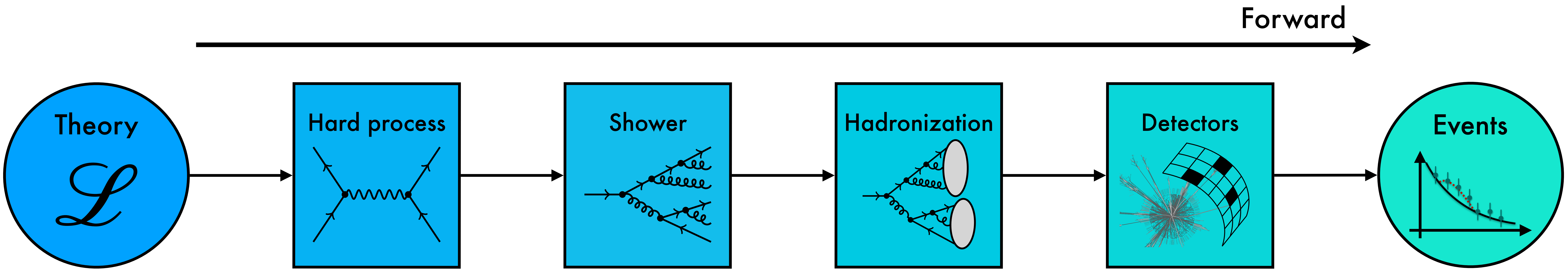
Forecasting Generative Amplification

Sascha Cassandra Diefenbacher, Heidelberg University

ML4Jets 2026, Vienna

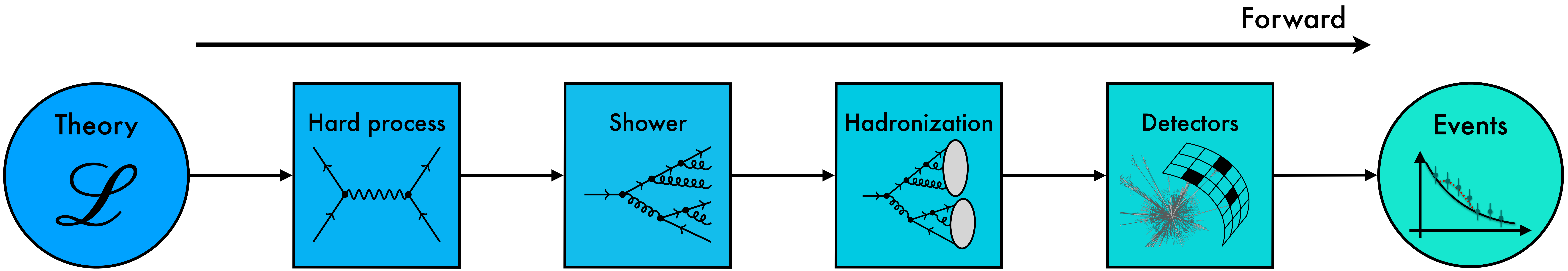
September 15th, 2026

Particle Physics Simulation



Particle Physics Simulation

Particle Physics Simulation

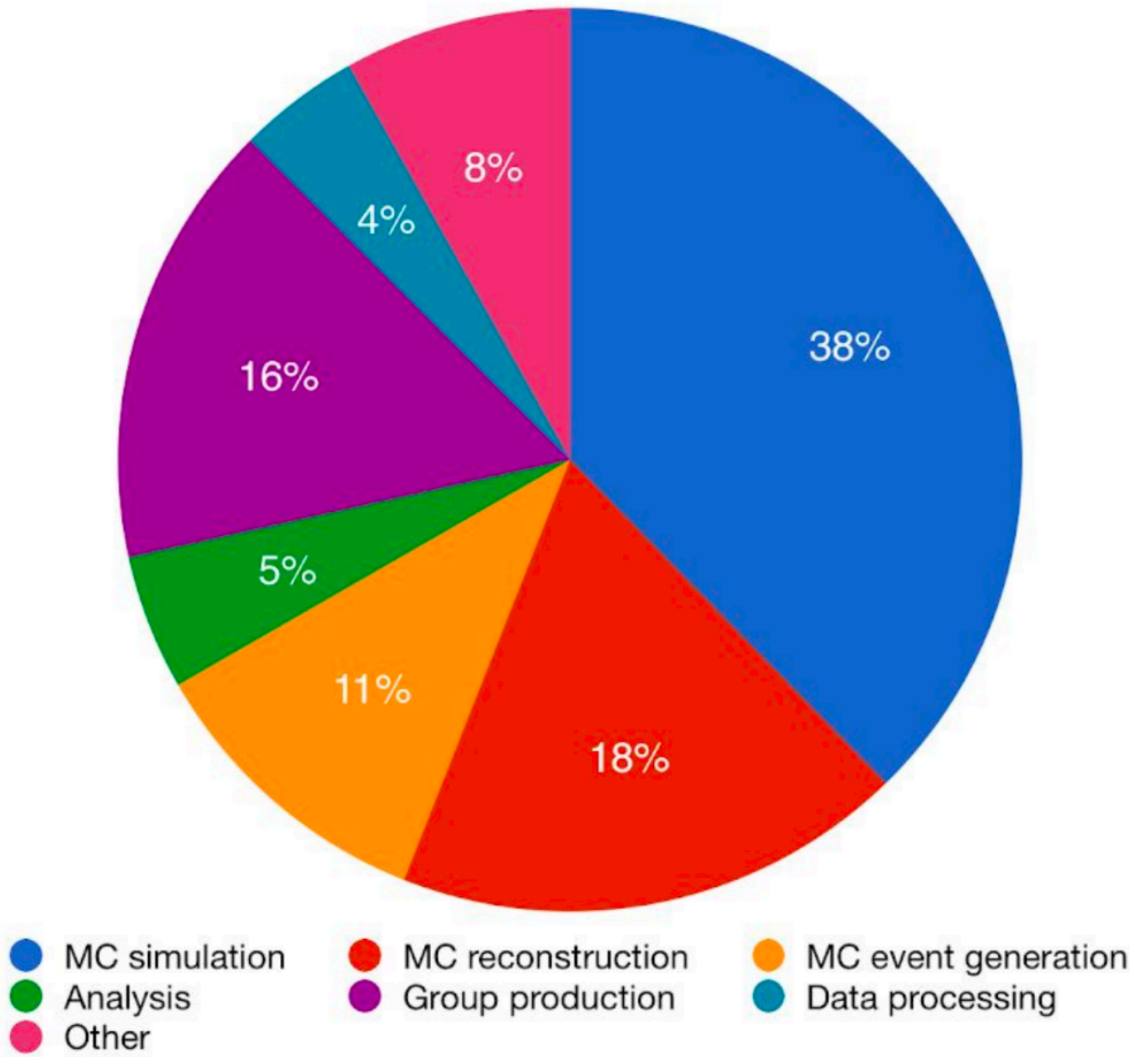


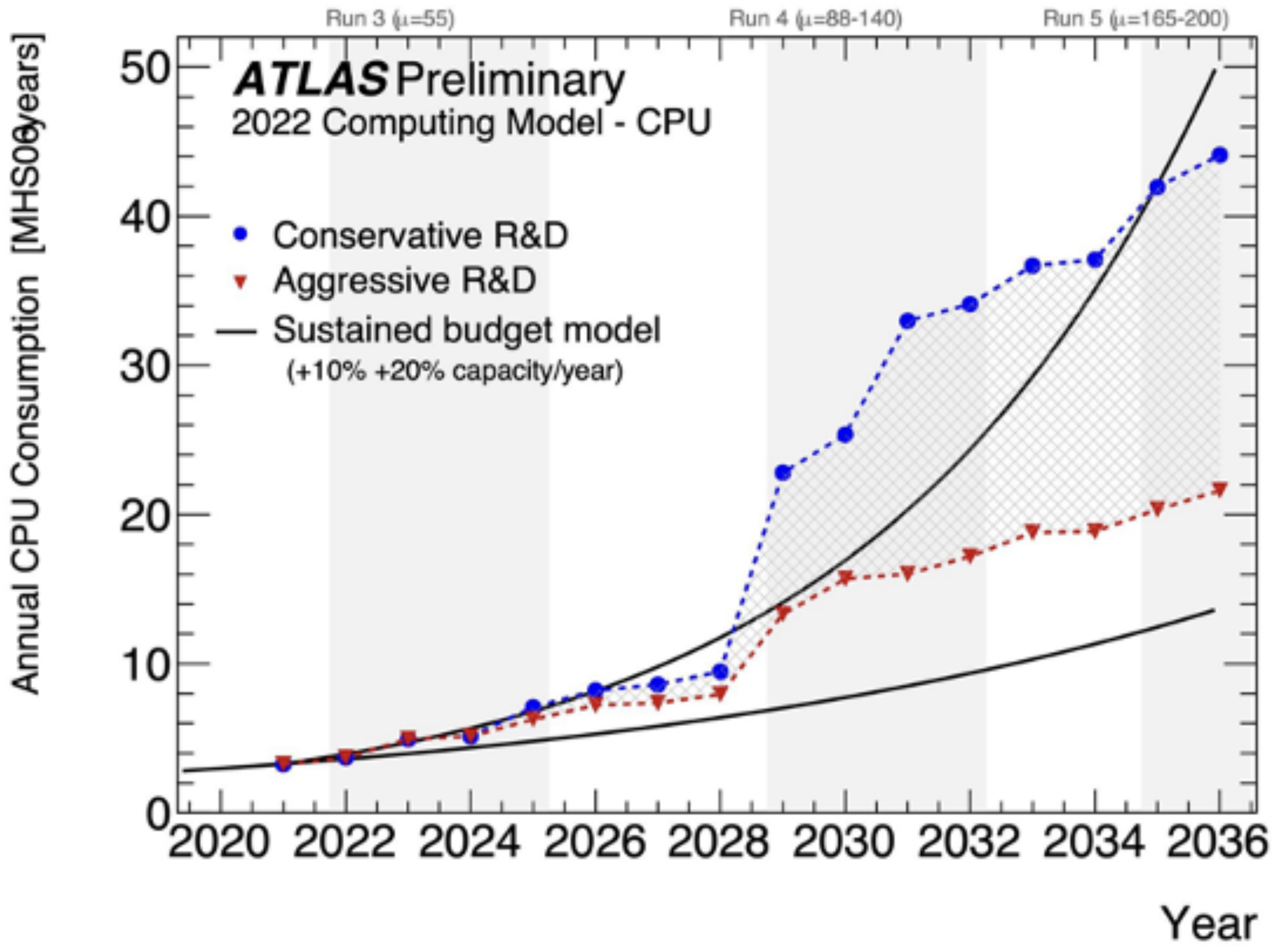
AllShowers

- First-principles precision simulations, accurate, expensive and slow
- Need more simulated data than recorded data
- Higher-rate experiments → need more and more precise simulation

Particle Physics Simulation

Forecasting Amplification



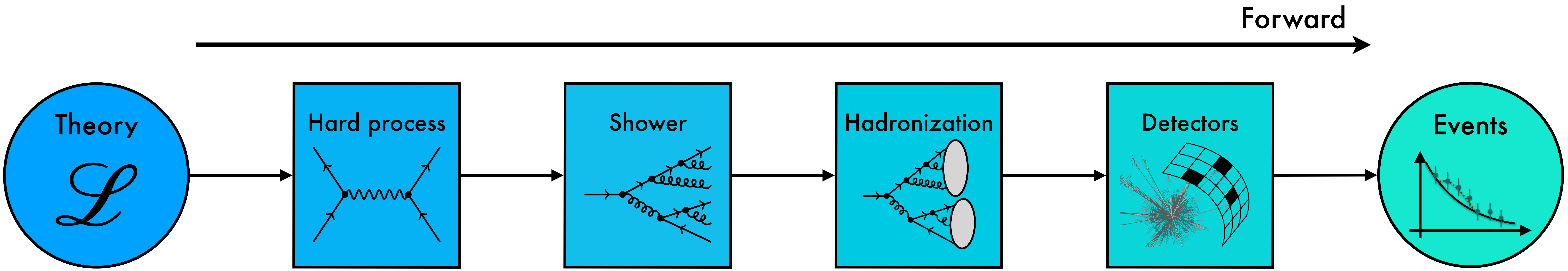


- Simulation cost is computational bottleneck for future HEP experiments
- Need accelerated methods

Al-Turany et. al. JENA Computing Initiative WP2 Report: Software and Heterogeneous Architectures, 2503.09213

Catmore et. al. ATLAS HL-LHC Computing Conceptual Design Report, CERN-LHCC-2020-015 ; LHCC-G-178

Generative Simulation



$p_{\text{scatter}}(x|\mathscr{L})$ $\quad p_{\text{shower}}(x|x_{\text{scatter}})$ $\quad p_{\text{hadron}}(x|x_{\text{shower}})$ $\quad p_{\text{detector}}(x|x_{\text{hadron}})$

$\Downarrow$

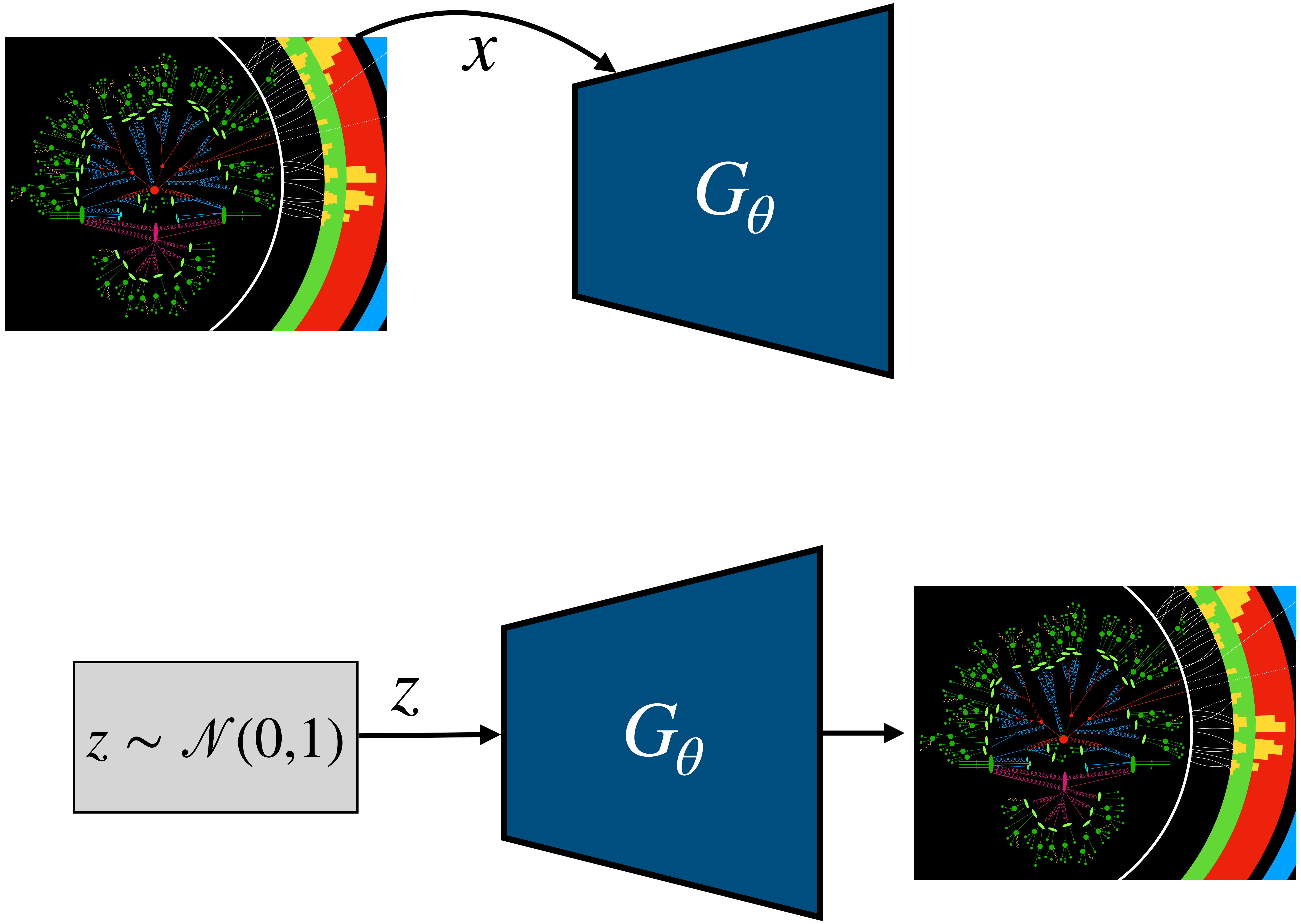
$p_{\text{gen, scatter}}(x|\mathscr{L})$ $\quad p_{\text{gen, shower}}(x|x_{\text{scatter}})$ $\quad p_{\text{gen, hadron}}(x|x_{\text{shower}})$ $\quad p_{\text{gen, detector}}(x|x_{\text{hadron}})$

- Series of conditional densities, approximate with conditional generative networks
- Faster, more efficient, leverage data

Generative Simulation

Fast simulation

- Train on simulated data
- Sample from generative network
- Faster than classical simulation
  - GAN
  - Normalizing flow
  - Diffusion models
  - Autoregressive transformers

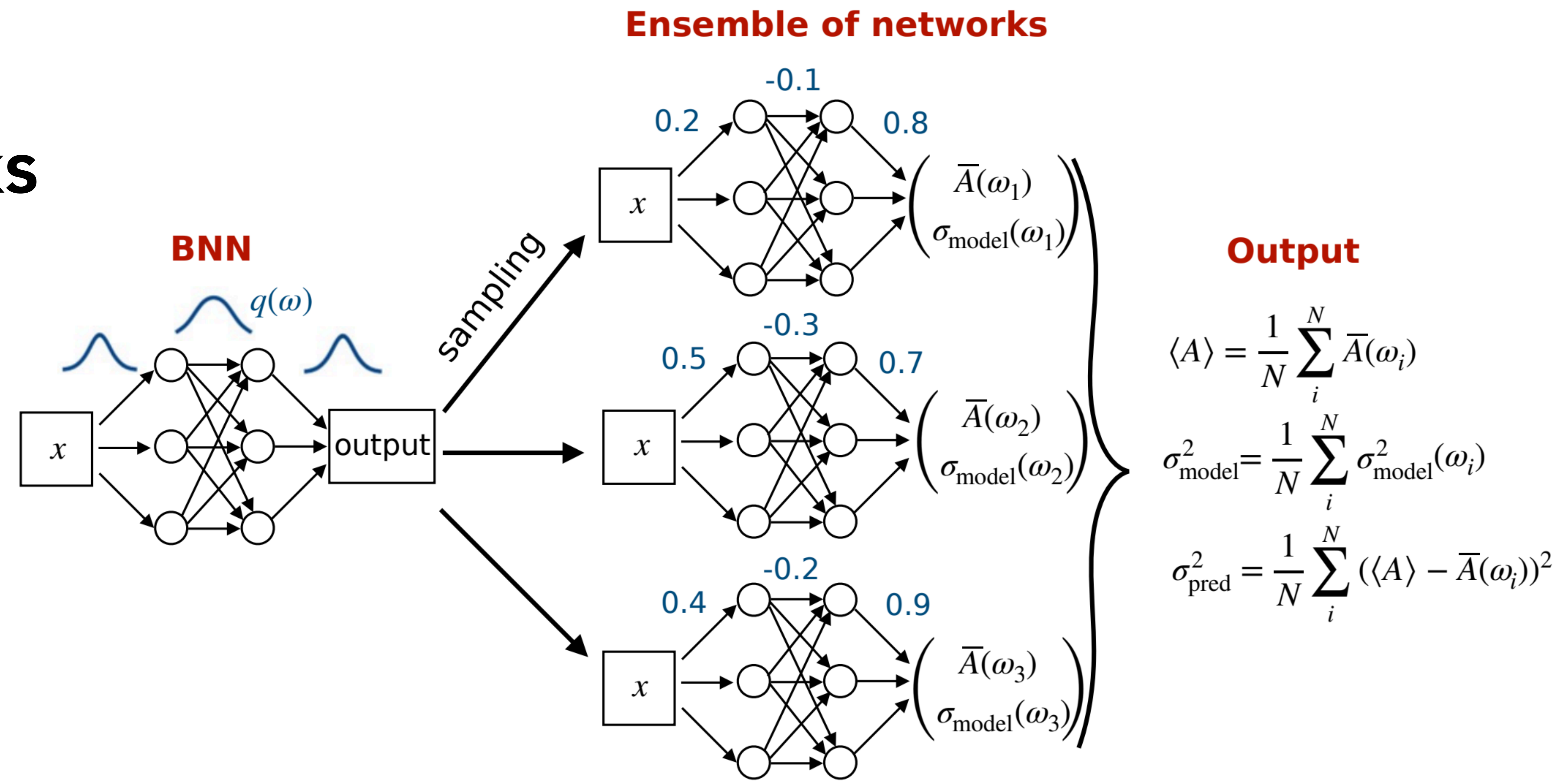


AI Uncertainties

Physics needs uncertainty estimates

ML uncertainty quantification:

- Heteroscedastic losses, repulsive ensemble, **bayesian neural networks**
- Transferable to generative models with likelihood losses
- Bayesian normalizing flows
- Uncertainty on **density estimation**

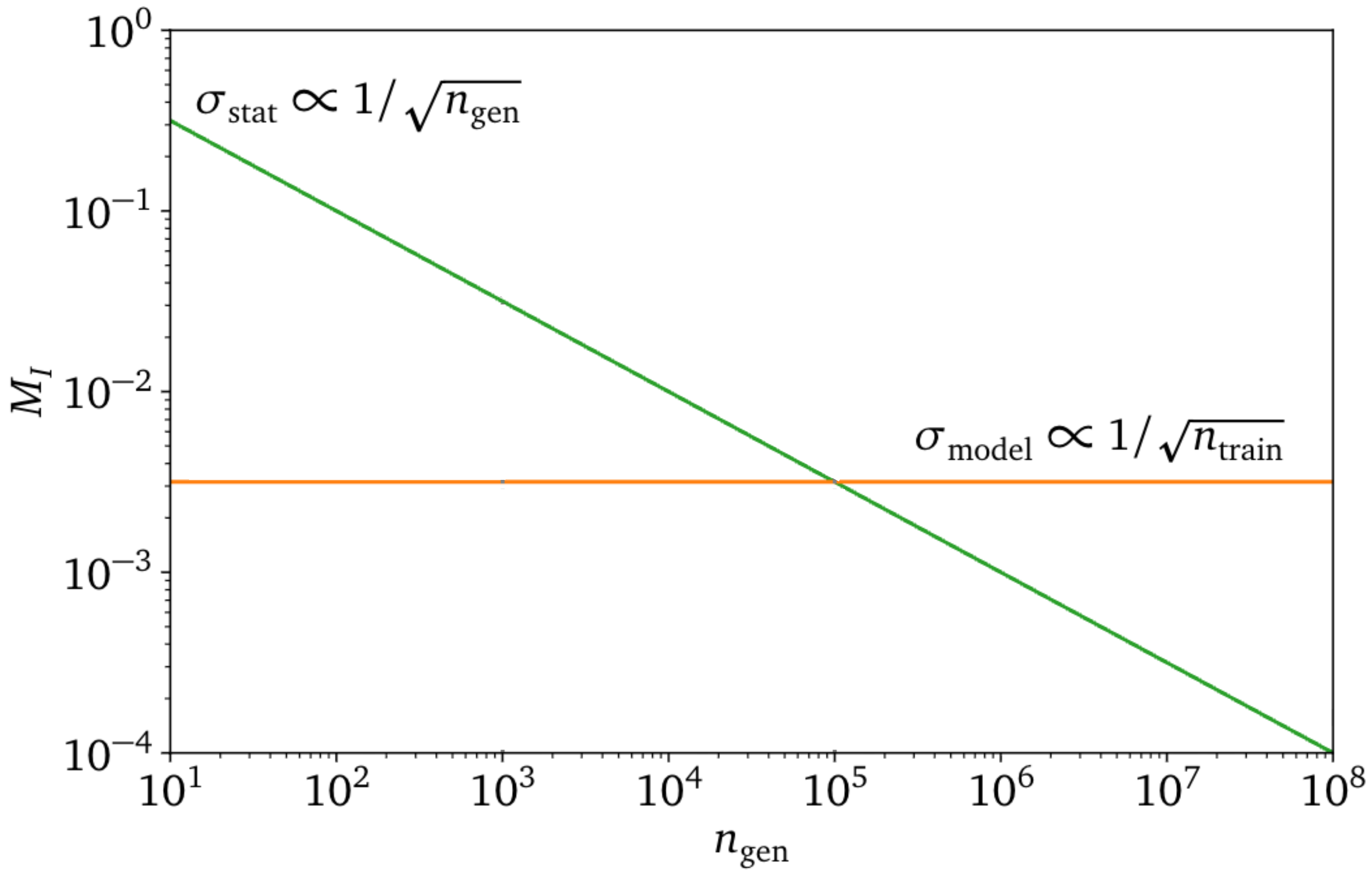


Butter et. al. Know What You Don't Flow, 2608.21509

AI Uncertainties

What uncertainty is learned?

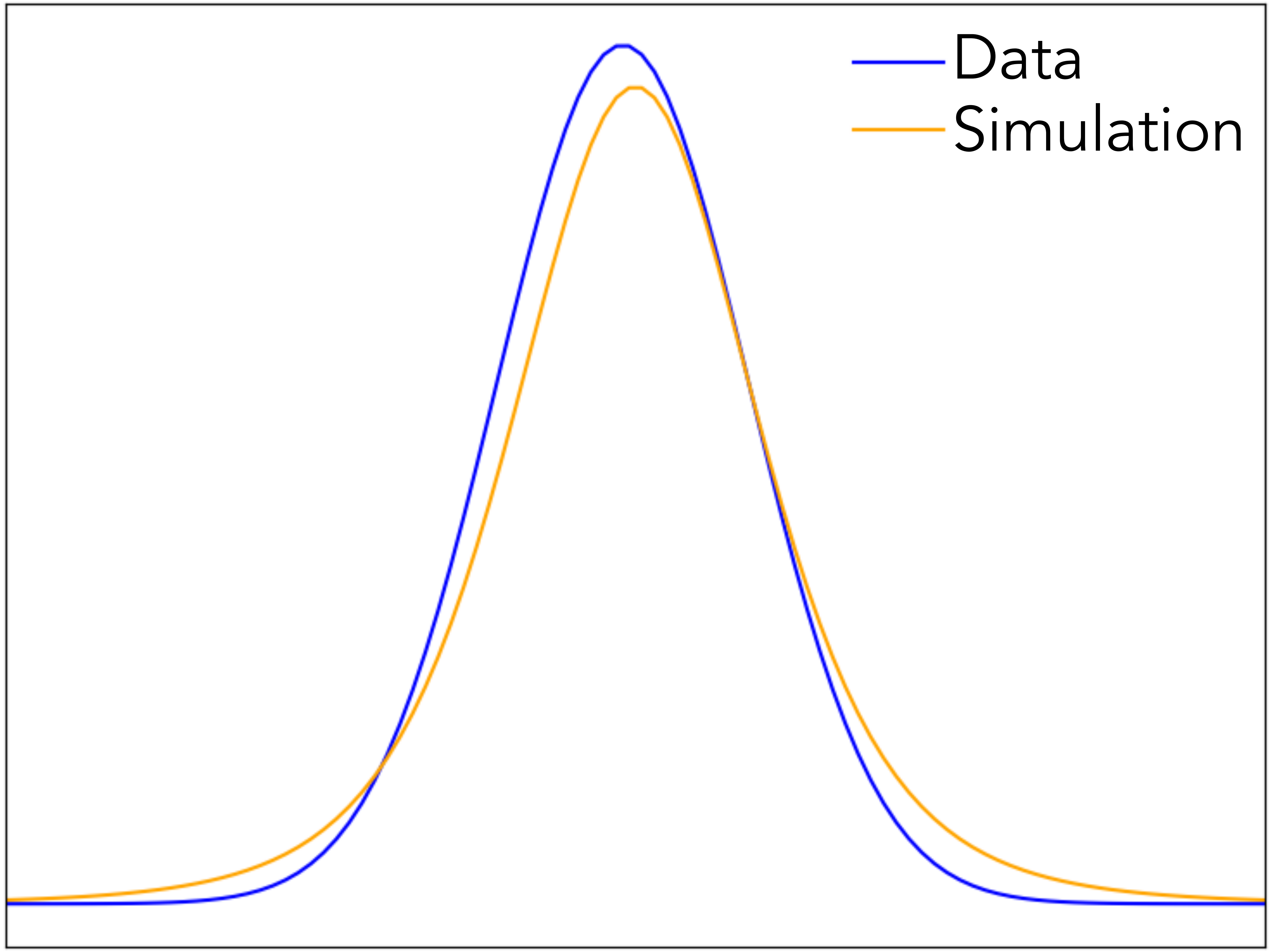
- For Generative models:
- **Statistical uncertainty** σ_{stat}
  - Scales $\dfrac{a}{\sqrt{n_{\text{gen}}}}$
  - Infinitely reducible
- **Generalization uncertainty** σ_{model}
  - Scales $\dfrac{b}{\sqrt{n_{\text{train}}}}$
  - Irreducible by sampling



AI Uncertainties

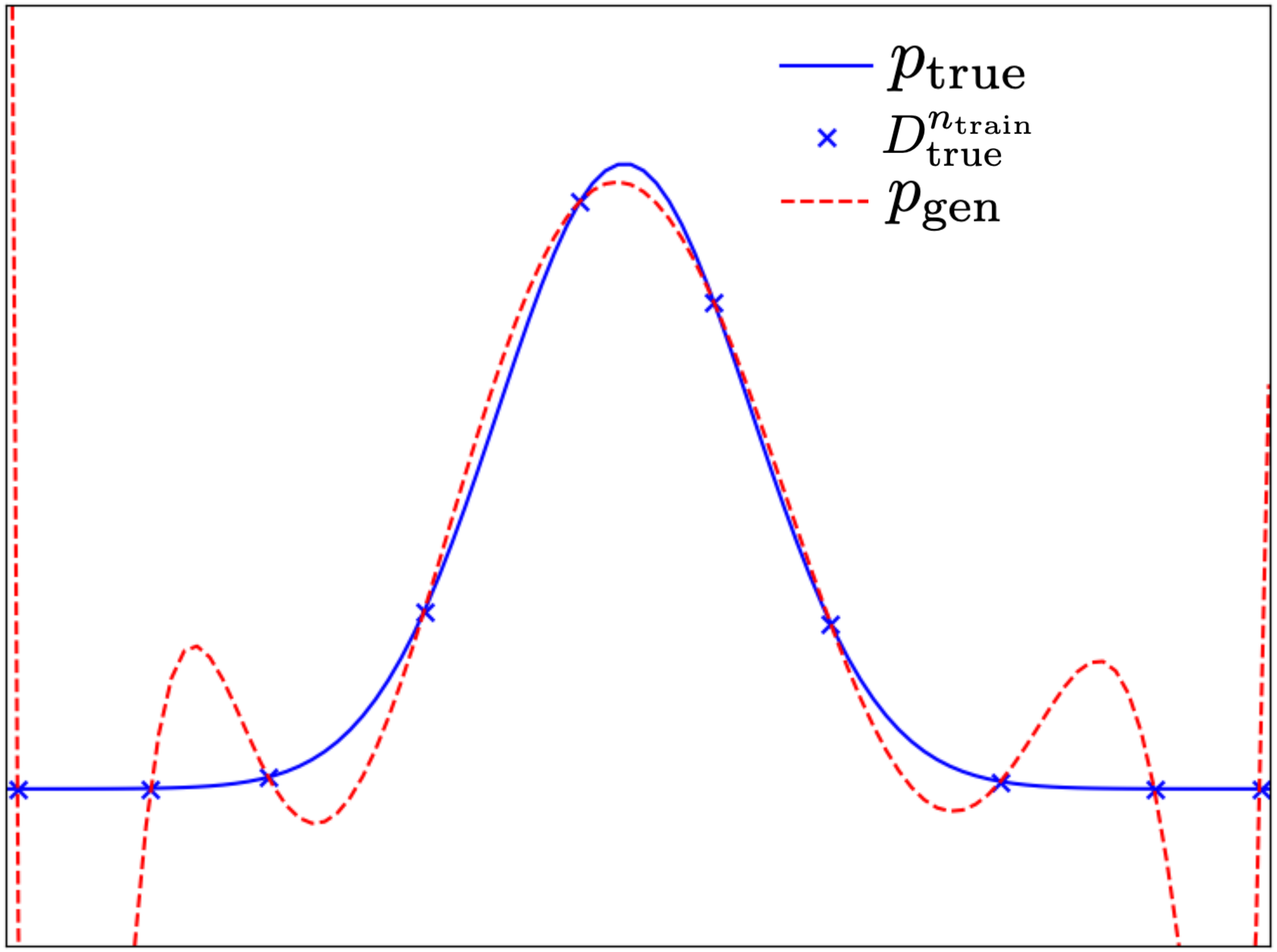
Generalization uncertainty

- Most applications:
  - Inference data ≠ training data
  - Extrapolation outside training



Generalization uncertainty

- Most applications:
  - Inference data ≠ training data
  - Extrapolation outside training
- Generative modeling:
  - Extrapolation in resolution



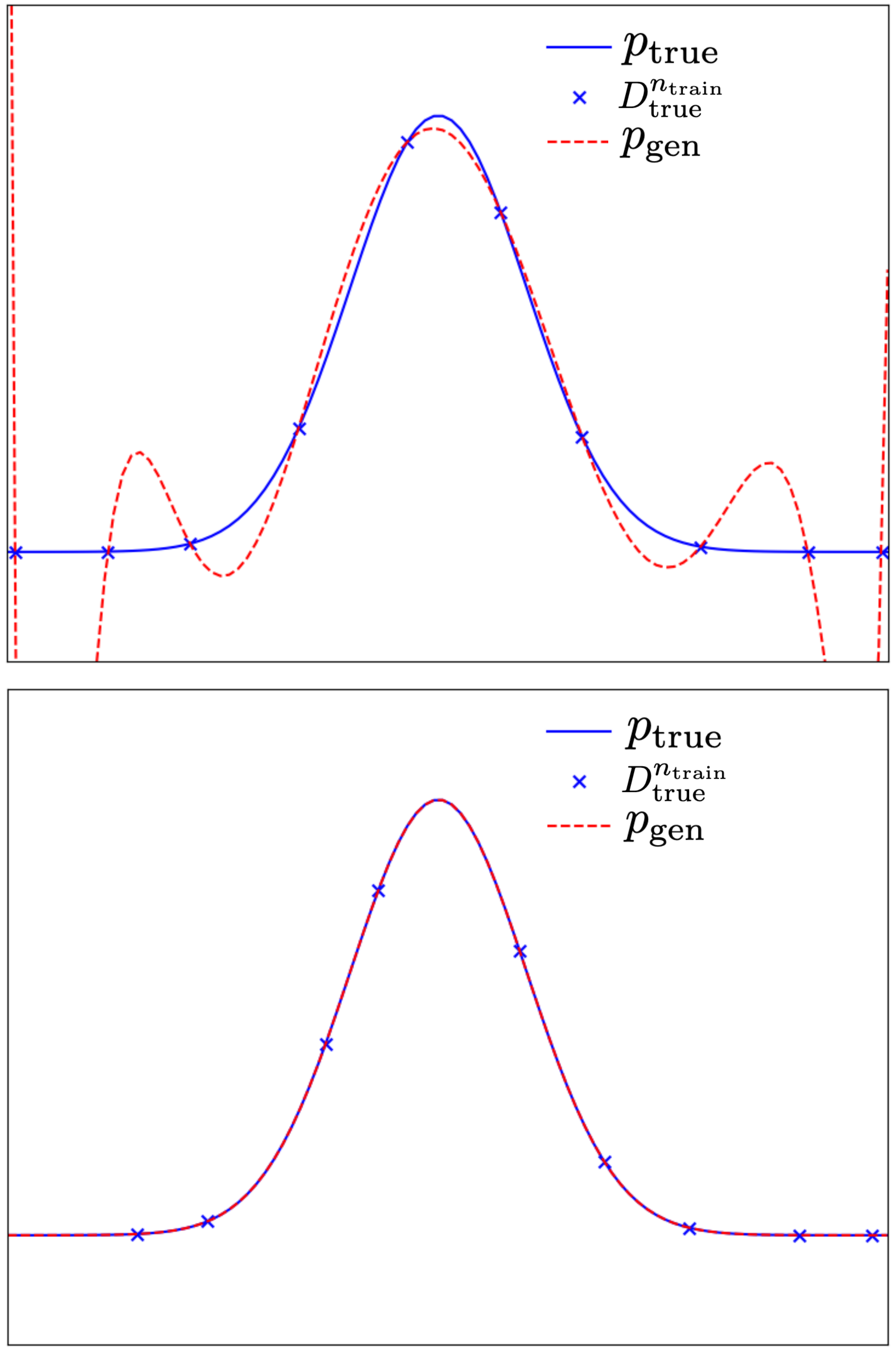
AI Uncertainties

Fundamental question

- What happens if I exceed the training resolution?
- How many points n_{gen} can I draw?
- Generative simulation assumes $n_{\text{gen}} > n_{\text{train}}$
- Is this really true?

"If I train a model on 1 million data points, how many can I draw from that model?"
—Louis Lyons

→**Amplification**



Amplification

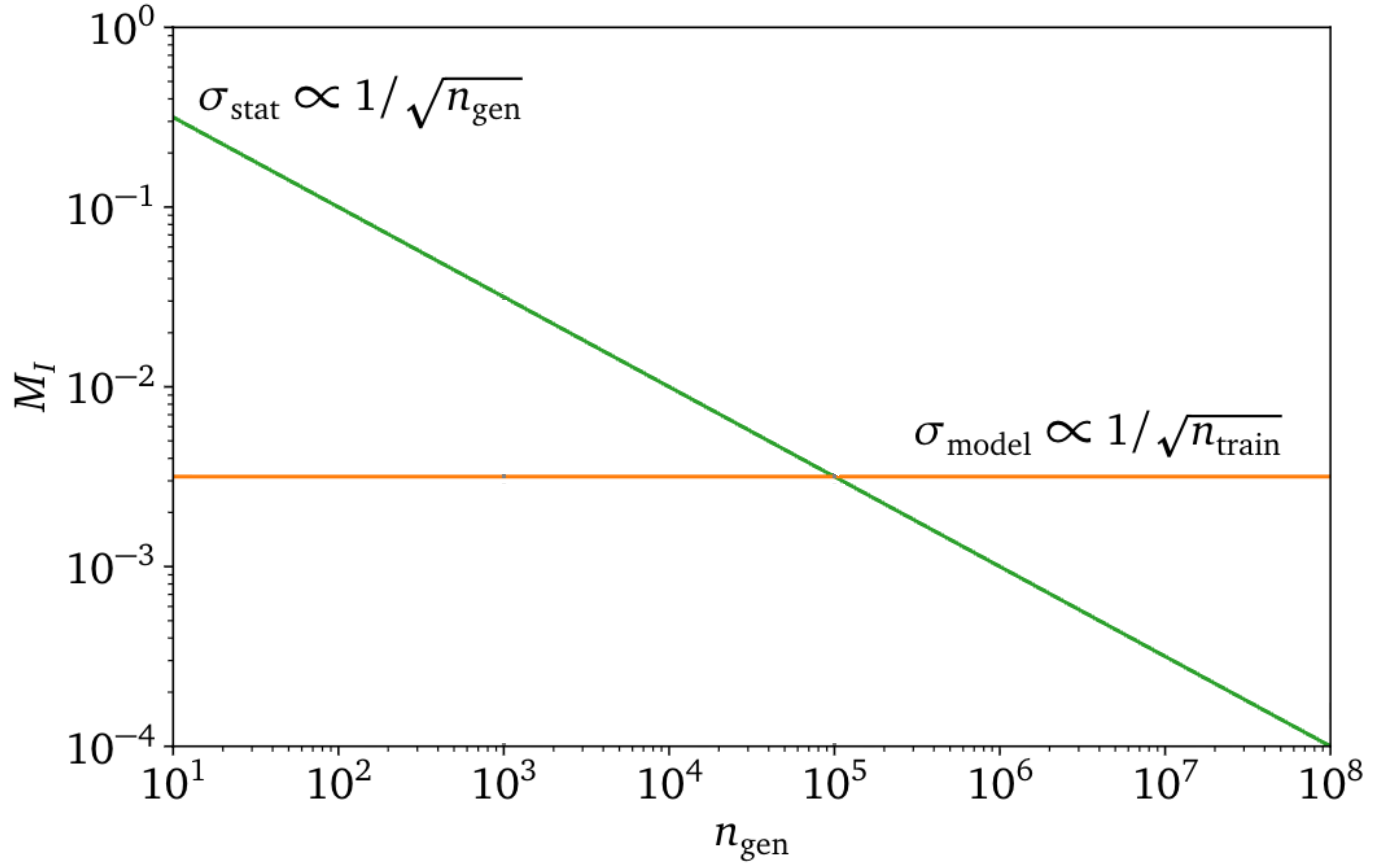
How many generative points can I sample from the model?

$$\sigma^2_{\text{gen}} = \sigma^2_{\text{model}} + \sigma^2_{\text{stat}}$$

$\sigma_{\text{model}} > \sigma_{\text{stat}}$: new points stop being useful

$$\sigma_{\text{stat}}(n)\Big|_{n=n_{\text{equiv}}} \stackrel{!}{=} \sigma_{\text{model}}$$

n_{equiv}: size of true data set with same stat. uncertainty as generalization uncertainty



Amplification

How many generative points can I sample from the model?

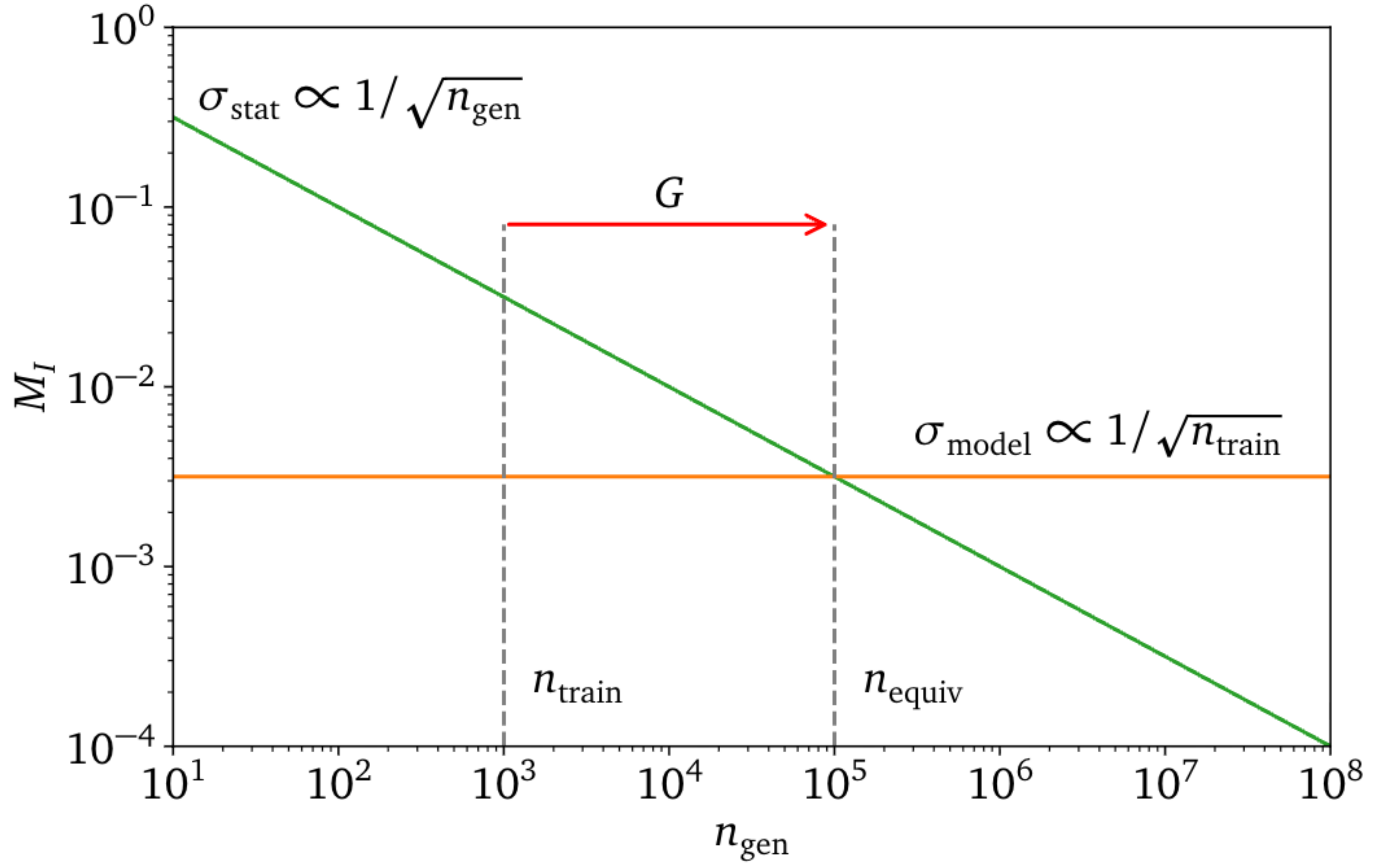
$$\sigma^2_{\text{gen}} = \sigma^2_{\text{model}} + \sigma^2_{\text{stat}}$$

$\sigma_{\text{model}} > \sigma_{\text{stat}}$: new points stop being useful

$$\sigma_{\text{stat}}(n)\Big|_{n=n_{\text{equiv}}} \stackrel{!}{=} \sigma_{\text{model}}$$

n_{equiv}: size of true data set with same stat. uncertainty as generalization uncertainty

$n_{\text{equiv}} \leq n_{\text{train}}$	: No Amplification
$n_{\text{equiv}} > n_{\text{train}}$	: Amplification
$G = \dfrac{n_{\text{eqiv}}}{n_{\text{train}}}$	: Amplification factor



Quantifying Amplification

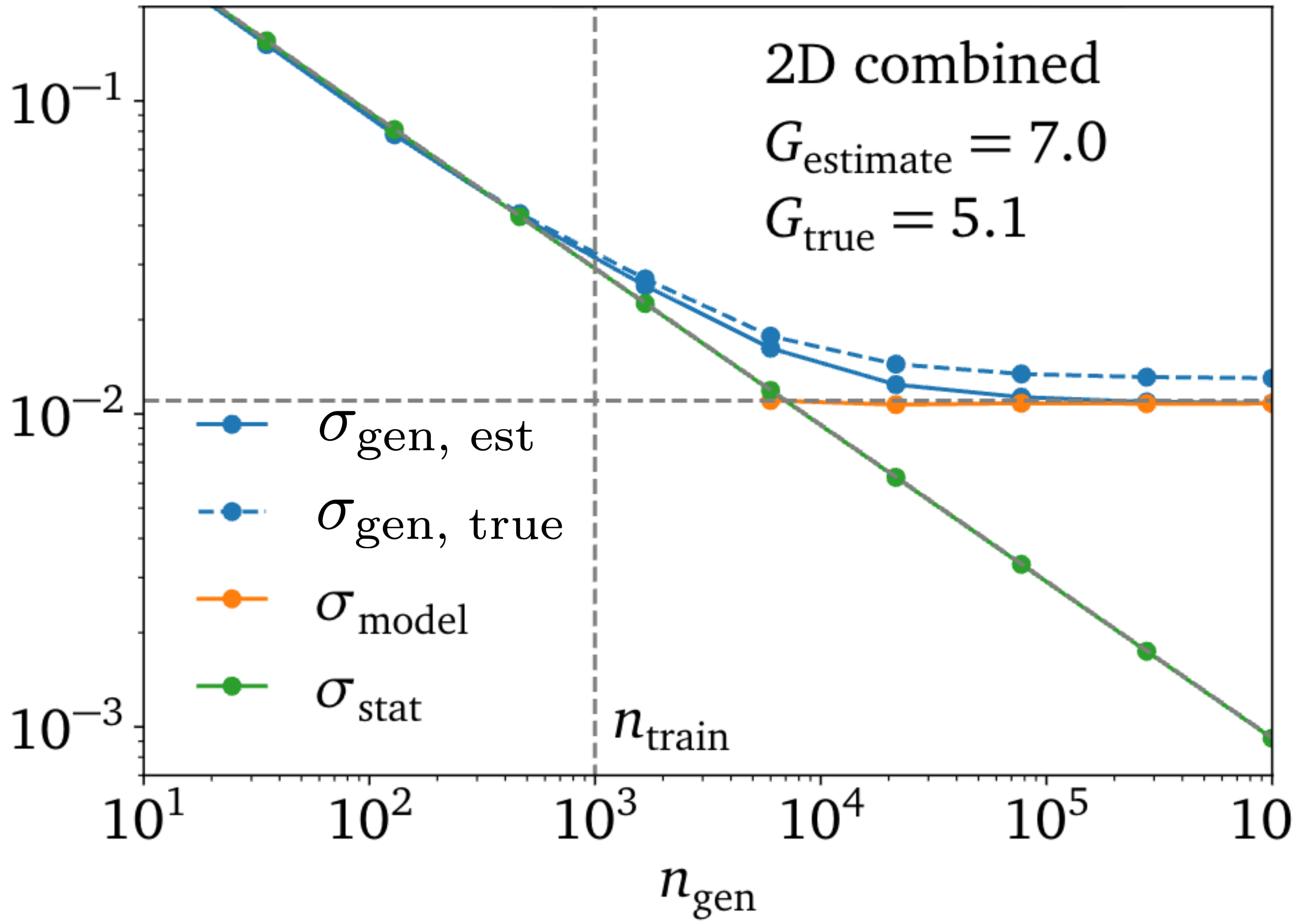
Example: Gaussian Ring Radius

- Plot σ_{gen}
- Get $\sigma_{\text{stat}}(n)$ from falling section
- Get σ_{model} from asymptotic part

$$\sigma_{\text{stat}}(n)\Big|_{n=n_{\text{equiv}}} \overset{!}{=} \sigma_{\text{model}}$$

$$G = \frac{n_{\text{eqiv}}}{n_{\text{train}}}$$

- Estimated G agrees with truth



Bahl et al. Forecasting Generative Amplification 2509.08048

Quantifying Amplification

How to get σ_{gen} ?

- Uncertainty quantification?
  - Only σ_{model} not σ_{stat}

→ Need metric $M\left[D^n_{\text{gen}}, p_{\text{true}}\right]$ agreement (generated) set vs. true density

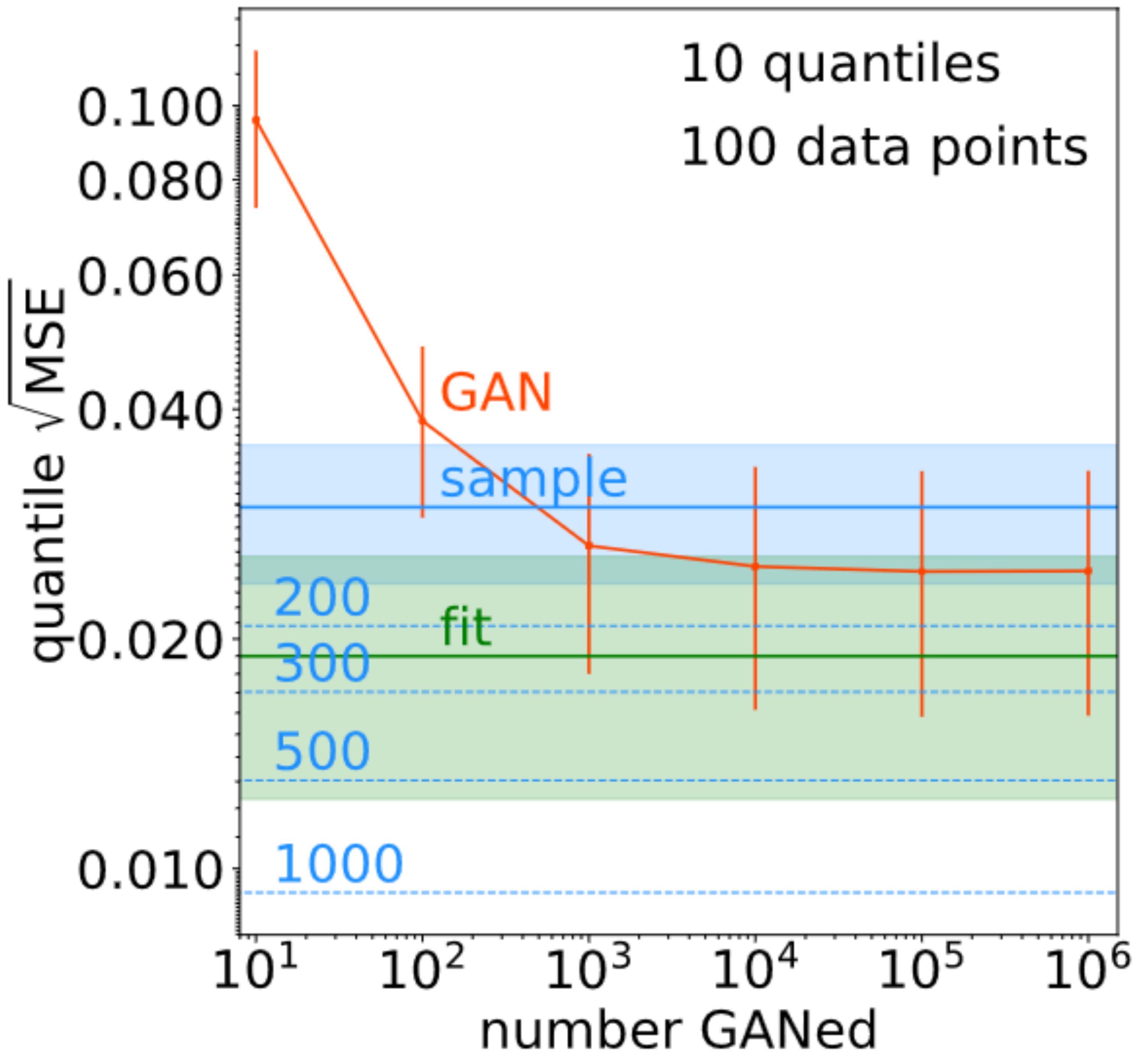
$$M\left[D^n_{\text{true}}, p_{\text{true}}\right] \approx \sigma_{\text{stat}} \qquad\qquad M\left[D^n_{\text{gen}}, p_{\text{true}}\right]\Big|_{n_{\text{gen}}\to\infty} \to \sigma_{\text{model}}$$

$$\boxed{\begin{array}{c} \sigma_{\text{stat}}(n)\Big|_{n=n_{\text{equiv}}} \overset{!}{=} \sigma_{\text{model}} \\ \Updownarrow \\ M\left[D^n_{\text{true}}, p_{\text{true}}\right]\Big|_{n=n_{\text{equiv}}} \overset{!}{=} M\left[D^n_{\text{gen}}, p_{\text{true}}\right]\Big|_{n_{\text{gen}}\to\infty} \end{array}}$$

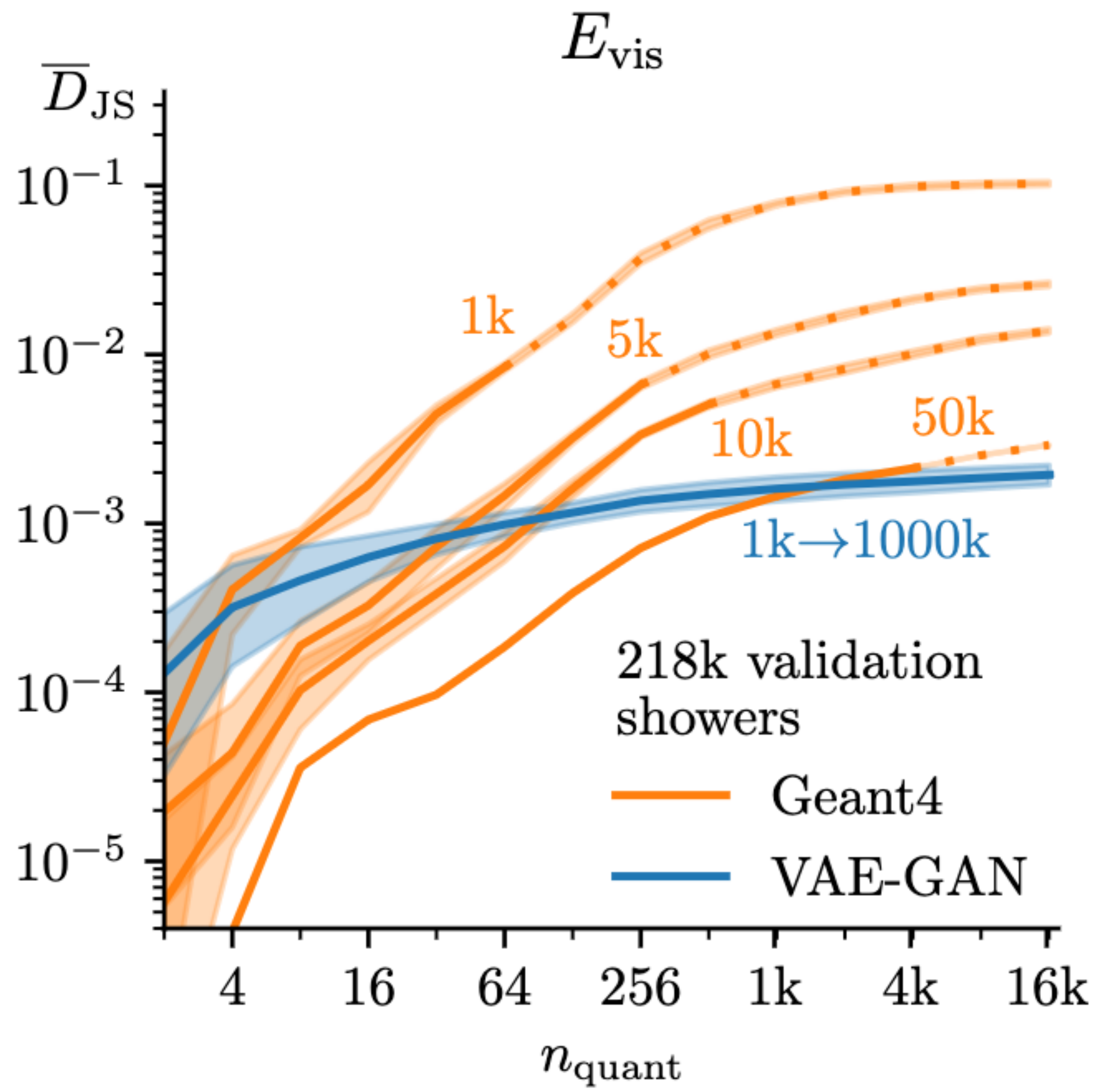
Quantifying Amplification

Known p_{true}

- Explicit



- Holdout set



- Past amplification results, proof of principle, hard to apply in practice

Butter et al. GANplifying Event Samples 2008.06545
Bieringer et al. Calomplification 2202.07352

Quantifying Amplification

First Metric: Averaging Amplification

- Fractions inside phase space region V

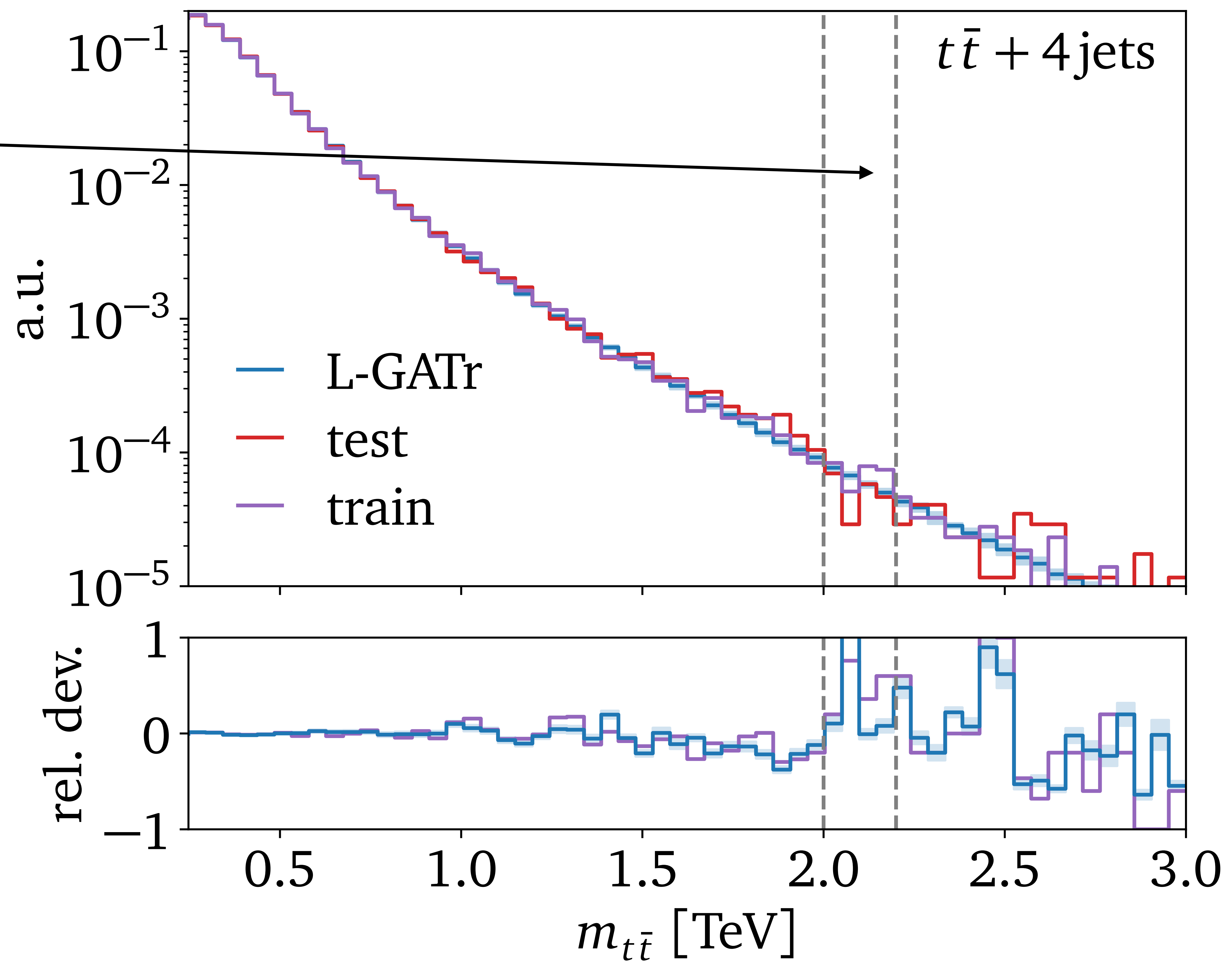
$$I = \int_V dx\, p_{\text{true}}(x)$$

$$\bar{I} = \frac{1}{n_{\text{gen}}} \sum_{x \in D_{\text{gen}}^{n_{\text{gen}}}} \mathbf{1}_{x \in V}$$

- Expected difference as metric

$$M_I\left[D_{\text{gen}}^{n_{\text{gen}}}, p_{\text{true}}\right] = \left\langle \left[I(p_{\text{true}}) - \bar{I}(D_{\text{gen}}^{n_{\text{gen}}}) \right]^2 \right\rangle$$

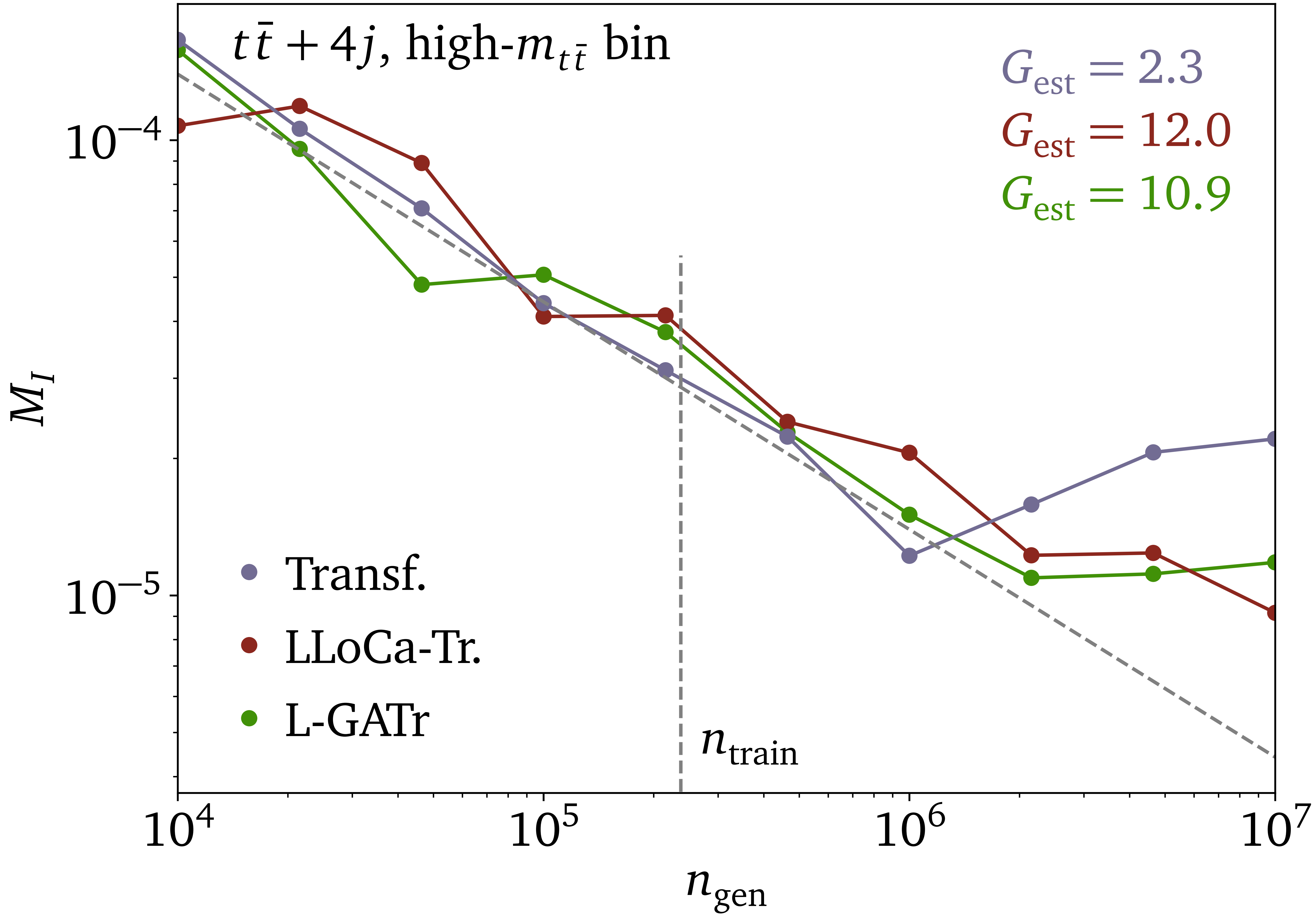
- Estimate this variance via BNN



Bahl et al. Forecasting Generative Amplification 2509.08048

Quantifying Amplification

Example: Di-Top + 4 jet events



Bahl et al. Forecasting Generative Amplification 2509.08048

Quantifying Amplification

Second Metric: Differential Amplification

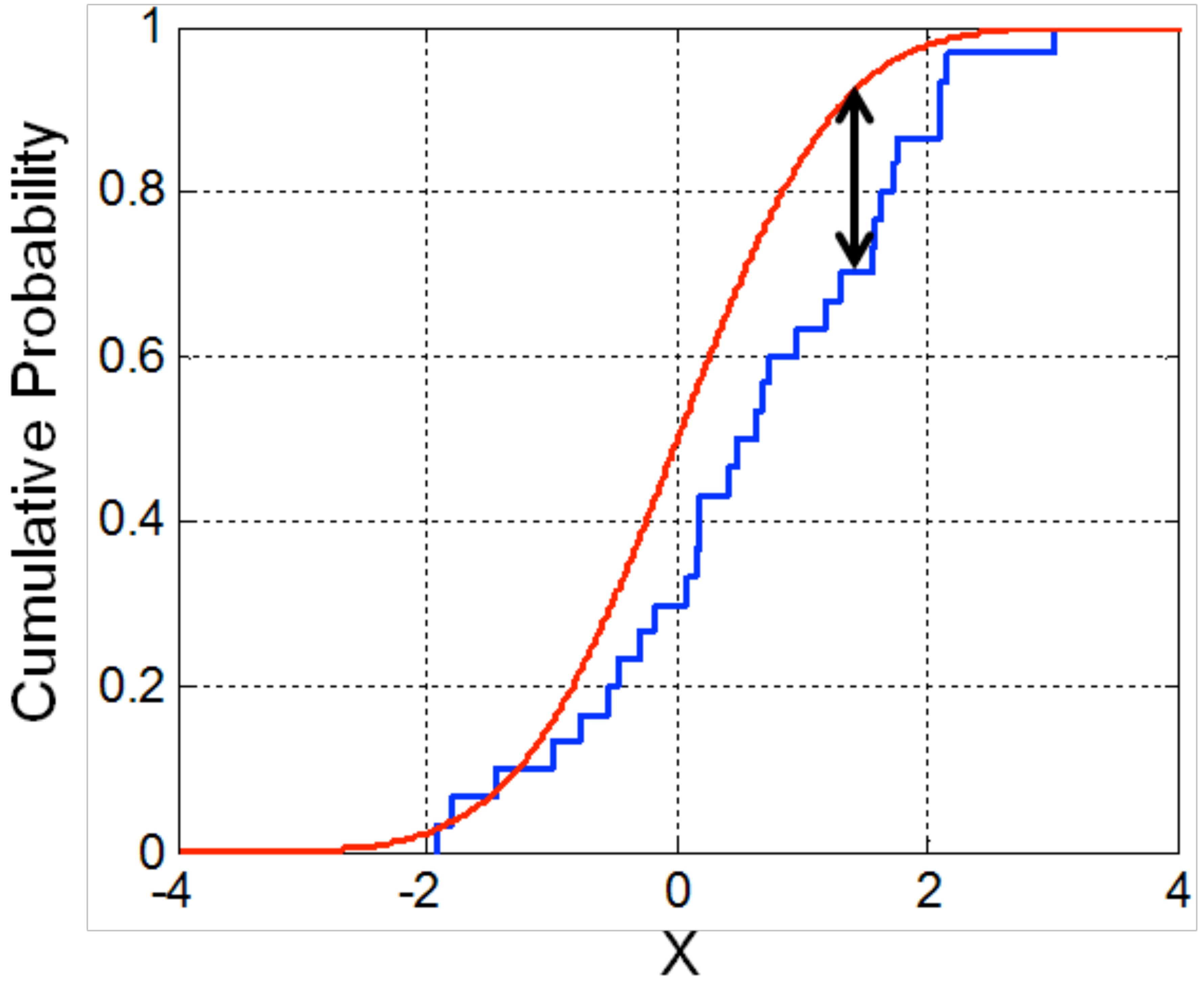
- Approximate unknown distribution with data sample

$$M\left[D^{n}_{\text{true}}, D^{n_{\text{train}}}_{\text{true}}\right] \approx M\left[D^{n}_{\text{true}}, p_{\text{true}}\right] \quad \text{with} \quad D^{n_{\text{train}}}_{\text{true}} \sim p_{\text{true}}$$

$$M\left[D^{n_{\text{gen}}}_{\text{gen}}, D^{n_{\text{train}}}_{\text{true}}\right] \approx M\left[D^{n_{\text{gen}}}_{\text{gen}}, p_{\text{true}}\right]$$

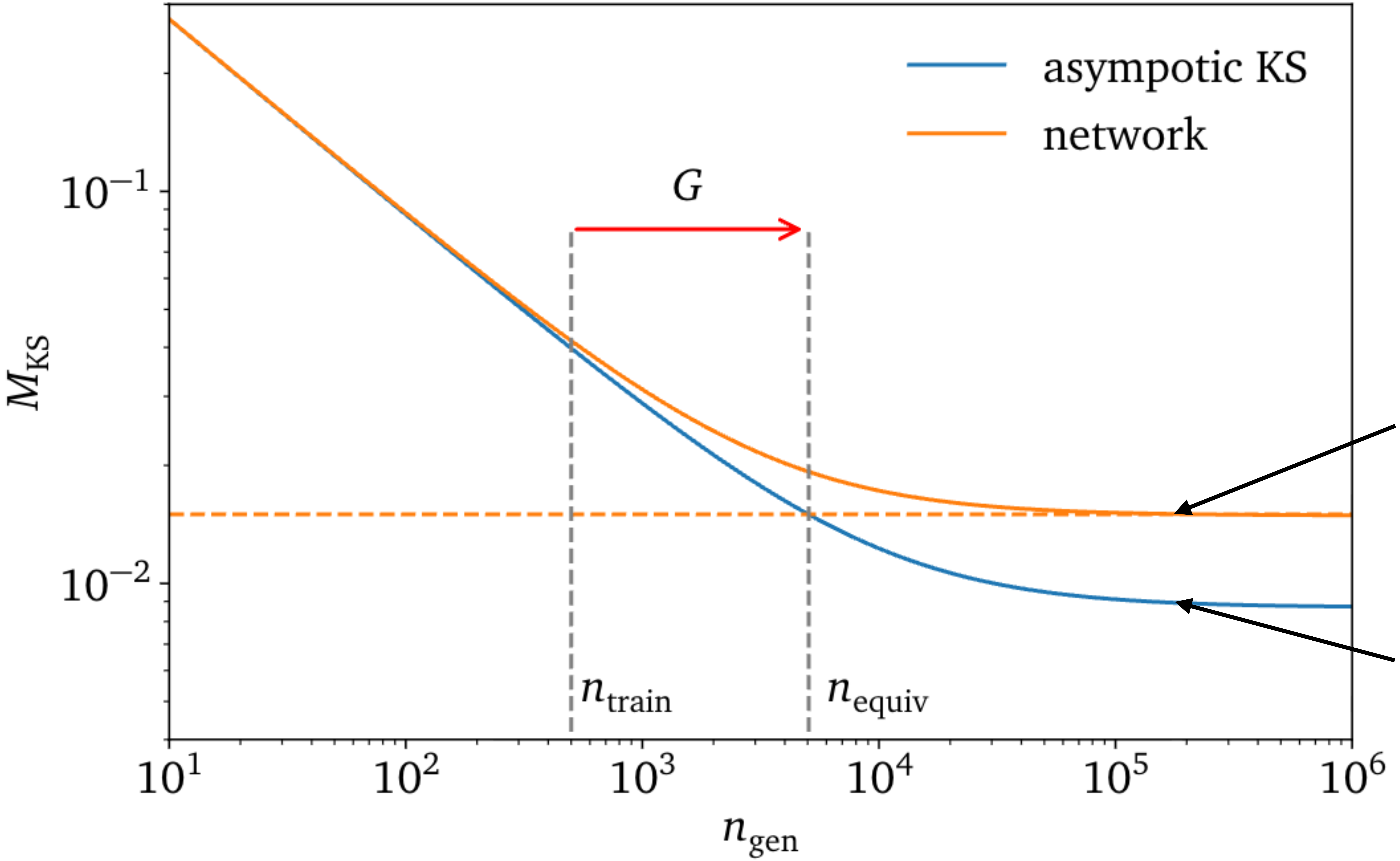
- Use two sample Kolmogorov Smirnov test
- Has analytical expectation value

$$\langle M_{\text{KS}}\left[D^{n}_{\text{true}}, D^{m}_{\text{true}}\right]\rangle = \sqrt{\frac{\pi}{2}}\log(2)\sqrt{\frac{n+m}{nm}}$$



Bahl et al. Forecasting Generative Amplification 2509.08048

Quantifying Amplification



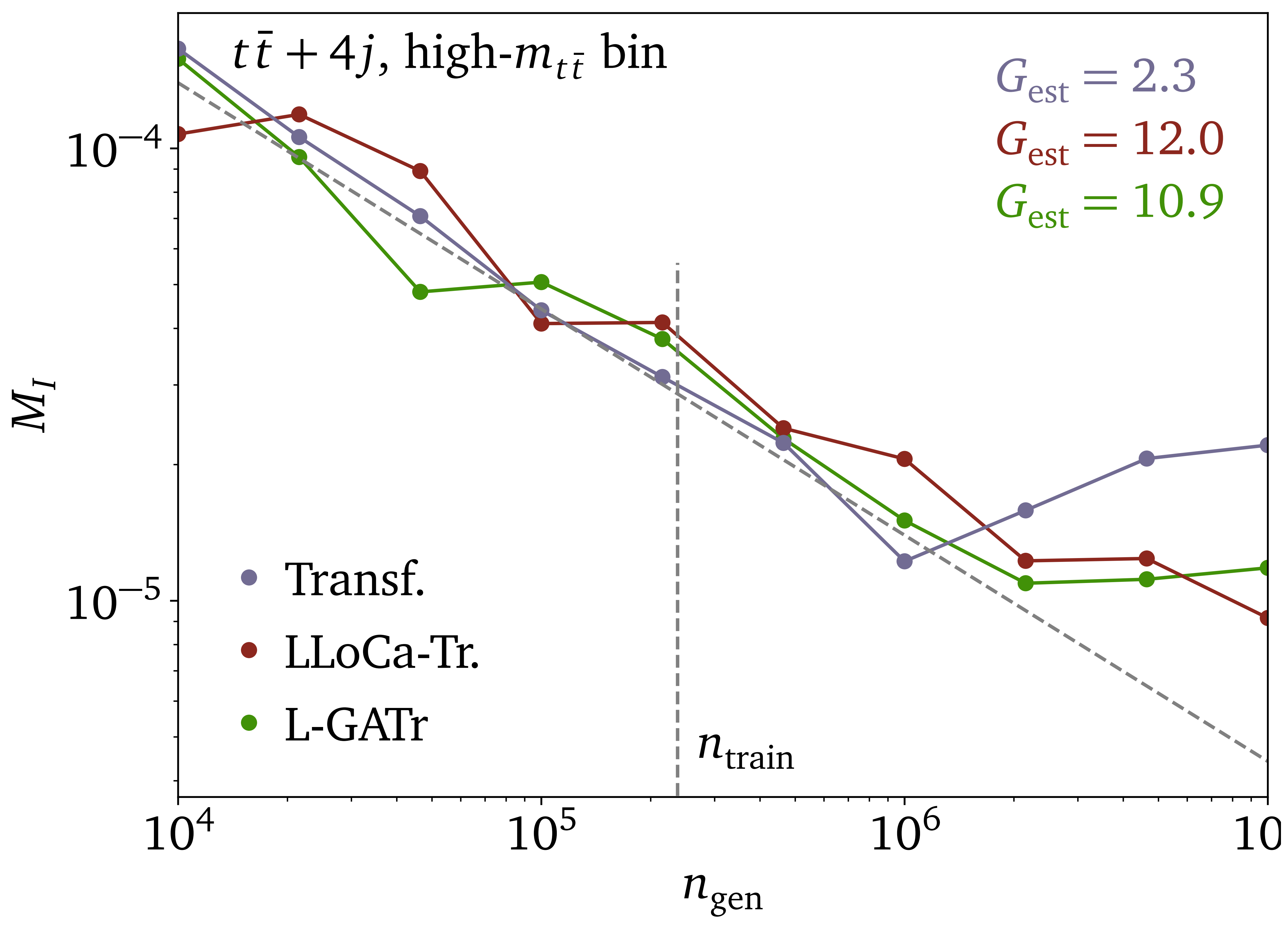
$$M_{KS}\left[D_{gen}^{n_{gen}}, D_{true}^{n_{train}}\right]$$

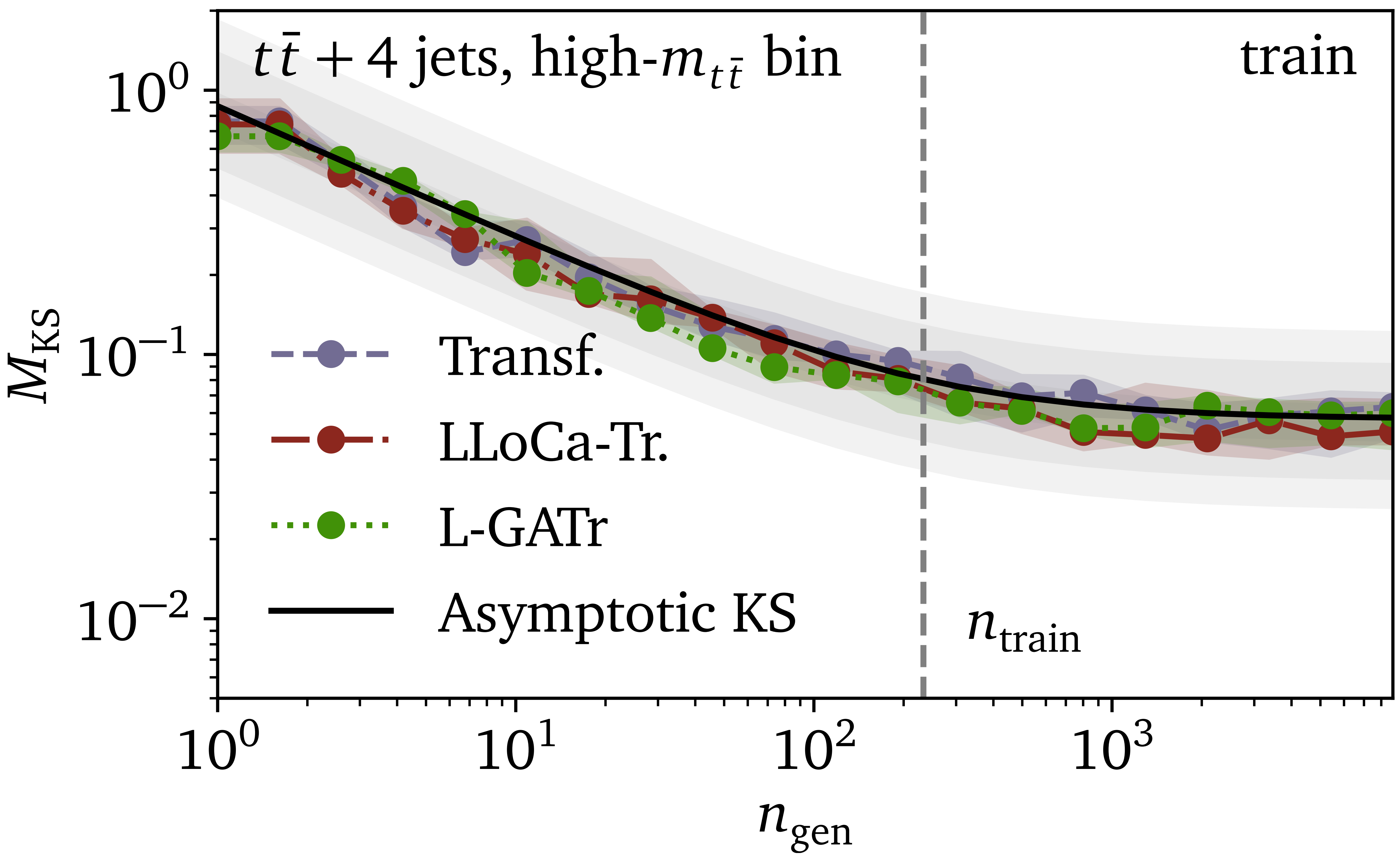
$$\sqrt{\frac{\pi}{2}}\log(2)\sqrt{\frac{n_{gen}+n_{train}}{n_{gen}n_{train}}}$$

Bahl et al. Forecasting Generative Amplification 2509.08048

Quantifying Amplification

Example: Di-Top + 4 jet events with both methods





Bahl et al. Forecasting Generative Amplification 2509.08048

Conclusion

- **Amplification Investigation**
  - Generative networks can sample beyond training data
  - Quantify by how much
  - Core to generative uncertainties
- **Enables generative approaches in HEP**
  - Amplification key for generative simulations
  - Simulation-based inference, generative unfolding need uncertainties

Conclusion

- **ML methods bring unprecedented precision**
  - LHC environment highly complex
  - $t\bar{t}H$, CP violation measurements impossible without
  - Directly enabled by ML

- **Fundamental physics understanding**
  - $t\bar{t}H$: early universe Higgs potential
  - CP violation: matter-antimatter asymmetry, baryogenesis

Thank You