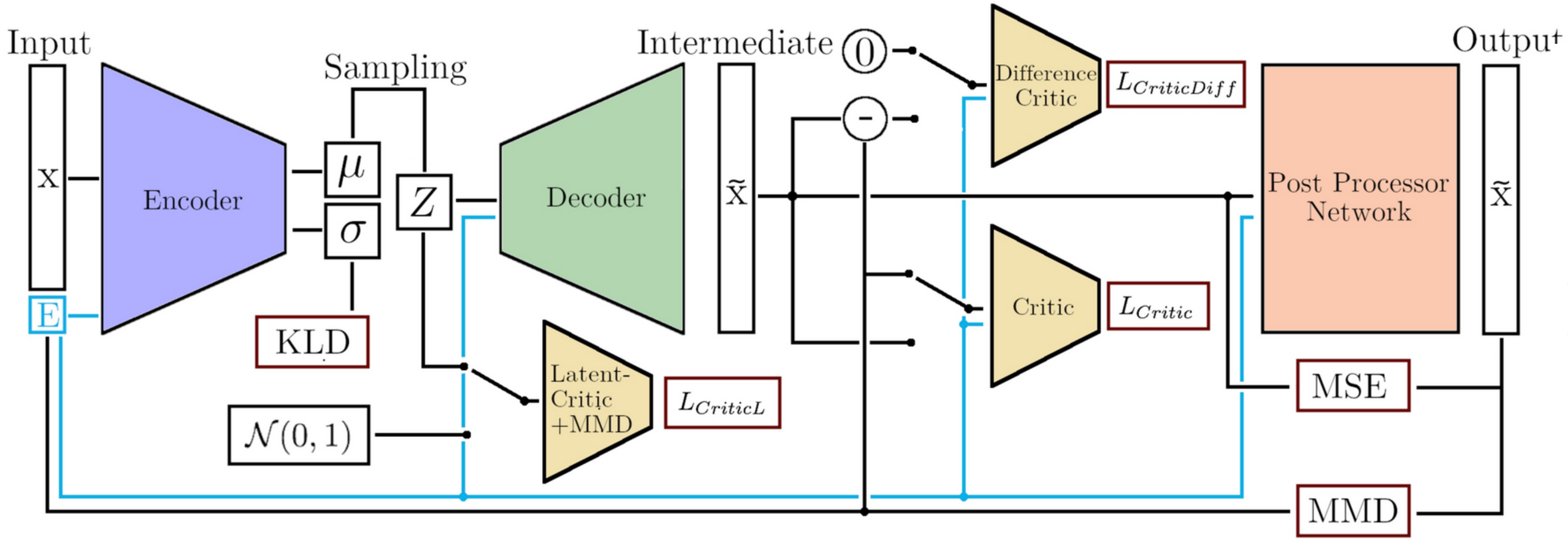


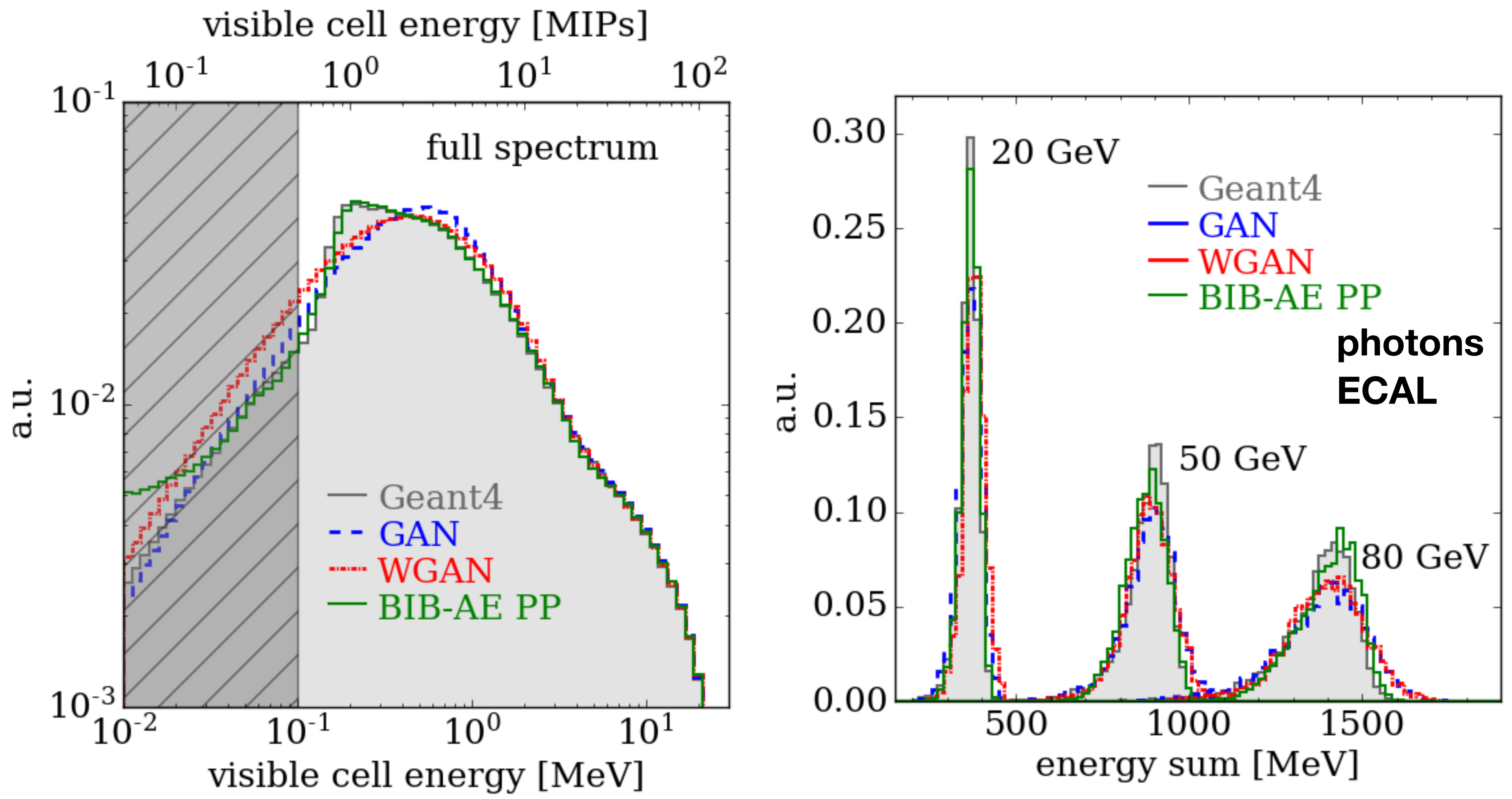
Image Based Models

BIB-AE: Photons and Pions

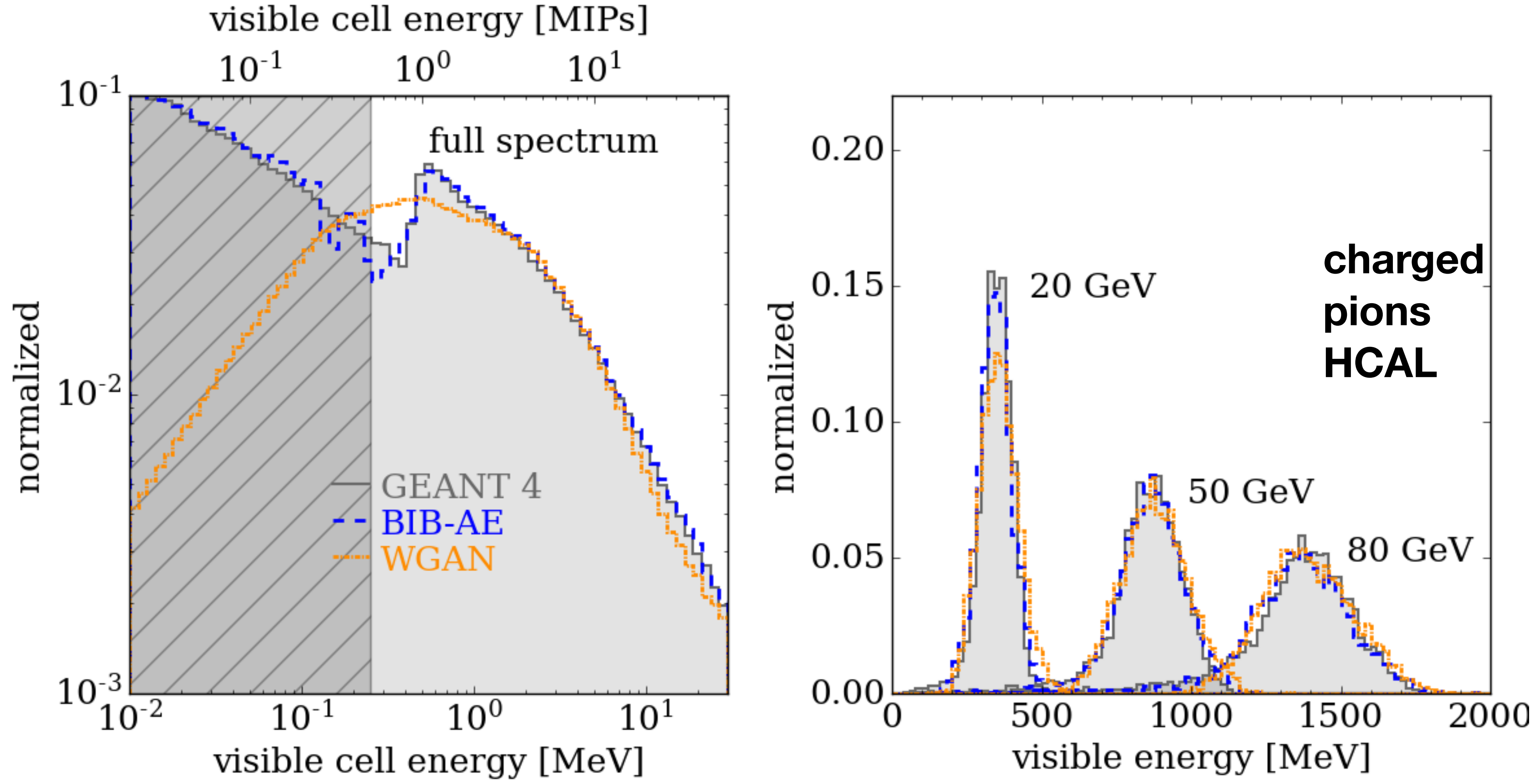
- Achieved **high fidelity** generation of **photon** and **pion** showers with **BIB-AE** architecture (and post processing)
  - 90 deg impact angle, fixed position in calorimeter
  - Fixed regular 3D grid geometry (O(10-100k) voxels)



BIB-AE: Bounded Information Bottleneck Auto-Encoder



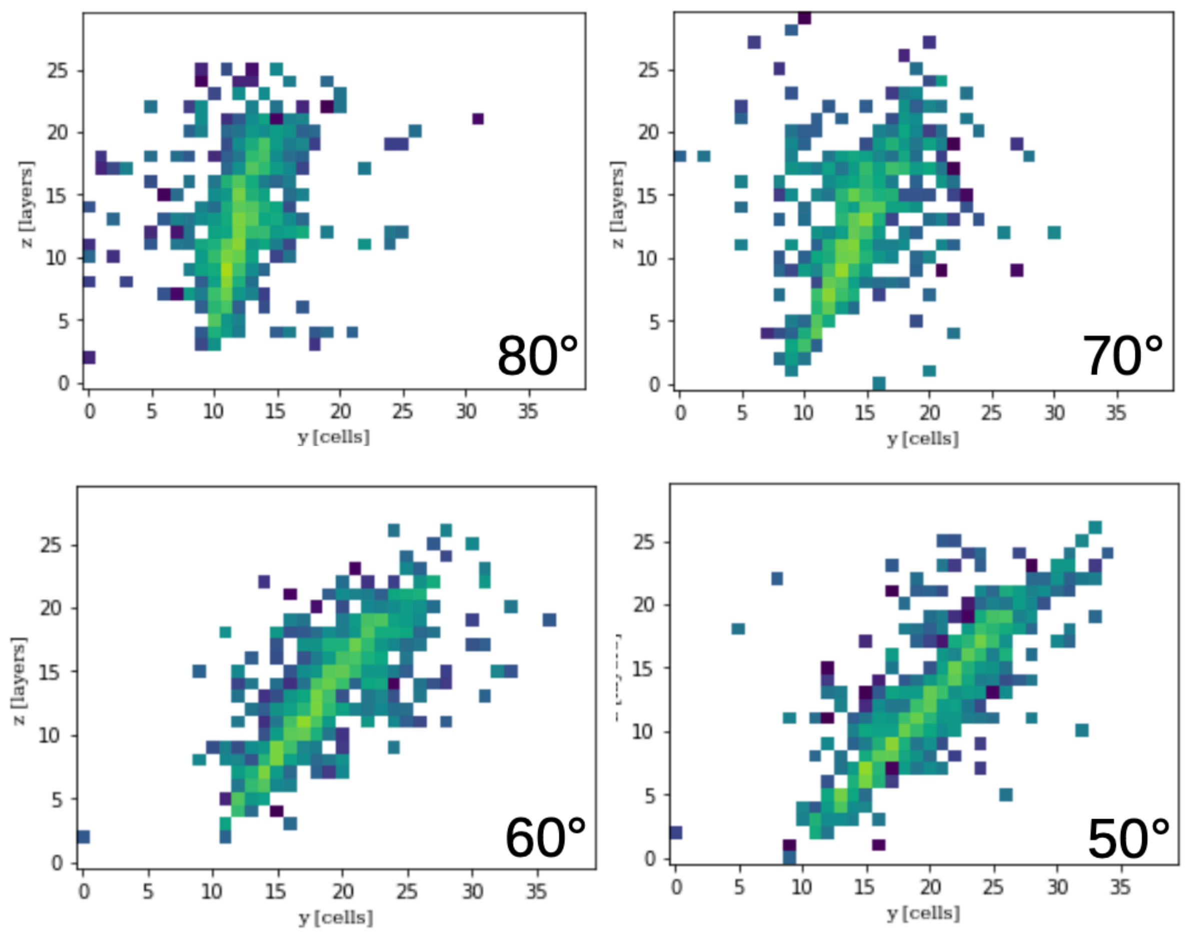
Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed
Buhmann et al., arXiv:2005.05334, Comput Softw Big Sci 5, 13 (2021)

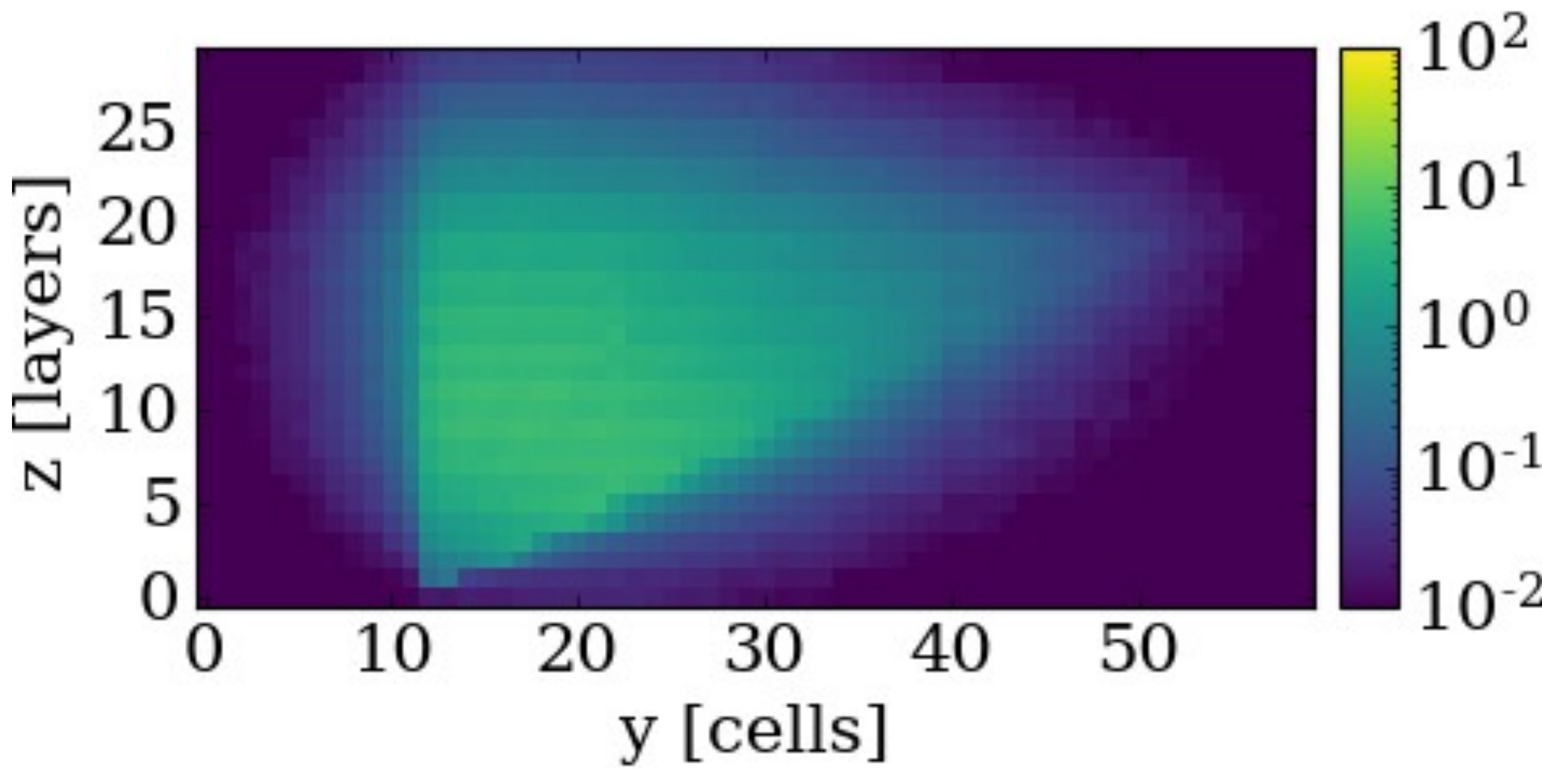


Hadrons, Better, Faster, Stronger
E. Buhmann, et al, arXiv:2112.09709, MLST 3 2, 025014 (2022)

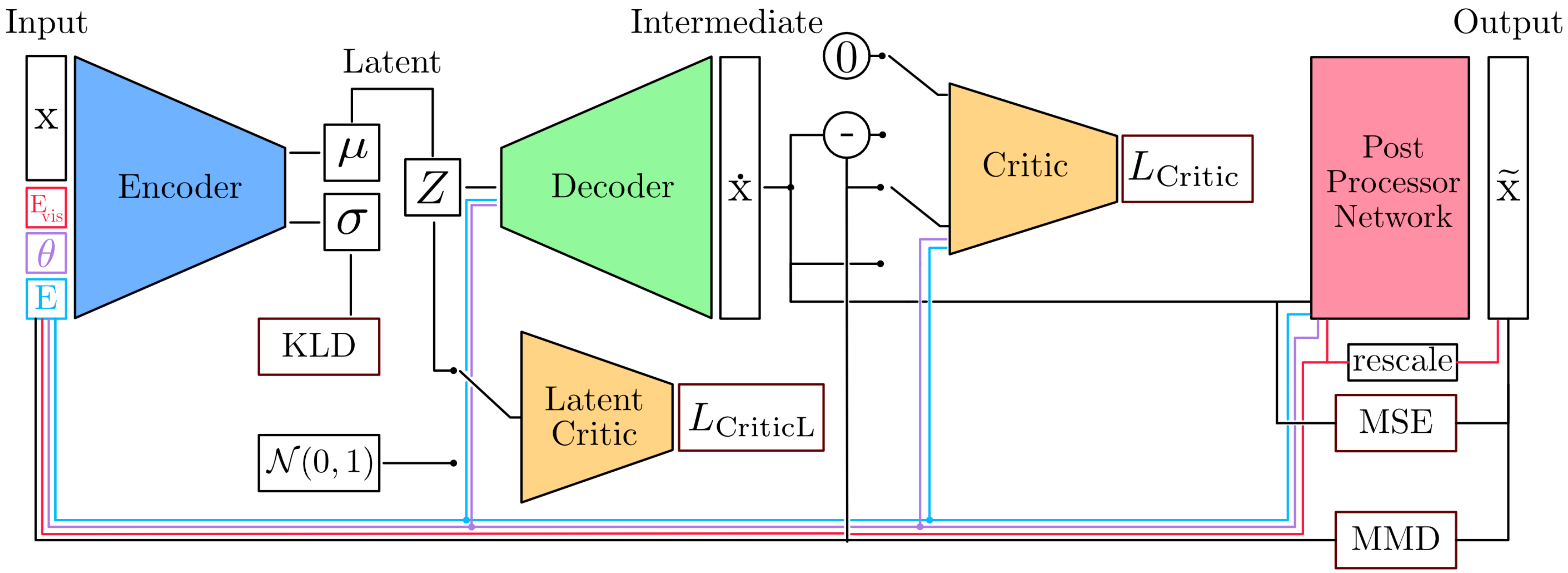
Energy and Angular Conditioning

- Extend BIB-AE architecture
- Vary incident energy and **polar angle**
  - Large training sample: 500k showers
  - Uniform in [10-100 GeV, 30-90 deg]
  - Test/validation samples at dedicated energies and angles
- **Full reconstruction** with PandoraPFA

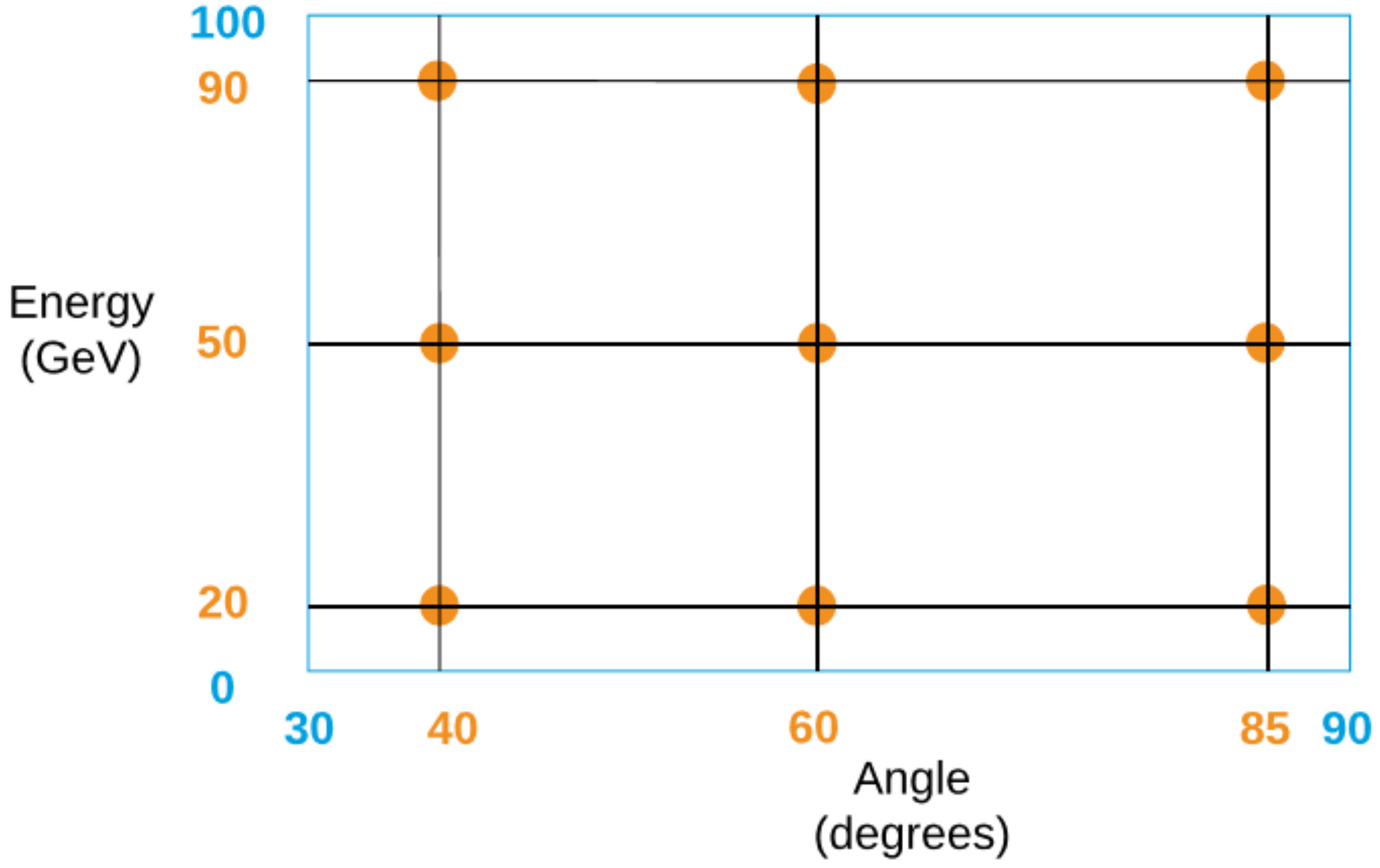




5k shower overlay



BIB-AE with angular conditioning



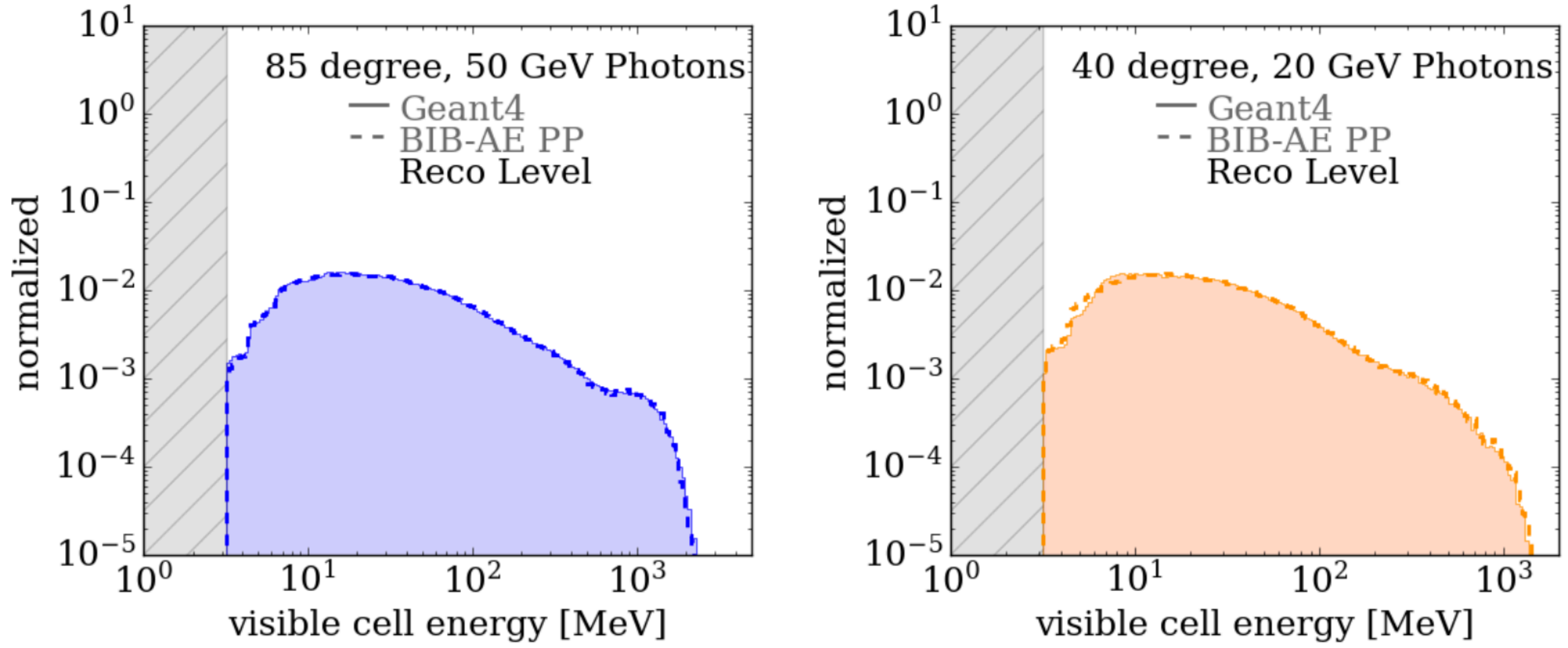
Combinations of incident energy and angle used for training

Performance after Reconstruction

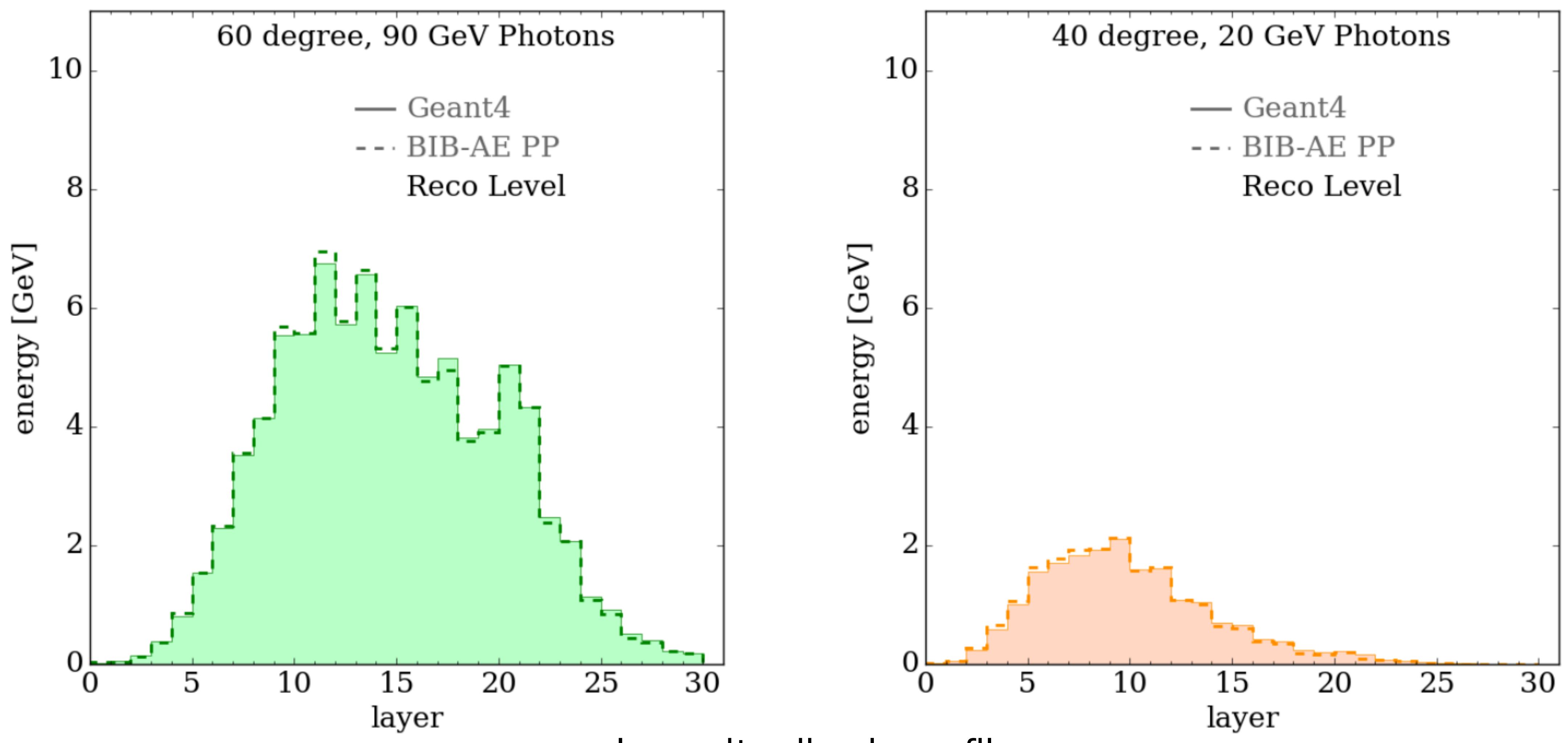
Best (left) and worst (right) test point

Excellent physics fidelity

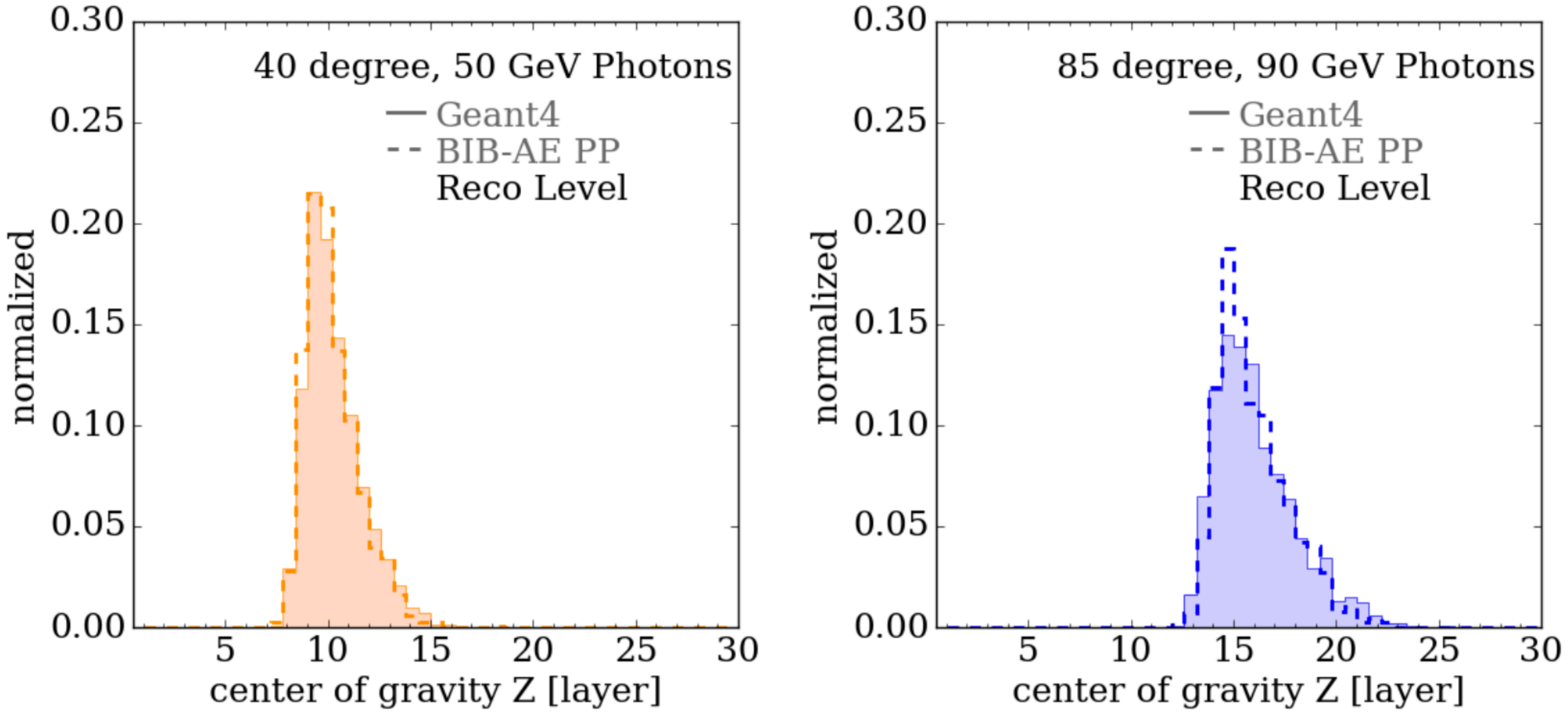
New Angles on Fast Calorimeter Shower Simulation, Diefenbacher, P.M. et al. 2023 MLST in press
DOI 10.1088/2632-2153/acefa9, arXiv: 2303.18150



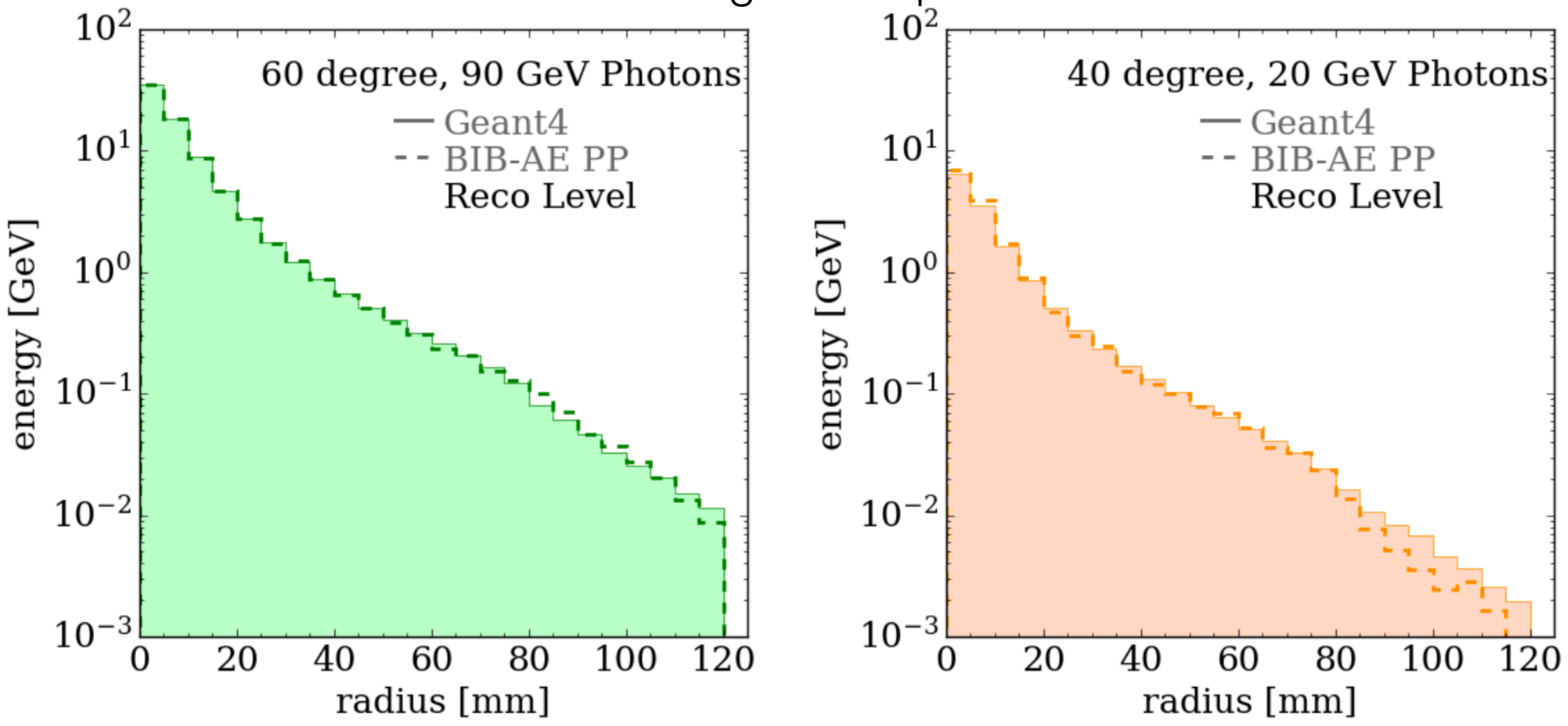
Hit energy spectrum



Longitudinal profile

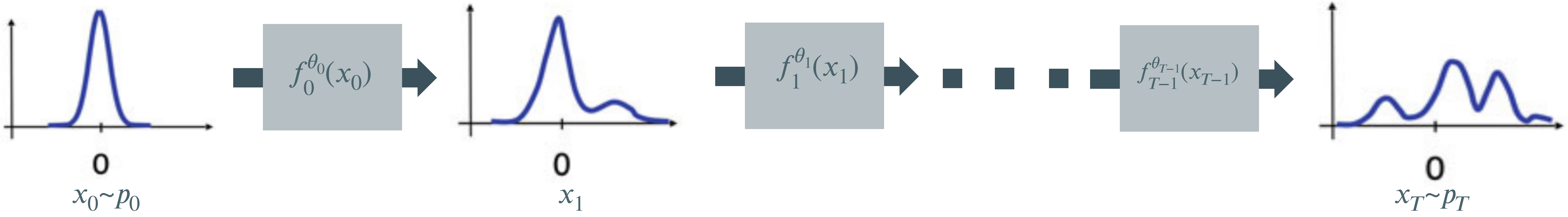


Centre of gravity in z



Radial profile

Normalizing Flows



Training: $$\log p_T(x_T) = \log p_0(x_0) - \log \left| \frac{\partial f_t^\theta}{\partial x_t} \right|$$

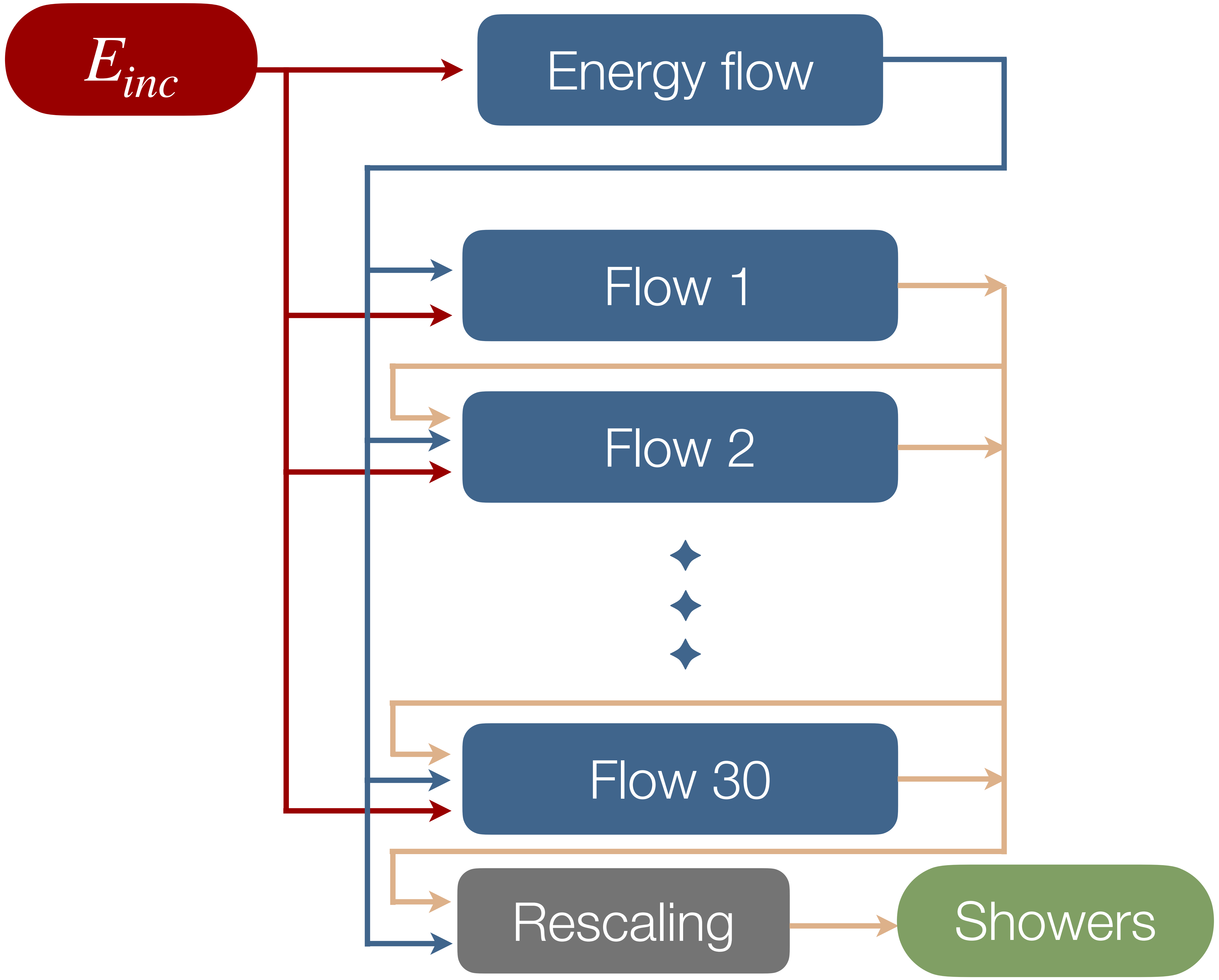
Sampling: $$x_T = f_{T-1} \circ \ldots \circ f_0(x_0)$$

- ***f*** must be invertible
- Determinant computationally expensive
  - Restricted transformations needed

Variational Inference with Normalizing Flows
Rezende et al.; arXiv:1505.05770

Layer-to-Layer Normalizing Flows Model

- Expands on CaloFlow[1] and L2LFlows[2]:
  - Changed MADE blocks to Convolutional Coupling blocks
  - Better scaling: can handle higher granularity
  - Faster generation
- One energy distribution flow
  - Learns distribution of layer energies
  - Conditioned on incident energy
- 30 causal flows
  - Learn shower shape in layer
  - Conditioned on:
    - Incident energy
    - Layer energy
    - Previous layers

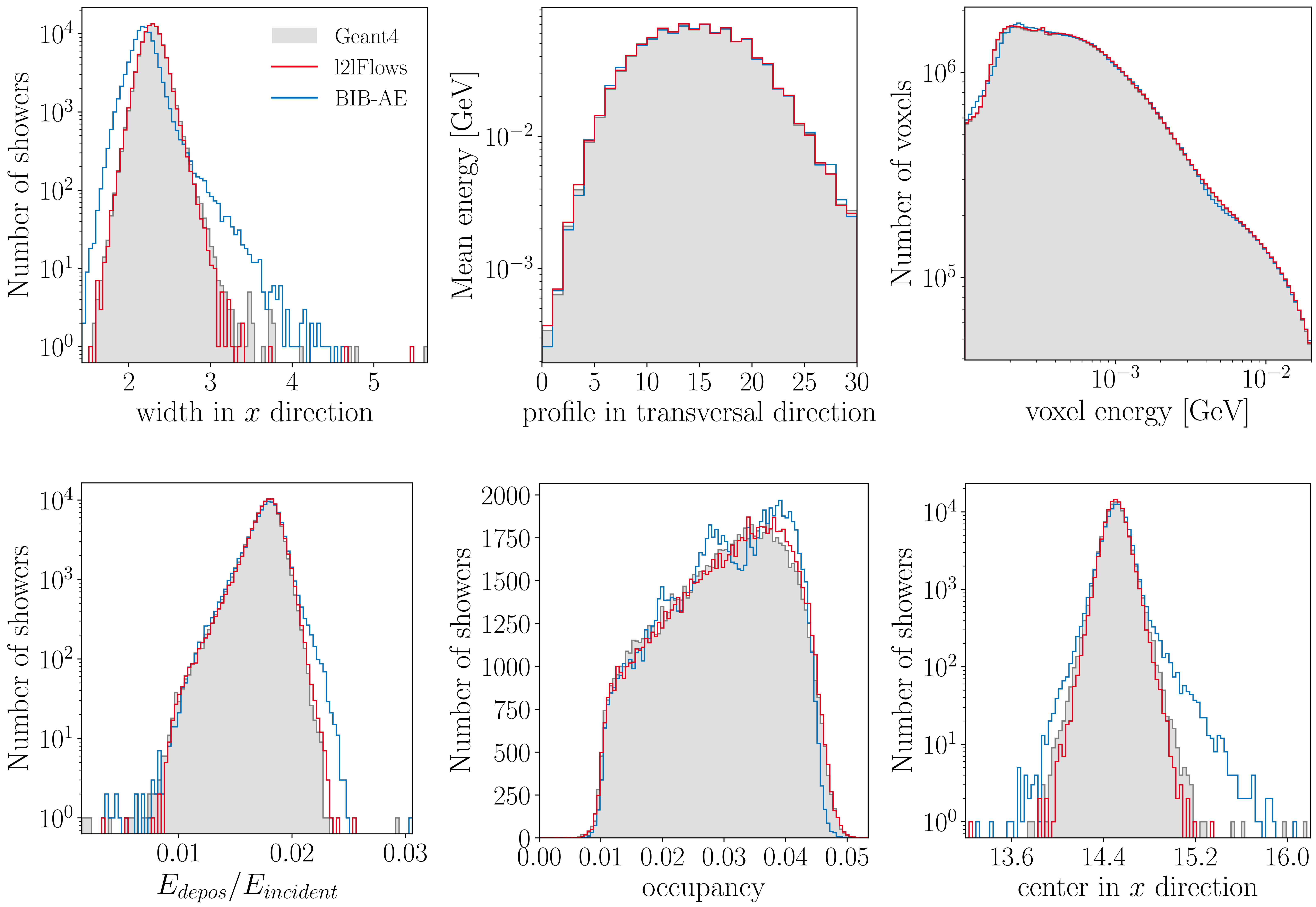


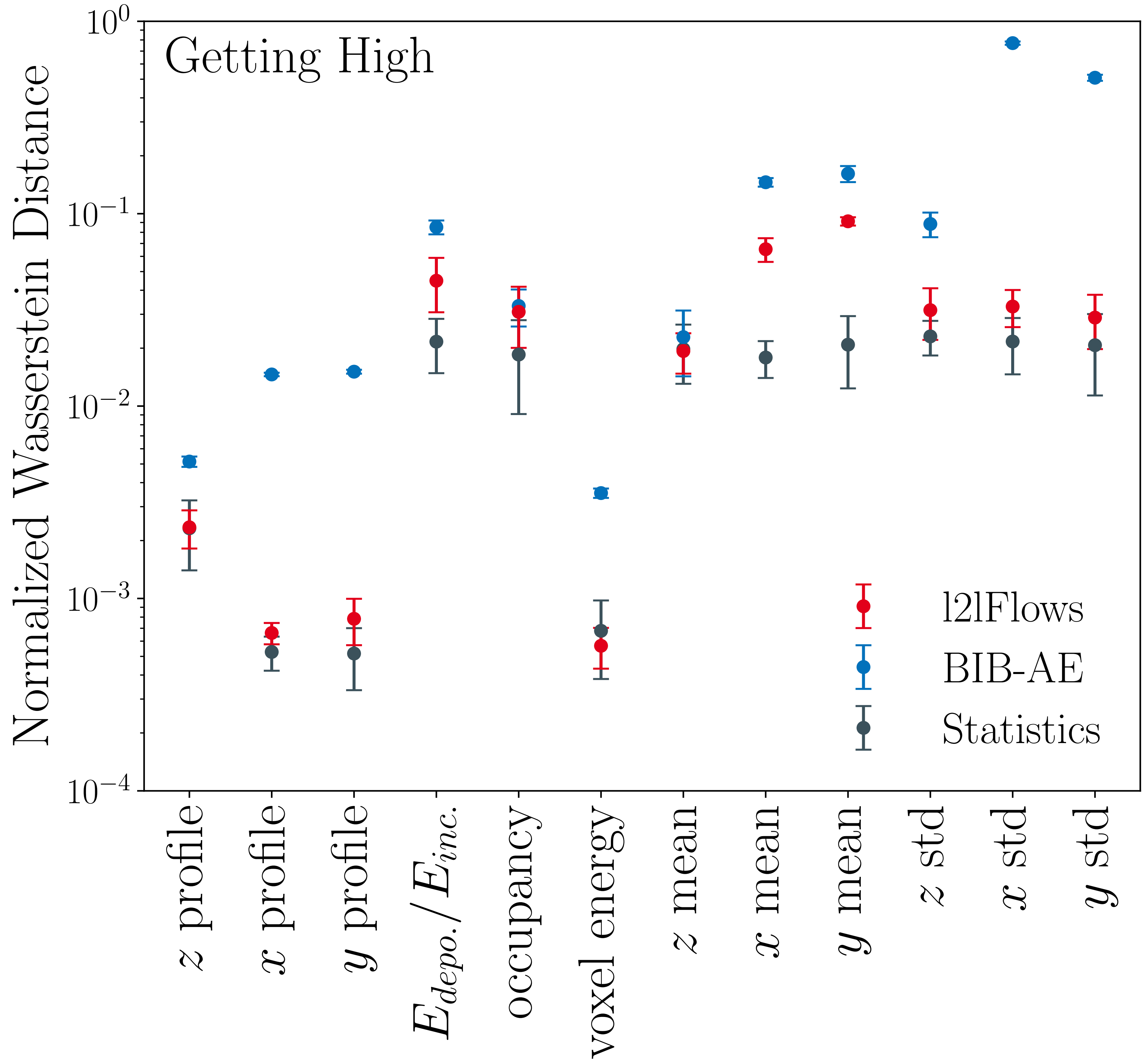
Paper coming soon!

[1]CaloFlow: Fast and Accurate Generation of Calorimeter Showers with Normalizing Flows
Claudius Krause and David Shih. arXiv: 2106.05285.

[2]L2LFlows: Generating High-Fidelity 3D Calorimeter Images
Sascha Diefenbacher et al; arXiv: 2302.11594.

L2L Flows Model Results





- Superior simulation-level performance vs a BIB-AE across a range of observables
- More work required to achieve competitive simulation speed

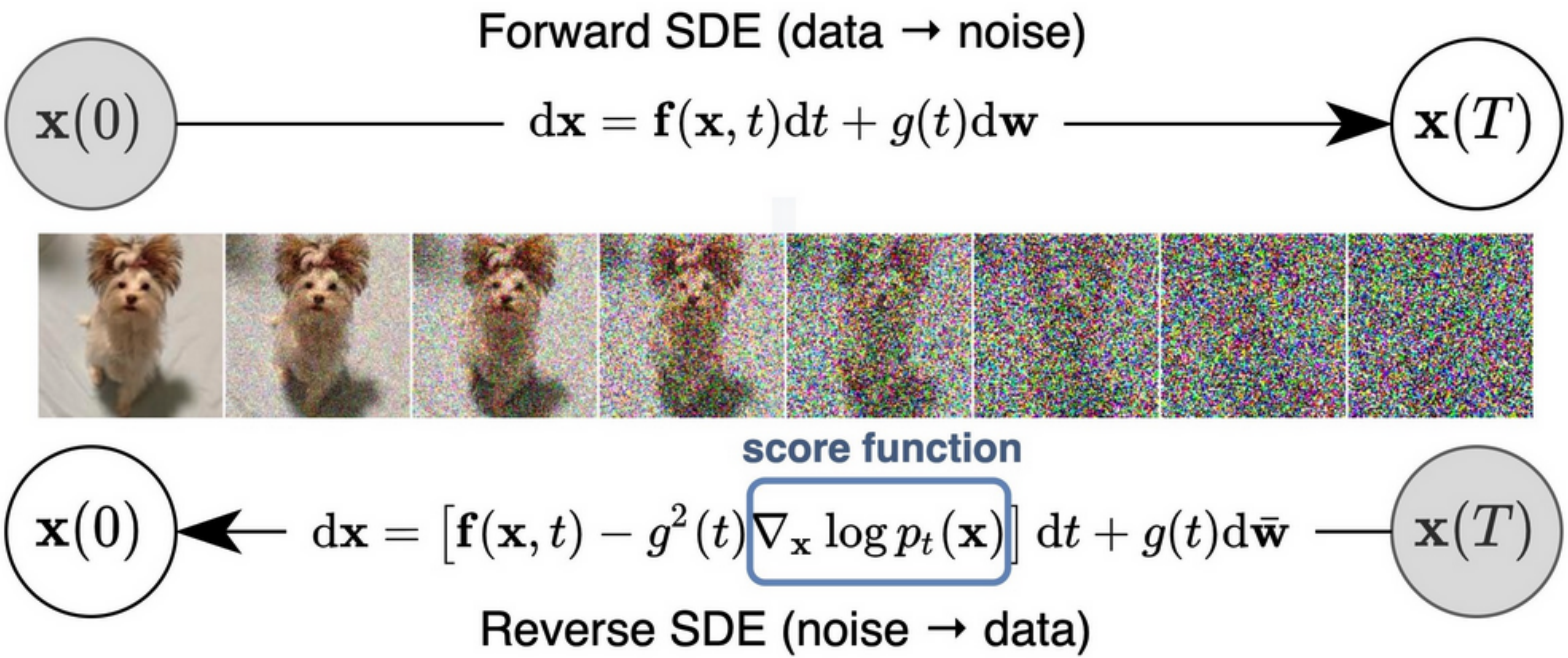
Point Cloud Models

Data Processing

- Photon showers (10-90 GeV) in the electromagnetic calorimeter (ECAL) of the International Large Detector (ILD) at the International Linear Collider (ILC)
- Point clouds of **clustered Geant4 steps**: 36x higher granularity than cell hits, 7x fewer points than full Geant4 steps
  - Multiple points per cell & geometry independent

	points / shower	Note
All GEANT4 steps	40 000	Initial output of GEANT4
Clustered GEANT4 steps	6 000	Input/output of CALOCLOUDS
Hits in calorimeter grid	1 500	Calculation of physics observables

Diffusion Models

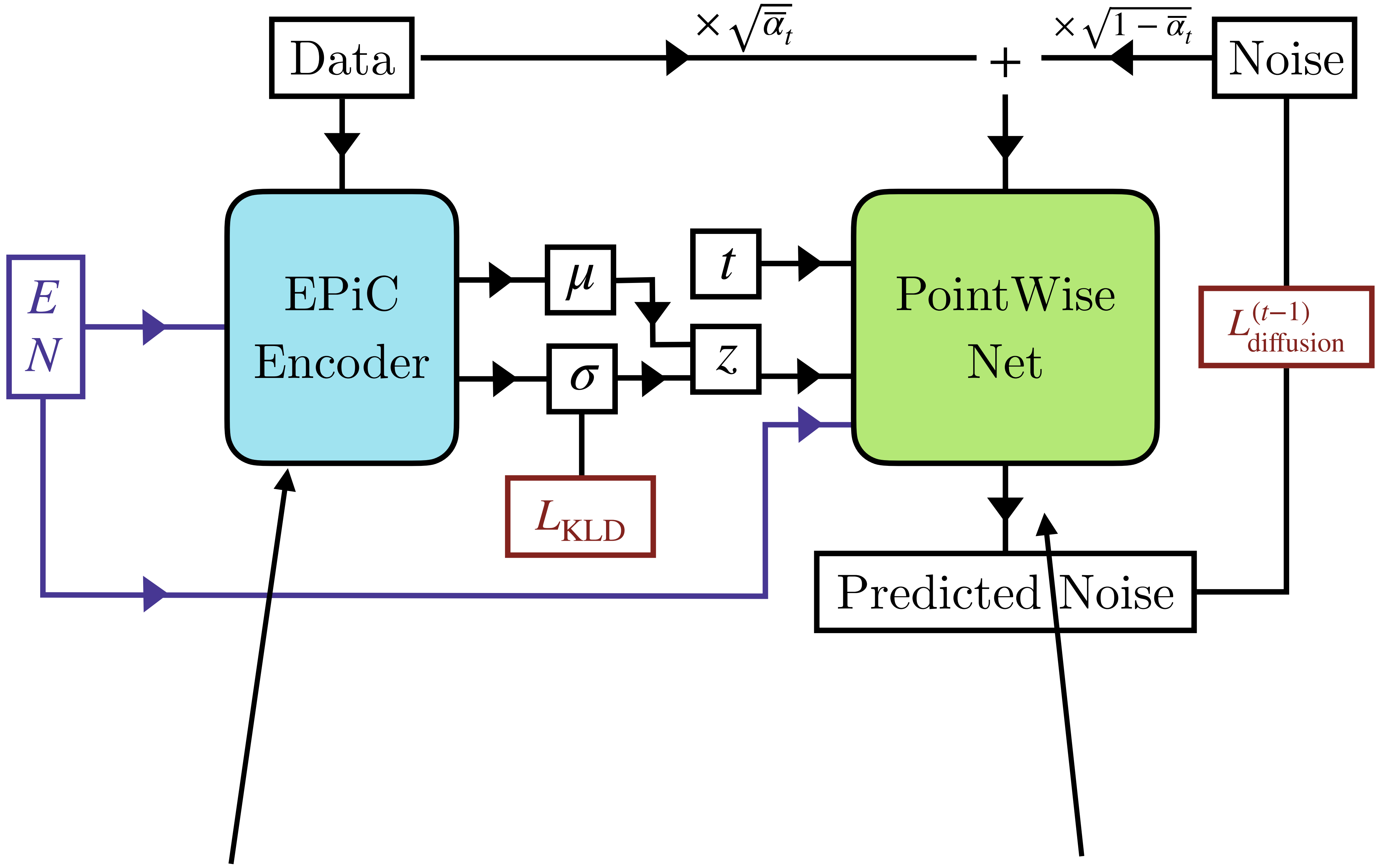


- Progressively perturb data with stochastic differential equation (SDE)
- Train by estimating the score function, fixed ***f*** and ***g***
- Sample by solving reverse SDE

Score-Based Generative Modeling through Stochastic Differential Equations
Song et al.; arxiv:2011.13456

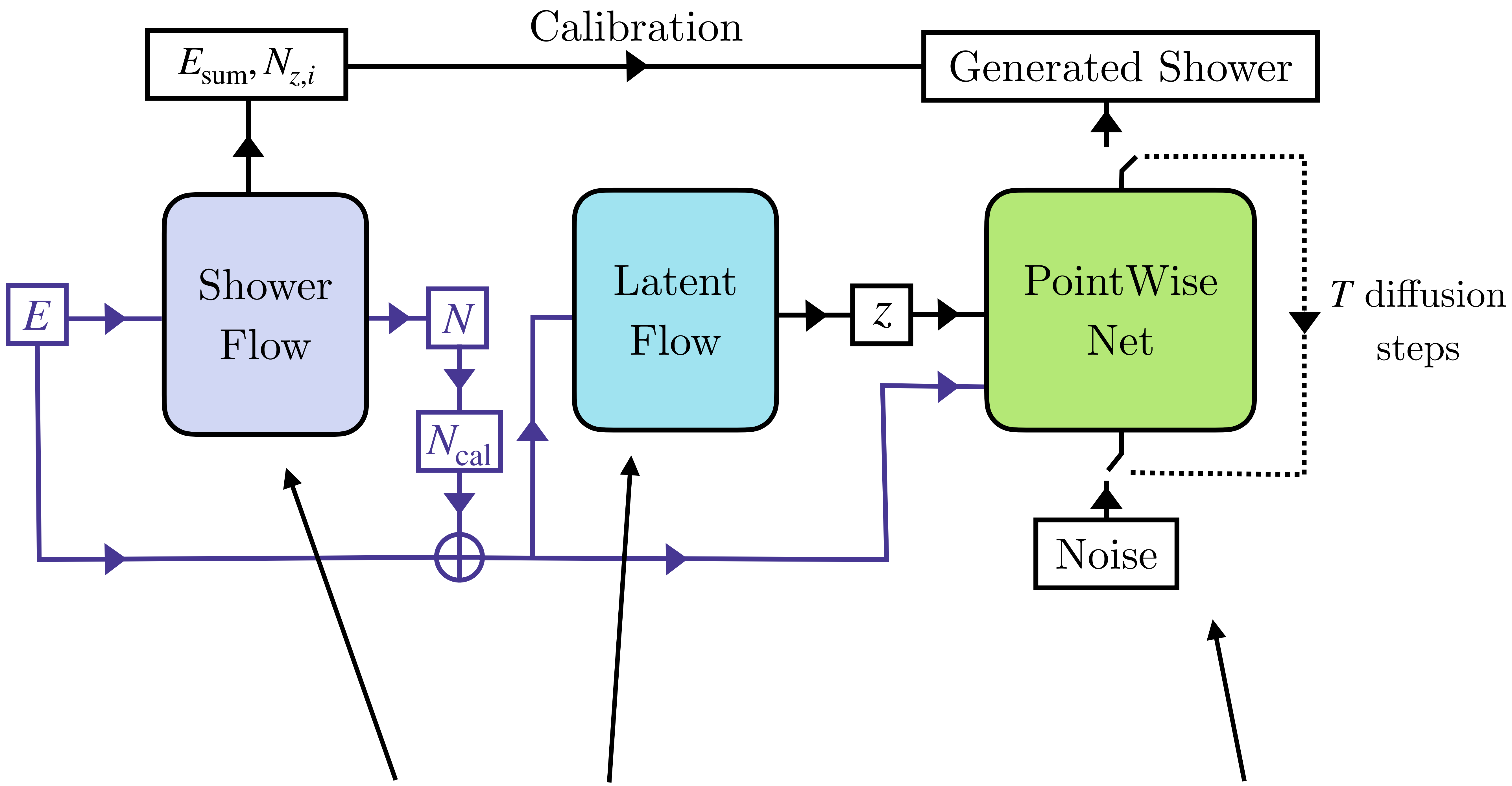
CaloClouds Model

Training



VAE-like encoder **Diffusion Model**

Sampling



Normalizing Flows **Diffusion Model**

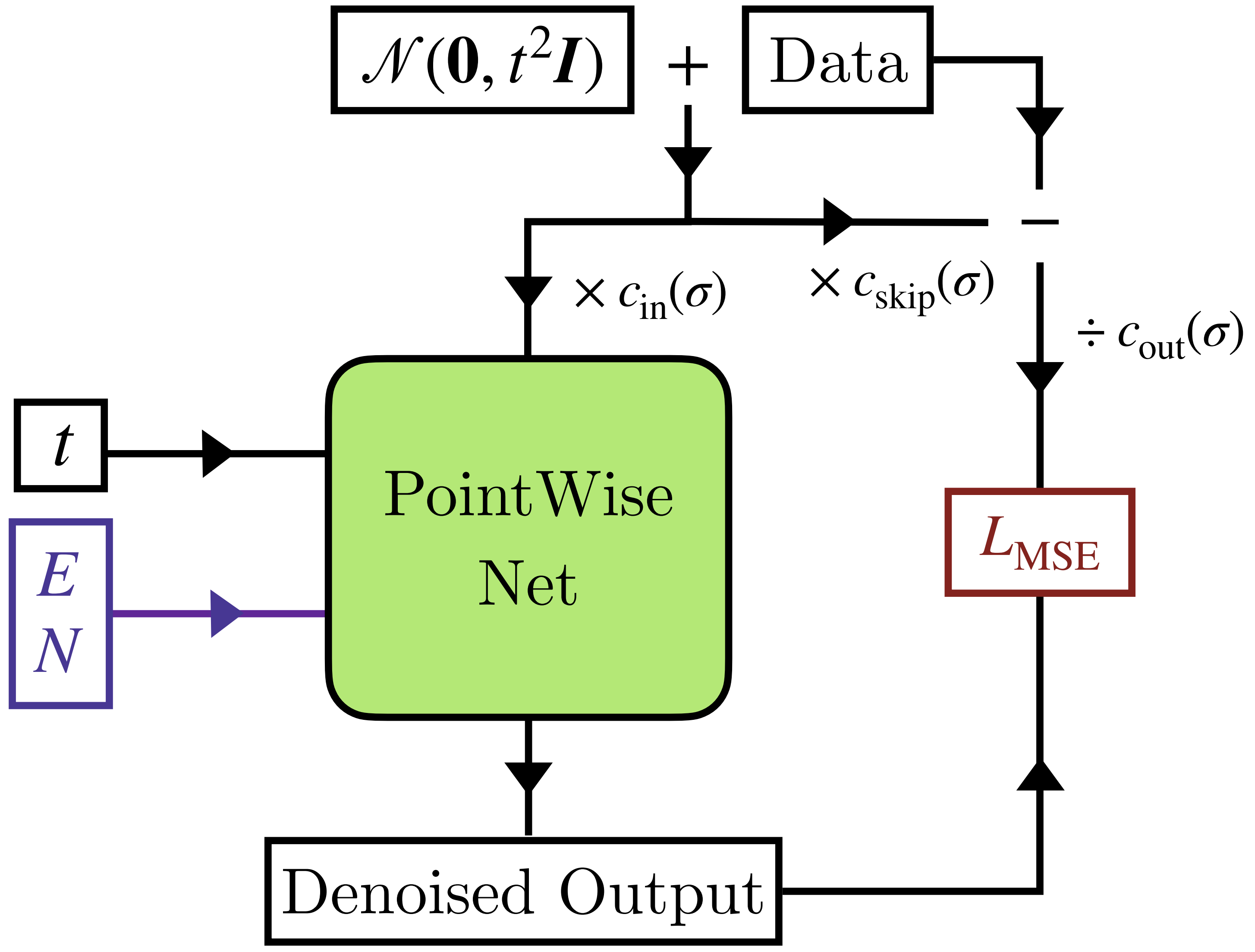
Equivariant Point Cloud Generation for Particle Jets
E. Buhmann, G Kasieczka, J Thaler:
EPiC-GAN

- Post-diffusion calibration: number of points per layer, energy per layer, center of gravity in X and Y-direction

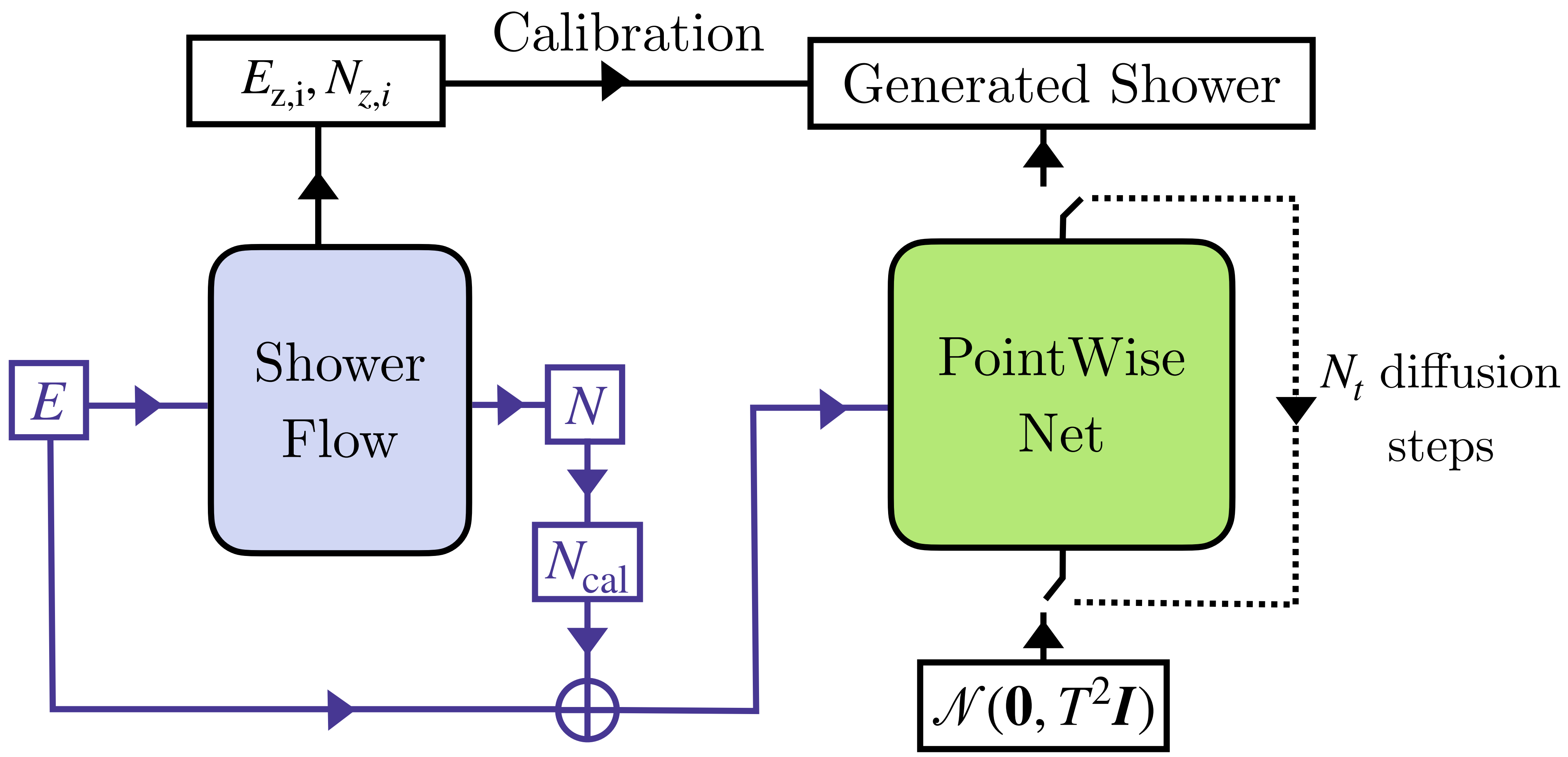
CaloClouds I & II

- **CaloClouds** diffusion model based on discrete-time Denoising Diffusion Probabilistic Models[3] (DDPM)
- Number of diffusion steps: training = sampling
    - Here: 100 denoising steps
- **CaloClouds II** diffusion model based on a continuous-time diffusion model[4]
    - Allows for a variety of stochastic and ordinary differential equation solvers (ODE / SDE solvers)
    - **Fewer & variable** number of steps during sampling
        - Here: Heun ODE solver with 25 model evaluations
    - **Allows for distillation** into a consistency model[5]
        - Here: Consistency model for single-shot generation
- CaloClouds II: No latent space (no encoder & latent flow)

Training (CaloClouds II)



Sampling (CaloClouds II)

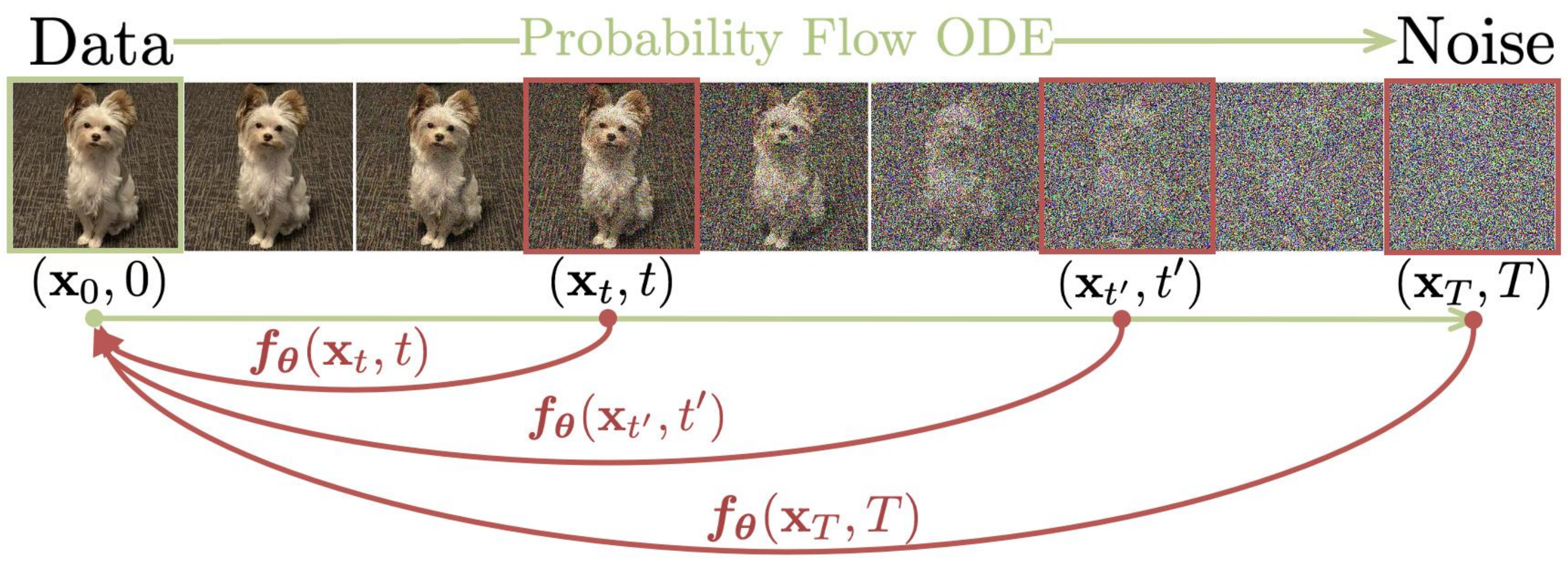


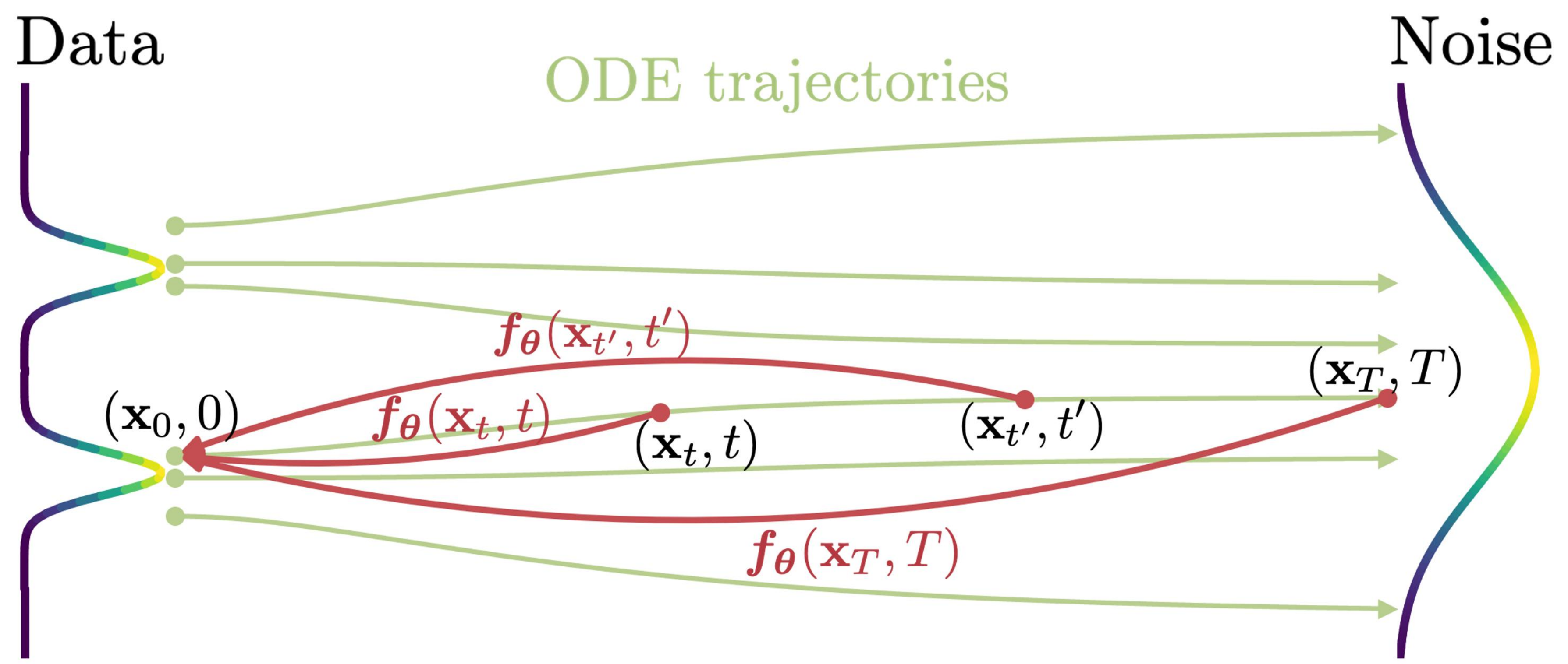
[3]Denoising Diffusion Probabilistic Models; J Ho et al. arxive: 2006.11239

[4]Elucidating the Design Space of Diffusion-Based Generative Models T. Karras et al: arxive: 2206.00364

[5]Consistency Models Y Song et al; arXiv: 2303.01469

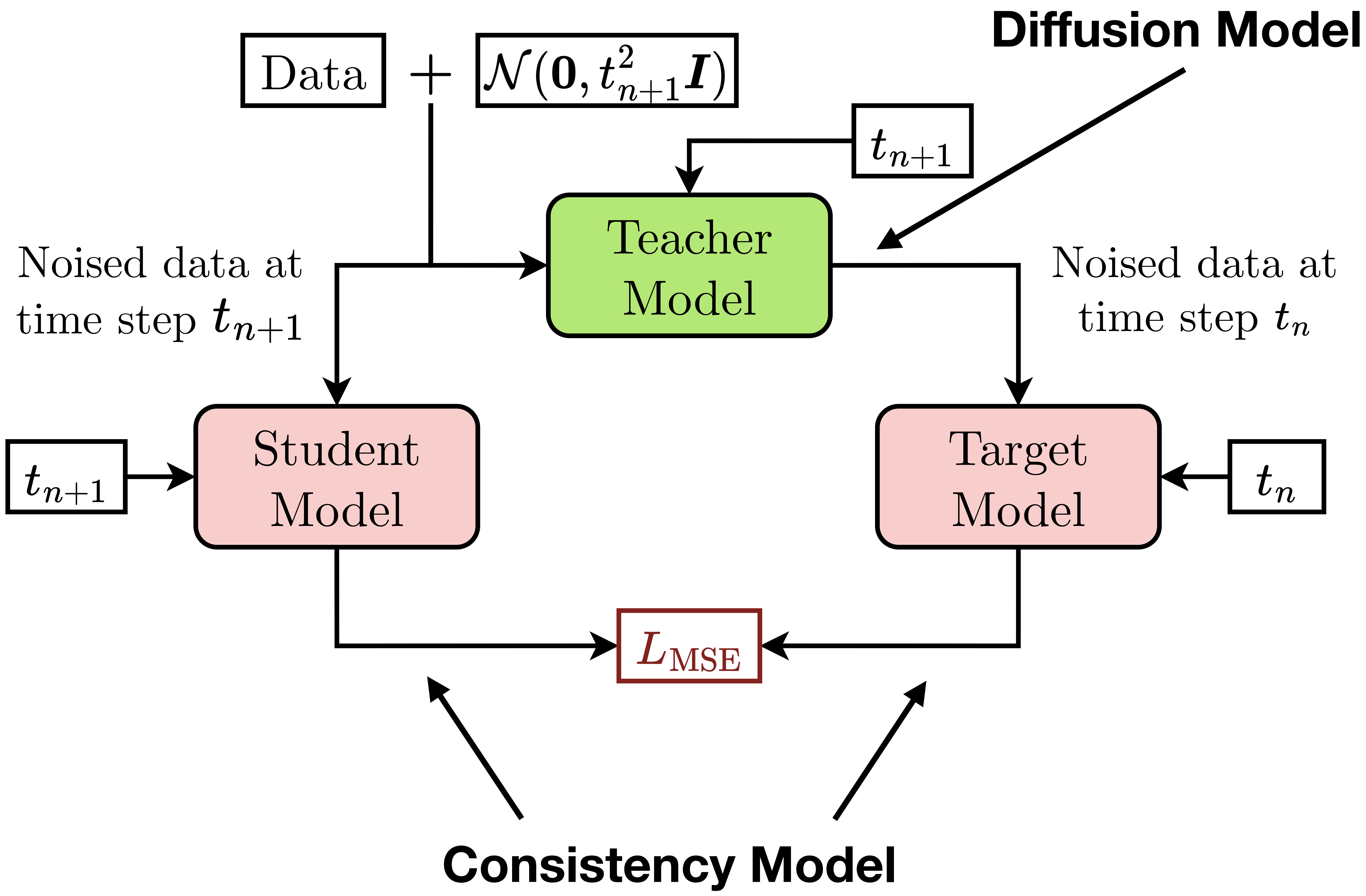
Consistency Models





- Consistency Models trainable standalone or distilled from a diffusion model
- Allow for **single-step** & **multi-step** generation

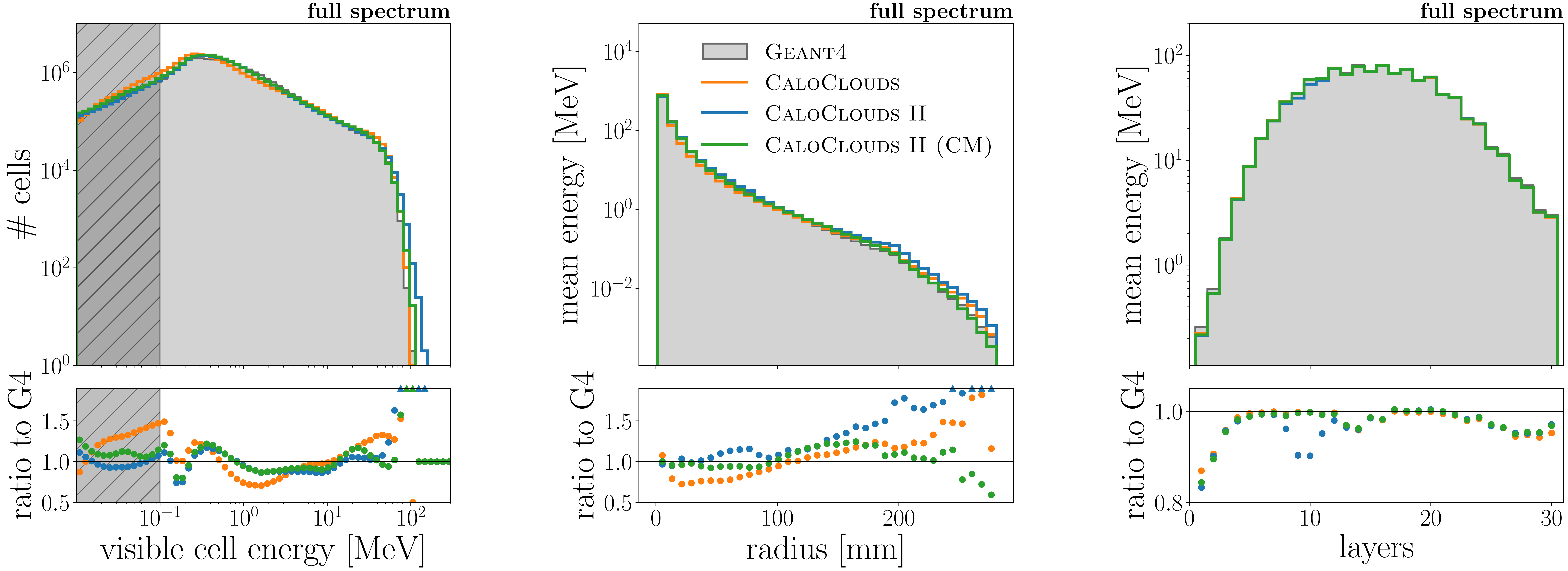
Consistency distillation:



Student model: updated via gradient descent
Target model: updated via weight average

Consistency Models
Y Song et al; arXiv: 2303.01469

CaloClouds Results



- All evaluations with point cloud showers projected to regular cell geometry
- Hit energy spectrum, radial energy profile, and longitudinal energy profile well modeled by all three CaloClouds variants (40,000 showers each)
- CaloClouds II models improve the radial energy profile

Generation Speed Comparison

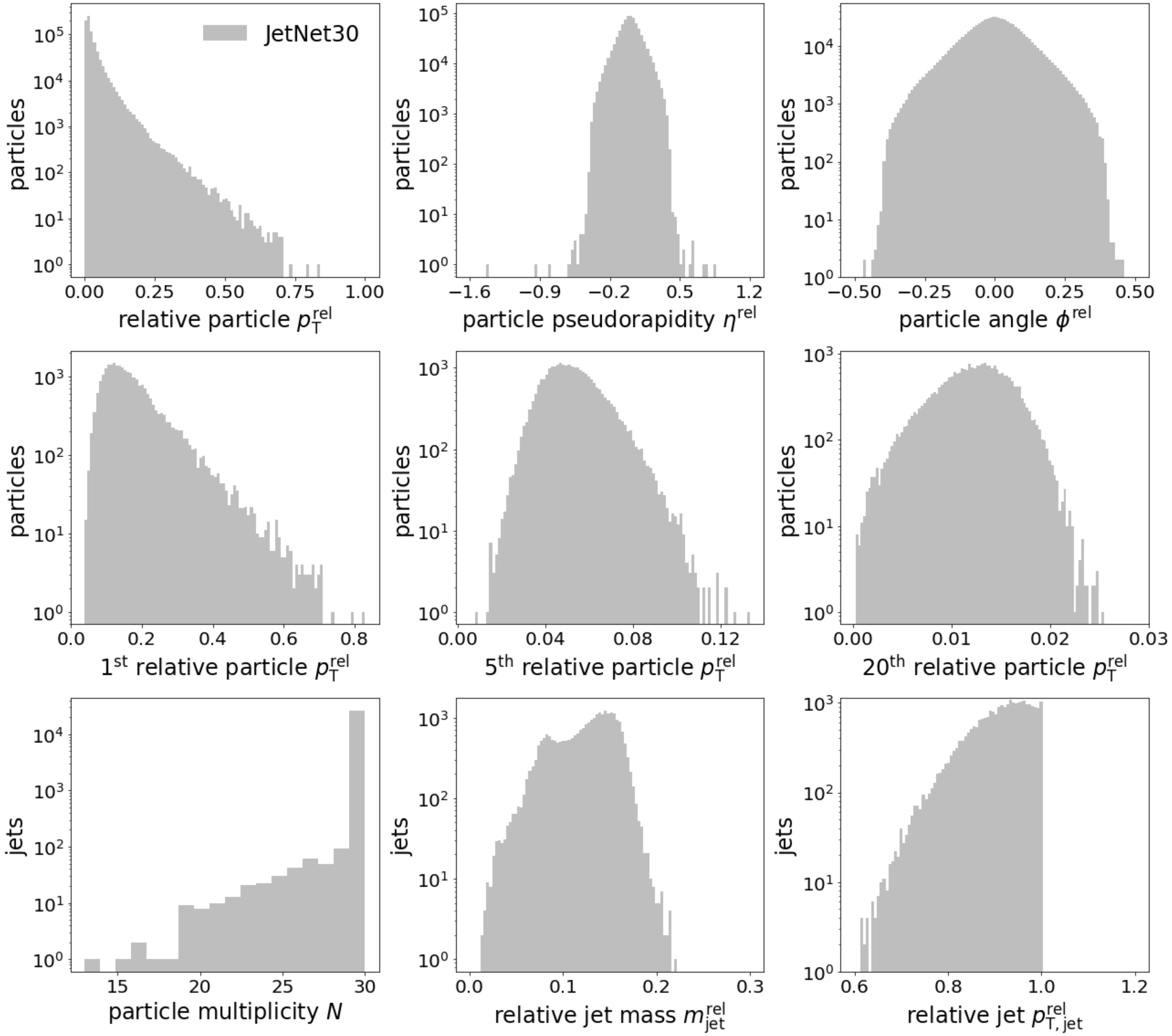
Hardware	Simulator	NFE	Batch Size	Time / Shower [ms]	Speed-up
CPU (Intel Xeon CPU E5-2640)	Geant4			3914.80 ± 74.09	$\times 1$
	CaloClouds	100	1	3146.71 ± 31.66	$\times 1.2$
	CaloClouds II	25	1	651.68 ± 4.21	$\times 6.0$
	CaloClouds II (CM)	1	1	84.35 ± 0.22	$\times 46$
GPU (NVIDIA A100 40 GB)	CaloClouds	100	64	24.91 ± 0.72	$\times 157$
	CaloClouds II	25	64	6.12 ± 0.13	$\times 640$
	CaloClouds II (CM)	1	64	2.09 ± 0.13	$\times 1873$

- Speed-up scales with the number of function evaluations (NFE)
- Largest speed-up for CaloClouds II (CM), even on CPU
- CPUs more widely available than GPUs, cheaper, and current simulation chain optimized on CPUs

More Point Clouds

JetNet30

- Benchmark dataset: JetNet30 [4]
- Simulated jets from proton-proton collisions
- Anti-k_T clustered with $R = 0.8$ and maximum particle multiplicity $N = 30$
- Particle collider coordinates normalized and centered
  - $p_T^{rel} = p_T^{particle} / p_T^{jet}$
  - $\eta^{rel} = \eta^{particle} - \eta^{jet}$
  - $\phi^{rel} = \phi^{particle} - \phi^{jet}$
- Jet types: Gluon, light quarks, <u>top quark</u>

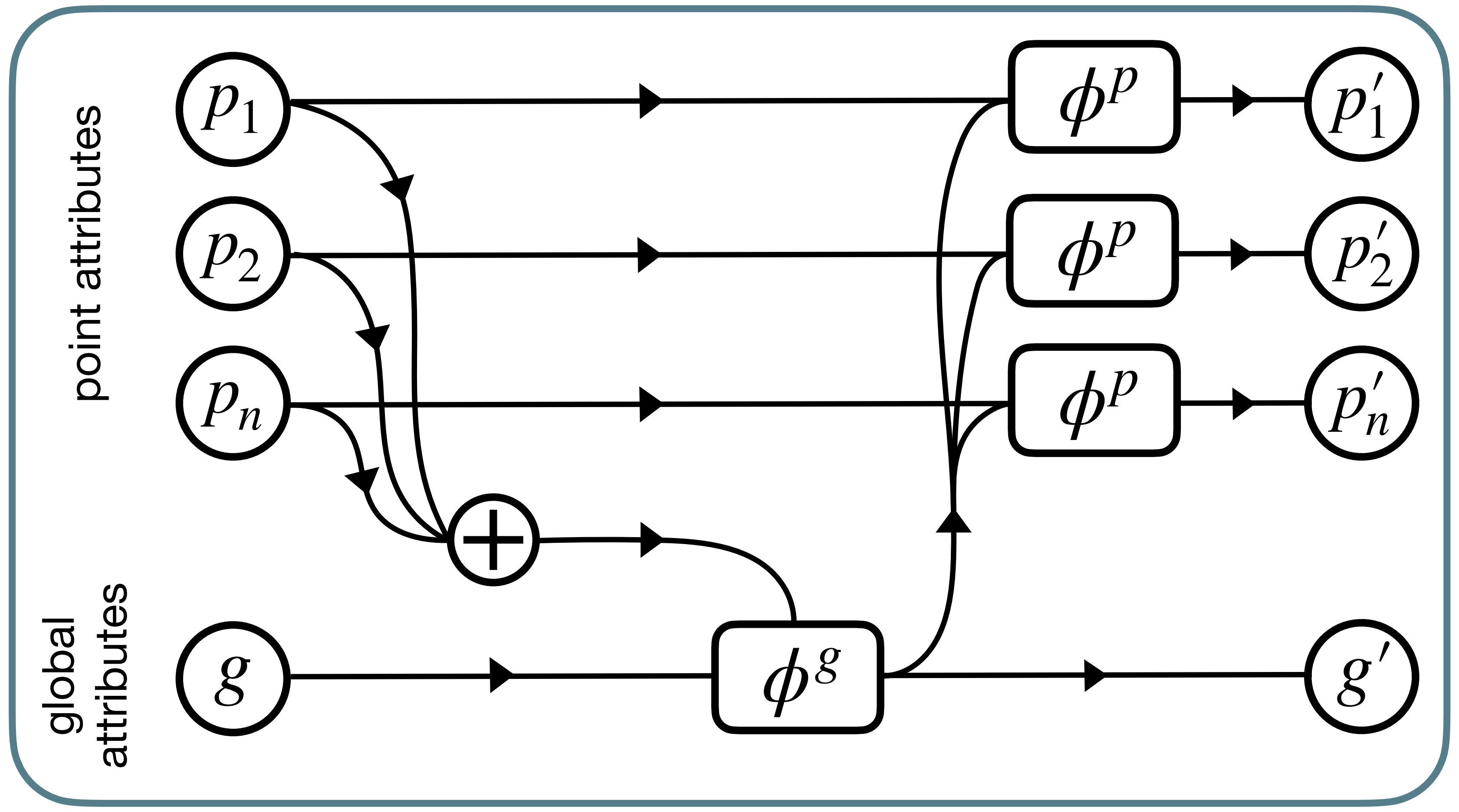


[4]R Kansal, et al: JetNet150 (2.0.0) [Data set]. Zenodo

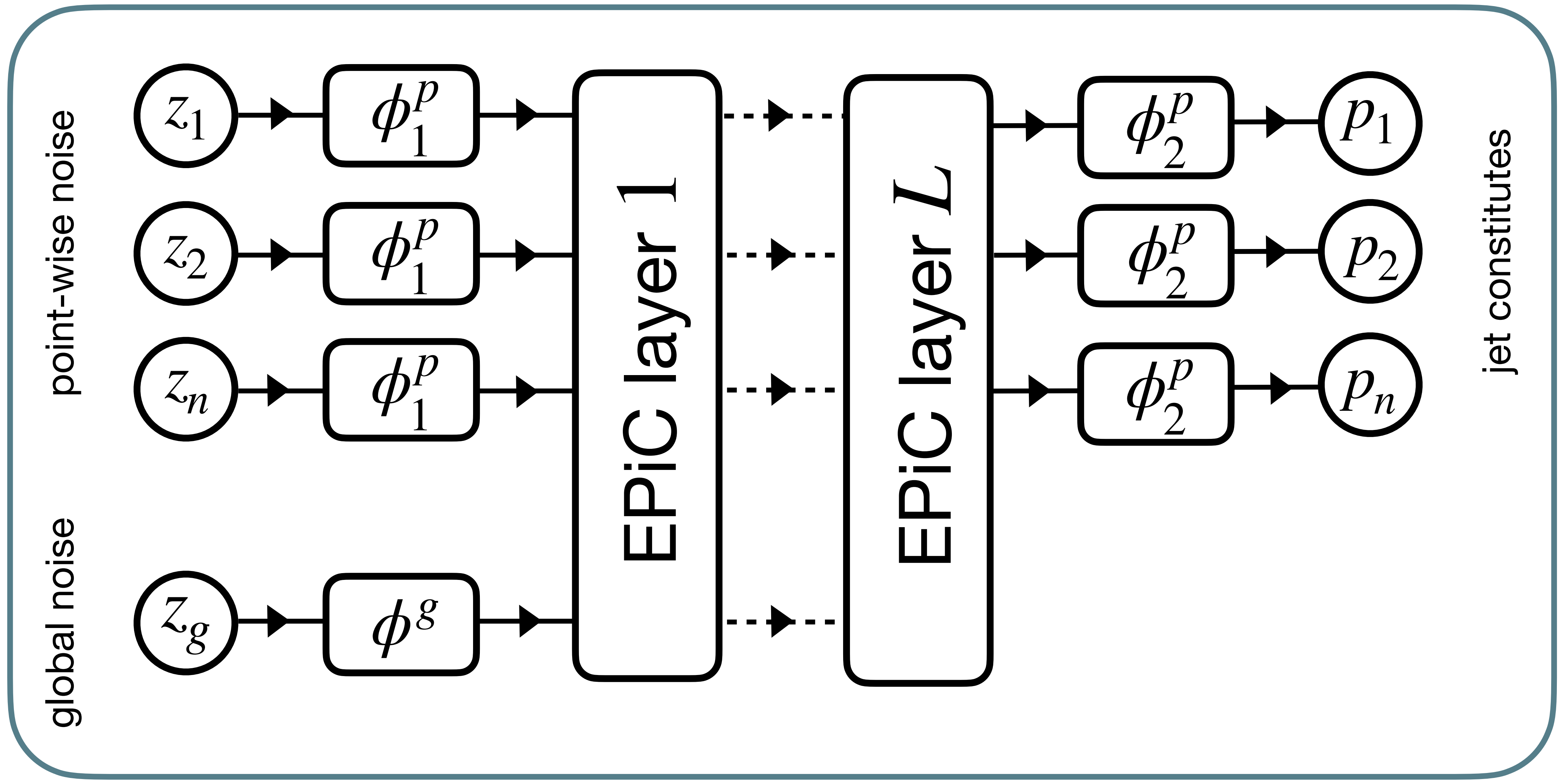
Equivariant Point Cloud (EPiC) GAN

- 2 step update process per EPiC layer:
  - Global attributes $\boldsymbol{g}$ are updated based on particle-wise attributes $\boldsymbol{p}_i$
  - Particle attributes $\boldsymbol{p}_i$ are updated base on the updated global attributes $\boldsymbol{g}'$
- Control of communication between local vectors via:
  - Length of global vector $dim(\boldsymbol{g})$
  - Number of stacked EPiC layers L

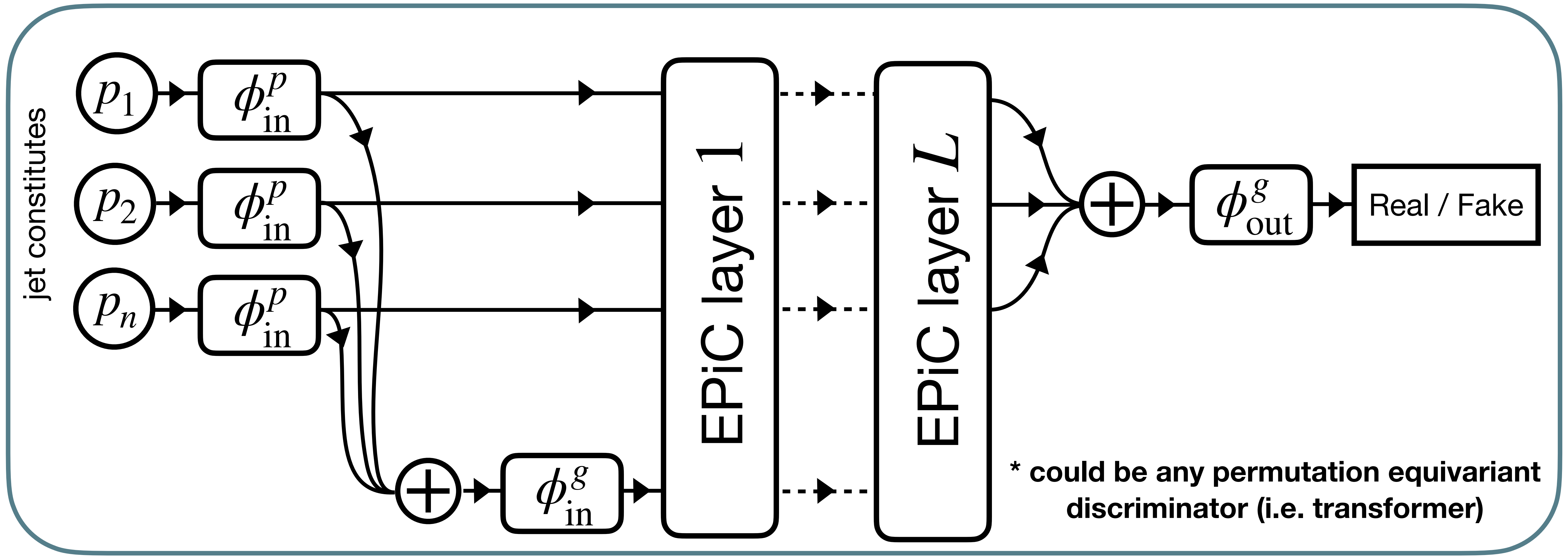
Equivariant Point Cloud (EPiC) Layer



Generator:

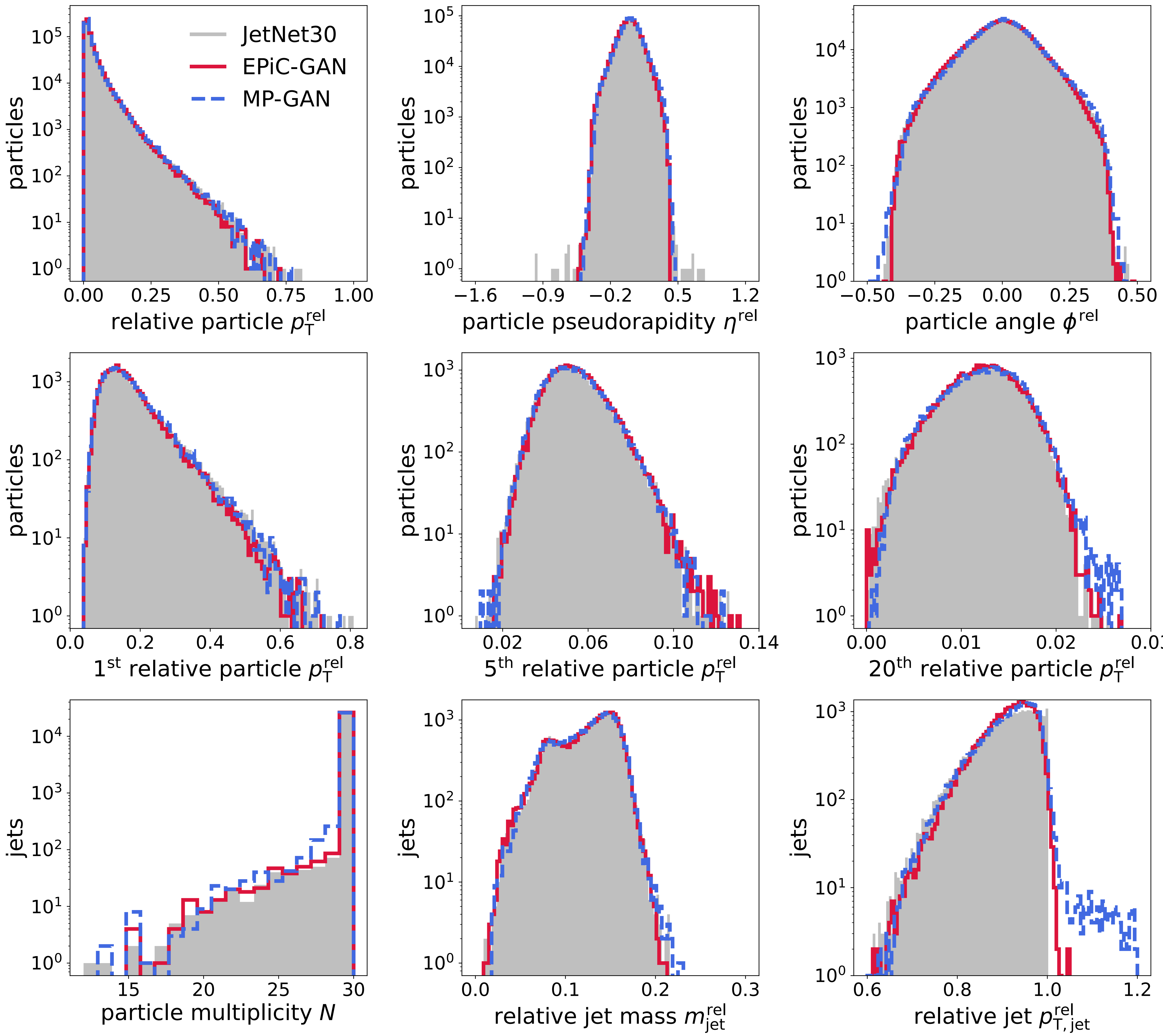


Discriminator:



Results on JetNet30 tops

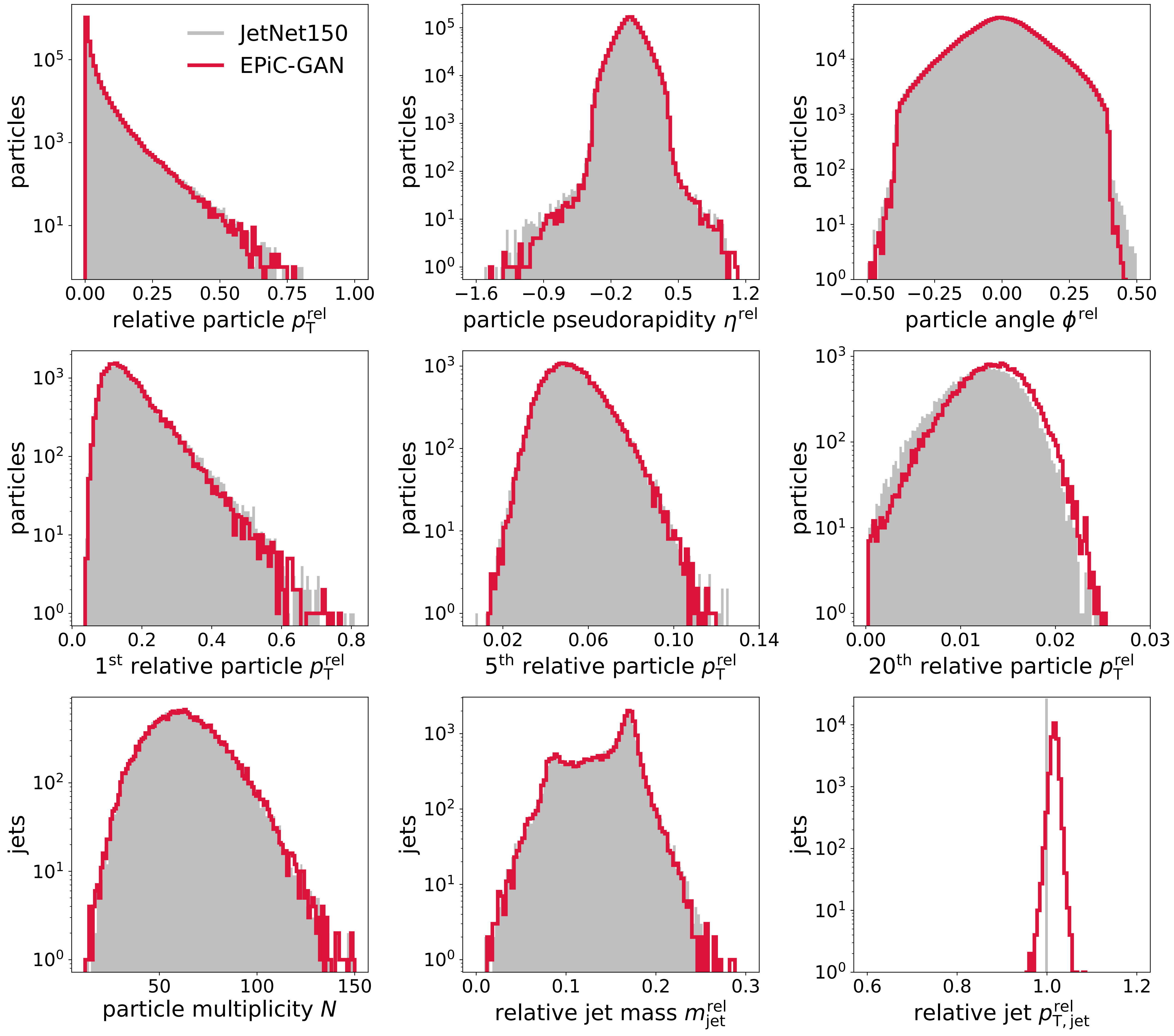
- EPiC-GAN setup:
    - $L_{\text{gen}} = 6, L_{\text{discr}} = 3$ EPiC layers
    - $dim(\boldsymbol{g}) = 10$ global attributes
- High generative fidelity after six equivariant update steps
- Distributions well represented by EPiC-GAN, jet mass distributions particularly challenging
- Very similar performance to state-of-the-art message passing (MP)-GAN[6]



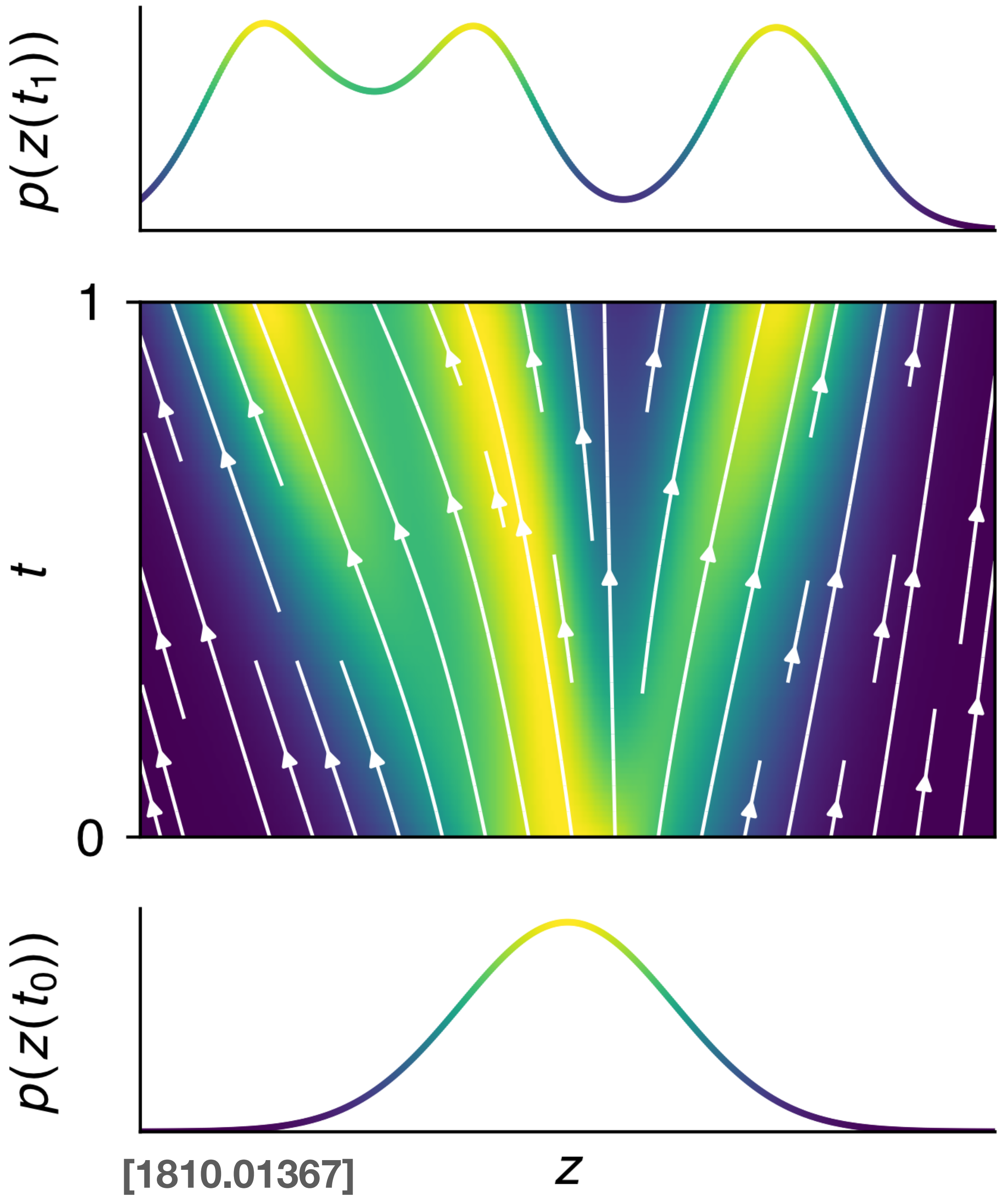
[6]Particle Cloud Generation with Message Passing Generative Adversarial Networks
R Kansal, et al.; arXiv: 2106.11535v3

Results on JetNet150 tops

- EPiC-GAN setup:
  - $L_{\text{gen}} = 6$, $L_{\text{discr}} = 3$ EPiC layers
  - $dim(\mathbf{g}) = 10$ global attributes
- Works well for up to 150 particles
- Distributions well represented by EPiC-GAN, jet mass distributions particularly challenging
- Sharp relative jet p_{T} distribution challenging; can be resolved by calibration



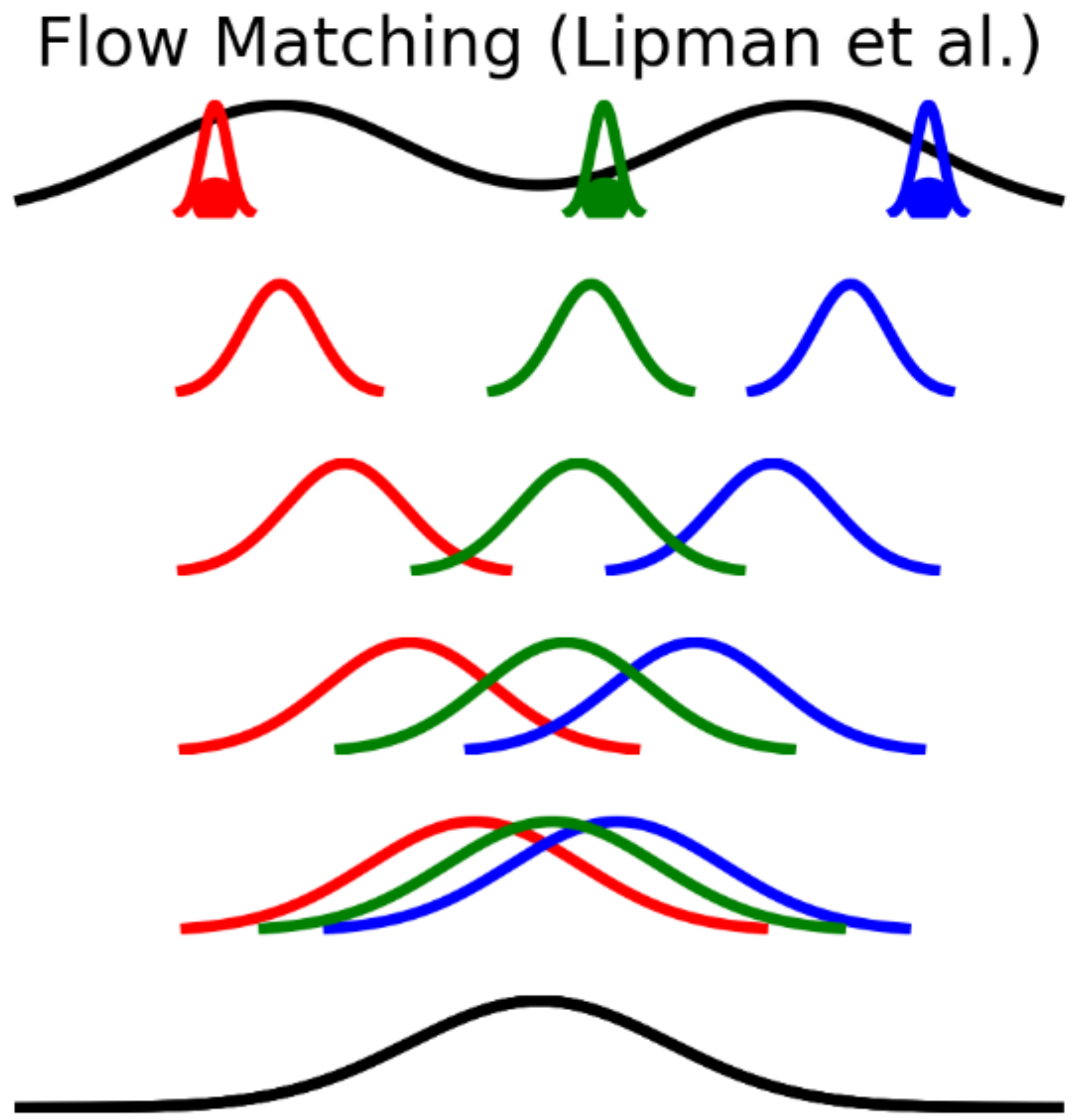
Flow Matching



Continuous Normalizing Flow (CNF)

- Training is difficult because ODE needs to be solved

$$\frac{\partial x_t}{\partial t} = v_\theta(x_t, t)$$



Flow Matching (FM)

- Simulation-free training objective (no ODE solving during training)
- Regressing against conditional flows
- Much faster training

$$L_{FM} = \left|\left| v_\theta(x_t) - u_t(x_t | x_0) \right|\right|^2$$

Flow Matching for Generative Modeling
Lipman et al.; arxiv:2210.02747

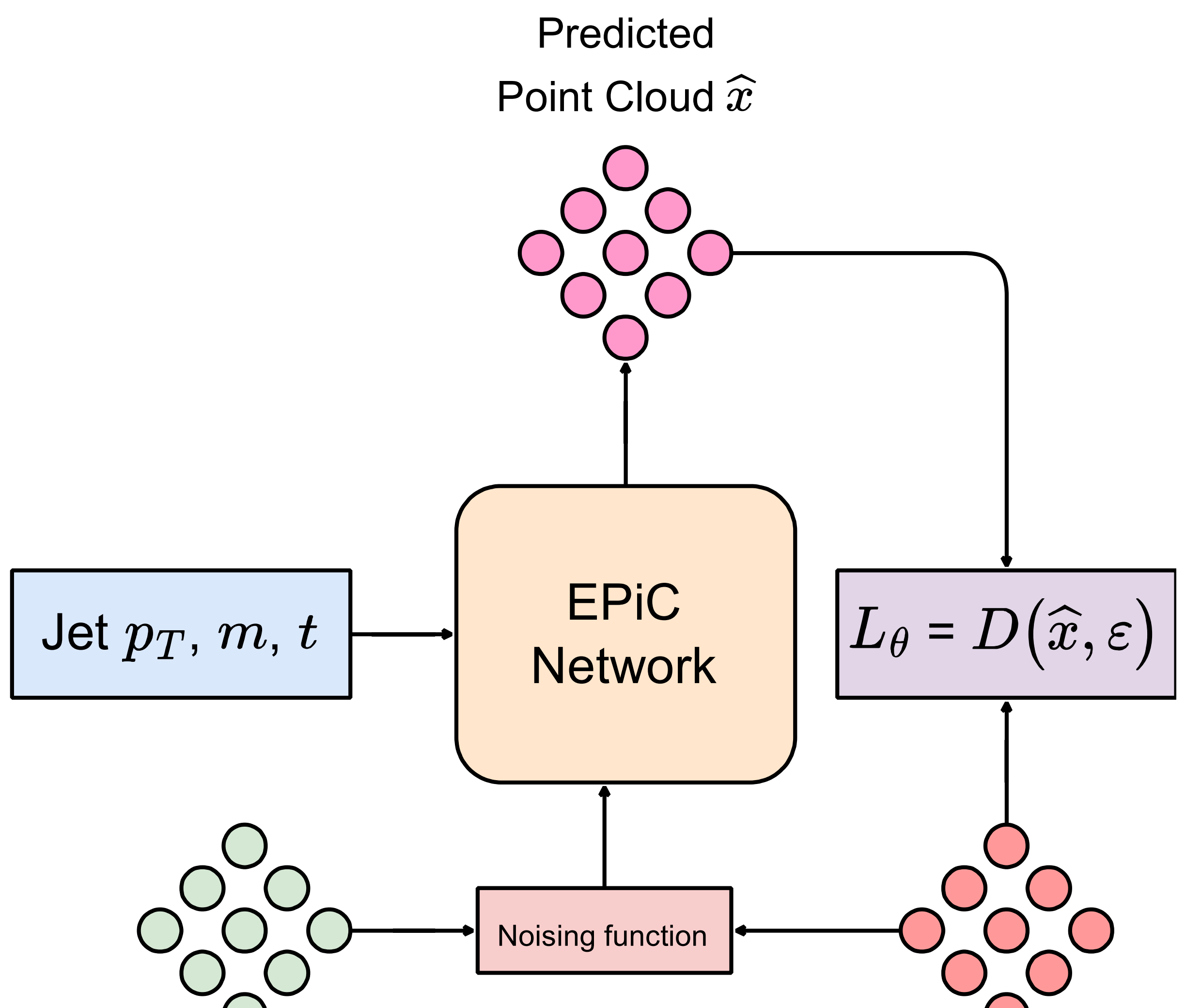
EPiC-FM & EpiC JeDi

- EPiC-FM: EPiC Architecture with Flow Matching

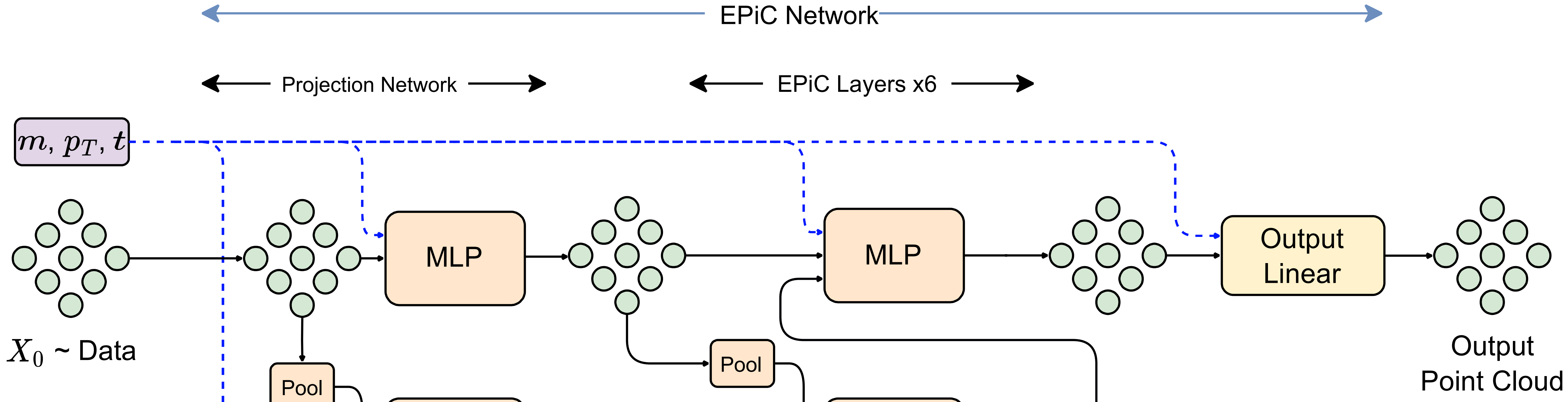
$$L_{FM}\left(v_\theta,\ u_t\left(x \middle| x_0\right)\right) = \left|\left| v_\theta(x_t, t) - \left(\left(1 - \sigma_{min}\right)\epsilon - x_0\right) \right|\right|^2$$

- EPiC-JeDi: EPiC Architecture with JeDi diffusion

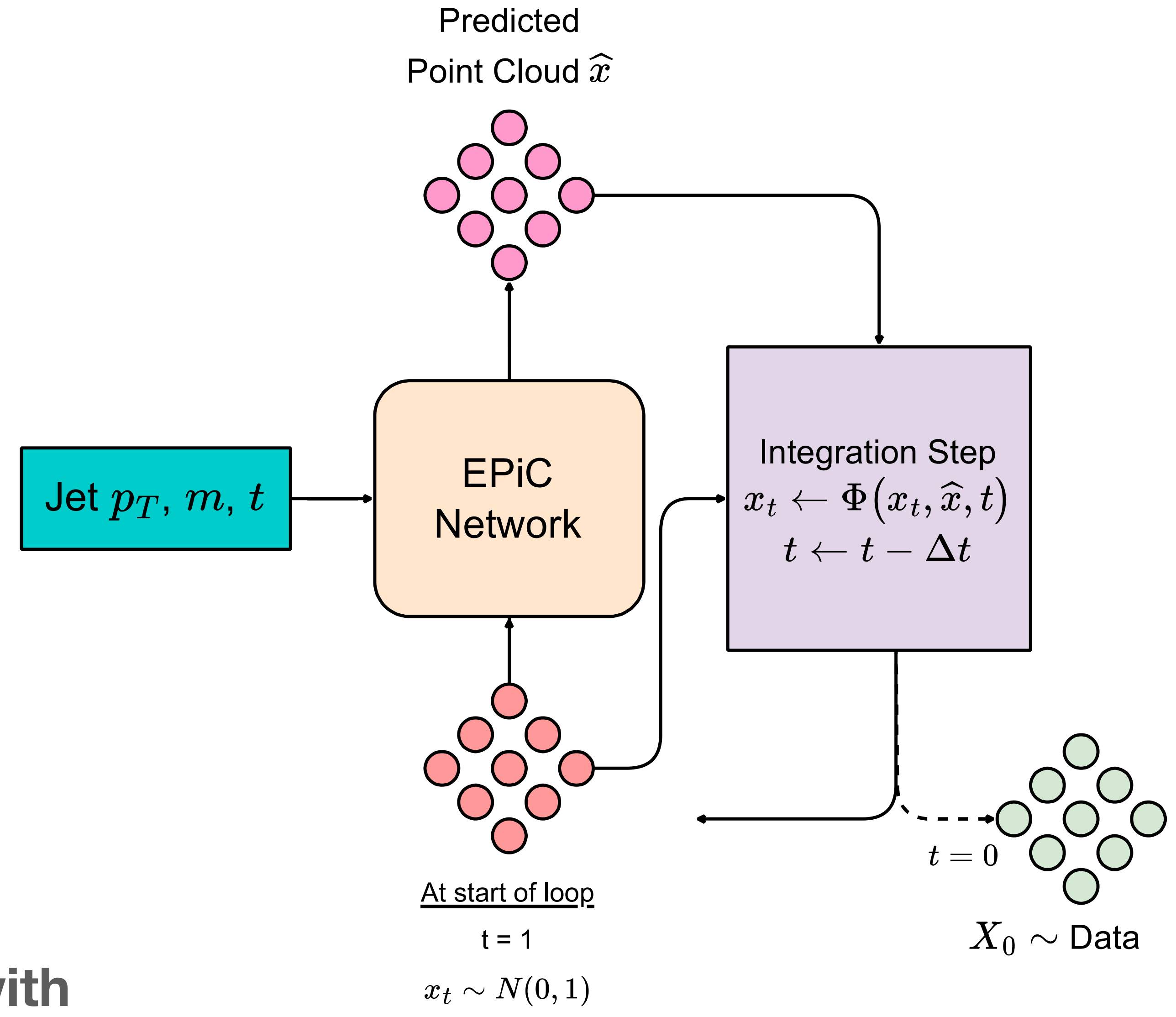
$$L_{JeDi}\left(v_\theta,\ s_t\left(x \middle| x_0\right)\right) = \left(1 - \alpha \frac{\beta(t)}{\sigma(t)^2}\right)\left|\left| v_\theta(x_t, t) - \epsilon \right|\right|^2$$



Training



EPiC Network



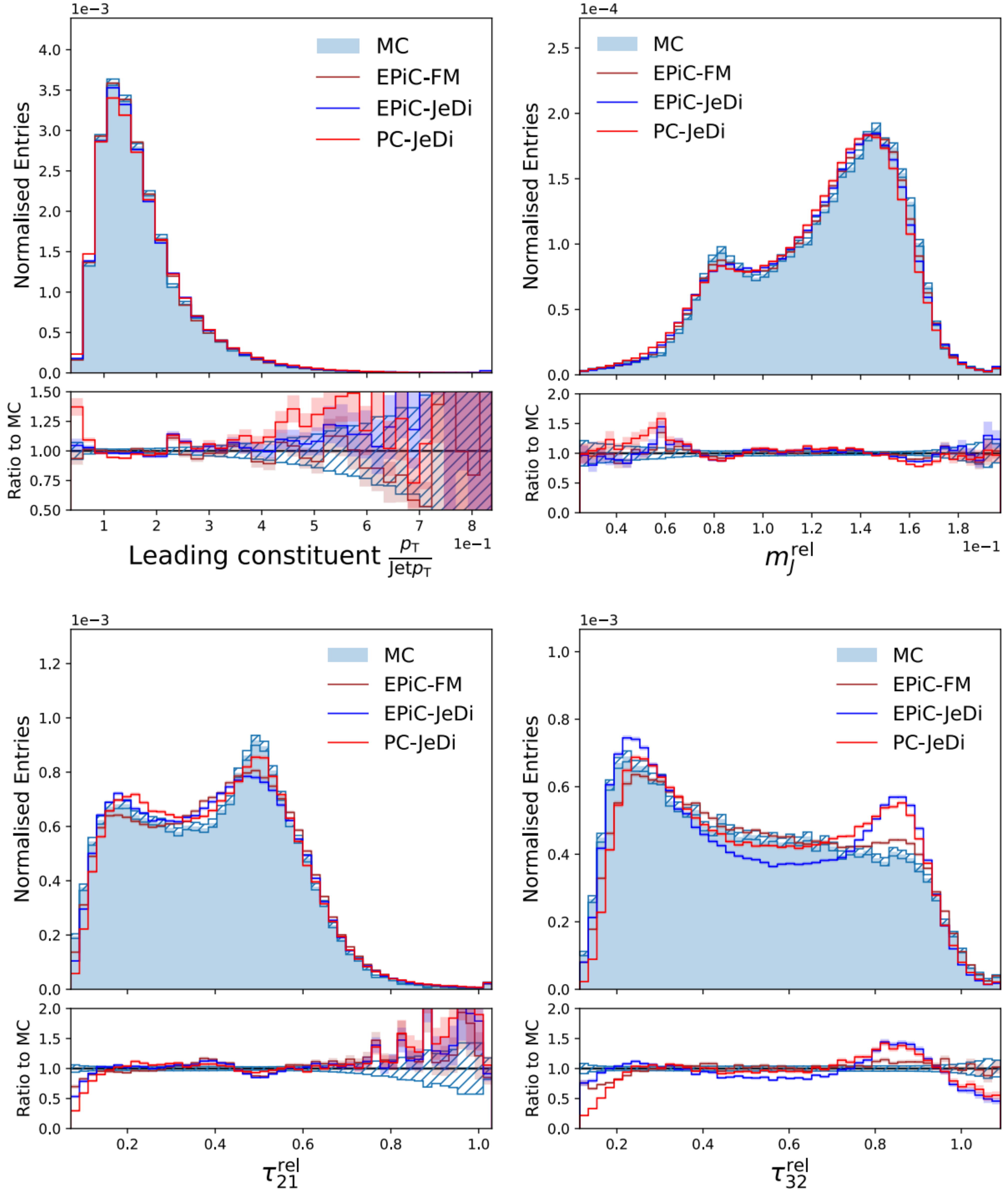
Sampling

EPiC-ly Fast Particle Cloud Generation with Flow-Matching and Diffusion
E. Buhmann, et al, arxiv: 2310.00049

Results on JetNet30

- Unconditioned and conditioned version (mass, p_T^{jet})
- Generate conditioning with normalizing flow
- Comparison to EPiC GAN and PC-JeDi
- Substructure most challenging to learn
- Similar results on JetNet150

Generation	Model	NLP	$\mathrm{KL}^{m}(\times 10^{-3})$	$\mathrm{KL}^{p_T^{\mathrm{const}}}(\times 10^{-3})$	$\mathrm{KL}^{\tau_{21}}(\times 10^{-3})$	$\mathrm{KL}^{\tau_{32}}(\times 10^{-3})$
Conditional	PC-JeDi	3.08	8.56 ± 0.75	3.25 ± 0.09	12.82 ± 1.16	27.08 ± 1.40
	EPiC-JeDi	3.1	5.26 ± 0.51	2.99 ± 0.05	$\mathbf{7.81 \pm 0.61}$	17.34 ± 1.08
	EPiC-FM	**1.35**	$\mathbf{3.77 \pm 0.50}$	$\mathbf{2.03 \pm 0.02}$	$\mathbf{7.40 \pm 0.64}$	$\mathbf{8.09 \pm 0.93}$
Unconditional	EPiC-GAN	3.43	$\mathbf{3.71 \pm 0.42}$	3.33 ± 0.03	$\mathbf{8.28 \pm 0.76}$	17.68 ± 0.91
	EPiC-JeDi	3.11	18.42 ± 1.12	3.73 ± 0.08	$\mathbf{8.00 \pm 0.80}$	15.27 ± 1.35
	EPiC-FM	1.38	5.80 ± 0.54	$\mathbf{2.03 \pm 0.01}$	$\mathbf{7.69 \pm 0.71}$	$\mathbf{9.24 \pm 1.00}$



PC-JeDi: Diffusion for Particle Cloud Generation in High Energy Physics
M. Leigh et al.; arXiv: 2303.05376

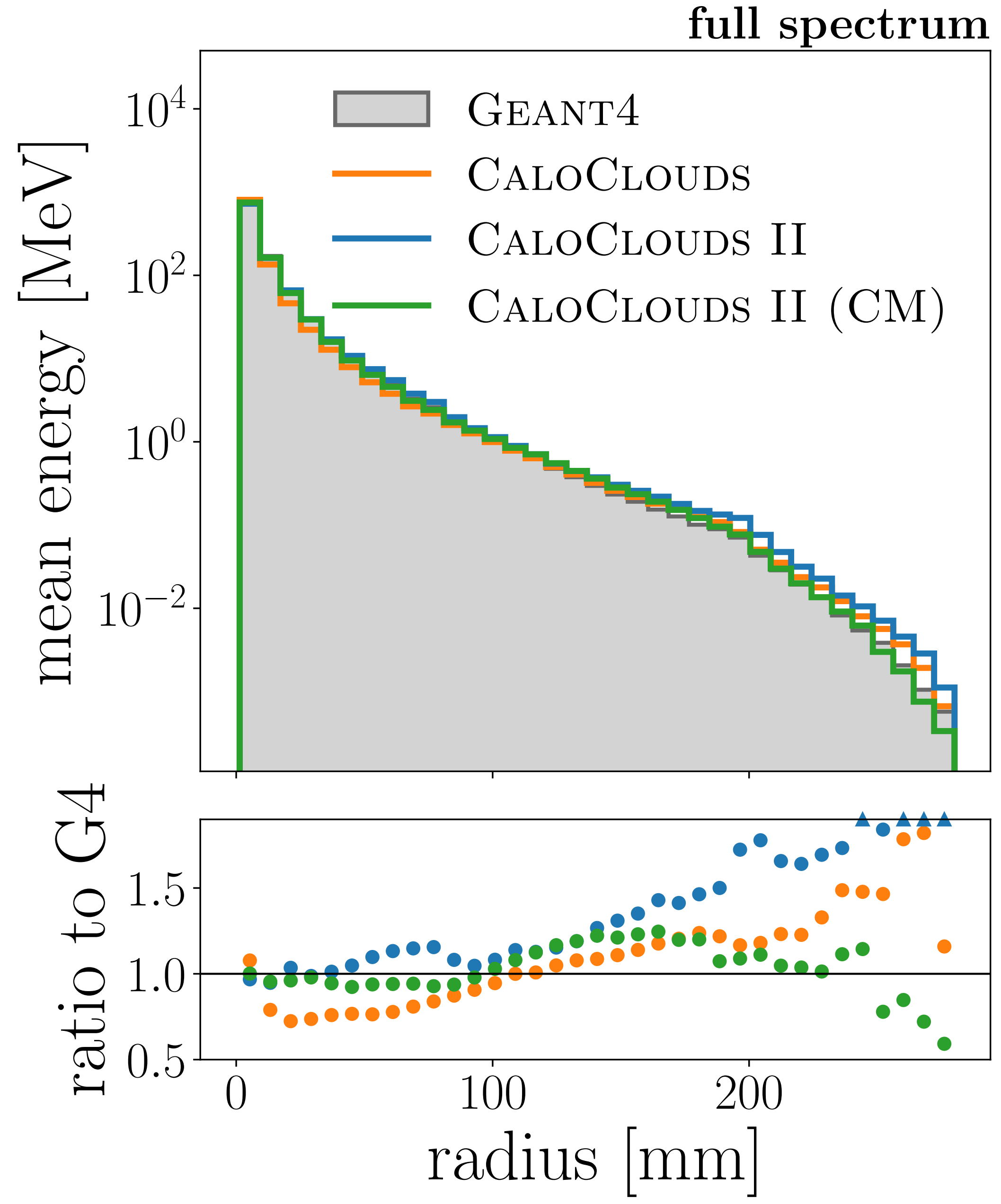
Summary

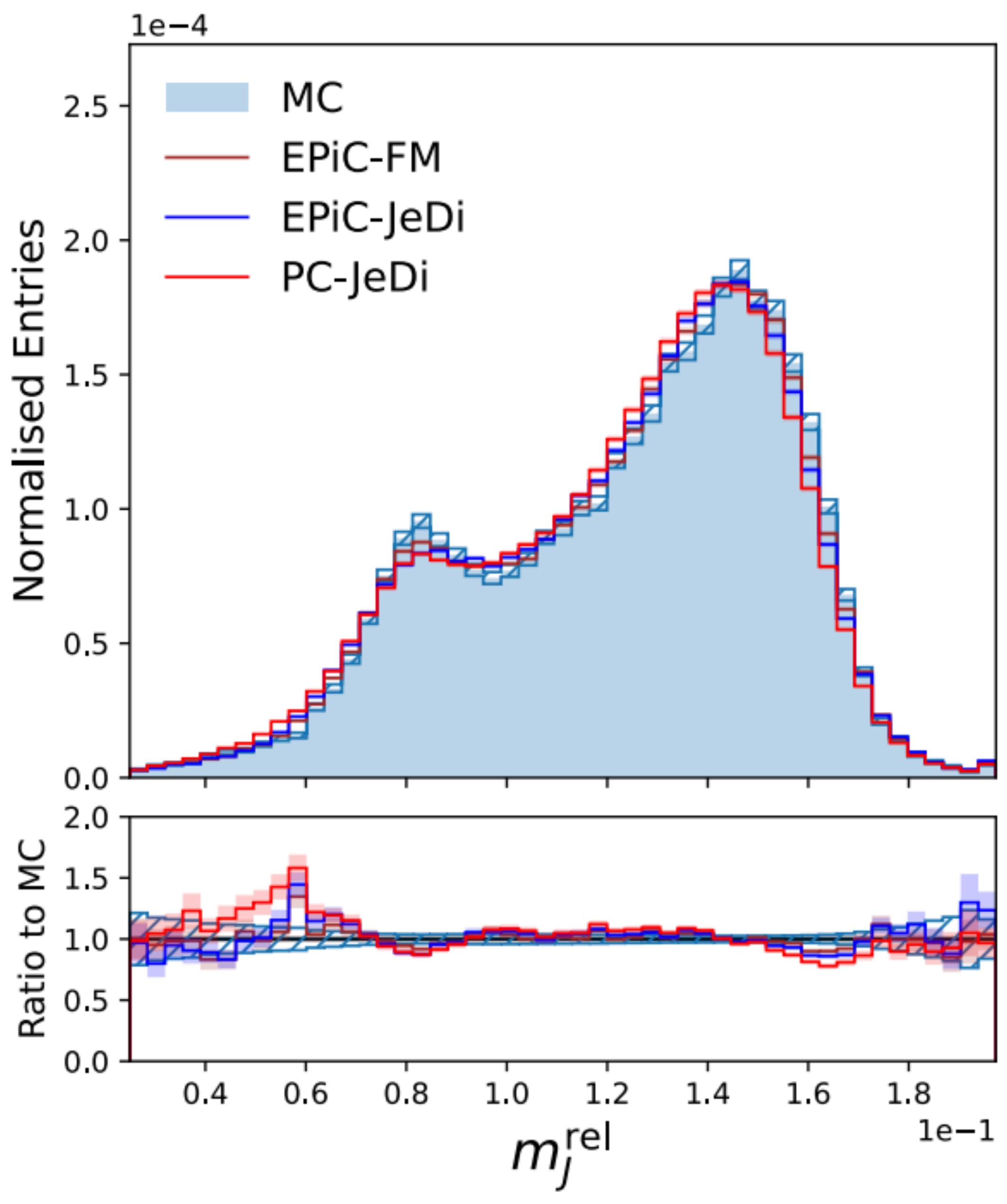
- **BIB-AE** with angular conditioning nice distributions after full reconstruction with PandoraPFA

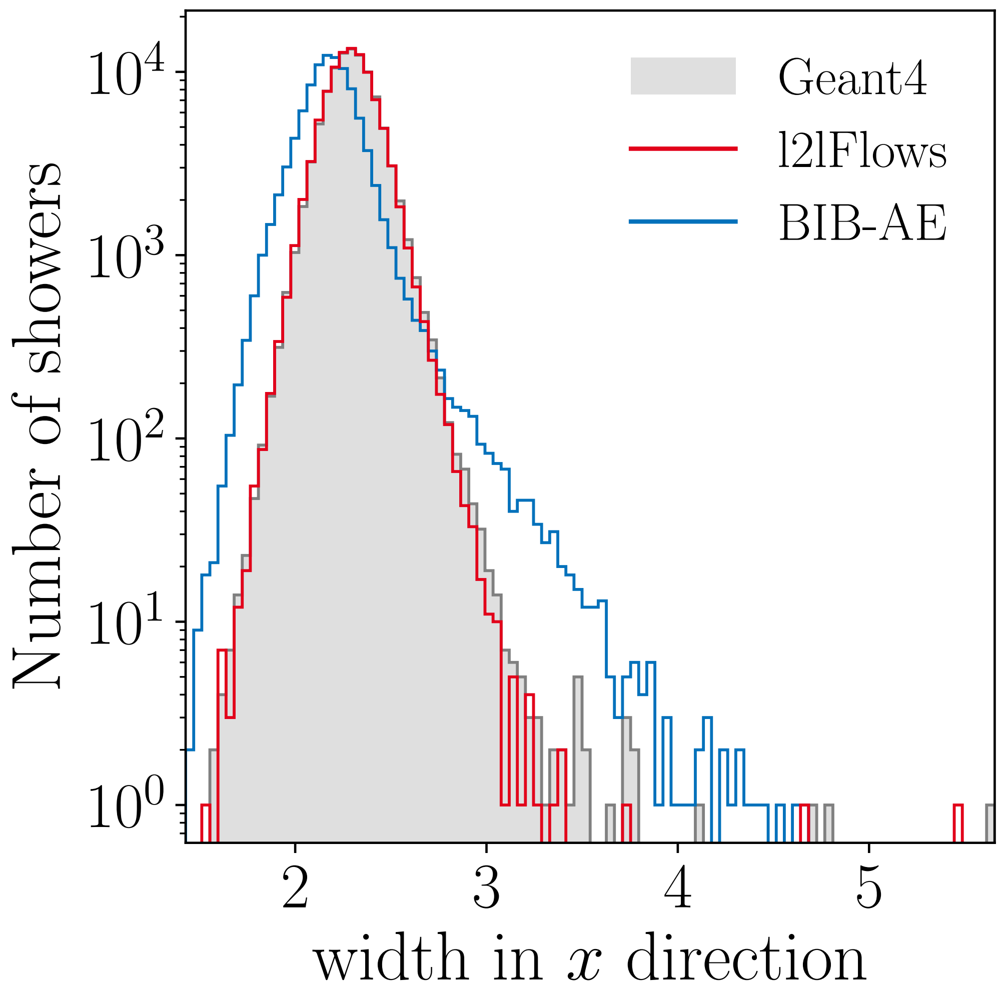
- **Layer-to-Layer Flows** with improved performance over BiB-AE

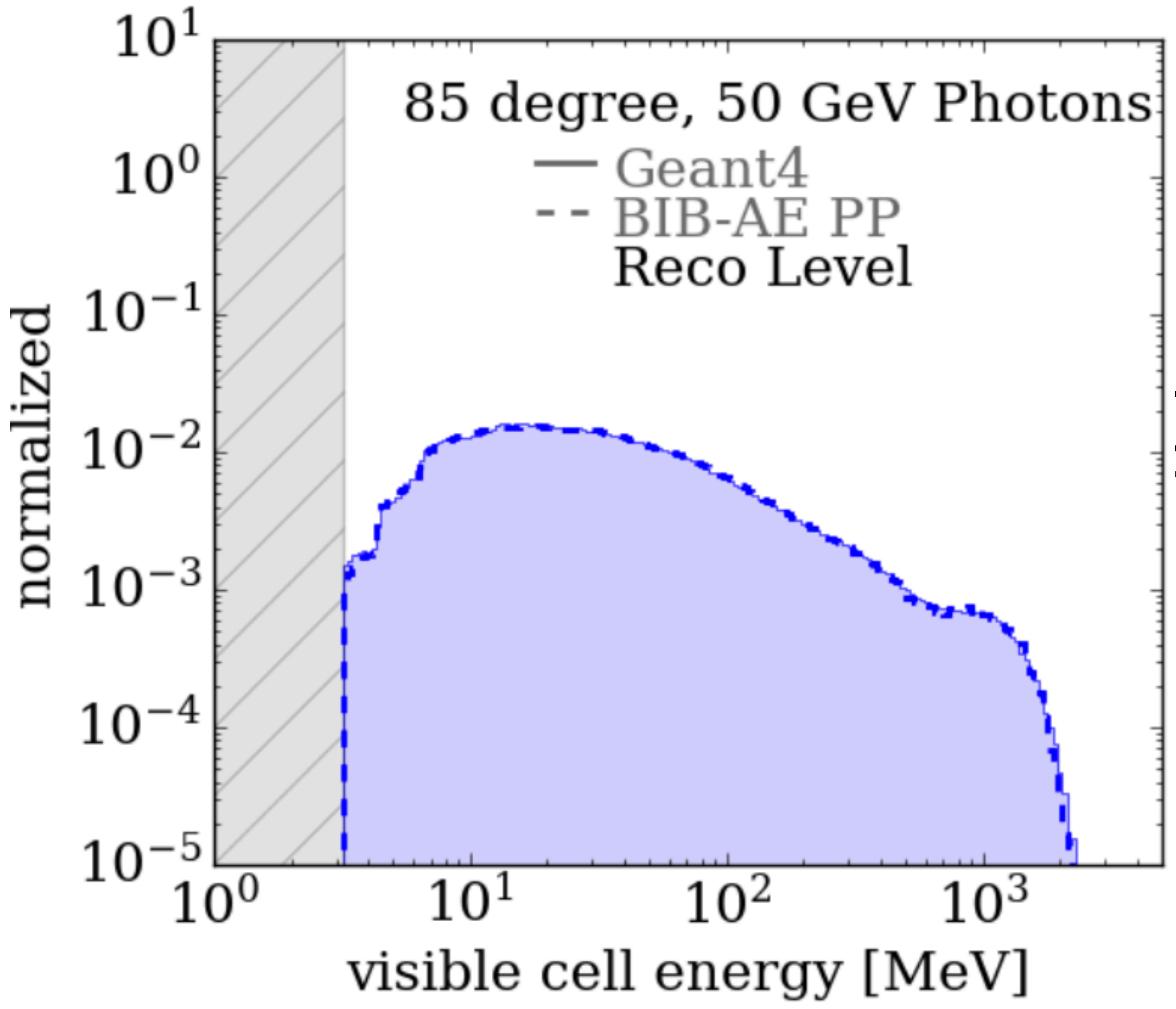
- **CaloClouds** produces high fidelity and geometry independent showers

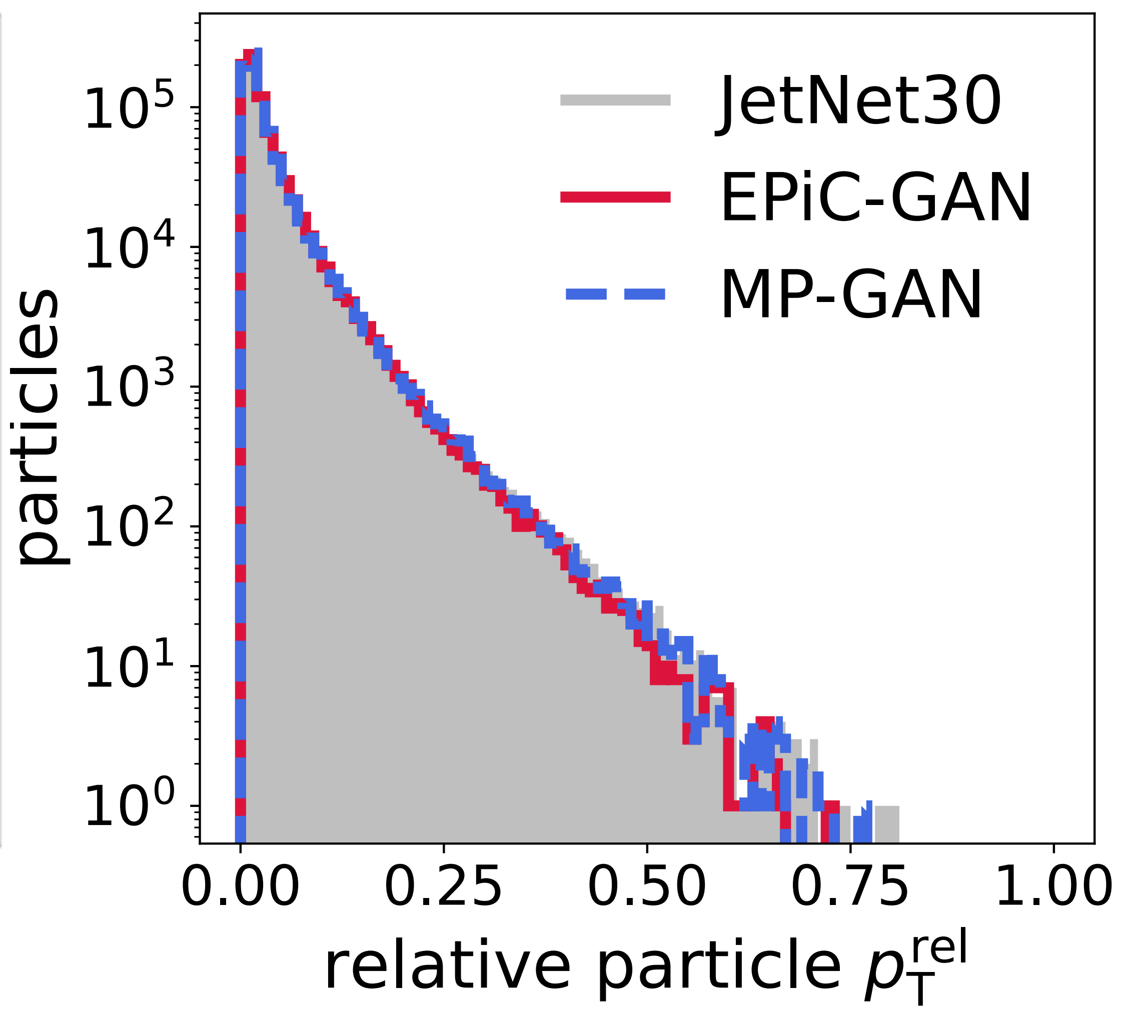
- **EPiC-GAN** and **EPiC-FM** good performances on JetNet30











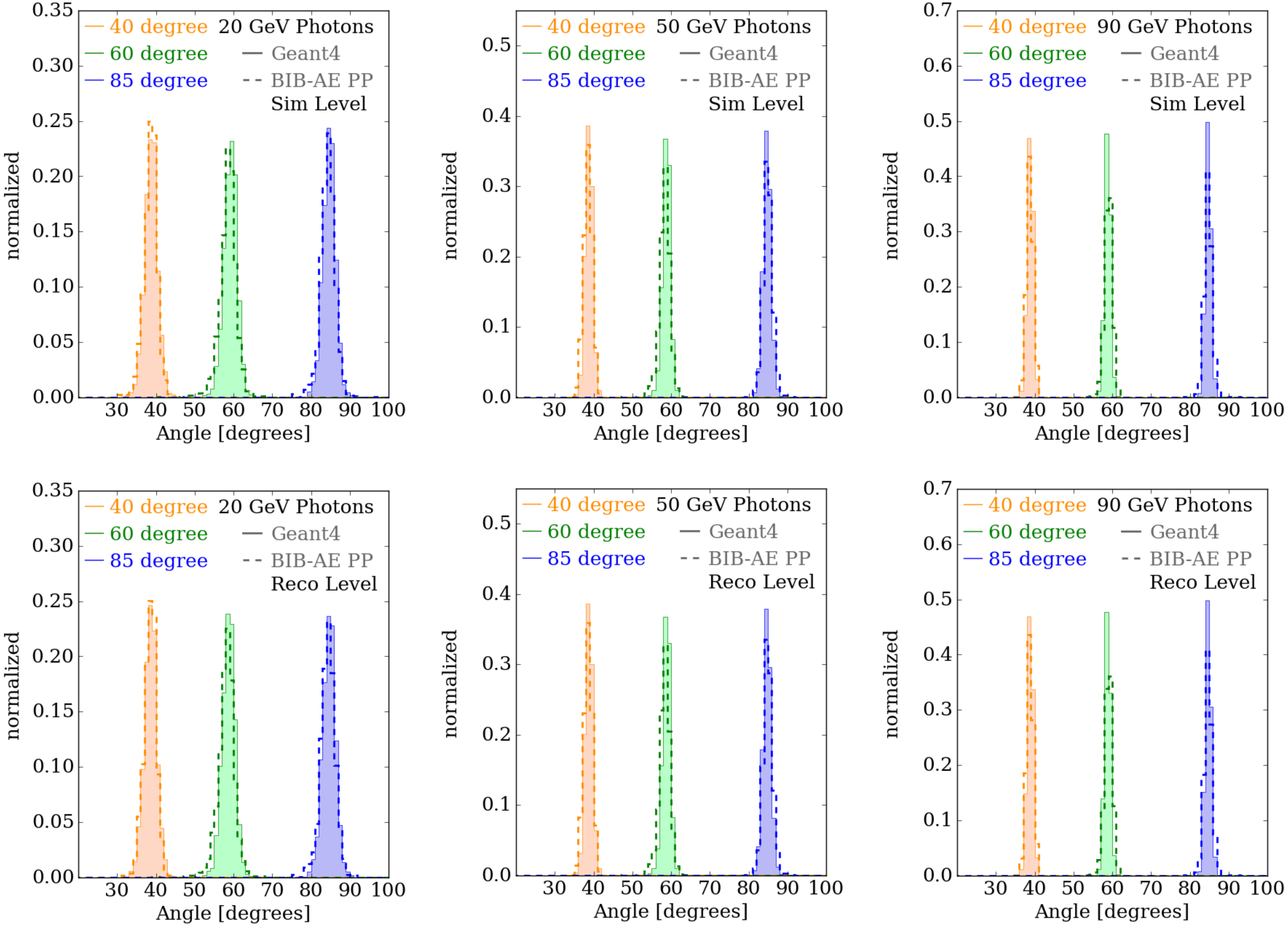
Thank you

Backup Slides

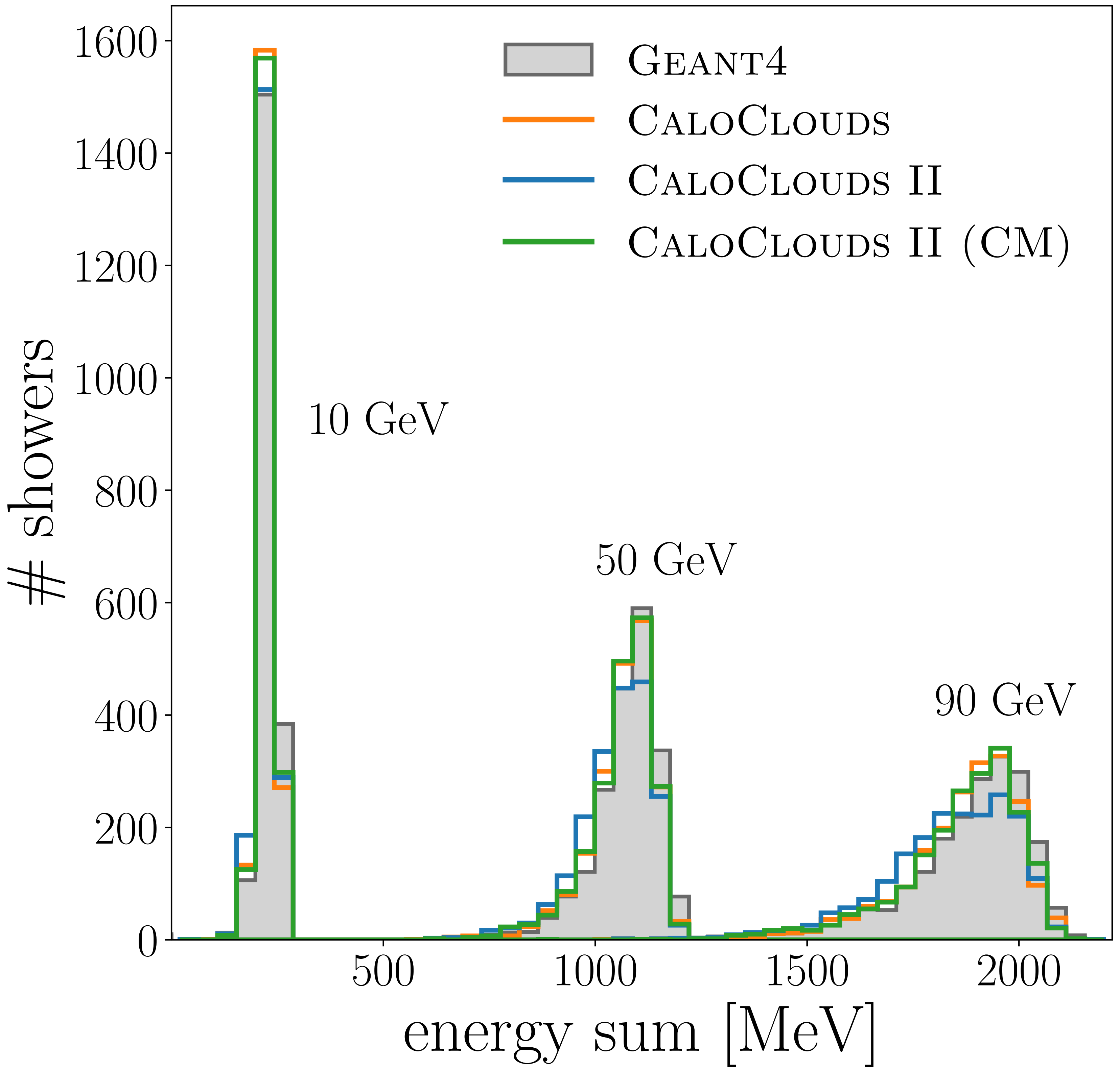
BiB-AE: Photons and Pions

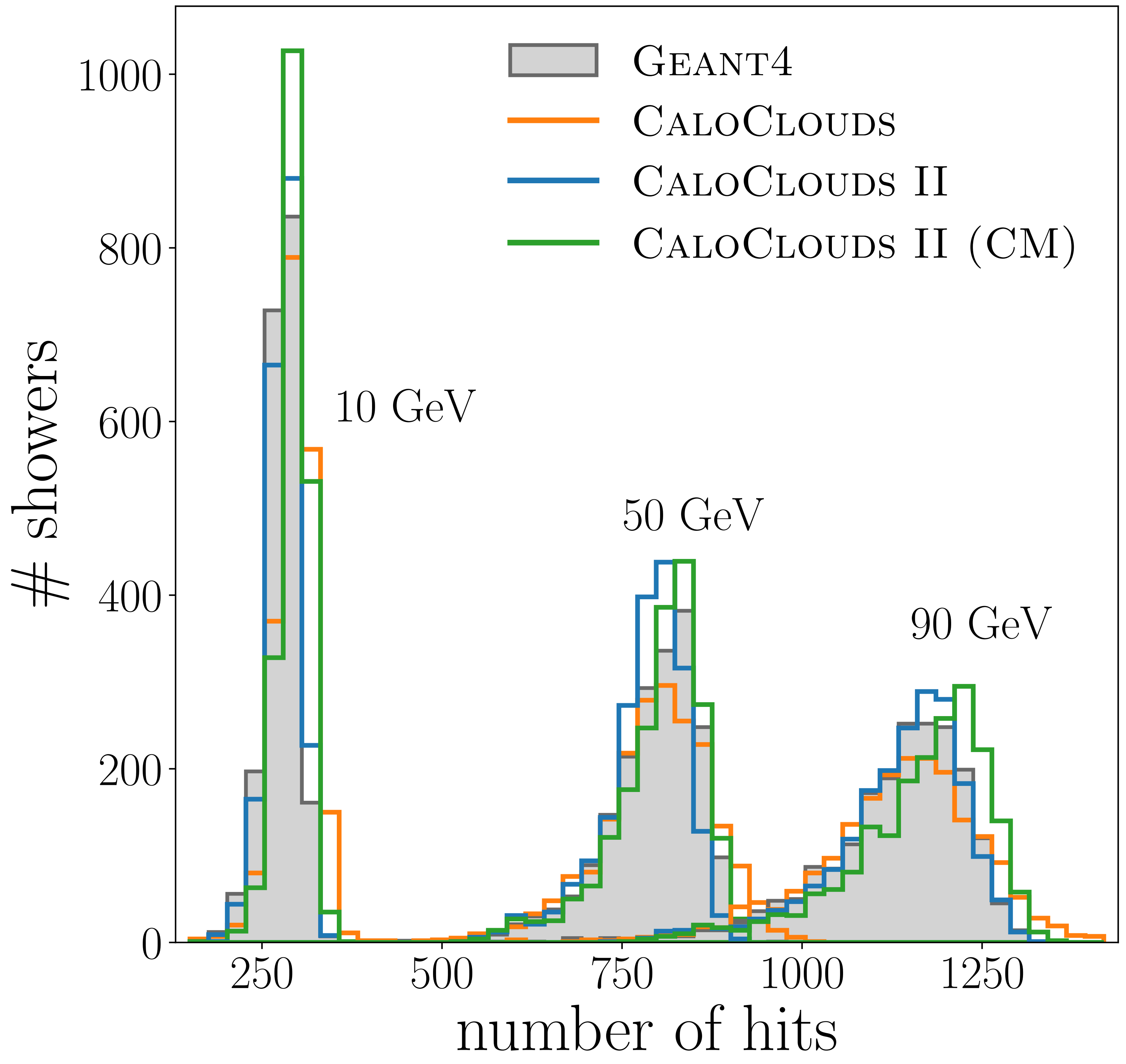
- **Sim** level angle reconstruction

- **Rec** level angle reconstruction
  - After full reconstruction with PandoraPFA



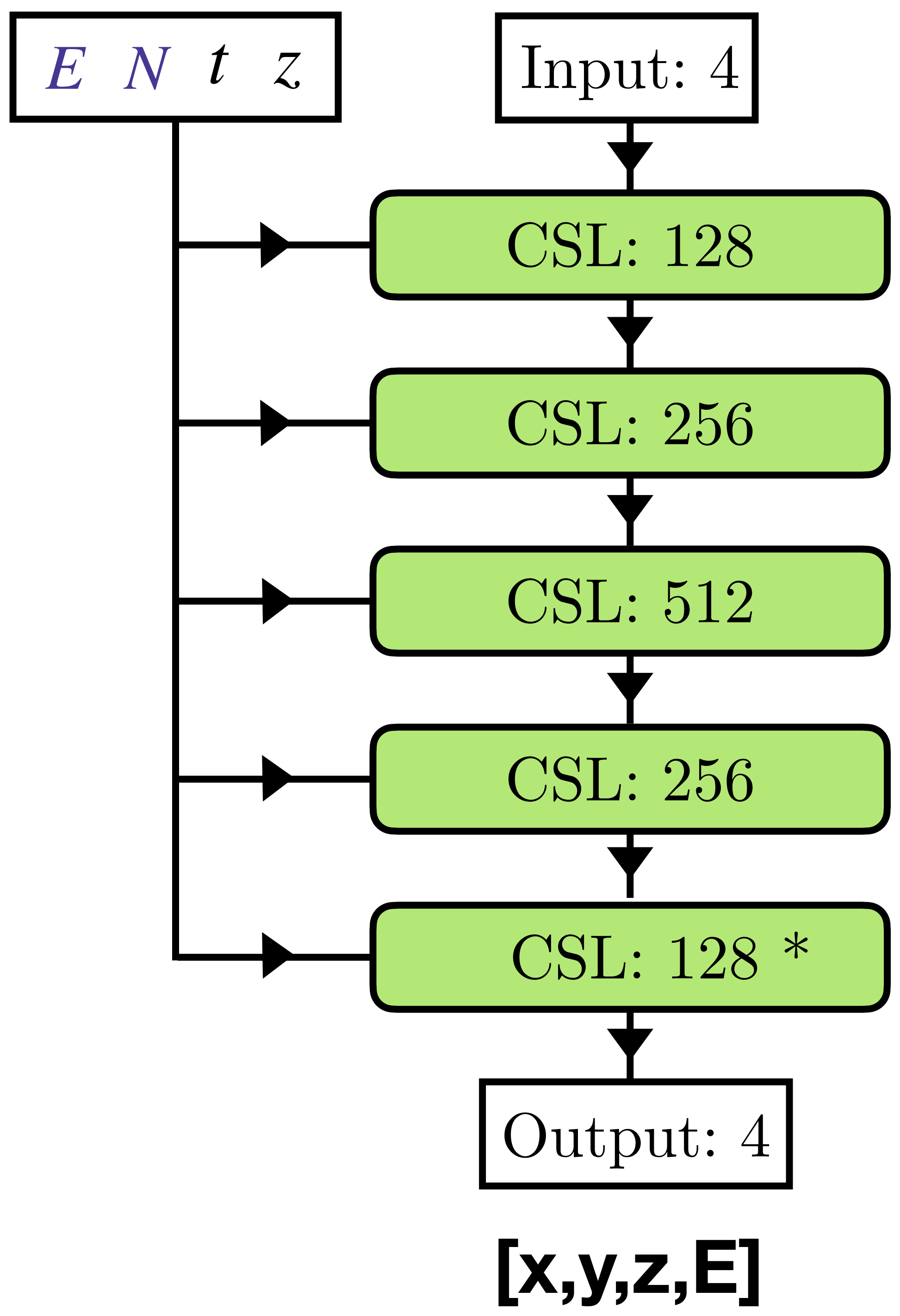
Results: Single Incident Energies

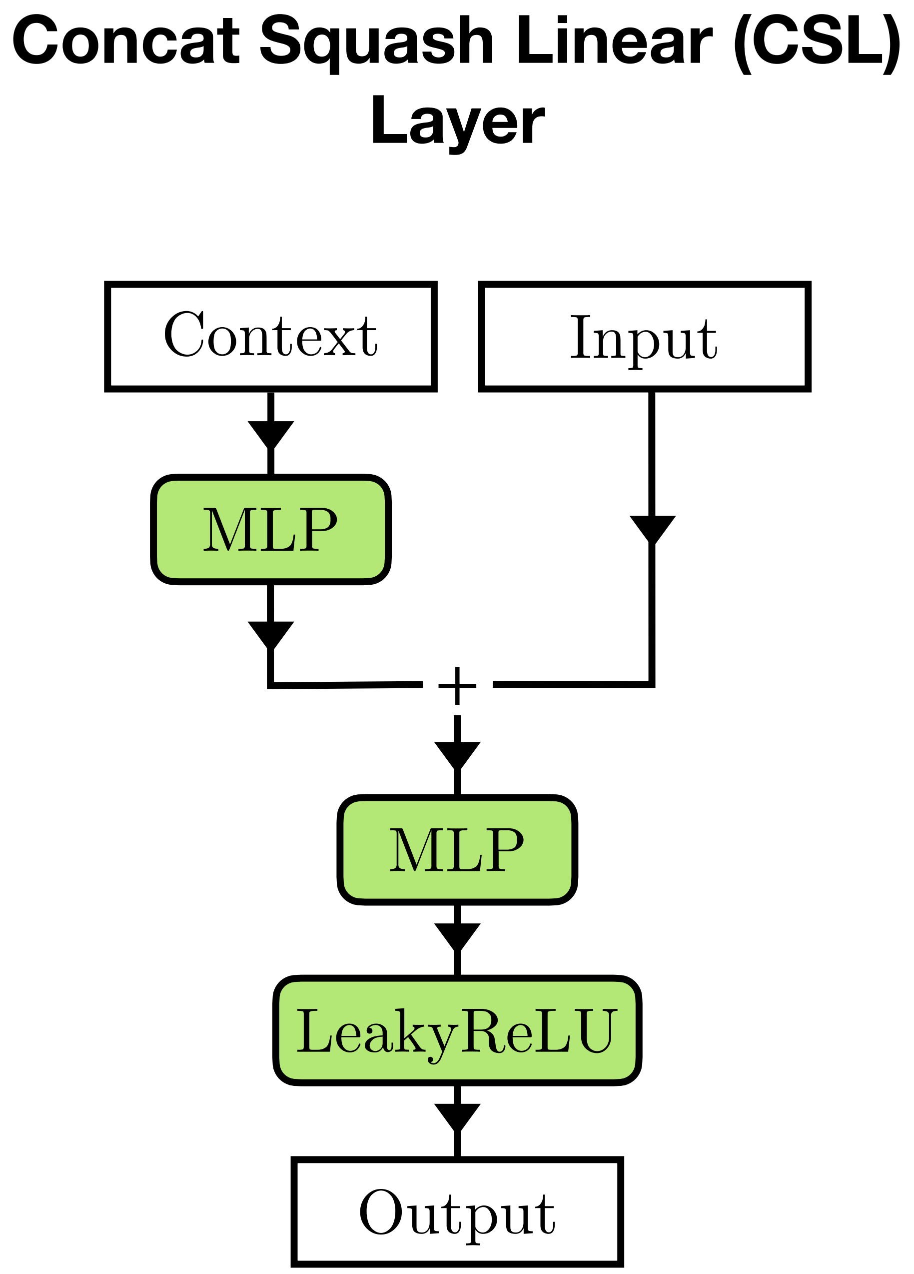


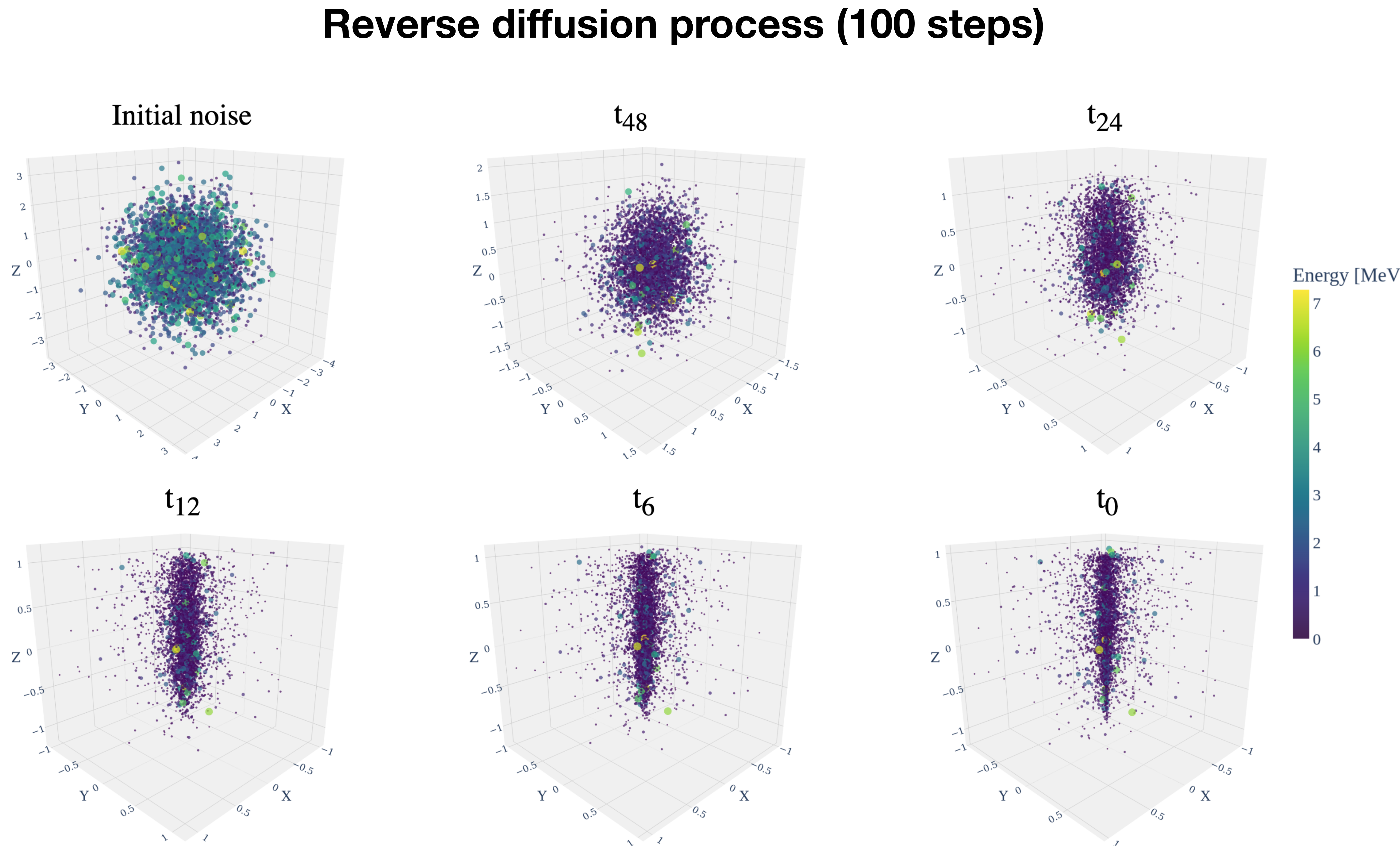


- Total visible energy & number of cell hits for single incident energies well modelled by all three CaloClouds variants (2,000 showers each)
- Number of hits better modeled by the CaloClouds II variants
- Expecting further improvements with wider energy range during training

CaloClouds Models







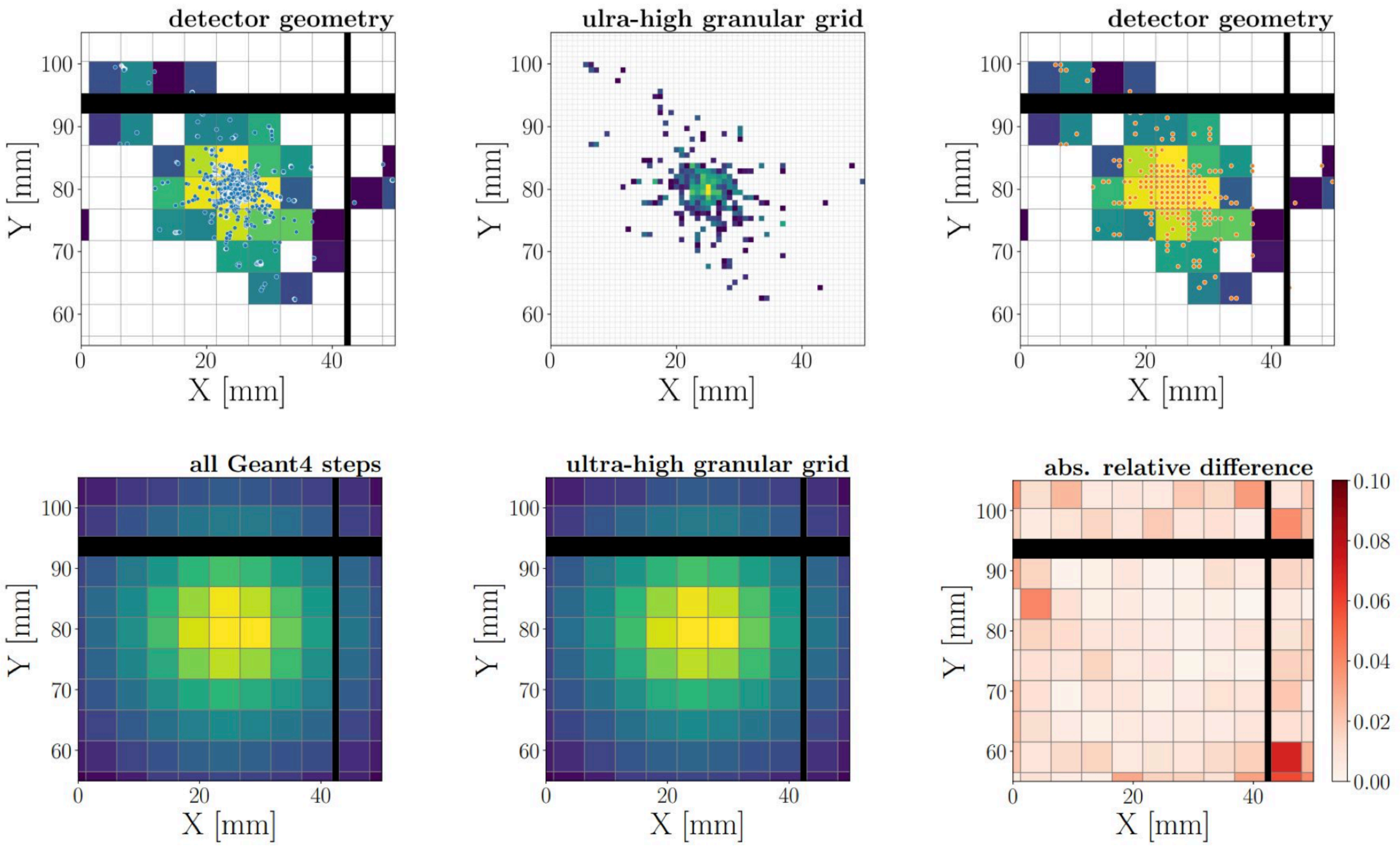
- Weight sharing across all points
- Each point independently sampled
- No interaction between points → very fast sampling

Architecture from:
Diffusion Probabilistic Models for 3D Point Cloud Generation
S Luo, W Hu; arXiv: 2103.0145

EM Showers: effects of pre-clustering

Single event of 90 GeV shower in 21th layer

2k events of 90 GeV showers in 21th layer, overlay



CaloClouds Results: Metrics & Classifier

Simulator	$W_1^{N_{\text{hits}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{vis}}/E_{\text{inc}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{cell}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{long}}}$ $(\times 10^{-3})$	$W_1^{E_{\text{radial}}}$ $(\times 10^{-3})$	$W_1^{m_{1,X}}$ $(\times 10^{-3})$	$W_1^{m_{1,Y}}$ $(\times 10^{-3})$	$W_1^{m_{1,Z}}$ $(\times 10^{-3})$
Geant4	0.7 ± 0.2	0.8 ± 0.2	0.9 ± 0.4	0.7 ± 0.8	0.7 ± 0.1	0.9 ± 0.1	1.1 ± 0.3	0.9 ± 0.3
CaloClouds	$\mathbf{2.5 \pm 0.3}$	11.4 ± 0.4	15.9 ± 0.7	$\mathbf{2.0 \pm 1.3}$	38.8 ± 1.4	4.0 ± 0.4	8.7 ± 0.3	1.4 ± 0.5
CaloClouds II	3.6 ± 0.5	26.4 ± 0.4	$\mathbf{15.3 \pm 0.6}$	3.7 ± 1.6	11.6 ± 1.5	$\mathbf{2.4 \pm 0.4}$	$\mathbf{7.6 \pm 0.2}$	3.9 ± 0.4
CaloClouds II (CM)	6.1 ± 0.7	$\mathbf{9.8 \pm 0.5}$	16.0 ± 0.7	$\mathbf{2.0 \pm 1.4}$	$\mathbf{8.3 \pm 1.9}$	3.0 ± 0.4	9.5 ± 0.6	$\mathbf{1.2 \pm 0.5}$

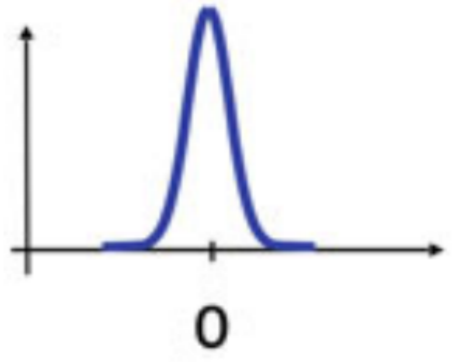
- Evaluation metrics based on 1-Wasserstein distance*
- Similar performance between all three CaloClouds versions
- High-level classifier "metric": CaloClouds II better than CaloClouds

High-level fully connected classifier:

Simulator	AUC
CaloClouds	0.999 ± 0.001
CaloClouds II	0.928 ± 0.001
CaloClouds II (CM)	$\mathbf{0.923 \pm 0.001}$

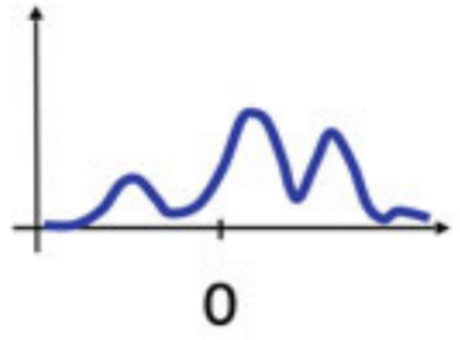
***Particle Cloud Generation with Message Passing Generative Adversarial Networks**
R Kansal, et al.; arXiv:2106.11535

Continuous Normalizing Flow



$x_0 \sim p_0$

$$x_t := f(x_t,\ t) \qquad \frac{\partial x_t}{\partial t} = v_\theta(x_t, t)$$



$x_1 \sim p_1$

Normalizing Flow (NF)

Training:

$$\log p_T(x_T) = \log p_0(x_0) - \log \left| \frac{\partial f_t^\theta}{\partial x_t} \right|$$

Sampling:

$$x_T = f_{T-1} \circ \ldots \circ f_0(x_0)$$

- f must be invertible
- Determinant computationally expensive
  - Restricted transformations needed

Continuous Normalizing Flow (CNF)

$$\log p_1(x_1) = \log p_0(x_0) - \int_{t_0}^{t} Tr\left(\frac{\partial v_\theta}{\partial x_t}\right) dt$$

Solve ODE (ordinary differential equation)

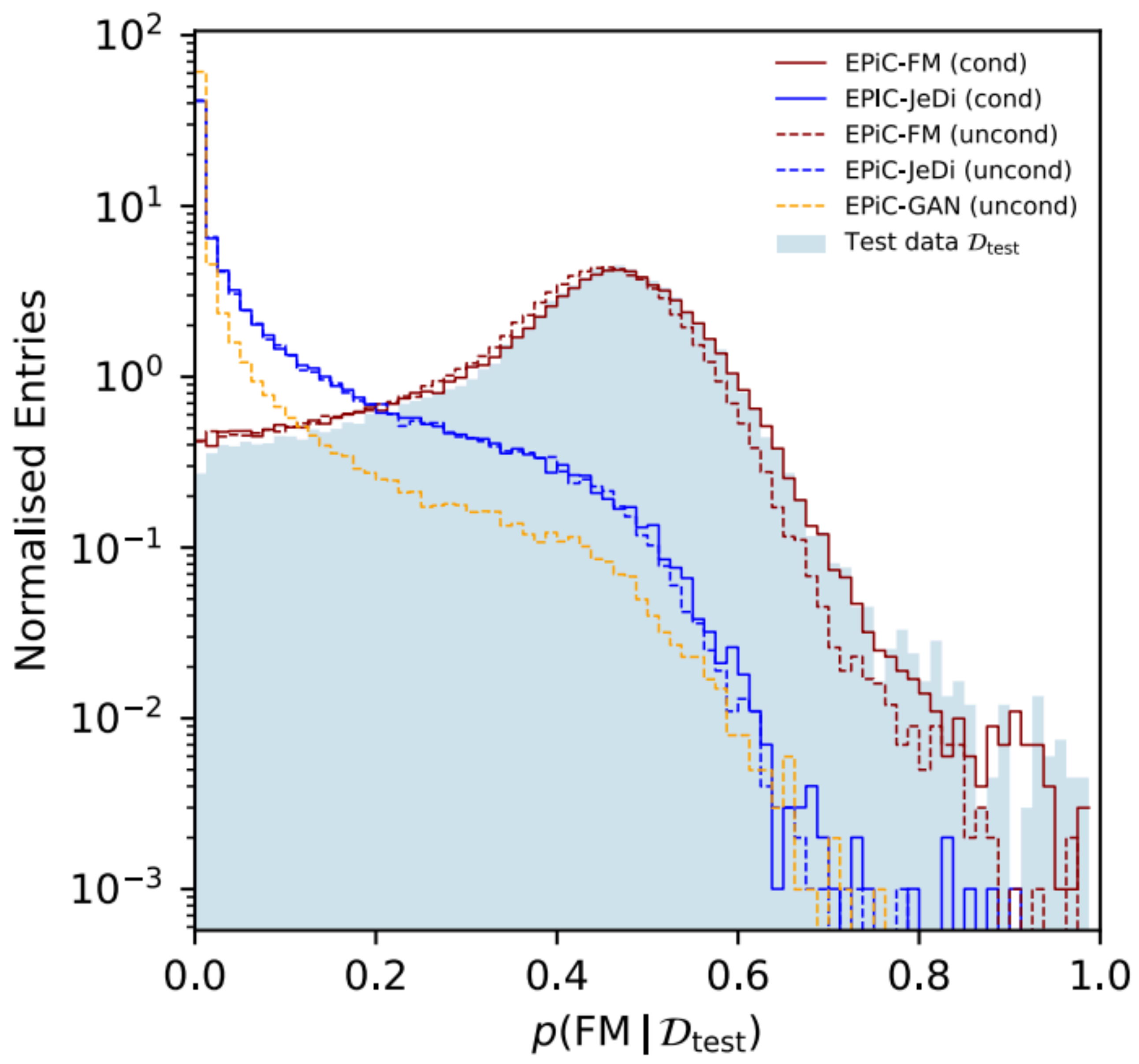
- f has no restrictions
- Trace is easier to calculate
- Still computationally expensive

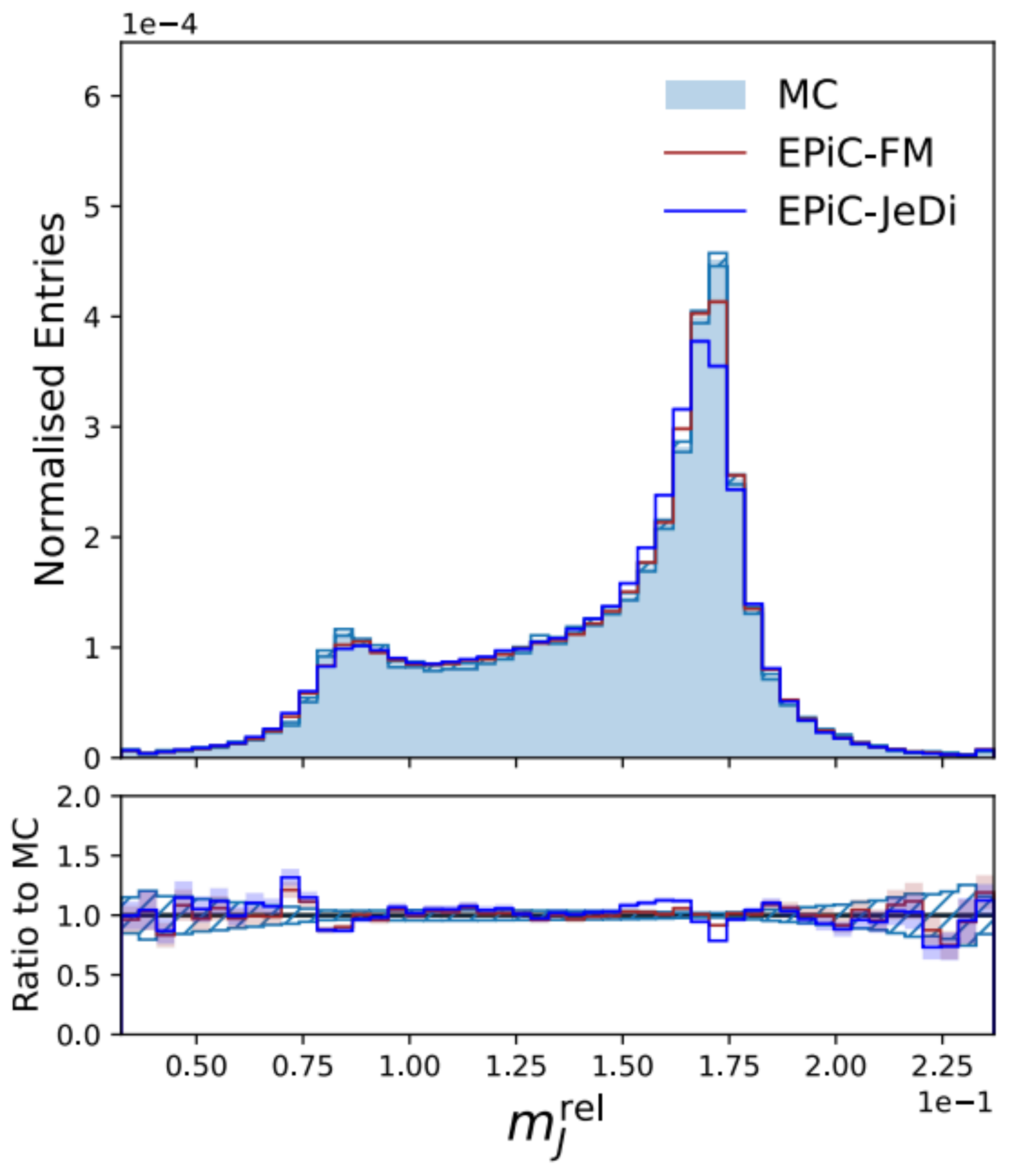
Neural Ordinary Differential Equations
Chen et al.; arxiv:1806.07366

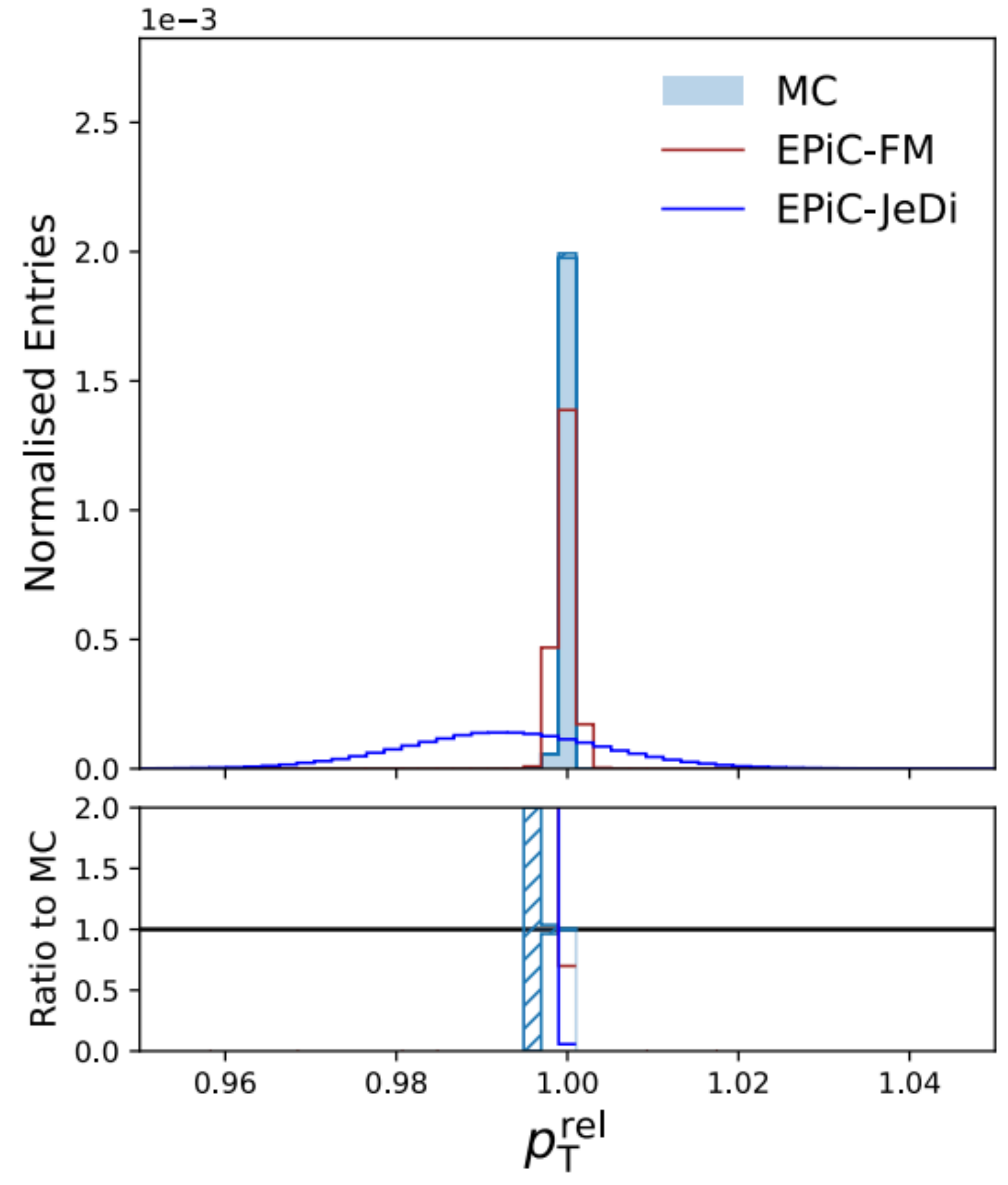
Results on JetNet150

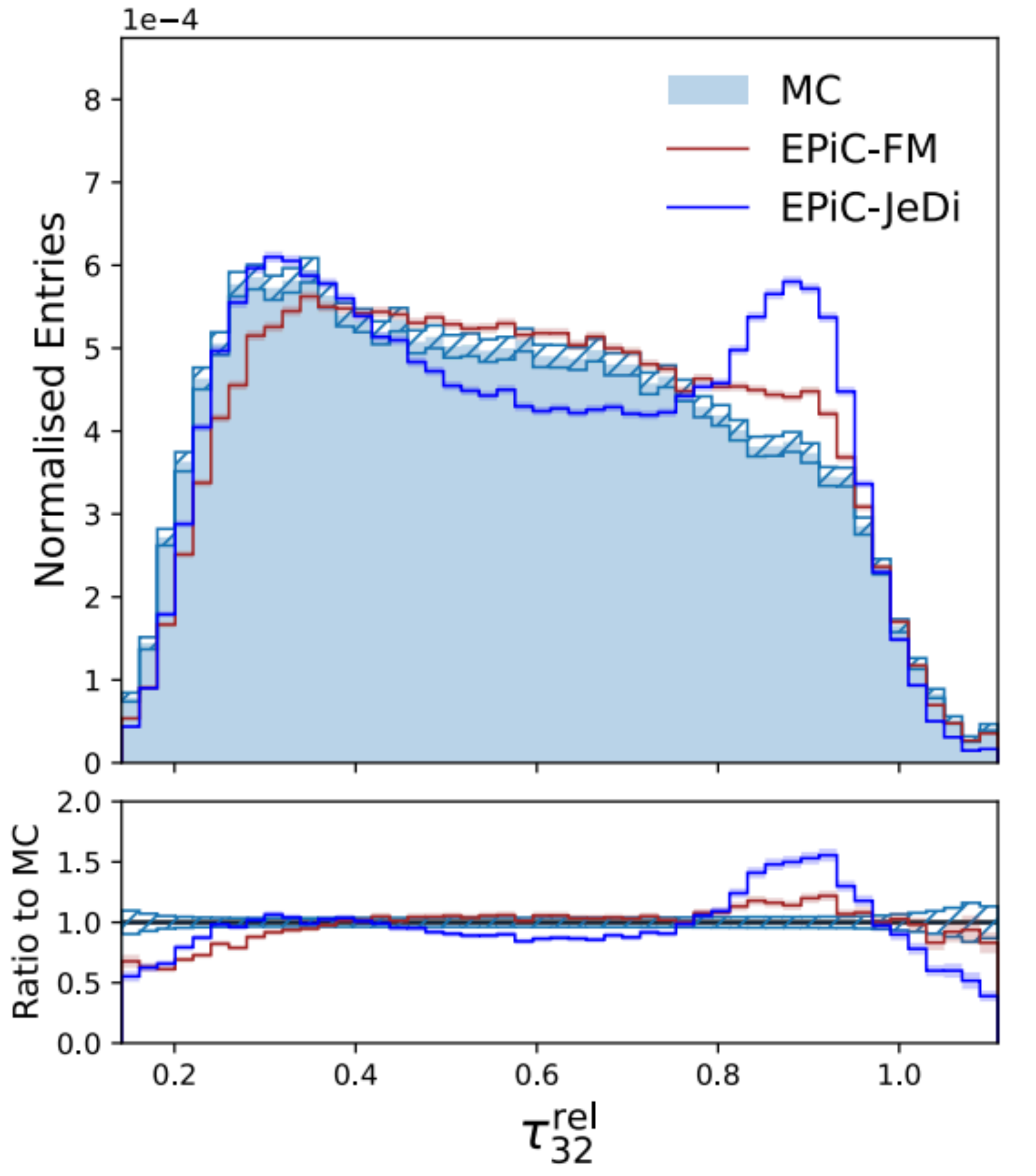
- Conditioned on mass and pT vs. unconditioned
- Comparison to EPiC GAN
- PC-JeDi too slow for 150 particles

- Similar behavior to JetNet30
- EPiC-FM outperforms all models
- Conditioned models are slightly better









Generation	Model	NLP	KL$^{m}(\times10^{-3})$	KL$^{p_T^{const}}(\times10^{-3})$	KL$^{\tau_{21}}(\times10^{-3})$	KL$^{\tau_{32}}(\times10^{-3})$
Conditional	EPiC-JeDi	5.67	9.10 ± 0.79	6.42 ± 0.76	14.32 ± 1.08	19.92 ± 1.21
	EPiC-FM	**0.12**	$\mathbf{4.30\pm0.53}$	$\mathbf{0.84\pm0.02}$	$\mathbf{9.43\pm0.61}$	$\mathbf{11.22\pm1.02}$
Unconditional	EPiC-GAN	11.6	$\mathbf{6.50\pm0.63}$	2.22 ± 0.09	20.60 ± 1.55	69.64 ± 3.30
	EPiC-JeDi	5.70	27.46 ± 1.24	6.39 ± 0.60	20.15 ± 1.25	36.50 ± 1.81
	EPiC-FM	0.98	12.95 ± 0.90	$\mathbf{0.87\pm0.02}$	$\mathbf{10.59\pm0.88}$	$\mathbf{12.14\pm0.97}$

Results on JetNet30

- Better scaling behavior for EPiC layers
  - 6.2x faster at 150 particles
- Effect increases for larger point clouds like calorimeter showers
- Slower than GANs
- Complementary to distillation approaches