Amplifying Photon Showers with Generative Models

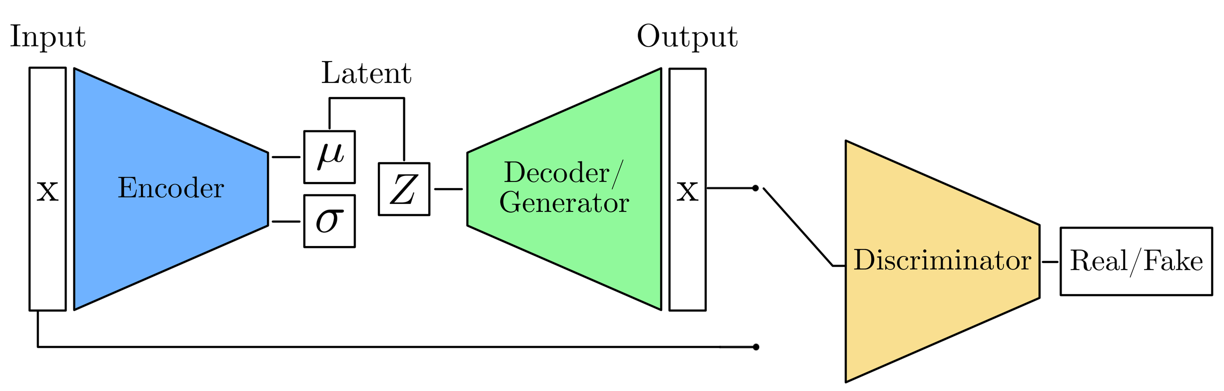
Sebastian Bieringer, Anja Butter, Sascha Diefenbacher, Engin Eren, Frank Gaede, **Daniel Hundhausen**,
Gregor Kasieczka, Benjamin Nachman, Tilman Plehn, and Mathias Trabs

DESY. | Universität Hamburg – DER FORSCHUNG | DER LEHRE | DER BILDUNG | DASHH Data Science in Hamburg HELMHOLTZ Graduate School for the Structure of Matter | CLUSTER OF EXCELLENCE QUANTUM UNIVERSE

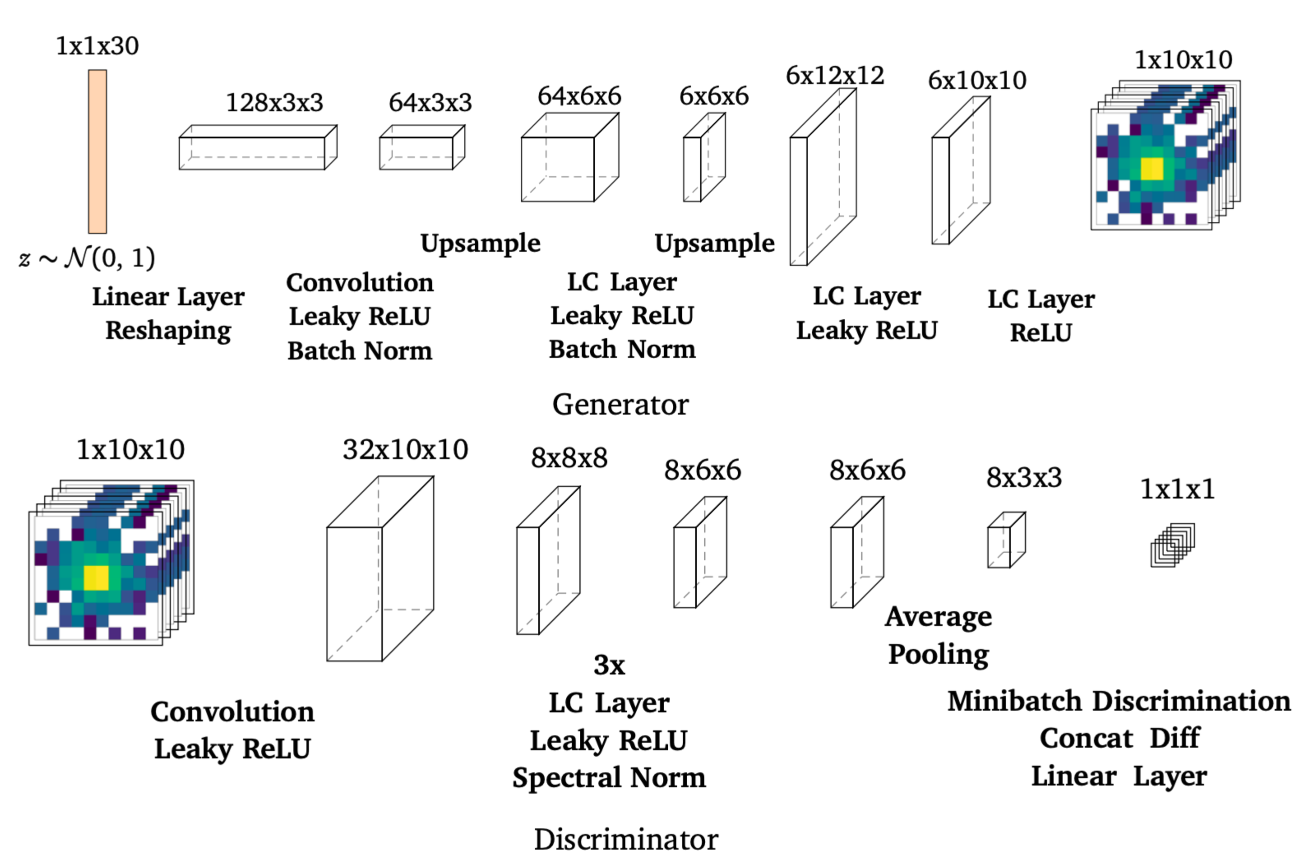
Introduction

- Deep generative neural networks (NNs) can accelerate and augment slow Monte Carlo detector simulators
- How many more events can a NN generate before being limited by training statistics?
- Analyze amplification of calorimeter images of photon showers in terms of their kinematic distributions
- Full paper on arXiv (2202.07352)

Generative Model



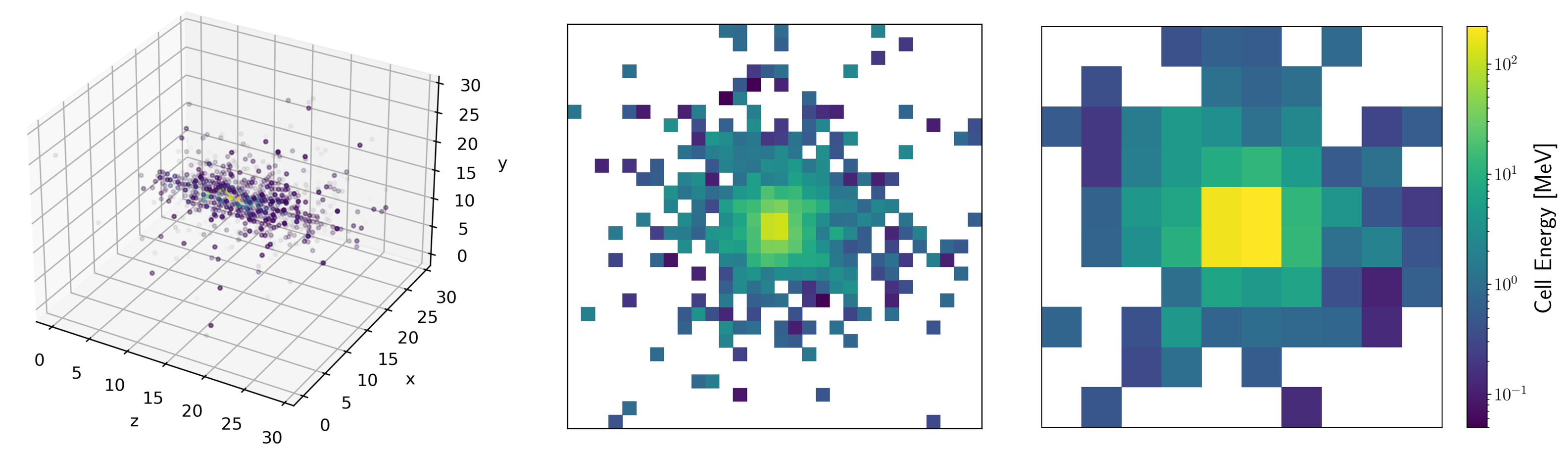
- Model: VAE-GAN using GAN-like Discriminator in place of VAE's element-wise reconstruction loss
- Train on set of only 1000 images
- Leverage ensembles: train 3 models on same training set
- Select the epoch with the best agreement between the generated and training distributions averaged over five kinematic observables



- Use locally connected layers to account for missing translational invariance
- Apply label smoothing to prevent vanishing gradients
- Utilize mini-batch discrimination to achieve better generalization

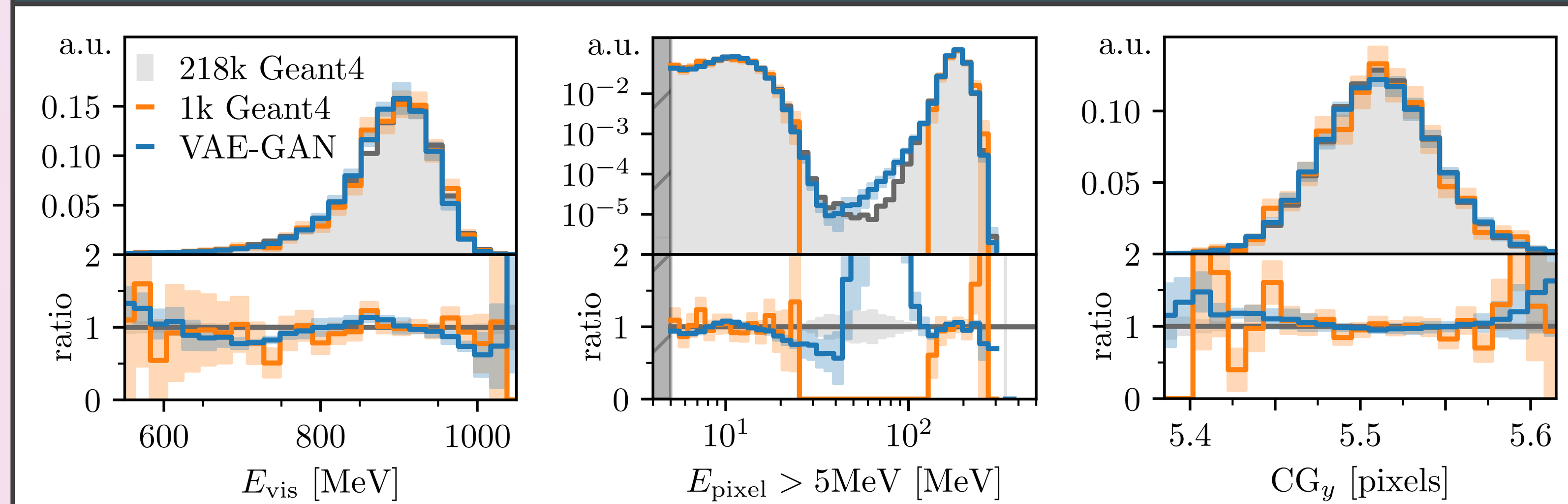
Dataset

- International Large Detector (ILD) electromagnetic calorimeter: 269k photon showers generated with GEANT4
- 50 GeV photons at perpendicular incident angle
- Dimensionality reduction for simplified training

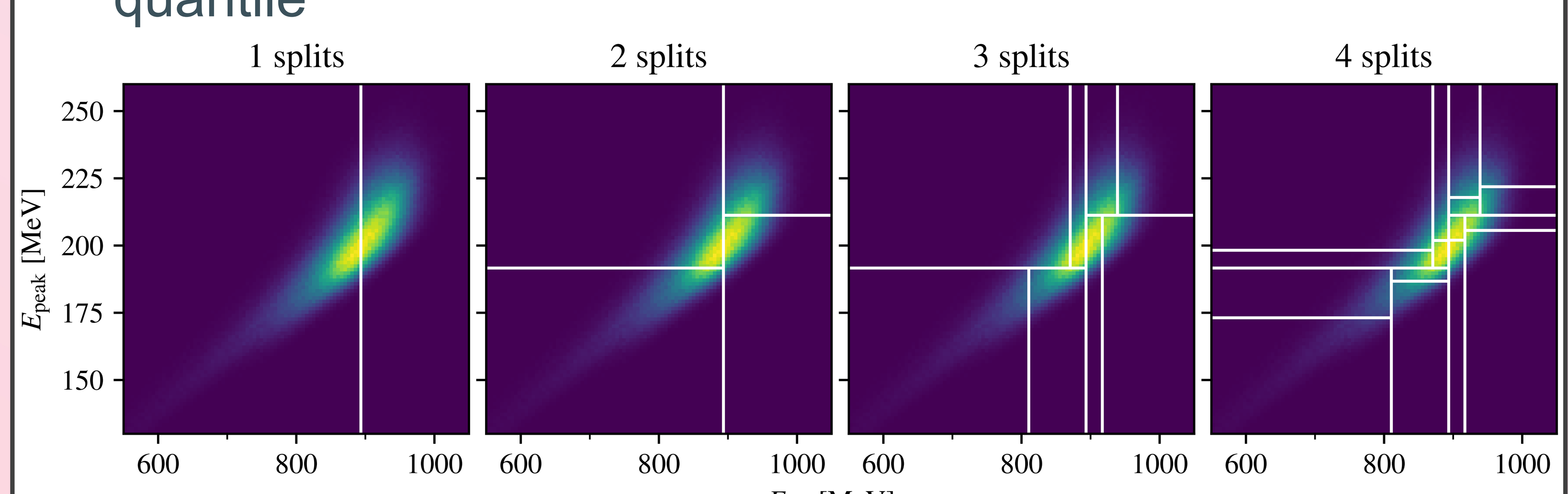


Transformation: Projection to 10x10 + Minimal Ionizing Particle cut (0.1 MeV)

Results

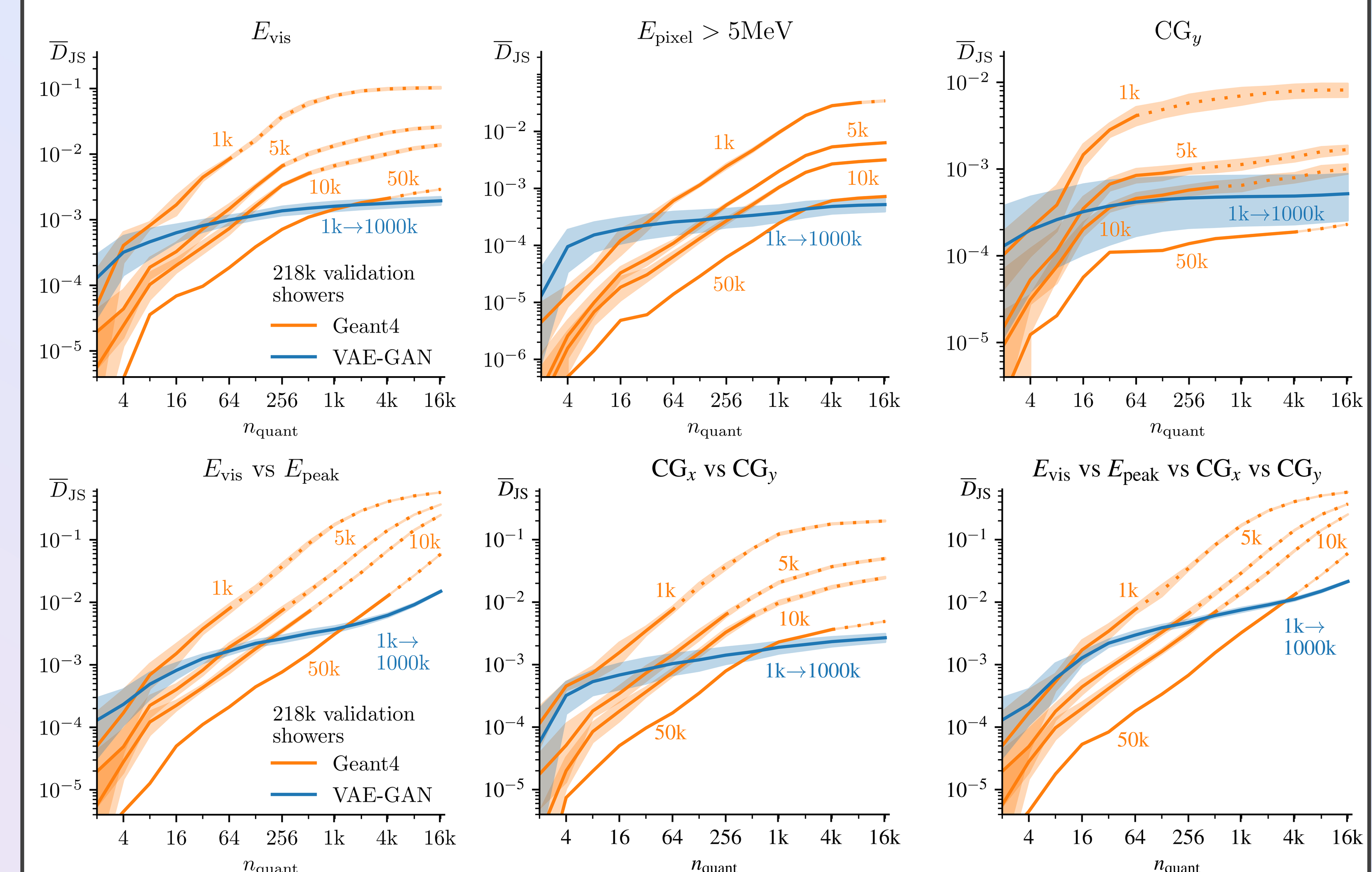


- Truth estimate: 218k GEANT4 shower samples
- VAE-GAN sample: 1M generated showers
- In *visible per-pixel energy*: VAE-GAN interpolates into the sparsely populated interval (2-120 MeV) even though the training set does not include a single pixel in this range
- Evaluate distributions of kinematic observables using quantiles Q_i of equal probability
- Avoid sparse quantiles by requiring at least 10 points per quantile

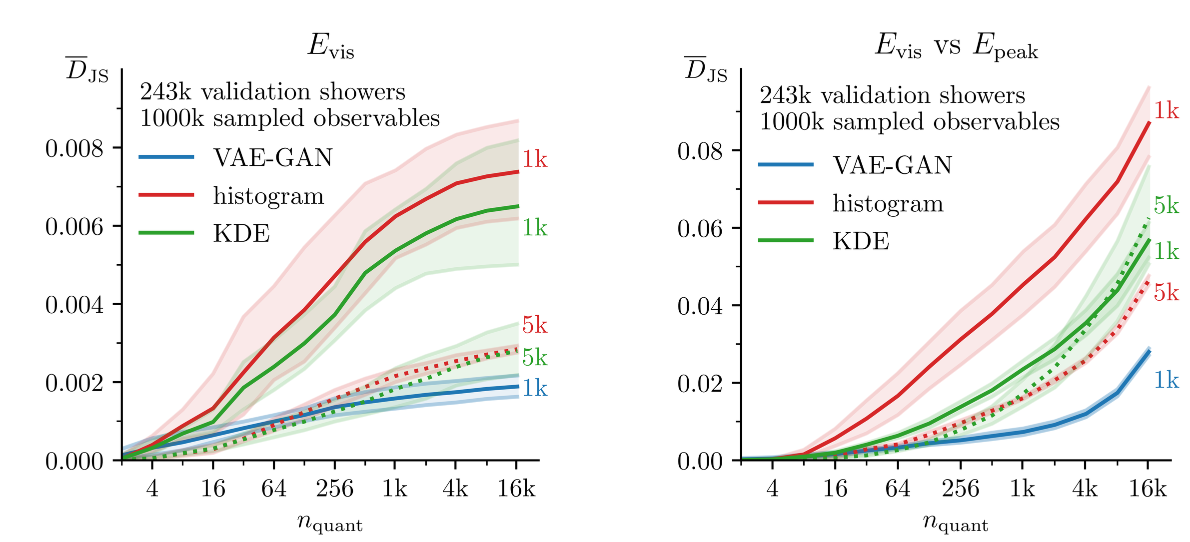


- Measure the similarity between two distributions defined by samples via the Jensen-Shannon divergence

$$\overline{D}_{\text{JS}}(g,p) = \frac{1}{2}\sum_{Q_i\in\mathbf{Q}}\left(g_i \log\frac{g_i}{\frac{1}{2}(g_i+p_i)} + p_i \log\frac{p_i}{\frac{1}{2}(g_i+p_i)}\right)$$



- At low numbers of quantiles: density estimation limited by training statistics, comparable performance
- At high number of quantiles: similarity of model generated sample comparable to up to 50k GEANT4 samples
- Similar behavior in multiple dimensions



- Classical density estimation techniques are
  - limited in functional shape, tend to overfit and
  - suffer drastically from *curse of dimensionality*

Conclusion

- Generative ML-models can
  - amplify scaling behavior of calorimeter image samples
  - be used to generate showers beyond limited training statistics to accelerate simulation