ÖAW AUSTRIAN ACADEMY OF SCIENCES

MARIETTA BLAU INSTITUTE FOR PARTICLE PHYSICS

Generative Models and FastSimulation

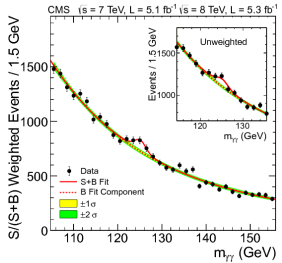
— Digital Twins for Nuclear and Particle Physics — NPTwins 2025 —

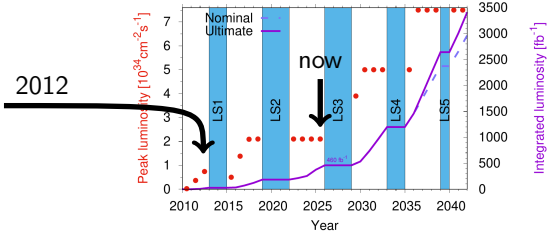
Claudius Krause

Marietta Blau Institute for Particle Physics (MBI Vienna)
formerly: Institute of High Energy Physics (HEPHY),
Austrian Academy of Sciences (OeAW)

October 7, 2025

We will have a lot more data in the near future.



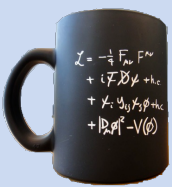


CMS Collaboration [arXiv:1207.7235, Phys.Lett.B]

https://lhc-commissioning.web.cern.ch/schedule/HL-LHC-plots.htm

- We will have 10× more data.

⇒ We want to understand every aspect of it based on 1st principles! (and find New Physics)



A (simplified) View on Particle Physics Analyses

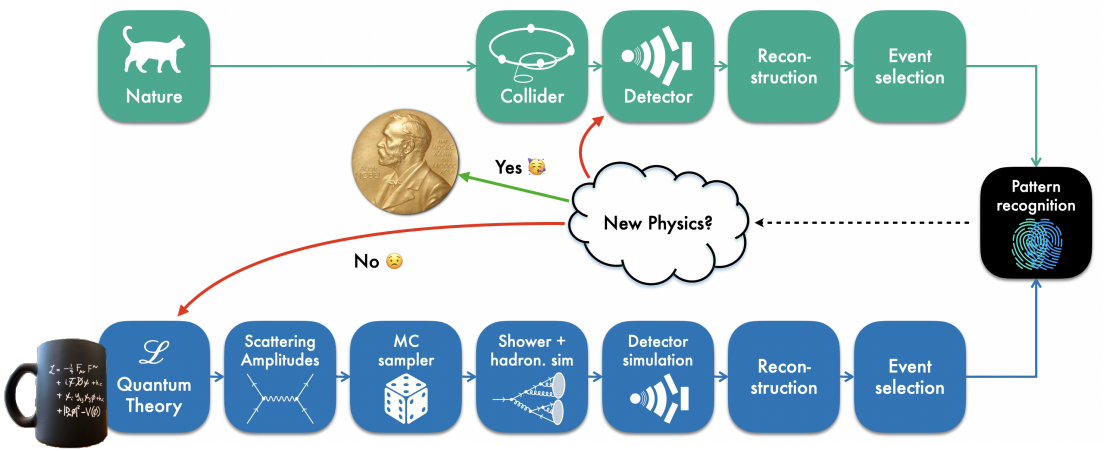
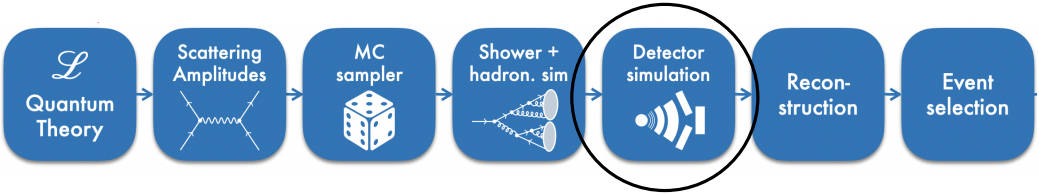
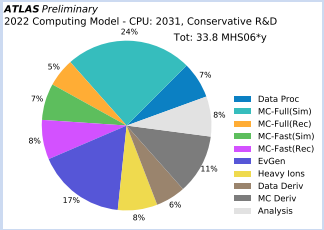


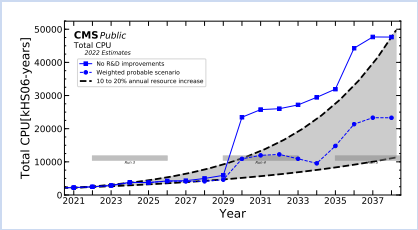
Figure by R. Winterhalder

Detector Simulation is the Bottleneck



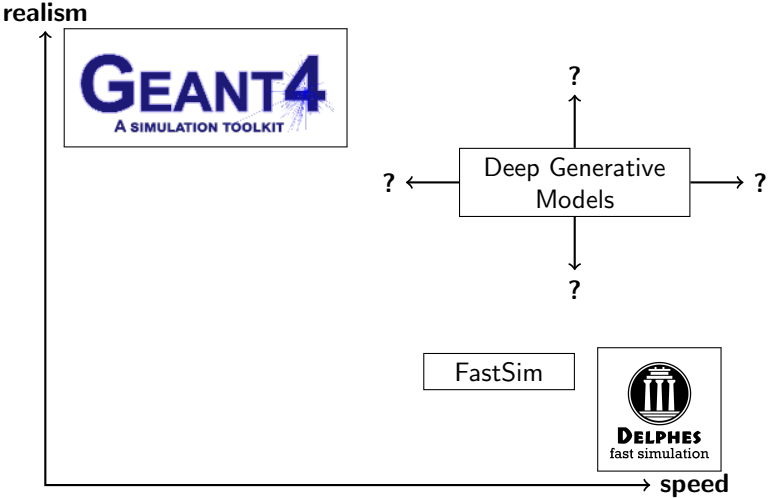


CERN-LHCC-2022-005



CMS-NOTE-2022-008

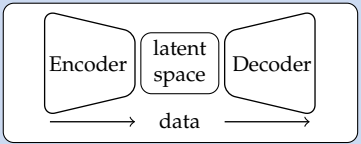
Generative AI can provide fast Surrogate Models.



Generative Models are very diverse.

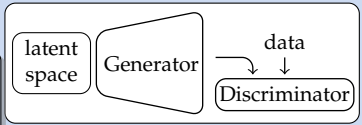
Variational Autoencoder (VAE)

⇒ Compressing data through a bottleneck.



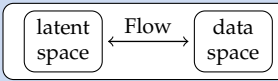
Generative Adversarial Network (GAN)

⇒ Generator and Discriminator play a game against each other.



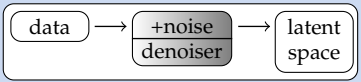
Conditional Flow Matching (CFM)

⇒ Continuous transformation into noise.



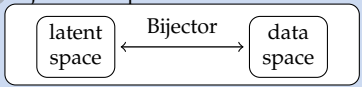
Diffusion Models

⇒ Gradually add noise and revert.



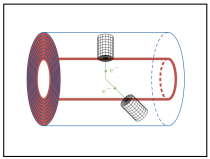
Normalizing Flows

Bijective map to a known distribution.

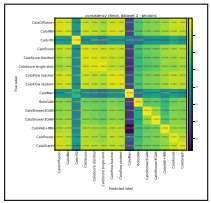


The Solution: Create the CaloChallenge 2022 as Benchmark!

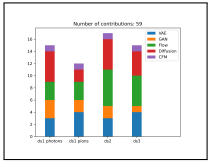
I: Datasets



II: Evaluation Metrics

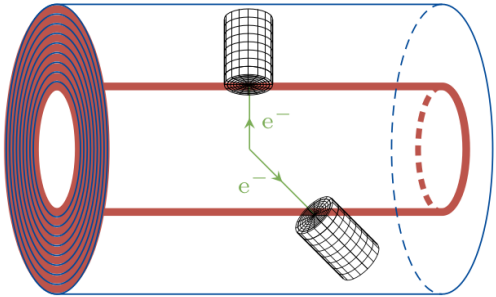


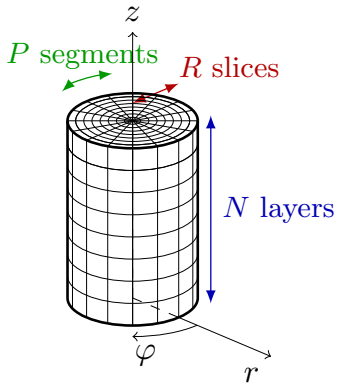
III: Results



CaloChallenge Showers are voxelized in cylindrical coordinates.

- There 4 datasets in increasing complexity / dimensionality.
- Particles enter perpendicular to front surface:

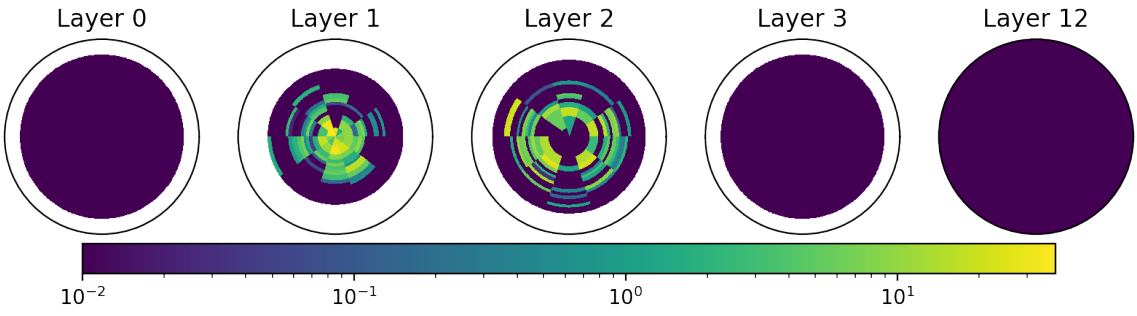




CaloChallenge Showers are voxelized in cylindrical coordinates.

- Showers are usually sparse.
- Energy depositions span several orders of magnitude.

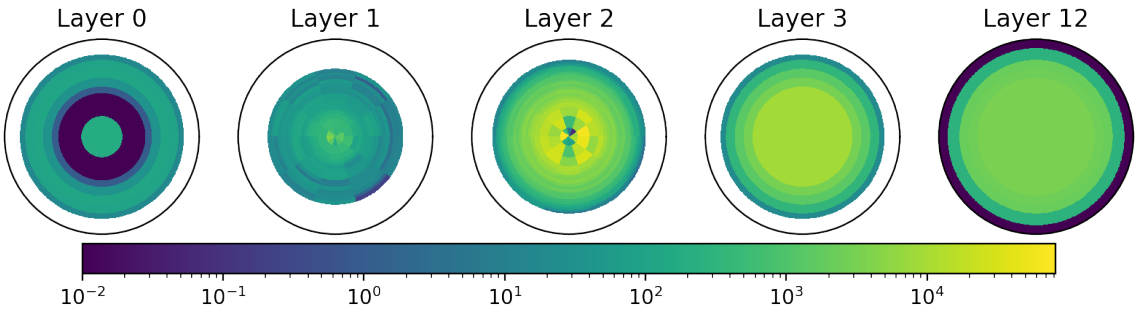
Photon shower at E = 1.0 GeV



CaloChallenge Showers are voxelized in cylindrical coordinates.

- Showers are usually sparse.
- Energy depositions span several orders of magnitude.

Photon shower at E = 1048.6 GeV



The Fast Calorimeter Simulation Challenge 2022

The main task: Develop a model that samples from $p(\text{shower}|E_{\text{incident}})$

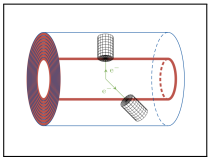
https://calochallenge.github.io/homepage/

Michele Faucci Giannelli, Gregor Kasieczka, **CK**, Ben Nachman, Dalila Salamani, David Shih, and Anna Zaborowska

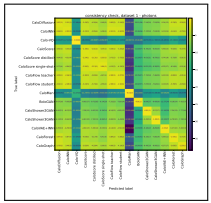
- Dataset 1: AtlFast3 trainig data (γ: 368, π: 533 voxels) $E_{\text{inc}} \in [256\ \text{MeV}, 4.2\ \text{TeV}]$
  [2109.02551, Comput.Softw.Big Sci.]
- Dataset 2: Par04 simulated detector (e^-: 6480 voxels) $E_{\text{inc}} \in [1\ \text{GeV}, 1\ \text{TeV}]$
- Dataset 3: Par04 simulated detector (e^-: 40500 voxels) $E_{\text{inc}} \in [1\ \text{GeV}, 1\ \text{TeV}]$

Generative Models and FastSimulation

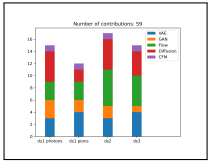
I: Datasets



II: Evaluation Metrics

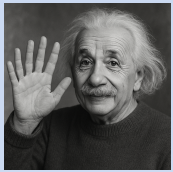


III: Results



How to evaluate generative models?

In text / image / video generation: "by eye".
⇒ Our brains are incredible good at this task, but it doesn't scale.



imagined with ChatGPT.

In high-energy physics: need to find something better!
⇒ We want to correctly cover the entire distribution of samples.

1. Can look at histograms of derived features / observables.

⇒ To quantify, we use the *separation power* of high-level feature histograms:

$$S(h_1, h_2) = \frac{1}{2} \sum_{i=1}^{n_{\text{bins}}} \frac{(h_{1,i} - h_{2,i})^2}{h_{1,i} + h_{2,i}}$$

But: this is just a 1-dim projection!

A Classifier provides the "ultimate metric".

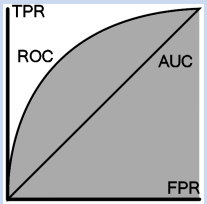
According to the Neyman-Pearson Lemma we have:

- The likelihood ratio is the most powerful test statistic to distinguish two samples.
- A powerful classifier trained to distinguish the samples should therefore learn (something monotonically related to) $w = \frac{p_{\text{data}}}{p_{\text{model}}}$.
- If this classifier is confused, we conclude $\Rightarrow \quad p_{\text{data}}(x) = p_{\text{model}}(x)$

$\Rightarrow$ This captures the full phase space incl. correlations. CK/D. Shih [2106.05285, PRD]

2. Now, the AUC provides a single number to compare different models.

   But: are AUCs of different models really comparable?



How to decide which model is closest to the reference: the Multiclass Classifier

A multi-class classifier:

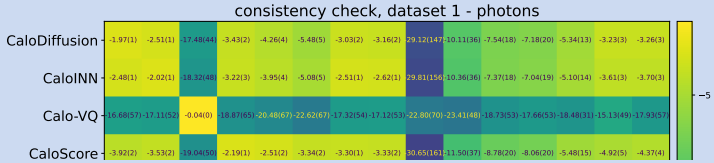
Train on submission 1 vs. submission 2 vs. ... vs. submission n and evaluate the *log posterior*:

$$L = \langle \log \left(p(x_{\in \text{class } i} | x_{\text{taken from } j}) \right) \rangle \qquad j \in \{\text{submission } k, \textsc{Geant4}\}$$

3. As metric: evaluate with GEANT4

Lim et al. [2211.11765, MNRAS]

As cross-check: validate with all submissions j



Other important metrics to look at.

⇒ The *generation time*.
- on CPU/GPU architectures
- for batch sizes 1 / 100 / 10000

⇒ The *number of trainable parameters*.
- as proxy for model size
- in training / generation

Other important metrics to look at.

⇒ The *generation time*.
- on CPU/GPU architectures
- for batch sizes 1 / 100 / 10000

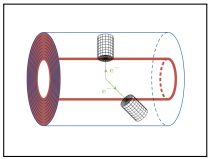
⇒ The *number of trainable parameters*.
- as proxy for model size
- in training / generation

- start singularity container
- load model weights + biases
- generate samples
- save them to `.hdf5`

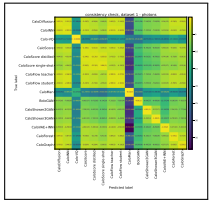
Powered by CLIP at the VBC.

Generative Models and FastSimulation

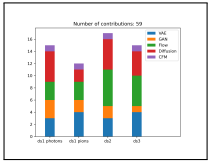
I: Datasets



II: Evaluation Metrics



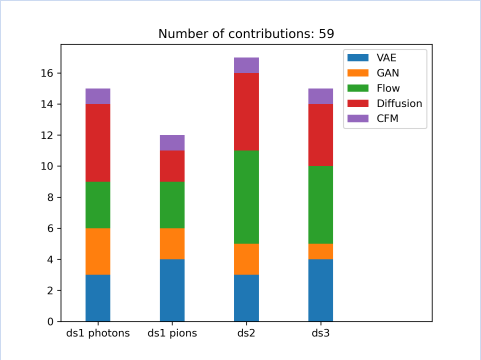
III: Results



We received many submissions.

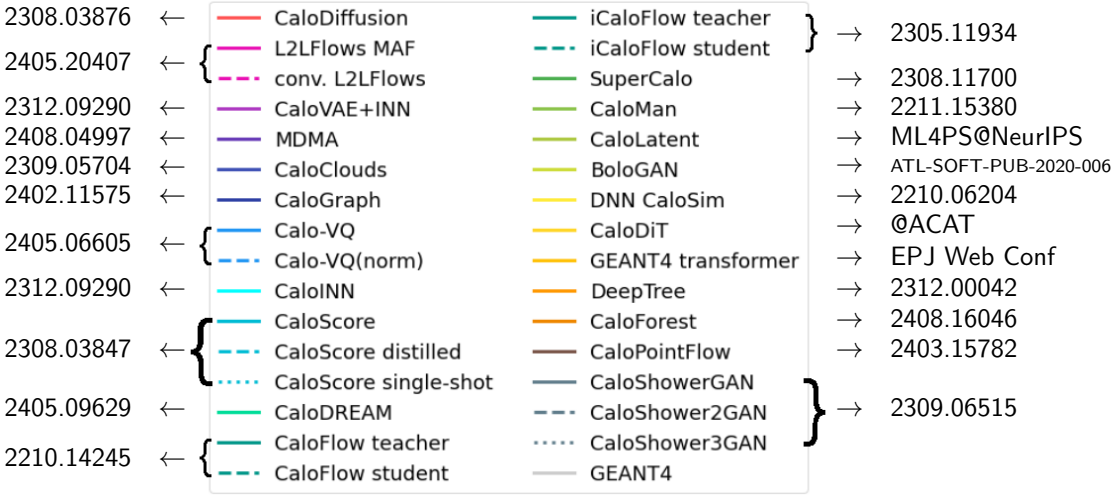
I will only be able to share some highlights of the results of the CaloChallenge.

The final write-up, **arXiv:2410.21611**, has a lot more content!

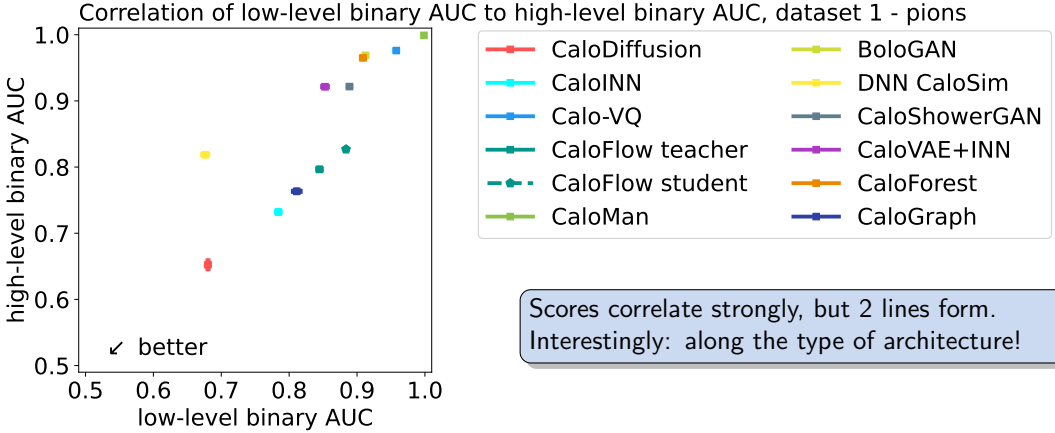


- We received 59 submissions for all datasets.
- They were generated by 23 different models.
- All types of generative AI architectures were used.

We received many submissions.

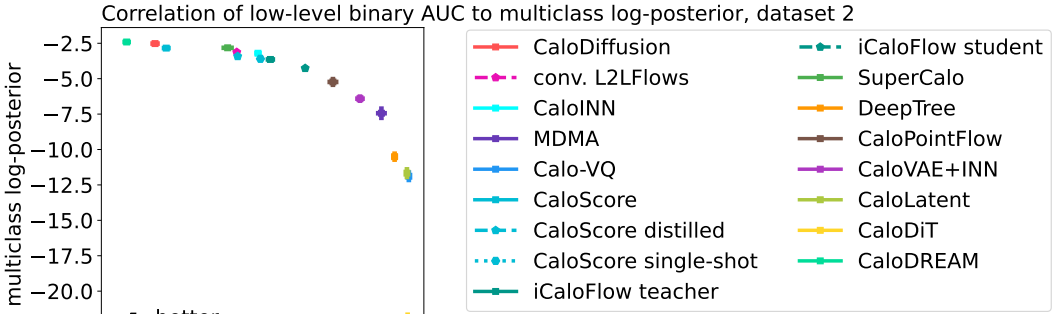


Comparing different quality metrics: classifier input



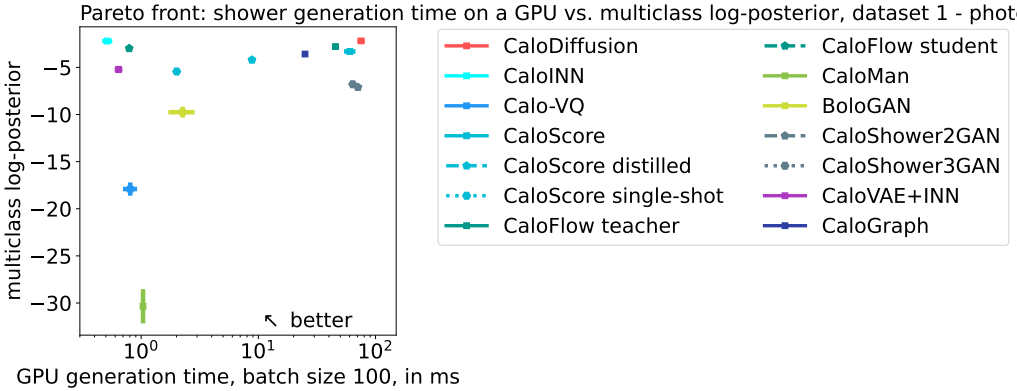
Scores correlate strongly, but 2 lines form.
Interestingly: along the type of architecture!

Comparing different quality metrics: binary vs. multiclass

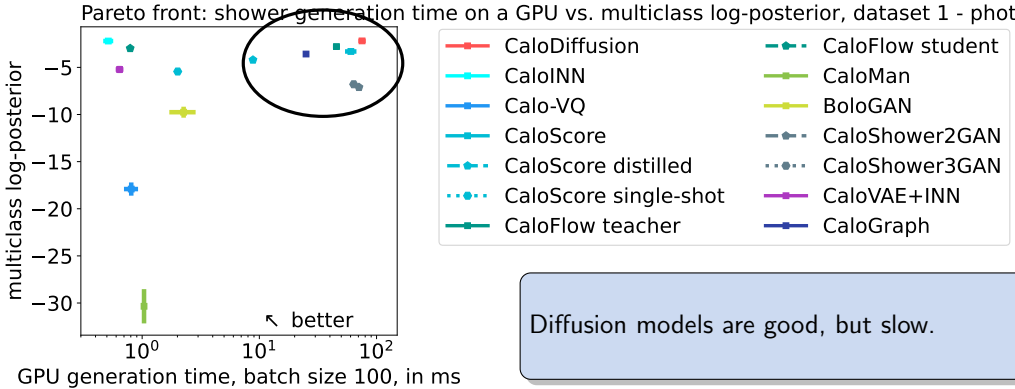


Very clear correlation, both seem to capture similar things!

Pareto Fronts: Quality vs. Generation Time

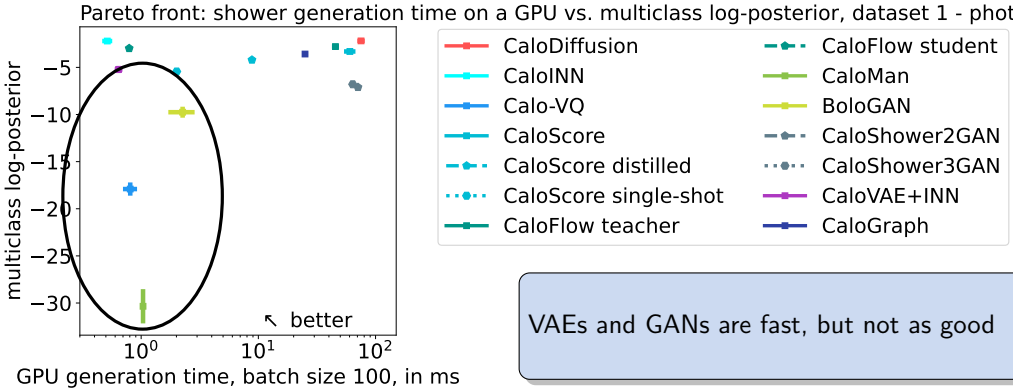


Pareto Fronts: Quality vs. Generation Time



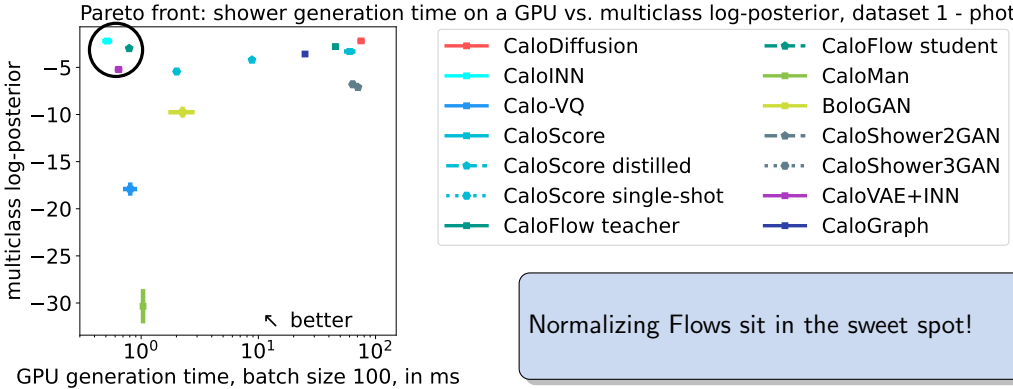
Diffusion models are good, but slow.

Pareto Fronts: Quality vs. Generation Time



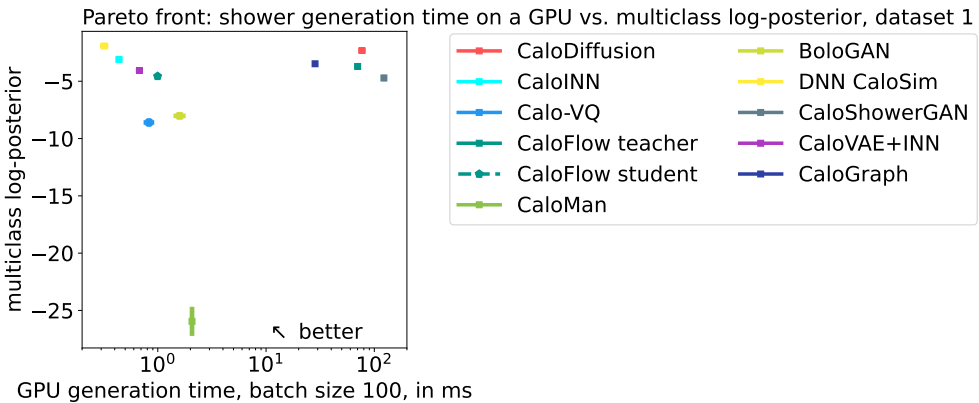
VAEs and GANs are fast, but not as good

Pareto Fronts: Quality vs. Generation Time

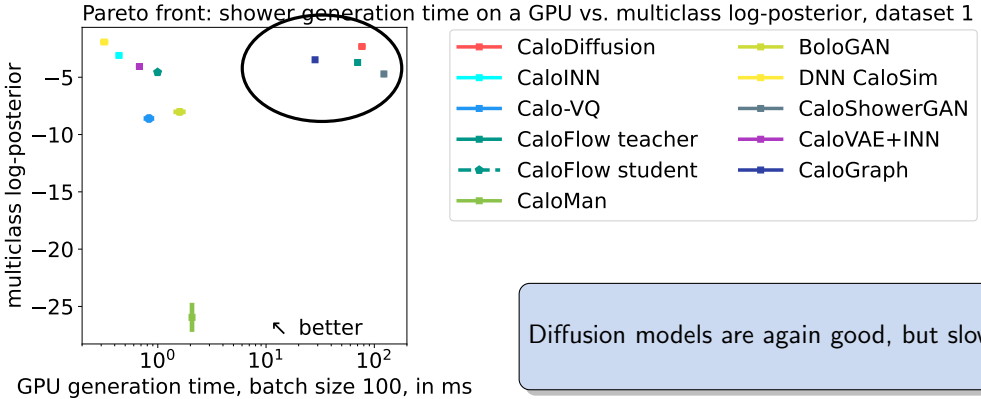


Normalizing Flows sit in the sweet spot!

Pareto Fronts: Quality vs. Generation Time

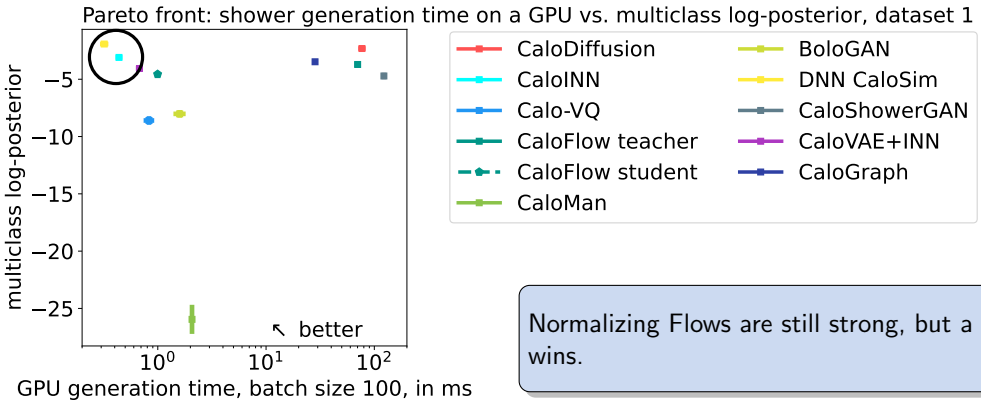


Pareto Fronts: Quality vs. Generation Time



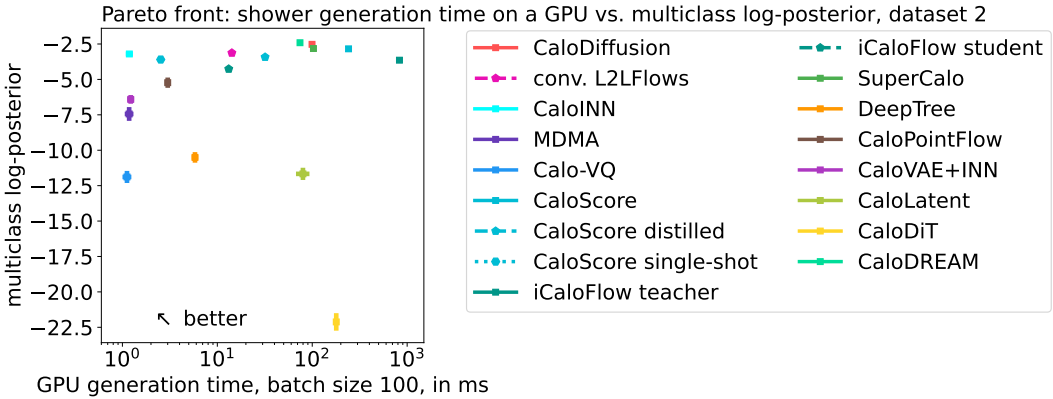
Diffusion models are again good, but slow.

Pareto Fronts: Quality vs. Generation Time

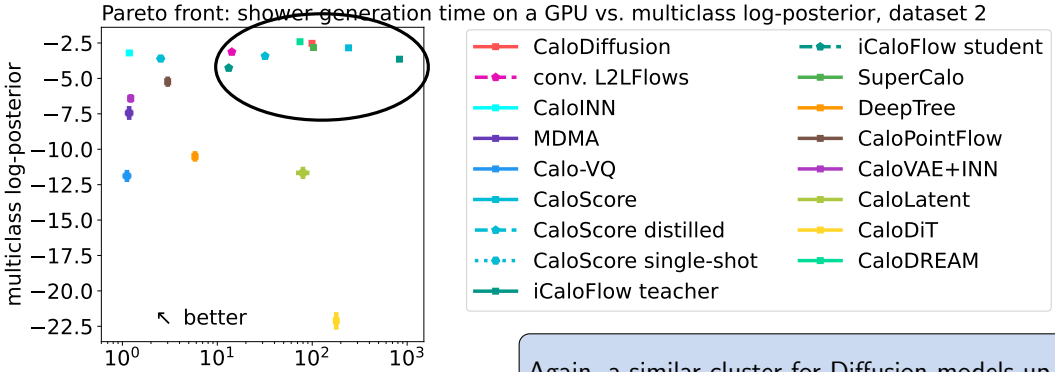


Normalizing Flows are still strong, but a VAE wins.

Pareto Fronts: Quality vs. Generation Time

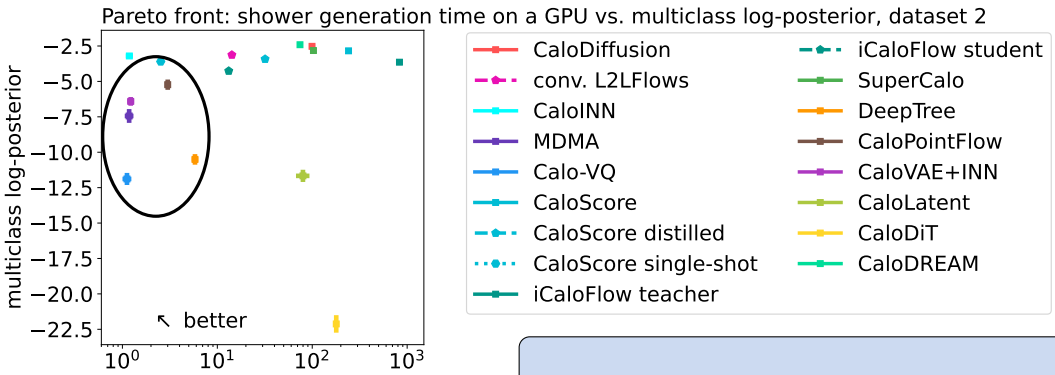


Pareto Fronts: Quality vs. Generation Time



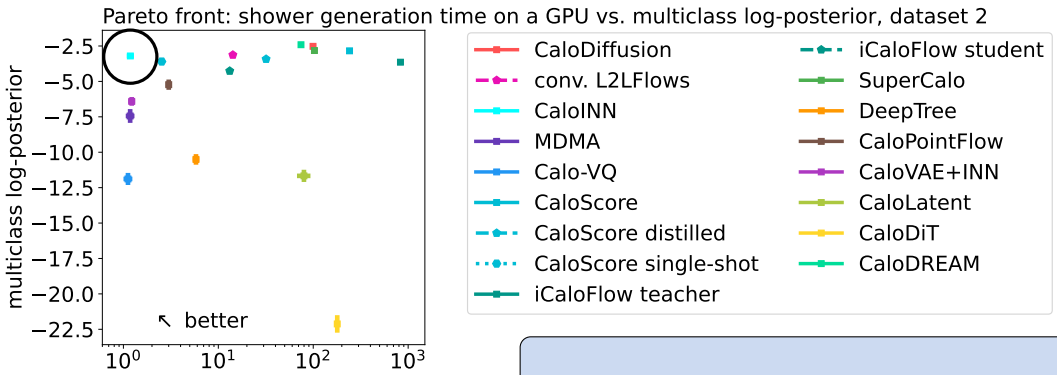
Again, a similar cluster for Diffusion models up here.

Pareto Fronts: Quality vs. Generation Time



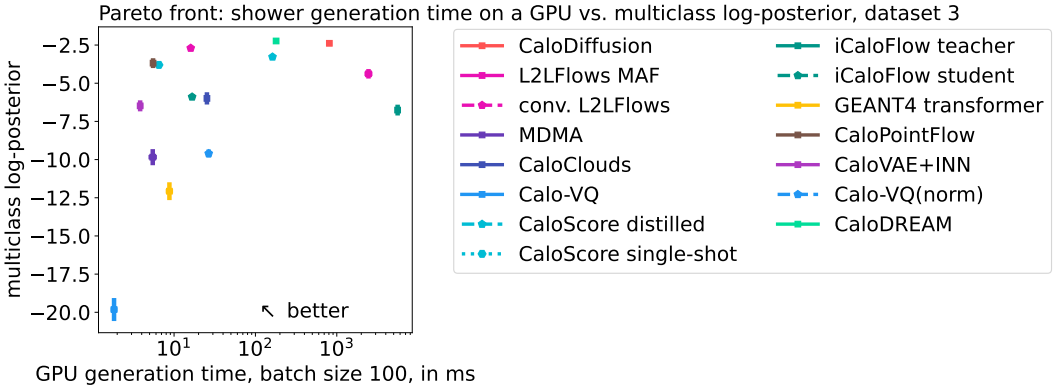
And a group of VAEs and GANs here

Pareto Fronts: Quality vs. Generation Time

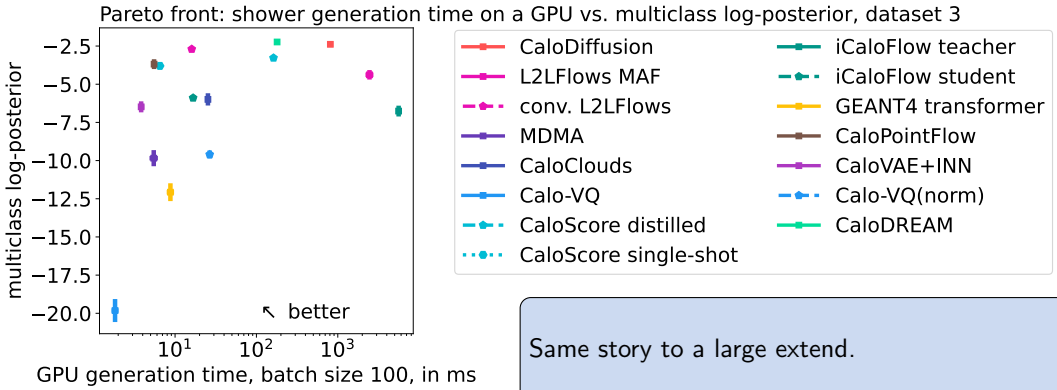


And a Normalizing Flows in the corner

Pareto Fronts: Quality vs. Generation Time

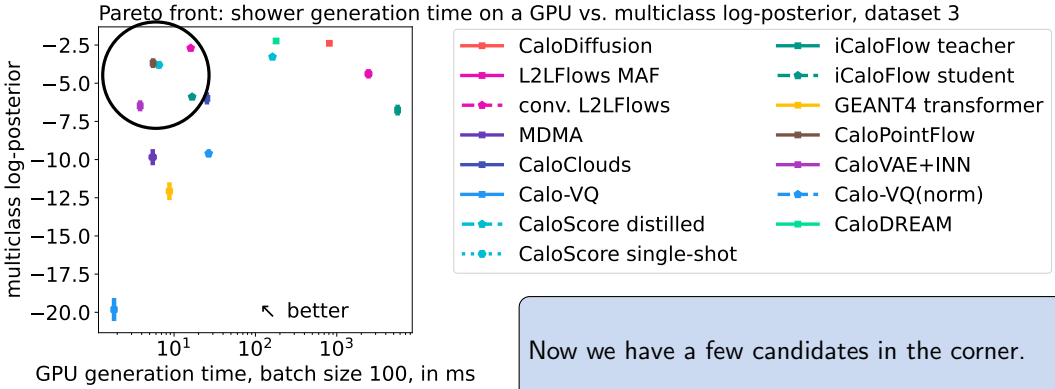


Pareto Fronts: Quality vs. Generation Time



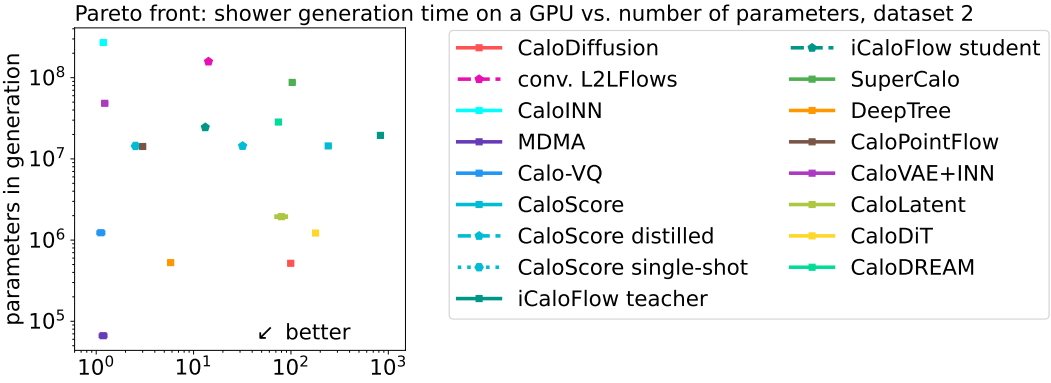
Same story to a large extend.

Pareto Fronts: Quality vs. Generation Time



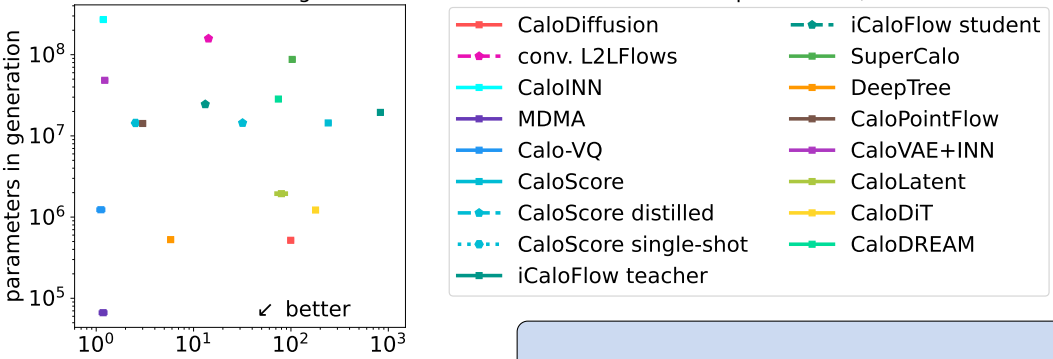
Now we have a few candidates in the corner.

What about a resource trade-off?



What about a resource trade-off?

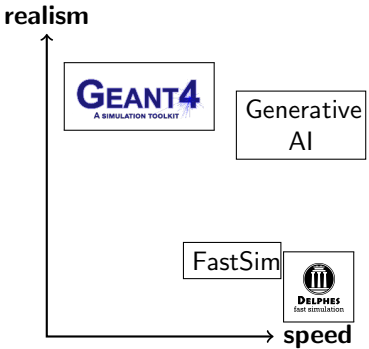
Pareto front: shower generation time on a GPU vs. number of parameters, dataset 2



Not really. It depends more on the architecture

Generative Models and FastSimulation

- Detector Simulation is a Bottleneck in the Simulation Chain.
- Generative AI offers to improve the quality of FastSim.

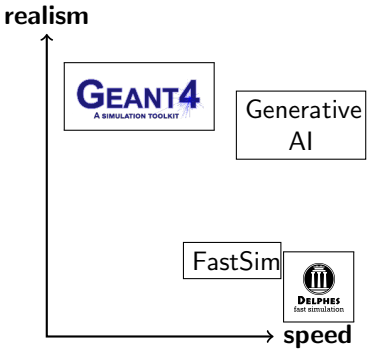


Generative Models and FastSimulation

- Detector Simulation is a Bottleneck in the Simulation Chain.
- Generative AI offers to improve the quality of FastSim.

We organized the CaloChallenge as Benchmark:

- 20+ original papers published, final write-up: 2410.21611
- Various correlations between quality metrics for all datasets.
- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.

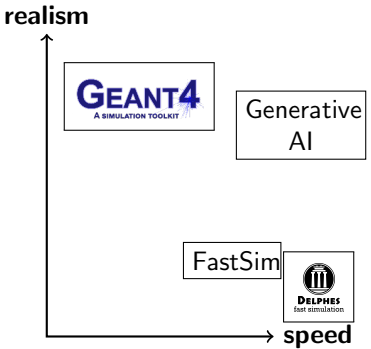


Generative Models and FastSimulation

- Detector Simulation is a Bottleneck in the Simulation Chain.
- Generative AI offers to improve the quality of FastSim.

We organized the CaloChallenge as Benchmark:

- 20+ original papers published, final write-up: 2410.21611
- Various correlations between quality metrics for all datasets.
- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.



Next steps:

- Embedding models in full fast simulation to see how trade-offs play out.
- Many ideas are now already finding their way into the experiments!

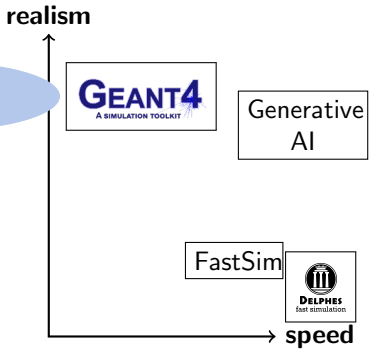
Generative Models and FastSimulation

- Detector Simulation is a Bottleneck in the Simulation Chain.
- Generative AI offers to improve the quality of FastSim

Thank you!

We organized the CaloChallenge as Benchmark:

- 20+ original papers published, final write-up: 2410.21611
- Various correlations between quality metrics for all datasets.
- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.



Next steps:

- Embedding models in full fast simulation to see how trade-offs play out.
- Many ideas are now already finding their way into the experiments!