CaloChallenge 2022 — Final Evaluation and Lessons Learned

— ML4Jets 2024 Paris —

Claudius Krause

Institute of High Energy Physics (HEPHY), Austrian Academy of Sciences (OeAW)

November 4, 2024

$\Rightarrow$ arXiv:2410.21611 $\Leftarrow$


arXiv:2410.21611v1 [cs.LG] 28 Oct 2024

Review Article

CaloChallenge 2022: A Community Challenge for Fast Calorimeter Simulation

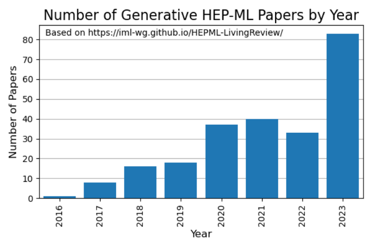
"Calorimeter Simulation", generated via midjourney, 2022

Claudius Krause[1,2] (main editor),
Michele Faucci Giannelli[3,4] (editor),
Gregor Kasieczka[5] (editor), Benjamin Nachman[6] (editor),
Dalila Salamani[7] (editor), David Shih[8] (editor),
Anna Zaborowska[7] (editor)

Oz Amram[9], Kerstin Borras[10,11], Matthew R. Buckley[8],
Erik Buhmann[5], Thorsten Buss[5,10],
Renato Paulo Da Costa Cardoso[7], Anthony L. Caterini[12],
Nadezda Chernyavskaya[7], Federico A.G. Corchia[13,14],
Jesse C. Cresswell[12], Sascha Diefenbacher[6],
Etienne Dreyer[15], Vijay Ekambaram[16], Engin Eren[10],
Florian Ernst[2,7], Luigi Favaro[2], Matteo Franchini[13,14],
Frank Gaede[10], Eilam Gross[15], Shih-Chieh Hsu[17],
Kristina Jaruskova[7], Benno Käch[5,10], Jayant Kalagnanam[18],
Raghav Kansal[9,19], Taewoo Kim[12], Dmitrii Kobylianskii[15],
Anatolii Korol[10], William Korcari[5]

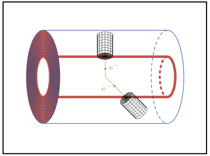
It all started in 2021 . . .

- . . . the LHC-Olympics had just concluded.
- Generative AI was kicking off in HEP in 2020.
- Applications to Detector Simulation, as major bottleneck, were gaining popularity.
- However, $\mathcal{O}(10)$ architectures used $\mathcal{O}(8)$ datasets.

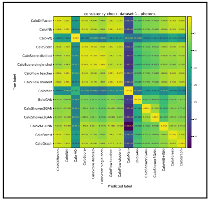


$\Rightarrow$ We created the CaloChallenge to benchmark and trigger new developments.

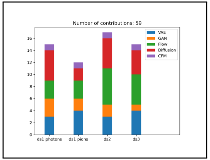
CaloChallenge 2022 — Final Evaluation and Lessons Learned



I: Datasets



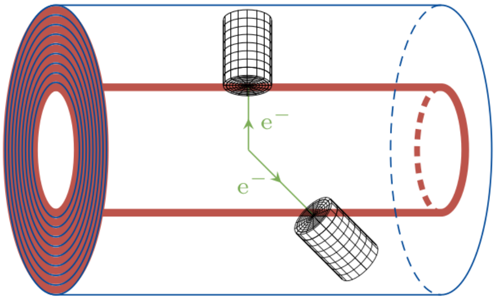
II: Evaluation Metrics

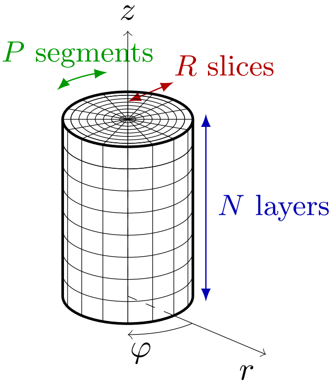


III: Results

CaloChallenge Showers are voxelized in cylindrical coordinates.

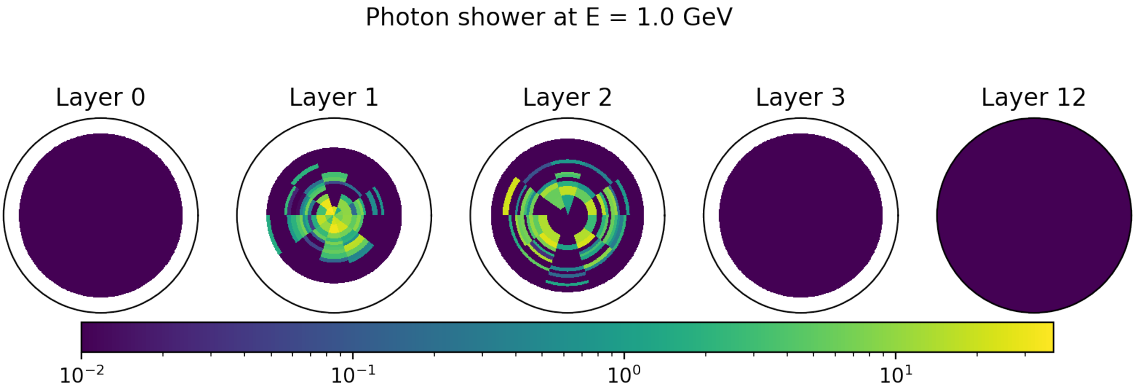
- There 4 datasets in increasing complexity / dimensionality.
- Particles enter perpendicular to front surface:





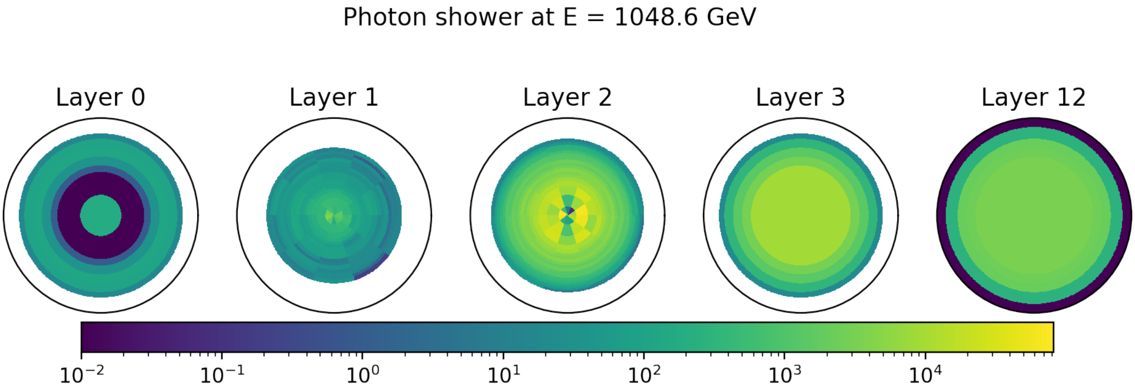
CaloChallenge Showers are voxelized in cylindrical coordinates.

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



CaloChallenge Showers are voxelized in cylindrical coordinates.

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



The Fast Calorimeter Simulation Challenge 2022

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

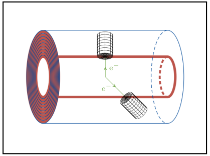
https://calochallenge.github.io/homepage/

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

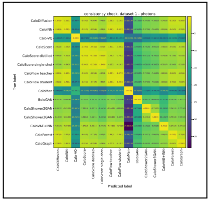
- Dataset 1: AtlFast3 trainig data (γ: 368, π: 533 voxels) $E_{\text{inc}} \in [256\text{ MeV}, 4.2\text{ TeV}]$
  [2109.02551, Comput.Softw.Big Sci.]
- Dataset 2: Par04 simulated detector (e^-: 6480 voxels) $E_{\text{inc}} \in [1\text{ GeV}, 1\text{ TeV}]$
- Dataset 3: Par04 simulated detector (e^-: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

CaloChallenge 2022 — Final Evaluation and Lessons Learned

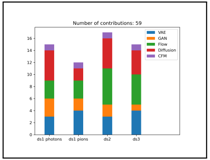
I: Datasets



<u>II: Evaluation Metrics</u>

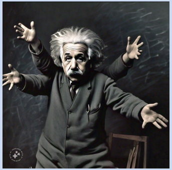


III: Results



How to evaluate generative models?

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



imagined with Meta AI.

In high-energy physics: need to find something better!
$\Rightarrow$ We want to correctly cover $p(x)$ of the entire phase space.

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

$\Rightarrow$ 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:

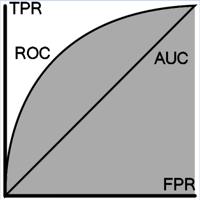
- The likelihood ratio is the most powerful test statistic to distinguish two samples.
- A powerful classifier trained to distinguish the samples should therefore learn (something monotonically related to) $w = \frac{p_{\text{data}}}{p_{\text{model}}}$.
- If this classifier is confused, we conclude $\Rightarrow$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

$\Rightarrow$ 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.

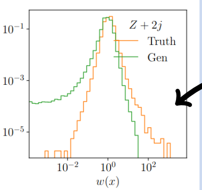
   But: are AUCs of different models really comparable?



A Classifier tells us much more about the model.

Failure modes of the model can now be seen in the $w = \frac{p_{\text{data}}}{p_{\text{model}}}$ histogram:

Data manifold over-populated by model:
$\Rightarrow$ missmodeled feature



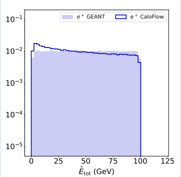
Data manifold not populated by model:
$\Rightarrow$ missed feature

R. Das et al. [2305.16774, SciPost]

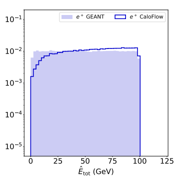
Cluster plots show where events lie in phase space:

figures by B. Schmidthaler / M. Rosendorf

small weights:



large weights:



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

A multi-class classifier:

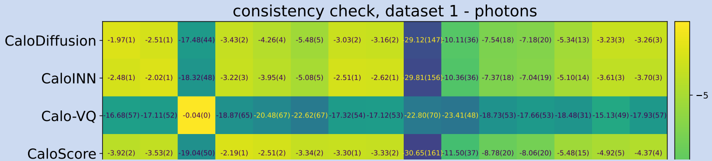
Train on submission 1 vs. submission 2 vs. … vs. submission n
and evaluate the *log posterior*:

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

3. As metric: evaluate with GEANT4

Lim et al. [2211.11765, MNRAS]

As cross-check: validate with all submissions j



Other quality metrics we looked at.

4. *KPD/FPD.* Kansal et al. [2211.10295, Phys.Rev.D]
   - Fréchet physics distance (FPD): Fréchet distance between Gaussian fits to obs.
   - Kernel physics distance (KPD): kernel-based MMD between observables.

5. *Pearson Correlation* between layer energies. Ahmad et al. [2406.12898]

6. *Precision / Recall / Density / Coverage.* Naeem et al. [2002.09797]
   - How many "real" samples are close to "fake" manifold.
   - How many "fake" samples are close to "real" manifold.

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

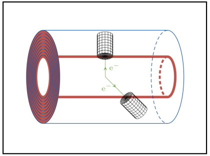
- start singularity container
- load model weights + biases
- generate samples
- save them to `.hdf5`

⇒ The *number of trainable parameters*.

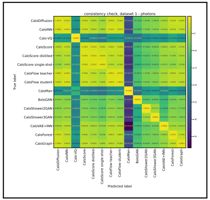
- as proxy for model size
- in training / generation

CaloChallenge 2022 — Final Evaluation and Lessons Learned

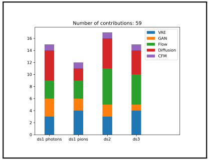
I: Datasets



II: Evaluation Metrics



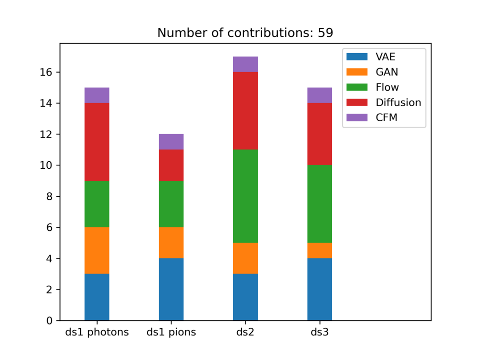
III: Results



The ~~preliminary~~ final! results of the CaloChallenge

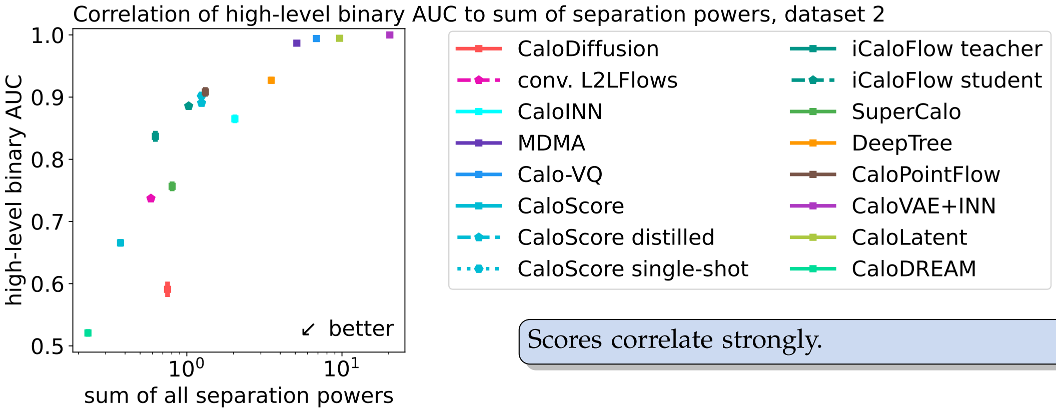
I will only be able to share some highlights of the results of the CaloChallenge.

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



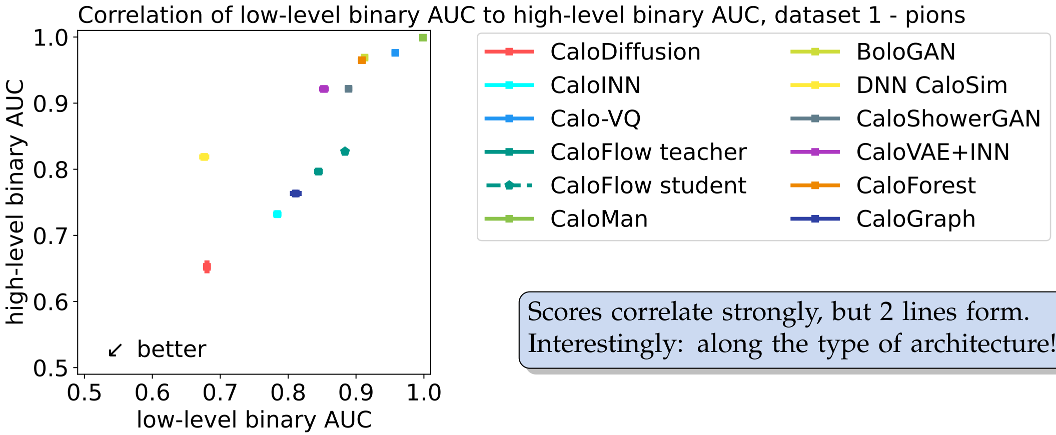
- We received 59 submissions for all datasets.
- They were generated by 23 different models.
- All types of generative AI architectures were used.

Comparing different quality metrics: high-level features



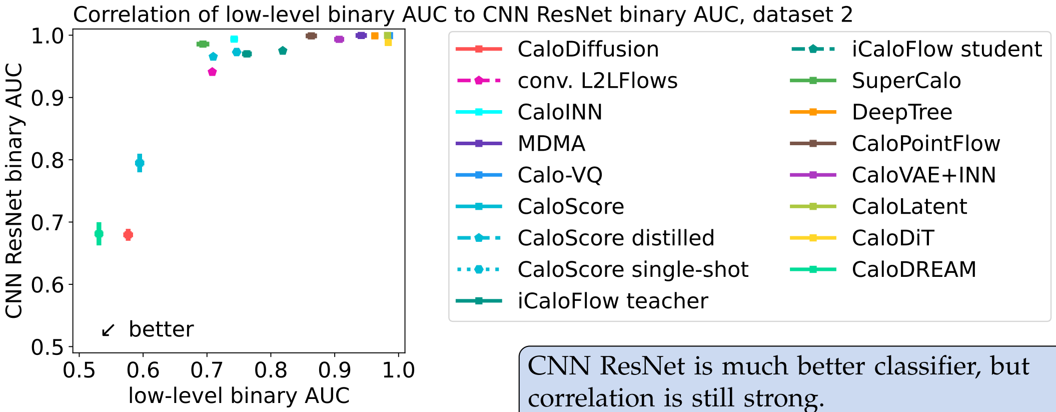
Scores correlate strongly.

Comparing different quality metrics: classifier input



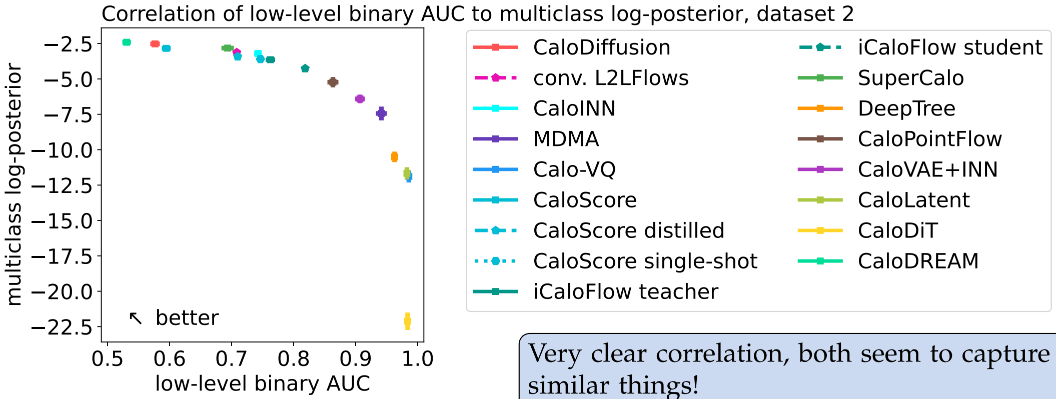
Scores correlate strongly, but 2 lines form.
Interestingly: along the type of architecture!

Comparing different quality metrics: classifier architecture



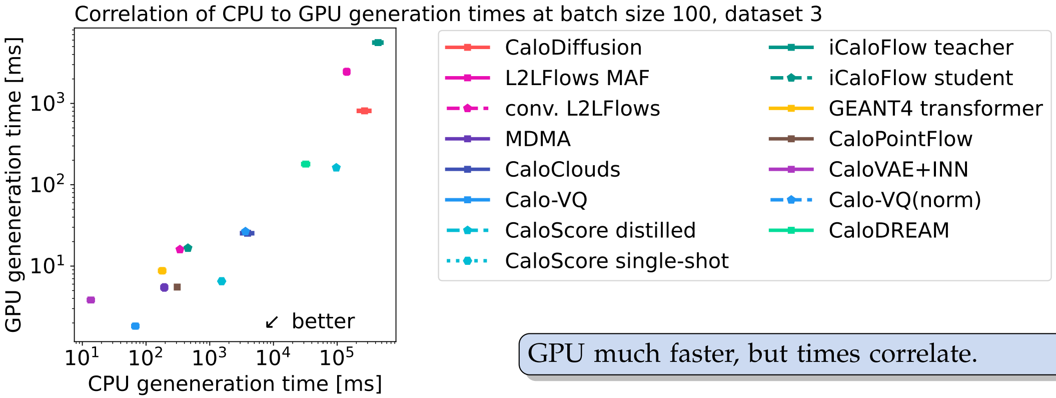
CNN ResNet is much better classifier, but correlation is still strong.

Comparing different quality metrics: binary vs. multiclass



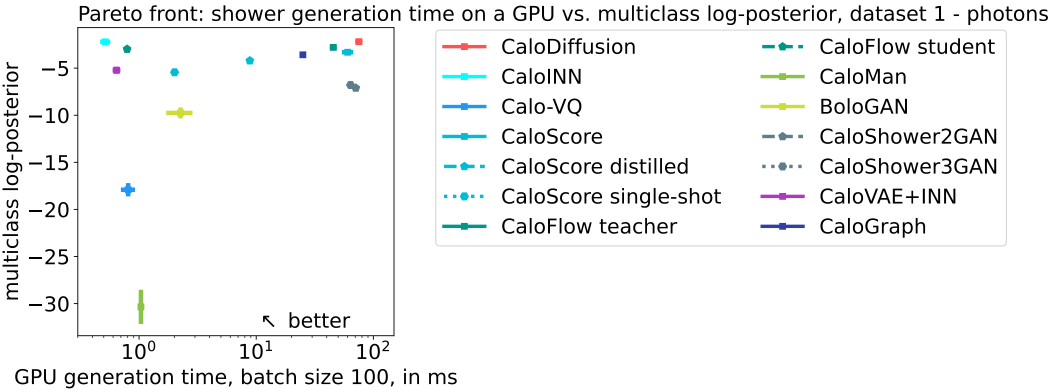
Very clear correlation, both seem to capture similar things!

Comparing different timing metrics: CPU vs. GPU

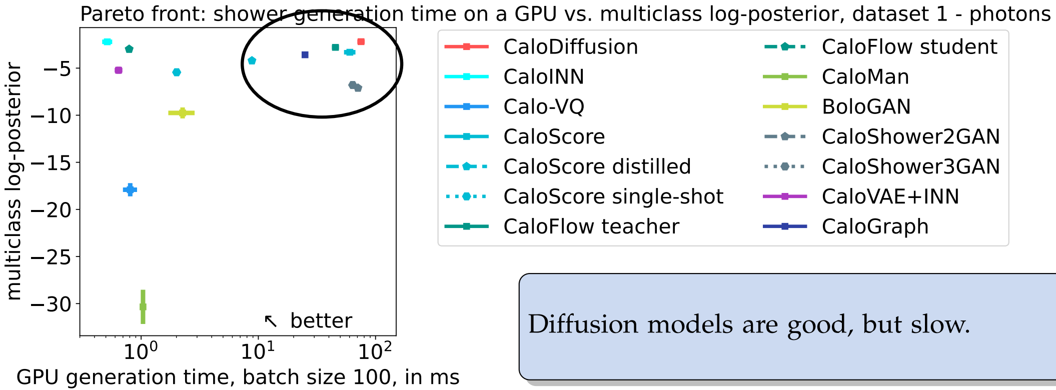


GPU much faster, but times correlate.

Pareto Fronts: Quality vs. Generation Time

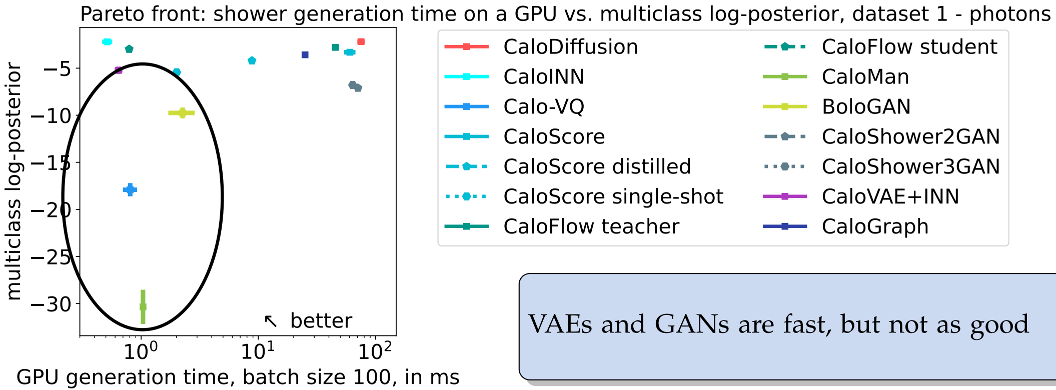


Pareto Fronts: Quality vs. Generation Time



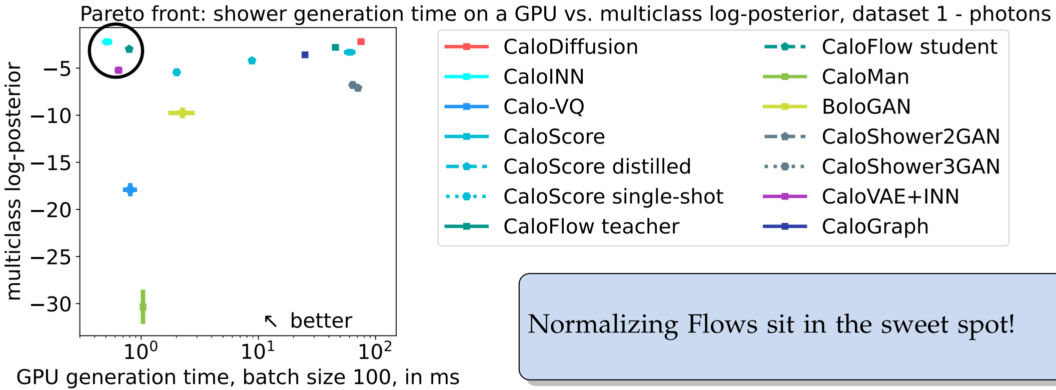
Diffusion models are good, but slow.

Pareto Fronts: Quality vs. Generation Time



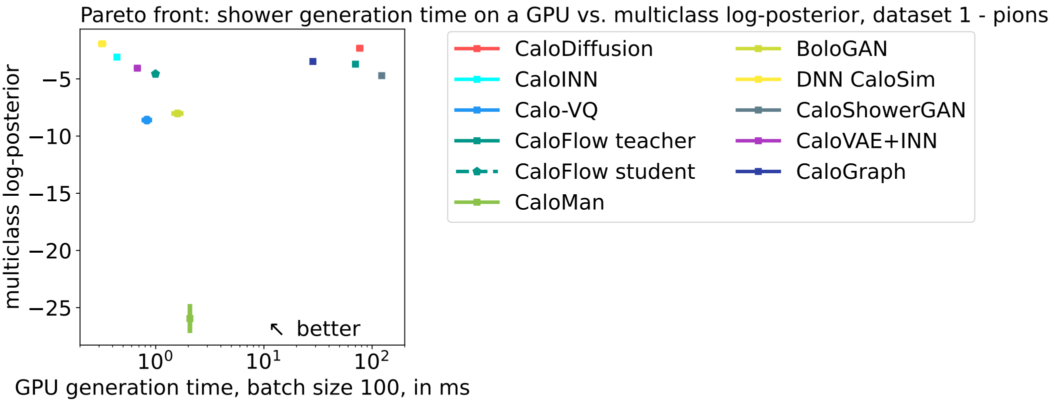
VAEs and GANs are fast, but not as good

Pareto Fronts: Quality vs. Generation Time

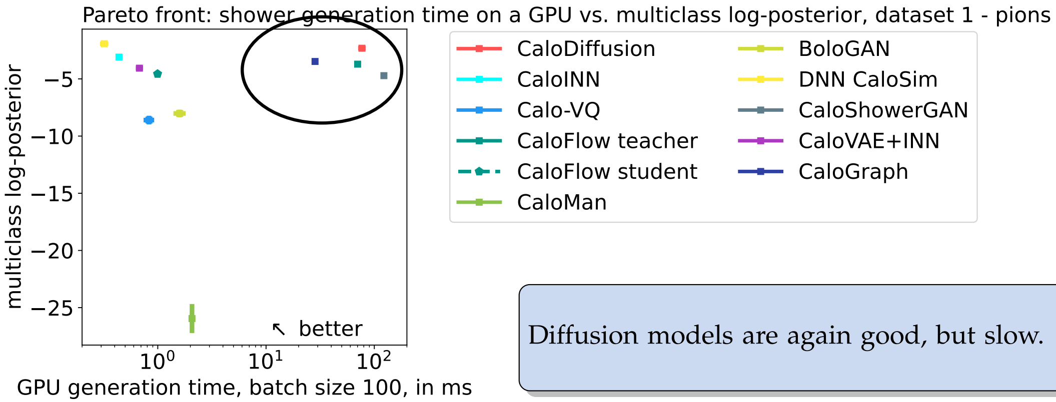


Normalizing Flows sit in the sweet spot!

Pareto Fronts: Quality vs. Generation Time

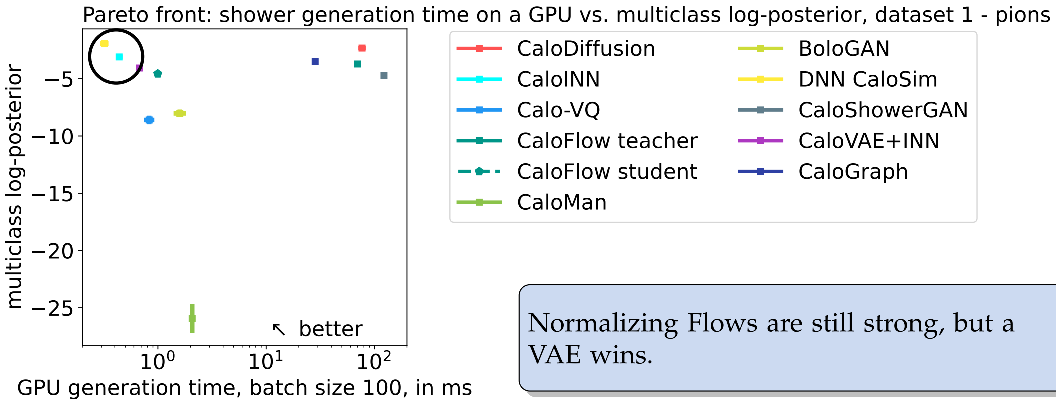


Pareto Fronts: Quality vs. Generation Time



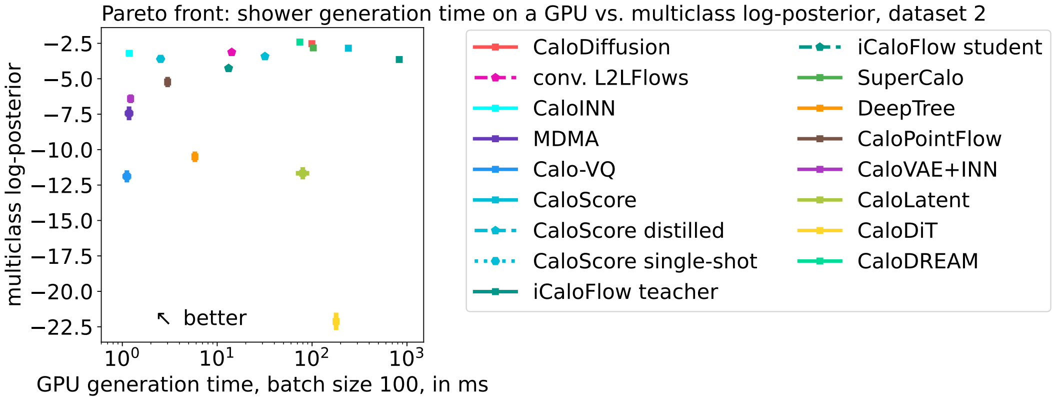
Diffusion models are again good, but slow.

Pareto Fronts: Quality vs. Generation Time

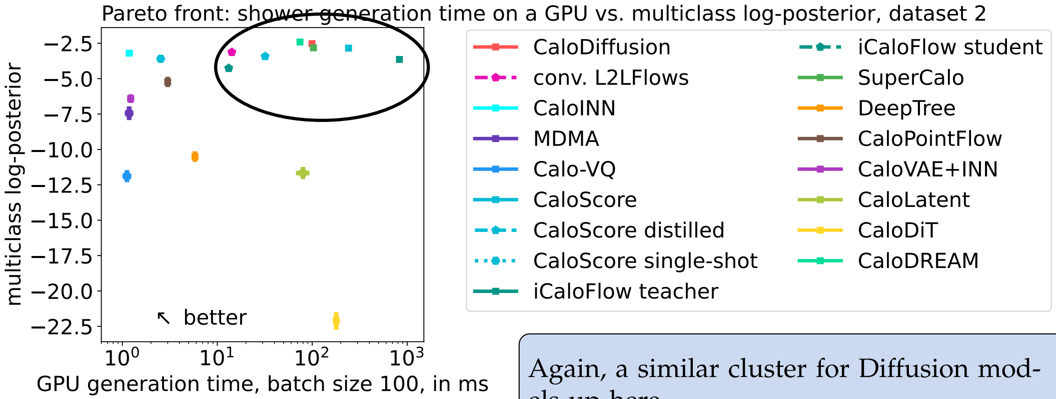


Normalizing Flows are still strong, but a VAE wins.

Pareto Fronts: Quality vs. Generation Time

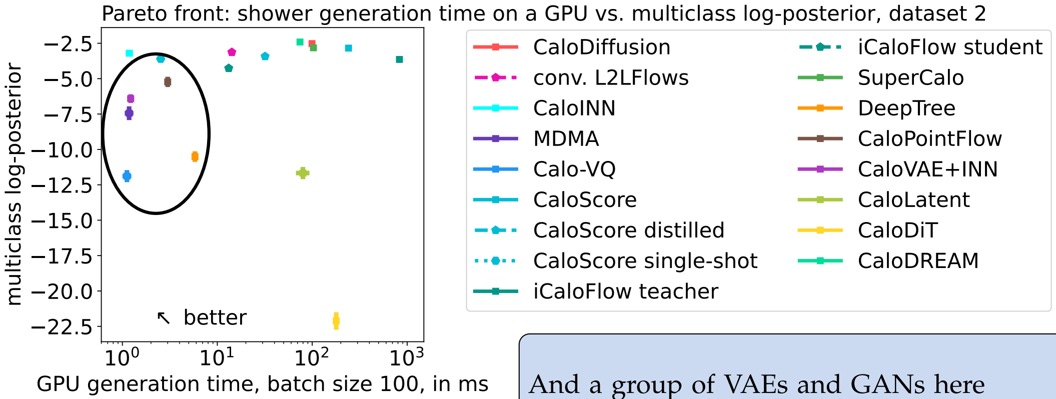


Pareto Fronts: Quality vs. Generation Time



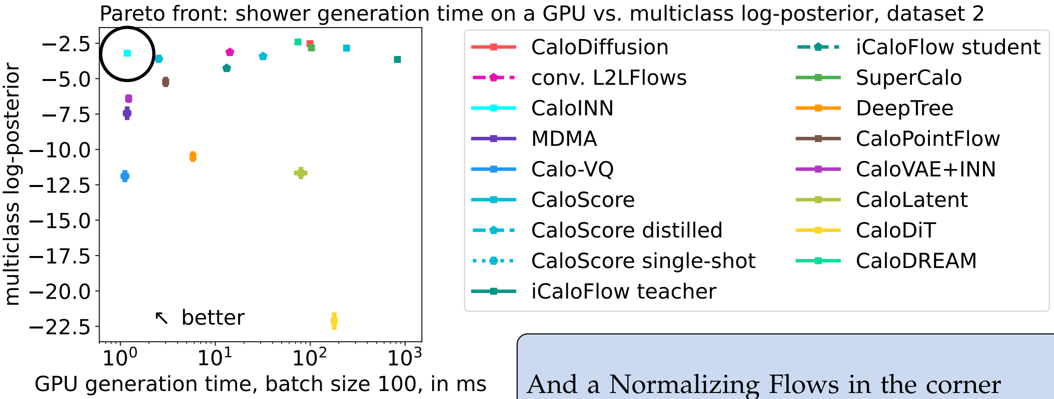
Again, a similar cluster for Diffusion models up here.

Pareto Fronts: Quality vs. Generation Time



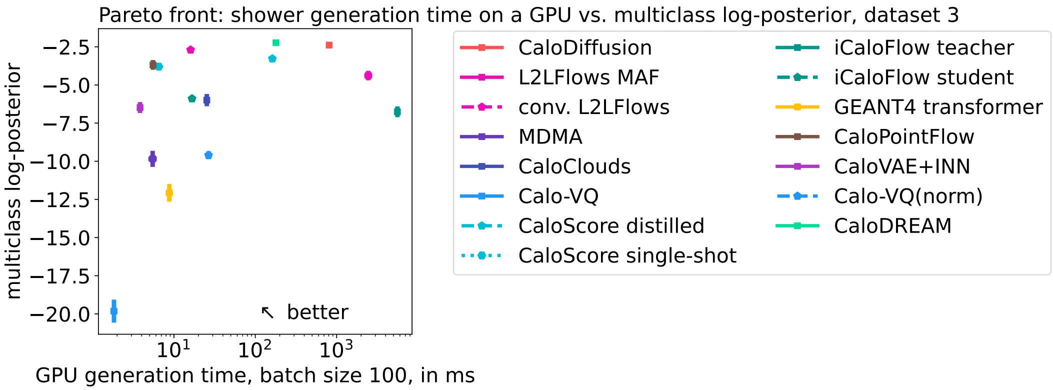
And a group of VAEs and GANs here

Pareto Fronts: Quality vs. Generation Time

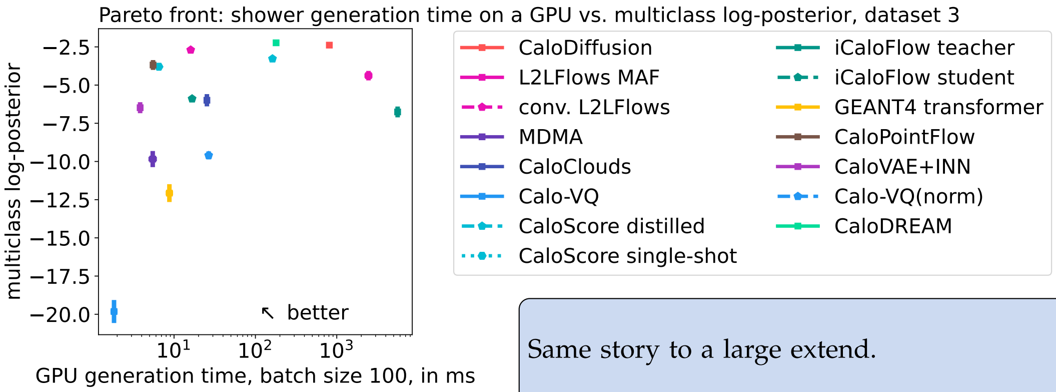


And a Normalizing Flows in the corner

Pareto Fronts: Quality vs. Generation Time

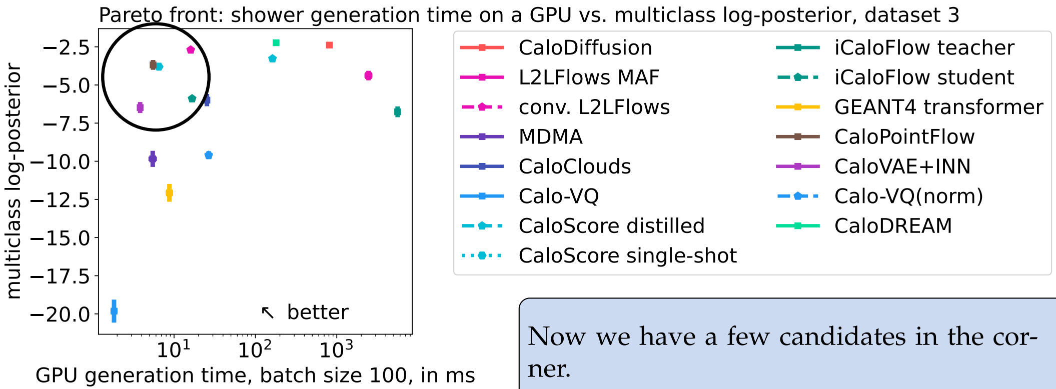


Pareto Fronts: Quality vs. Generation Time



Same story to a large extend.

Pareto Fronts: Quality vs. Generation Time



Now we have a few candidates in the corner.

CaloChallenge 2022 — Final Evaluation and Lessons Learned

The CaloChallenge was well-received in the community:

- 20+ papers
- Even more talks at ML4Jets / ML 4 Physical Sciences@NeurIPS / CHEP / ...
- Many discussions and feedback on evaluation metrics etc.
- All repositories are public!

CaloChallenge 2022 — Final Evaluation and Lessons Learned

The CaloChallenge was well-received in the community:

- 20+ papers
- Even more talks at ML4Jets / ML 4 Physical Sciences@NeurIPS / CHEP / …
- Many discussions and feedback on evaluation metrics etc.
- All repositories are public!

Final evaluation:

- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.

CaloChallenge 2022 — Final Evaluation and Lessons Learned

The CaloChallenge was well-received in the community:

- 20+ papers
- Even more talks at ML4Jets / ML 4 Physical Sciences@NeurIPS / CHEP / ...
- Many discussions and feedback on evaluation metrics etc.
- All repositories are public!

Final evaluation:

- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.

Lessons Learned:

- Various correlations between quality metrics for all datasets.
- Next step: embedding models in full fast simulation to see how trade-offs play out.

CaloChallenge 2022 — Final Evaluation and Lessons Learned

The CaloChallenge was well-received in the community:

- 20+ papers
- Even more talks at ML4Jets / ML 4 Physical Sciences@NeurIPS / CHEP / ...
- Many discussions and feedback on evaluation metrics etc.
- All repositories are public!

Thank you!

Final evaluation:

- Quality: Diffusion and CFM better than NF better than GAN/VAE.
- Speed: GAN/VAE faster than NF faster than Diffusion and CFM.

Lessons Learned:

- Various correlations between quality metrics for all datasets.
- Next step: embedding models in full fast simulation to see how trade-offs play out.