Estimating Support Size of the 3DGAN

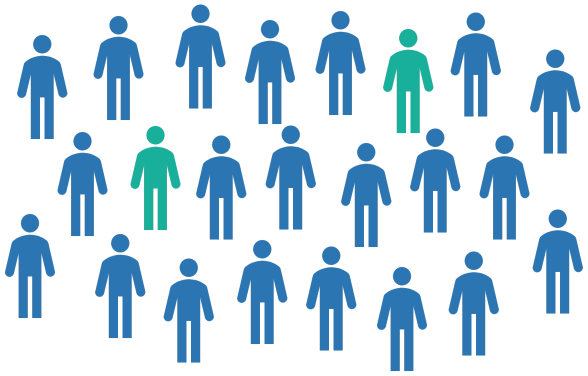
Kristina Jaruskova, Czech Technical University in Prague
(2020 CERN openlab online summer project)
***Sofia Vallecorsa*, CERN openlab, IT Dept.**

IML Meeting, November 23rd 2020

Support size of a GAN

- Comparison of true distribution and learnt distribution support sizes
    - New information about the GAN behavior
- Low support size of the GAN
    - Learnt distribution not representative enough, not covering all possible samples
    - Similar images appears more frequently

- Test based on the **birthday paradox** to measure the support size
    - First conducted in Arora and Sanjeev, 2017
    - Empirical evidence that even with the training objective approaching the optimum, the support size can be still small



Outline

1. Birthday paradox
2. Use case: 3DGAN
3. Measuring similarity of samples
   - SSIM
   - Shower shapes
   - Deposited energy
4. Estimates of the support size
5. Conclusions

Birthday paradox

How many people need to be in one room so that
P(*at least two people were born on the same day of the year*) > 0.5 ?

- 365 days in a year → 23 people is enough
- Generalized problem – a year with d **days** → approx. $\sqrt{d}$ people are needed
  - Brink, 2012

How many samples is it necessary to generate to have
P*(at least one pair of duplicates among the samples)* > 0.5 ?

- **(The answer)^2** = estimate of the support size
- The same question for the training data

Birthday paradox for GANs

Original birthday paradox problem

- Days in a year – finite set of possible values with discrete uniform distribution
- Duplicates – people born on the same day

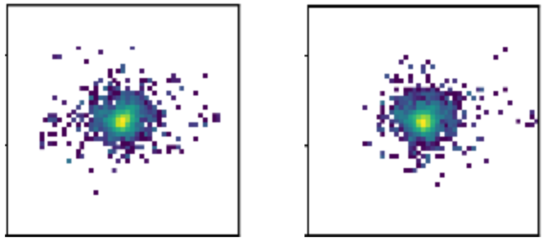
GAN distribution

- Images – pixels of continuous values
- Multivariate continuous distribution → occurrence of exact duplicates has zero probability
- Duplicates – images "similar enough"
- Metrics of similarity depend on the use case and the type of data

Exact duplicates

July 4 **July 4**

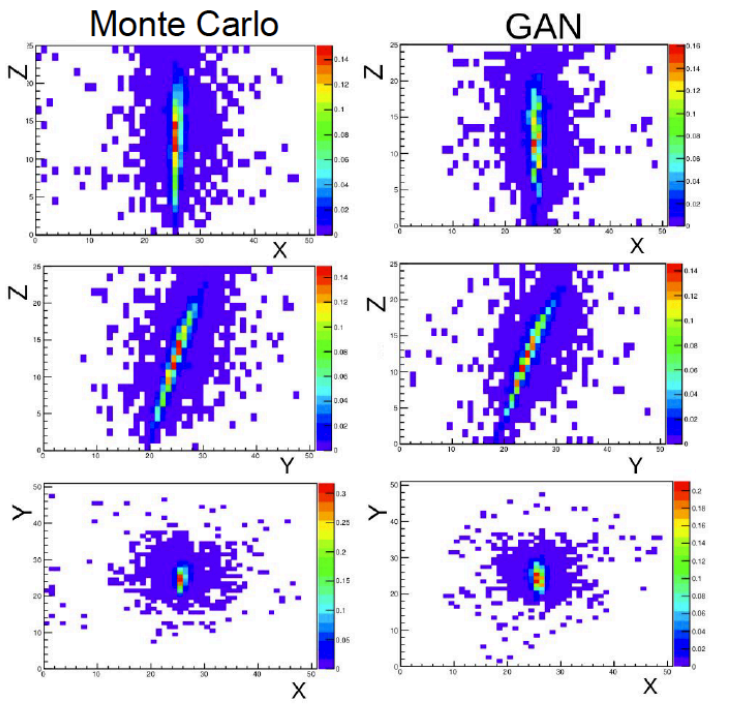
Not exact duplicates
But similar enough?

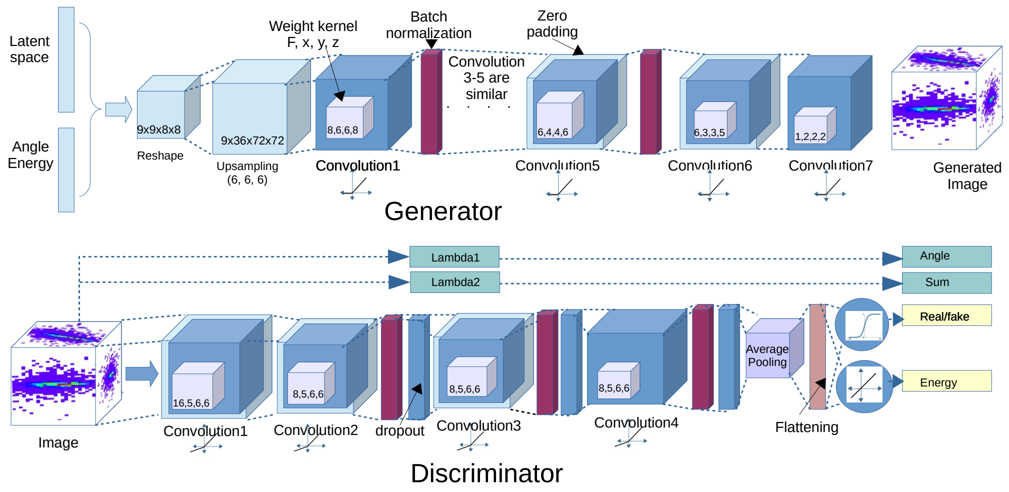


Use case: 3DGAN

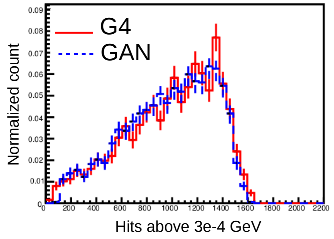
[G. Khattak, ICMLA 2019](#)

- Convolutional GAN architecture
- Simulates 3D output (51x51x25) of high granularity EM calorimeter
- Remarkable agreement to Monte Carlo
  - can we quantify similarity?

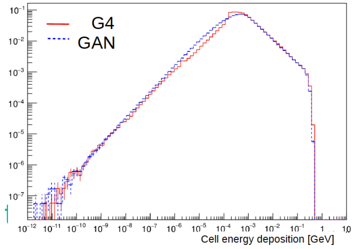




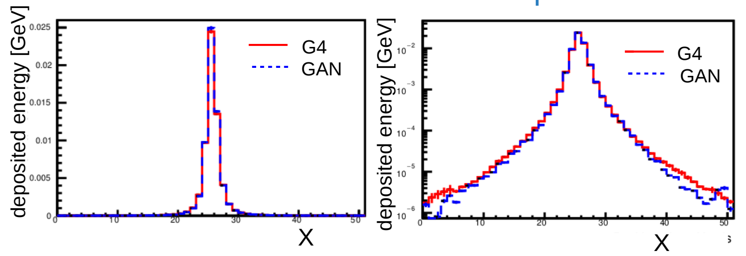
N_{cells} above 300 keV



Cell energy



Transverse shower shape



Longitudinal shower shape

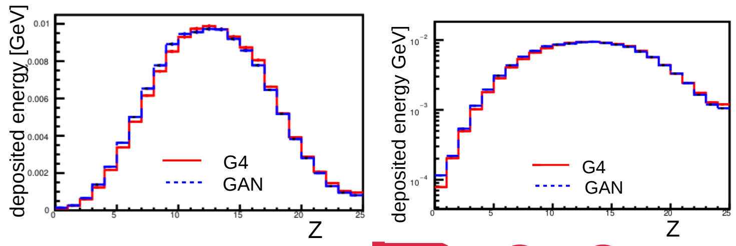


Image similarity

Pixel-based metric

- GAN suffer from well known mode-collapse and mode-drop problems
- How much diversity there is in the generated sample?
- Structural Similarity Index

$$\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

$$\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y}) = 1 \Leftrightarrow \boldsymbol{x} = \boldsymbol{y}$$

where $\boldsymbol{x}, \boldsymbol{y}$ are two samples to be compared

- Calculated for an area given by a smooth sliding window, then averaged.
- Computed in *xy* planes
  - 3rd dimension is channel

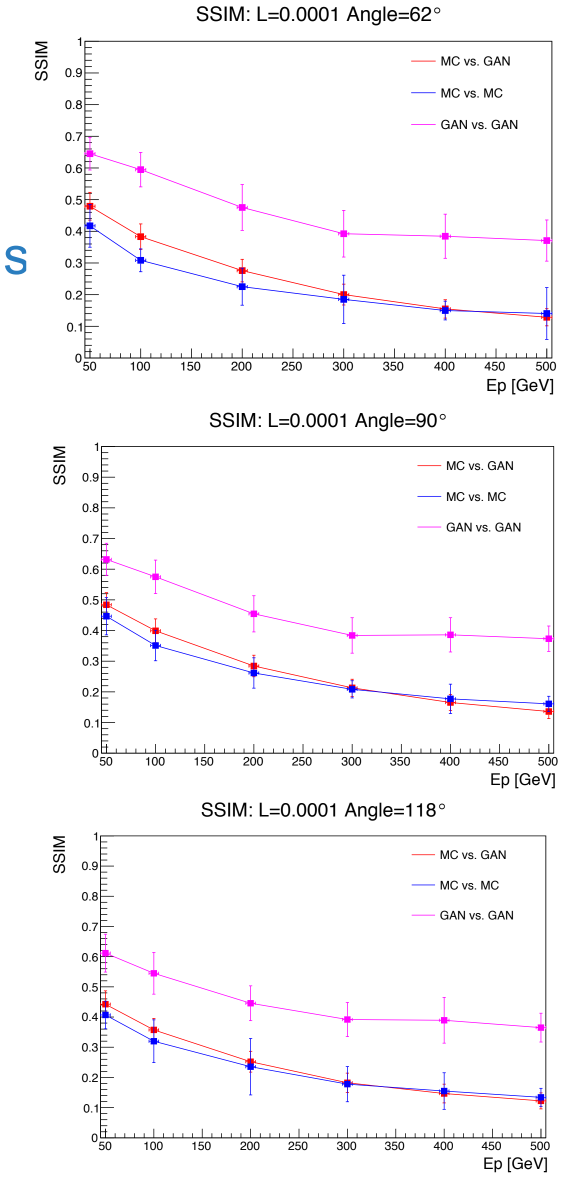


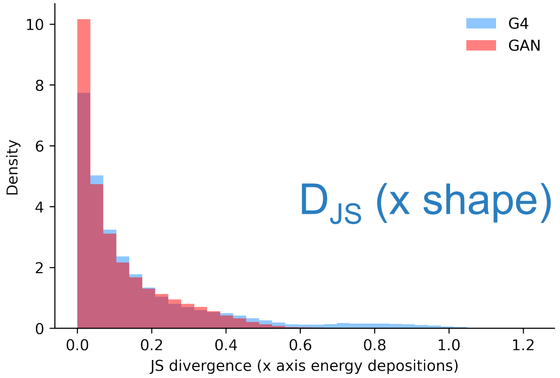
Image similarity

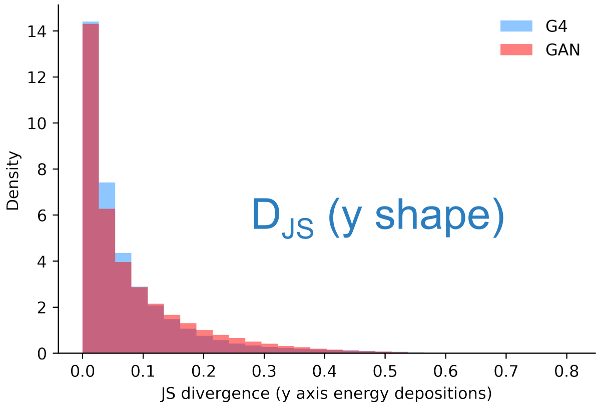
High-level features

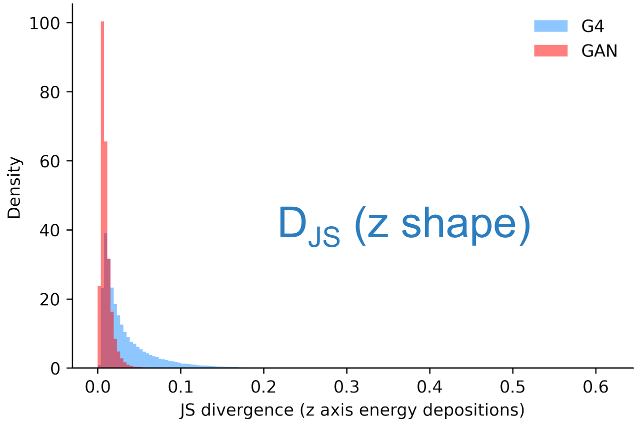
$$D_{JS}(P,Q) = \frac{1}{2} \cdot D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} \cdot D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

$$D_{KL}(P,Q) = P \cdot \ln\left(\frac{P}{Q}\right) - P + Q$$

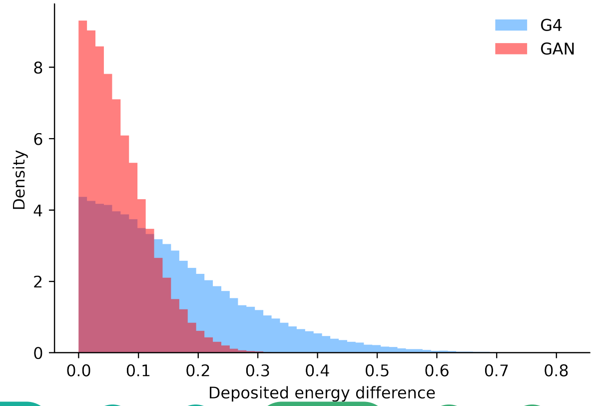
- Shower shapes along axes *x*, y and *z*
  - Jensen-Shannon divergence







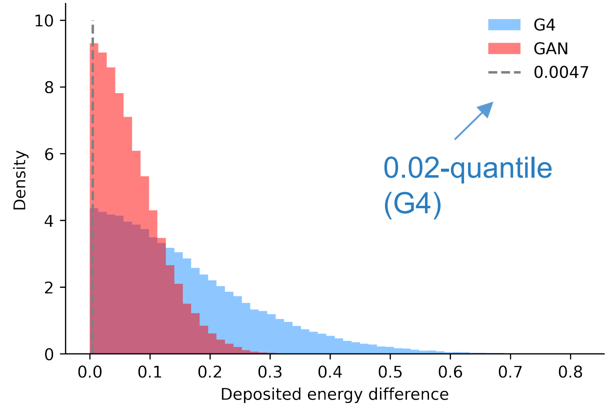
- Total deposited energy
  - Absolute difference

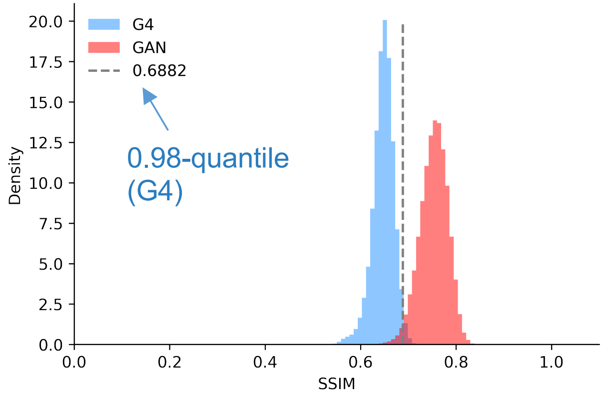


CERN openlab

Definition of duplicates

1. Compute similarity metrics between samples on GEANT4 data
2. Find α-quantile of the distances computed on training data.
   - α = 0.02 (shower shapes, deposited energy)
   - α = 0.98 (SSIM)
3. Use the α-quantile as a threshold value for the definition of duplicates.
   - Distances below (or above) the threshold indicate duplicate samples.
4. Compute similarity metrics between the generated samples.
5. Combine the threshold conditions for all features.





Estimates of support size

Choose a subset → **Compute distances** → **Check for duplicates** → **Repeat 1 000 x** → **Calculate probability**

Choose a subset: Randomly select a subset of *s* samples.

Compute distances: Determine distances between all possible pairs of samples in the given subset.

Check for duplicates: Use the thresholds to find pairs of samples that are similar enough in terms of all selected metrics.

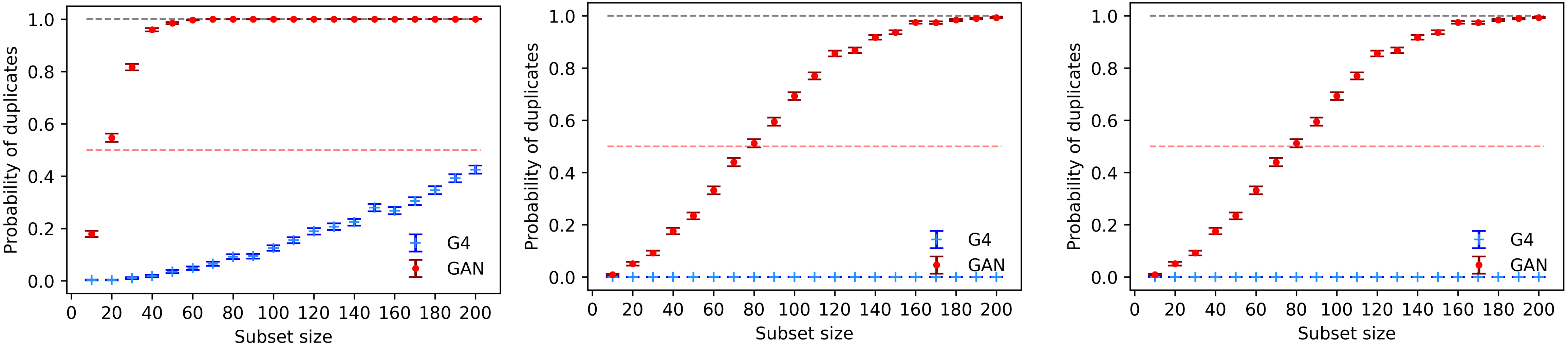
Repeat 1 000 x: Repeat the subset selection and duplicates detection.

Calculate probability: Determine the probability of encountering at least one pair of duplicates in a subset of *s* samples.

Estimates of support size

Probabilities of encountering duplicates for sets of different sizes

Probability exceeds 0.5 → estimate of the support size



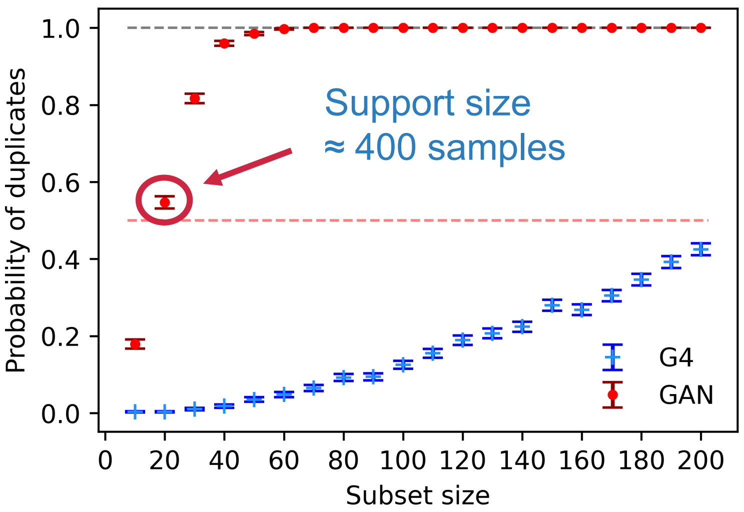
a) Shower shapes

b) Shower shapes and deposited energy

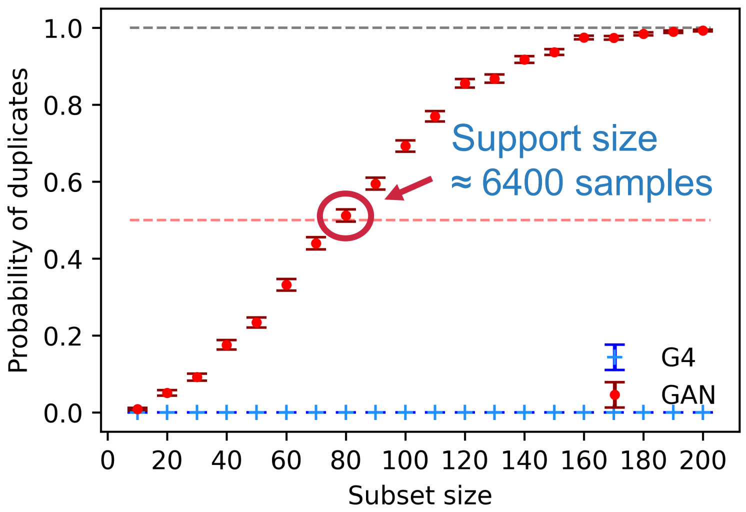
c) Shower shapes, deposited energy and SSIM

Estimates of support size

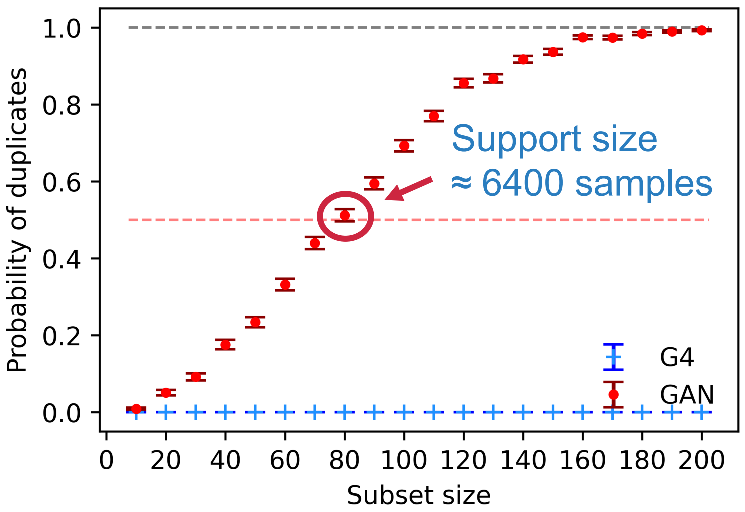
- GAN samples significantly more similar → smaller support size
- Adding SSIM as an additional metrics did not change the results



a) Shower shapes



b) Shower shapes and deposited energy



c) Shower shapes, deposited energy and SSIM

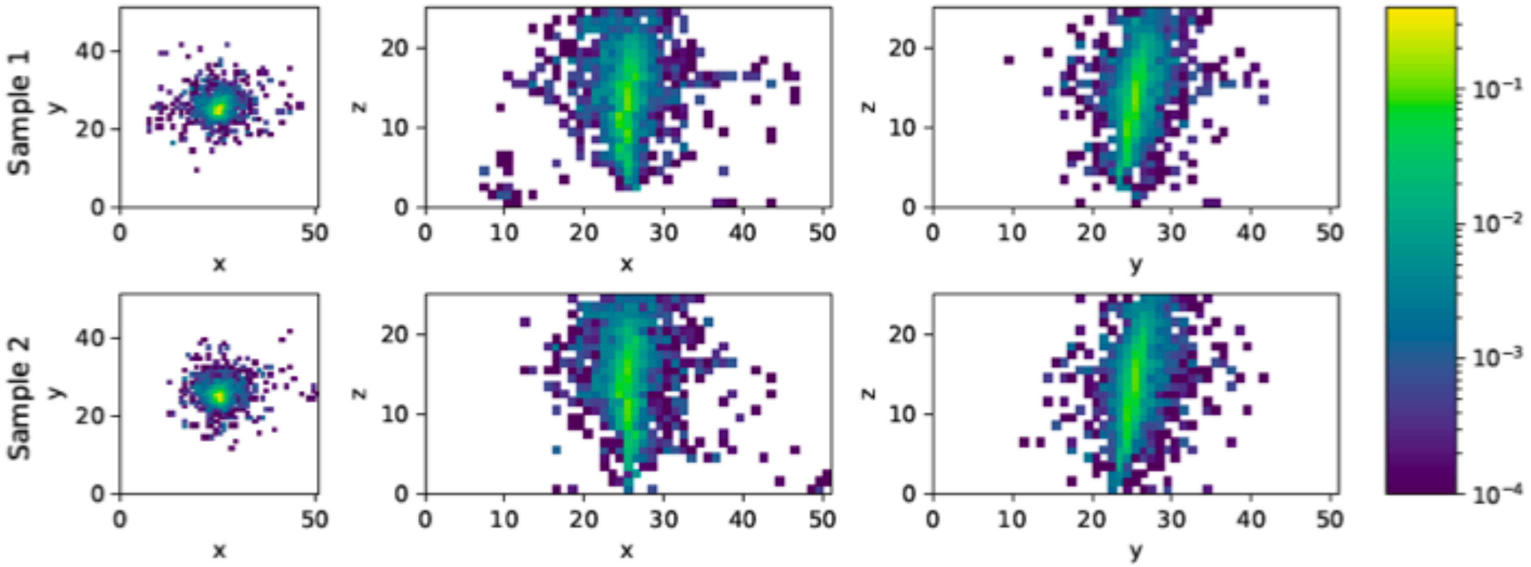
Accepted at NeurIPS 2020 workshop

Summary

… and discussion

- 3DGAN produces significantly more similar images than Monte Carlo
  - Much smaller support size → room for improvement
- This test depends strongly on duplicates definition.
  - Features: High-level physics variables, pixel-based features
  - Metrics: How do we measure similarity of these features
  - Not adapted to our problem?

When are two samples "similar enough" ?



Thank you

Questions?