CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Amplifying Statistics with Ensembles of Generative Models

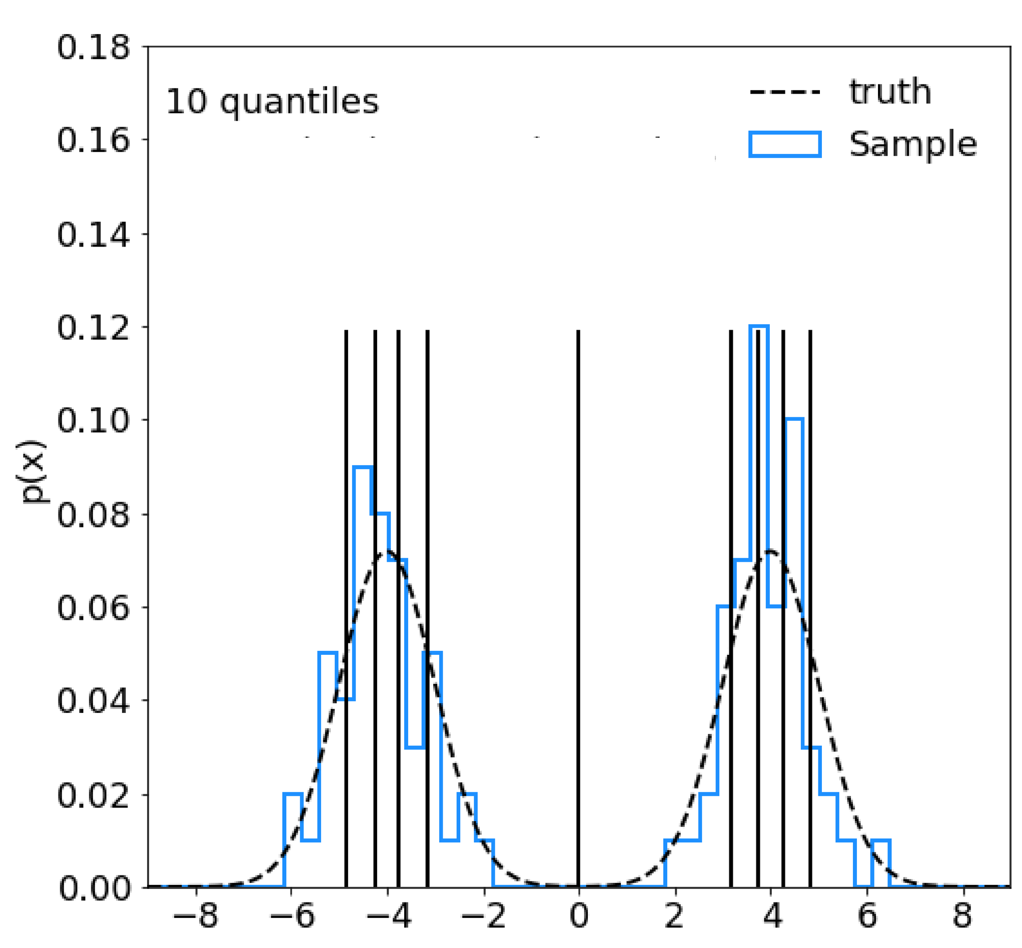
Sascha Diefenbacher, Anja Butter (Heidelberg), Gregor Kasieczka, Ben Nachman (Berkeley), Tilman Plehn (Heidelberg)

Introduction

- Deep generative models are used to accelerate or augment slow physics simulators.
- If you train a generator using N examples and produce M, what is the statistical power of the M examples?
- Test this using simplified dataset
- For more information see the full paper on arxiv

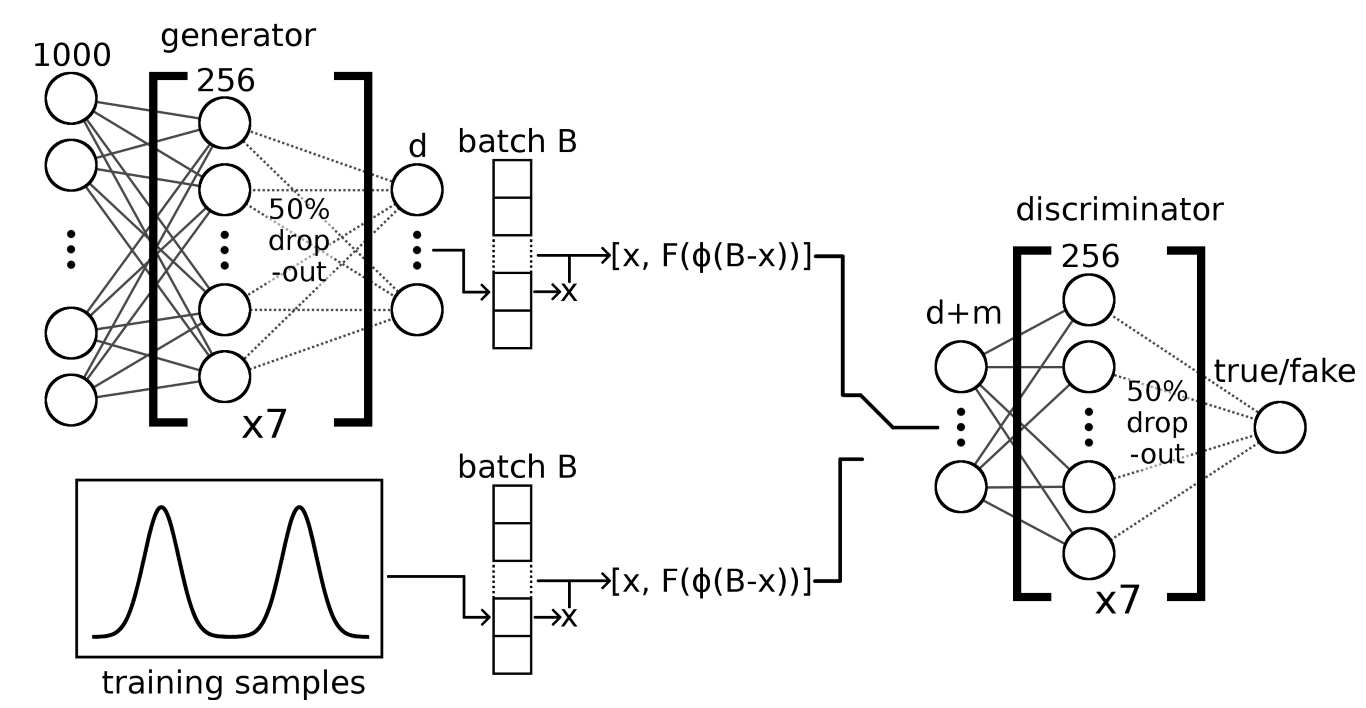
Dataset

- Camel back function
- Smooth multimodal distribution, common in physics
- Define K quantiles for comparison
  - Intervals that each contains equal probability



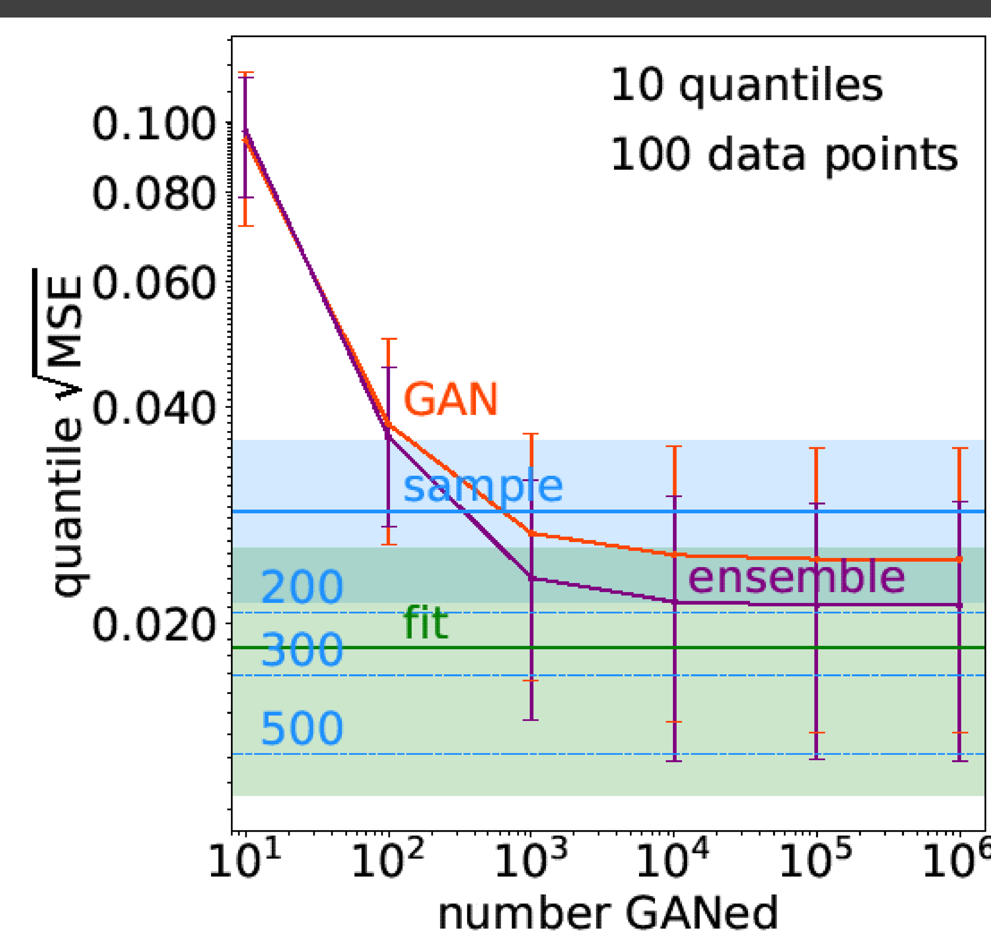
- Training sample: 100 camel back points
- Count points per quantile
- Compare quantile fractions to true values
- True fractions given by
- Baseline for comparison $\frac{1}{K_{\text{quant}}}$

Generative Model



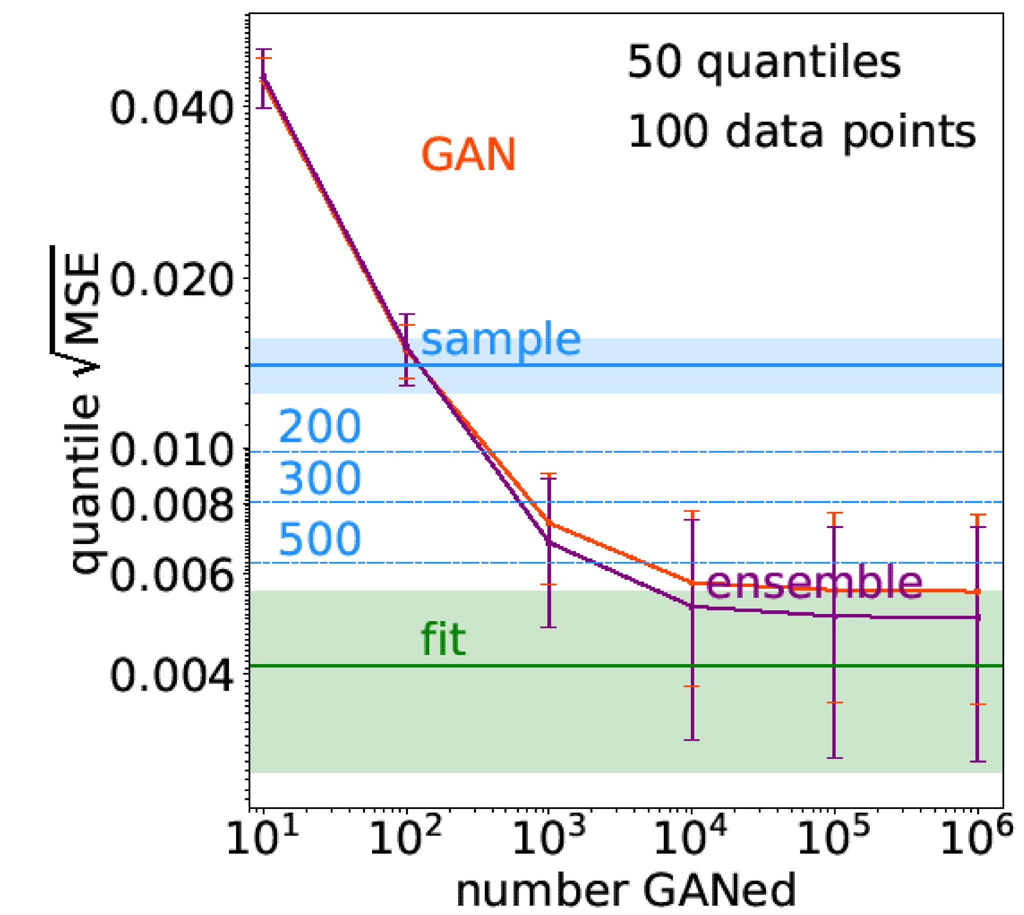
- Trained on 100 data points from training samples
  - Use regularization methods against overfitting (dropout, training noise, batch-statistics)
- Ensemble approach: train 10 independent GANs on exactly identical training data
- Calculate quantile fraction from GANed points
- Quantile MSE: $\frac{1}{K_{\text{quant}}}\sum_{j=1}^{K_{\text{quant}}}\left(x_j-\frac{1}{K_{\text{quant}}}\right)^2$

Results



- 50 training samples, 50 fits, 50 GANs
- 10,000 GANed points equal to 150 training points
- Interpretation in terms of information:
  - Sample: only data points
  - Fit: data + true function
  - GAN/ensemble: data + smooth function

- Assumption about smoothness adds information
- GAN can interpolate between points
- Interpolation allows for amplification
- Ensemble smooths out overfitting effects
- Better interpolation leading to better amplification
- More quantiles ➔ less points per quantile



- GAN interpolation more impactful for sparse data
- High dimensional data often sparse
- Promising approach for higher dimensions

Conclusion

- It makes sense to GAN significantly more events than we have in the training sample,
- Individual events carry less information than a training sample event.
- Net benefit, if the GAN sampling is sufficiently fast

Acknowledgments

We further thank Ramon Winterhalder for helpful input during the early phases of the project. The research of AB and TP is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant 396021762 -- TRR~257 *Particle Physics Phenomenology after the Higgs Discovery*. GK and SD acknowledge support by the DFG under Germany's Excellence Strategy – EXC 2121 *Quantum Universe – 390833306*. BN is supported by the U.S.~Department of Energy, Office of Science under contract DE-AC02-05CH11231.