CERN openlab

# Estimating Support Size of the 3DGAN

Kristina Jaruskova, Czech Technical University in Prague

Sofia Vallecorsa, CERN openlab

4th IML Machine Learning Workshop

October 2020

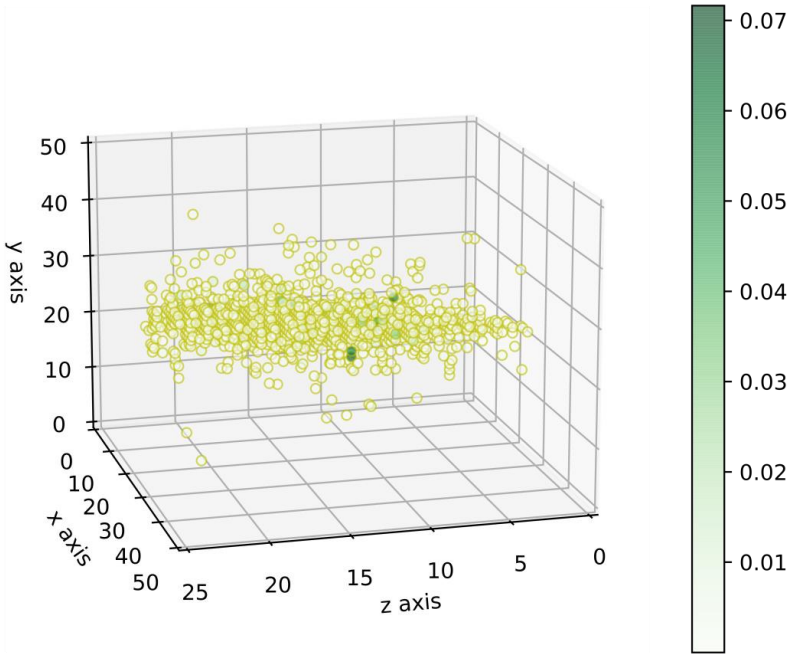
# Support size of 3DGAN

Support size of GAN

- Size of the support space of the learnt distribution
- Low support size → generated samples does not represent the target distribution

3DGAN prototype

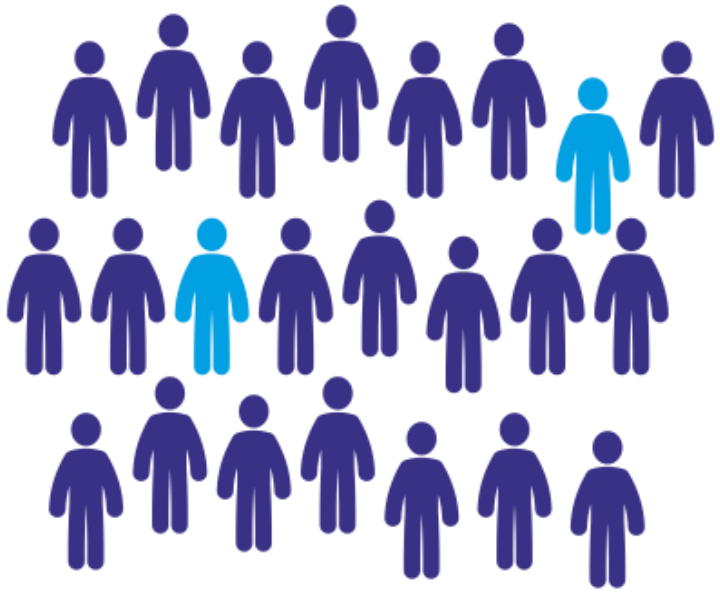
- Convolutional GAN architecture
- Calorimeter's energy response
- Output: 3D image (51x51x25) representing the deposited energy



# Birthday Paradox

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 ?



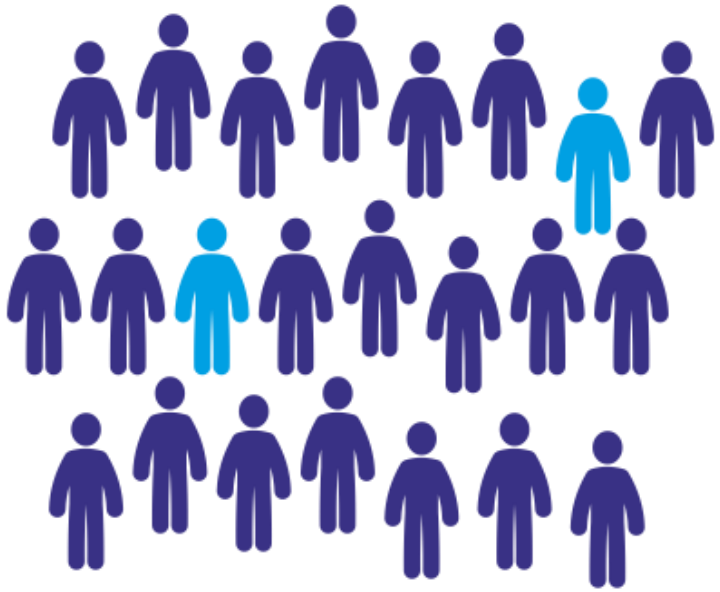
- 365 (366) days in a year
  -> 23 people is enough
- For a year with $d$ days, approx. $\sqrt{d}$ people are needed.

# Birthday Paradox and 3DGAN

For 3DGAN:

- How many samples do I need to generate to have at least one pair of duplicate samples with the probability of 50 %?
  - (The answer)^2 = estimate of the support size
- How many training data do I need to take to encounter duplicates?

- Enables comparison:
  Support size of GAN vs. support size of training data



# Birthday Paradox and 3DGAN

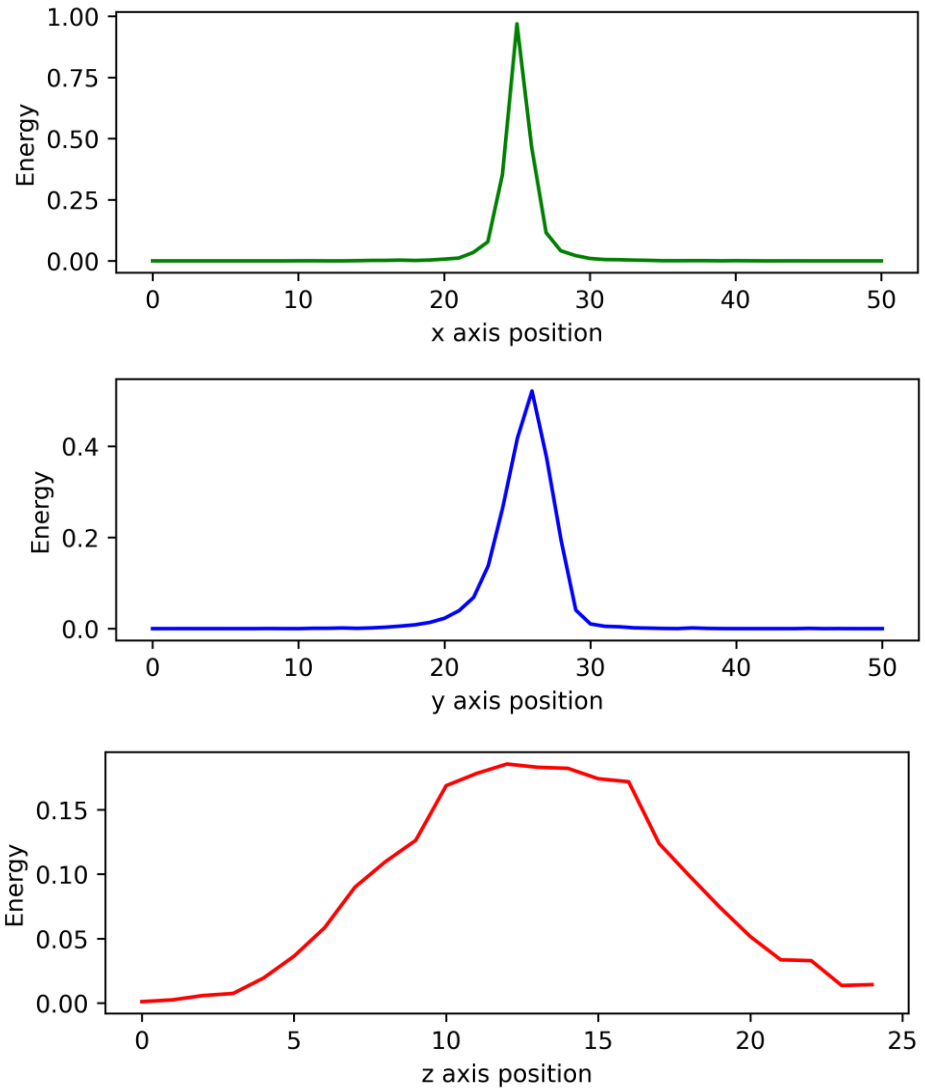
*Definition of duplicates*

High-level physics features

- Energy distributions along the main axes *x*, *y*, and *z*
- Total deposited energy

Pixel-based metric

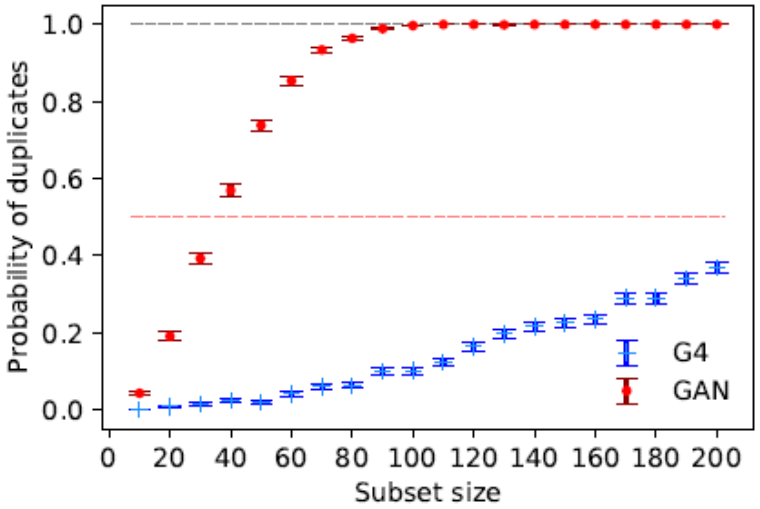
- Structural Similarity Index (SSIM)



# Results

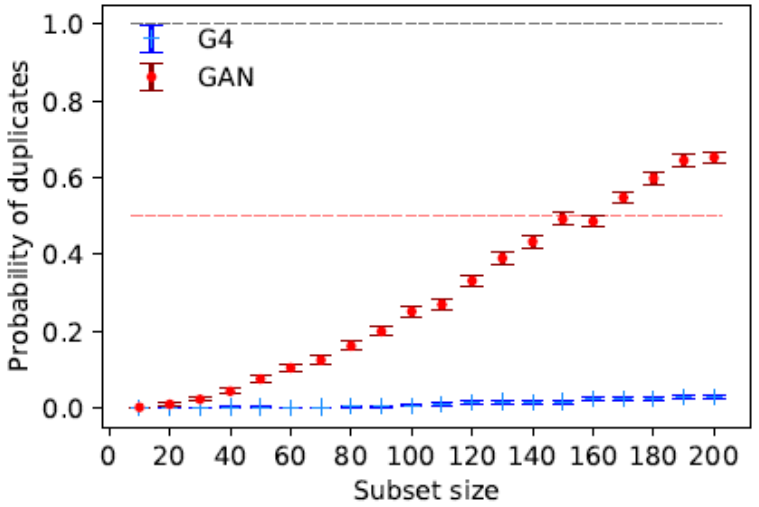
*For a year with $d$ days, approx. $\sqrt{d}$ people are needed.*

Probabilities of encountering duplicates for sets of different sizes (denoted as subset size). The first subset size for which the probability of 0.5 is exceeded gives an estimate of the support size.

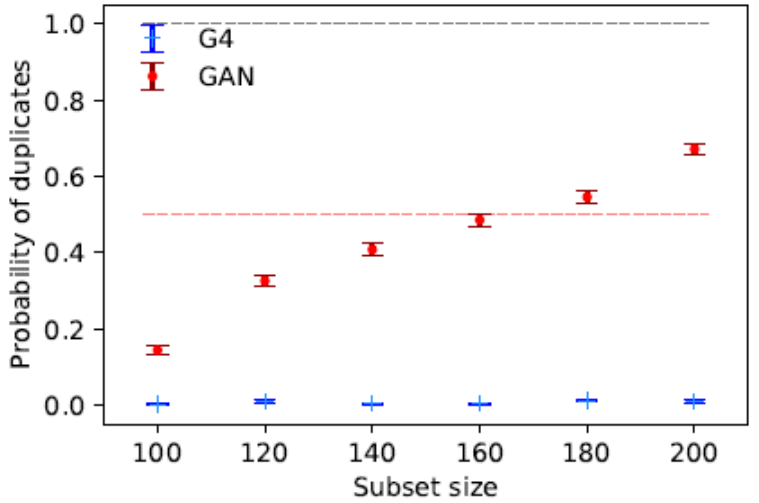



a) Shower shapes

G4 – GEANT4 (training data)




b) Shower shapes and deposited energy

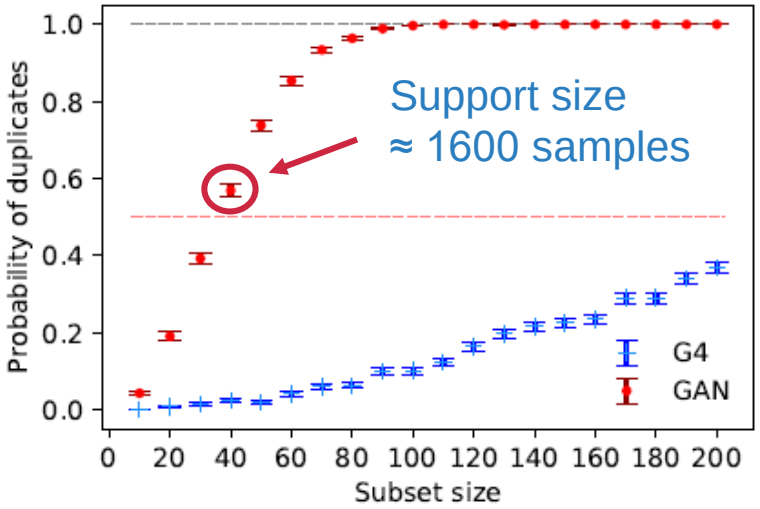



c) Shower shapes, deposited energy and SSIM

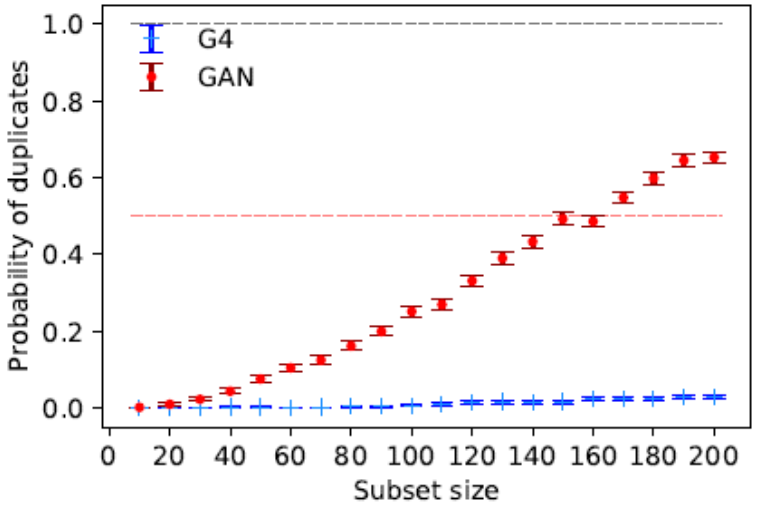
# Results

*For a year with $d$ days, approx. $\sqrt{d}$ people are needed.*

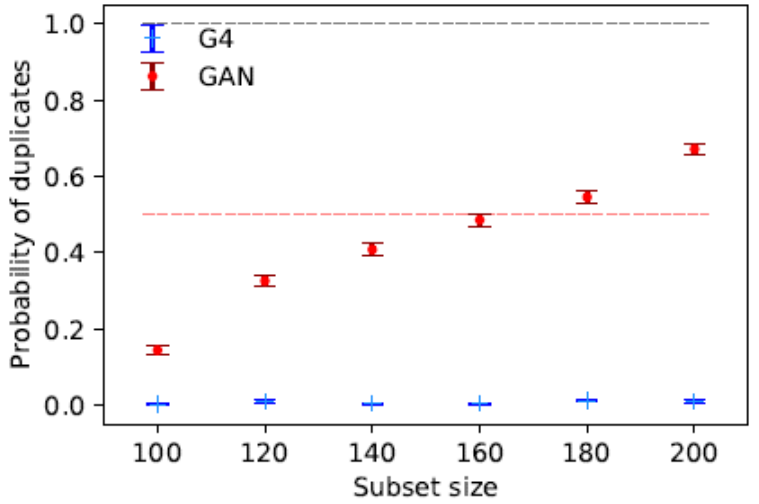
Probabilities of encountering duplicates for sets of different sizes (denoted as subset size). The first subset size for which the probability of 0.5 is exceeded gives an estimate of the support size.




a) Shower shapes




b) Shower shapes and deposited energy




c) Shower shapes, deposited energy and SSIM

G4 – GEANT4 (training data)

# Conclusions

- 3DGAN produces significantly more similar images than the Monte Carlo GEANT4 toolkit.
  - In terms of shower shapes, deposited energy and SSIM

- The estimates of support size depend strongly on duplicates definition.
  - Features: High-level physics variables, pixel-based metrics
  - Metrics: How do we measure similarity of these features
  - Strongest limitation

# Thank you for your attention. Any questions?

*Kristina Jaruskova*

*kristinajaruskova@gmail.com*

CERN openlab

# Estimating Support Size of the 3DGAN

Kristina Jaruskova, Czech Technical University in Prague

Sofia Vallecorsa, CERN openlab

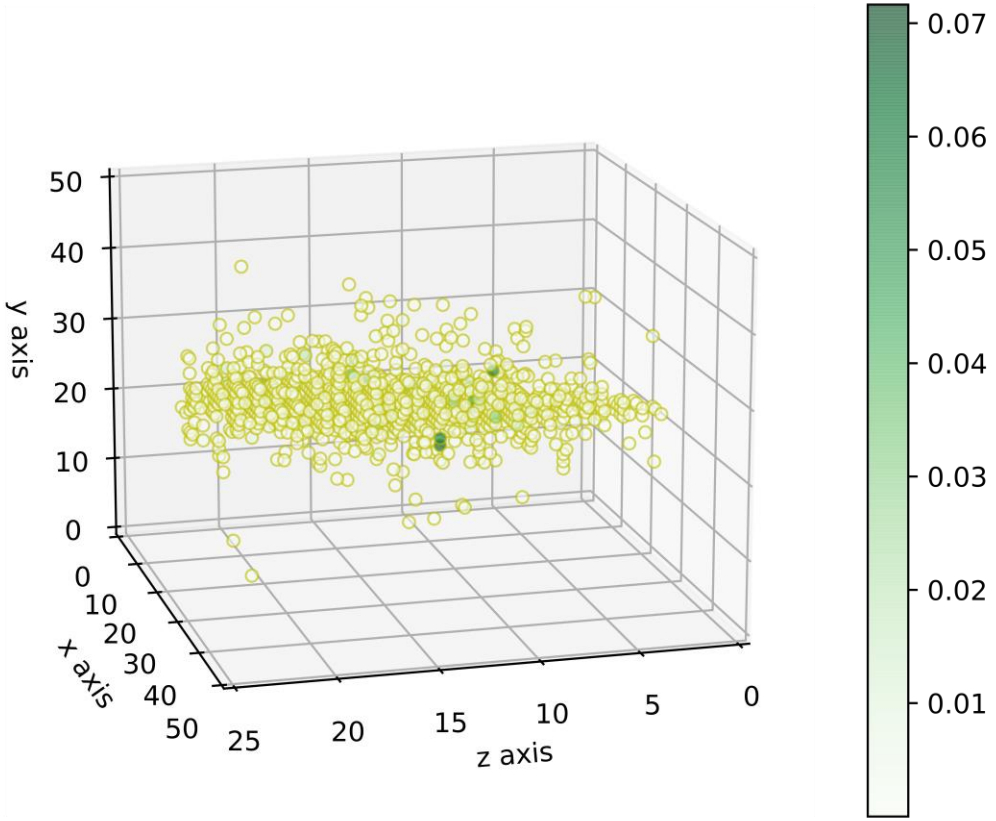
4th IML Machine Learning Workshop

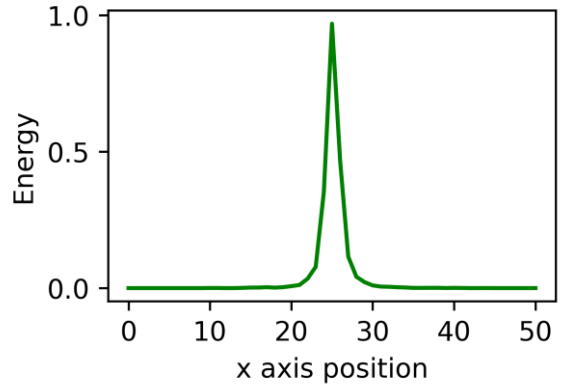
October 2020

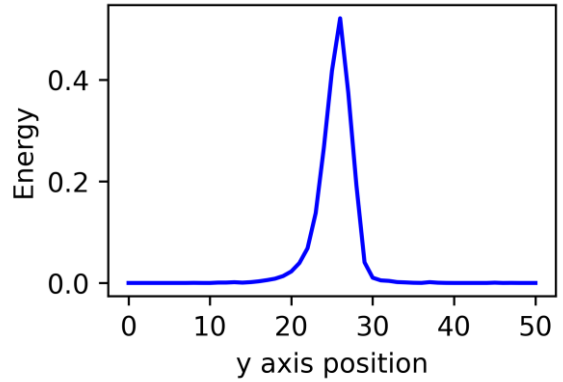
# Definition of duplicates

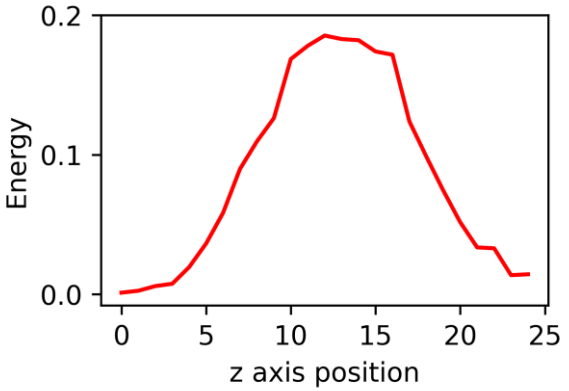
1. Compute distances between samples on GEANT4 data (training data)
2. Find $\alpha$-quantile of the distance computed on training data.
   - $\alpha = 0.02$
3. Use the $\alpha$-quantile as a threshold value for the definition of duplicates.
   - Distances below (or above) the threshold indicate duplicate samples
4. Compute distances between all generated samples for all selected features.
5. Combine the threshold conditions for all features
   - Be below threshold for distances between shower shapes along all axes, deposited energy and above the threshold for the SSIM

# Shower shapes



# Shower shapes

*Measuring the distance*

- Shower shape along *x* = energy distribution along the *x* axis
- Jensen-Shannon divergence

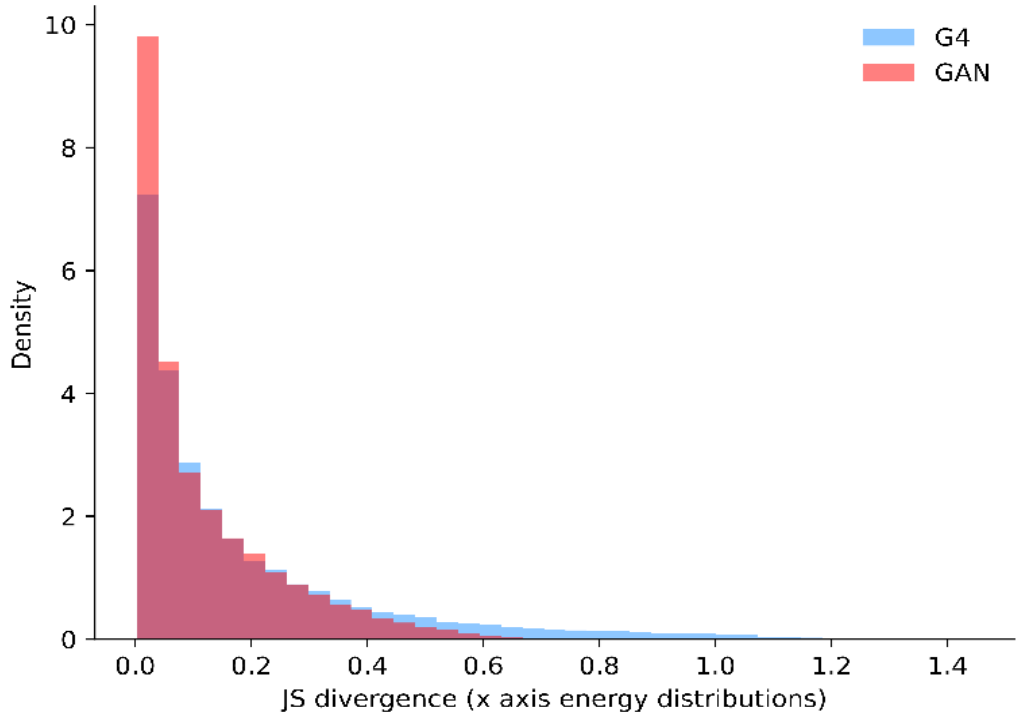
$$D_{JS}(P,Q) = \frac{1}{2} \cdot D_{KL}\left(P, \frac{P+Q}{2}\right) + \frac{1}{2} \cdot D_{KL}\left(Q, \frac{P+Q}{2}\right)$$

- Kullback-Leibler divergence (unnormalized)

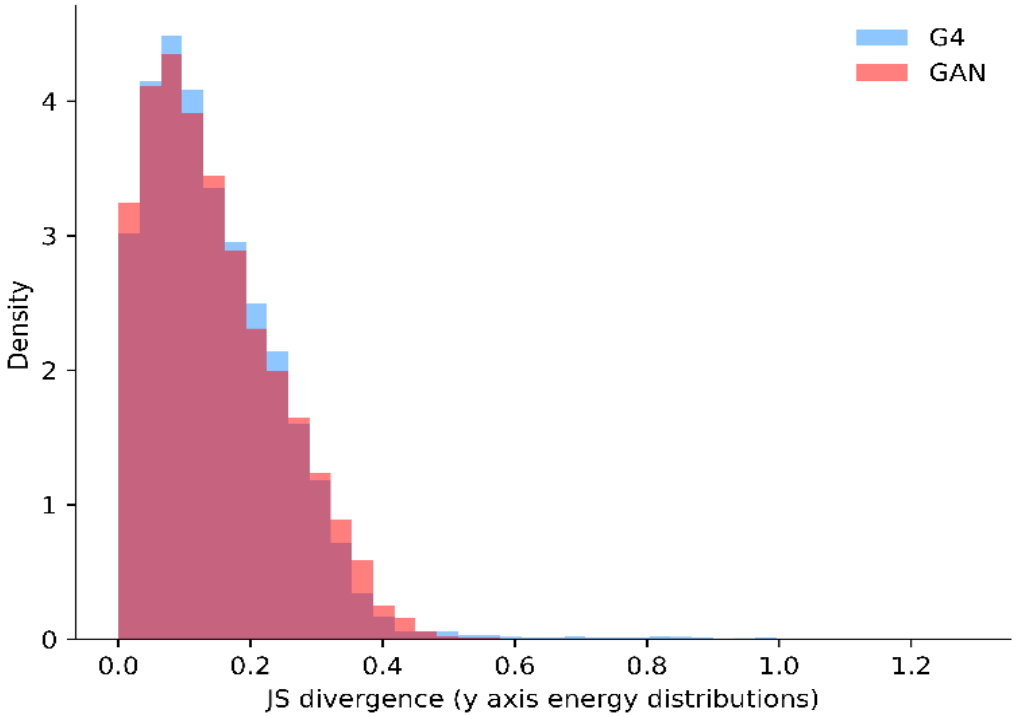
$$D_{KL}(P,Q) = P \cdot \ln\left(\frac{P}{Q}\right) - P + Q$$

# Shower shapes

## *Measuring the distance*



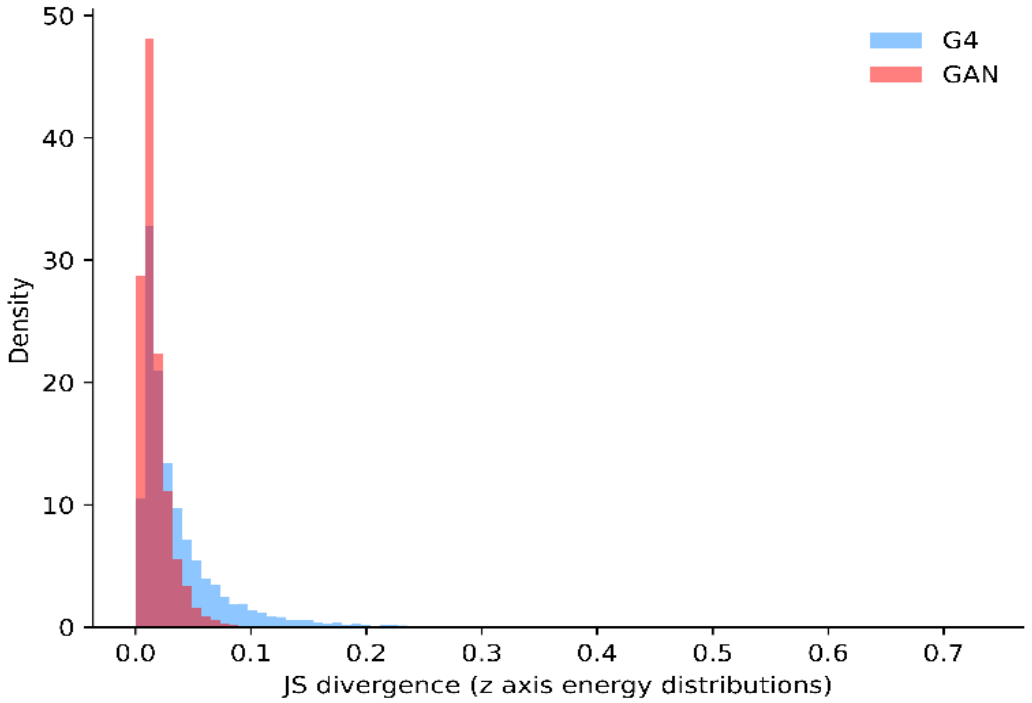

*x* axis




*y* axis

# Shower shapes

*Measuring the distance*




*z* axis

# SSIM

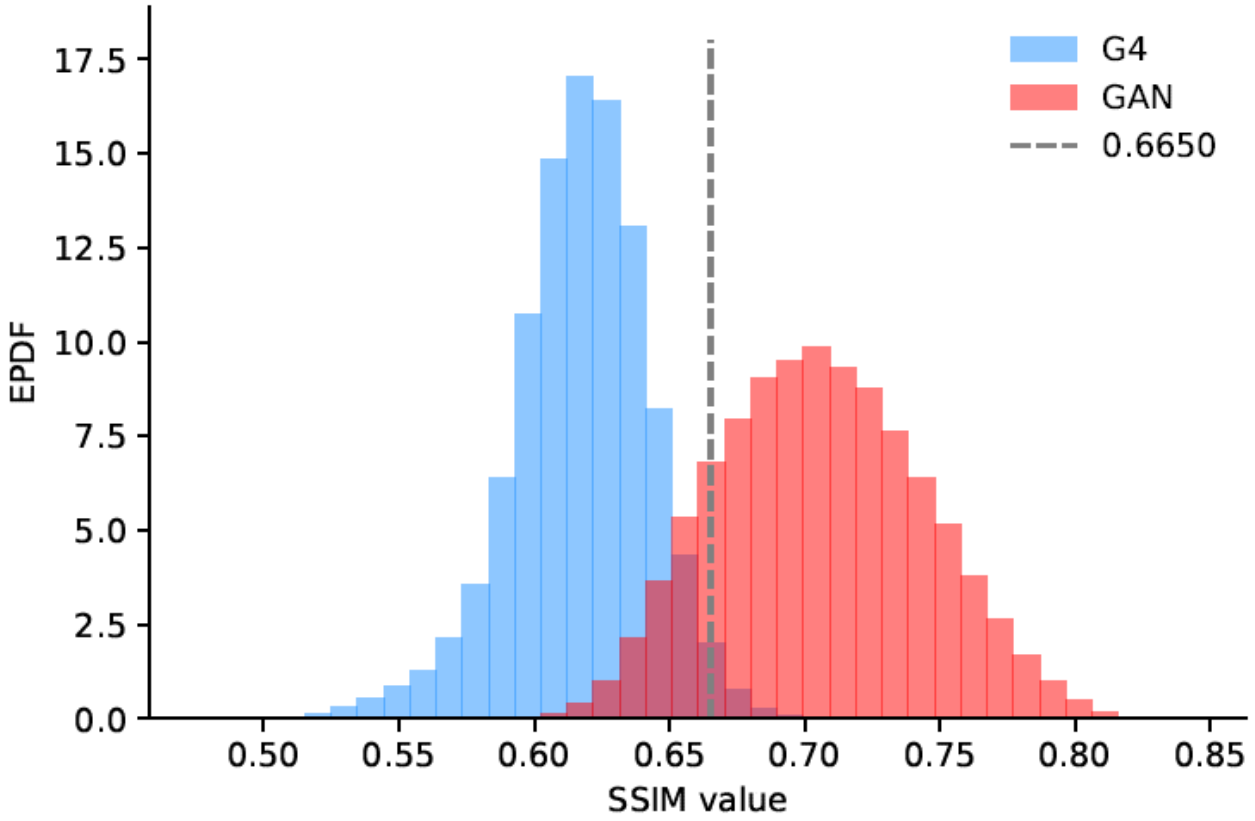
*Structural Similarity Index*

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

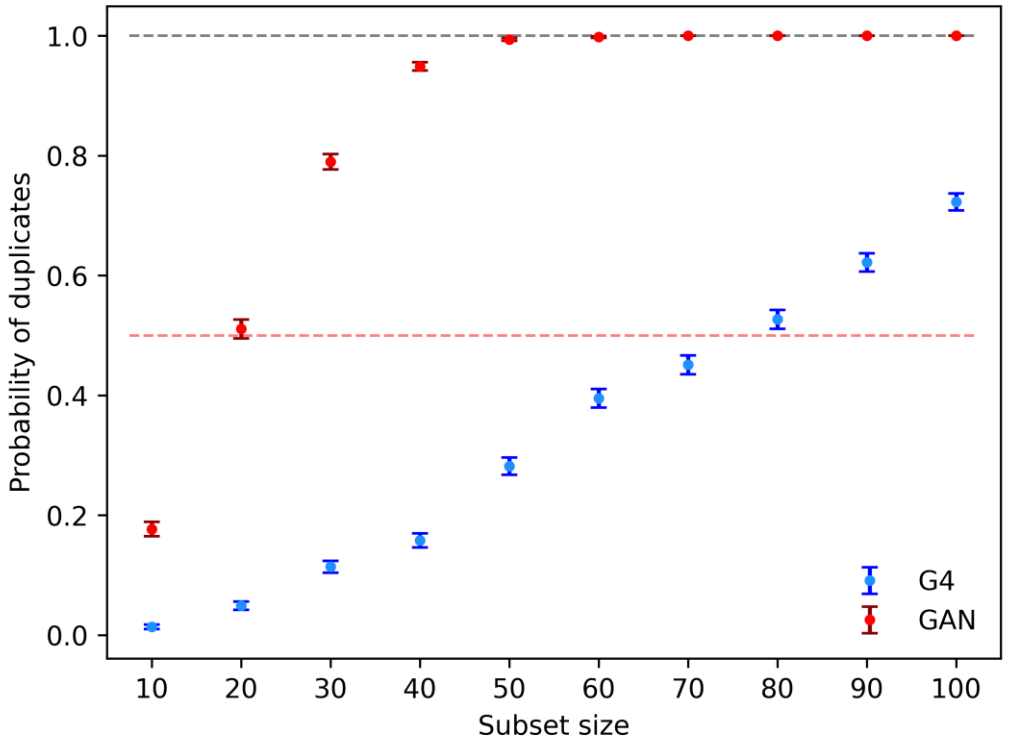
$$\mu_x = \sum_{i-1}^{N} w_i x_i \,, \sigma_x = \left( \sum_{i-1}^{N} w_i (x_i - \mu_x)^2 \right)^{\frac{1}{2}} \,, \sigma_{xy} = \sum_{i=1}^{N} w_i (x_i - \mu_x)(y_i - \mu_y)$$

$$\boldsymbol{w} = \{w_i | i = 1,2, \dots, N\}, w_i \sim \mathrm{N}(0,1.5), \sum_{i=1}^{N} w_i = 1$$

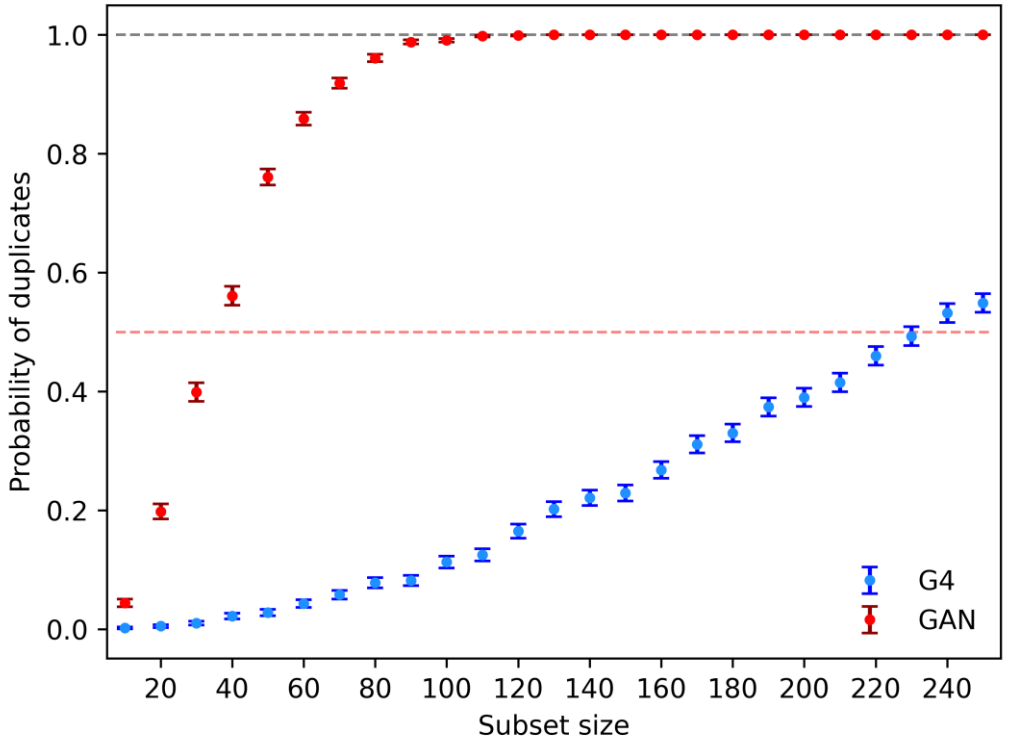
# SSIM


G4
GAN
0.6650
EPDF
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
SSIM value

# 0.02 vs. 0.05-quantiles




a) 0.05-quantile threshold

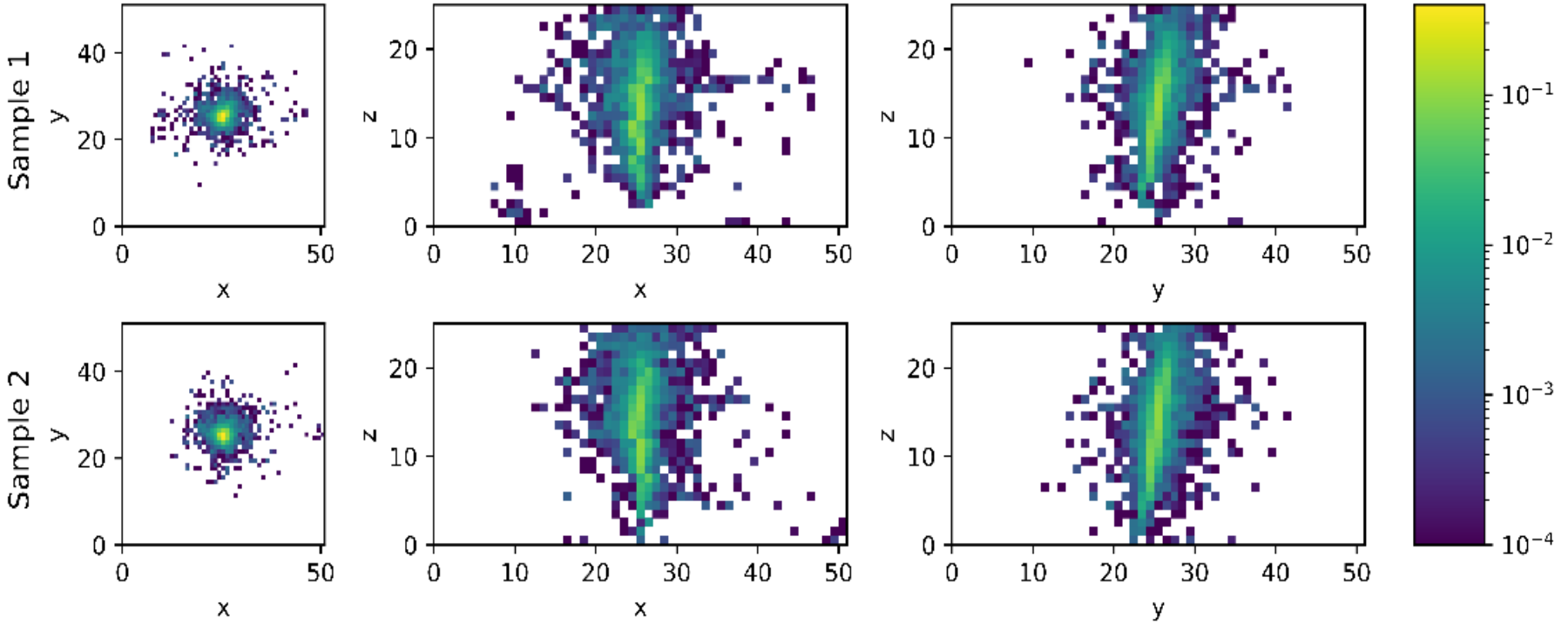



b) 0.02-quantile threshold

Probability of getting at least one duplicate in a set of '*subset size*' samples.

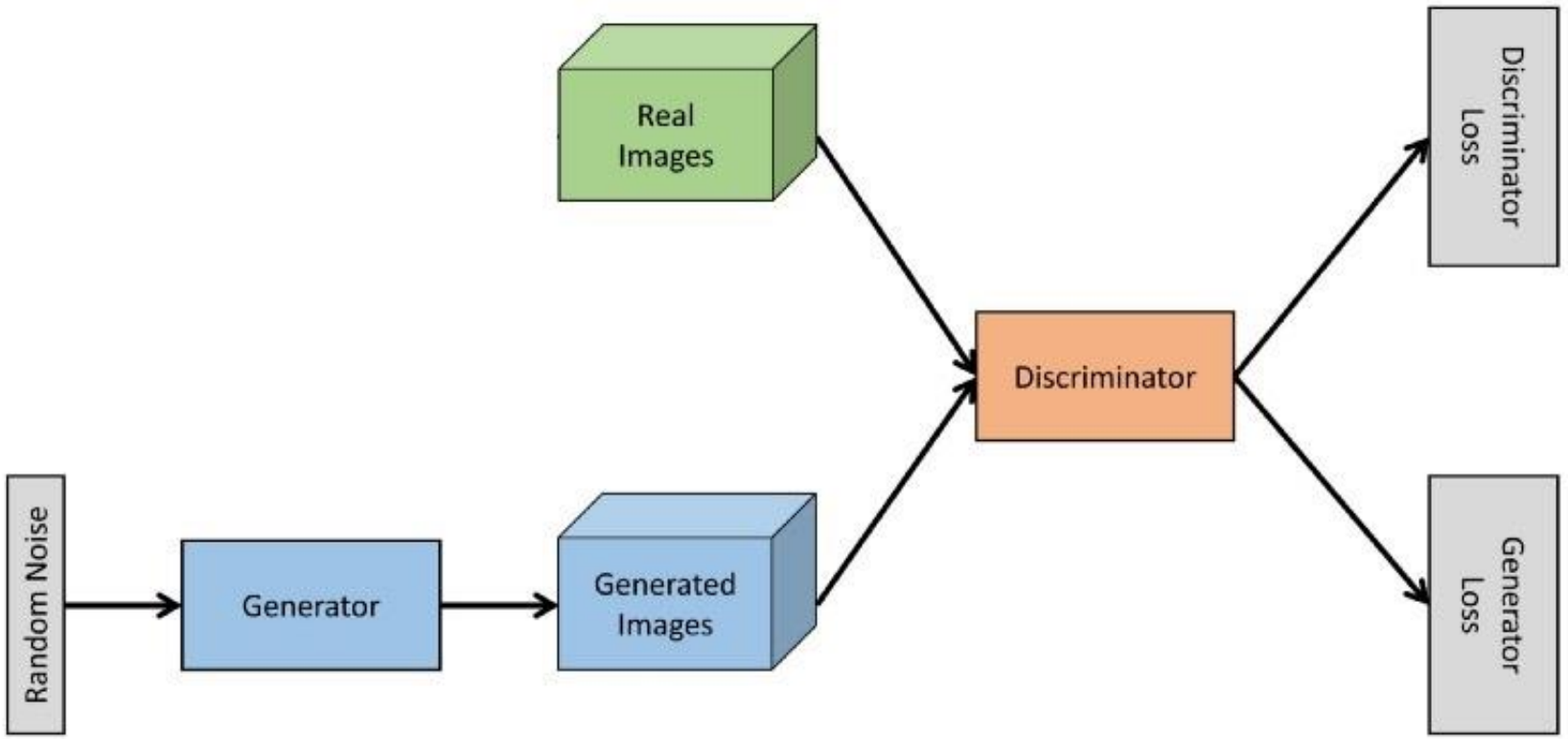
(1 000 replications)

# 3DGAN duplicates



# GAN

*Generative Adversarial Network*



# 3DGAN architecture