How to GAN Event Subtraction

Anja Butter

ITP, Universität Heidelberg

arXiv:1912.08824

with Ramon Winterhalder and Tilman Plehn

Physics case

- Theory uncertainties have become a limiting factor for LHC analyses

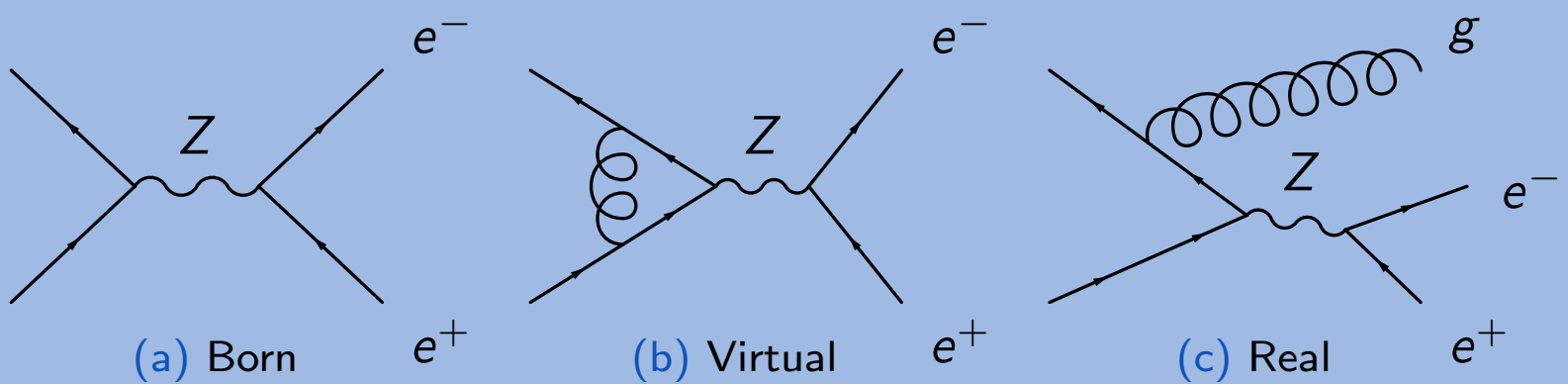
$\rightarrow$ Need for better accuracy

Physics case

- Theory uncertainties have become a limiting factor for LHC analyses

→ Need for better accuracy

NLO in a nutshell



(a) Born (b) Virtual (c) Real

$$\sigma_{NLO} = \int d\Phi_B (B + V) + \int d\Phi_R R$$

Subtracting divergencies

- Virtual and real corrections diverge individually (eg. IR divergence)
- Sum of divergent contributions is finite

$\rightarrow$ Introduce dipoles D_i to cancel divergencies

Dipole subtraction

$$\sigma_{NLO} = \int d\Phi_B (B + V + \sum_i d\Phi_{R|B} D_i) + \int d\Phi_R (R - \sum_i D_i)$$

Subtracting divergencies

- Virtual and real corrections diverge individually (eg. IR divergence)
- Sum of divergent contributions is finite

→ Introduce dipoles D_i to cancel divergencies

Dipole subtraction

$$\sigma_{NLO} = \int d\Phi_B (B + V + \sum_i d\Phi_{R|B} D_i) + \int d\Phi_R (R - \sum_i D_i)$$

- Analytic solution only possible for simple processes
- Numeric subtraction of samples:
  - → large statistic uncertainties
  - → limits efficiency
- Other use cases: eg. on-shell subtractions, multi-jet merging

Why GAN?

- A lot of experience as a community!
  - Jet Images - de Oliveira et al. [1701.05927], Carazza et al. [1909.01359],
  - Particle shower in Calorimeters - Paganini et al. [CaloGAN, 1705.02355, 1712.10321], Musella et al. [1805.00850], Erdmann et al. [1807.01954], ATLAS [ATL-SOFT-PUB-2018-001, ATL-SOFT-PROC-2019-007]
  - Event generation - Otten et al. [1901.00875], Hashemi et al. [1901.05282], Di Sipio et al. [1903.02433], Butter et al. [1907.03764], Martinez et al. [1912.02748]
  - Unfolding - Datta et al. [1806.00433], Bellagente et al. [1912.0047]
  - Templates for QCD factorization - Lin et al. [1903.02556]
  - EFT models - Erbin et al. [1809.02612]
  - ...

- Demonstrated that GANs
  - learn underlying distributions from samples
  - have excellent interpolation properties

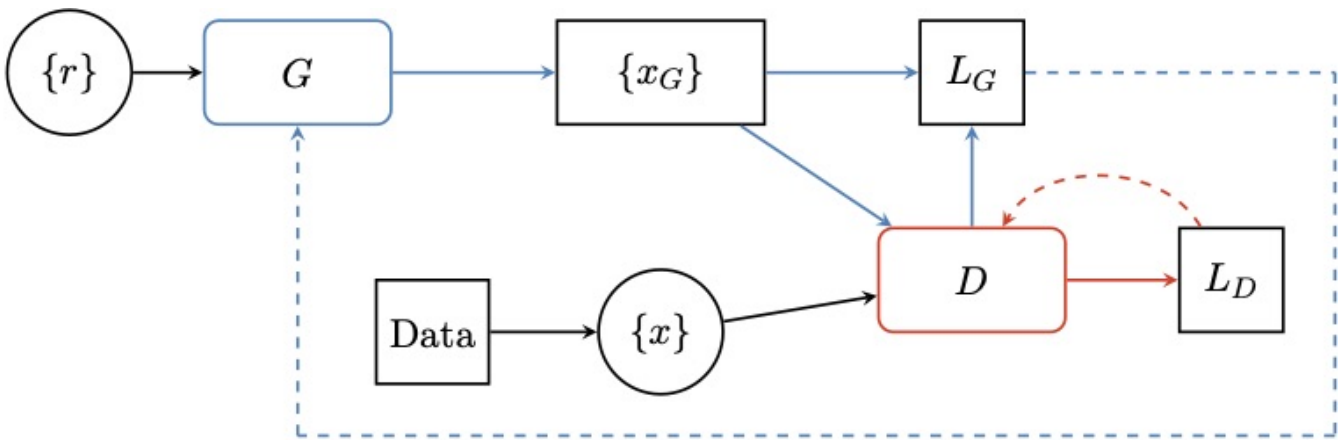
Sample based subtraction of distributions

- Use GAN to subtract distribution P_S (subtract) from P_B (base)
- Distributions represented by samples
- GAN output: samples following P_{B-S}
- Idea:
  - One discriminator per sample distribution
  - Generate label vector c to identify subtraction events
  - $0 \leq c_i \leq 1$, $\sum_i c_i = 1 \rightarrow$ softmax

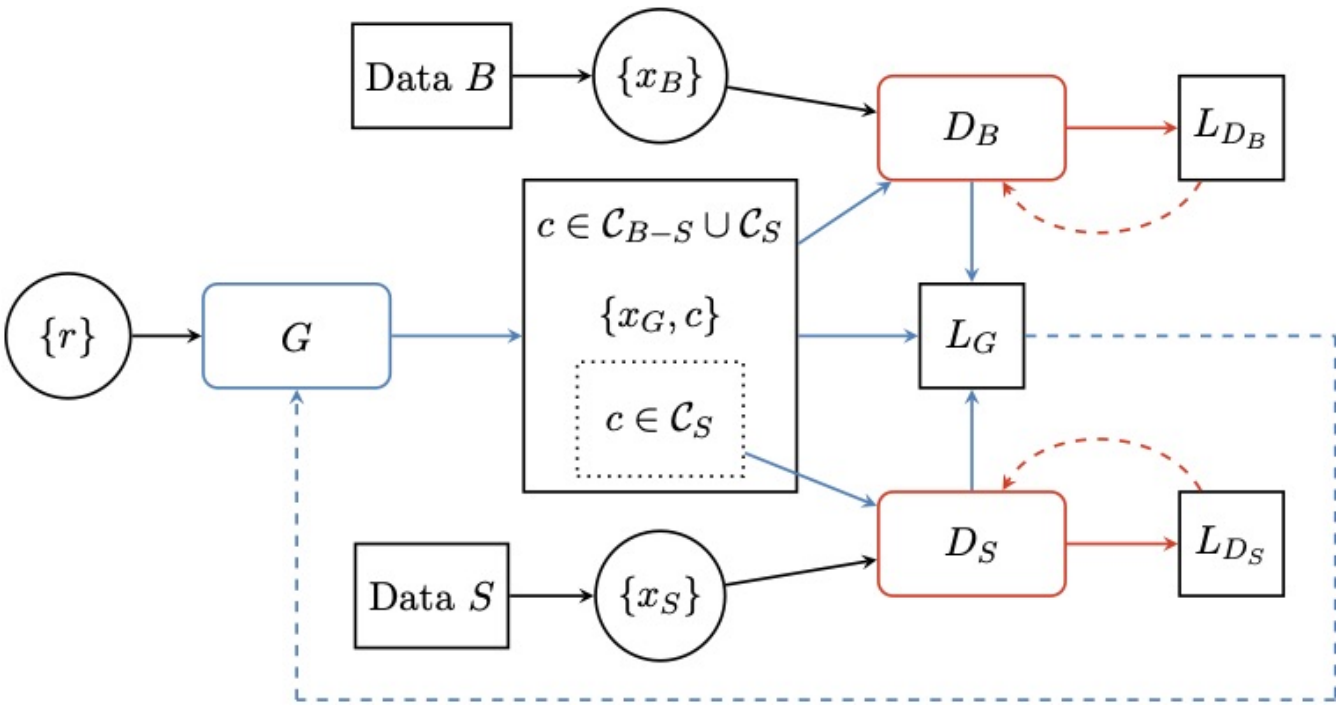
$$c = \begin{pmatrix} \mathcal{C}_S \\ \mathcal{C}_{B-S} \end{pmatrix}$$

	$\mathcal{C}_{B-S}$	$\mathcal{C}_S$
Data B	1	1
Data S	0	1
B-S	1	0

From a standard GAN ...



... to a subtraction GAN



Building the loss function

- Standard GAN loss for each discriminator

Building the loss function

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

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

Building the 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

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

Building the 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

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

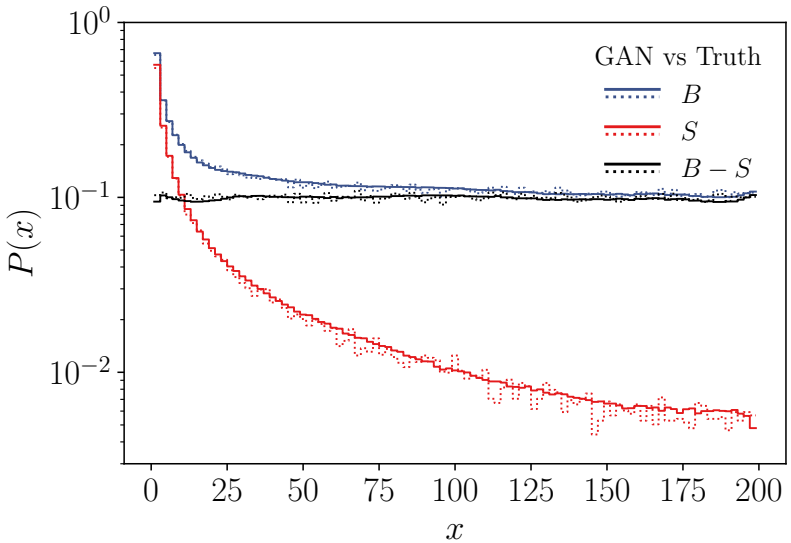
- Fix normalization

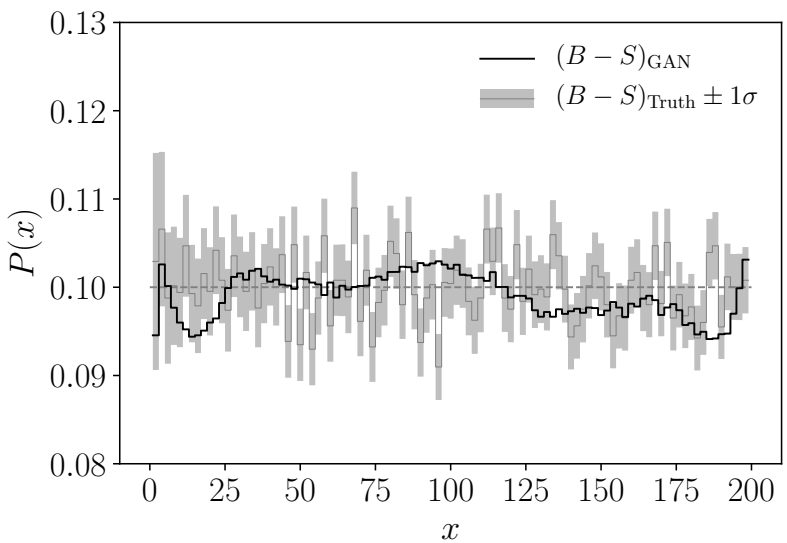
$$L_{G_i}^{(\text{norm})} = \left(\frac{\sum_{c\in\mathcal{C}_i} f(c)}{\sum_{c\in\mathcal{C}_B} f(c)} - \frac{\sigma_i}{\sigma_0}\right)^2$$

Toy example

- Toy example:

$$P_B(x) = \frac{1}{x} + 0.1$$
$$P_S(x) = \frac{1}{x}$$
$$P_{B-S}(x) = 0.1$$





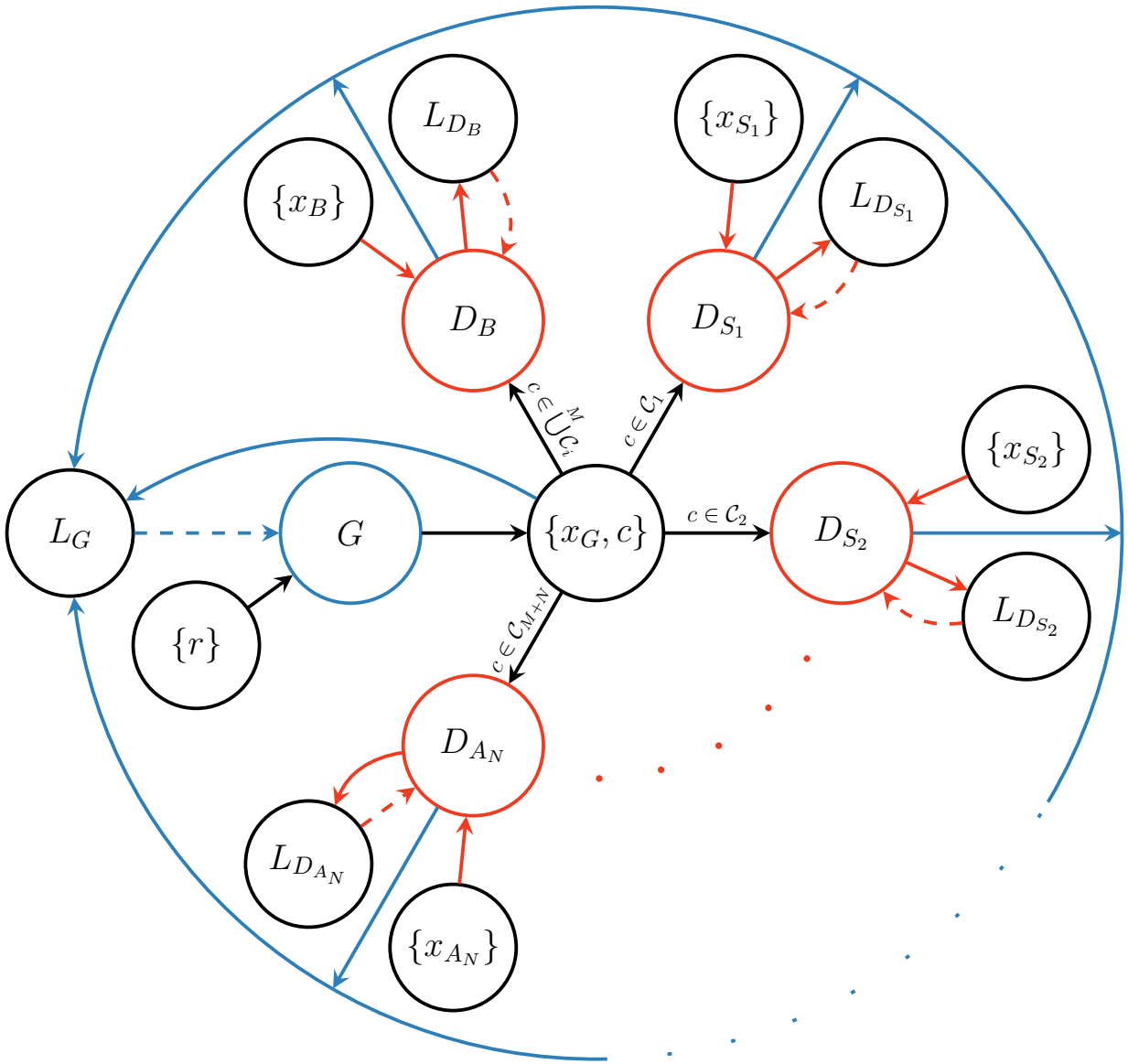
Generalizing the setup

- Include addition

	$\mathcal{C}_{B-S}$	$\mathcal{C}_S$	$\mathcal{C}_A$
Data B	1	1	0
Data S	0	1	0
Data A	0	0	1
B-S+A	1	0	1

- Use case:
  - One distribution is represented by significantly smaller dataset
- Allow for more datasets

Generalizing the setup

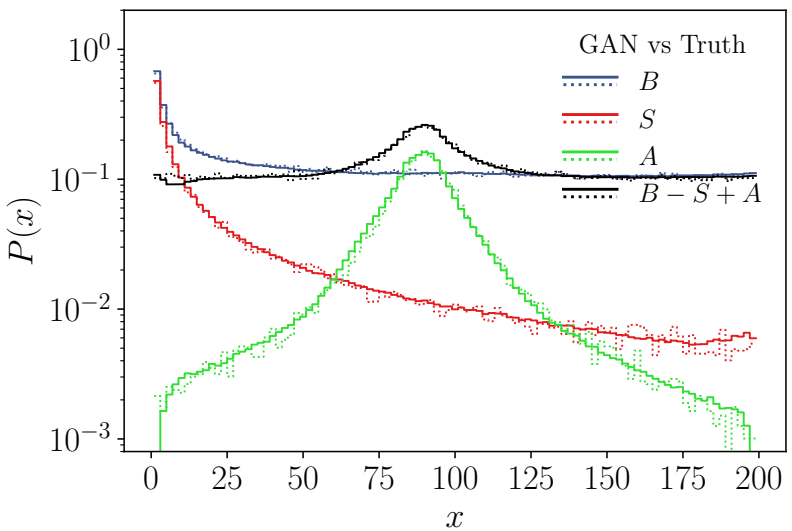


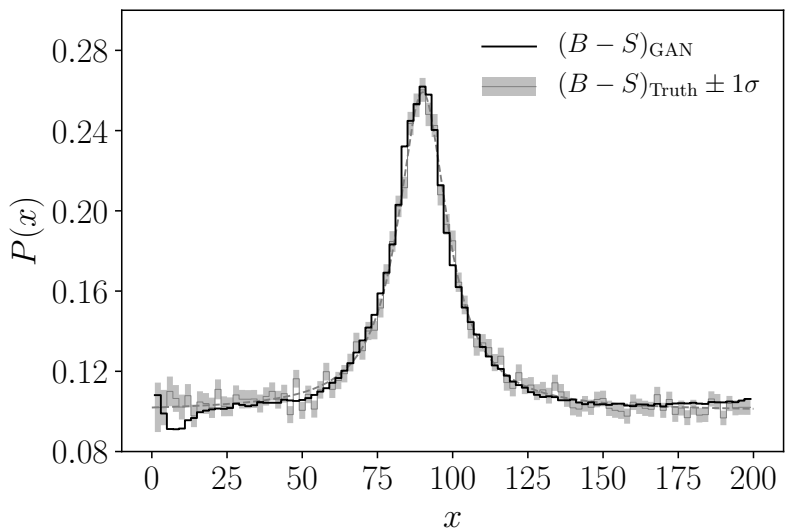
Include addition

$$P_B(x) = \frac{1}{x} + 0.1$$

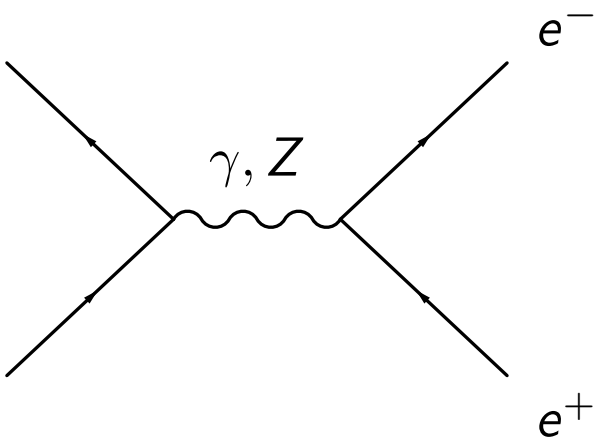
$$P_S(x) = \frac{1}{x}$$

$$P_A(x) = \frac{5}{\pi} \frac{10}{10^2 + (x - 90)^2}$$

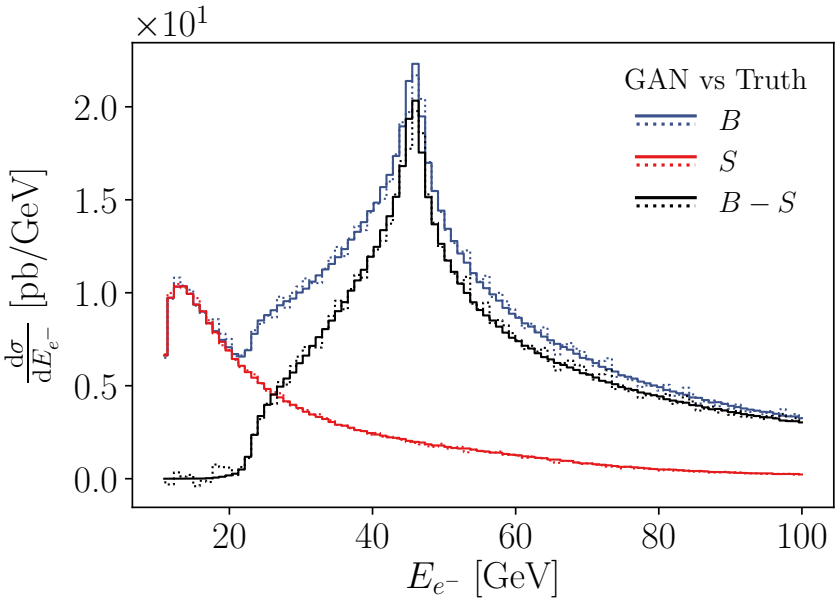




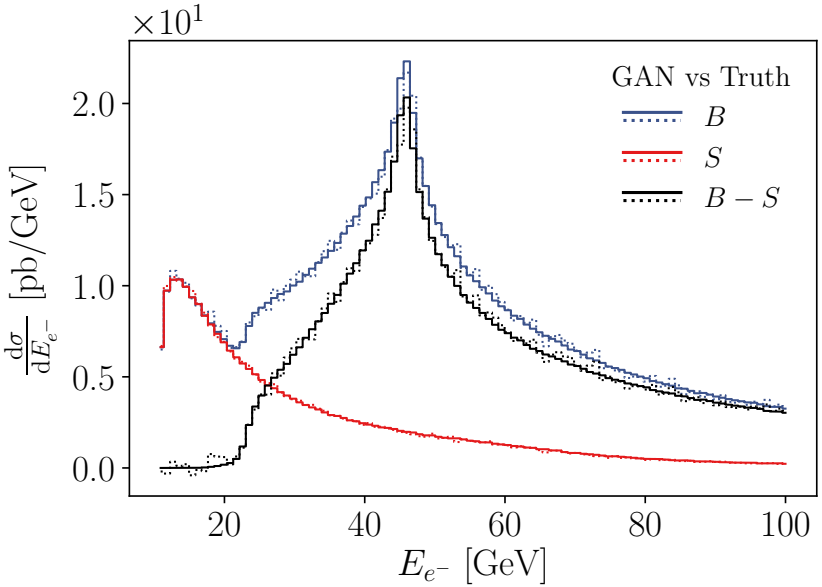
Subtracting LHC events

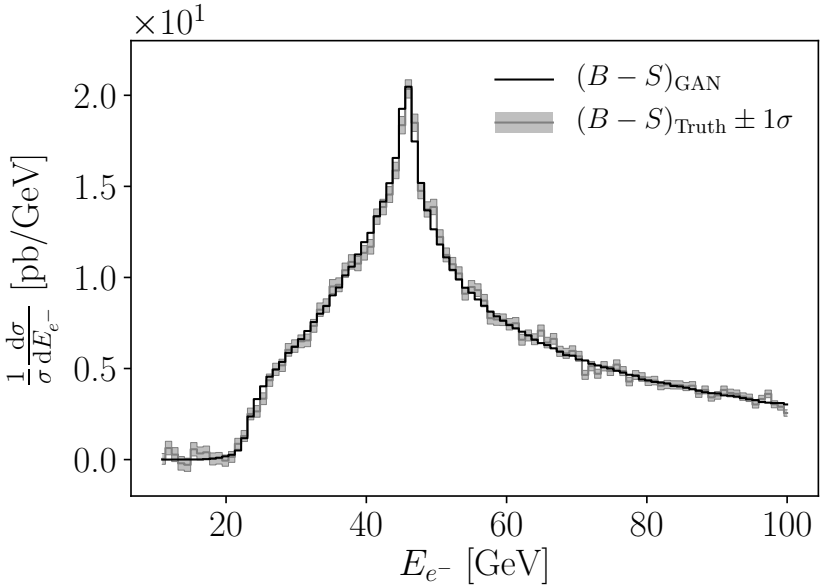


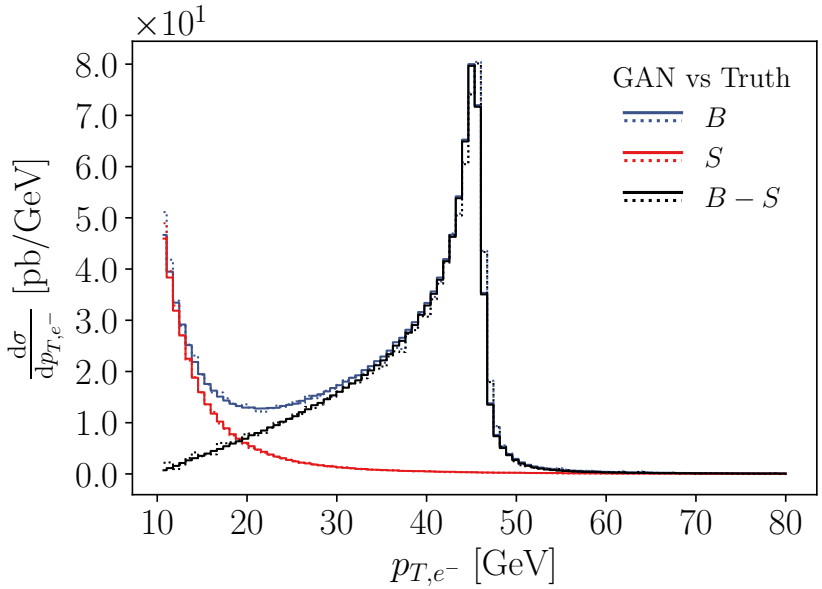
- P_B: $pp \to e^+e^-$
- P_S: $pp \to \gamma \to e^+e^-$
- $p_T > 10$ GeV
- on-shell final state:
  6 dimensional output

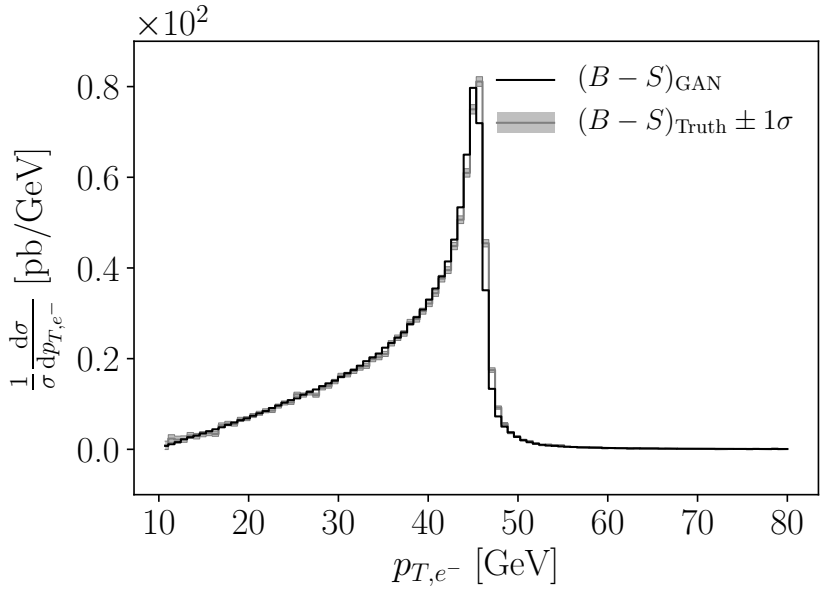


Subtracting LHC events

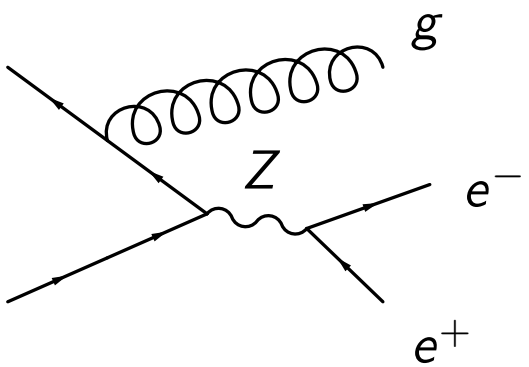






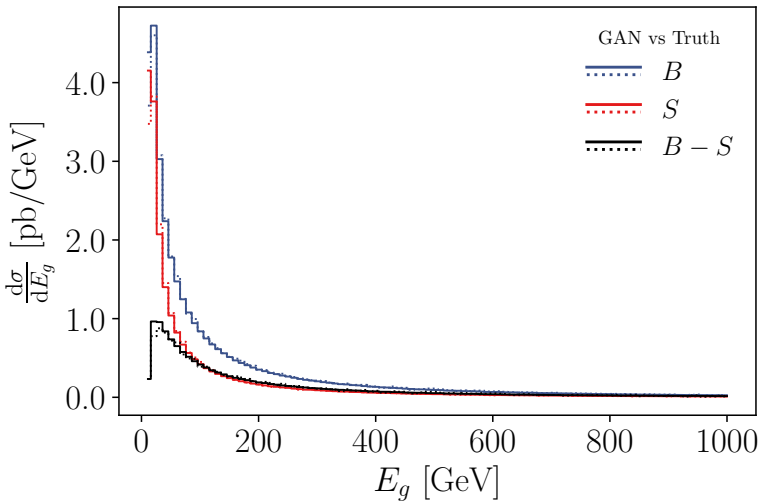


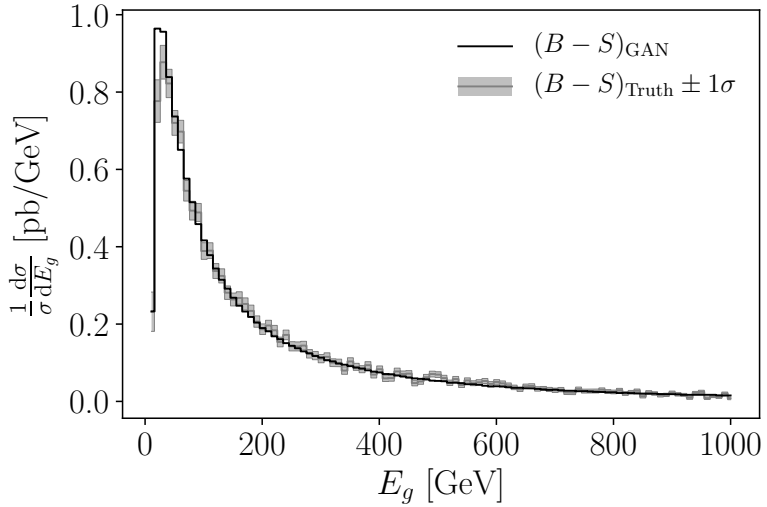
Back to the original problem

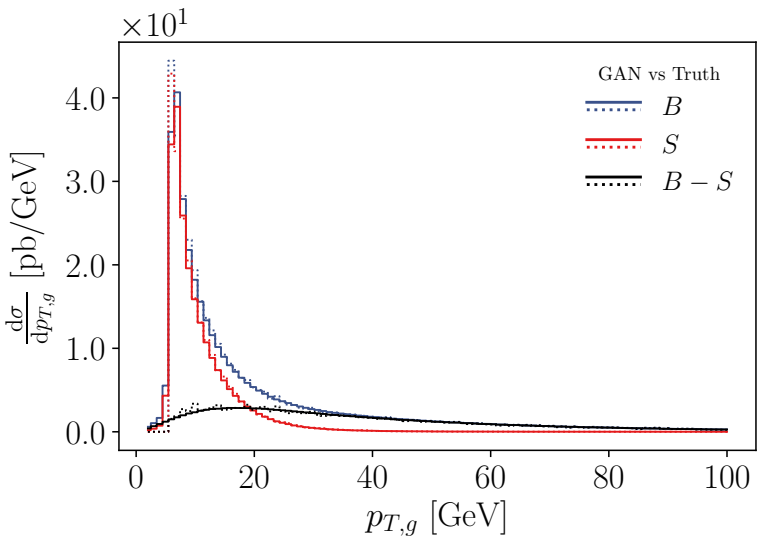


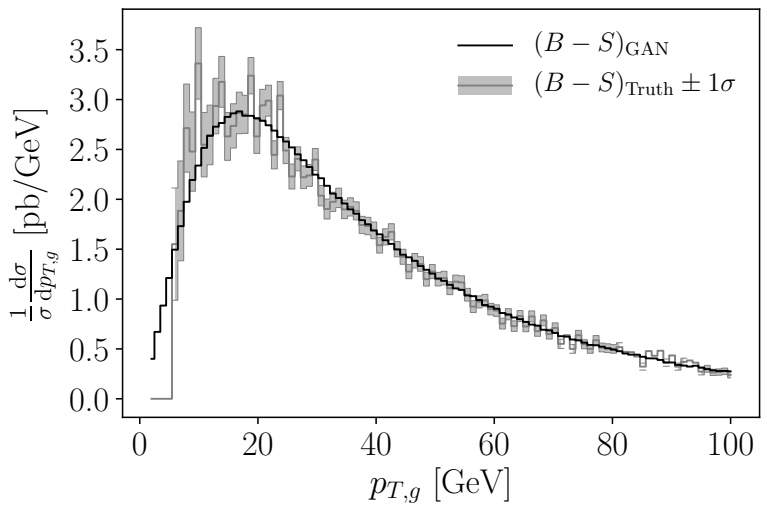
- Subtract the Catany Seymour Dipole from the real emission term
- For proof of concept we use a slightly modifed Catany Seymour kernel $\rightarrow$ increase difference
- Training
  - 10^5 samples per distribution
  - 4-vector representation of Z and g
  - $E_g > 5$ GeV

Results I

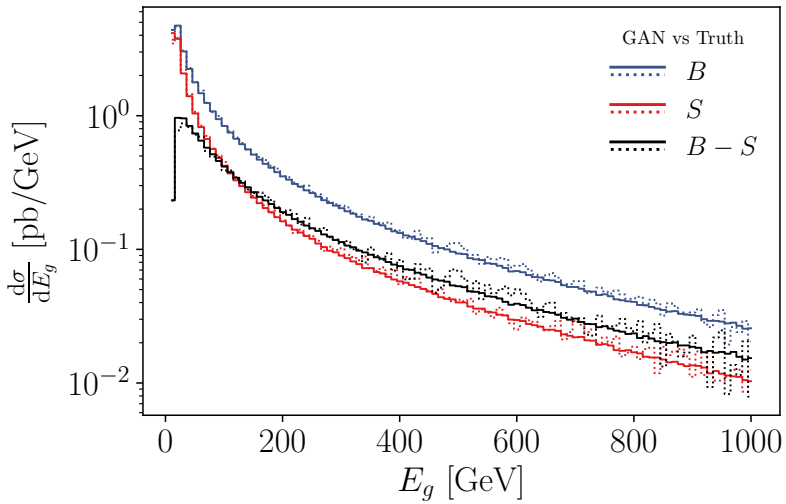

GAN vs Truth
B
S
$B - S$
$\frac{d\sigma}{dE_g}$ [pb/GeV]
E_g [GeV]

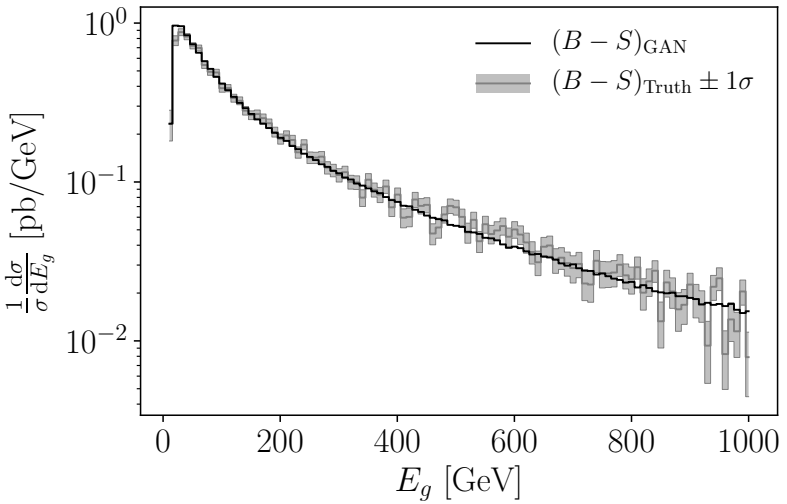

$(B - S)_{\text{GAN}}$
$(B - S)_{\text{Truth}} \pm 1\sigma$
$\frac{1}{\sigma}\frac{d\sigma}{dE_g}$ [pb/GeV]
E_g [GeV]

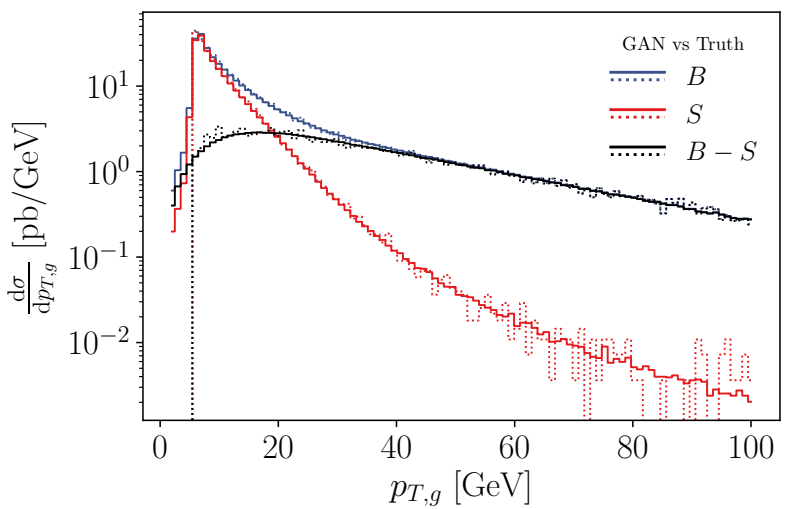

$\times 10^1$
GAN vs Truth
B
S
$B - S$
$\frac{d\sigma}{dp_{T,g}}$ [pb/GeV]
$p_{T,g}$ [GeV]

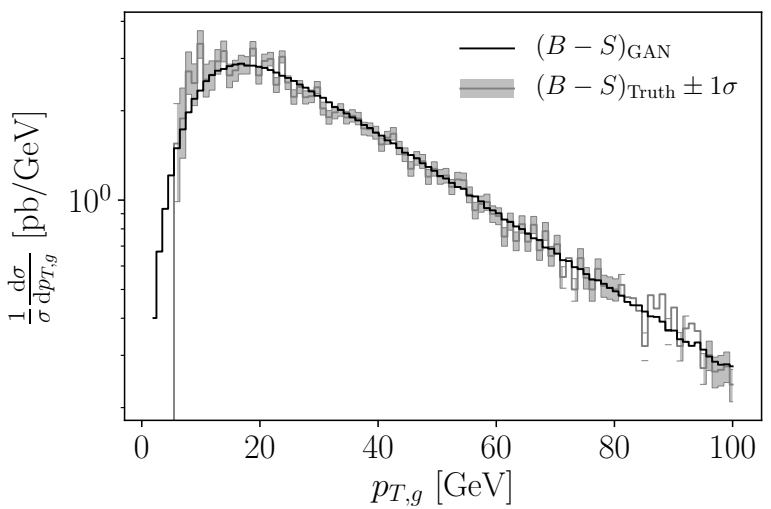

$(B - S)_{\text{GAN}}$
$(B - S)_{\text{Truth}} \pm 1\sigma$
$\frac{1}{\sigma}\frac{d\sigma}{dp_{T,g}}$ [pb/GeV]
$p_{T,g}$ [GeV]

Results II


GAN vs Truth
B
S
$B - S$
$\frac{\mathrm{d}\sigma}{\mathrm{d}E_g}$ [pb/GeV]
E_g [GeV]


$(B - S)_{\mathrm{GAN}}$
$(B - S)_{\mathrm{Truth}} \pm 1\sigma$
$\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}E_g}$ [pb/GeV]
E_g [GeV]


GAN vs Truth
B
S
$B - S$
$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{T,g}}$ [pb/GeV]
$p_{T,g}$ [GeV]


$(B - S)_{\mathrm{GAN}}$
$(B - S)_{\mathrm{Truth}} \pm 1\sigma$
$\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}p_{T,g}}$ [pb/GeV]
$p_{T,g}$ [GeV]

Outlook

- HL-LHC results limited by uncertainty on theory prediction
- Need to improve efficiency of computing the subtracted real-emission corrections
- GAN for sample based subtraction
  $\rightarrow$ successful proof of concept!
- Work with Monte Carlo community to test efficiency
- New tool for our ML toolbox
  $\rightarrow$ other use cases?

Hyperparameters - Toy1

Parameter	Value
training size	10^5
layers	5
units	128
batch size	1024
learning rate	$3 \cdot 10^{-4}$
decay generator	$5 \cdot 10^{-3}$
decay discriminator	$2 \cdot 10^{-2}$
epochs	4000
discriminator updates	20
α	10
gradient penalty λ_{D_i}	$5 \cdot 10^{-5}$

Hyperparameters - Toy2

Parameter	Value
training size	10^5
layers	7
units	128
batch size	1024
learning rate	$8 \cdot 10^{-4}$
decay generator	$2 \cdot 10^{-2}$
decay discriminator	$2 \cdot 10^{-2}$
epochs	1000
iterations	4
discriminator updates	20
α	5
gradient penalty λ_{D_i}	$5 \cdot 10^{-5}$

Hyperparameters - Resonance

Parameter	Value
training size	10^5
layers	8
G units	160
D units	80
batch size	1024
learning rate	10^{-3}
decay generator	10^{-2}
decay discriminator	10^{-2}
epochs	1000
iterations	5
discriminator updates	2
α	5
gradient penalty λ_{D_i}	10^{-5}

Hyperparameters - Dipole

Parameter	Value
training size	10^5
layers	8
G units	512
D units	256
batch size	1024
learning rate	0.001
decay generator	0.01
decay discriminator	0.01
epochs	20000
iterations	5
discriminator updates	2
α	5
gradient penalty λ_{D_i}	0.001