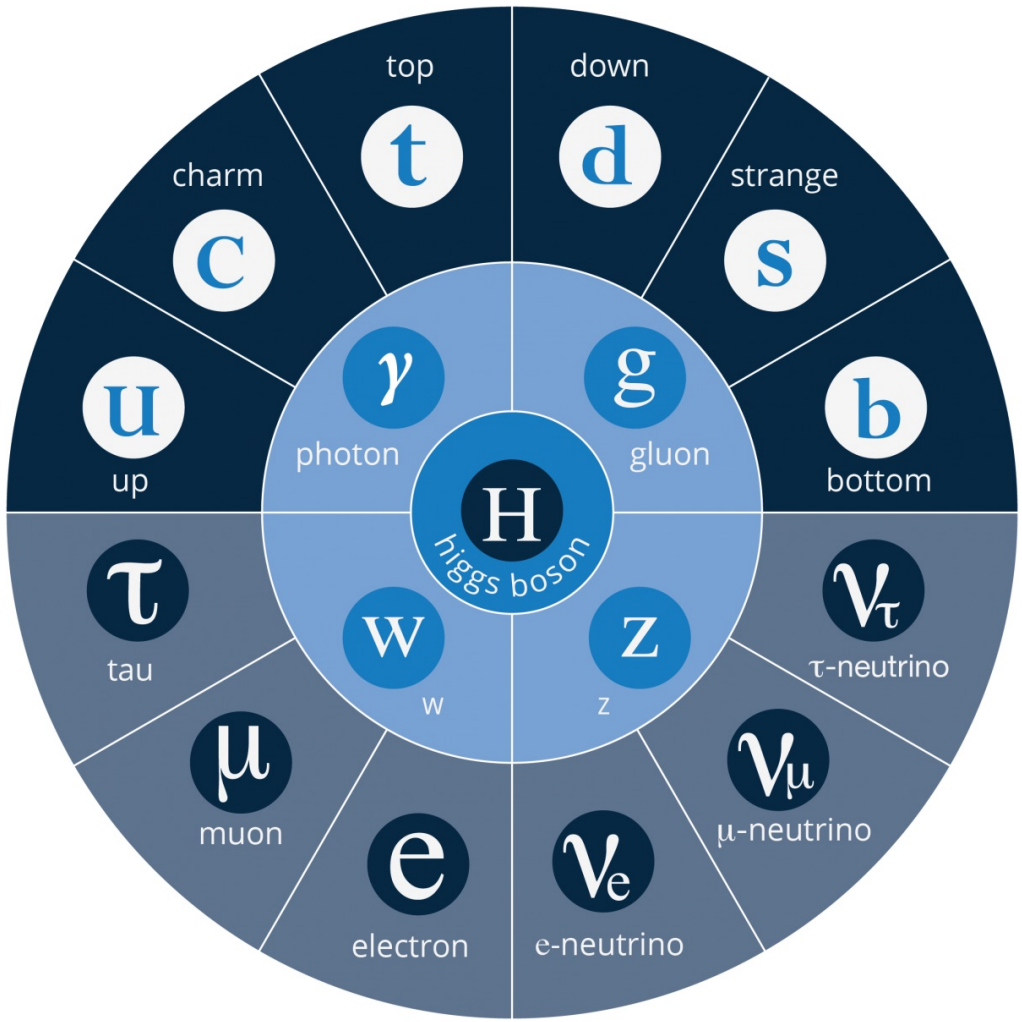
Simulating and unfolding LHC events with generative networks

ATLAS joint Statistics Forum & Machine Learning meeting

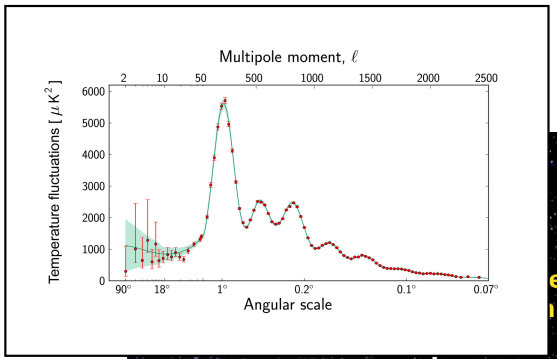
Anja Butter

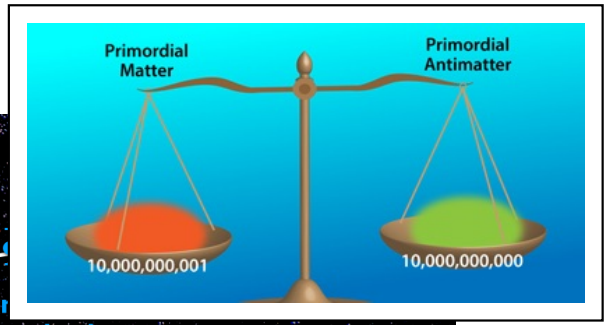
ITP, Universität Heidelberg

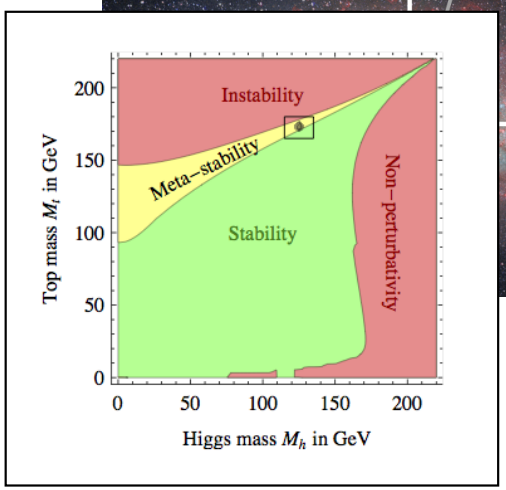
A structurally complete theory

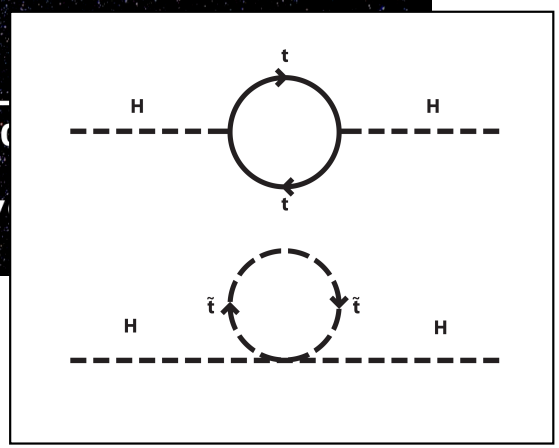


The need for new physics

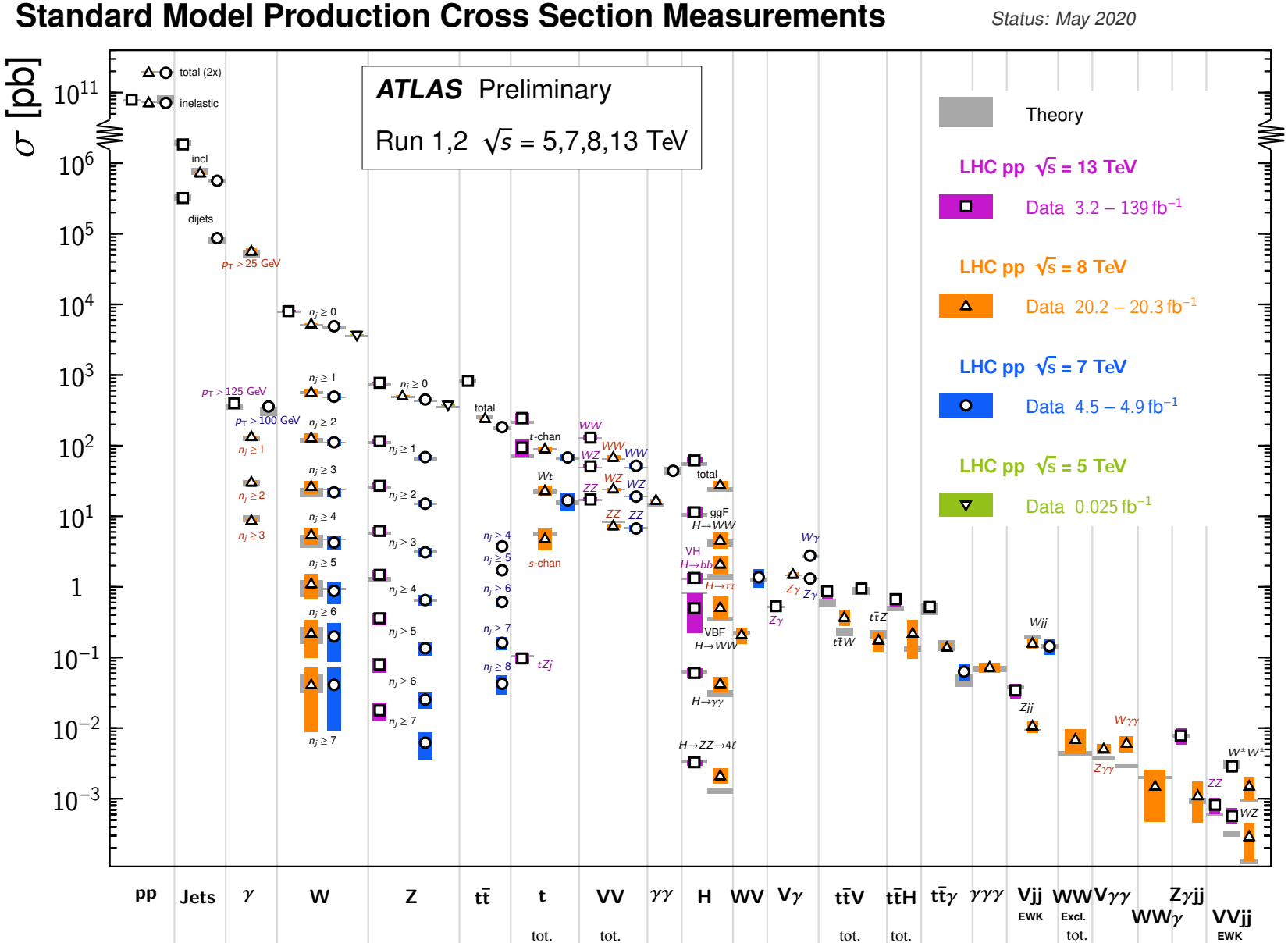




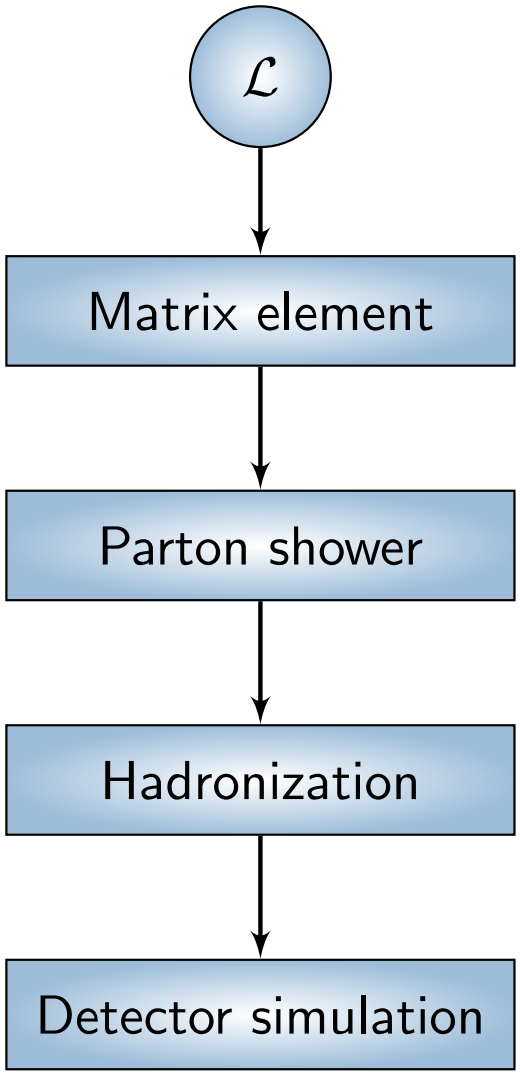


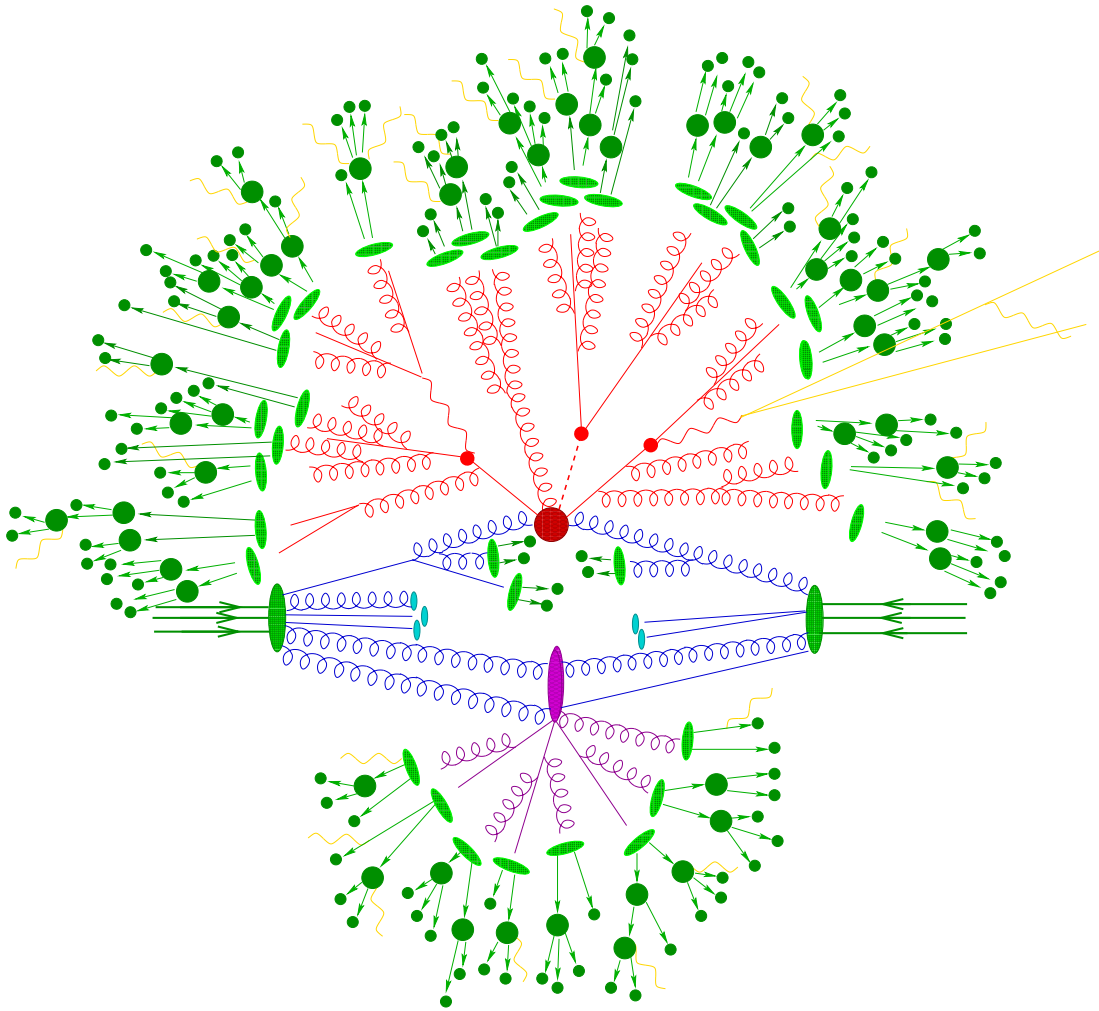


Era of data


Standard Model Production Cross Section Measurements
Status: May 2020
ATLAS Preliminary
Run 1,2 √s = 5,7,8,13 TeV
σ [pb]
Theory
LHC pp √s = 13 TeV
Data 3.2 – 139 fb⁻¹
LHC pp √s = 8 TeV
Data 20.2 – 20.3 fb⁻¹
LHC pp √s = 7 TeV
Data 4.5 – 4.9 fb⁻¹
LHC pp √s = 5 TeV
Data 0.025 fb⁻¹
pp
Jets
γ
W
Z
tt̄
t
VV
γγ
H
WV
Vγ
tt̄V
tt̄H
tt̄γ
γγγ
Vjj EWK
WW Excl.
Vγγ
Zγjj
WWγ
VVjj EWK

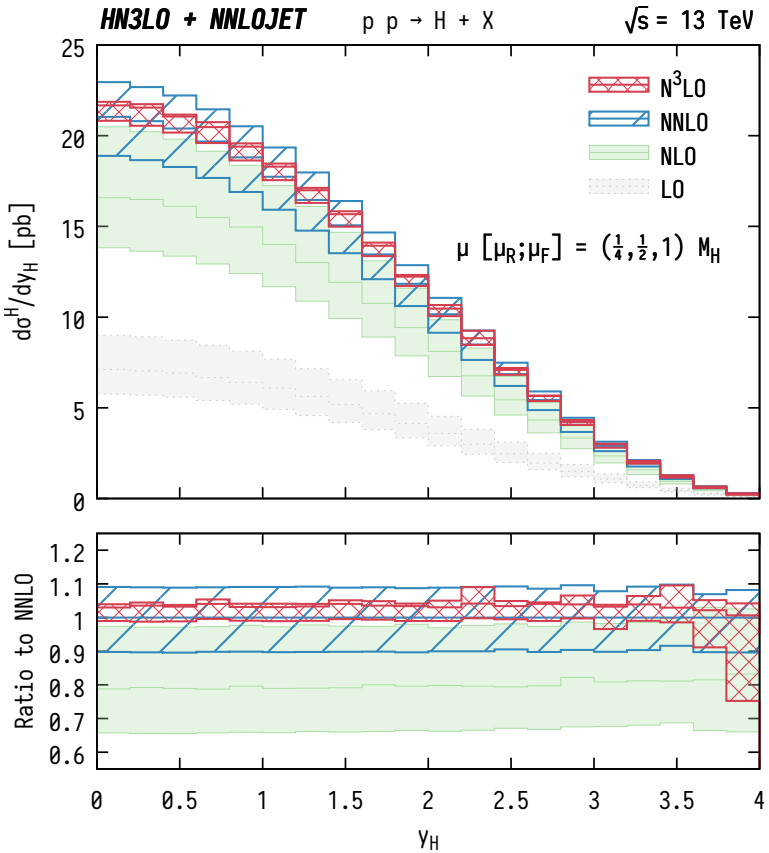
First principle based event generation





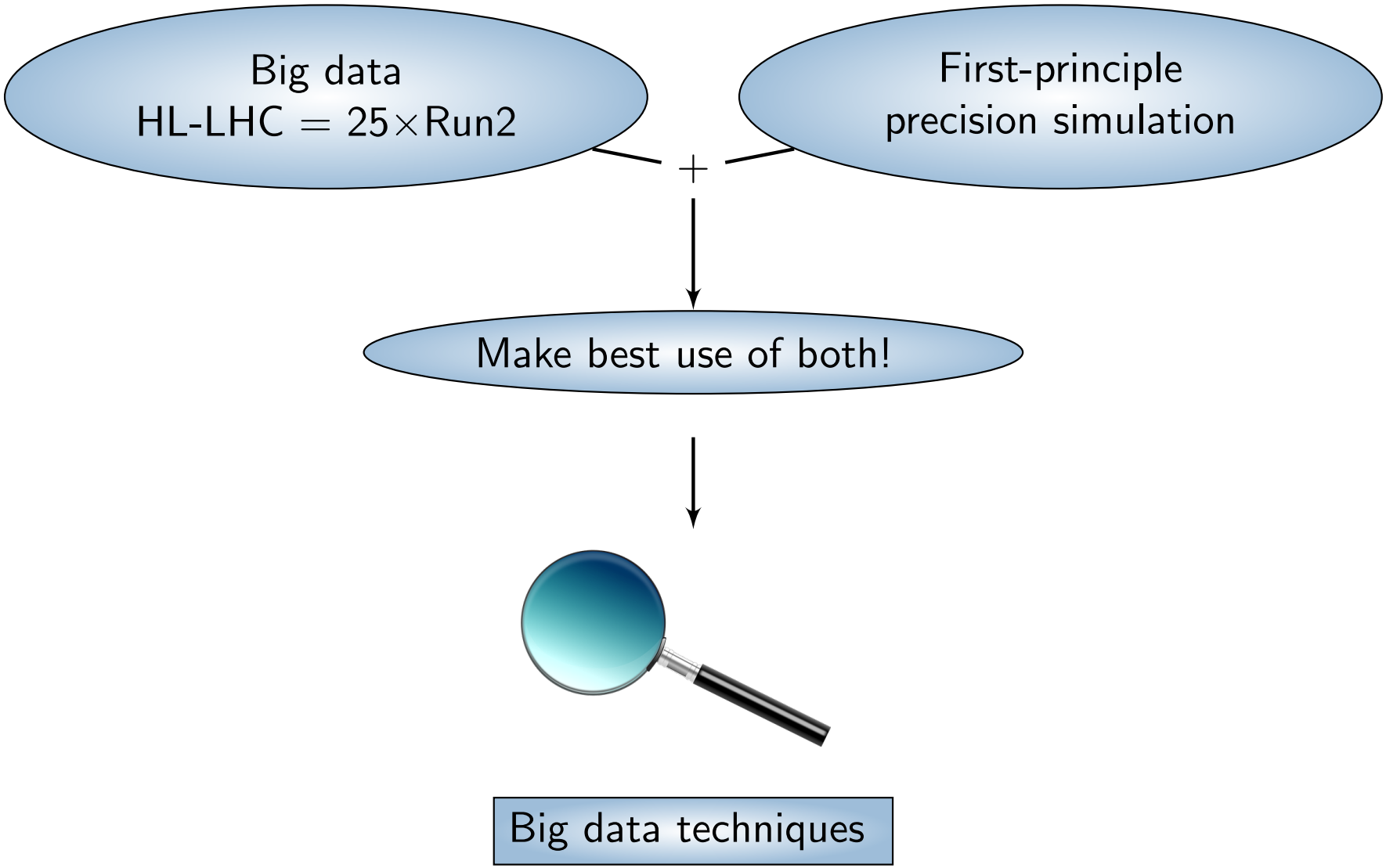
a sherpa artist

Precision simulations



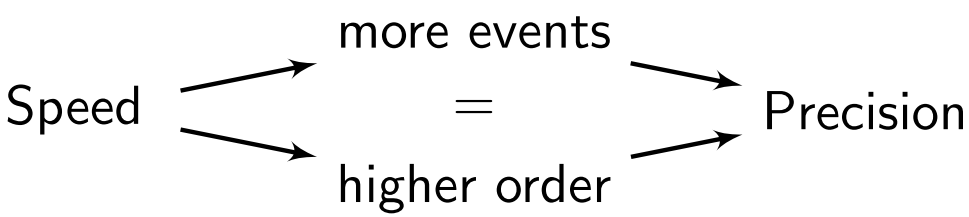
[1807.11501] Cieri, Chen, Gehrmann, Glover, Huss

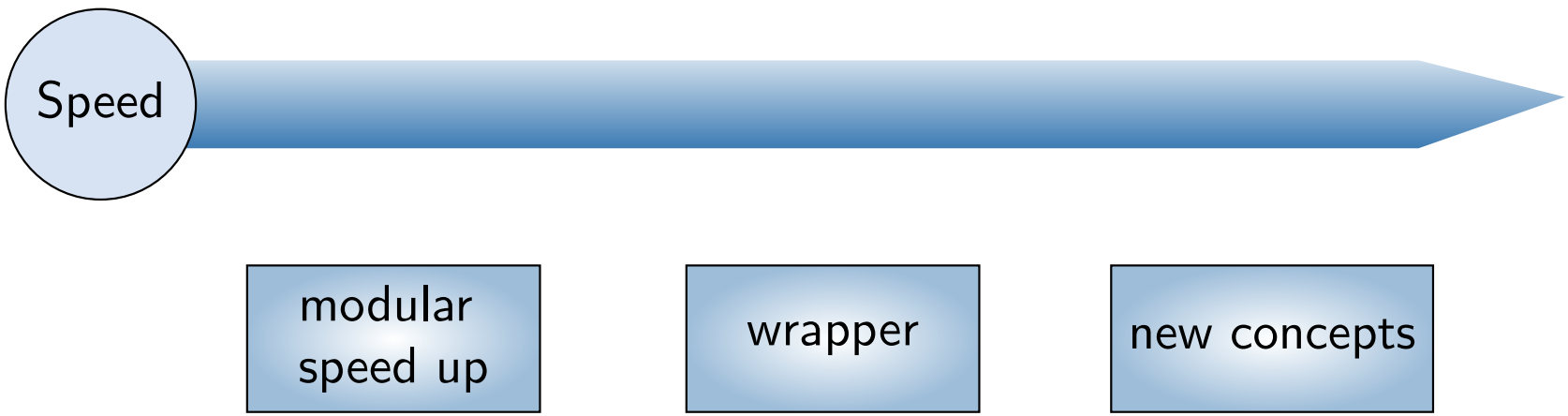
New physics is hidden



Precision in forward simulations

- ML 2.0 Generative models
  → Can we simulate new data?





Boosting standard event generation...

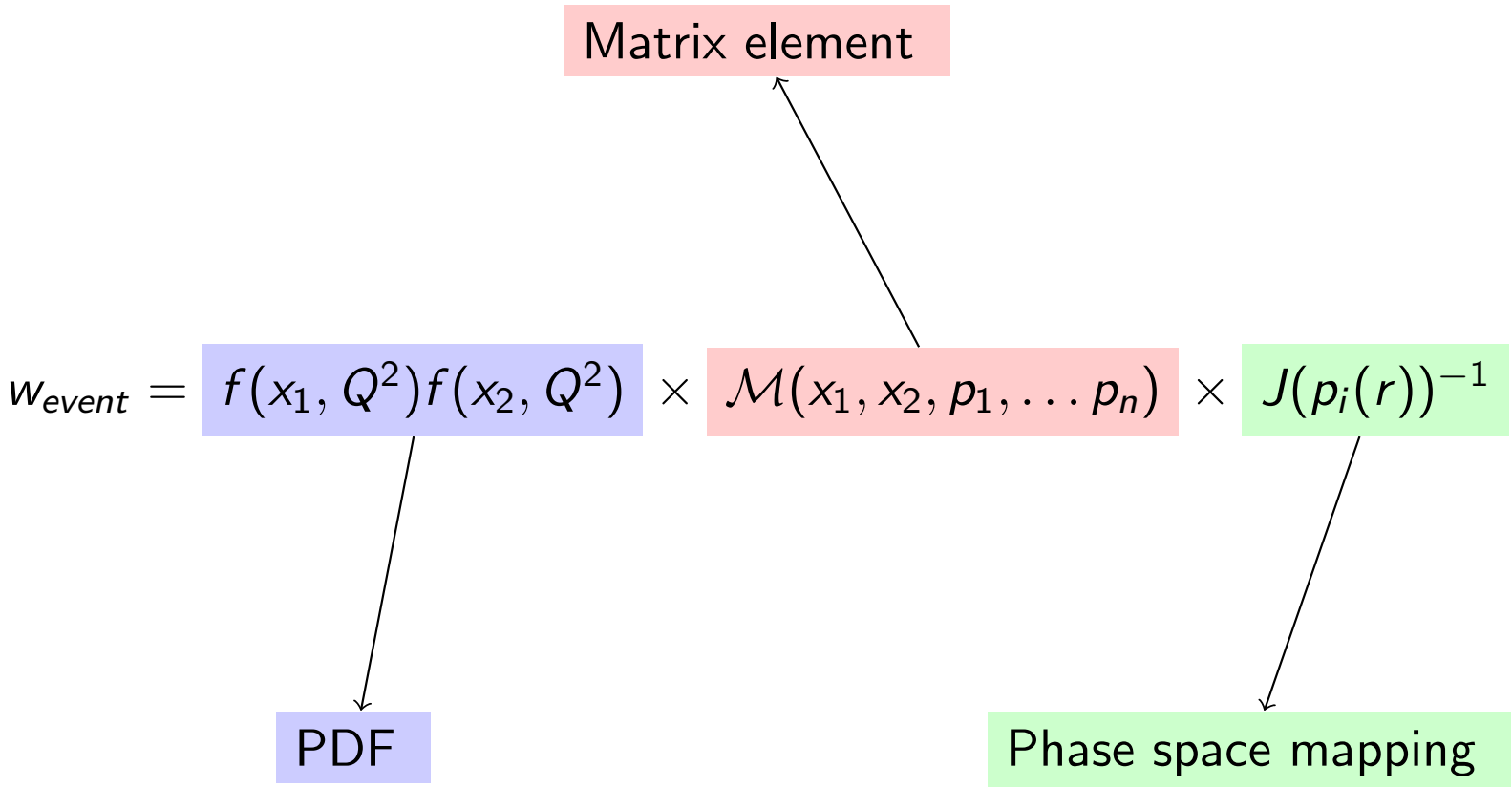
1. Generate phase space points

2. Calculate event weight

$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$

3. Unweighting via importance sampling
$\rightarrow$ optimal for $w \approx 1$

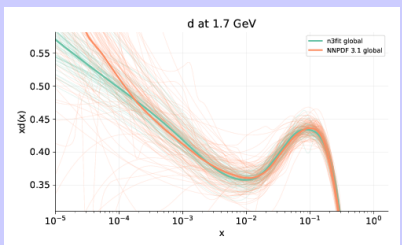
Boosting standard event generation...



Boosting standard event generation...

Matrix element

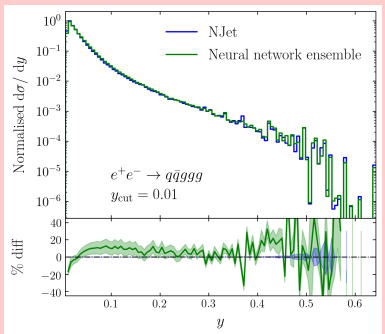
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



- NNPDF since 2002(!)
- S. Carrazza, J. Cruz-Martinez [1907.05075]

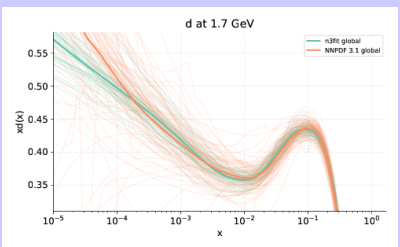
Phase space mapping

Boosting standard event generation...



- Amplitude estimation
- S. Badger, J. Bullock [2002.07516]
- J. Bendavid [1707.00028]

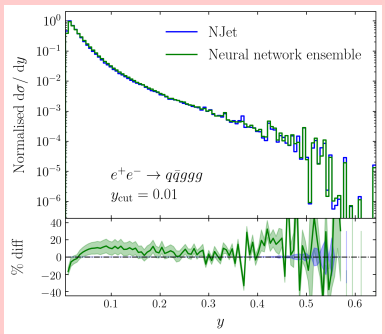
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



- NNPDF since 2002(!)
- S. Carrazza, J. Cruz-Martinez [1907.05075]

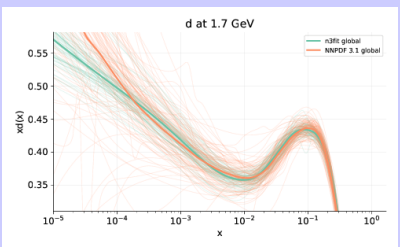
Phase space mapping

Boosting standard event generation...

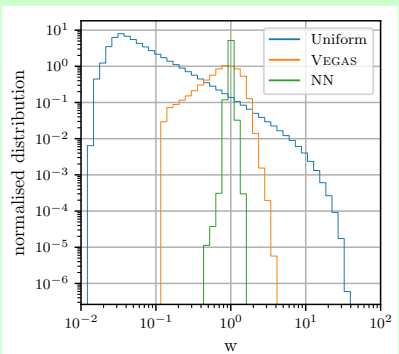


- Amplitude estimation
- S. Badger, J. Bullock [2002.07516]
- J. Bendavid [1707.00028]

$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



- NNPDF since 2002(!)
- S. Carrazza, J. Cruz-Martinez [1907.05075]



- Learn phase space mapping ($\to w \approx 1$)
- Gao et al. [2001.10028]
- Bothmann et al. [2001.05478]

... or training directly on event samples

Event generation

- Generating 4-momenta
- $Z > ll$, $pp > jj$, $pp > t\bar{t}$+decay

[1901.00875] Otten et al. **VAE** & **GAN**

[1901.05282] Hashemi et al. **GAN**

[1903.02433] Di Sipio et al. **GAN**

[1903.02556] Lin et al. **GAN**

[1907.03764, 1912.08824] Butter et al. **GAN**

[1912.02748] Martinez et al. **GAN**

[2001.11103] Alanazi et al. **GAN**

[2011.13445] Stienen et al. **NF**

[2012.07873] Backes et al. **GAN**

[2101.08944] Howard et al. **VAE**

Detector simulation

- Jet images
- Fast calorimeter simulation

[1701.05927] de Oliveira et al. **GAN**

[1705.02355, 1712.10321] Paganini et al. **GAN**

[1802.03325, 1807.01954] Erdmann et al. **GAN**

[1805.00850] Musella et al. **GAN**

[ATL-SOFT-PUB-2018-001, ATLAS-SIM-2019-004, ATL-SOFT-PROC-2019-007] ATLAS **VAE** & **GAN**

[1909.01359] Carazza and Dreyer **GAN**

[1912.06794] Belayneh et al. **GAN**

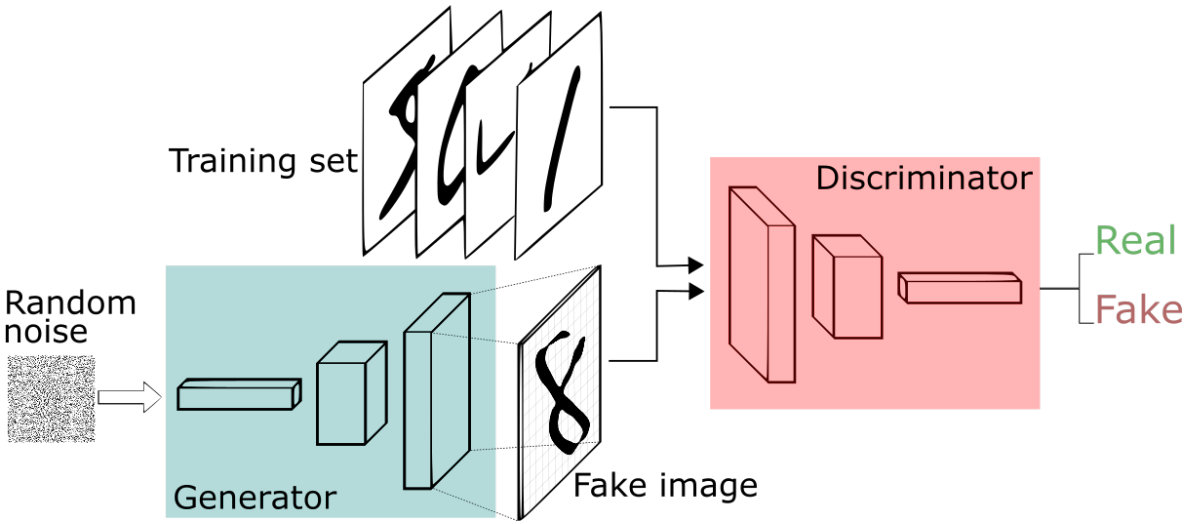
[2005.05334] Buhmann et al. **VAE**

[2009.03796] Diefenbacher et al. **GAN**

[2009.14017] Lu et al.

NO claim to completeness!

Generative Adversarial Networks



Discriminator $[D(x_T) \to 1, D(x_G) \to 0]$

$$L_D = \left\langle -\log D(x) \right\rangle_{x \sim P_{Truth}} + \left\langle -\log(1 - D(x)) \right\rangle_{x \sim P_{Gen}} \to -2 \log 0.5$$

Generator $[D(x_G) \to 1]$

$$L_G = \left\langle -\log D(x) \right\rangle_{x \sim P_{Gen}}$$

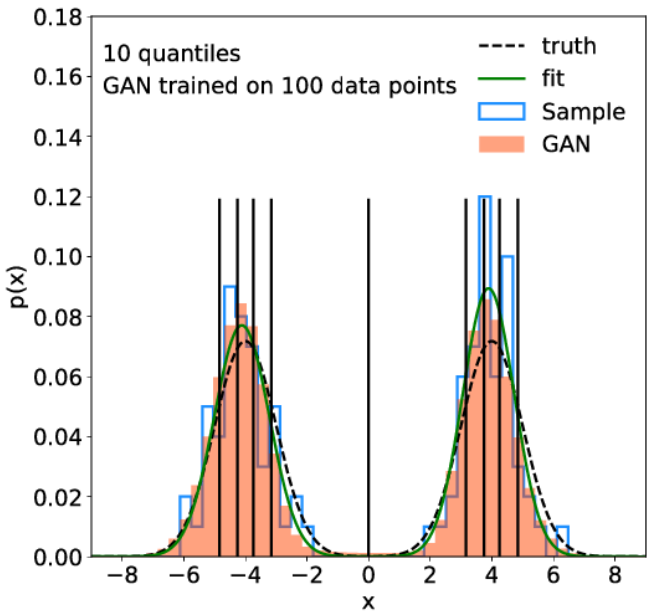
$\Rightarrow$ **New statistically independent samples**

What is the statistical value of GANned events? [2008.06545]

- Camel function
- Sample vs. GAN vs. 5 param.-fit

Evaluation on quantiles:

$$\text{MSE}^* = \sum_{j=1}^{N_\text{quant}} \left(p_j - \frac{1}{N_\text{quant}} \right)^2$$

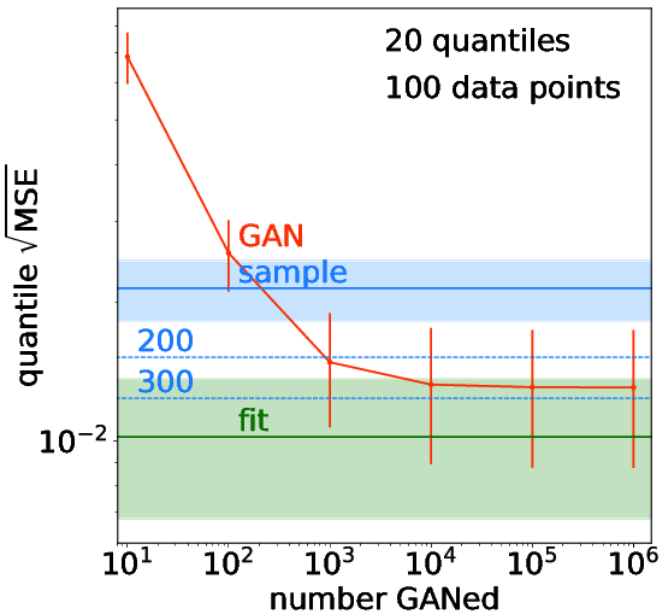


What is the statistical value of GANned events? [2008.06545]

- Camel function
- Sample vs. GAN vs. 5 param.-fit

Evaluation on quantiles:

$$\text{MSE}^* = \sum_{j=1}^{N_{\text{quant}}} \left(p_j - \frac{1}{N_{\text{quant}}} \right)^2$$

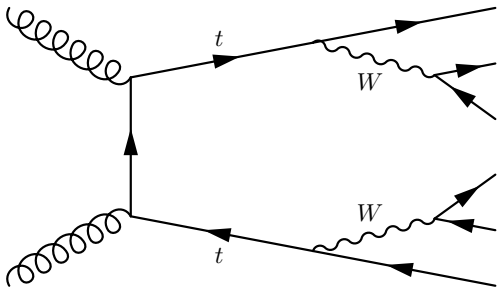


$\rightarrow$ Amplification factor 2.5

Sparser data $\rightarrow$ bigger amplification

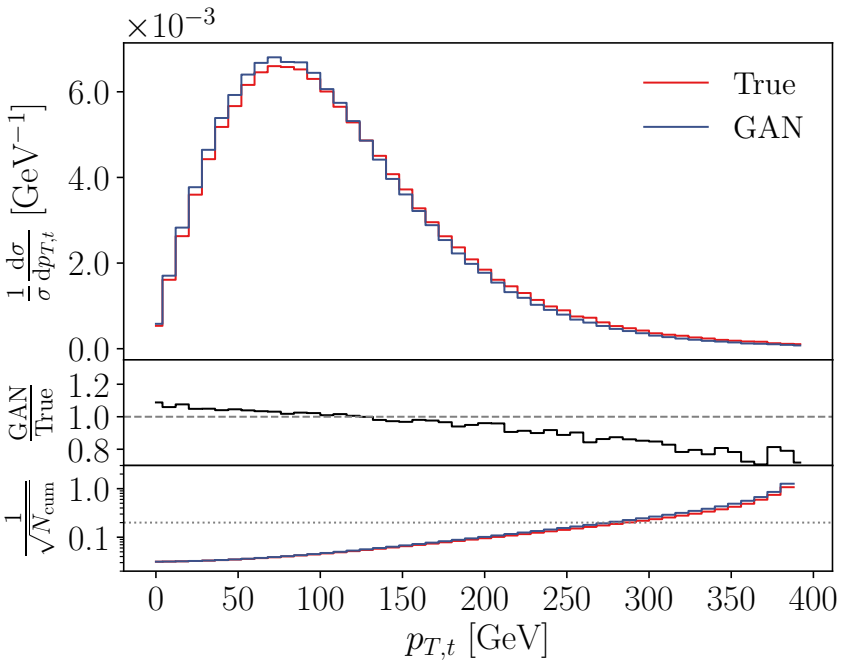
How to GAN LHC events [1907.03764]

- $t\bar{t} \to 6$ quarks
- 18 dim output
  - external masses fixed
  - no momentum conservation

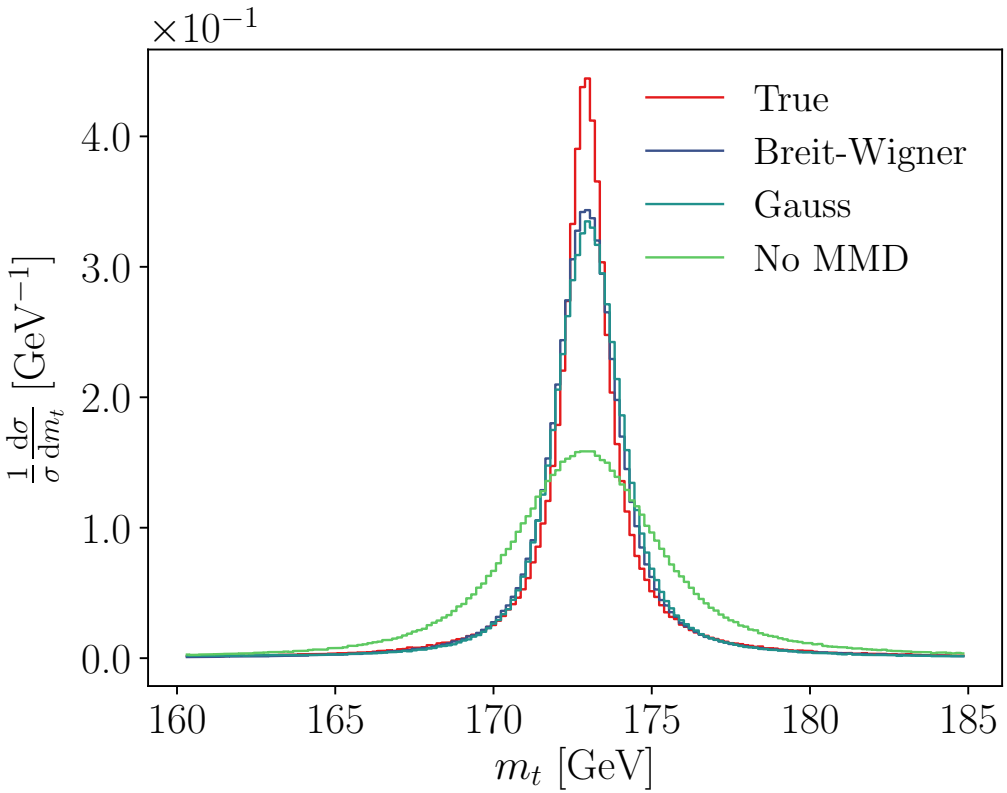


\+ Flat observables ✓

– Systematic undershoot in tails [10-20% deviation]



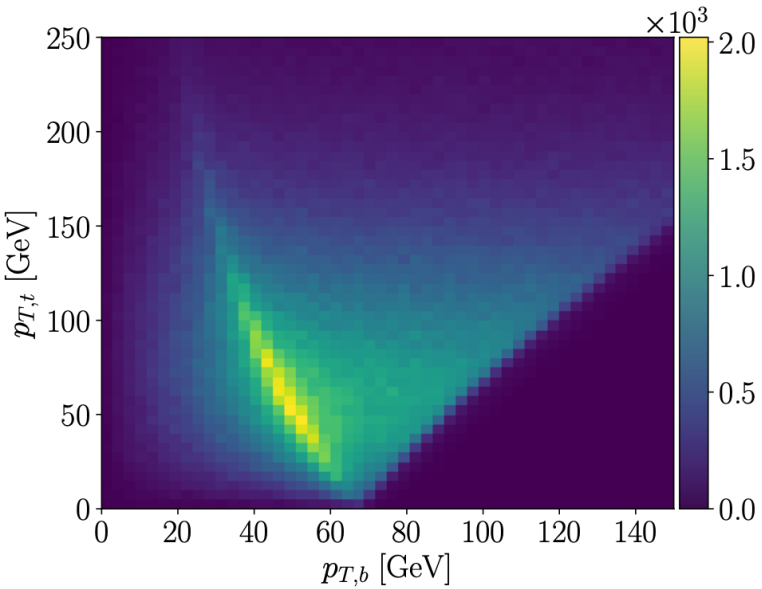
Special features

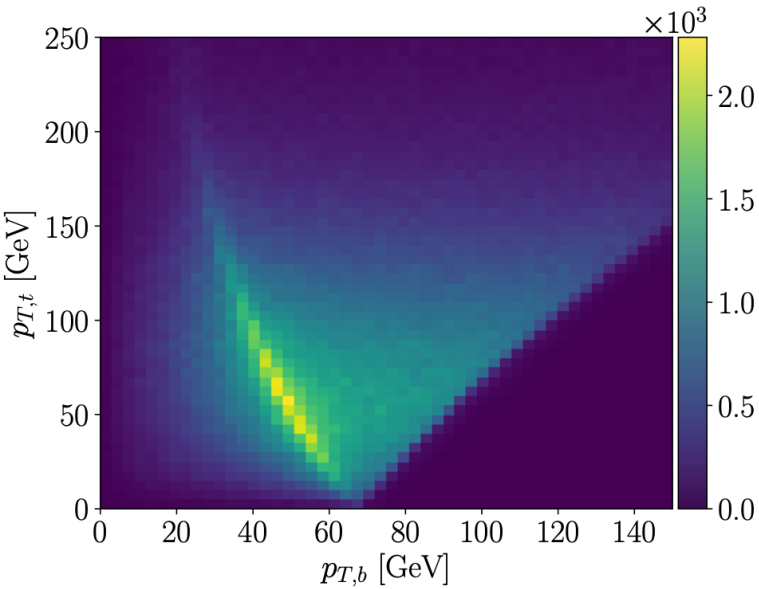


Solution: MMD kernel

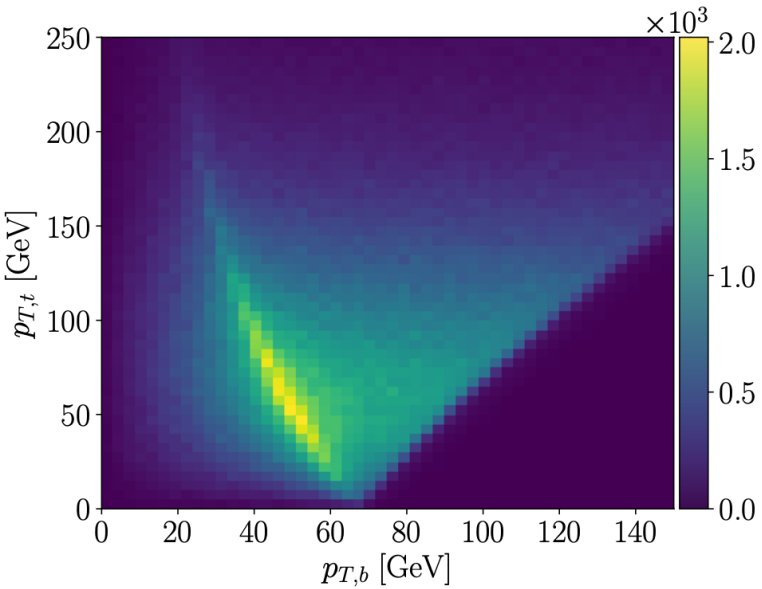
$$\mathrm{MMD}^2(P_T, P_G) = \big\langle k(x,x') \big\rangle_{x,x' \sim P_T} + \big\langle k(y,y') \big\rangle_{y,y' \sim P_G} - 2\big\langle k(x,y) \big\rangle_{x \sim P_T, y \sim P_G}$$

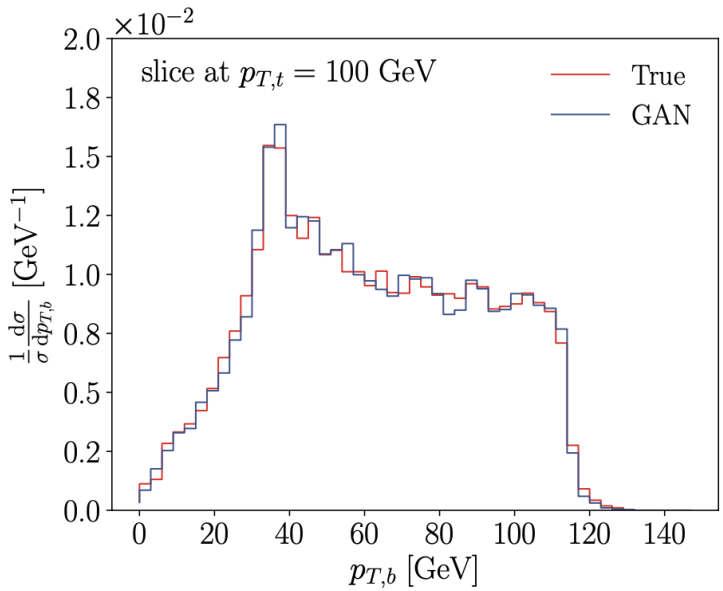
Correlations





Correlations





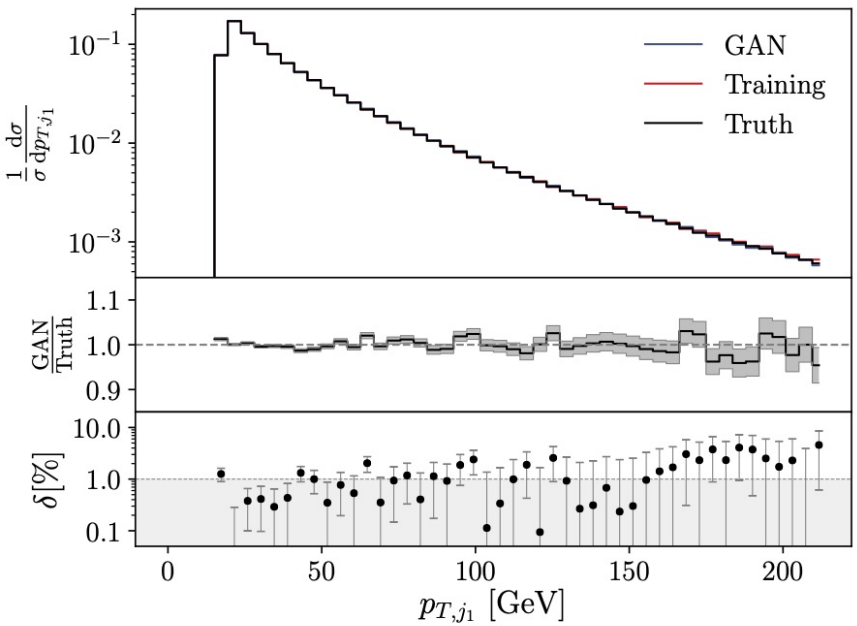
Reaching precision (preliminary)

1. Representation p_T, η, ϕ
2. Momentum conservation
3. Resolve $\log p_T$
4. Regularization: spectral norm
5. Batch information

$\rightarrow$ 1% precision ✓

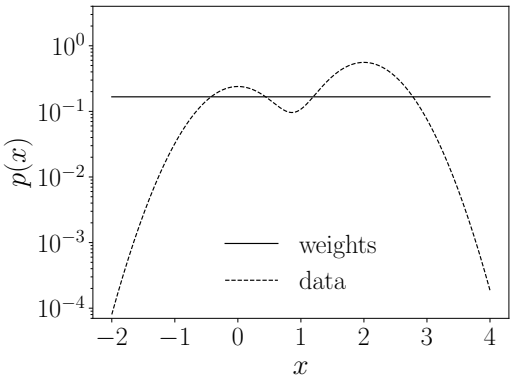
Automization?

W + 2 jets

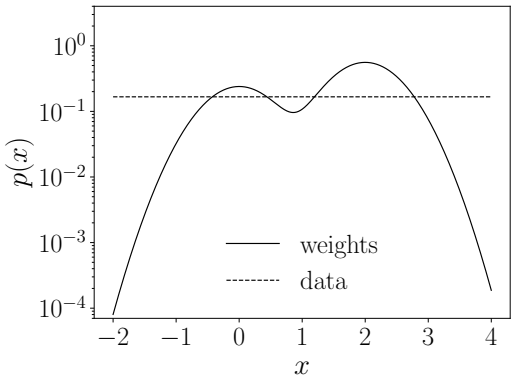


Training on weighted events

Information contained in distribution or event weights

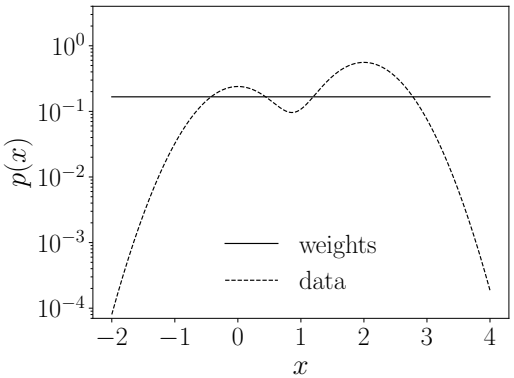


Train on
weighted events

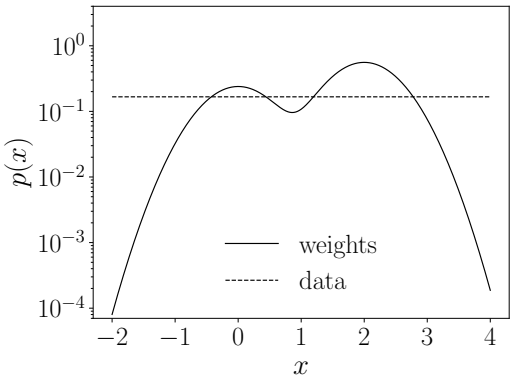


Training on weighted events

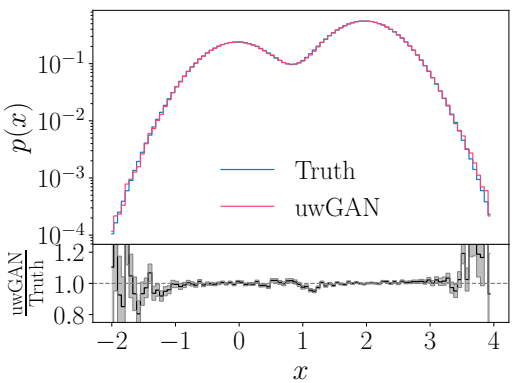
Information contained in distribution or event weights

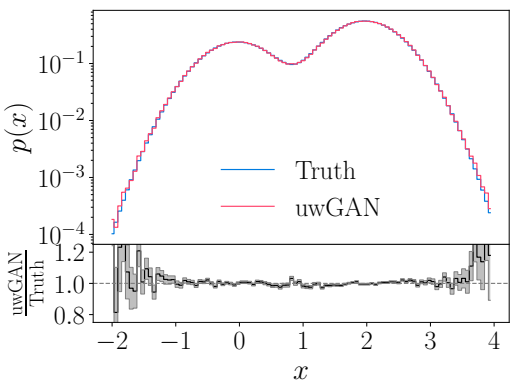


Train on
weighted events



Generate
unweighted events

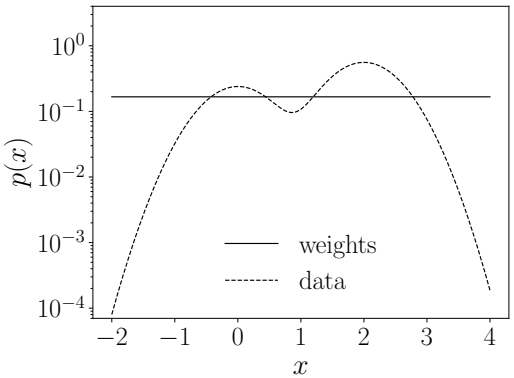




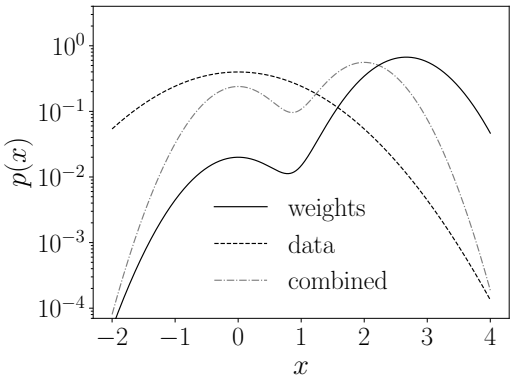
$$L_D = \left\langle -w \log D(x) \right\rangle_{x \sim P_{Truth}} + \left\langle -\log(1 - D(x)) \right\rangle_{x \sim P_{Gen}}$$

Training on weighted events

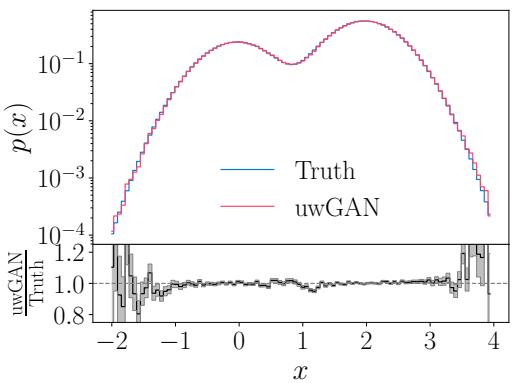
Information contained in distribution or event weights

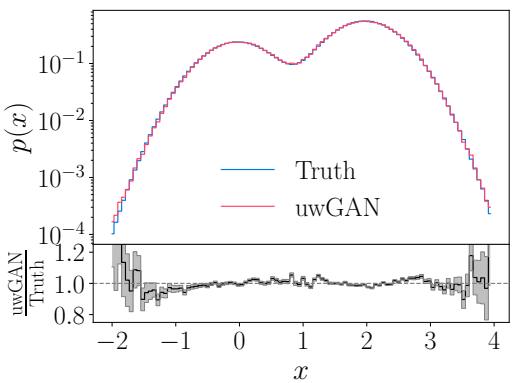


Train on
weighted events



Generate
unweighted events





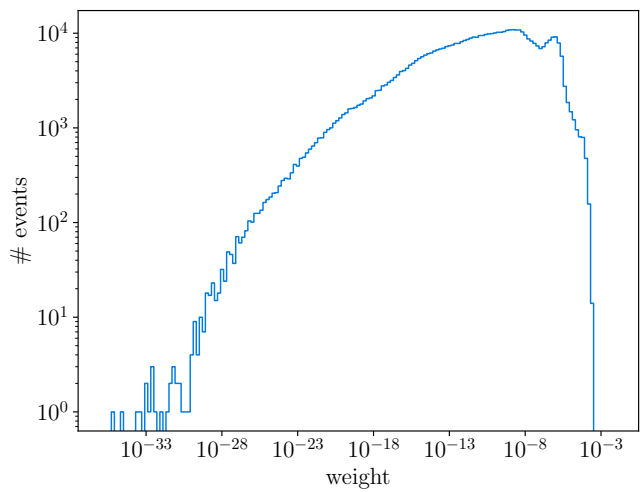
$$L_D = \left\langle -w \log D(x) \right\rangle_{x \sim P_{Truth}} + \left\langle -\log(1 - D(x)) \right\rangle_{x \sim P_{Gen}}$$

The unweighting bottleneck

- High-multiplicity processes & higher-order calculations
  $\rightarrow$ unweighting efficiency below 1%

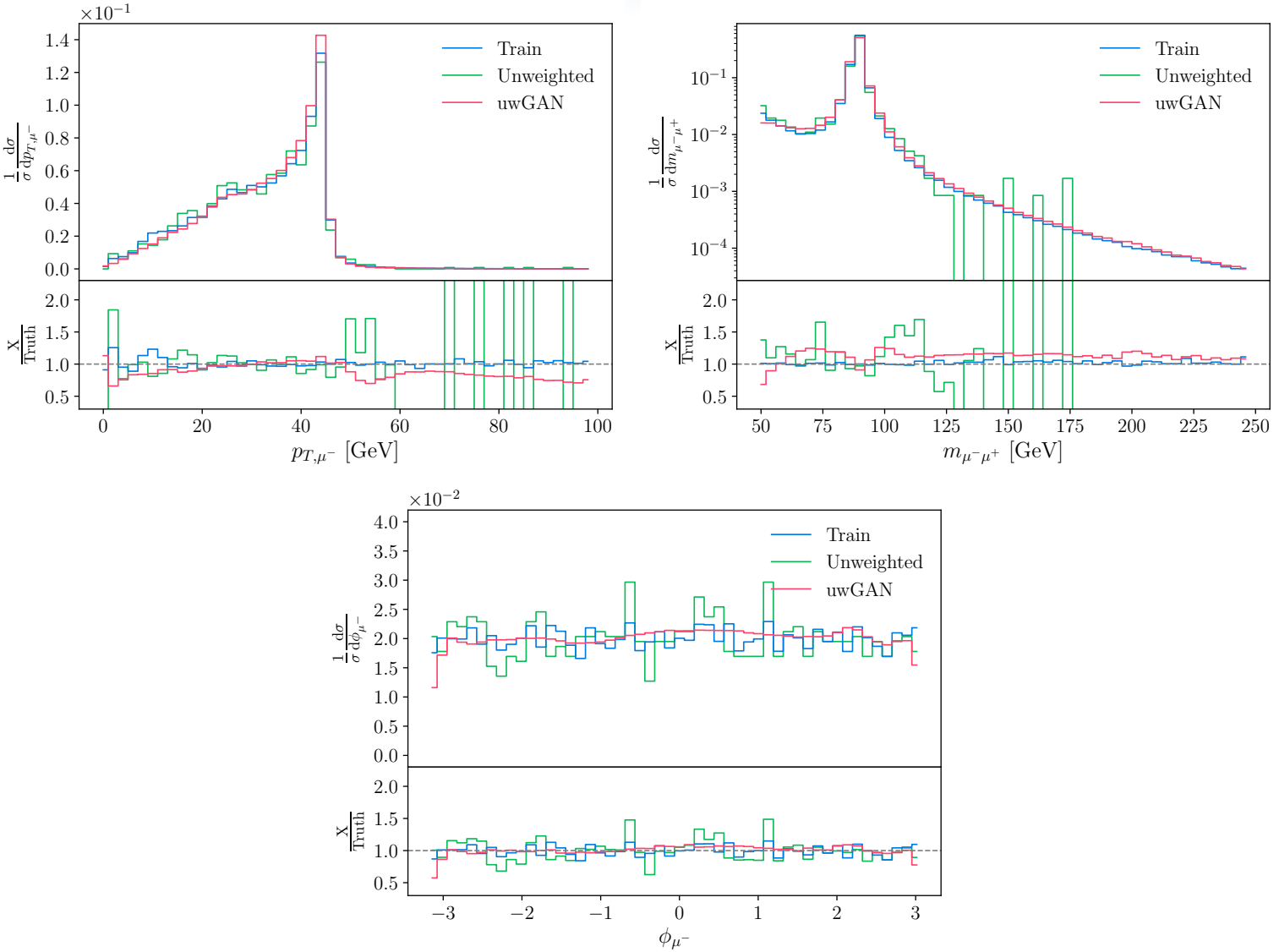
$\rightarrow$ Simulate conditions with naive Monte Carlo generator
[ME by `Sherpa`, parton densities from `LHAPDF`, Rambo-on-diet]

$pp \to \mu^+\mu^-$ with $m_{\mu\mu} > 50$ GeV



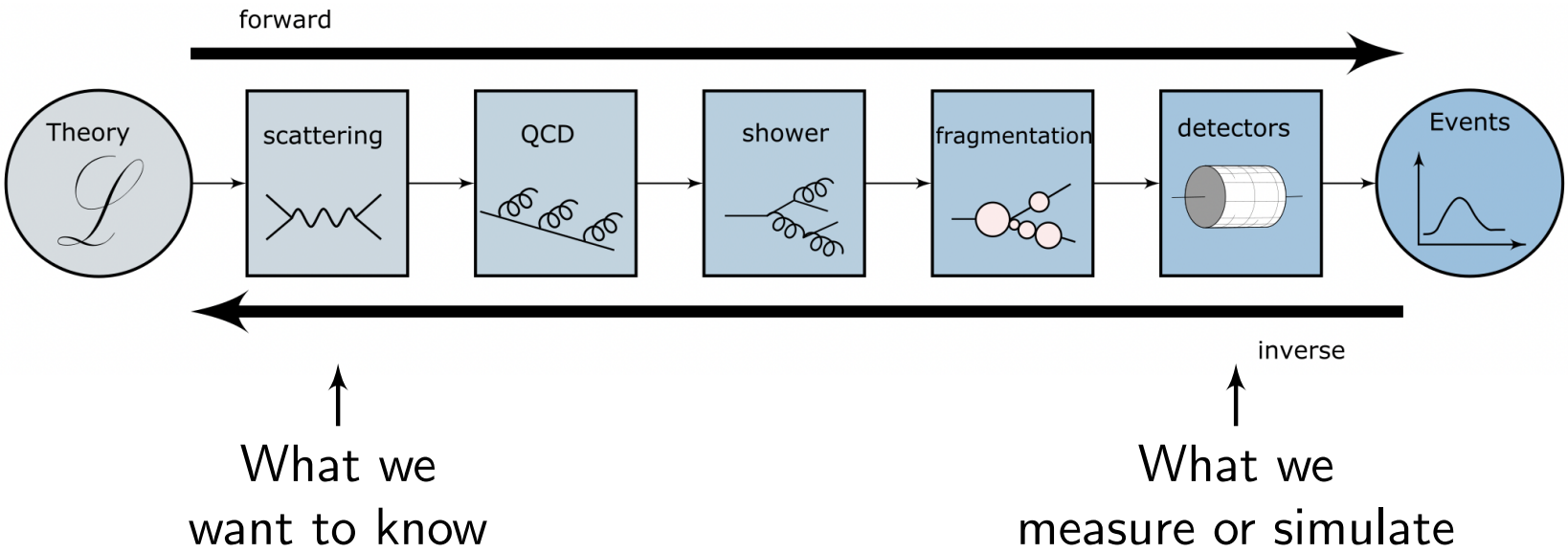
$\rightarrow$ unweighting efficieny $4 \cdot 10^{-3}$

uwGAN results

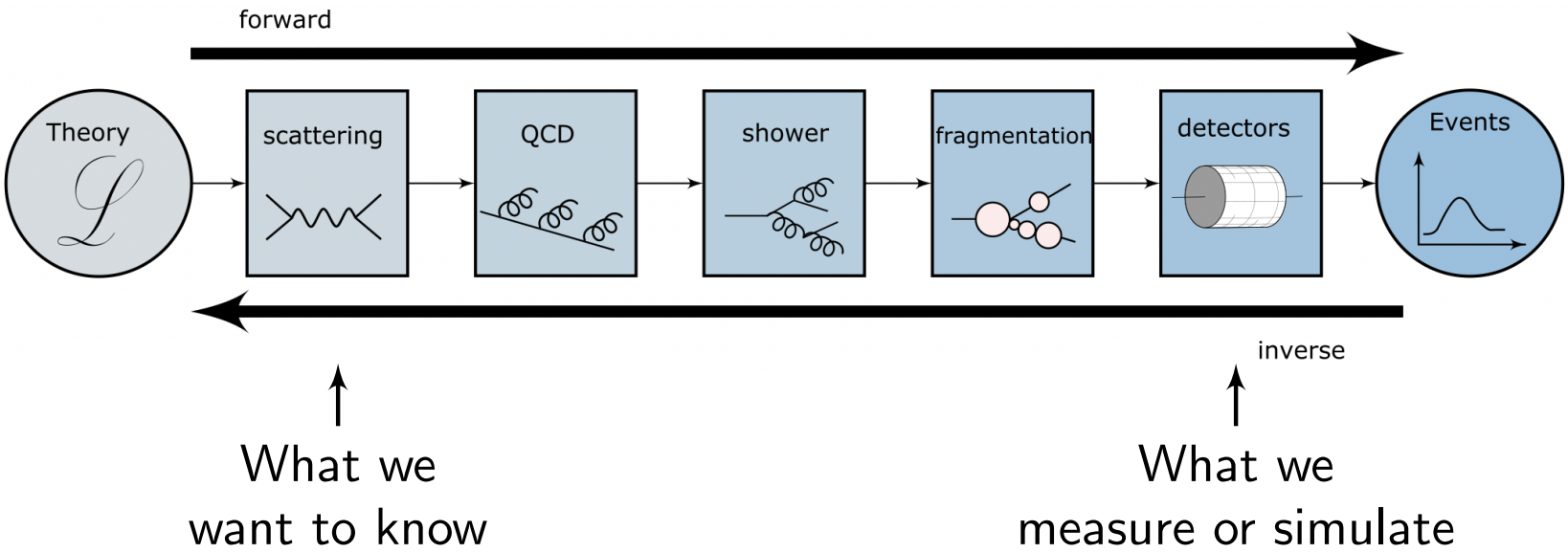


Large amplification factor

Can we invert the simulation chain?



Can we invert the simulation chain?

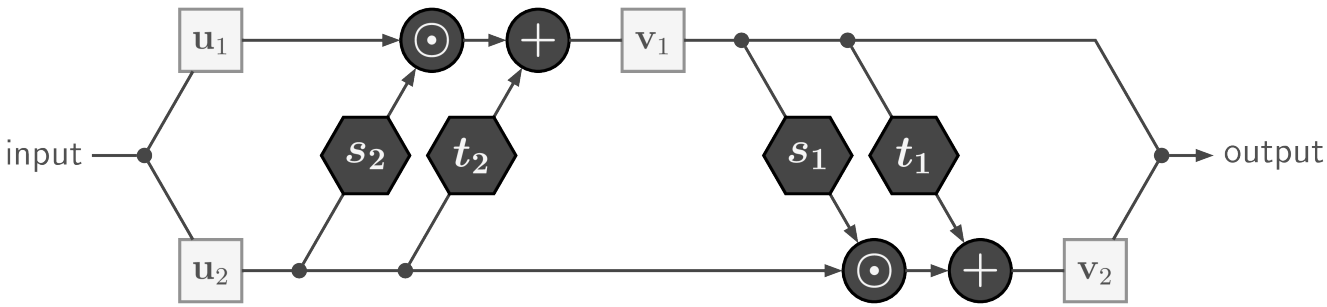


$\rightarrow$ Unfold high-dimensional distributions

$\rightarrow$ Get high-dimensional probability distribution at parton level

Invertible networks

$$(x_p) \xleftarrow[\leftarrow \text{unfolding:} \bar{g}]{\text{PYTHIA,DELPHES:} g \rightarrow} (x_d)$$



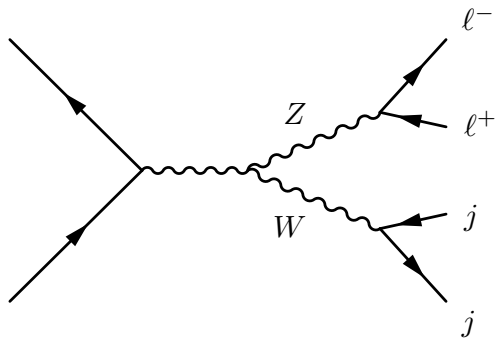
[1808.04730] L. Ardizzone, J. Kruse, S. Wirkert, D. Rahner,

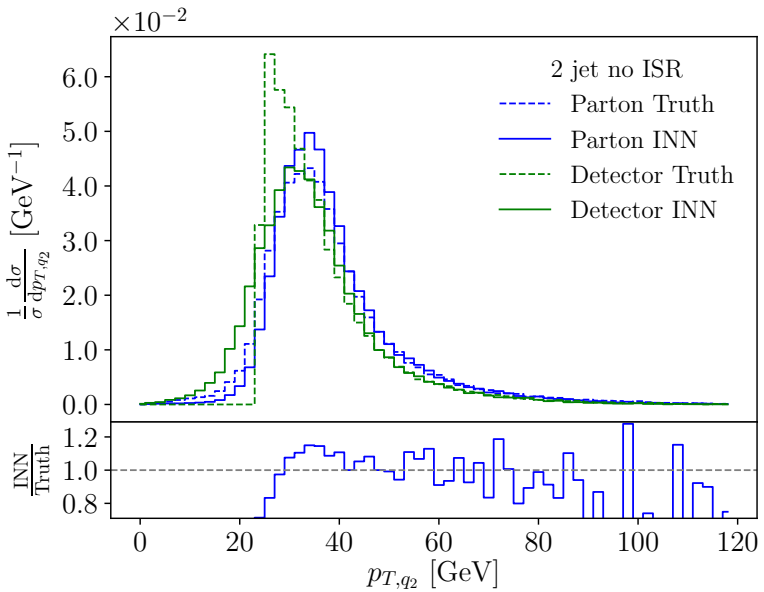
E. W. Pellegrini, R. S. Klessen, L. Maier-Hein, C. Rother, U. Köthe

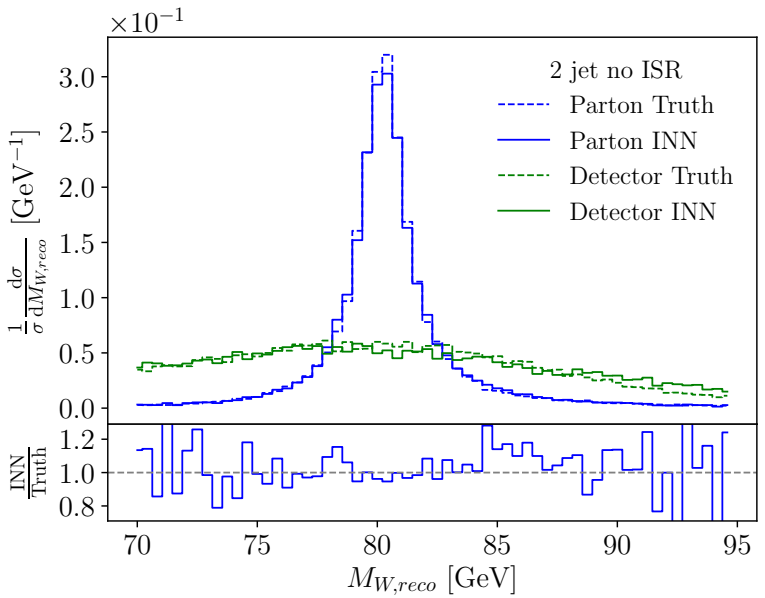
- Fast evaluation in both directions
- Tractable Jacobian
- Arbitrary networks s and t
- Equal input and output dimension

Inverting detector effects

- $pp \to ZW \to (ll)(jj)$
- Train parton $\to$ detector
- Evaluate detector $\to$ parton







Including stochastical effects

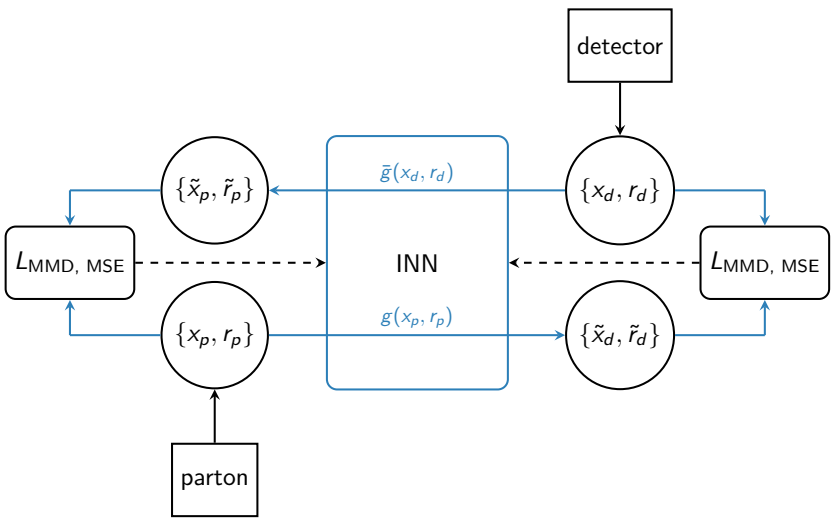
- So far: only mapping of mean values
- Noise extended INN to include probabilistic nature

$$\begin{pmatrix} x_p \\ r_p \end{pmatrix} \underset{\leftarrow \text{ unfolding:} \bar{g}}{\overset{\text{PYTHIA,DELPHES:} g \rightarrow}{\longleftrightarrow}} \begin{pmatrix} x_d \\ r_d \end{pmatrix}$$

- Training in both directions

→ Improved stability

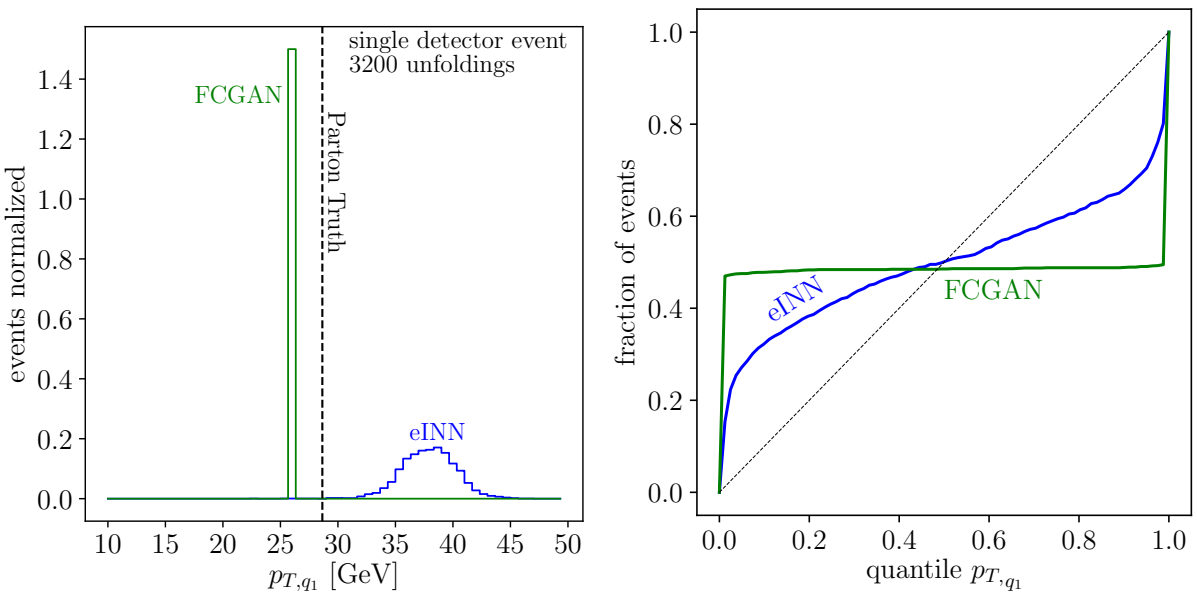
- MSE fixes mean values
- MMD fixes distributions



Calibration curves

$$\begin{pmatrix} x_p \\ r_p \end{pmatrix} \xleftarrow[\leftarrow \text{ unfolding:} \bar{g}]{\text{PYTHIA,DELPHES:} g \rightarrow} \begin{pmatrix} x_d \\ r_d \end{pmatrix}$$

Fix detector level event & sample over r_d
How often is Truth included in distribution quantile?

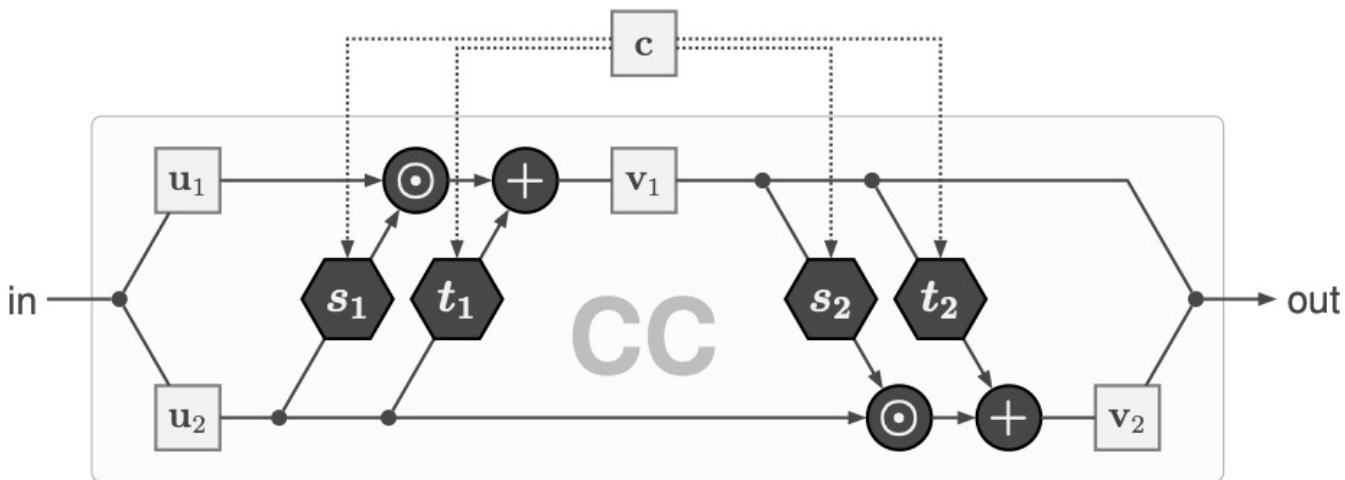


- Mean correct, distribution too narrow
- Problem: arbitrary balance of many loss functions

Taking a different angle

Given an event x_d, what is the probability distribution at parton level?
$\rightarrow$ sample over r, condition on x_d

$$x_p \underset{\leftarrow \text{ unfolding: } \bar{g}(r, f(x_d))}{\overset{g(x_p, f(x_d)) \rightarrow}{\longleftrightarrow}} r$$



Taking a different angle

Given an event x_d, what is the probability distribution at parton level?
$\rightarrow$ sample over r, condition on x_d

$$x_p \xleftarrow[\leftarrow \text{ unfolding: } \bar{g}(r, f(x_d))]{g(x_p, f(x_d))\rightarrow} r$$

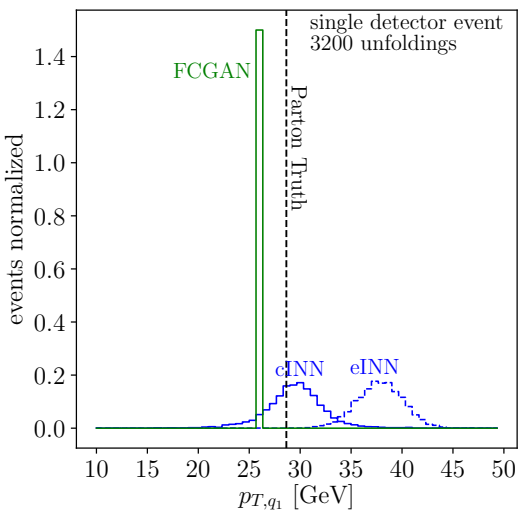
$\rightarrow$ Training: Maximize posterior over model parameters

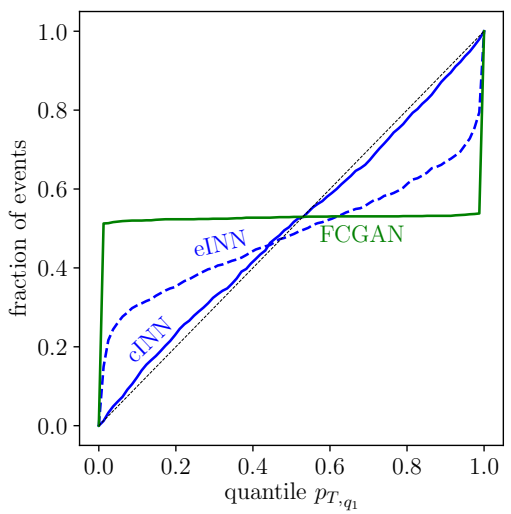
$$\begin{aligned}
L &= -\left\langle \log p(\theta|x_p, x_d) \right\rangle_{x_p\sim P_p, x_d\sim P_d} \\
&= -\left\langle \log p(x_p|\theta, x_d) + \log p(\theta|x_d) - \log p(x_p|x_d) \right\rangle_{x_p\sim P_p, x_d\sim P_d} \leftarrow \text{Bayes} \\
&= -\left\langle \log p(x_p|\theta, x_d) \right\rangle_{x_p\sim P_p, x_d\sim P_d} - \log p(\theta) + \text{const.} \\
&= -\left\langle \log p(\bar{g}(x_p, x_d)) + \log \left| \frac{\partial \bar{g}(x_p, x_d)}{\partial x_p} \right| \right\rangle - \log p(\theta) \leftarrow \text{change of variable} \\
&= \left\langle 0.5||\bar{g}(x_p, f(x_d))||_2^2 - \log |J| \right\rangle_{x_p\sim P_p, x_d\sim P_d} - \log p(\theta)
\end{aligned}$$

Condition INN on detector data [2006.06685]

$$x_p \xleftarrow[\leftarrow \text{ unfolding: } \bar{g}(r, f(x_d))]{g(x_p, f(x_d)) \rightarrow} r$$

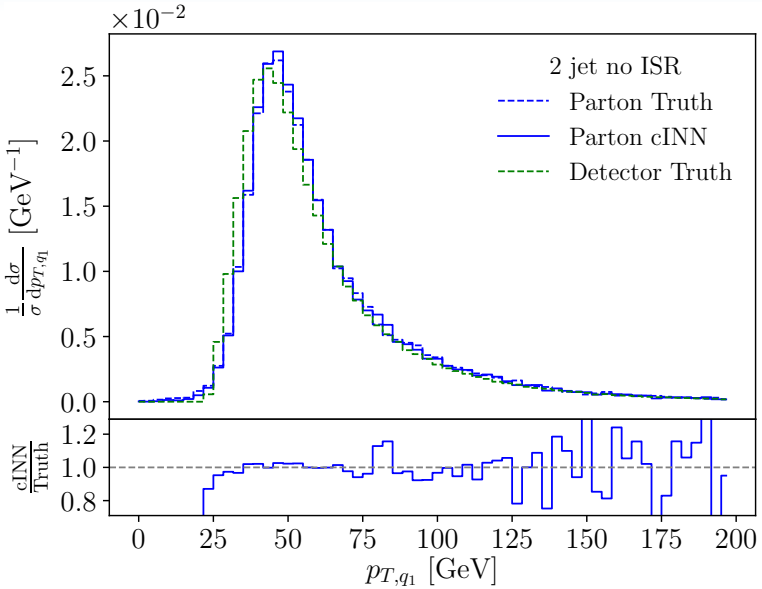
Minimizing $L = \left\langle 0.5 ||\bar{g}(x_p, f(x_d)))||_2^2 - \log |J| \right\rangle_{x_p \sim P_p, x_d \sim P_d} - \log p(\theta)$

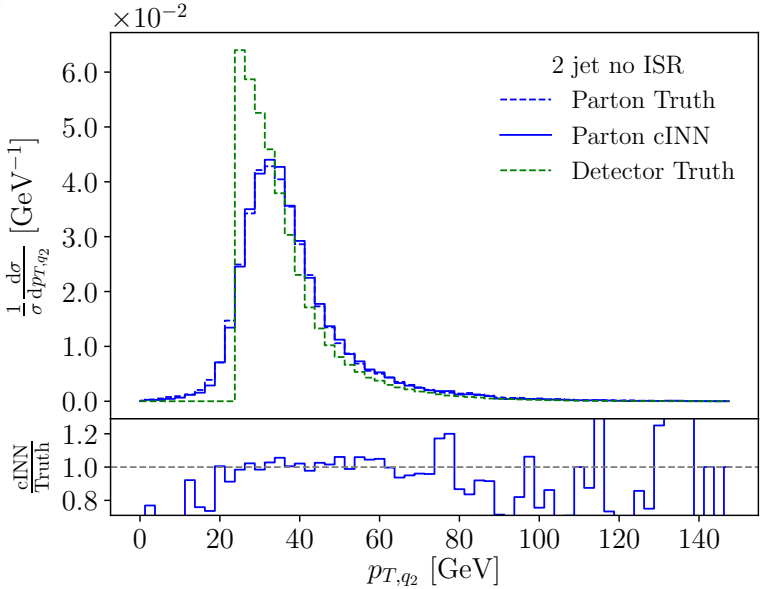


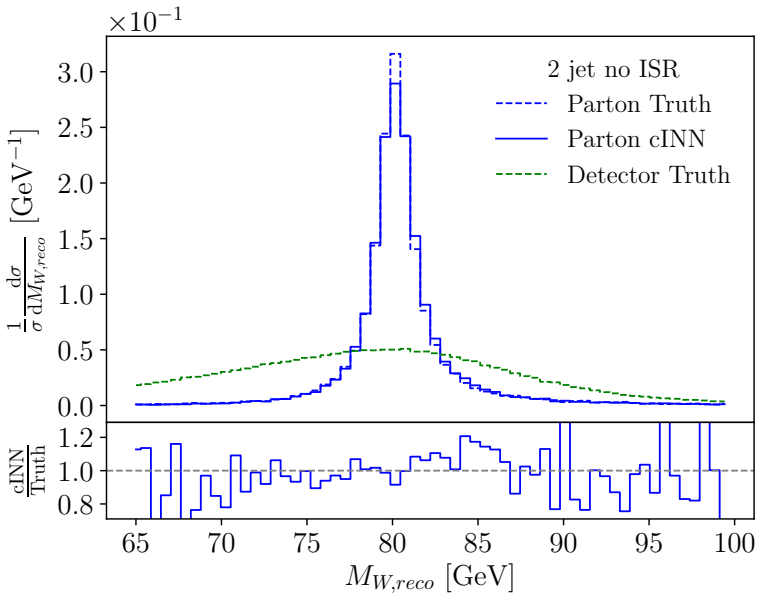


$\rightarrow$ calibrated parton level distributions

Cross check distributions





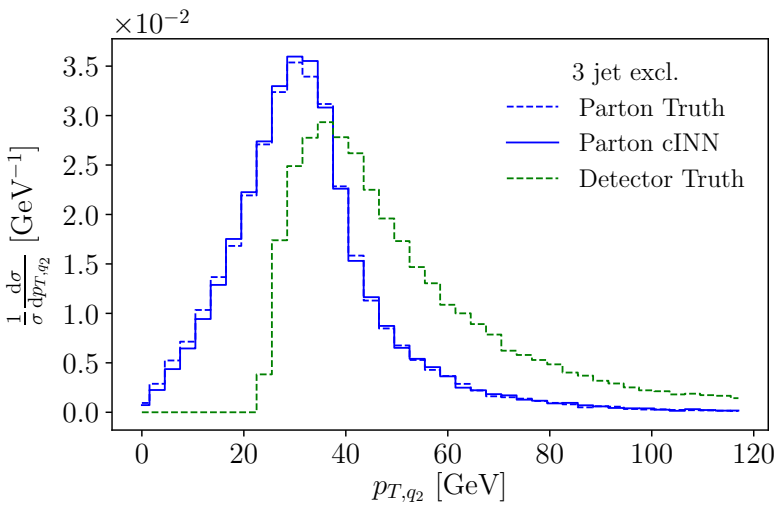


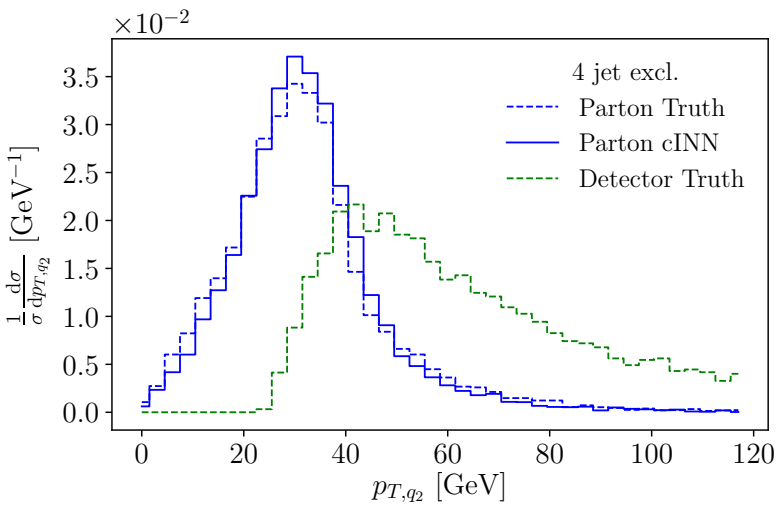
Inverting the full event I

- $pp > WZ > q\bar{q}l^+l^- +$ ISR

$\rightarrow$ ISR leads to large fraction of 2/3/4 jet events

- Conditional information can have arbitrary dimension
- Train and test on exclusive channels

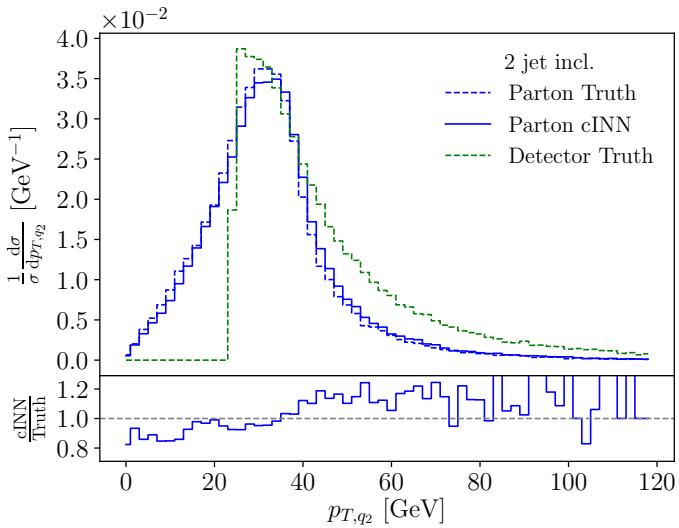




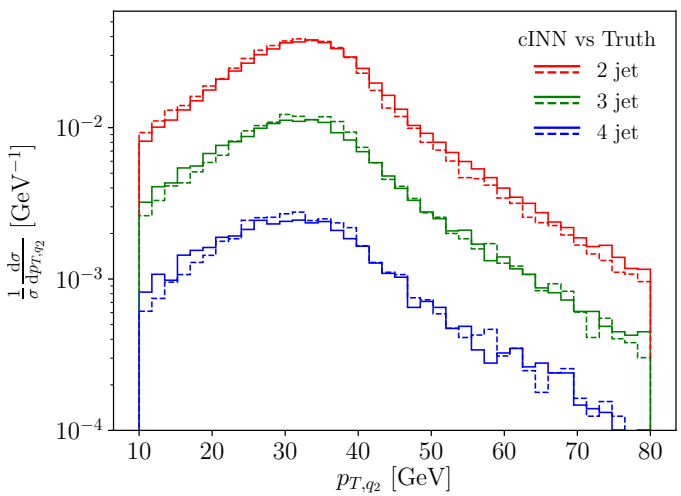
Inverting the full event II

$pp > WZ > q\bar{q}l^+l^- + \text{ISR}$

Train on inclusive dataset



Evaluate
exclusive 2/3/4 jet channels



We can use ML to ...

... enable precision simulations in forward direction

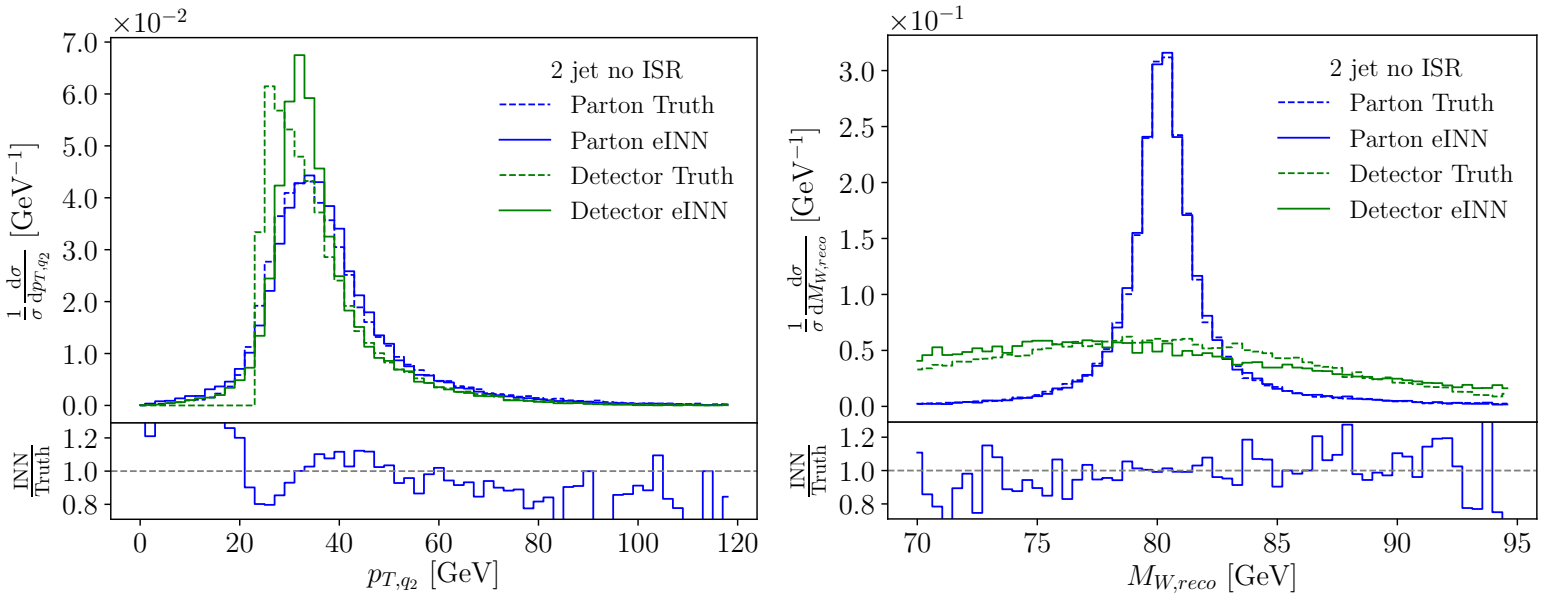
... invert the simulation chains statistically

... unfold high dimensions

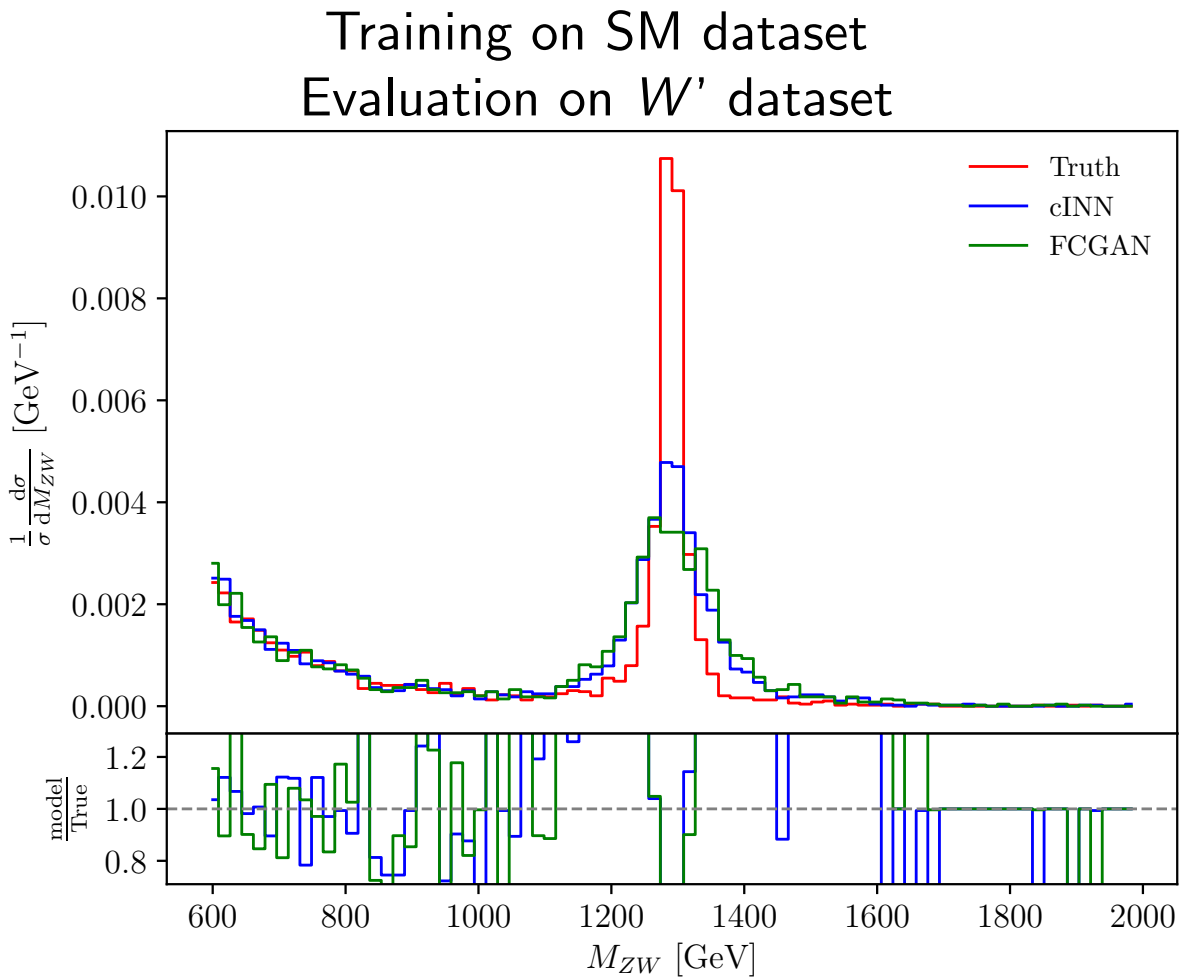
... learn more about particle physics!

BACK UP

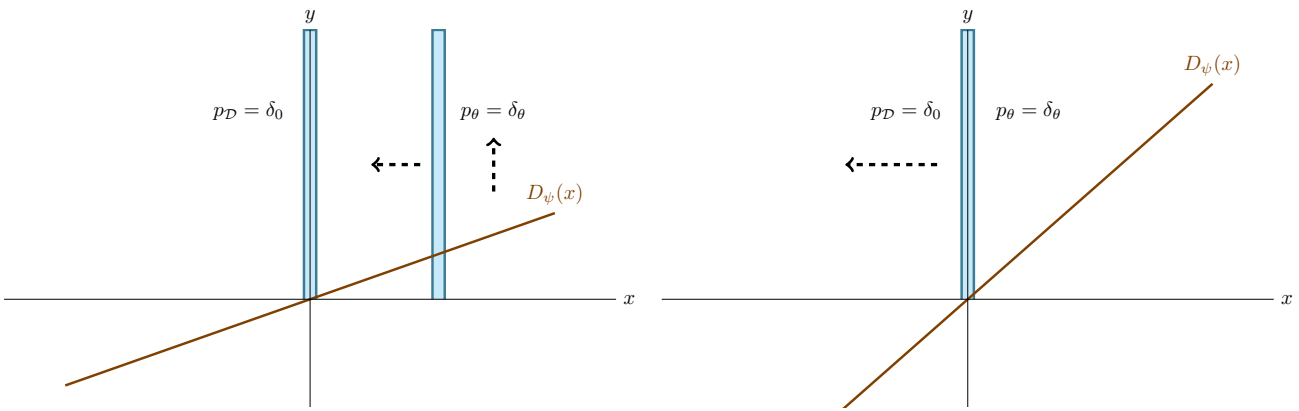
Noise extended INN



Model dependence



The GAN challenge
or
Why do we need regularization?



(a) $t = t_0$ (b) $t = t_1$

Solutions:
Additional loss or restricted network parameters

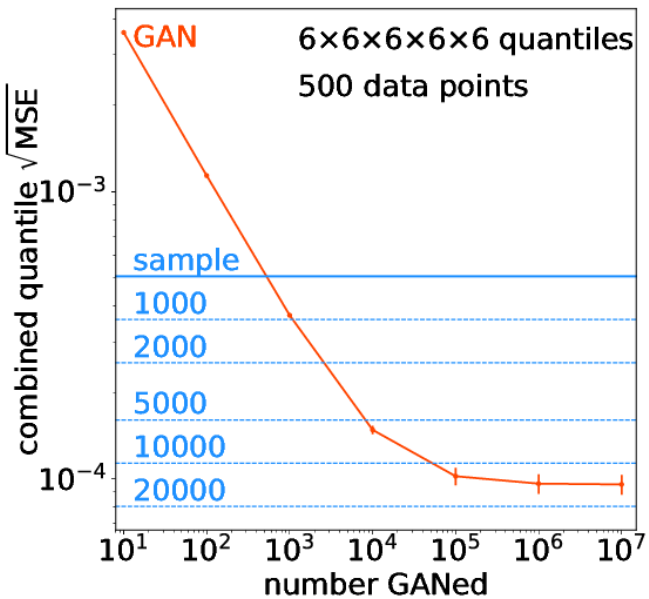
Improving GAN training

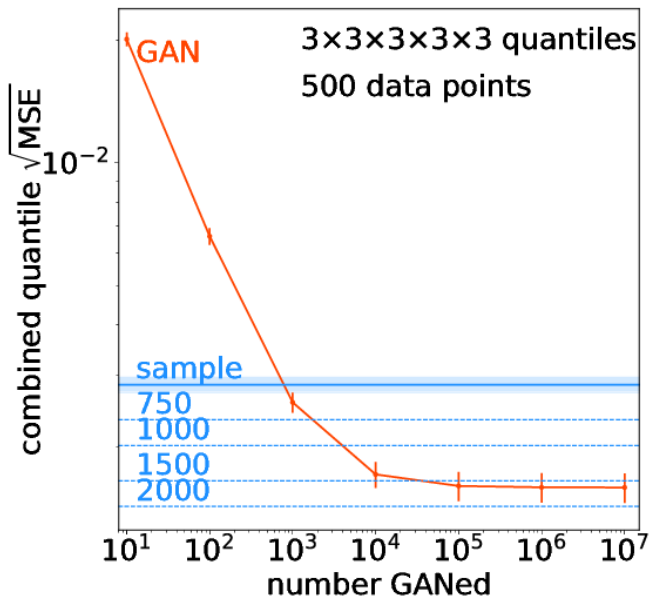
Solutions

- Regularization of the discriminator, eg. gradient penalty [Ghosh, Butter et al., ...]
- Modified training objective:
  - Wasserstein GAN (incl. gradient penalty) [Lin et al., Erdmann et al., ...]
  - Least square GAN (LSGAN) [Martinez et al., ...]
  - MMD-GAN [Otten et al., ...]
  - MSGAN [Datta et al., ...]
  - Cycle GAN [Carazza et al., ...]
- Use of symmetries [Hashemi et al., ...]
- Whitening of data [Di Sipio et al., ...]
- Feature augmentation [Alanazi et al., ...]

Amplification

5-dim sphere





The subtraction loss function

- Standard GAN loss for each discriminator
- Differentiable function to count events of one type

$$f(c) = e^{-\alpha(\max(c)^2-1)^{2\beta}} \in [0,1] \qquad \text{for} \qquad 0 \leq c_i \leq 1 \,.$$

- Reward clear class assignment

$$L_G^{(\text{class})} = \left(1 - \frac{1}{b} \sum_{c \in batch} f(c)\right)^2$$

- Fix normalization

$$L_{G_i}^{(\text{norm})} = \left(\frac{\sum_{c \in \mathcal{C}_i} f(c)}{\sum_{c \in \mathcal{C}_B} f(c)} - \frac{\sigma_i}{\sigma_0}\right)^2$$