GANplifying Event Samples

Anja Butter, **S. D.**, Gregor Kasieczka,
Benjamin Nachman, Tilman Plehn

based on 2008.06545

IDT-UM Meeting

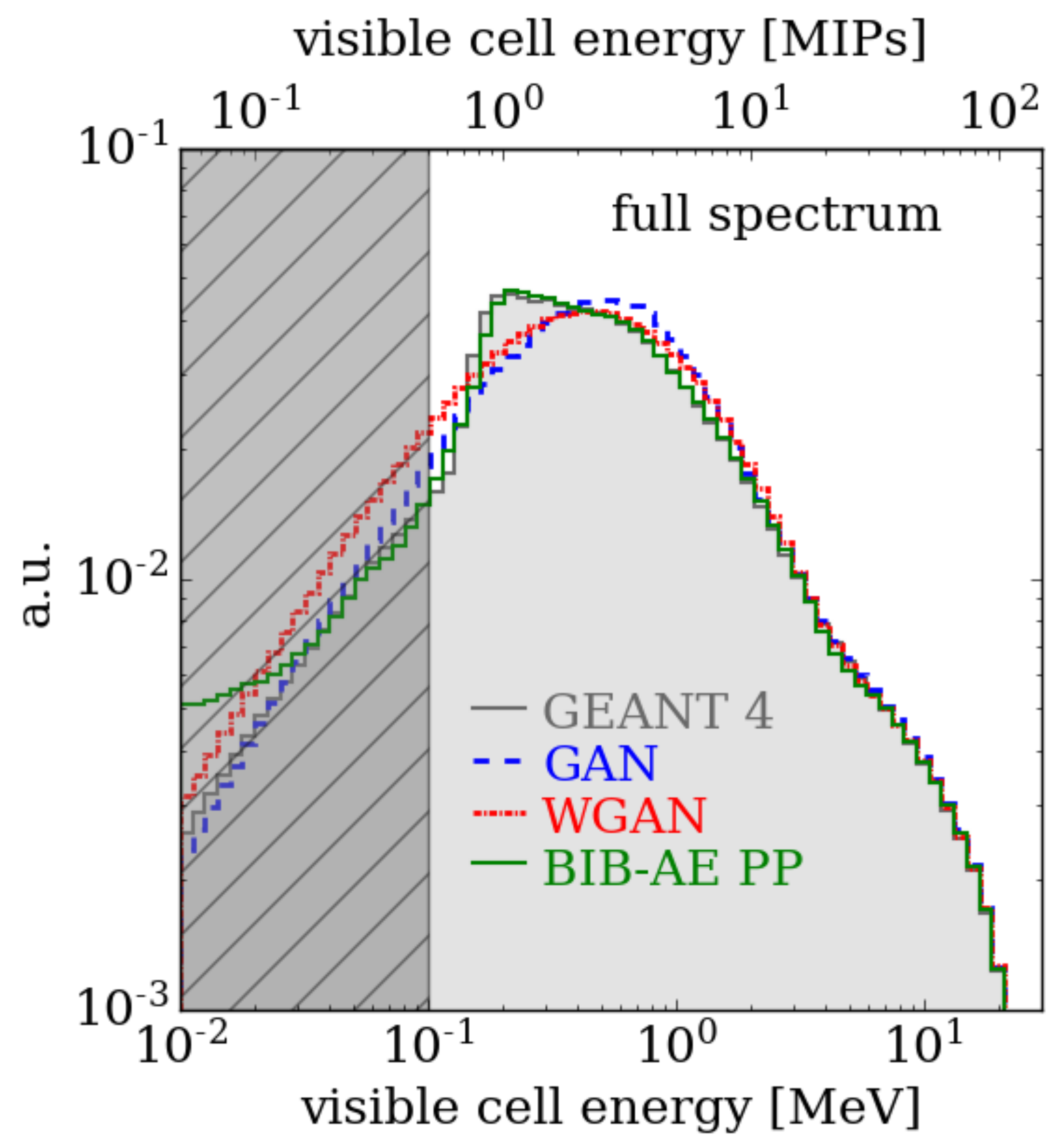
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

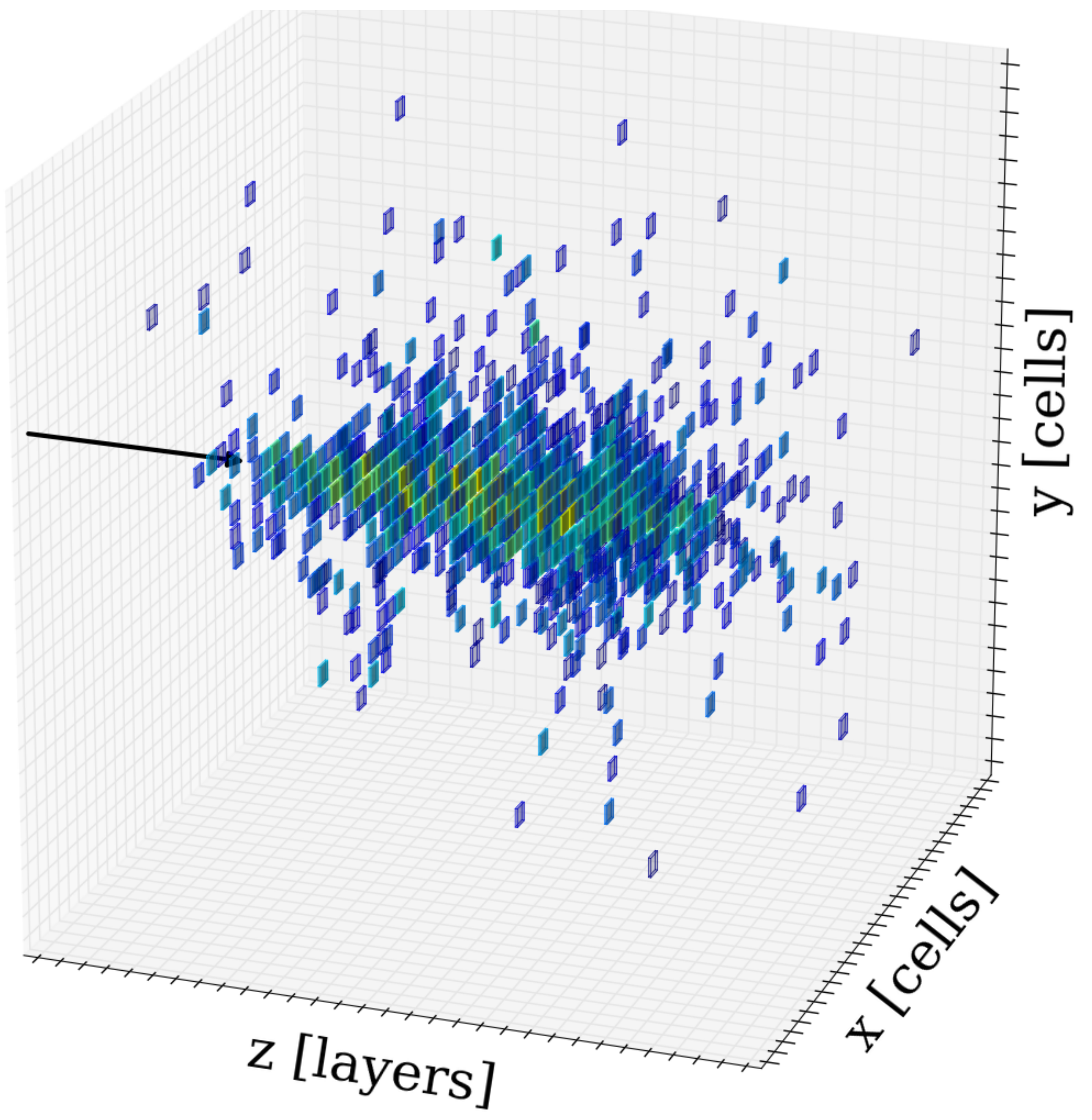
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Introduction

- Generative machine learning models are increasingly common in physics
- Most commonly **G**enerative **A**dversarial **N**etworks (GANs)
- Applied to:
  - Event generation
  - Calorimeter simulation
  - Cosmology
  - Environmental physics





Introduction

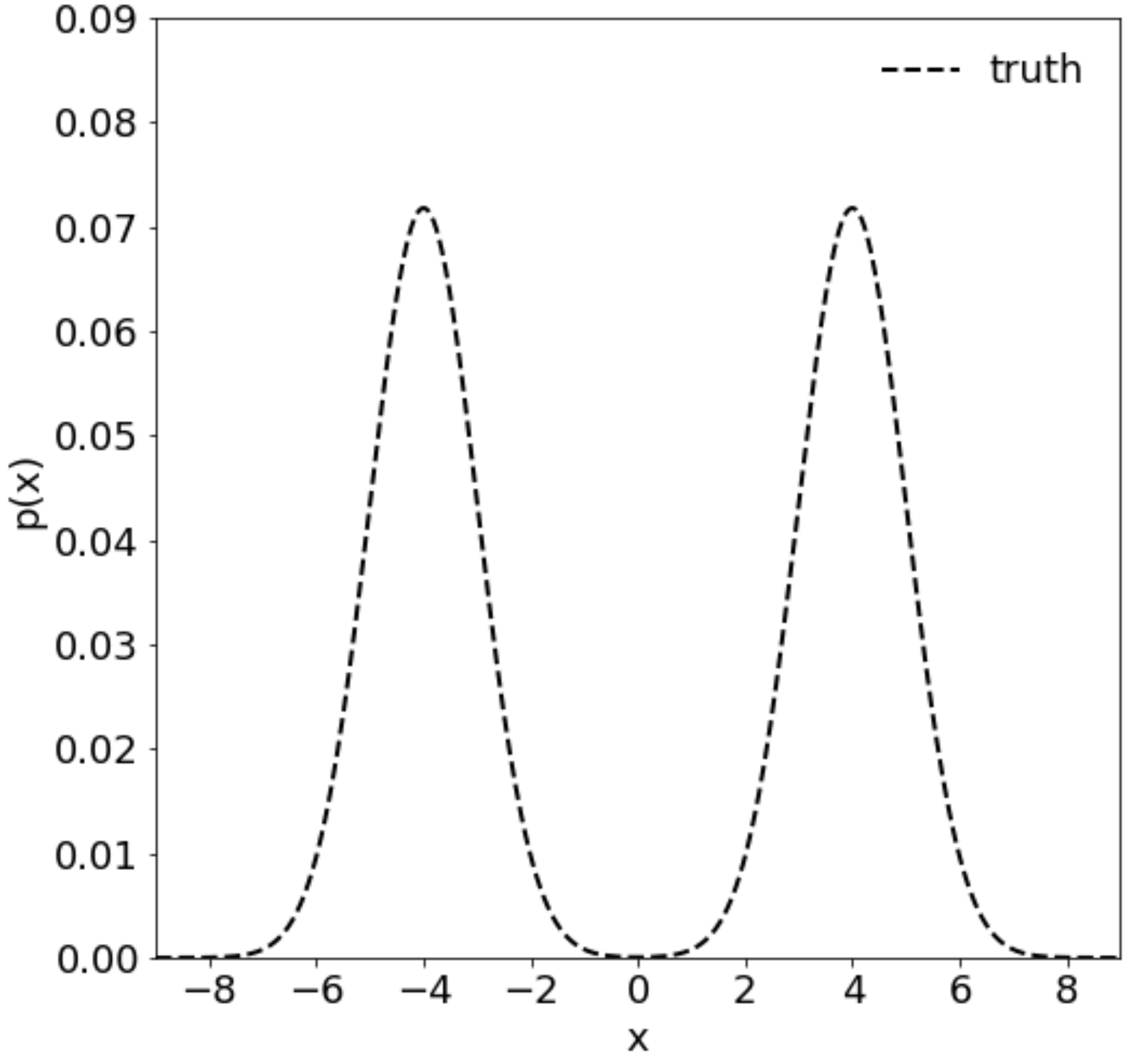
- Potential problem
- If a GAN is trained on N data points, how many new points can I draw from the GAN?
- Standard assumption: no more than N new points
- Is that really true?

→ Run tests using toy example

1-D Toy Model

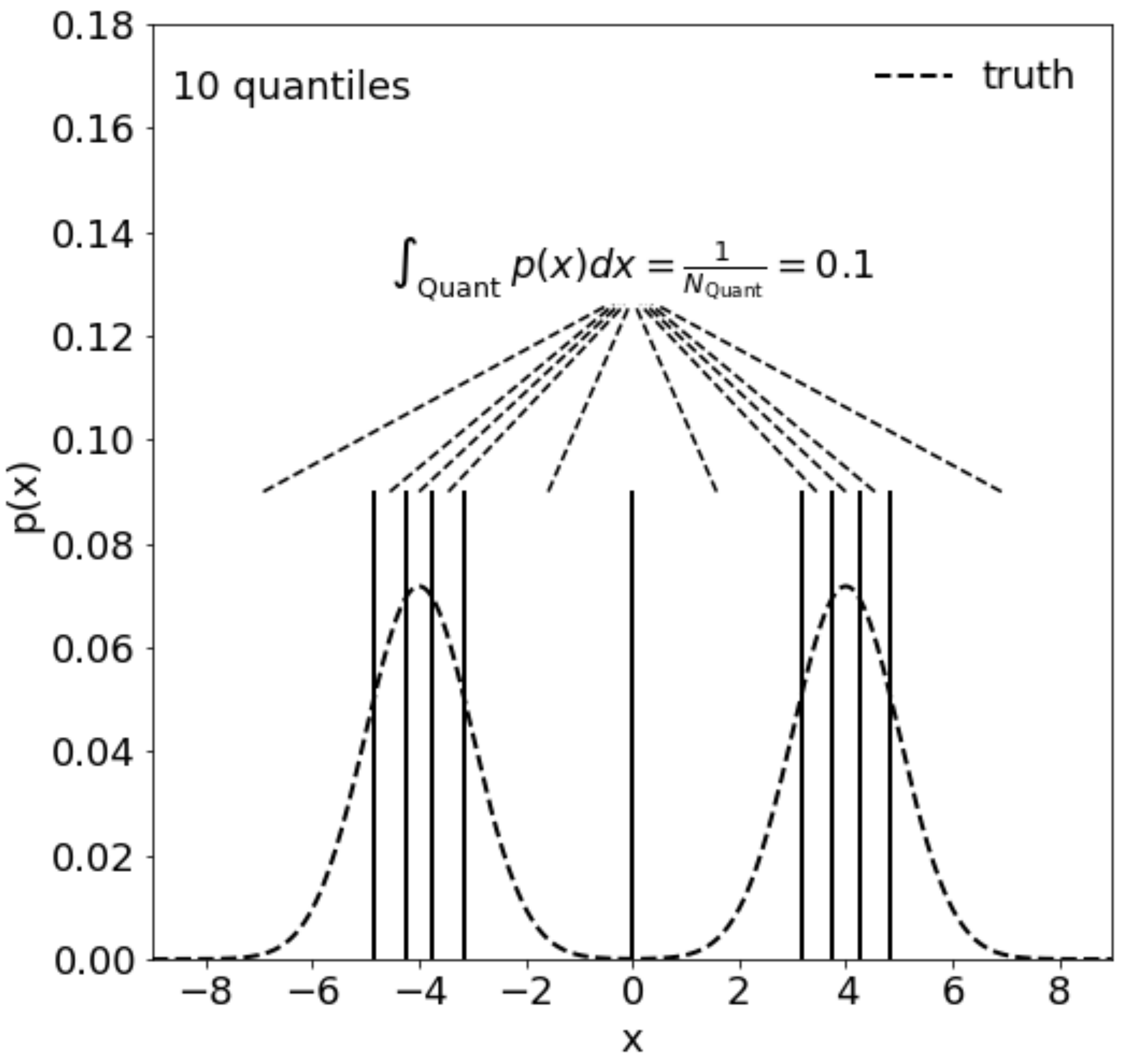
- Camel back function: double peak Gaussian

$$p(X) = \frac{1}{2}(N_{-4,1}(x) + N_{4,1}(x))$$



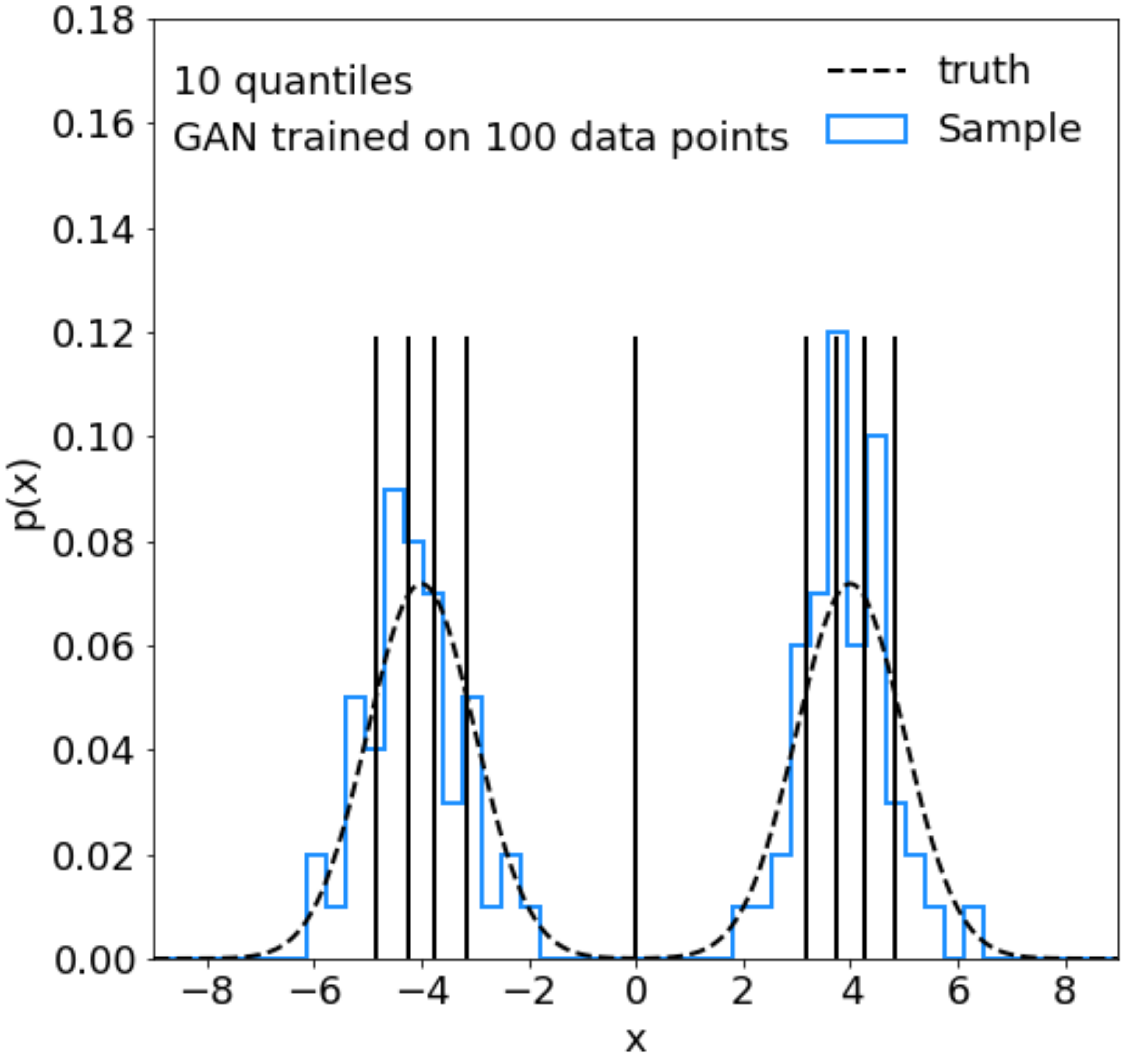
Quantiles

- Measurement how well function is described
- Define N quantiles on true distribution
- Each quantile contains equal probability



Training Sample

- Draw 100 points from true camel back distribution
- This is designated as the (training) sample
- Calculate fraction of points in each quantile

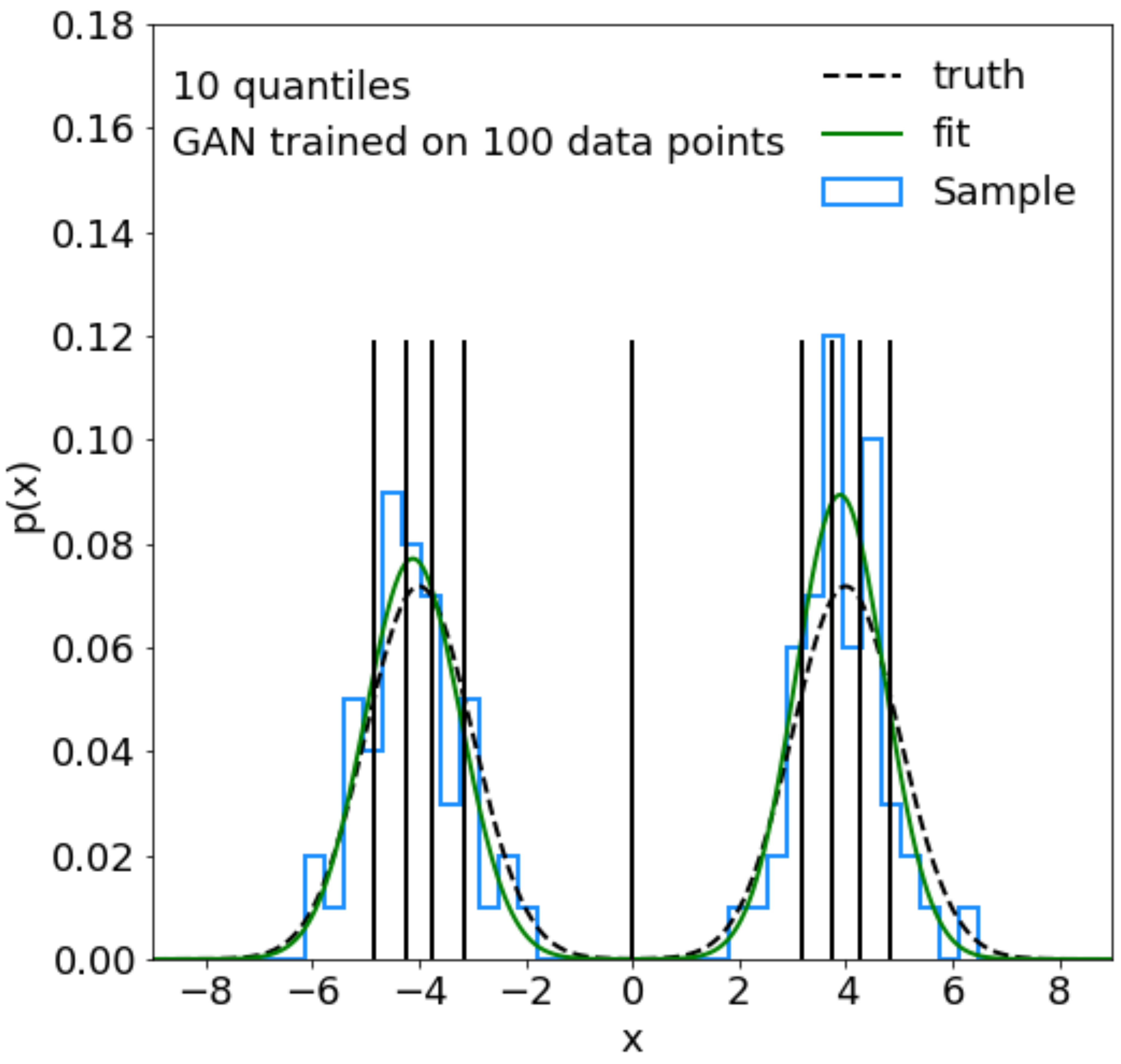


Parameter Fit

- Fit 5 parameter camel back function to training samples

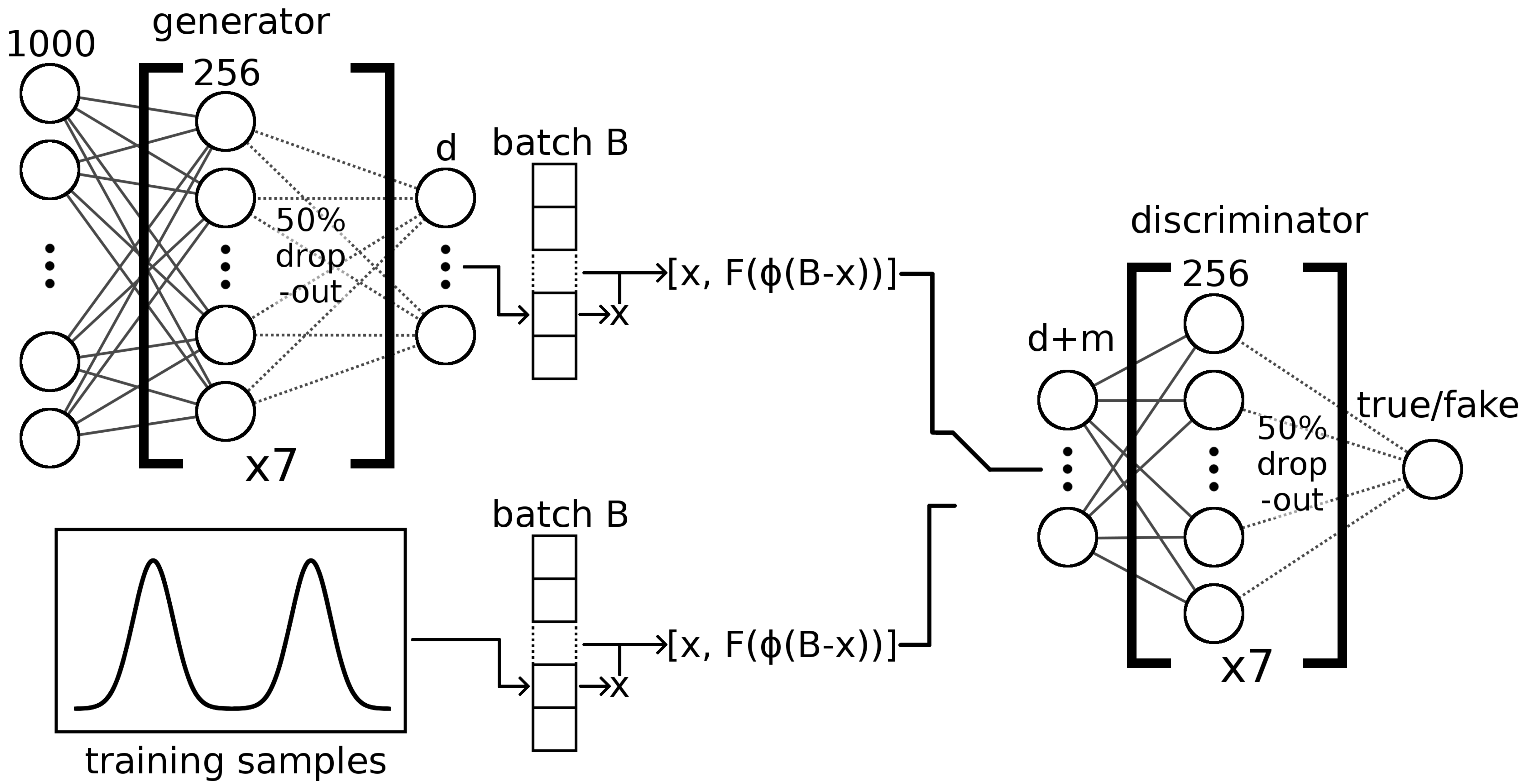
$$p(X) = a\, N_{\mu_1,\sigma_1}(x) + (1-a) N_{\mu_2,\sigma_2}(x)$$

- Analytically calculate integral for each quantile
- Gives upper performance benchmark



Generative Network

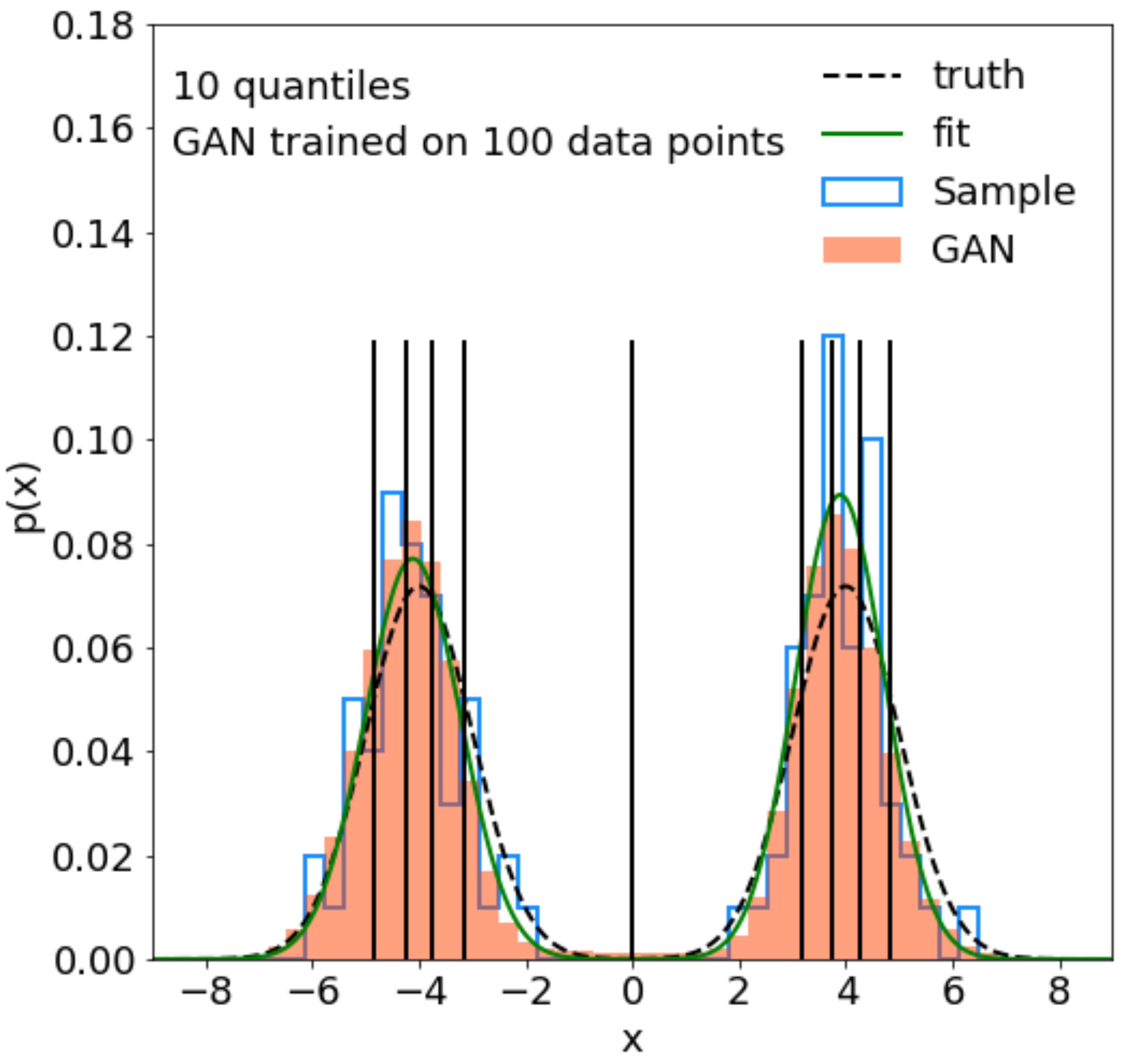
- Train GAN on 100 data points from training sample
- Mode-collapse and overfitting problematic
  - Dropout
  - Added training noise
  - Batch-statistics



Generative Network

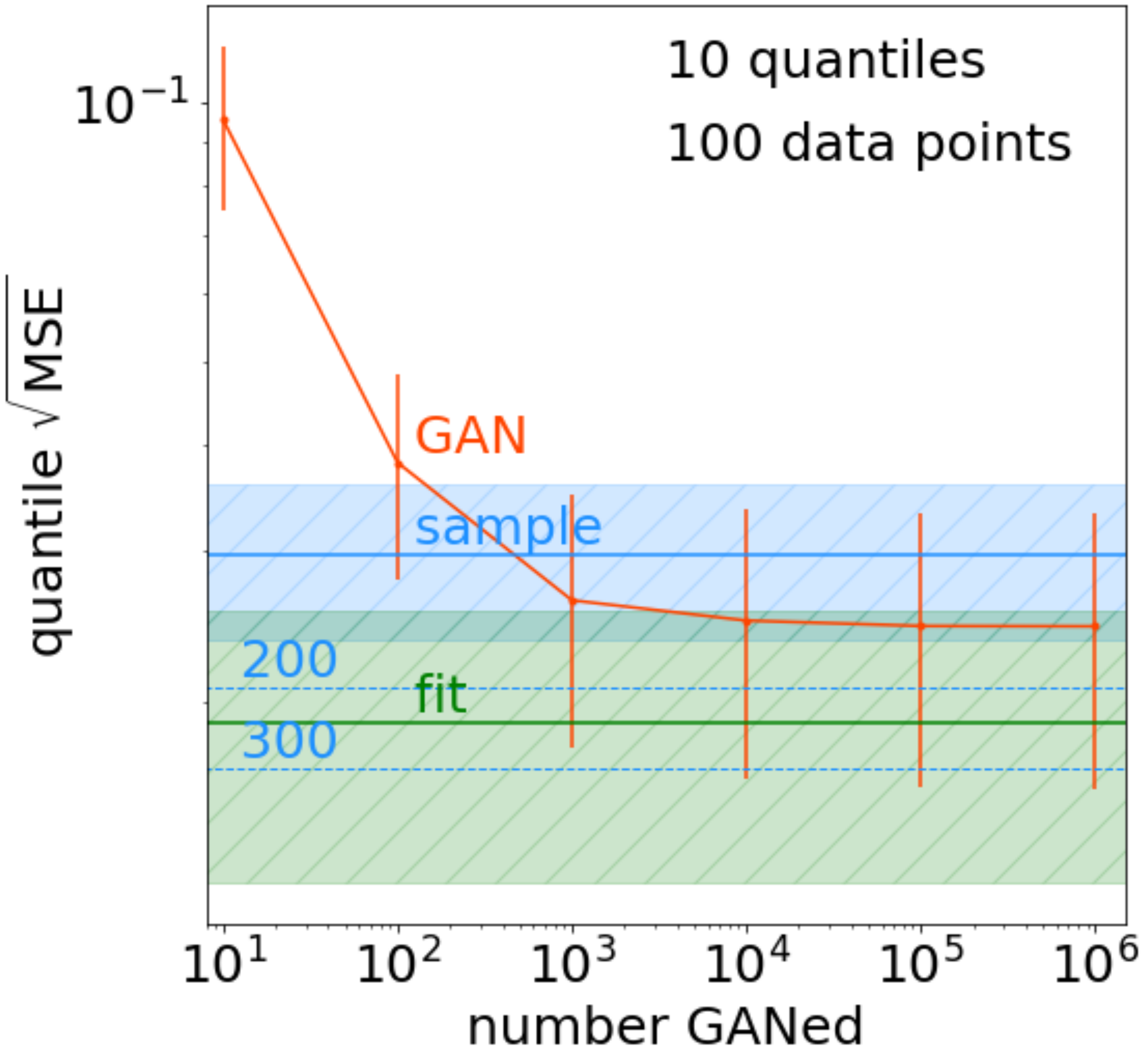
- Generate $O(10^7)$ data points using GAN
- Calculate fraction of points in each quantile
- Define quantile MSE:

$$\mathrm{MSE} = \frac{1}{N_{\mathrm{quant}}} \sum_{j=1}^{N_{\mathrm{quant}}} \left(x_j - \frac{1}{N_{\mathrm{quant}}} \right)^2$$



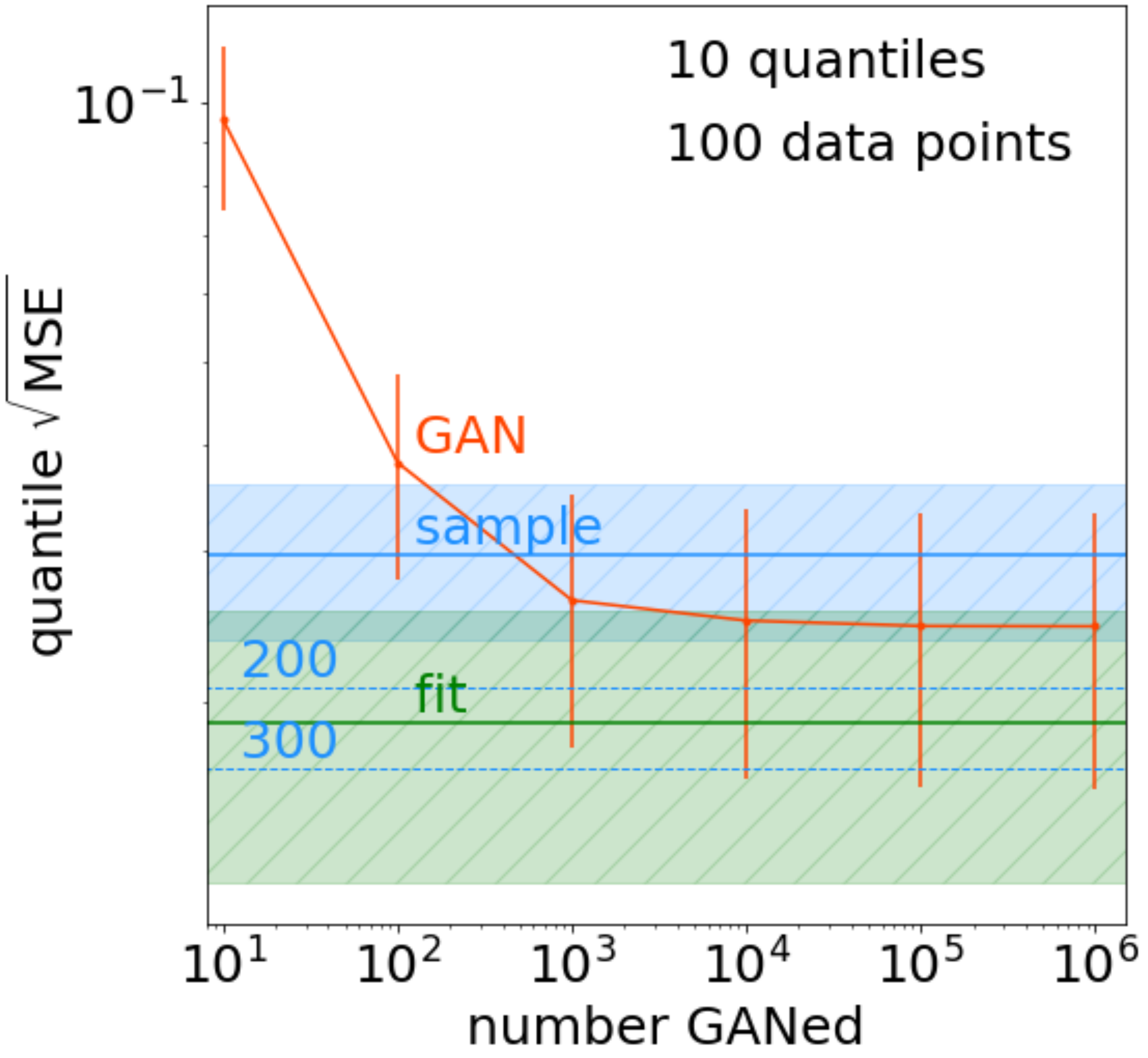
Generative Network

- For 100 training samples, 100 fits and 100 GANs compare MSE
- GAN describes distribution better than training data
- Needs 10,000 GANed points to match 150 true points
- Shifts statistical uncertainty to systematic uncertainty



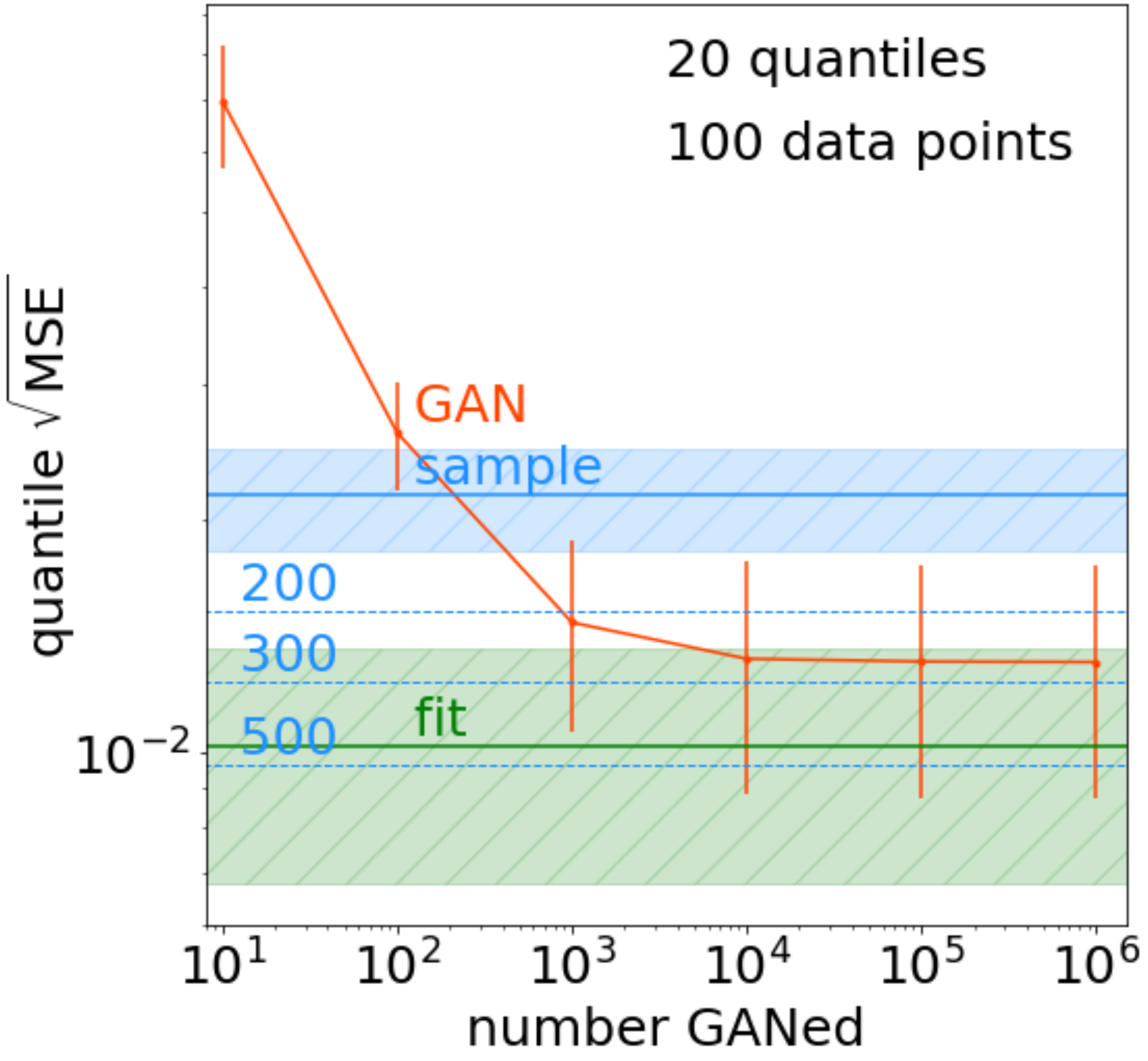
Generative Network

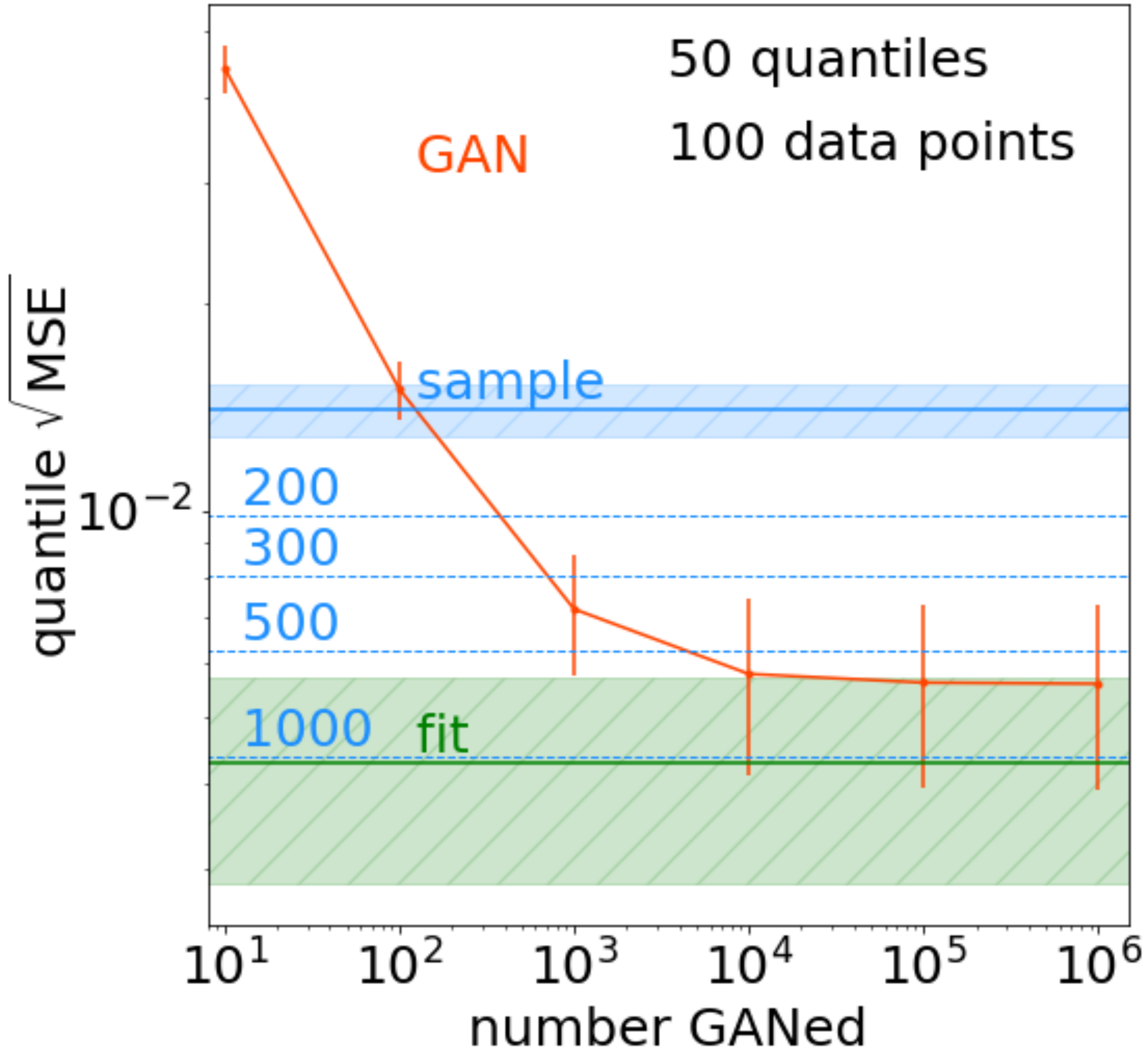
- How is this possible?
- In terms of information:
  - sample: only data points
  - fit: data + true function
  - GAN: data + smooth, continuous function
- This allows the GAN to interpolate



Generative Network

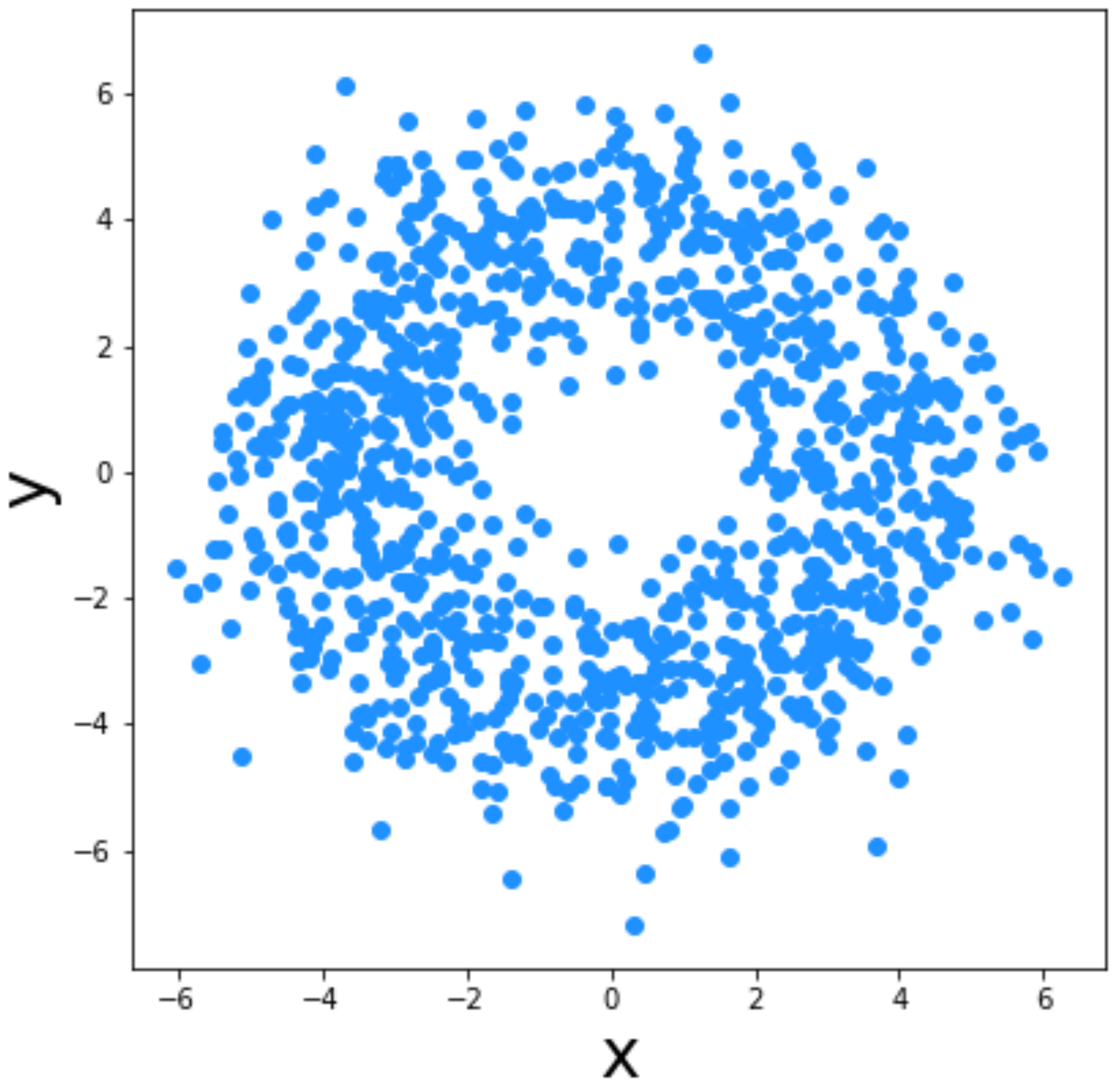
- Interpolation more noticeable for sparser data





2-D Toy Model

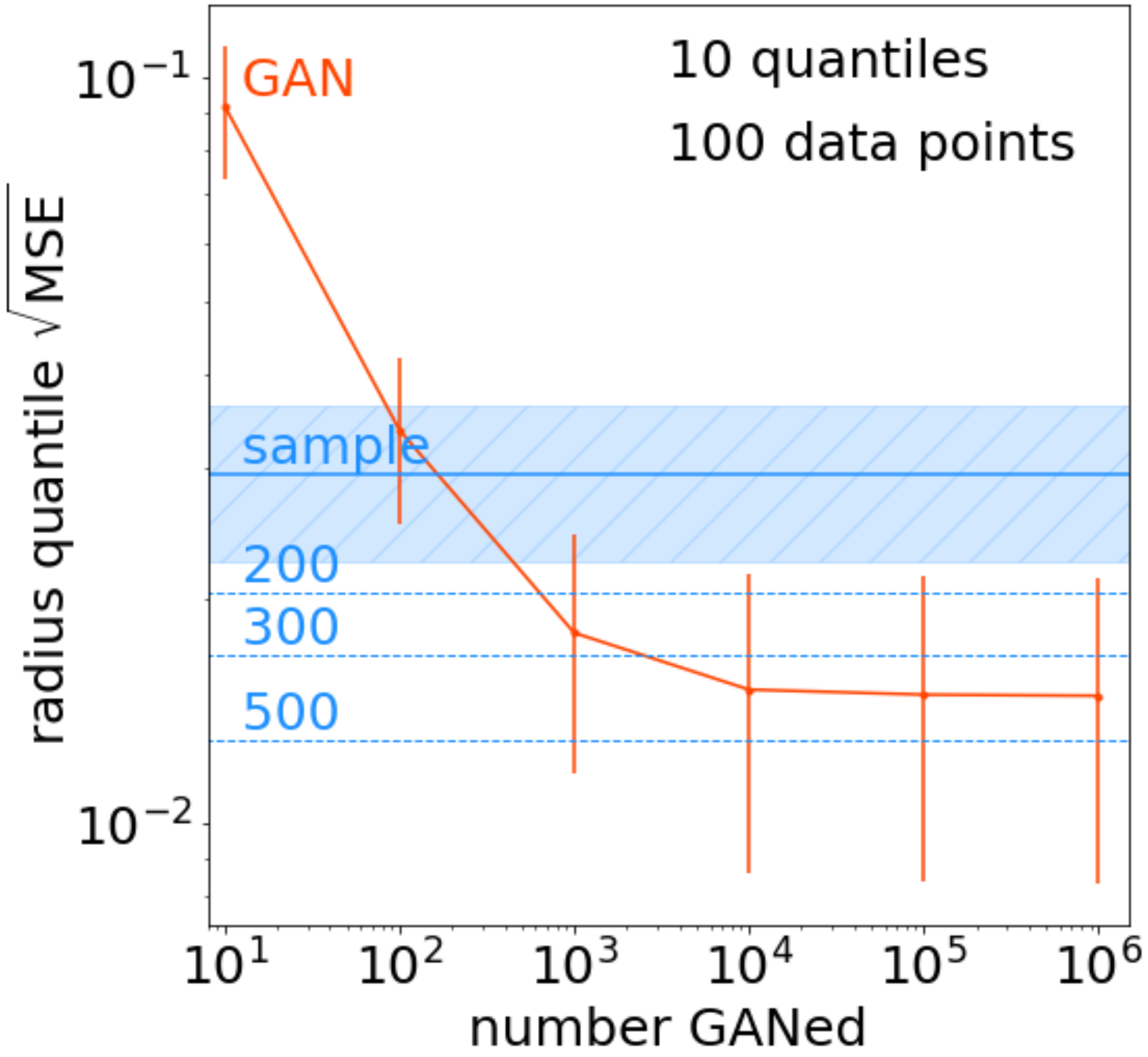
- Extend setup into two dimension
- Ring with gaussian radius
- 2-D analogue of camel back
- GAN is trained on cartesian coordinates
- Quantiles are calculate in polar coordinates

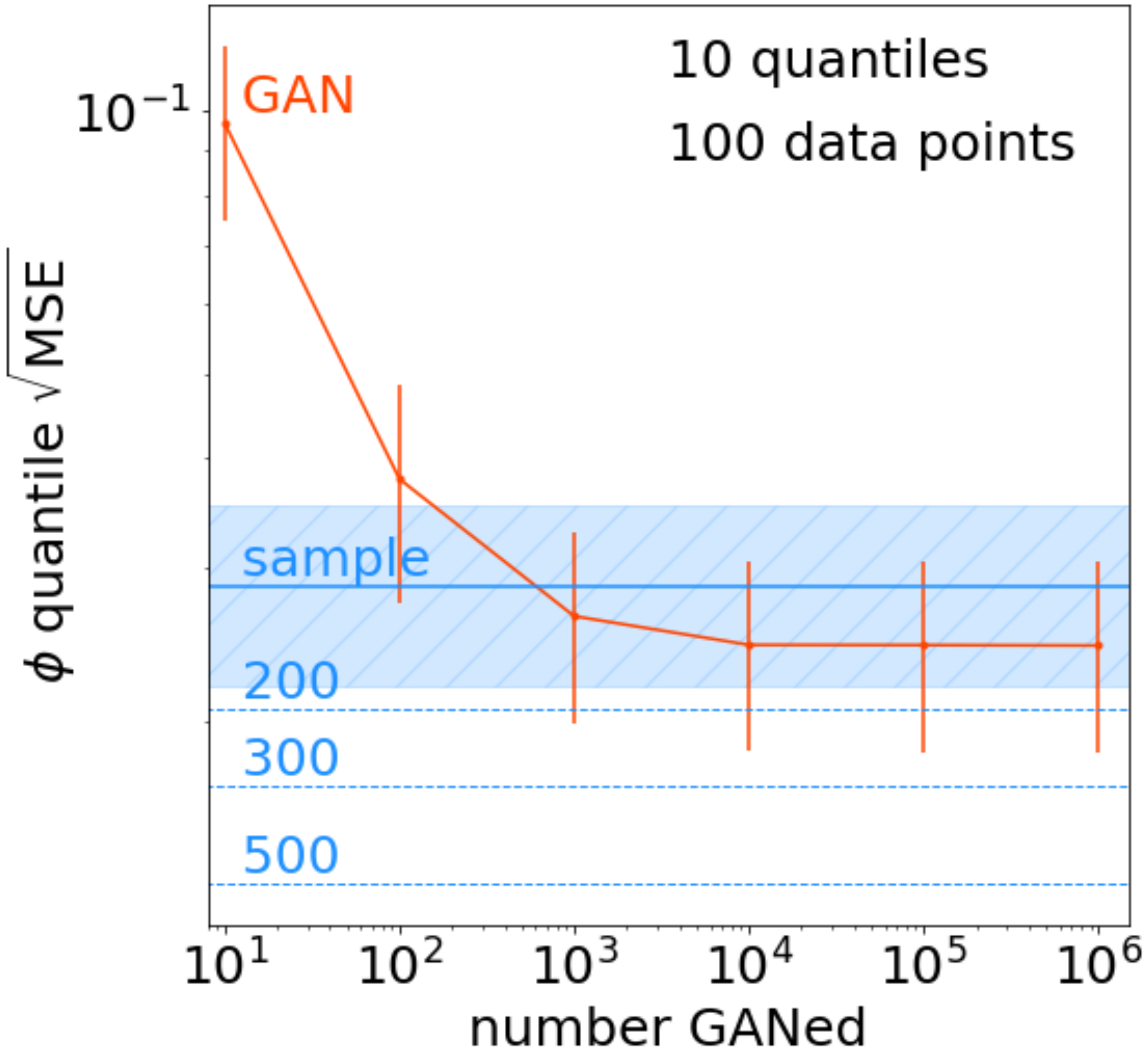


➡ GAN has to learn correlations

2-D Toy Model

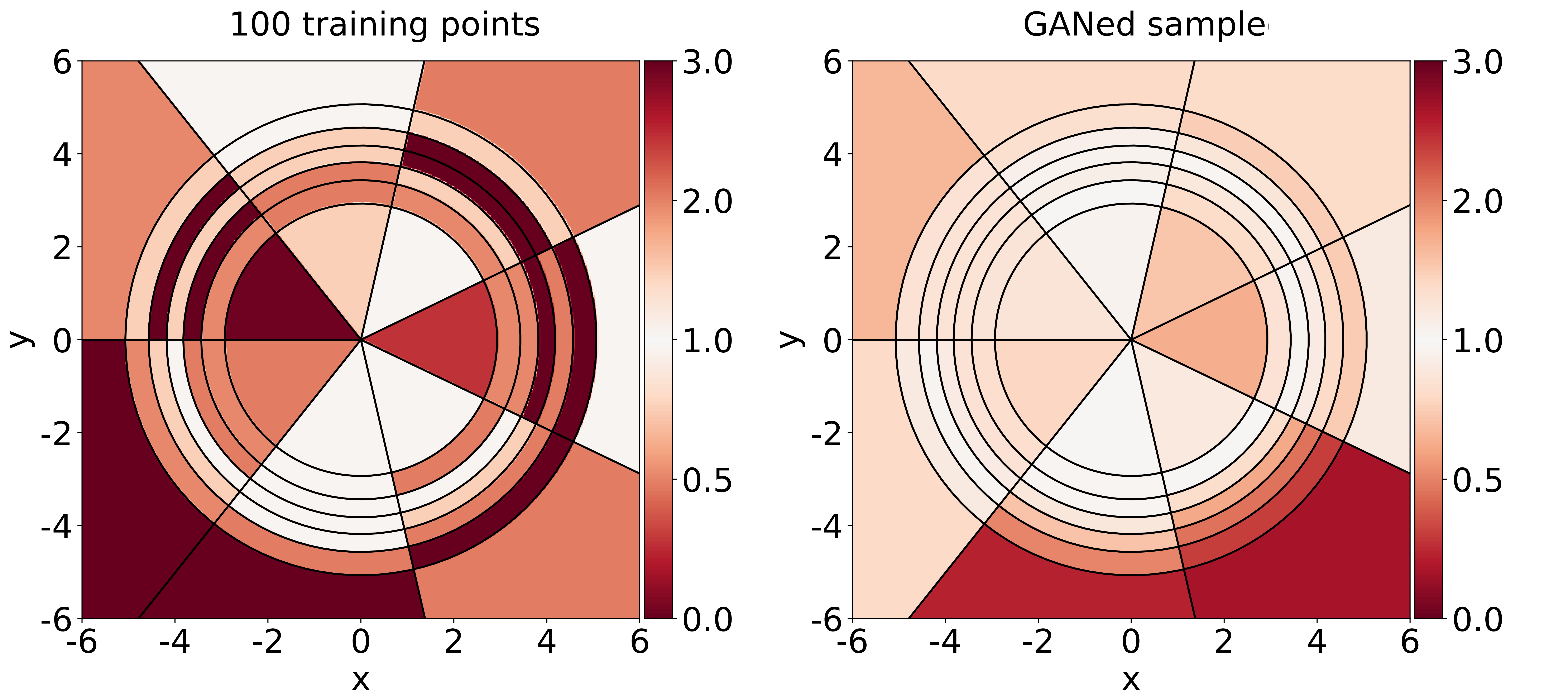
- Once again: compare quantile MSE





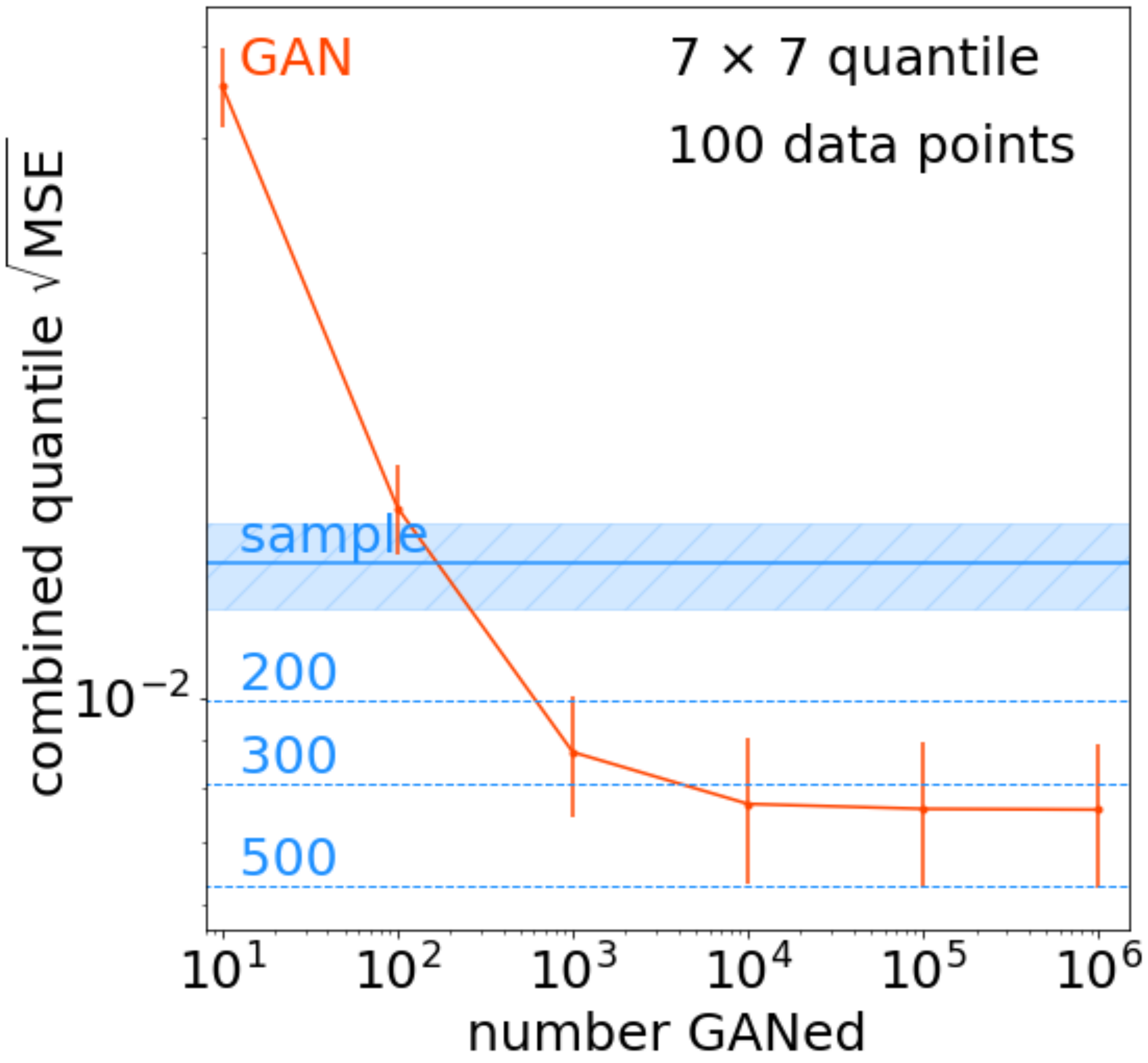
2-D Combined Quantile

- Combine quantiles in radius and angle direction
  - 2-D histogram with quantiles as bins



2-D Combined Quantile

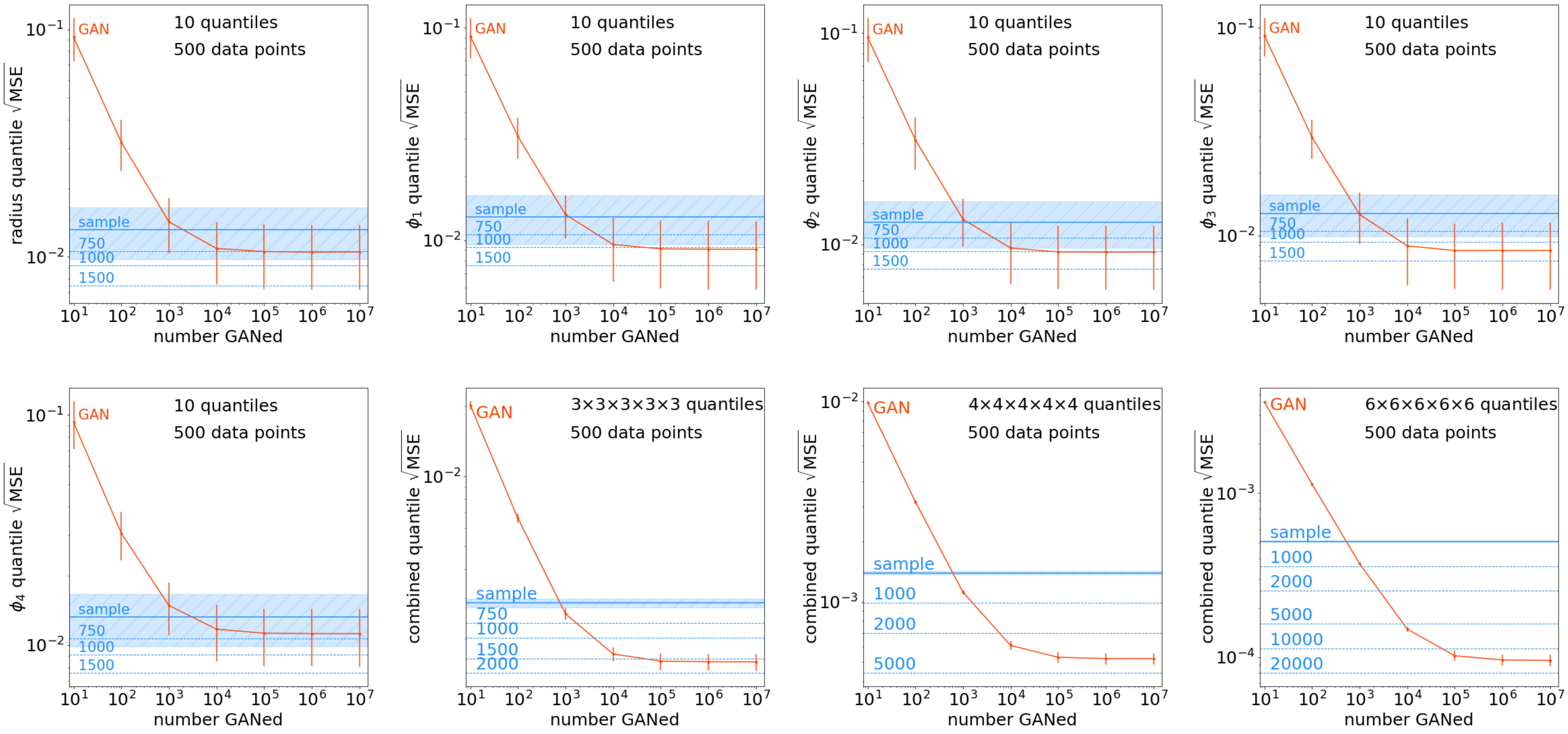
- Similar behaviour to 50 quantile case in 1-D
- GAN manages to interpolate in 2-D space as well
- Indicates use beyond simple toy model



5-D Toy Model

- Further extend setup into five dimensions
- Surface of a 5-sphere with Gaussian radius
- Angles sampled uniform on 5-sphere
- Increased size of training samples to 500
- Perform similar quantile MSE comparison as before

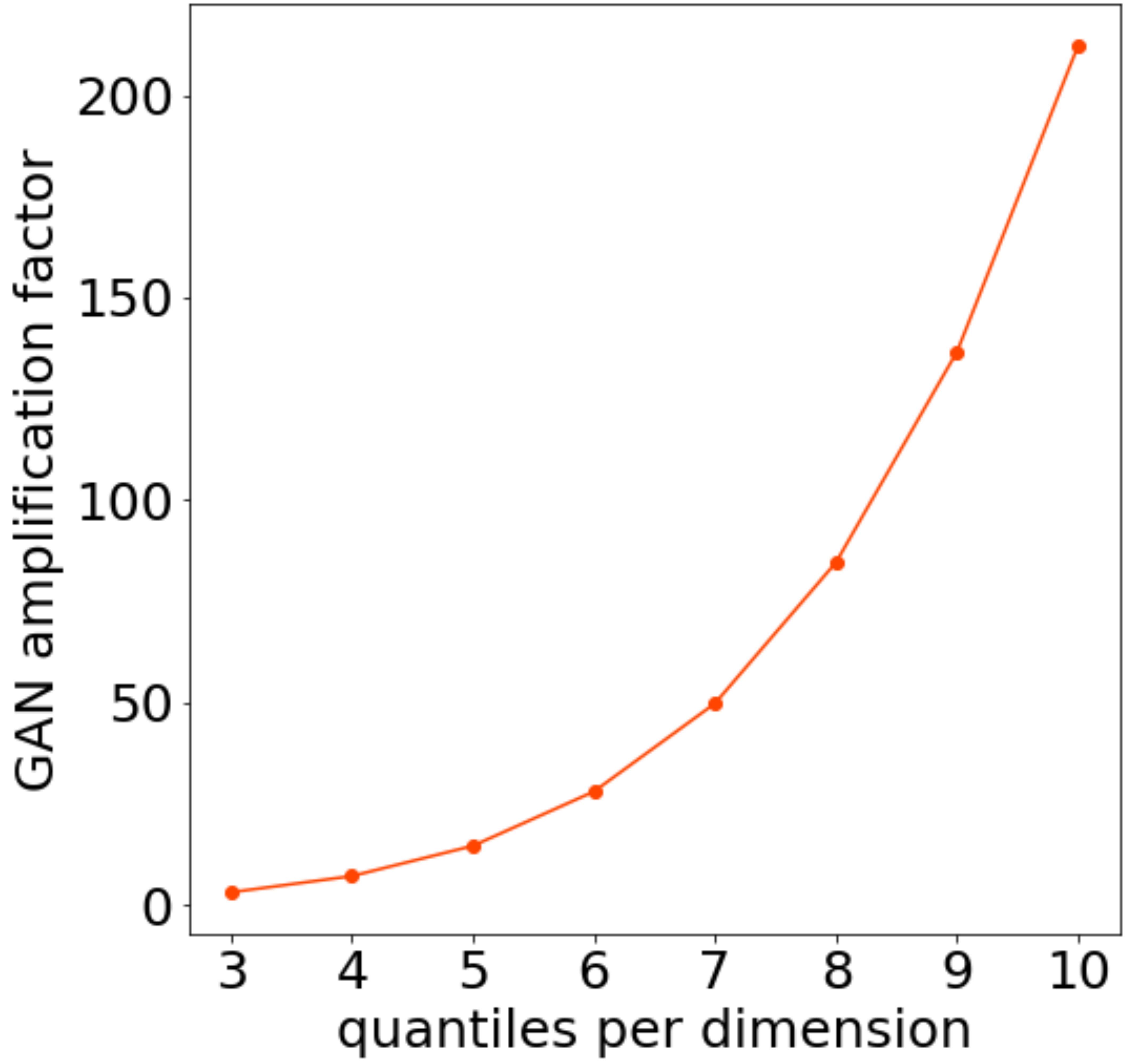
5-D Toy Model



radius quantile √MSE
GAN
10 quantiles
500 data points
sample
750
1000
1500
number GANed
ϕ_1 quantile √MSE
GAN
10 quantiles
500 data points
sample
750
1000
1500
number GANed
ϕ_2 quantile √MSE
GAN
10 quantiles
500 data points
sample
750
1000
1500
number GANed
ϕ_3 quantile √MSE
GAN
10 quantiles
500 data points
sample
750
1000
1500
number GANed
ϕ_4 quantile √MSE
GAN
10 quantiles
500 data points
sample
750
1000
1500
number GANed
combined quantile √MSE
GAN
3×3×3×3×3 quantiles
500 data points
sample
750
1000
1500
2000
number GANed
combined quantile √MSE
GAN
4×4×4×4×4 quantiles
500 data points
sample
1000
2000
5000
number GANed
combined quantile √MSE
GAN
6×6×6×6×6 quantiles
500 data points
sample
1000
2000
5000
10000
20000
number GANed

5-D Toy Model

- Plot amplification factor as function of N quantiles
- Interpolation power again gets greater for sparser data
- Very sparse data not commonly encountered
- Although still possible for high enough dimensions



Conclusion

- If a GAN is trained on N data points, how many new points can I draw from the GAN?
- Of course dependant on GAN setup and dataset
- If dataset allows for smooth interpolation:

➡ More then N points

- Condition is fulfilled for a lot of physics cases

➡ Promising for physics application

Thank you