Fast Generative Simulation for HEP

Thorsten Buss[1,2], **Henry Day-Hall**[1], Frank Gaede[1], Gregor Kasieczka[2], Katja Krüger[1], Anatolii Korol[1], Thomas Madlener[1], Peter McKeown[3], Martina Mozzanica[2], Lorenzo Valente[2]

[1] DESY, [2] University of Hamburg [3] CERN

Sustainable HEP 2026, 08.07.2026

HELMHOLTZ

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Overview

- Jevon’s paradox

Overview

- Jevon’s paradox
- What and Why
  - Need for Fast Simulation
  - Simulation Stages

Overview

- Jevon’s paradox
- What and Why
  - Need for Fast Simulation
  - Simulation Stages
- How
  - CaloClouds 3
  - CaloClouds 3 performance

Overview

- Jevon’s paradox
- What and Why
  - Need for Fast Simulation
  - Simulation Stages
- How
  - CaloClouds 3
  - CaloClouds 3 performance
- Benchmarks
  - Benchmark quantities
  - Benchmark results
  - Jevon’s paradox revisited

Overview

- Jevon's paradox
- What and Why
  - Need for Fast Simulation
  - Simulation Stages
- How
  - CaloClouds 3
  - CaloClouds 3 performance
- Benchmarks
  - Benchmark quantities
  - Benchmark results
  - Jevon's paradox revisited
- Future Directions

Jevon’s Paradox

Efficiency Increases Consumption

In 1865 William Stanley Jevons observed that as coal use became more efficient, total coal consumption increased rather than decreased.

- Efficiency lowers the cost of use.
- Lower cost drives greater demand.
- Net resource consumption grows.

As simulation becomes faster and cheaper, will we simply want to simulate more?

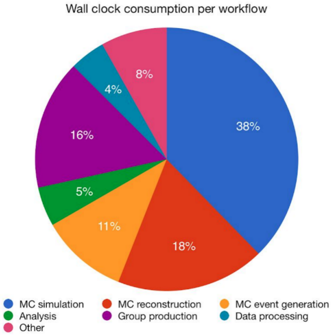
William Stanley Jevons

WHAT AND WHY

Need for Fast Simulation
Analysis Workflow

Analysis in HEP; compare recorded collisions ("events") to $\times 10$ simulated events.

- In ATLAS triggers select $\mathcal{O}(1000)$ events per second.
- Total $\mathcal{O}(10^{10})$ recorded events so far.
- Future colliders will record $\mathcal{O}(10^{12})$ events at the Z-pole alone...



In pure Monte Carlo

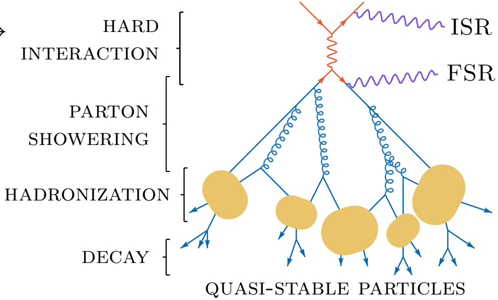
Single event for today's detectors $\approx 10^3$ CPU seconds, future detectors are significantly more complex.

Simulation Stages

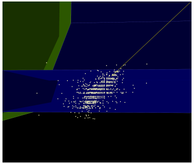
Why We Target Detector Simulation

Lightning fast $\approx 10^{-6}$s per event $\rightarrow$

Fast $\approx 10^{-3}$s per event



Slow $\approx 10^{2}$s to $\approx 10^{4}$s per event $\rightarrow$ GEANT4 Simulation



HOW

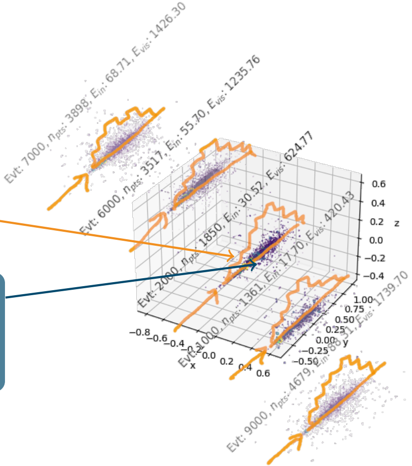
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

CaloClouds 3

Photons Only – Anatolii Korol and Henry Day-Hall

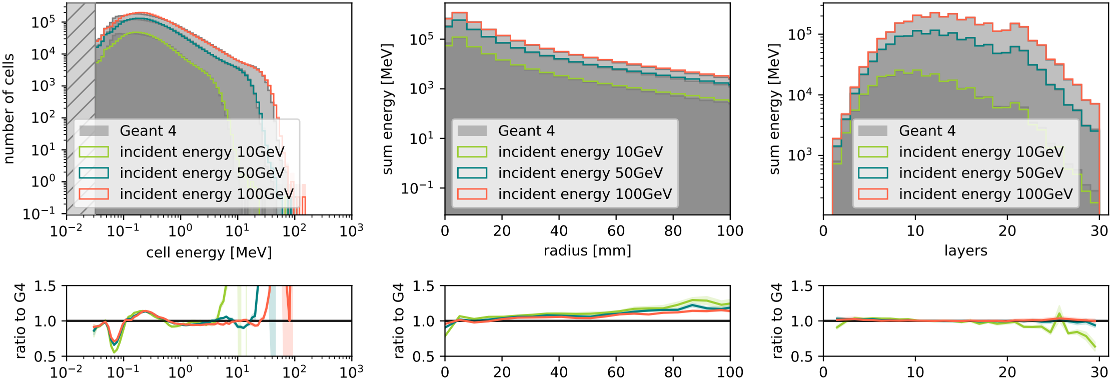
Normalising flow;
creates macro features
points & energy per detector layer

Diffusion model;
generates points
no communication between points



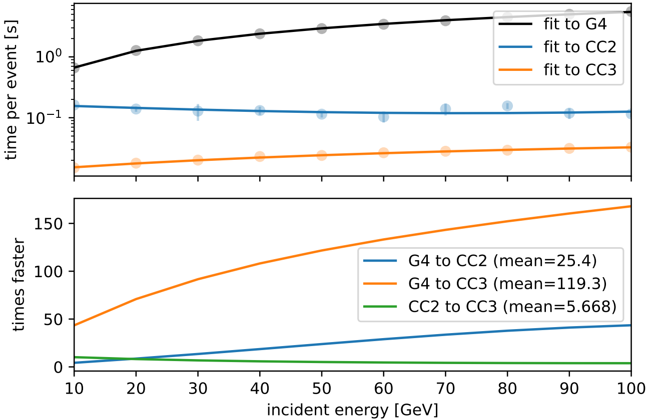
CaloClouds 3; Performance

Kinematics at fixed energies



Good results across a range of energies.
High energies are better, because iid assumption holds better.

CaloClouds 3; Performance
Inference Speed



On CPU $\approx \mathcal{O}(10^2)$ faster than G4. On GPU more like $\approx \mathcal{O}(10^3)$.

BENCHMARKS

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

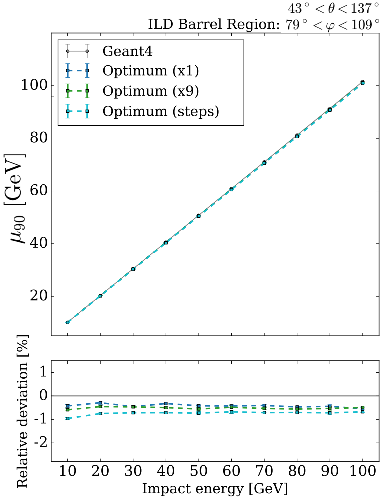
DESY.

FTX

UH

Reconstructions considered

Energy of a reconstructed photon

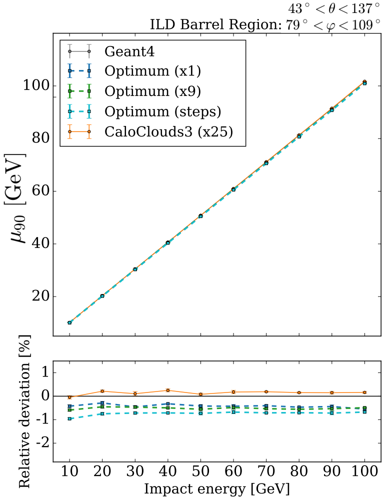


Using a G4 simulation reconstructed energy of the photon is very linear with the true energy of the photon.

Reconstructions considered

Energy of a reconstructed photon

Using a G4 simulation reconstructed energy of the photon is very linear with the true energy of the photon.

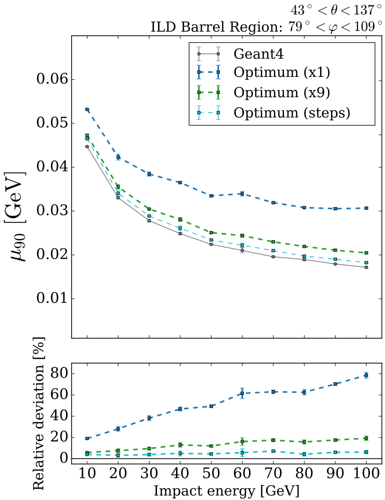


Reconstructions considered

Resolution of a reconstructed photon

- σ_{90} is the width of the distribution of reconstructed energies at 90% of the true energy.
- μ_{90} is the mean of the distribution of reconstructed energies at 90% of the true energy.

The resolution, $\frac{\sigma_{90}}{\mu_{90}}$, falls with increasing energy using a G4 simulation.

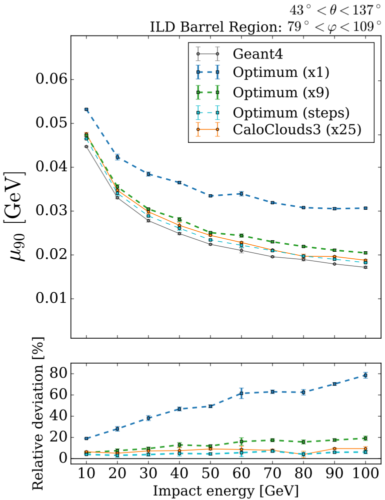


Reconstructions considered

Resolution of a reconstructed photon

- σ_{90} is the width of the distribution of reconstructed energies at 90% of the true energy.
- μ_{90} is the mean of the distribution of reconstructed energies at 90% of the true energy.

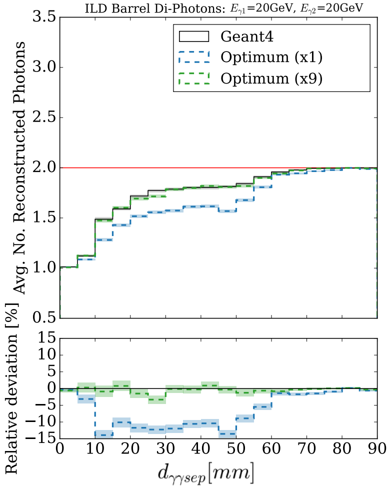
The resolution, $\frac{\sigma_{90}}{\mu_{90}}$, falls with increasing energy using a G4 simulation.



Reconstructions considered

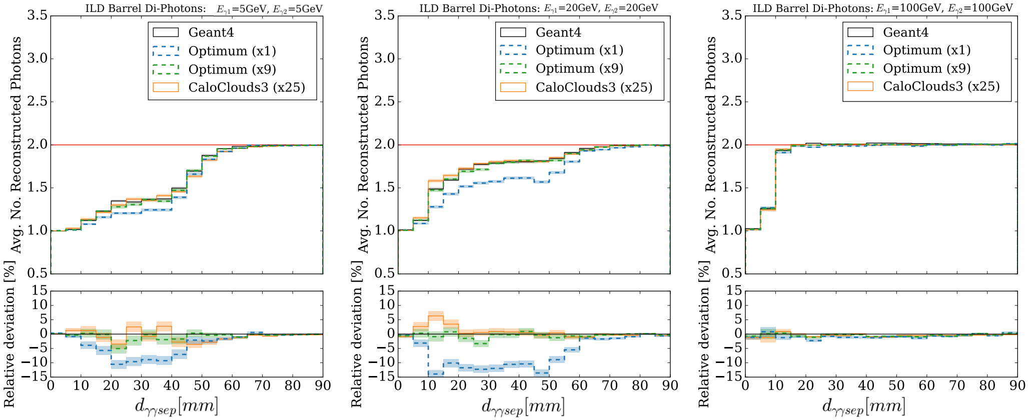
Multiplicity of reconstructed photons

The number of reconstructed photons varies with separation. We must reproduce the quite non linear behaviour of the G4 simulation.



Reconstructions

CaloClouds 3 multiplicities



Jevon’s Paradox revisited

For simulation in physics

We have a bounded objective. If we are smart about how we construct simulations this bounded objective saves us from Jevon’s paradox.

- Total volume of simulation is bounded by number of real events observed (and some factor for number of groups doing analysis).
- Quality needed is bounded by the fidelity of reconstructed quantities. If we benchmark properly, we will be only as accurate as needed.

William Stanley Jevons

Conclusions and the future

- ML can simulate fast and accurately – this is needed to do any physics at all.
- Good benchmarking ensures that we properly optimise the complexity of the models to meet needed accuracy with minimal compute.

Further work

- Further benchmarks for newer models.
- Mixture of specialist model, to leverage the best all models.
- Experiments in transfer learning.

Thank you!

Contact

Deutsches Elektronen-
Synchrotron DESY

ftx.desy.de

Henry Day-Hall
0000-0002-7881-2506
FTX-SFT
henry.day-hall@desy.de
https://github.com/FLC-QU-hep/CaloClouds-3

BACKUP

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

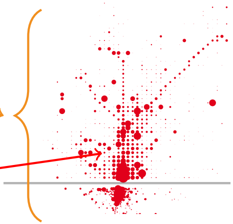
AllShowers

12 particle types – Thorsten Buss

Continuous normalising flow;
creates points per layer

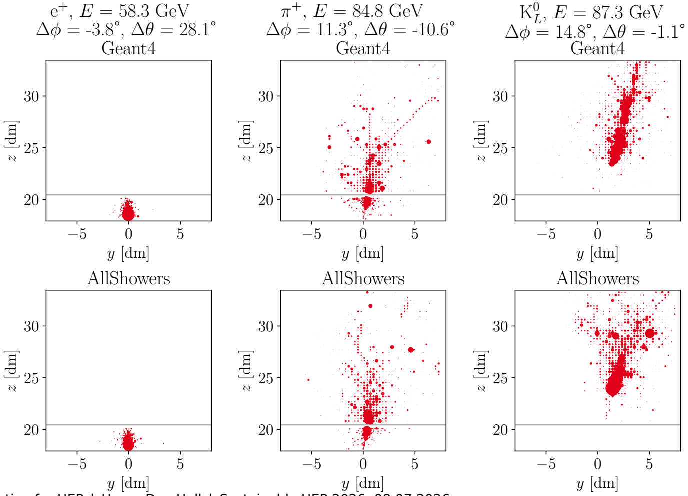
Transformer Encoder;
converts to/from latent space

Continuous normalising flow;
generates points
uses masked attention



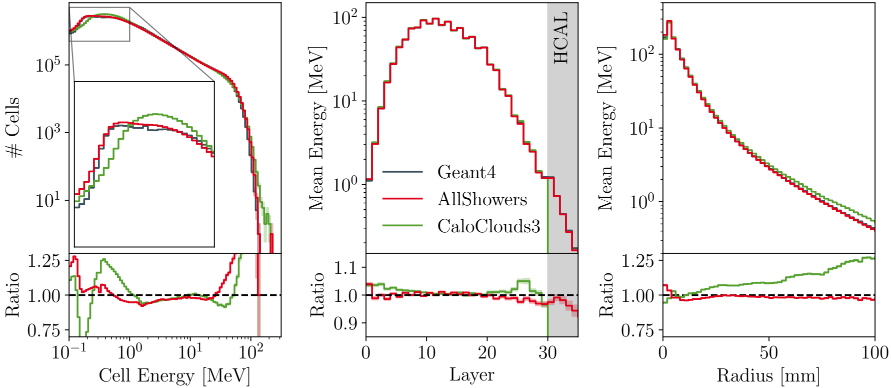
AllShowers; Performance

Sample Detail Replication



AllShowers; Performance

Comparison to CaloClouds 3



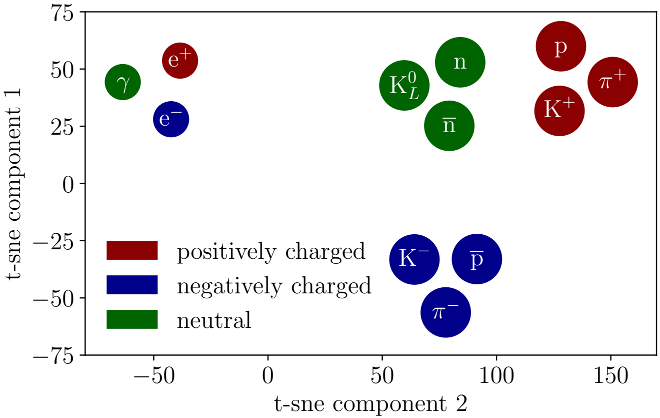
Significantly more accurate than previous models.

AllShowers v.s. CaloClouds 3

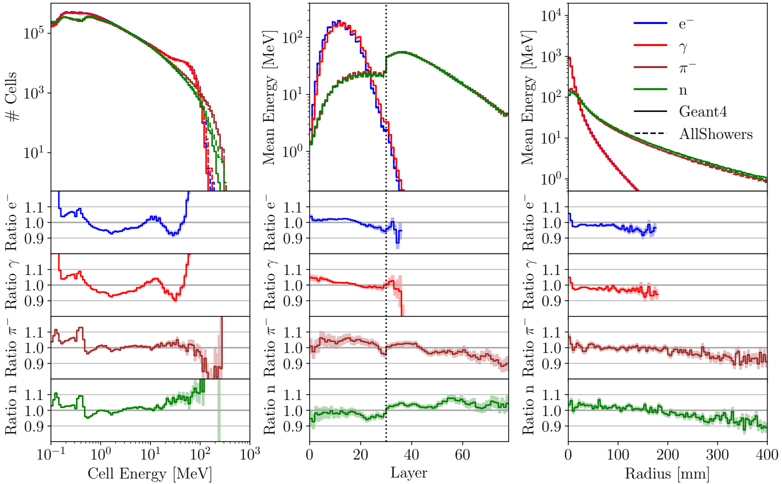
Timing

Hardware	Model	NFE	Batch Size	Time / Sample [s]	Speed-factor
CPU	Geant4	-	1	2.88	1.0x
	CaloClouds3	1	1	0.014	194.3x
			16	0.0041	654.5x
	AllShowers	1	1	0.17	16.7x
		32	1	5.0	0.6x
GPU	CaloClouds3	1	1	0.014	208.3x
			16	0.00088	3256. x
	AllShowers	1	1	0.018	164.0x
			16	0.0014	2114. x
		32	1	0.054	53.2x
			16	0.0066	440.0x

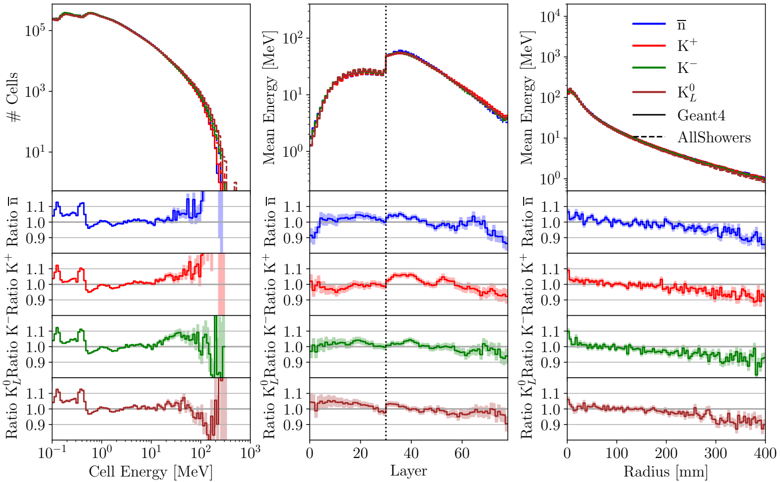
AllShowers; embeding



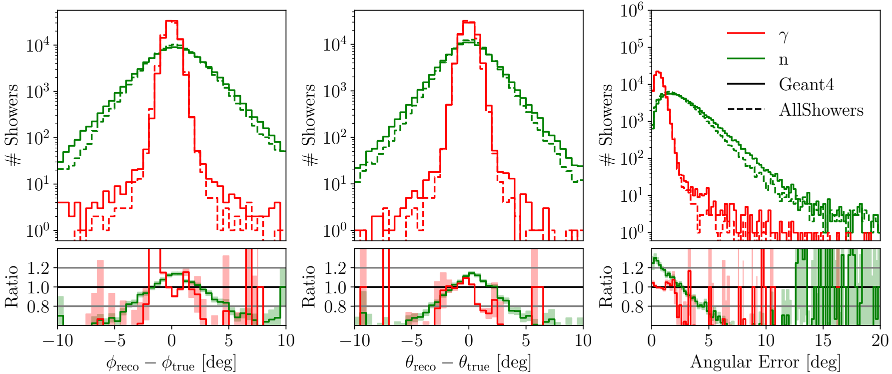
AllShowers; performance



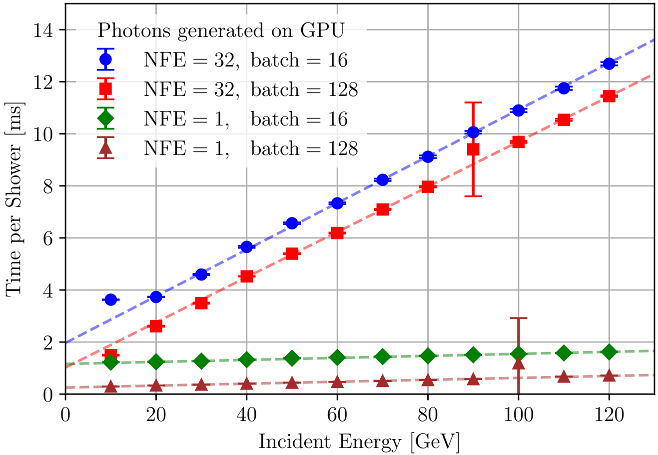
AllShowers; performance



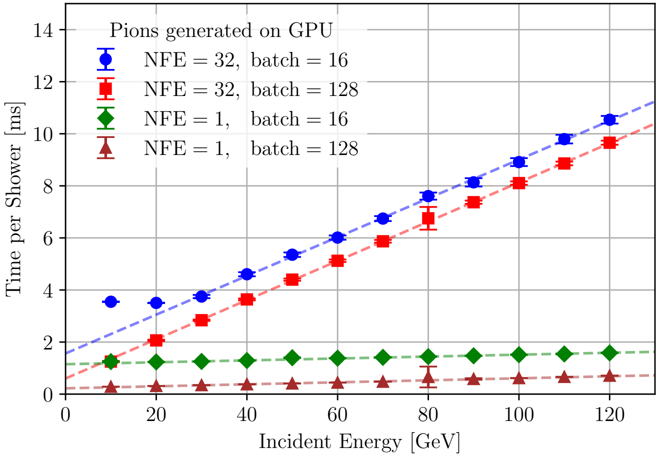
AllShowers; performance



AllShowers; timing

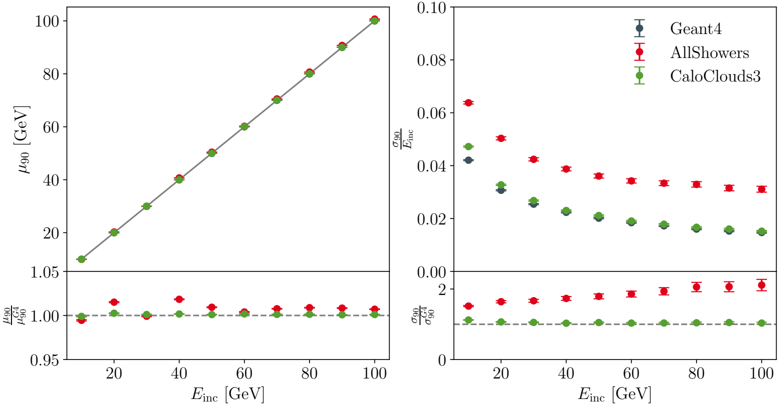


AllShowers; timing



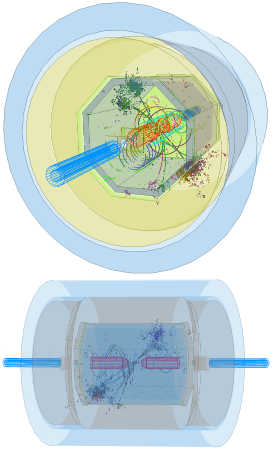
Reconstructions

AllShowers v.s. CaloClouds 3



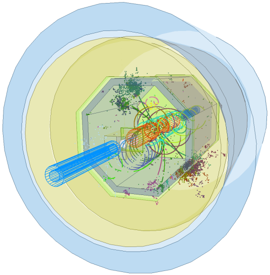
Chosen detector example

Geometry of the ILC EM calorimeter



- The ILC EM calorimeter is a high granularity calorimeter, we consider SiECAL variant.
- Calorimeter naturally has supports, creating blind spots.
- Layers of sensitive silicon detectors are sandwiched either side of passive Tungsten absorber, creating alternating patterns.
- Silicon cells are $5 \times 5 \times 0.5$ mm, Tungsten is 2.1 mm in first 20 layers, then 4.2 mm in last 10.
- Octagonal structure means layers start at varying offset, such that cells are staggered one layer to the next.
- Magnet field of 3.5 T.

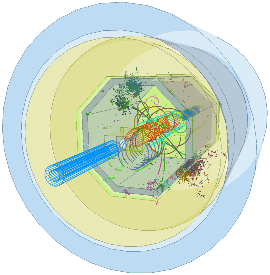
PreProcessing
Regularising the geometry



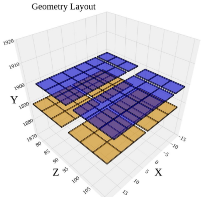
Take the ILD (SiECAL) as an example. It has a octagonal ECAL with 30 layers.
We want one model for the whole calorimeter, so we need to make some position agnostic training data.

PreProcessing

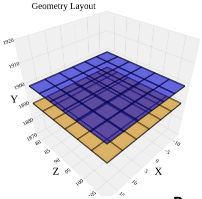
Regularising the geometry



Take the ILD (SiECAL) as an example. It has a octagonal ECAL with 30 layers.
We want one model for the whole calorimeter, so we need to make some position agnostic training data.



⇓ *Regularise* ⇓



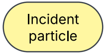
- Gaps due to supporting structures are removed.
- The modules are aligned, so each cell is centred above the one radially below.

End up with ≈ 1000 of (e_i, x_i, y_i, z_i) per event.

CaloClouds 3; Network architecture

Normalising flow and diffusion model

Incident particle is specified by (e, p_x, p_y, p_z)

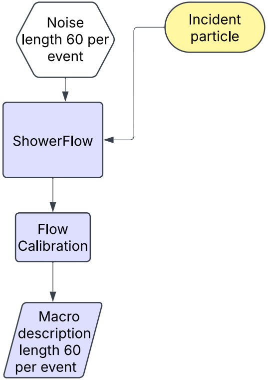


(x, y, z) handled later.

CaloClouds 3; Network architecture

Normalising flow and diffusion model

ShowerFlow is a normalising flow - accurate, but can't do high dimensions.



Predicts

$$N_{\text{layer 1}}, N_{\text{layer 2}} \cdots N_{\text{layer 30}}$$

where N is points in layer

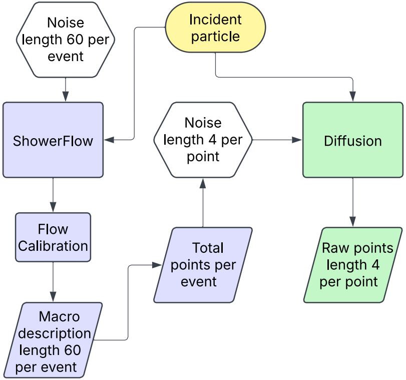
and

$$E_{\text{layer 1}}, E_{\text{layer 2}} \cdots E_{\text{layer 30}}$$

where E is total energy in layer

CaloClouds 3; Network architecture

Normalising flow and diffusion model

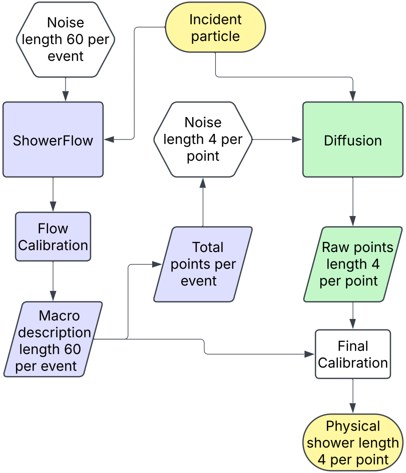


Diffusion model creates batch of iid points.

(e_1, x_1, y_1, z_1),
(e_2, x_2, y_2, z_2),
$\ldots$
$(e_{\sum N}, x_{\sum N}, y_{\sum N}, z_{\sum N})$

CaloClouds 3; Network architecture

Process of generation

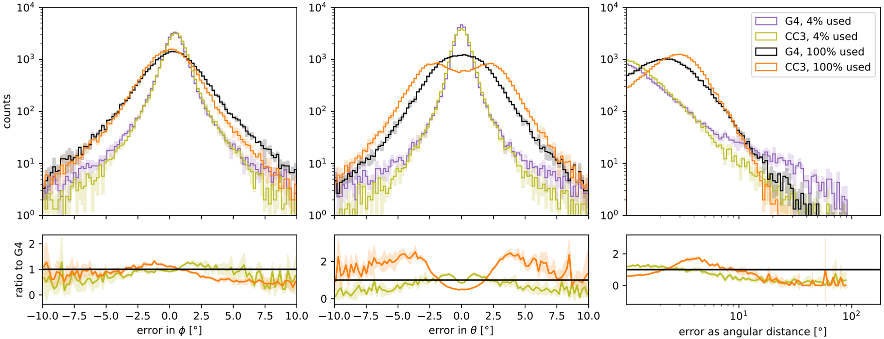


Split the points between layers, as dictated by ShowerFlow.

Enforce energy per layer predicted by ShowerFlow.

CaloClouds 3; Reconstructing internal angles

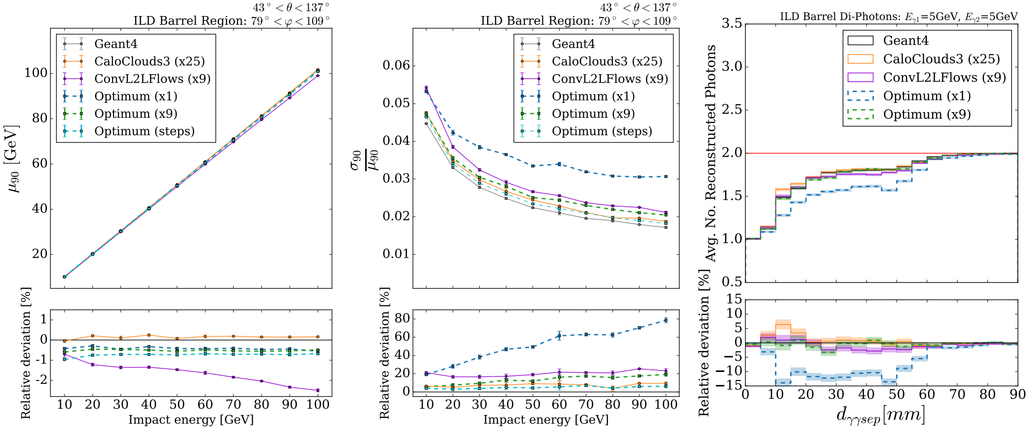
With all points, and highest energy 4%



Agreement is poor for all hits, but good when using the improved metric, with the 4% of highest energy hits.

CaloClouds 3; Reconstructed quantities

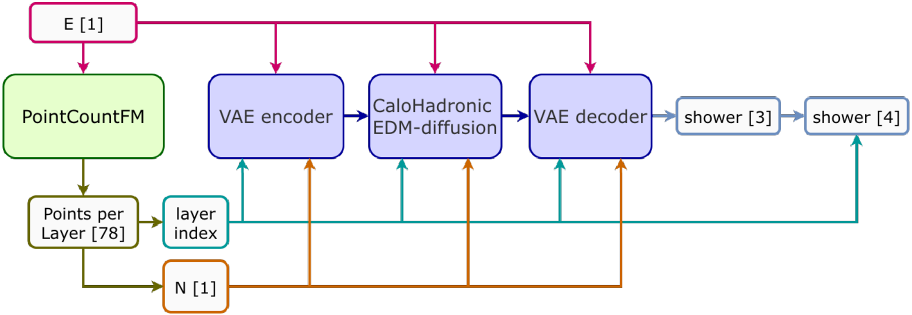
Ultimate performance evaluation



Comparing the behaviour of models after reconstruction.

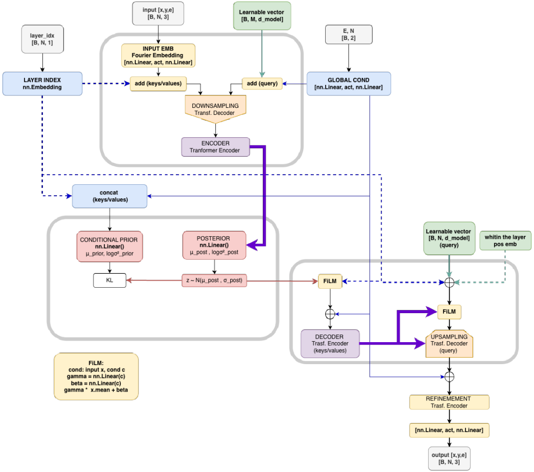
CaloHadronic; Network architecture

Process of generation

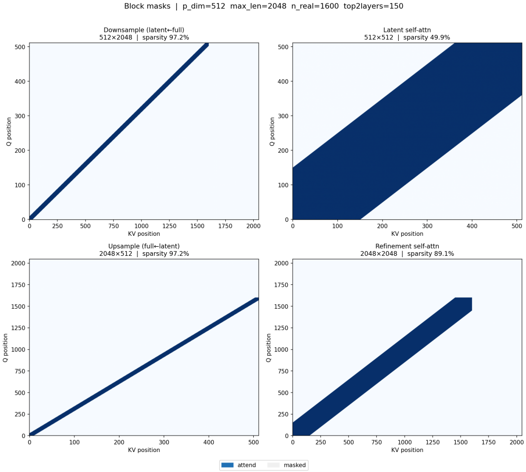


CaloHadronic; Network architecture

Process of generation

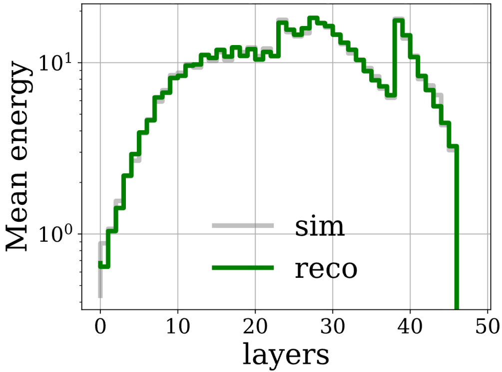


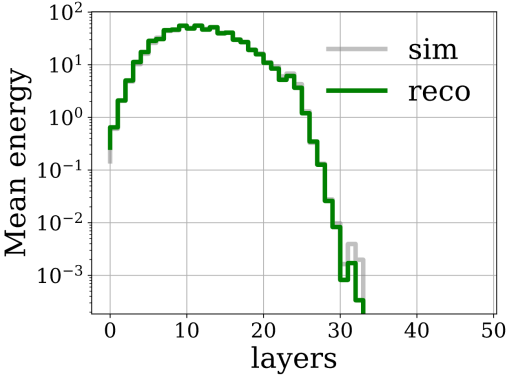
CaloHadronic; Network architecture
Masks



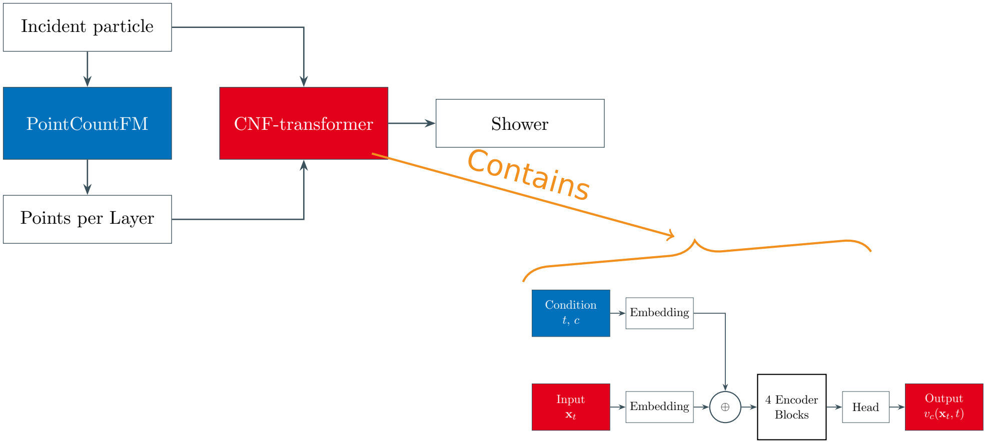
CaloHadronic; Performance

Energy per layer



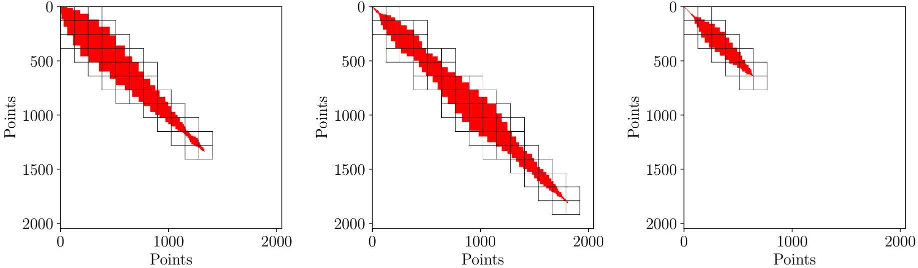


AllShowers; Network architecture
Process of generation



AllShowers; Network architecture

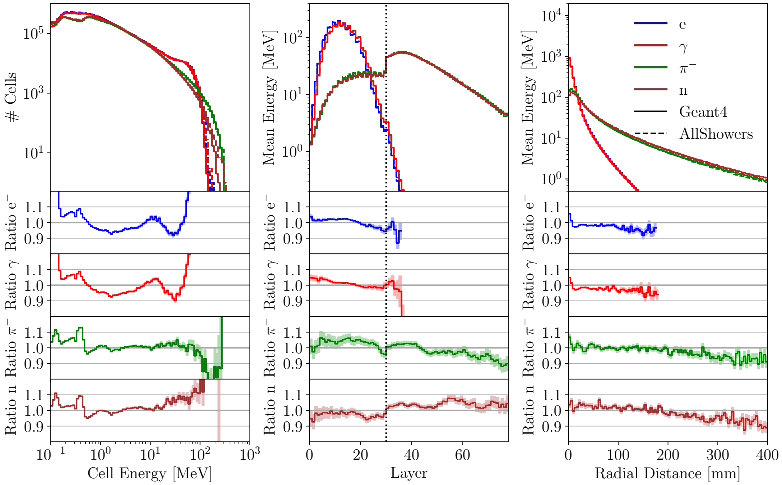
Masks



Examples of attention masks.
Particles up to ± 2 layers attend to each other.

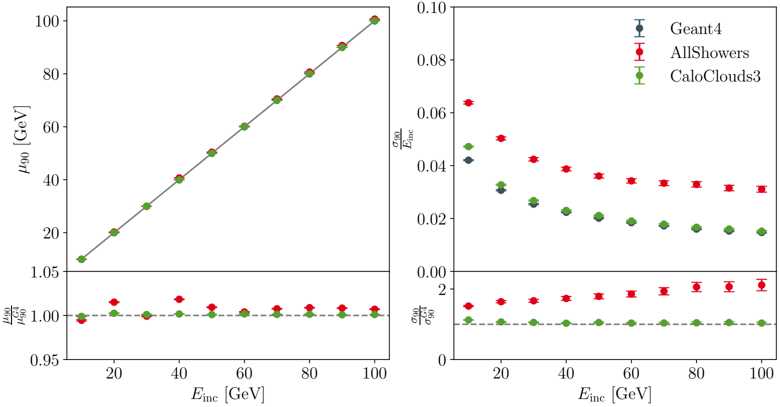
AllShowers; Performance

Multiple particle types



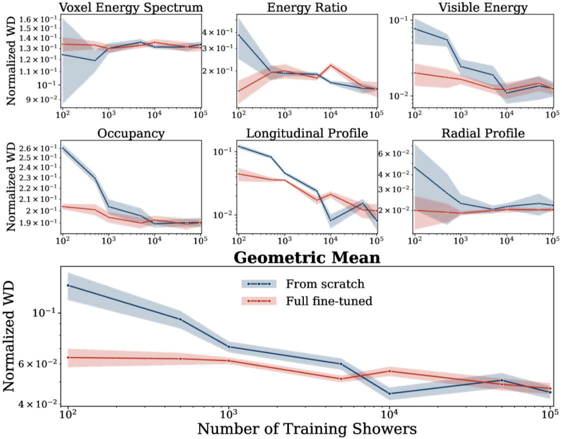
AllShowers; Performance

Photon Linearity and Resolution



Transfer Learning

Fine Tuning Performance Summary



Transfer Learning

Comparison of Parameter Efficient Fine Tuning Methods