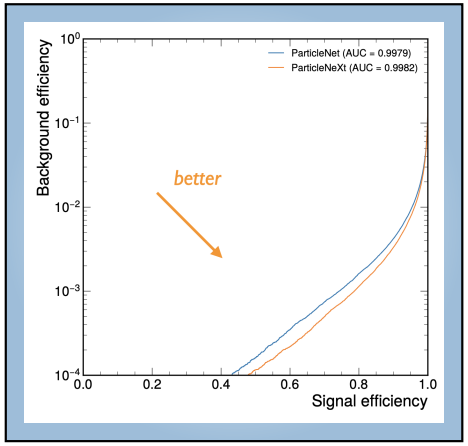
Machine Learning for LHC Theory

DPNC - Université de Genève

Anja Butter

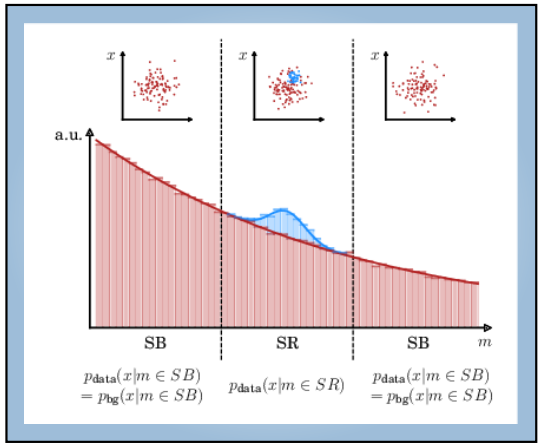
ITP, Universität Heidelberg

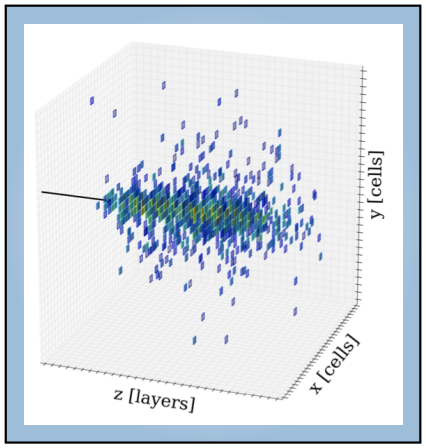
Performance boosts and new developments for many applications



←Top tagging

Anomaly detection →





←Detector simulation

What about ML4Theory?

What about ML4Theory?

Better predictions?

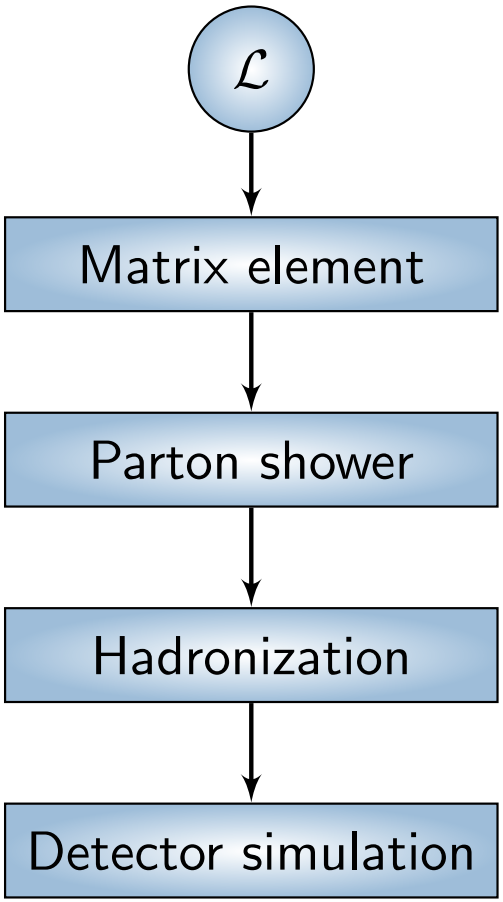
Better understanding?

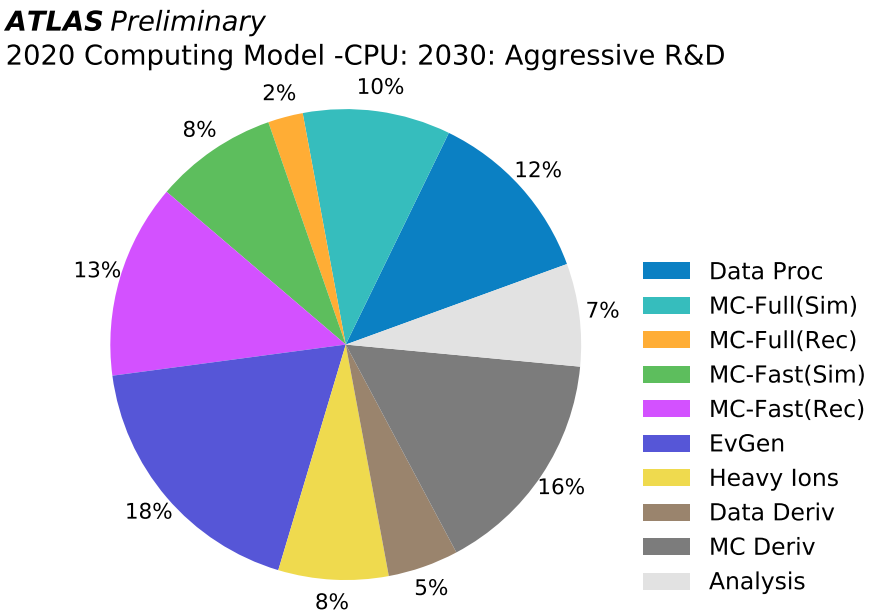
What about ML4Theory?

Better predictions?
$\rightarrow$ ML for precision simulations

Better understanding?
$\rightarrow$ Turn data into theory?

Precision simulations with limited resources





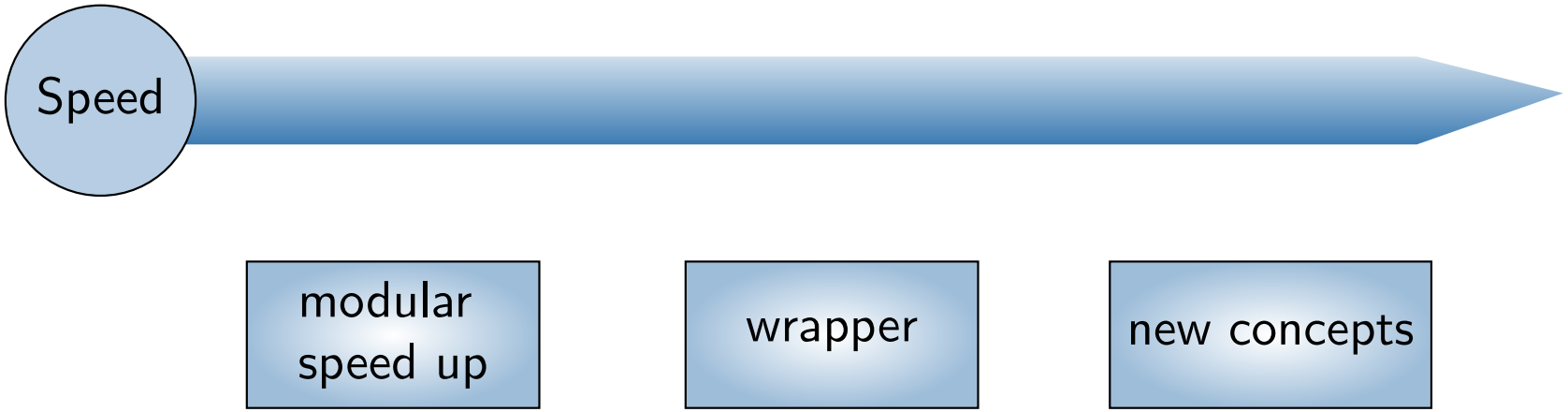
Speed = Precision

How can ML help increasing precision

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

Fast evaluation → more events / higher order → Precision

more events = higher order



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...

Matrix element

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

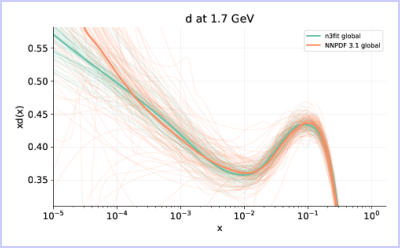
PDF

Phase space mapping

Boosting standard event generation...

Matrix element

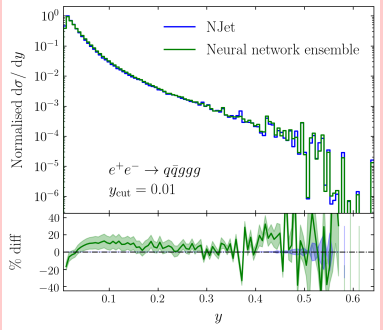
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



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

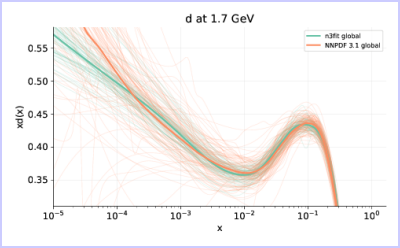
Phase space mapping

Boosting standard event generation...



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

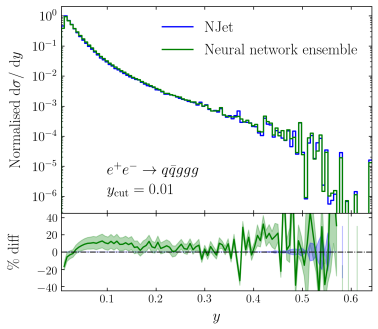
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



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

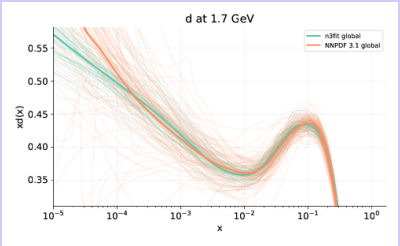
Phase space mapping

Boosting standard event generation...

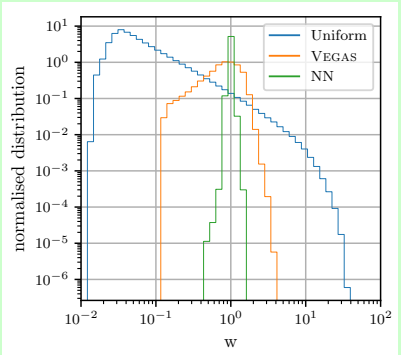


- Amplitude estimation
- S. Badger, J. Bullock [2002.07516]
- J. Bendavid [1707.00028]

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



- 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**

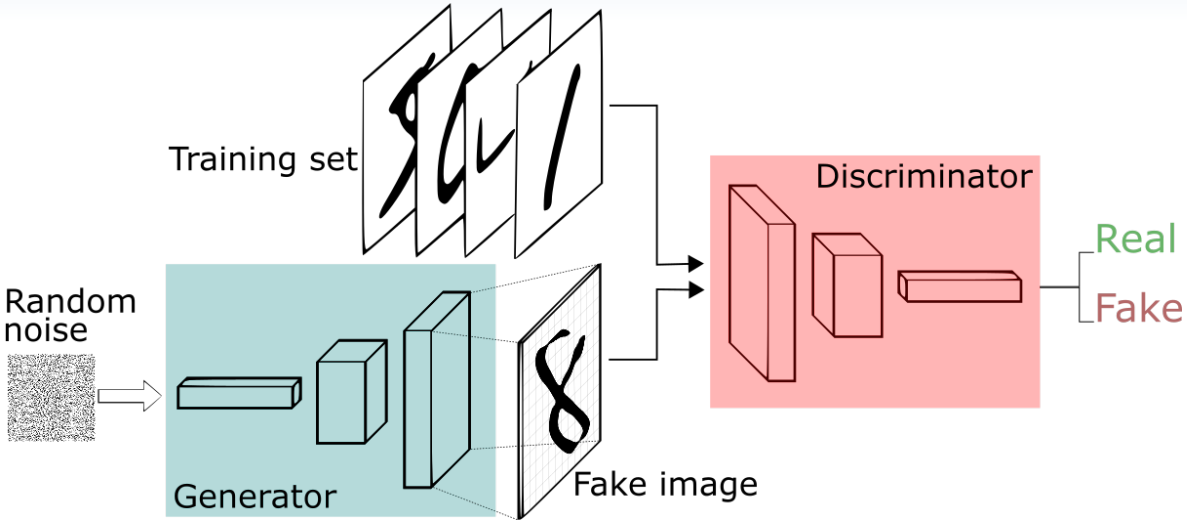
[2005.05334] Buhmann et al. **VAE**

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

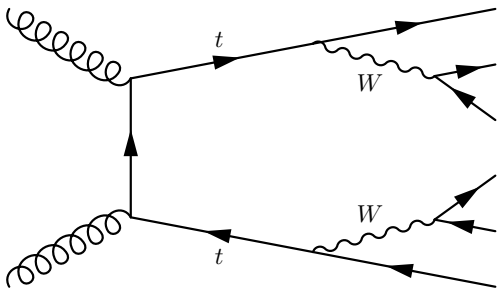
$$L_G = \left\langle -\log D(x) \right\rangle_{x \sim P_{Gen}}$$

$\Rightarrow$ **Equilibrium**

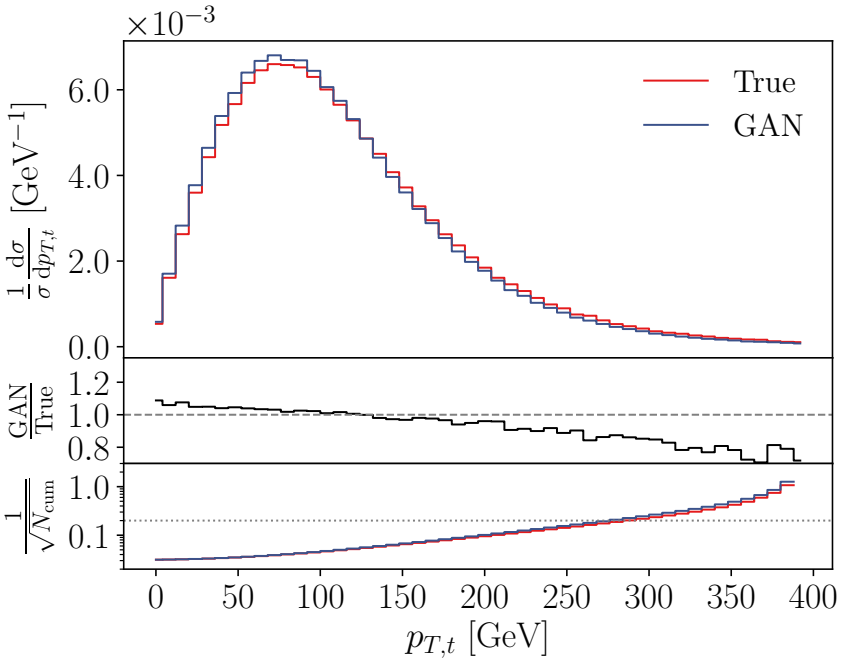
$\Rightarrow$ **New statistically independent samples**

How to GAN LHC events [1907.03764]

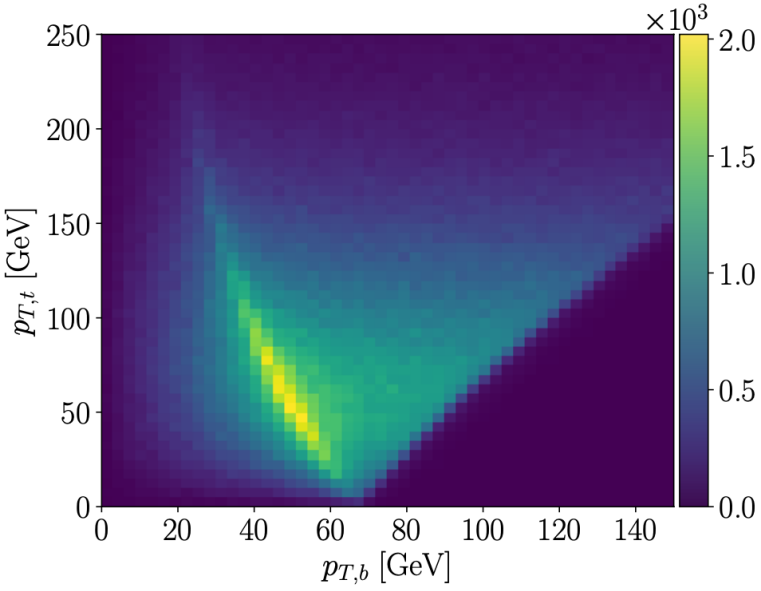
- $t\bar{t} \to 6$ quarks
- 18 dim output
  - external masses fixed
  - no momentum conservation

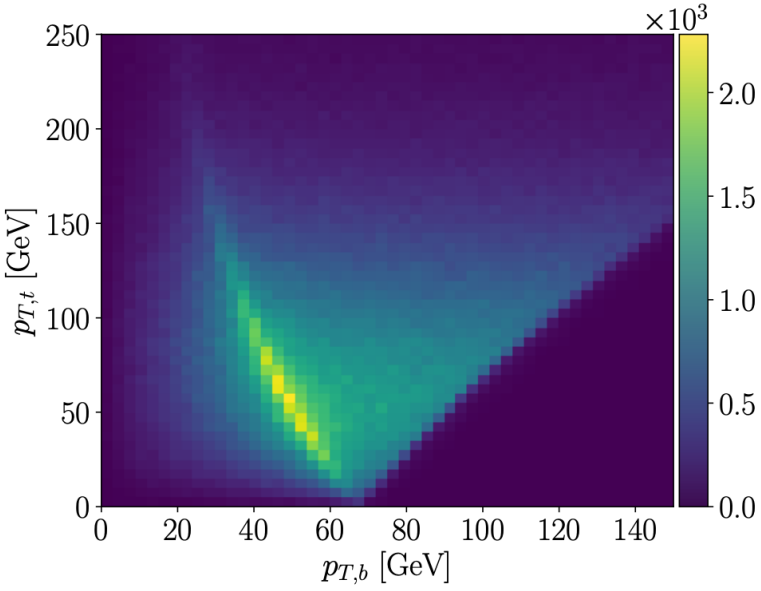


+ Flat observables ✓
− Systematic undershoot in tails [10-20% deviation]



Correlations





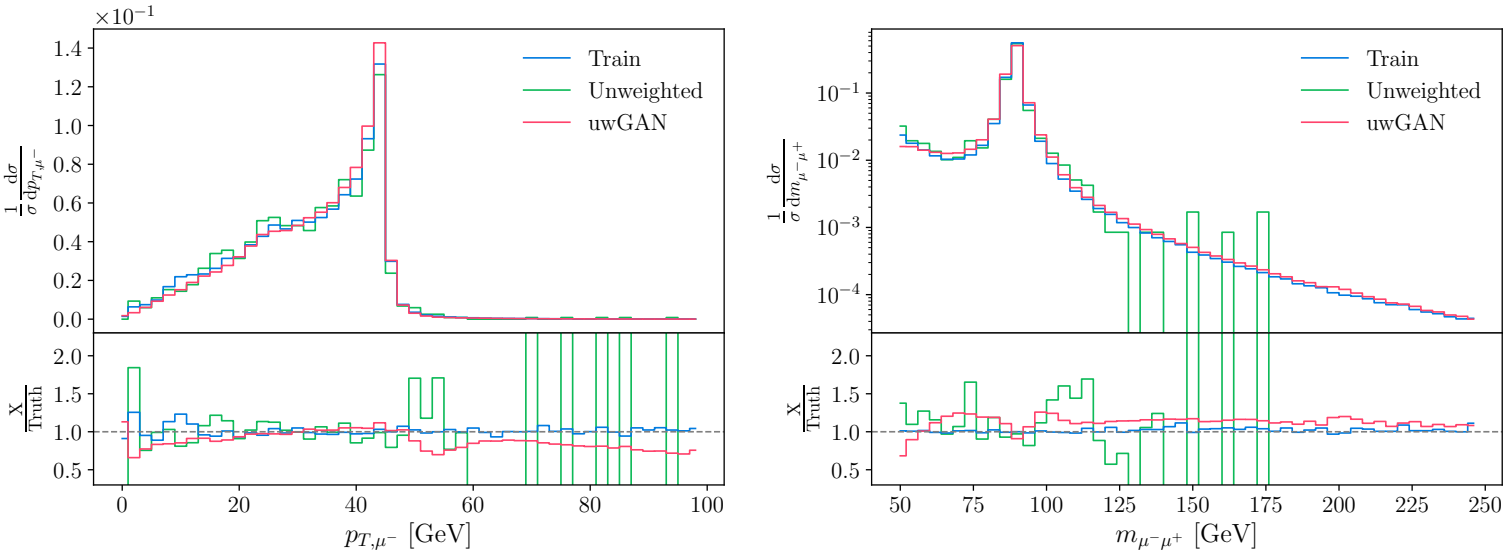
✓

Training on weighted events

Low unweighting efficiencies → bottleneck before training

→ Train on weighted events

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



Populates high energy tails

Large amplification wrt. unweighted data!

Short summary

GANs can ..

$\rightarrow$ learn event distributions and correlations

$\rightarrow$ amplify underlying statistics

$\rightarrow$ train on weighted events

Open questions:

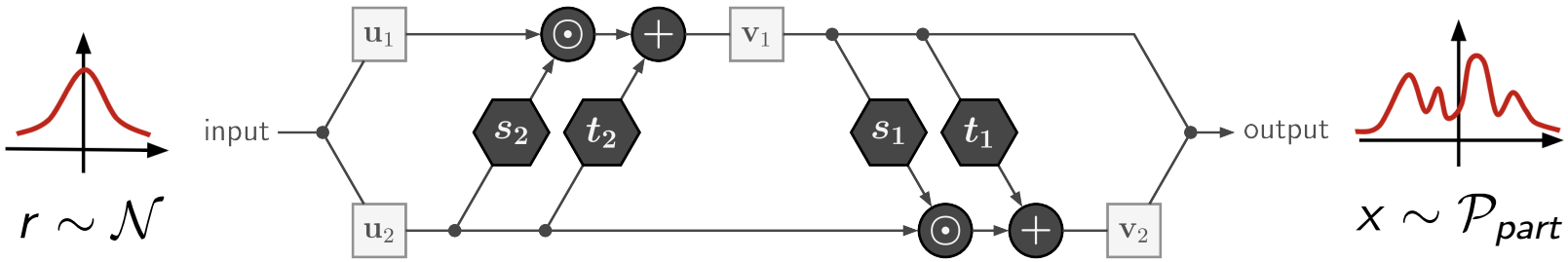
$\rightarrow$ Enough flexibility for inclusive production?

$\rightarrow$ Control?

$\rightarrow$ How to achieve precision?

$\rightarrow$ How to quantify uncertainties?

Invertible networks



+ Tractable Jacobian

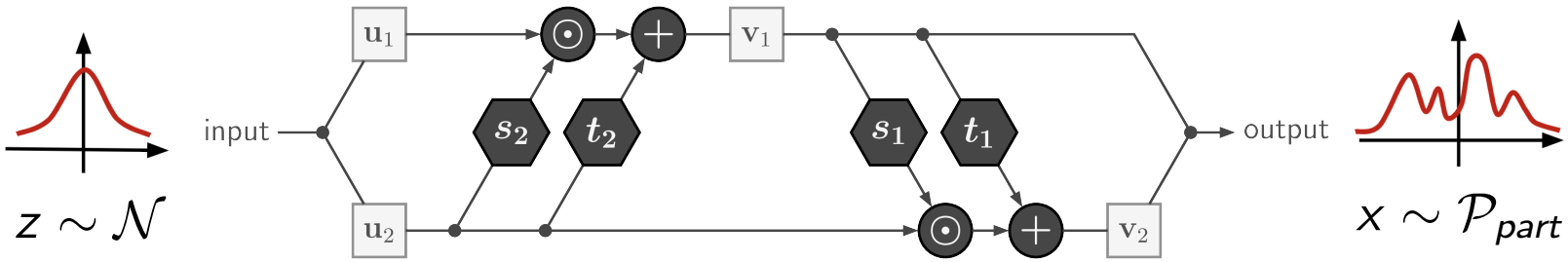
+ Enable correction for perfect precision

+ Fast evaluation in both directions

+ Extendable to Bayesian invertible networks

Training on density

Sherpa [2001.05478, 2001.10028]



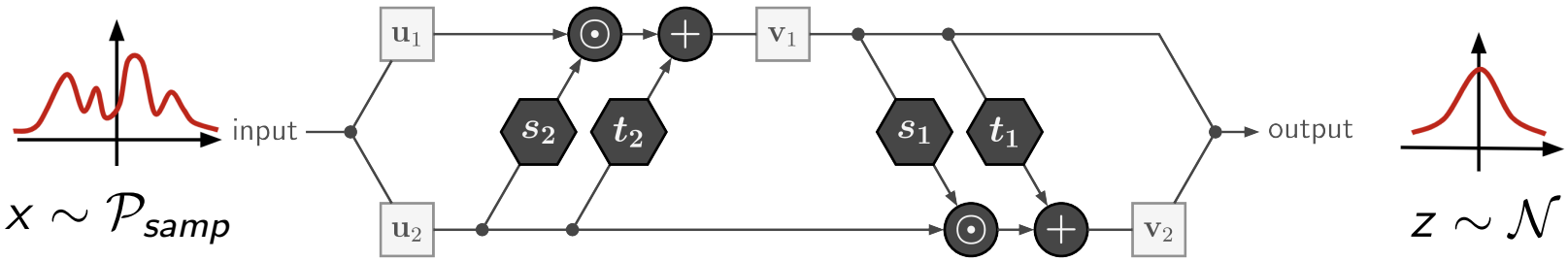
- $z \sim \mathcal{N} \rightarrow$ NN $\rightarrow x \sim p_x$
- $p_x(x) = p_z(z) \cdot J_{\text{NN}}$
- Given target density $t(x)$

$\rightarrow$ Train NN to minimize $\log(p_z(z) \cdot J_{\text{NN}}/t(x))$

- Problem: Calculate $f(x)$ each time

Training on samples

A.B., T. Heimel, S. Hummerich, T. Krebs, T. Plehn, A. Rousselot, S. Vent [arXiv:2110.XXXXX]



- $x \sim p_{\text{samples}} \rightarrow$ NN $\rightarrow z$

$\rightarrow$ Train NN to ensure $z \sim \mathcal{N}$

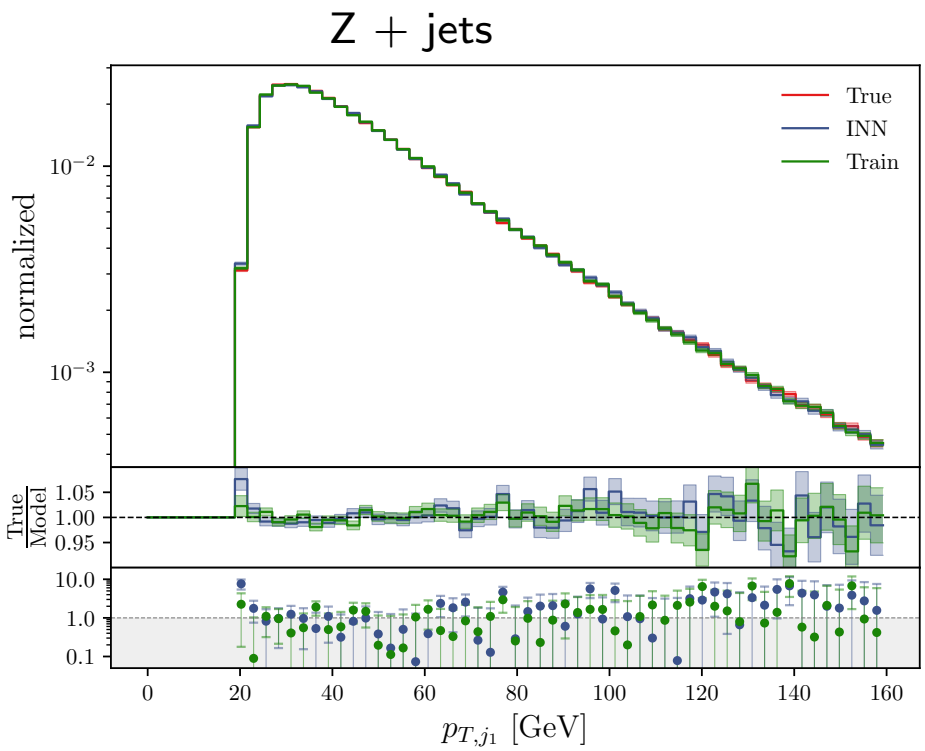
- Loss: Maximize posterior over network weights:

$$
\begin{aligned}
-\log(p(\theta|x)) &= -\log(p(x|\theta)) - \log(p(\theta)) + \text{const.} \\
&= -\log(p(z|\theta)) - \log(J) - \log(p(\theta)) + \text{const.}
\end{aligned}
$$

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline
- Challenges:



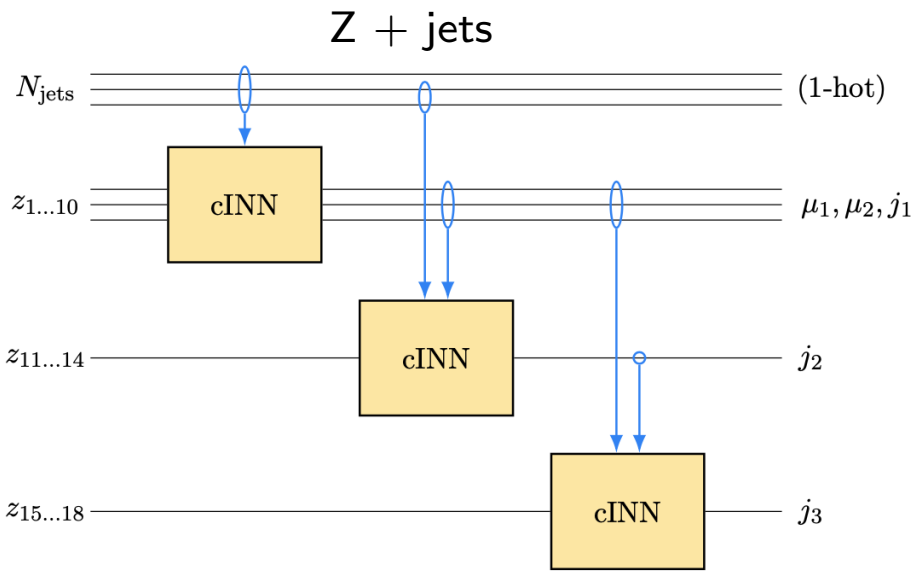
arXiv:2110.XXXXX

A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline
- Challenges:
  - Variable number of jets



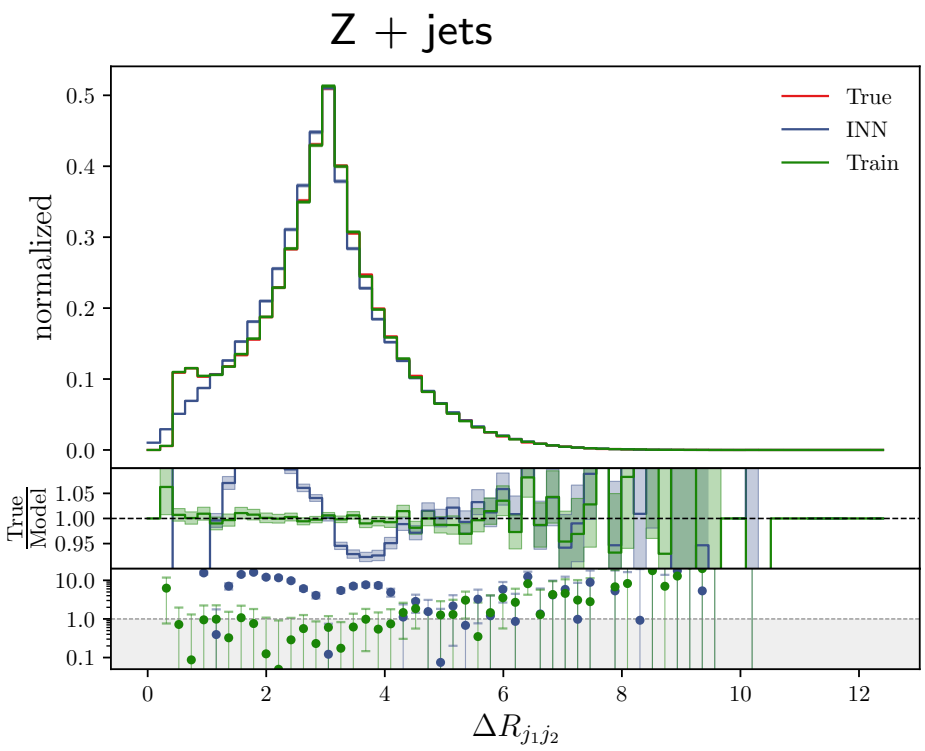
arXiv:2110.XXXXX

A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline
- Challenges:
  - Variable number of jets
  - Topological holes



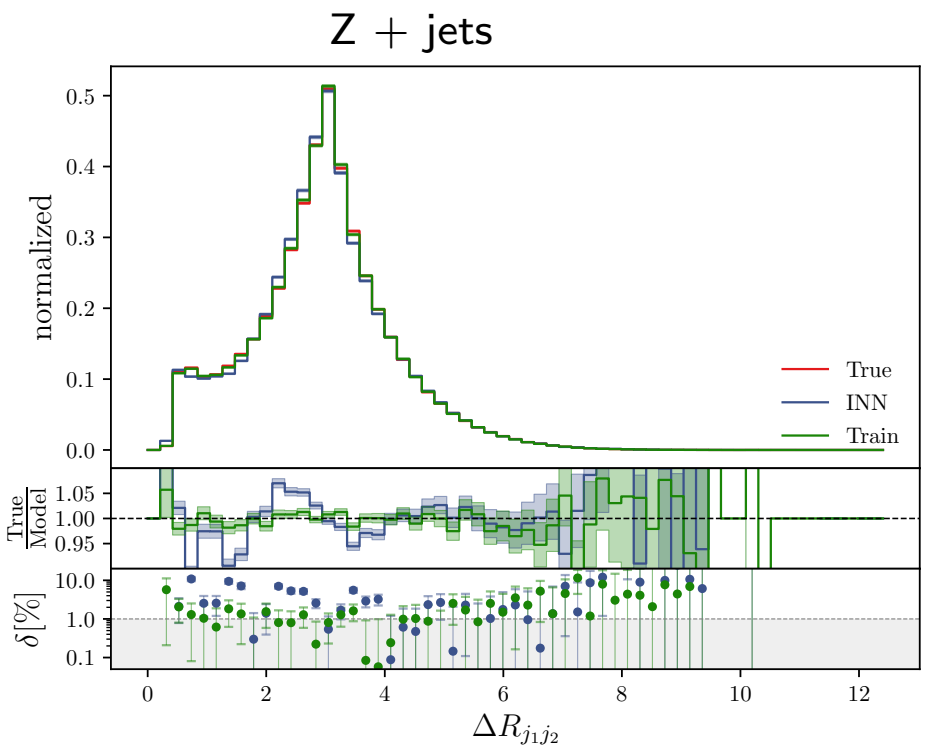
arXiv:2110.XXXXX

A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline
- Challenges:
  - Variable number of jets
  - Topological holes
  - 1% precision



arXiv:2110.XXXXX

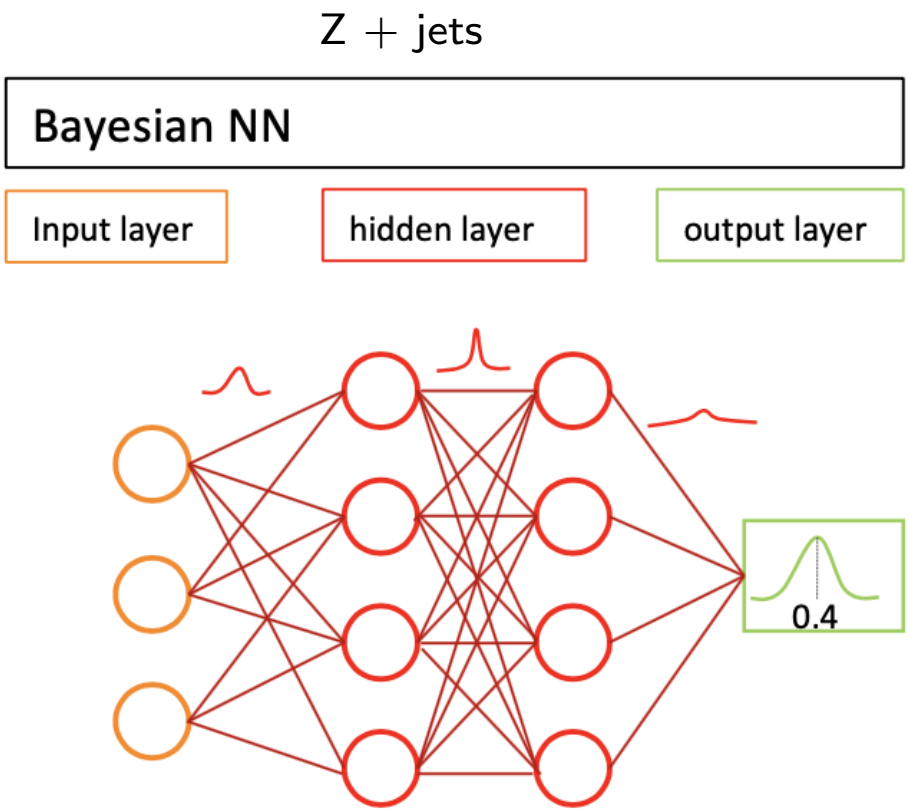
A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline

- Challenges:
  - Variable number of jets
  - Topological holes
  - 1% precision
  - Associated uncertainties



arXiv:2110.XXXXX

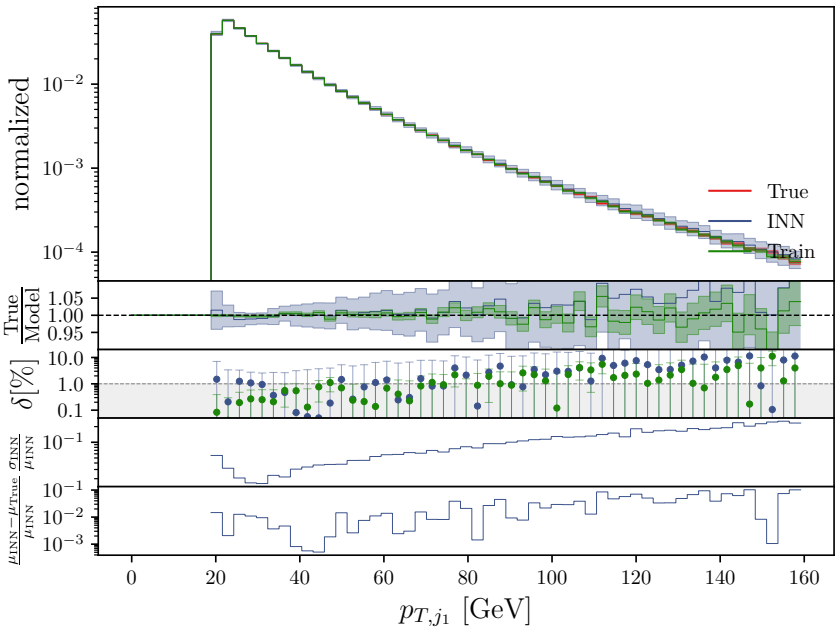
A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

Preliminary

Inclusive Z+jets production

- INN easy trainable, powerful baseline
- Challenges:
  - Variable number of jets
  - Topological holes
  - 1% precision
  - Associated uncertainties

Z + jets



arXiv:2110.XXXXX

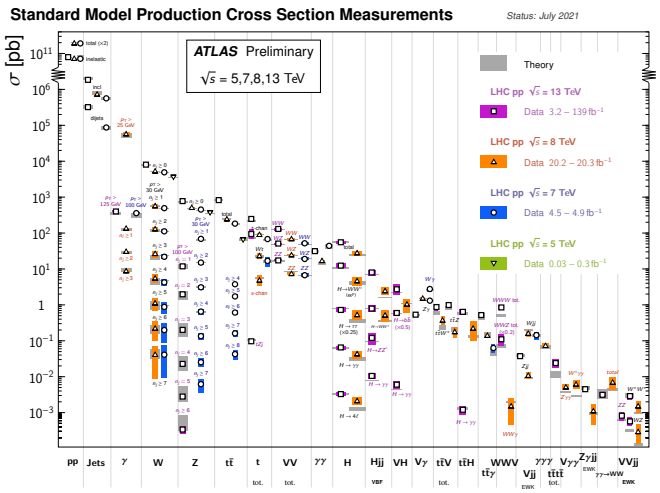
A.B., Theo Heimel, Sander Hummerich, Tobias Krebs, Tilman Plehn, Armand Rousselot, Sophia Vent

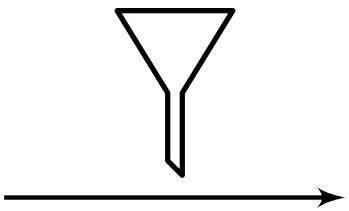
What about ML4Theory?

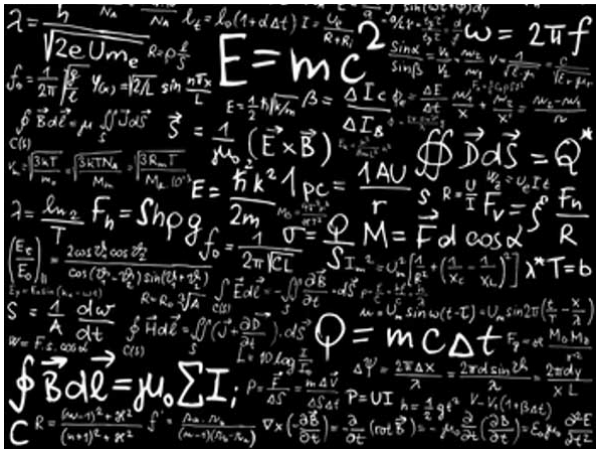
Better predictions?
$\rightarrow$ ML for precision simulations
✓

Better understanding?
$\rightarrow$ Turn data into theory?

Can we learn theory from data?

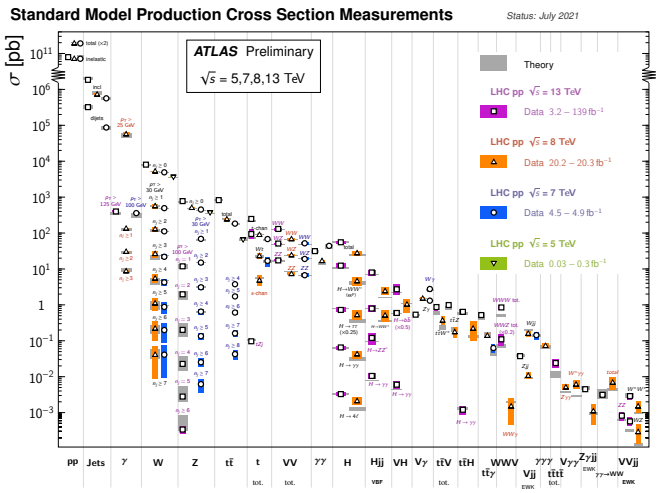


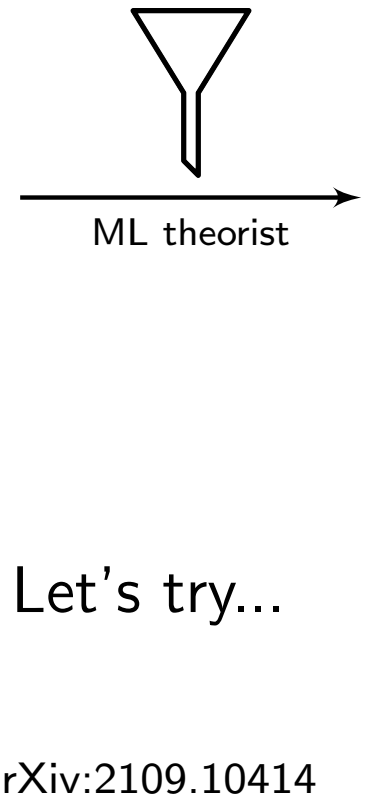


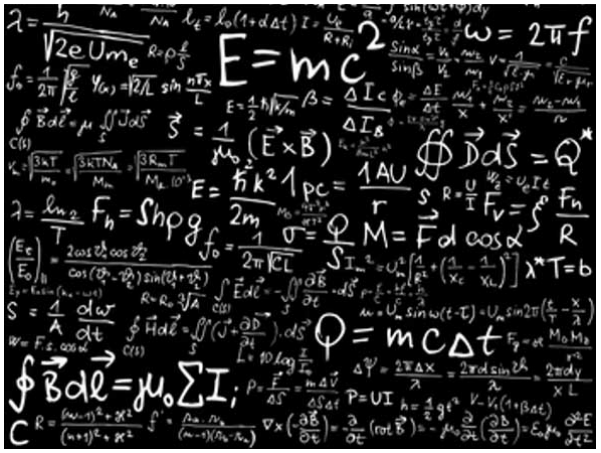


Let's try...

Can we learn theory from data?





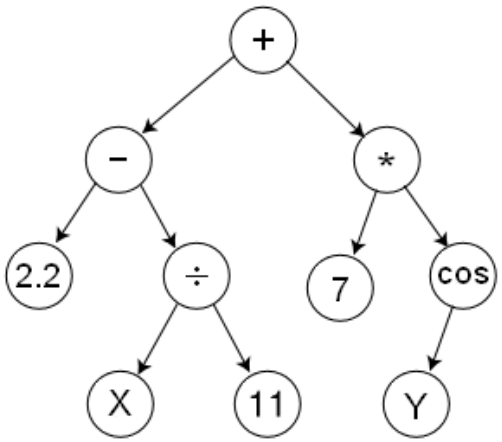


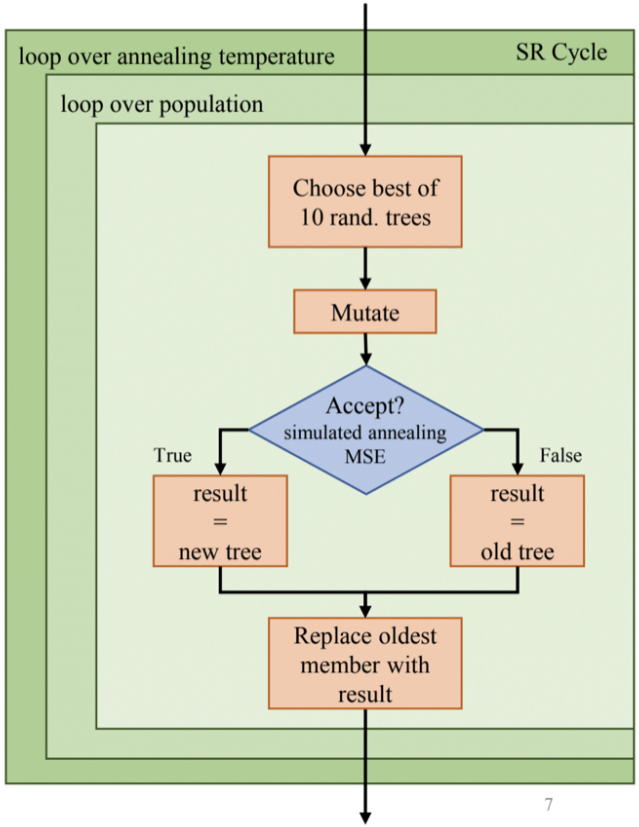
Let's try...

arXiv:2109.10414
Johann Brehmer, A.B., Tilman Plehn, Nathalie Soybelman

Symbolic regression with PySR Miles Cranmer, et al.

Tree representation





$$\texttt{pysr score} = \frac{\texttt{MSE}}{\texttt{baseline}} + \texttt{parsimony} \cdot \texttt{complexity}$$

$$p_{\text{accept}} = \exp\left(-\frac{\texttt{score}_{\text{new}} - \texttt{score}_{\text{old}}}{\texttt{alpha} \cdot T \cdot \texttt{score}_{\text{old}}}\right)$$ ← modified wrt. original PySR

Optimal observable to measure a parameter θ?

- High-dimensional event representation
- 1D representation (p_T, m_{jj}) looses information
- Optimal observable

$$\mathcal{O}_i^{\text{opt}}(x) \equiv t(x|\theta_0) = \left.\frac{\partial \log p(x|\theta)}{\partial \theta_i}\right|_{\theta_0}$$

$\rightarrow$ contains all information on θ

- Problem: $p(x|\theta)$ is untractable

$$p(x_{reco}|\theta) = \int \mathrm{d}z \; p(x_{reco}|z_{det}) p(z_{det}|z_{shower}) p(z_{shower}|z_{parton}) p(z_{parton}|\theta)$$

How to compute the optimal observable

- Solution:
  - Consider joint score

$$t(x,z|\theta) = \nabla_\theta \log p(x,z|\theta)$$

$$= \frac{p(x|z_{det})p(z_{det}|z_{shower})p(z_{shower}|z_{parton})\nabla_\theta p(z_{parton}|\theta)}{p(x|z_{det})p(z_{det}|z_{shower})p(z_{shower}|z_{parton})p(z_{parton}|\theta)}$$

How to compute the optimal observable

- Solution:
  - Consider joint score

$$t(x,z|\theta) = \nabla_\theta \log p(x,z|\theta)$$

$$= \frac{\nabla_\theta p(z_{parton}|\theta)}{p(z_{parton}|\theta)} \qquad \text{with } p(z_{parton}|\theta) = \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}\theta}$$

How to compute the optimal observable

- Solution:
  - Consider joint score

$$
\begin{aligned}
t(x,z|\theta) &= \nabla_\theta \log p(x,z|\theta) \\
&= \frac{\nabla_\theta p(z_{parton}|\theta)}{p(z_{parton}|\theta)} \qquad \text{with } p(z_{parton}|\theta) = \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}\theta} \\
&= \frac{\nabla_\theta |\mathcal{M}(z|\theta)|^2}{|\mathcal{M}(z|\theta)|^2} - \frac{\nabla_\theta \sigma_{\text{tot}}(\theta))}{\sigma_{\text{tot}}(\theta)}
\end{aligned}
$$

How to compute the optimal observable

- Solution:
  - Consider joint score

$$
\begin{aligned}
t(x,z|\theta) &= \nabla_\theta \log p(x,z|\theta) \\
&= \frac{\nabla_\theta p(z_{parton}|\theta)}{p(z_{parton}|\theta)} \qquad \text{with } p(z_{parton}|\theta) = \frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}\theta} \\
&= \frac{\nabla_\theta |\mathcal{M}(z|\theta)|^2}{|\mathcal{M}(z|\theta)|^2} - \frac{\nabla_\theta \sigma_{\text{tot}}(\theta))}{\sigma_{\text{tot}}(\theta)}
\end{aligned}
$$

  - Score is given by

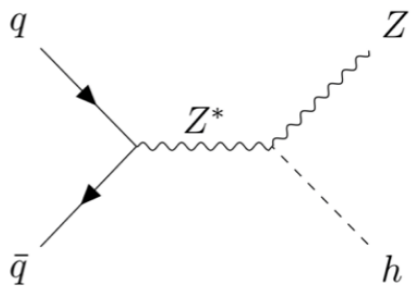
$$
t(x|\theta) = \arg\min_{g(x)} \mathcal{E}_{x,z\sim p(x,z|\theta)} \left|g(x) - t(x,z|\theta)\right|^2
$$

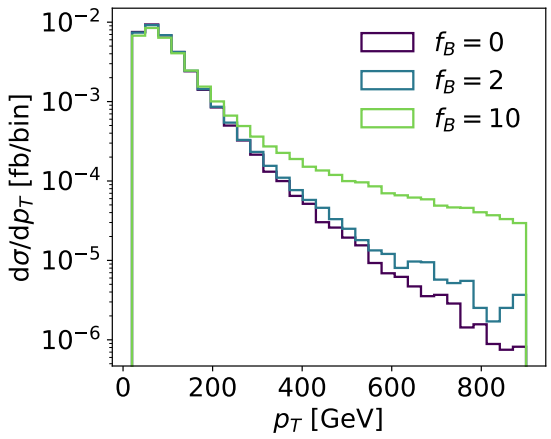
- Option 1: Minimization with NN $\rightarrow$ SALLY
- *new* Option 2: Learn formula to minimize $g(x)$

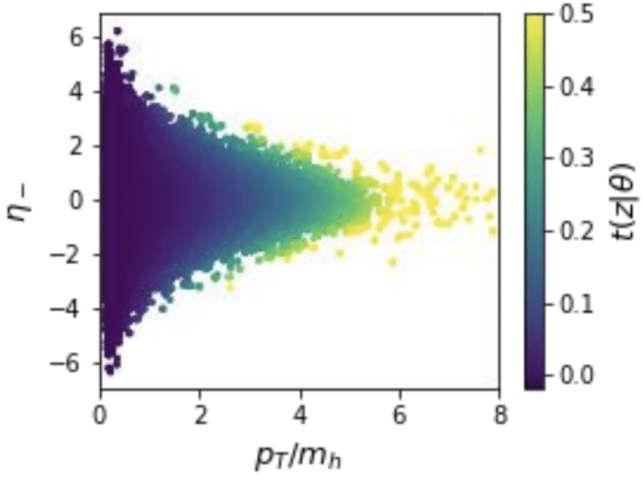
ZH production

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \frac{f_B}{\Lambda^2}\mathcal{O}_B \qquad \text{with}$$

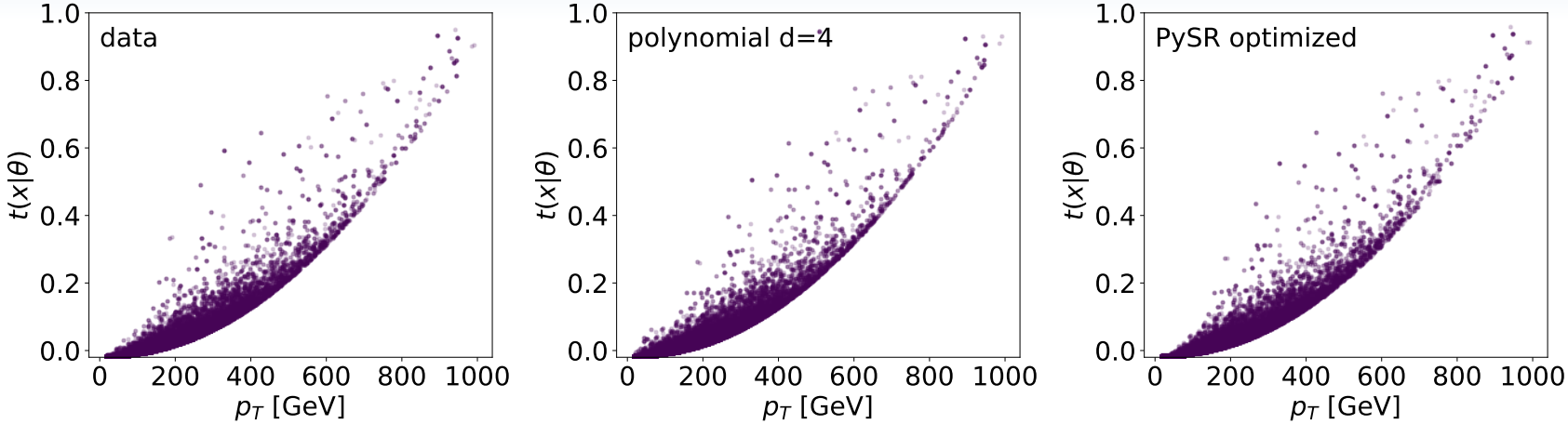
$$\mathcal{O}_B = \frac{ig'}{2}(D^\mu\phi)^\dagger D^\nu\phi B_{\mu\nu}$$







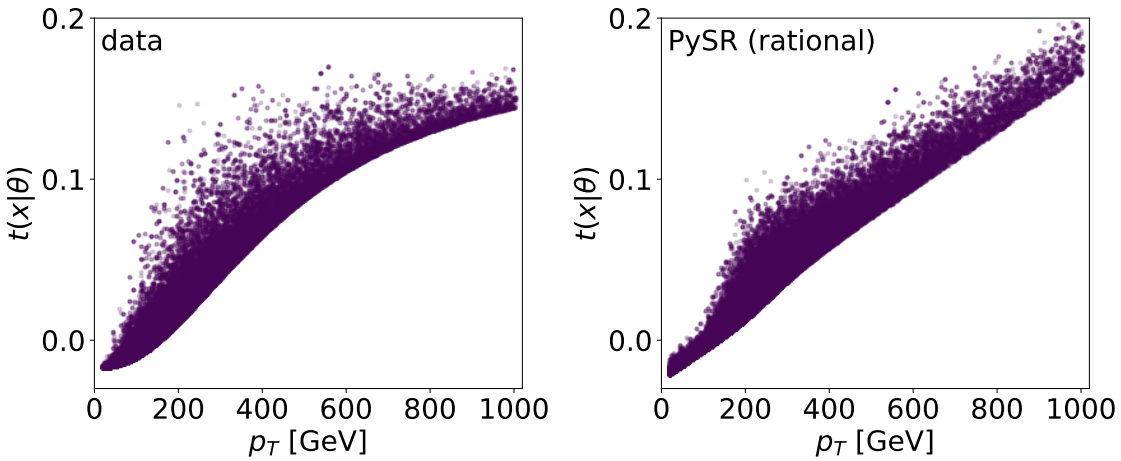
Results for simplified process $f_B = 0$



	polynomial $d = 2$	polynomial $d = 4$	PySR	PySR optimized
MSE	$3.49 \cdot 10^{-3}$	$1.28 \cdot 10^{-4}$	$1.23 \cdot 10^{-4}$	$7.65 \cdot 10^{-5}$
dof	6	15	9	9

cmpl	dof	function	MSE
7	1	$ax_p(x_p + x_\eta)$	$3.81 \cdot 10^{-2}$
10	3	$ax_p^2(b + x_\eta) - c$	$2.49 \cdot 10^{-3}$
14	3	$ax_p^2 + bx_p^2x_\eta^2 - c$	$6.64 \cdot 10^{-4}$
22	4	$ax_p^2 + bx_p^2x_\eta^2 - cx_px_\eta - d$	$3.09 \cdot 10^{-4}$
32	6	$a(x_p^2 + x_\eta) + bx_p^2x_\eta - (cx_p - d)^2 + ex_p^2x_\eta^3 - f$	$2.06 \cdot 10^{-4}$
34	7	$a(x_p^2 + x_\eta) + bx_p^2x_\eta - (cx_p - d)^2 + ex_\eta^3(x_p - f)^2 - g$	$7.77 \cdot 10^{-5}$
49	9	$ax_p^2 + bx_p^2x_\eta - cx_\eta(x_p - d) + ex_\eta^3(x_p - f)^2 + gx_p^2x_\eta^2 - hx_p - i$	$7.65 \cdot 10^{-5}$

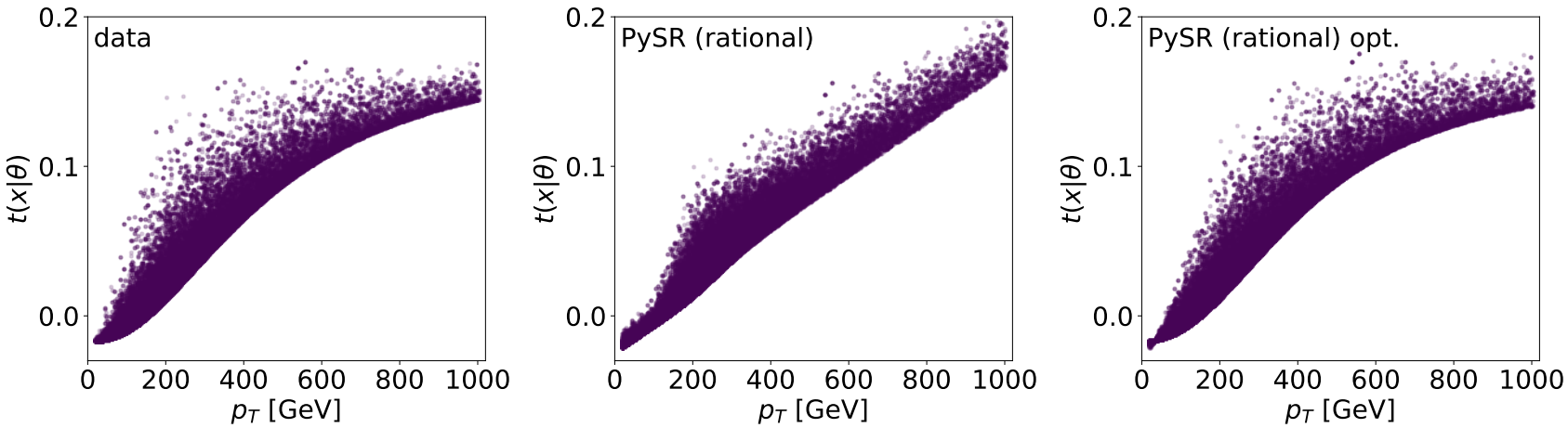
Learning complex functions for $f_B = 10$



Hall of fame „winner“:

$$t(x_p, x_\eta | f_B = 10) = a x_p - b + \frac{c(x_\eta + d)}{e + \frac{f}{x_p \left((x_p - g)^4 + h\right)\left(i\,(x_\eta - j)^2 + k\right)}}$$

Learning complex functions for $f_B = 10$



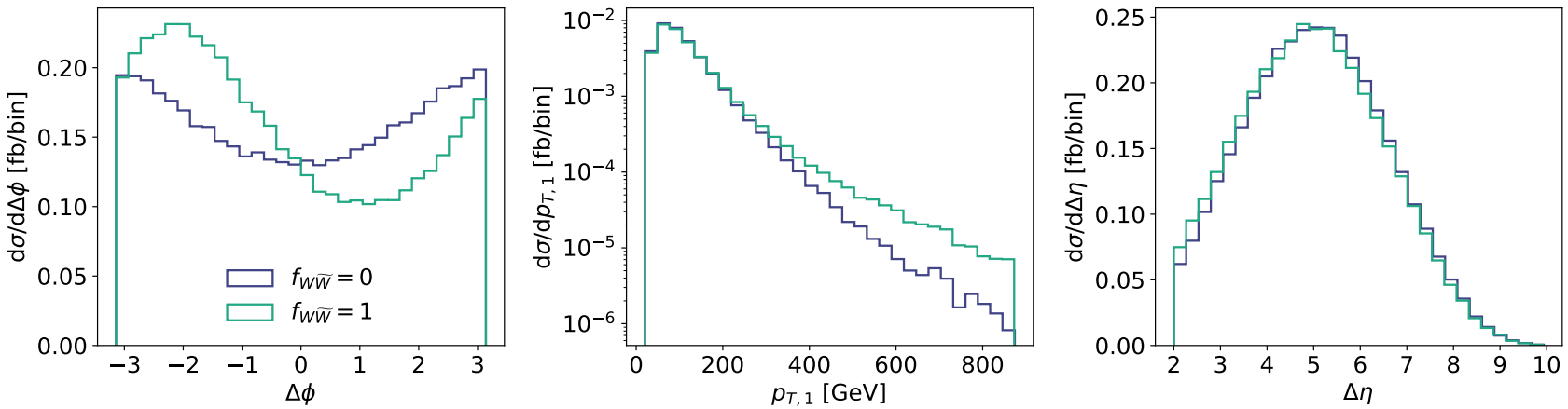
Hall of fame „winner“:

$$t(x_p, x_\eta | f_B = 10) = a x_p - b + \frac{c' x_\eta + d'}{1 + \frac{f'}{x_p \left(h' \left(x_p + g\right)^4 - 1\right) \left(i' \left(x_\eta - j\right)^2 + 1\right)}}$$

Remove flat direction for optimization stability

WBF Higgs production with CP violation

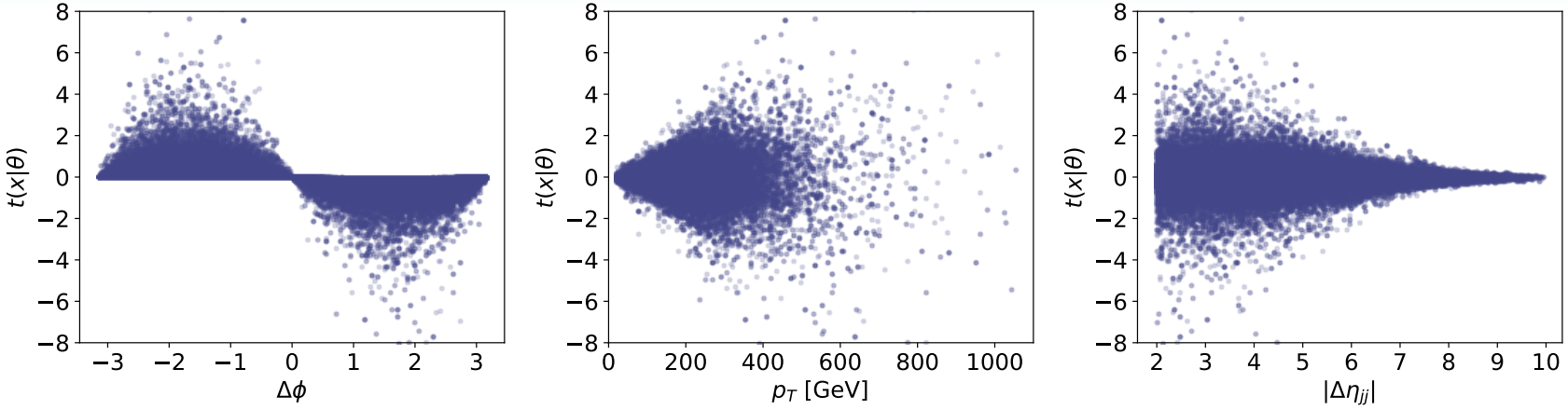
$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \frac{f_{W\widetilde{W}}}{\Lambda^2}\mathcal{O}_{W\widetilde{W}} \qquad \text{with} \qquad \mathcal{O}_{W\widetilde{W}} = -(\phi^\dagger\phi)\ \widetilde{W}^k_{\mu\nu}W^{\mu\nu k}$$



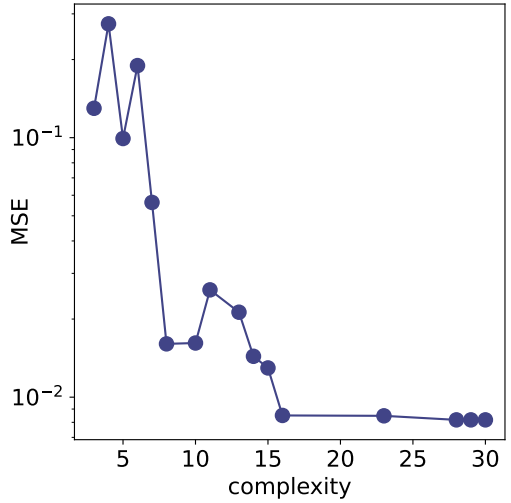
Approximation for leading partonic contribution:

$$t(x|f_{W\widetilde{W}} = 0) \approx -\frac{8v^2}{m_W^2}\frac{2E_+E_- + (p_+p_-)}{(p_+p_-)}p_{T+}p_{T-}\ \sin\Delta\phi$$

Result $f_{W\widetilde{W}} = 0$



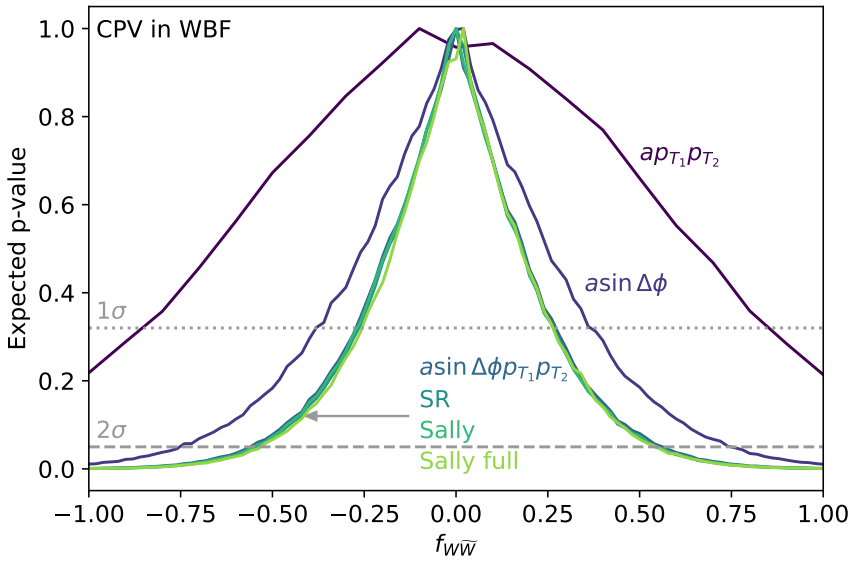
compl	function	MSE
3	$a\,\Delta\phi$	$1.30 \cdot 10^{-1}$
4	$\sin(a\Delta\phi)$	$2.75 \cdot 10^{-1}$
5	$a\Delta\phi x_{p,1}$	$9.93 \cdot 10^{-2}$
6	$-x_{p,1}\sin(\Delta\phi + a)$	$1.90 \cdot 10^{-1}$
8	$(a - x_{p,1})x_{p,2}\sin(\Delta\phi)$	$1.61 \cdot 10^{-2}$
16	$-x_{p,1}(a - b\Delta\eta)(x_{p,2} + c)\sin(\Delta\phi + d)$	$8.50 \cdot 10^{-3}$
28	$(x_{p,2} + a)(bx_{p,1}(c - \Delta\phi)$ $-x_{p,1}(d\Delta\eta + ex_{p,2} + f)\sin(\Delta\phi + g))$	$8.18 \cdot 10^{-3}$



$$t(p_{T,j_1}, p_{T,j_2}, \Delta\phi, \Delta\eta | f_{W\widetilde{W}} = 0) = -p_{T,j_1}\left(p_{T,j_2} + c\right)(a - b\Delta\eta)\sin(\Delta\phi + d)$$

with $a = 1.086(11)$ $b = 0.10241(19)$ $c = 0.24165(20)$ $d = 0.00662(32)$

Including detector effects

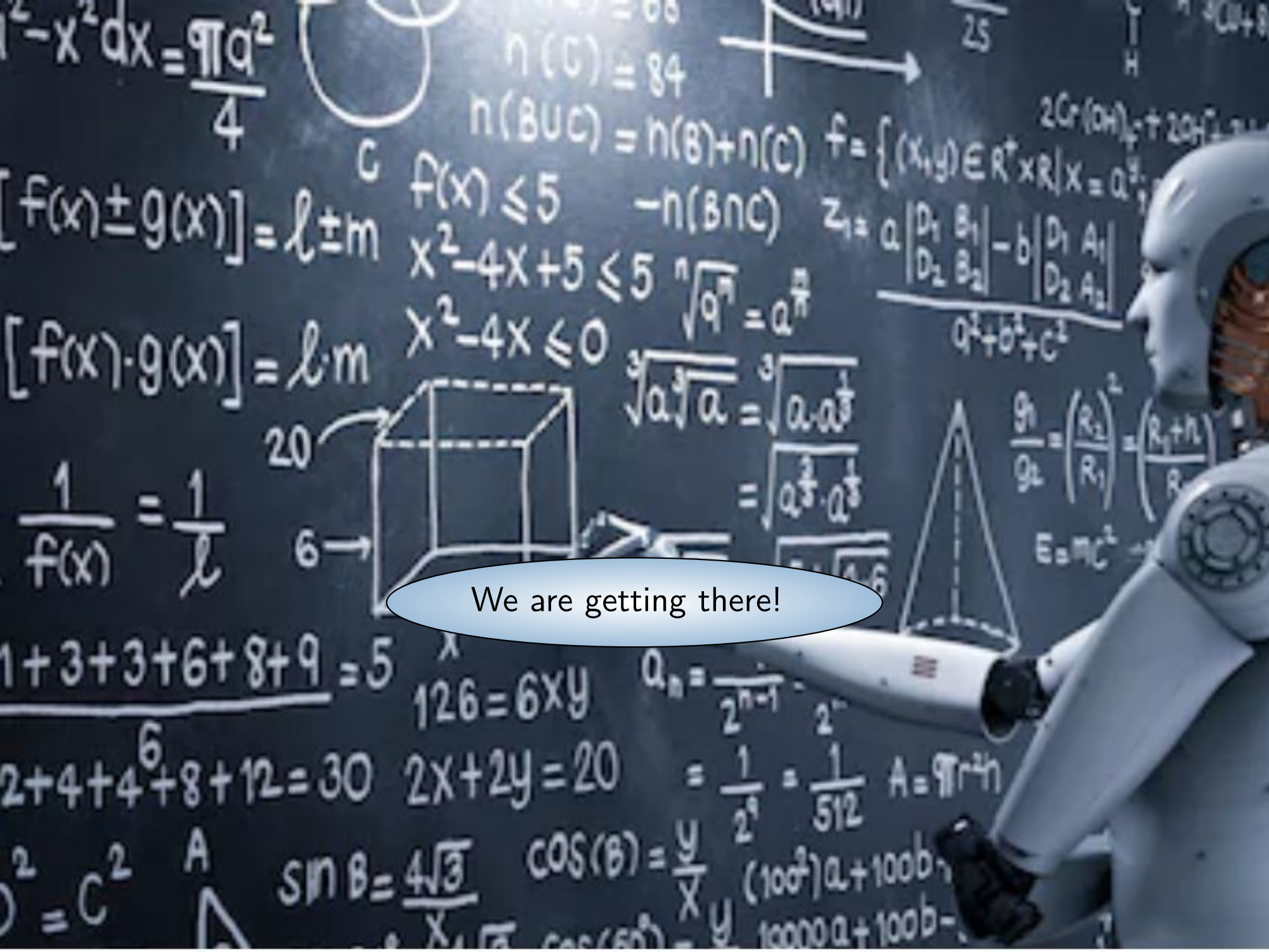


(optimal) observable	MSE all	reach 1 σ	reach 2 σ
$ap_{T1}p_{T2}$	0.1576	[-0.86,0.86]	—
$a\sin\phi$	0.0885	[-0.38,0.36]	[-0.76,0.74]
$a\sin\phi p_{T1}p_{T2}$	0.0217	[-0.28,0.28]	[-0.56,0.56]
SR complexity 16	0.0145	[-0.26,0.26]	[-0.54,0.54]
SALLY	0.0129	[-0.26,0.26]	[-0.56,0.54]
SALLY full	0.0048	[-0.26,0.26]	[-0.54,0.54]

What about ML4Theory?

Better predictions?
$\rightarrow$ ML for precision simulations
✓

Better understanding?
$\rightarrow$ Turn data into theory
(✓)


We are getting there!