Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Advancing Generative Modelling of Calorimeter Showers

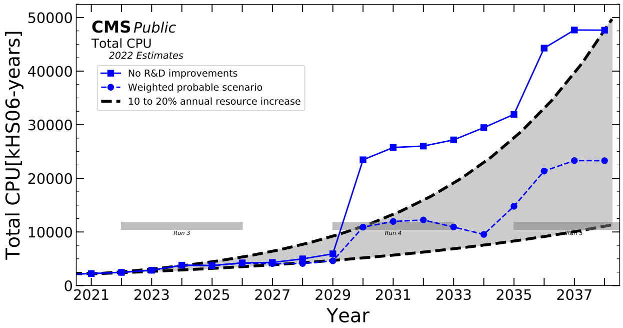
Erik Buhmann, Thorsten Buss, Sascha Diefenbacher, Engin Eren, Frank Gaede, Gregor Kasieczka, William Korcari, Anatolii Korol, Katja Krüge, Peter McKeown, Martina Mozzanica, Lorenzo Valente

April 30, 2024
European AI for Fundamental Physics Conference 2024

thorsten.buss@uni-hamburg.de

Detector Simulation

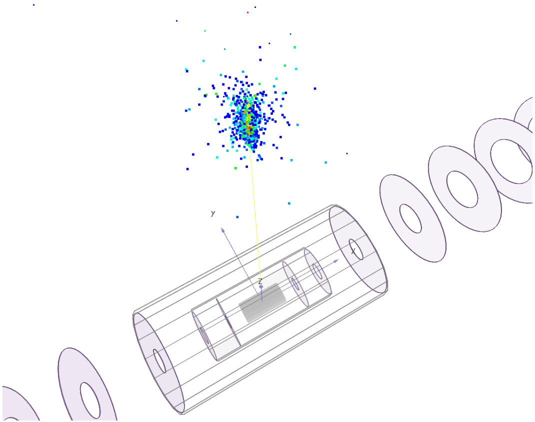
- Monte Carlo (MC) necessary to compare theory and measurements
- computational requirements expected to exceed available resources soon
- detector simulation most expensive part of simulation chain



[1] CMS Offline Software and Computing. *CMS Phase-2 Computing Model: Update Document*. 2022. URL: https://cds.cern.ch/record/2815292

International Large Detector (ILD)

- proposed detector for the International Linear Collider ILC
- has two sampling calorimeters
- electromagnetic calorimeter (ECAL)
  - 30 layers, 5mm x 5mm cells
- hadronic calorimeter (HCAL)
  - 48 layers, 30mm x 30mm cells
- dataset:
  - photon showers in ECAL
  - uniform distribution of incident energies
    - between 10 and 90 GeV



[2] Erik Buhmann et al. *Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed*. 2021. arXiv: 2005.05334

[3] ILD Concept Group. *International Large Detector: Interim Design Report*. 2020. arXiv: 2003.01116

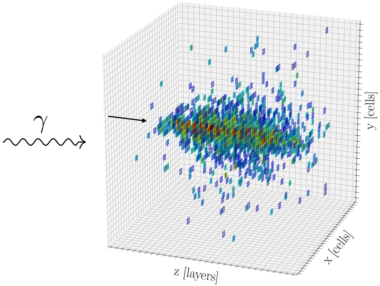
Data Representation of Showers

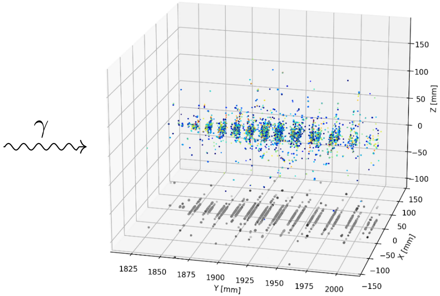
Fixed Grid

- 3D array filled with energy values
- entries correspond to calorimeter cells
- allows for convolutional networks
- needs bounding box

Point Clouds

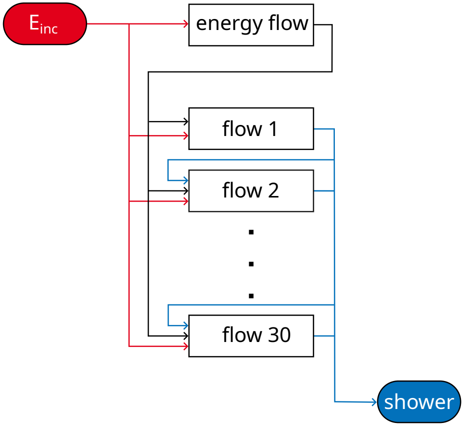
- variable-length, permutation-invariant sets
- only c.a. 4% of cells are non-zero
- more economically represented
- only generation of non-zero points





Convolutional L2LFlows

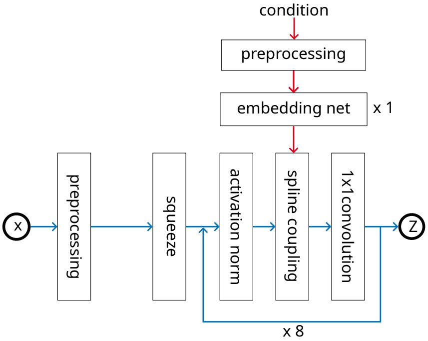
- based on CaloFlow[4] and L2LFlows[5]
- one energy distribution flow
  - learns distribution of layer energies
  - conditioned on incident energy
- 30 causal flows
  - learn shower shape in layer
  - conditioned on
    - incident energy
    - layer energy
    - previous layers
- generation
  - sample layer energies using energy distribution flow
  - sample shower shape using causal flows



[4] Claudius Krause and David Shih. *CaloFlow: Fast and Accurate Generation of Calorimeter Showers with Normalizing Flows*. 2021. arXiv: 2106.05285

[5] Sascha Diefenbacher et al. *L2LFlows: Generating High-Fidelity 3D Calorimeter Images*. 2023. arXiv: 2302.11594

Flow Architecture



- energy distribution flow
  - masked autoregressive flow[6]
- causal flows
  - spline coupling flow[7]
    - allows for efficient sampling
  - convolutional U-Nets[8] as sub networks
    - better scaling properties
  - architecture similar to Glow[9]

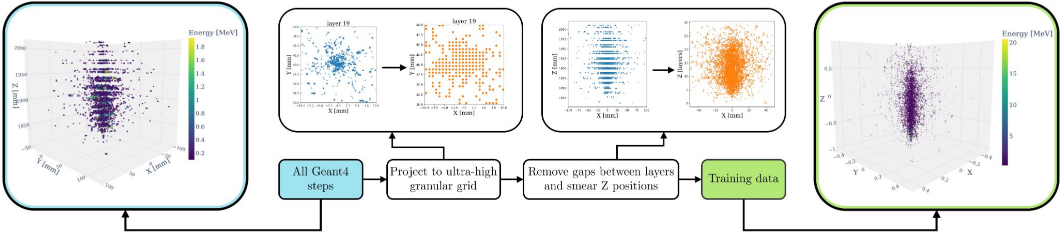
[6] Mathieu Germain et al. *MADE: Masked Autoencoder for Distribution Estimation*. 2015. arXiv: 1502.03509

[7] Conor Durkan et al. *Neural Spline Flows*. 2019. arXiv: 1906.04032

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597

[9] Diederik P. Kingma and Prafulla Dhariwal. *Glow: Generative Flow with Invertible 1x1 Convolutions*. 2018. arXiv: 1807.03039

Point Cloud Representation Pre-Processing

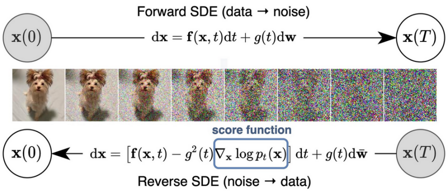


- point clouds of clustered Geant4 steps
- 36x higher resolution than detector cells
- 7x fewer points than full Geant4 steps

	points per shower
all Geant4 steps	40 000
clustered Geant4 steps	6 000
hits in calorimeter grid	1 500

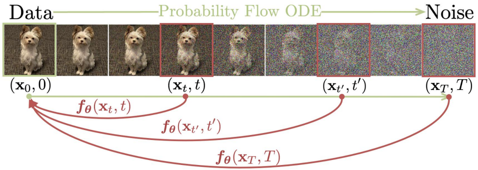
Diffusion Models

- score-based model[11]
  - continuous time diffusion process
  - stochastic differential equation (SDE)
  - sample by solving reverse SDE
- probability flow ODE
  - remove stochasticity
  - SDE $\rightarrow$ ODE
- consistency model distillation[12]
  - allows for single step sampling



$$\mathcal{L} = \| s_\theta(x_t, t) - \nabla_x \log p_t(x_t) \|_2^2$$

$$dx = [f(x, t) - \frac{1}{2} g(x, t)^2 \nabla_x \log p_t(x)]dt$$



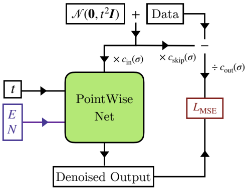
[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239

[11] Yang Song et al. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456

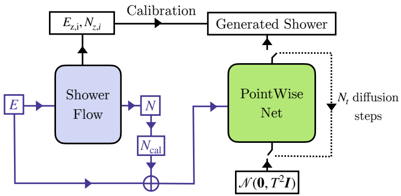
[12] Yang Song et al. *Consistency Models*. 2023. arXiv: 2303.01469

Calo Clouds II

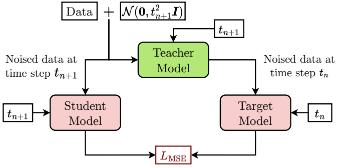
training



sampling



distillation

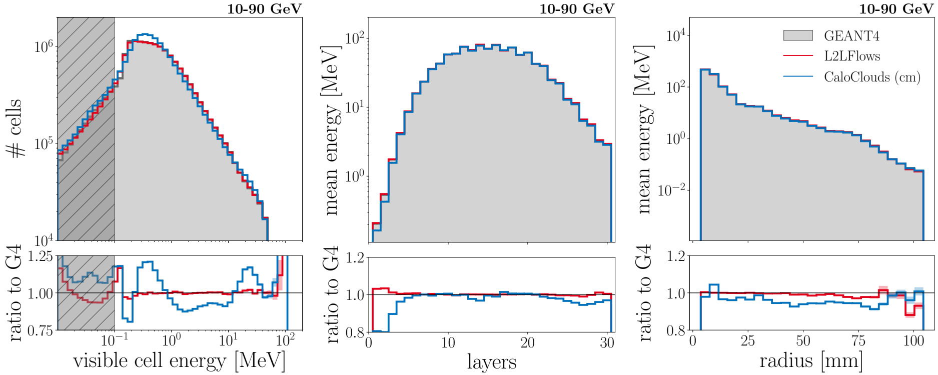


- score-based model
  - continuous time diffusion process
  - probability flow ODE
- distillation into a consistency model (cm)
  - allows for single step sampling

[13] Erik Buhmann et al. *CaloClouds: fast geometry-independent highly-granular calorimeter simulation*. 2023. arXiv: 2305.04847

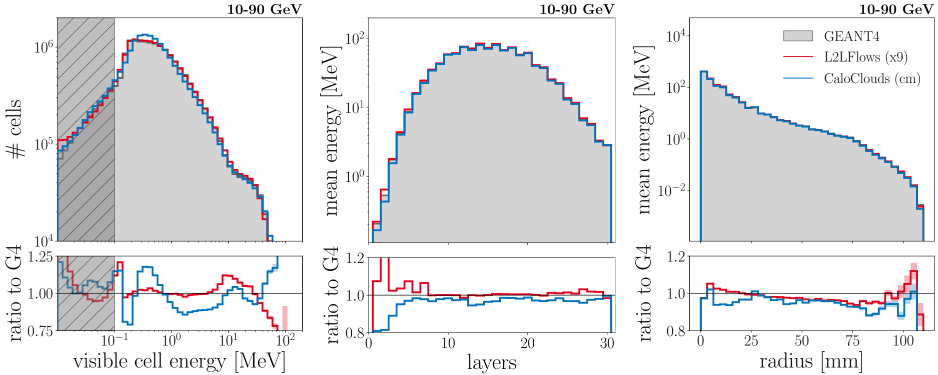
[14] Erik Buhmann et al. *CaloClouds II: Ultra-Fast Geometry-Independent Highly-Granular Calorimeter Simulation*. 2023. arXiv: 2309.05704

Results with Box Cut



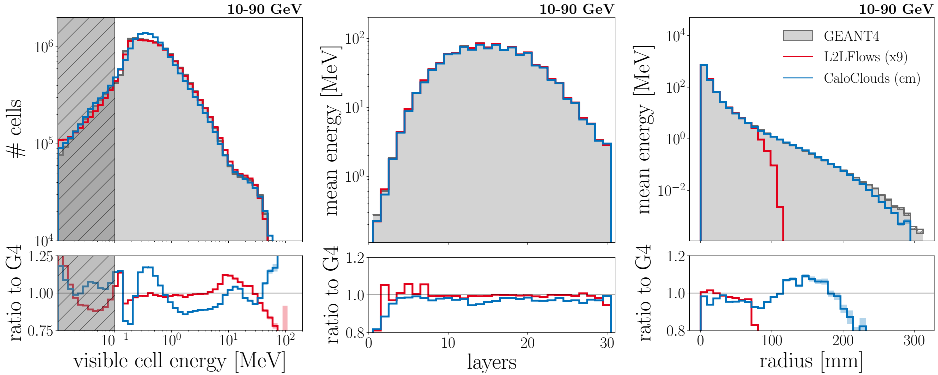
- evaluation at same incident point
- evaluated with 30x30 cell box cut
- good agreement with data

Shifting the Showers



- shift the showers in the calorimeter
- still apply 30x30 box cut
- need to train L2LFlows with nine times higher granularity

No Box Cut



- shift the showers in the calorimeter
- no box cut applied

Speedup over GEANT4

- comparison of generation times
- hardware: Intel® Xeon® E5-2640
- #threads: 1
- on GPU speed up of several thousands

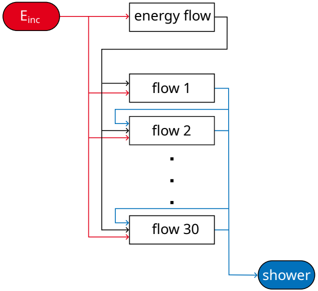
Simulator	Batch size	time [ms]	speed up
GEANT4	1	3915	x1.0
CaloClouds II		652	x6.0
CaloClouds (cm)		84	x46.6
L2LFlows		1203	x3.3
L2LFlows (x9)		4210	x0.9
L2LFlows	100	371	x10.6
L2LFlows (x9)		2775	x1.4

timing on single CPU thread

Summary

L2LFlows

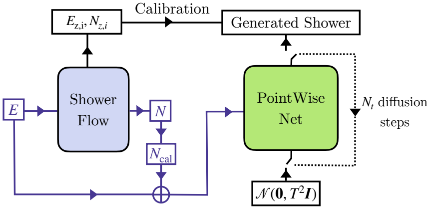
- fixed grid representation
- convolutional networks
- very good agreement within box



[paper coming soon]

CaloClouds II

- point cloud representation
- geometry independent
- no bounding box necessary
- fast sampling



[arXiv: 2309.05704]