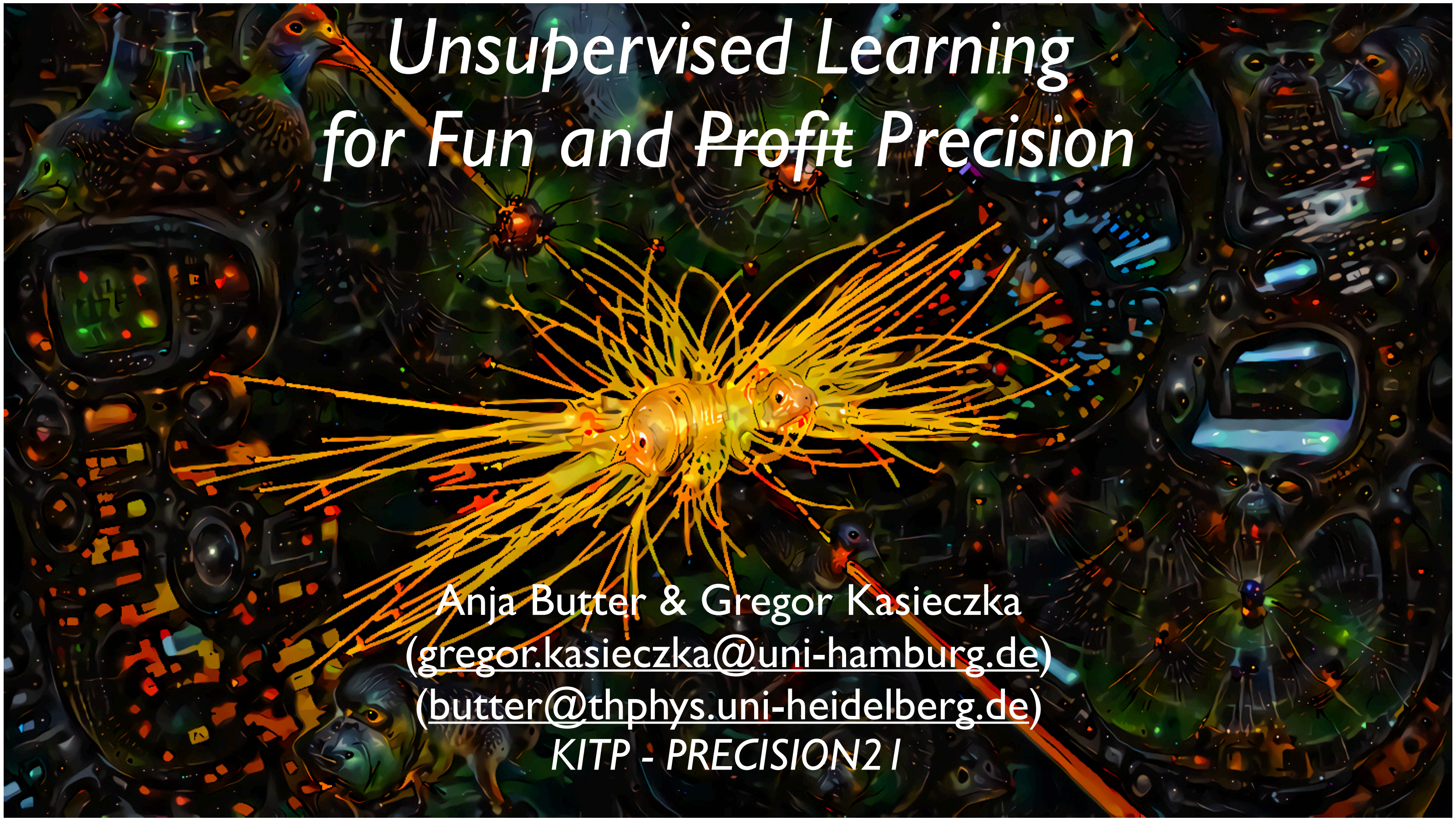


Unsupervised Learning for Fun and ~~Profit~~ Precision

Anja Butter & Gregor Kasieczka
(gregor.kasieczka@uni-hamburg.de)
(butter@thphys.uni-heidelberg.de)

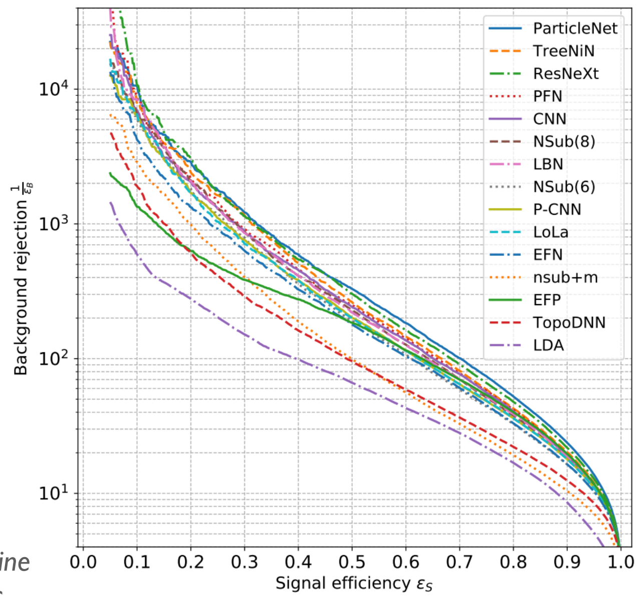
KITP - PRECISION21

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE
PIER
Partnership of Universität Hamburg and DESY
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft DFG
Bundesministerium für Bildung und Forschung

Motivation

- Large excitement for machine learning in particle physics:
  - Particle tagging / signal selection
  - Low level reconstruction / calibration
  - Simulation

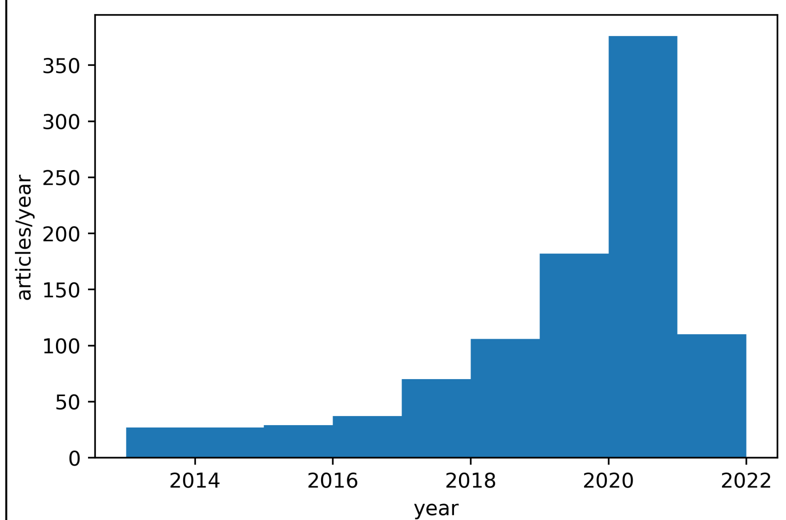
… and many more



GK, Plehn (eds), et al, *The Machine Learning Landscape of Top Taggers*, 1902.09914

Inspire Search:

("machine learning" or "deep learning" or neural) and (hep-ex or hep-ph or hep-th)

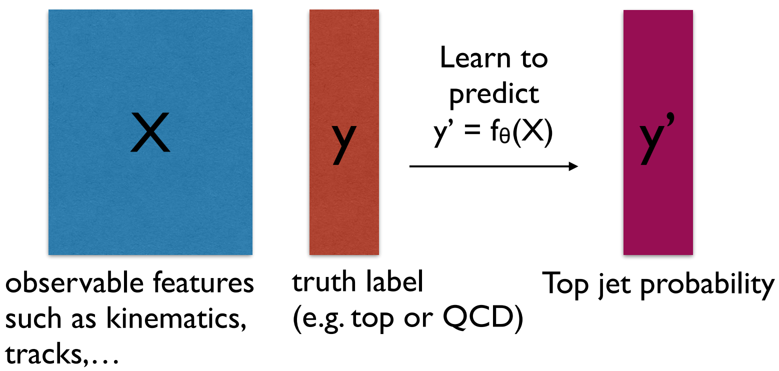


Two* types of problem:

Supervised Learning

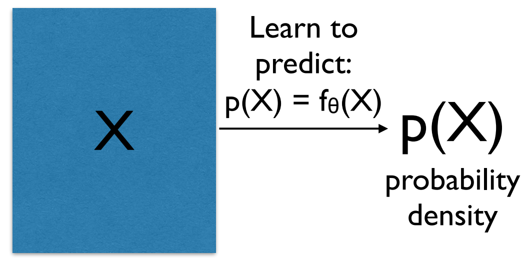
Attempt to infer some target *(truth label)*: classification (*jet flavour tagging*) or regression (*energy calibration*)

Use training data with known labels (often from Monte Carlo simulation)



Unsupervised

No target, learn the probability distribution (directly from data)



Maximize likelihood p(X)
(minimize -log p(x))

There also exists a number of other less-than-supervised approaches (weakly supervised learning, semi-supervised learning, …) Not so important for now.

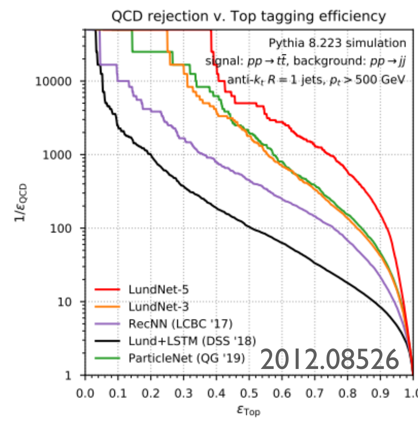
Supervised

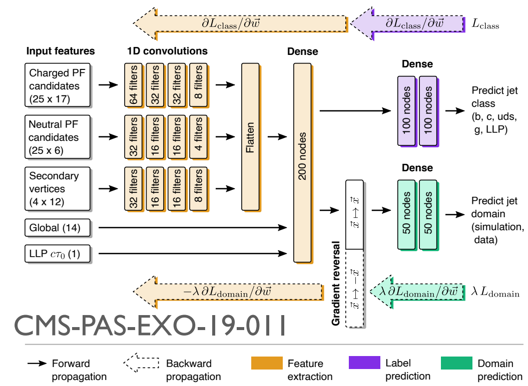
Most early works fall under this category.

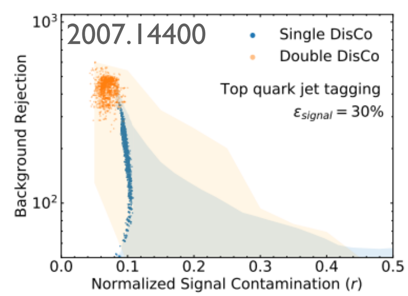
Crucially important for large number of tasks.

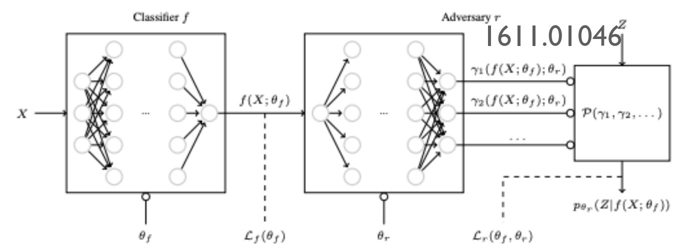
Need:

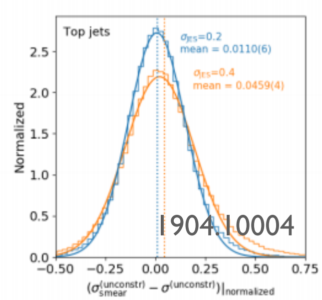
- Higger accuracy (easy to measure, many results)
- Better stability (domain adaptation issue)
- More control over uncertainties
- Resource efficient implementations
- Experimental integration
- …

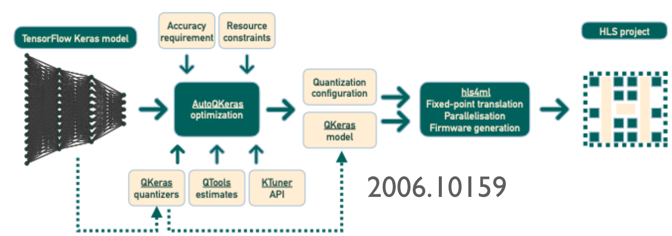












Unsupervised

Exciting space for developing new ideas
(also including all other forms of less-than-supervised learning).

Topic of our talks today.

Part I (Gregor):
How can we use a learned $p(X)$ to find new physics?
Anomaly detection // Model independent searches

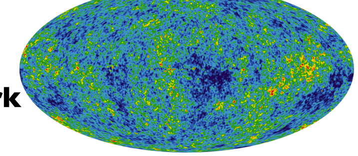
Part II (Anja):
How can we efficiently sample from $p(X)$?
Generative Models // Fast simulation

Anomaly Searches

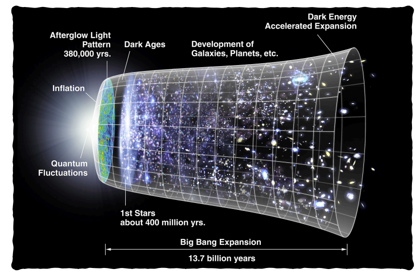
Physics Motivation

- Theoretical and experimental reasons to expect new physics beyond the Standard Model
- However, only negative results in searches
- Make sure that we do not miss potential discoveries at the LHC:
  Supplement traditional searches with model-independent anomaly searches*

Why are neutrinos massive?

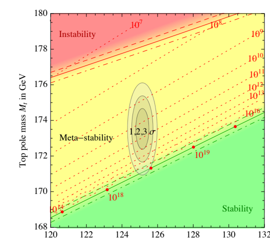


What is the nature of dark matter & dark energy?



What are the origins of the LHCb flavour anomaly?

Why is there more matter than anti-matter?



Why is there more matter than anti-matter?

Is the electroweak vacuum stable?

How can the Higgs boson be light when the mass receives large quantum corrections?

What are the details of cosmic inflation?

****Tricky term, will discuss meaning of model-independence later***

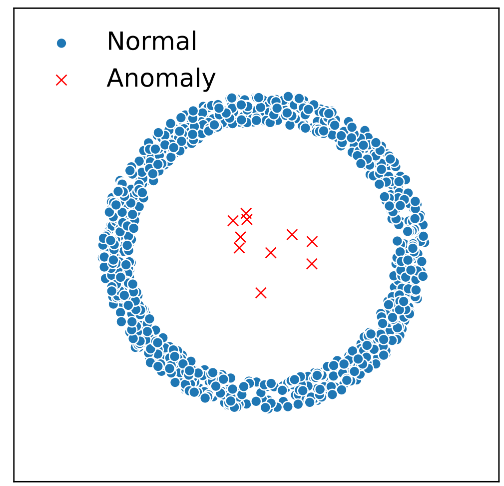
Dissecting the problem

- Zero: *What are anomalies*
- First: *Build an anomaly scoring function* ***a****(X)*
- Second: *Design analysis strategy*
- Third: *Interpret result*

What is an anomaly?

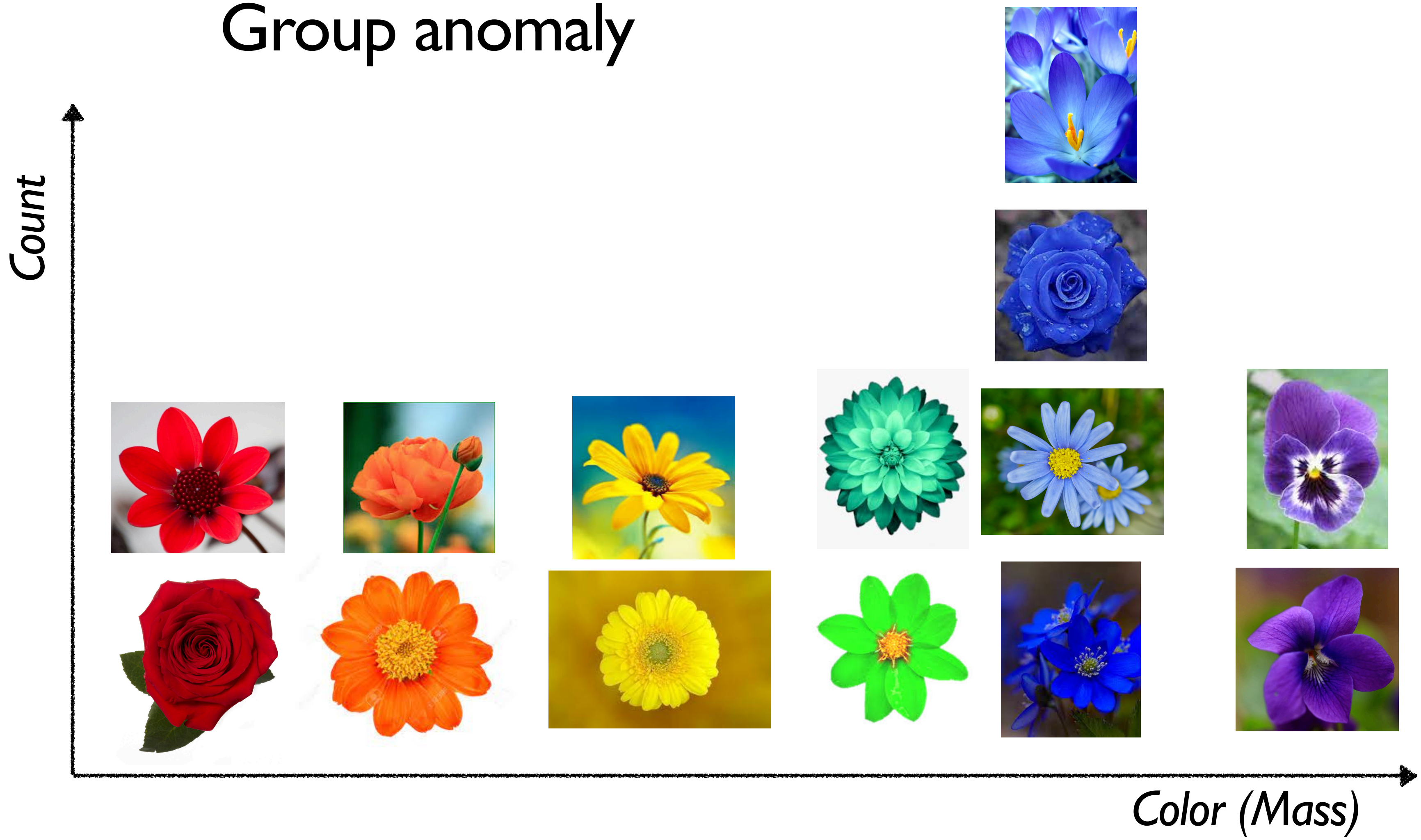
Point anomaly

- Outliers: Datapoints far away from regular distribution
- Examples:
  - Background free searches (e.g. long lived particles)
  - Detector malfunctions



And now?

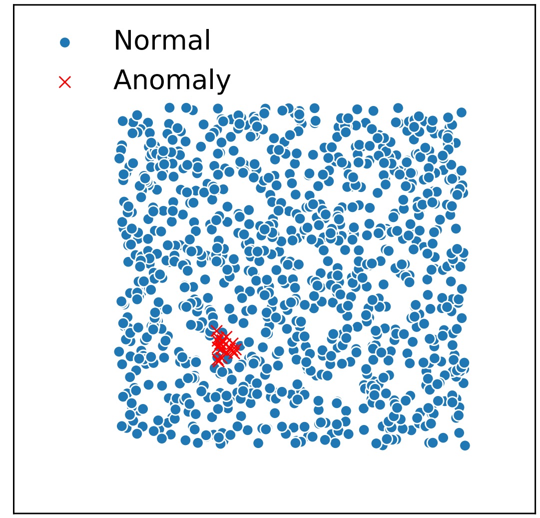
Group anomaly

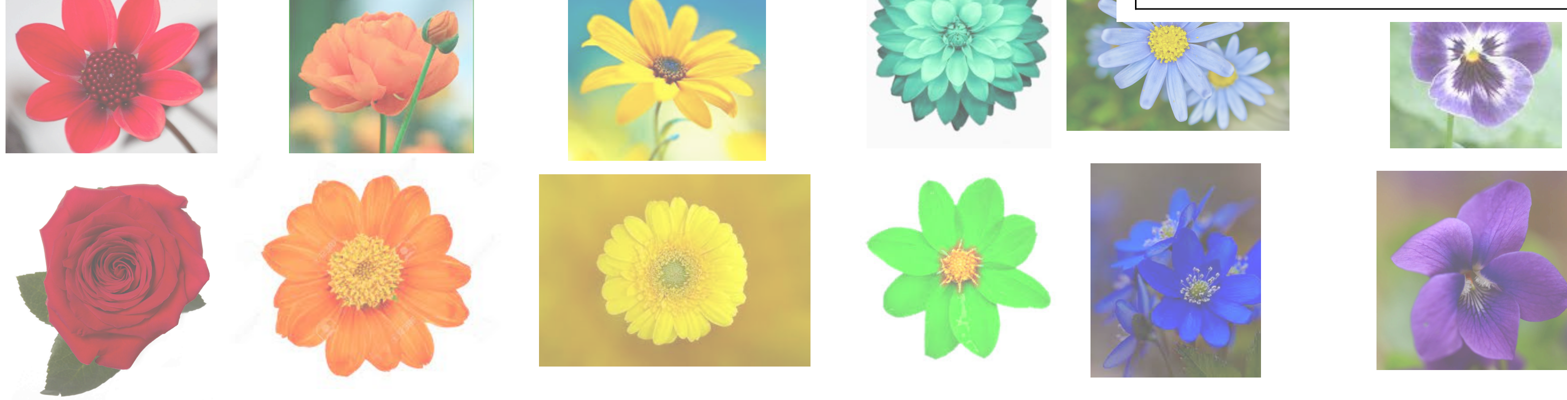


Group anomaly

Count

- Individual examples not anomalous, but interesting collective behaviour
- Examples:
  - New physics searches, e.g. resonances





Color (Mass)

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events

- Some options:

  - **a(x) = (Semi-) Supervised**

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like)
  and low for background-like events

- Some options:

  - **a(x) = (Semi-) Supervised**

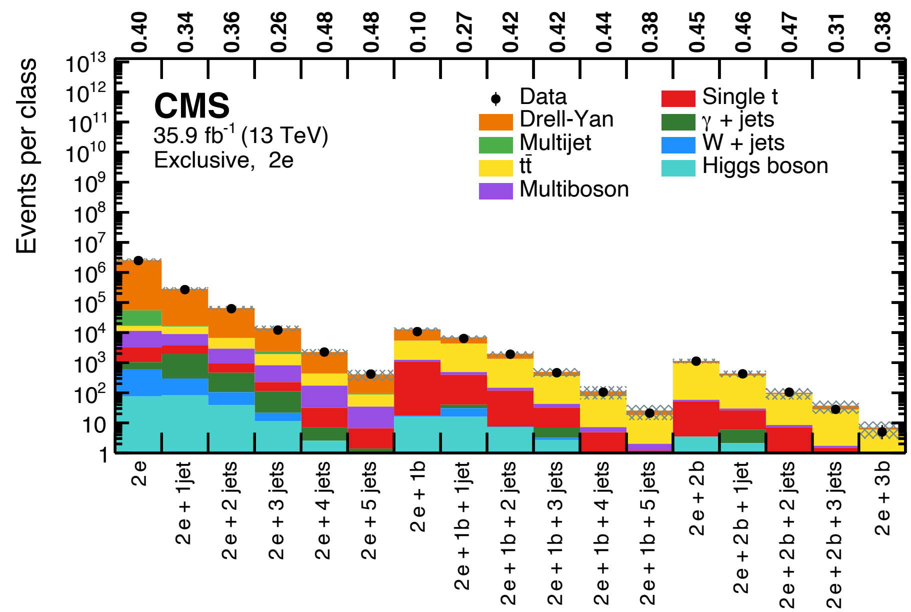
- Train binary classifier network (using simulation) to discriminate:
  - Standard Model background
    vs
  - Cocktail of new physics models

- Pros:
  - Close to known methods, simple training
  - Clear trade-off: width vs sensitivity
- Cons:
  - Ambiguity on mixture choice
  - Needs to account for residual difference between data/simulation

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like)
  and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - **a(x) = 1 / p(x|Background) *(from simulation)***

- Systematically look for differences between background simulation and data
- MUSIC / General search



CMS Collaboration, *MUSiC: a model unspecific search for new physics in proton-proton collisions at sqrt(s)=13 TeV*, 2010.02984

How to build anomaly score?

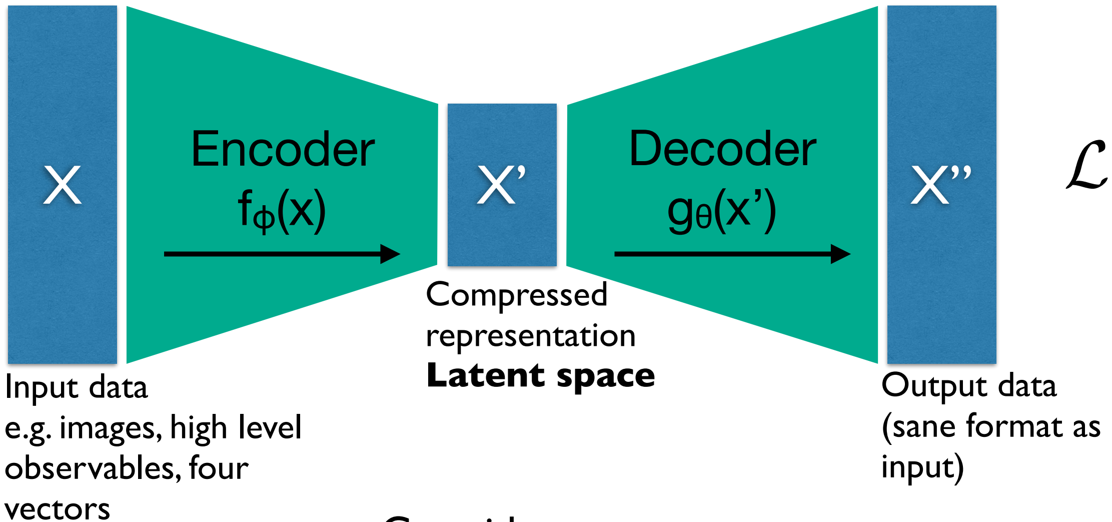
- Anomaly score **a** should be high for anomalous (signal-like)
  and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - **a(x) = 1 / p(x|Background)** ***(from simulation)***

- Systematically look for differences between background simulation and data
- MUSIC / General search
  - Use histograms of many variables in many dimensions to estimate
  - Potentially also improve via ML
- Pros:
  - Very signal model independent
  - Already delivering results
- Cons:
  - Strongly depends on background simulation
  - Large penalty from many histogram bins

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - **a(x) = 1 / p(x|Background)**
    (from simulation)
    (from data)
- Search differences between different phase space regions in data
- Show example using **autoencoders**

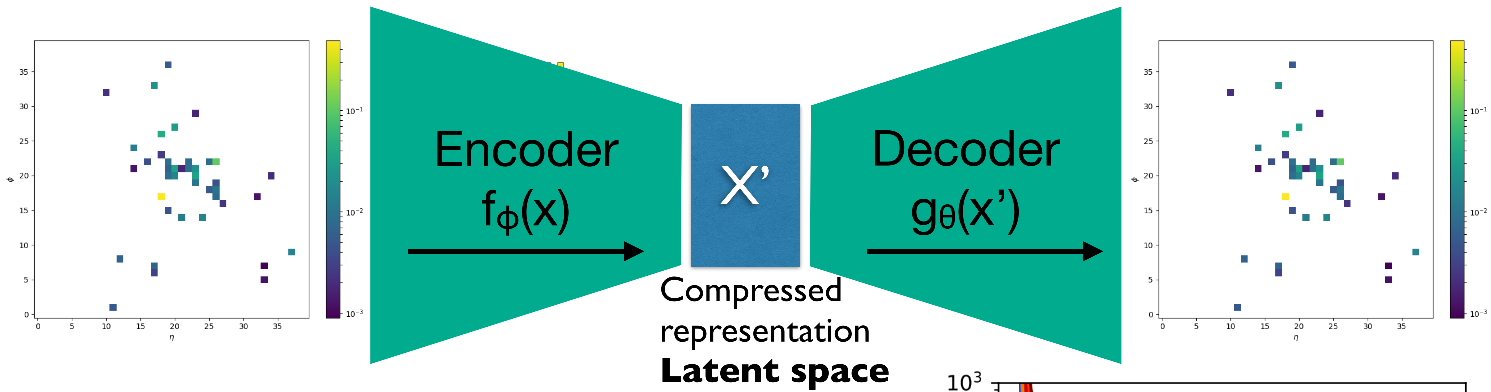
Example: Autoencoder



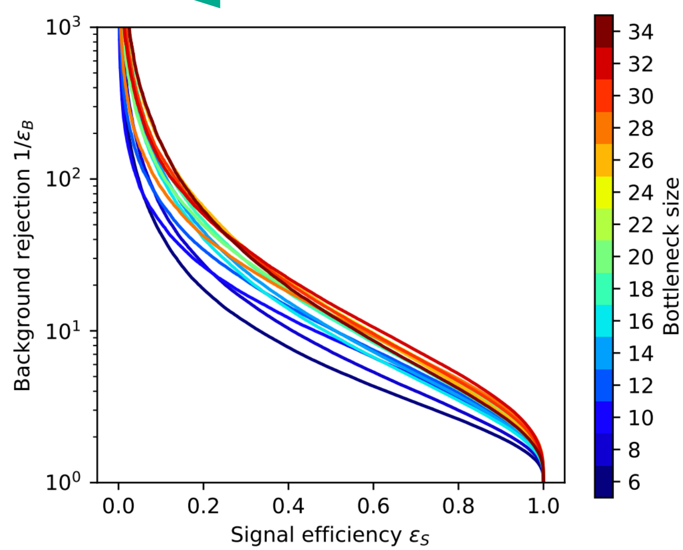
$$\mathcal{L}(x) = ||x - g_\theta(f_\phi(x))||_2$$
$$a(x) = \mathcal{L}(x)$$

- Core idea:
  - Train lossy compression algorithm on anomaly-free data (minimise *L*)
  - Apply to data containing potential anomalies
- Expect quality to decrease for atypical examples: anomaly score

Apply to jet images



- Represent data as images
  - Boosted top vs QCD jets (~600 GeV)
    1 jet = 1 image (40x40 pixels, color=energy)
- Train QCD only sample
- Evaluate on mixed top/QCD jet sample
  - **Tops detected as anomaly**

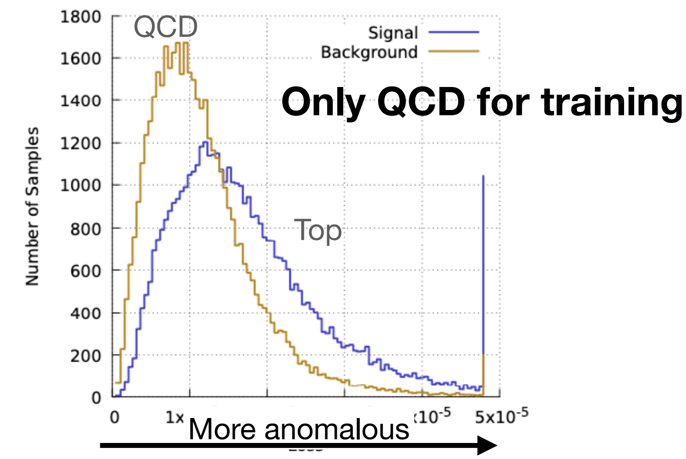


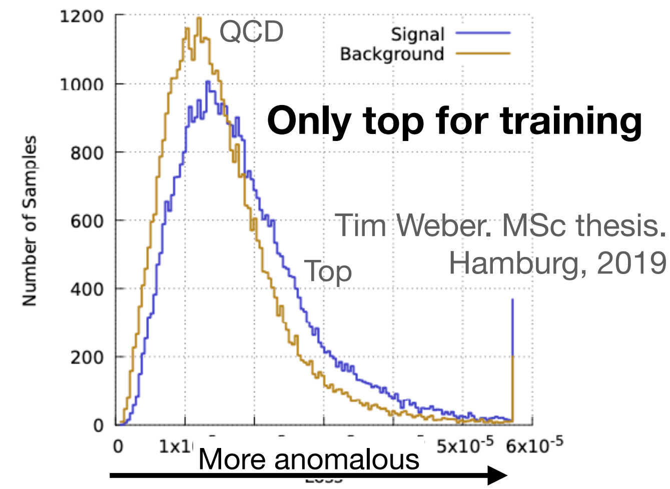
Heimel, GK, Plehn, Thompson, *QCD or What?*, 1808.08979
Farina, Nakai, Shih, *Searching for New Physics with Deep Autoencoders, 1808.08992*

Limitations

Complexity

- If anomalies are much simpler (therefore easier to reconstruct):
  a(x) will still be lower, despite never encountered in training
- Observed with naive AE in QCD vs top
  - Train on tops only; top still considered anomaly wrt/ QCD



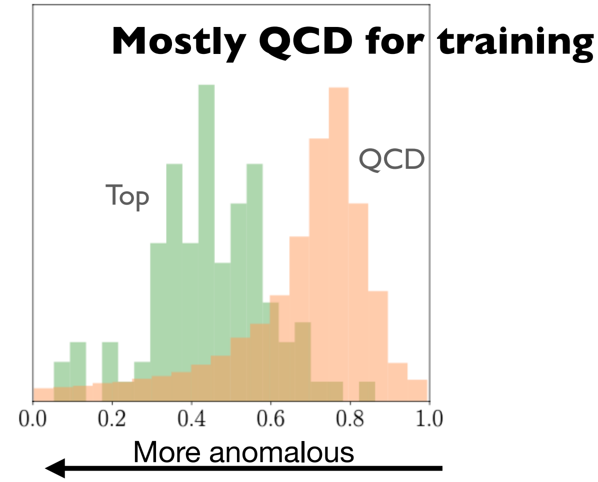


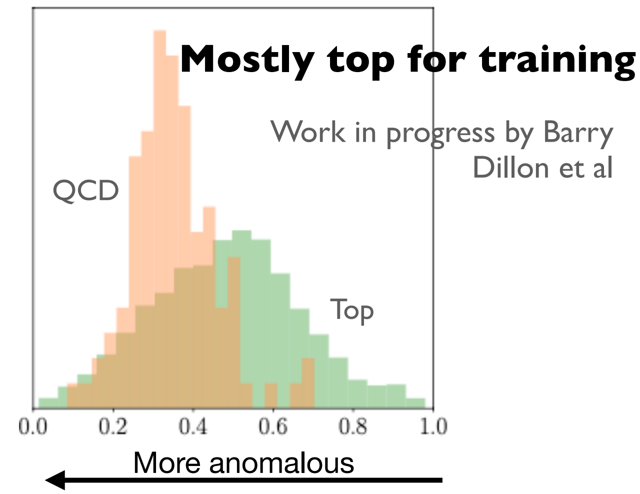
Limitations

Complexity

- If anomalies are much simpler (therefore easier to reconstruct):
  a(x) will still be lower, despite never encountered in training
- Observed with naive AE in QCD vs top
  - Train on tops only; top still considered anomaly wrt/ QCD

Hope that this can be overcome with alternative AE trainings: Stay tuned for update by Heidelberg group using mixture model latent space!





Limitations

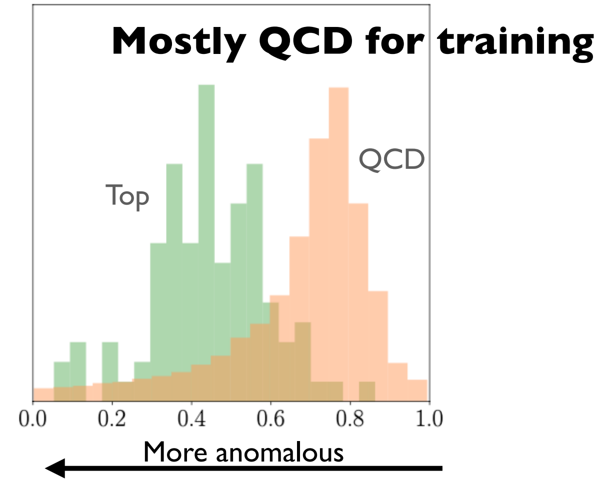
Complexity

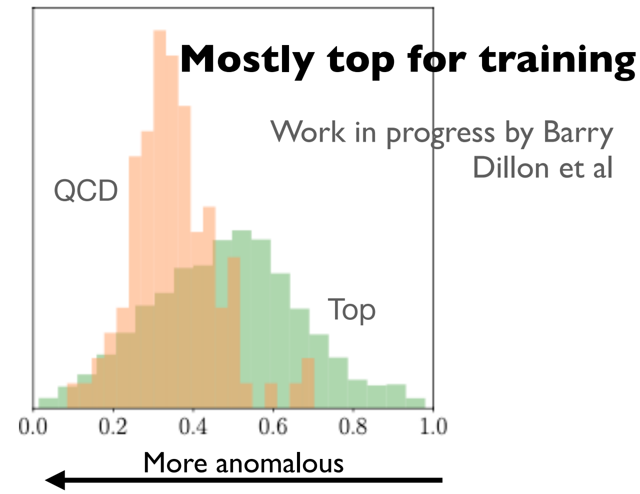
- If anomalies are much simpler (therefore easier to reconstruct):
  $a(x)$ will still be lower, despite never encountered in training
- Observed with naive AE in QCD vs top
  - Train on tops only; top still considered anomaly wrt/ QCD

Hope that this can be overcome with alternative AE trainings: Stay tuned for update by Heidelberg group using mixture model latent space!

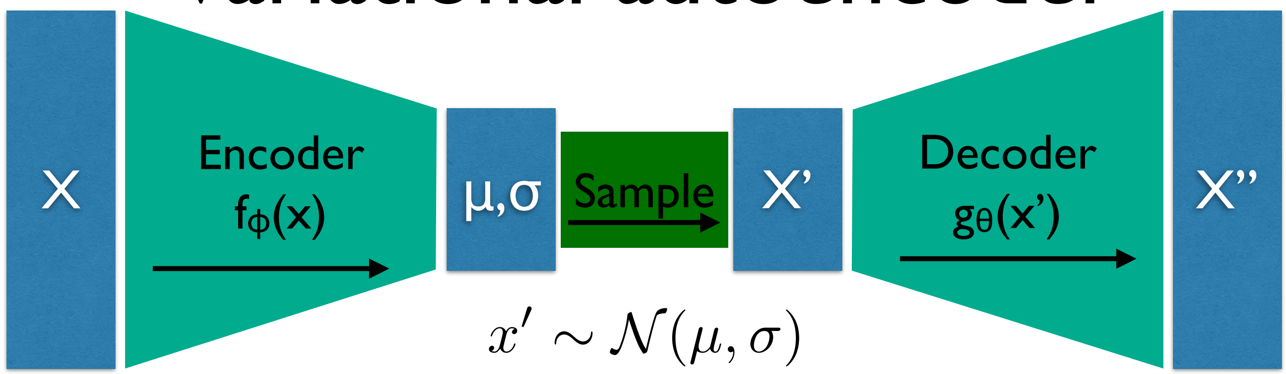
Topology

- Additional potential difficulty if data space has a non-trivial global topology.
  See 2102.08380 for more





Brief aside on generative models: Variational autoencoder



$$x' \sim \mathcal{N}(\mu, \sigma)$$

- The decoder maps a latent space distribution X' to realistic examples
- Control over latent space:
  - Decode X' to generate new examples

- Achieve by:
  - Make X' Gaussian, encoder learns paramaters μ,σ
  - Add term to loss so that (μ,σ) approach standard normal (0,1)

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - **a(x) = 1 / p(x|Background)**
    (from simulation)
    (from data)

- Search differences between different phase space regions in data
- Show example using **autoencoders**
- Pros:
  - Relatively signal model independent
  - Intuitive to construct and train
- Cons:
  - Little control over sensitivity
  - Some model assumptions needed for construction

How to build anomaly score?

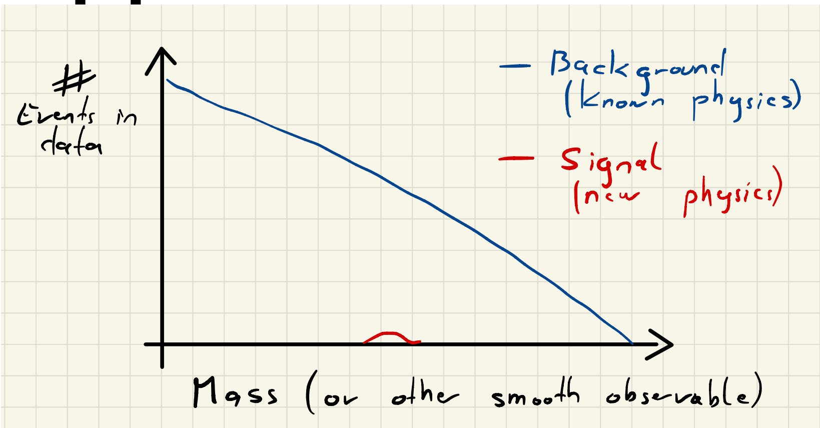
- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - a(x) = 1 / p(x|Background)
    (from simulation)
    (from data)
  - **a(x) = p(x|Signal) / p(x|Background)**

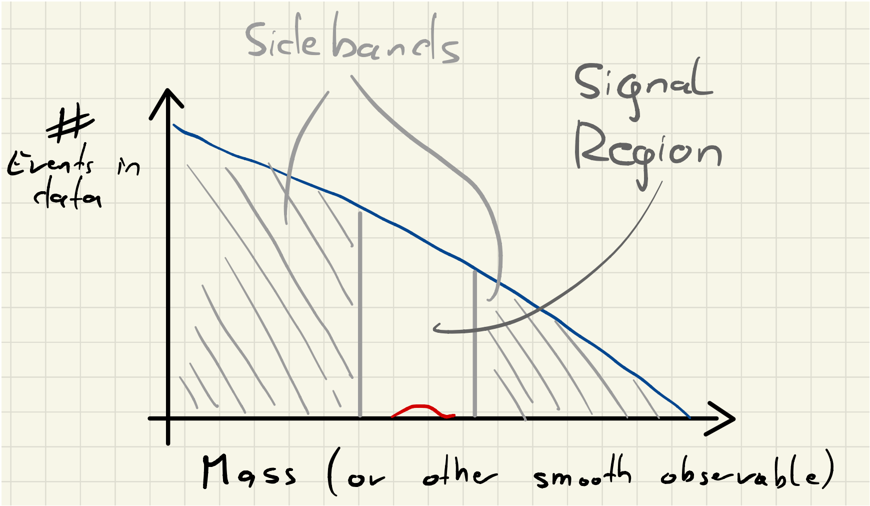
- Systematically look for differences between different phase space regions in data
- Show two examples:
  - **Mixed sample training**

Sideband approach

Key assumptions:

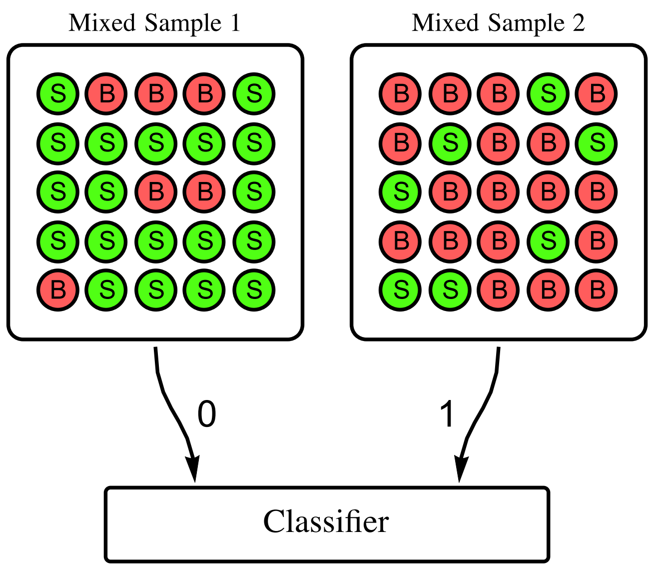
- There exists one feature so that:
  - Background distribution is smooth
  - Signal distribution is localised (and very small wrt/ background)





- Use sidebands to train anomaly score.
- Test signal region for new physics.
- Scan over different signal regions (trial factor)
- (Other ways to define anomaly-free regions in data possible as well. Not thoroughly explored yet)

Example: Mixed Sample training (aka CWola hunting)



$$L_{M_1/M_2} = \frac{p_{M_1}}{p_{M_2}} = \frac{f_1\, p_S + (1-f_1)\, p_B}{f_2\, p_S + (1-f_2)\, p_B} = \frac{f_1\, L_{S/B} + (1-f_1)}{f_2\, L_{S/B} + (1-f_2)}$$

- Distinguishing mixed samples is equivalent to signal/background classification assuming
  - Signal/background in both mixed samples are from same source
  - Sufficiently different mixed samples
- Translated to anomaly detection:
  - Train to distinguish signal region and sideband
  - Only use inputs independent of variable used to define these regions

Metodiev, Nachman, Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, 1708.02949
Collins, Howe, Nachman, *Anomaly Detection for Resonant New Physics with Machine Learning*, 1805.02664

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events

- Some options:
  - a(x) = (Semi-) Supervised
  - a(x) = 1 / p(x|Background)
    (from simulation)
    (from data)
  - **a(x) = p(x|Signal) / p(x|Background)**

- Systematically look for differences between different phase space regions in data
- Show two examples:
  - **Mixed sample training**
  - density estimation

- Pros:
  - Relatively signal model independent
  - Cheap to train
- Cons:
  - Sensitive to correlations
  - Some model assumptions needed for construction

How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - a(x) = 1 / p(x|Background)
    (from simulation)
    (from data)
  - **a(x) = p(x|Signal) / p(x|Background)**

- Systematically look for differences between different phase space regions in data
- Show two examples:
  - Mixed sample training
  - **density estimation**

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

Expand $$p(x) = f_{\text{normal}}\, p(x|\text{normal}) + f_{\text{anomaly}}\, p(x|\text{anomaly})$$

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

Expand $$p(x) = f_{\text{normal}}\, p(x|\text{normal}) + f_{\text{anomaly}}\, p(x|\text{anomaly})$$

And insert: $$L_{D/B} \approx f_{\text{normal}} + f_{\text{anomaly}} \frac{p(x|\text{anomaly})}{\tilde{p}(x|\text{normal})}$$

Example: Density Estimation

Per Neyman-Pearson: Likelihood-ratio is optimal test statistic
Unfortunatly, p(x|anomaly) is not available

$$L_{S/B} = \frac{p(x|\text{anomaly})}{p(x|\text{normal})}$$

Build data/background ratio:

$$L_{D/B} = \frac{p(x)}{p(x|\text{normal})}$$

Approximate background density using control measurement (e.g. sideband)

$$L_{D/B} \approx \frac{p(x)}{\tilde{p}(x|\text{normal})}$$

- Data-Background likelihood is monotonous to Signal-Background likelihood if we can approximate background.
- **We can use this to construct an anomaly score**

Expand $p(x) = f_{\text{normal}}\, p(x|\text{normal}) + f_{\text{anomaly}}\, p(x|\text{anomaly})$

And instert:

$$L_{D/B} \approx f_{\text{normal}} + f_{\text{anomaly}} \frac{p(x|\text{anomaly})}{\tilde{p}(x|\text{normal})}$$

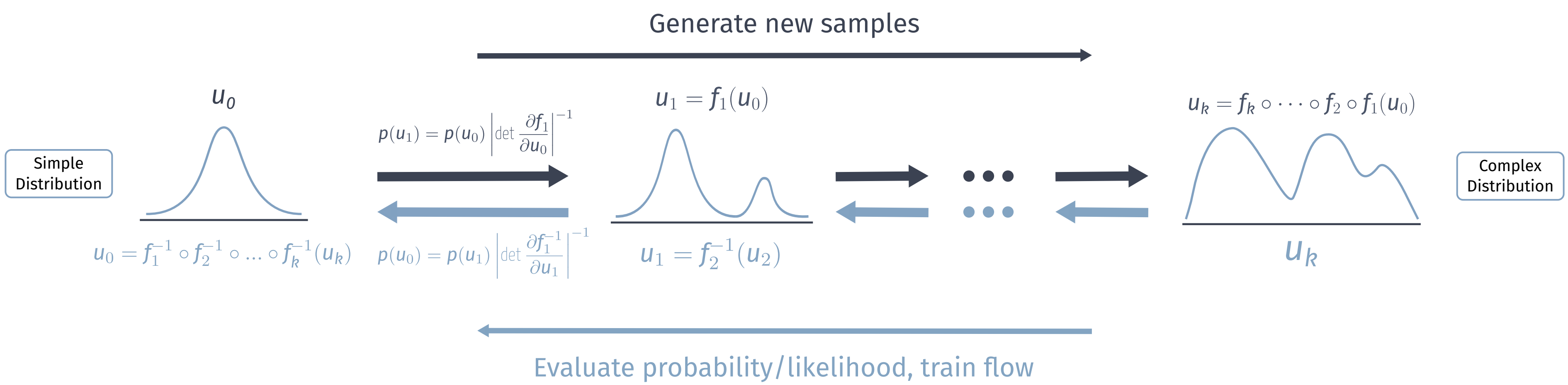
Thanks to T. Loesche

Normalising Flows

- **Goal:** assign probability density to each datapoint
- Learn bijective transformation between data and a latent space with tractable probability
- Build from simple invertible transformations, tractable Jacobian

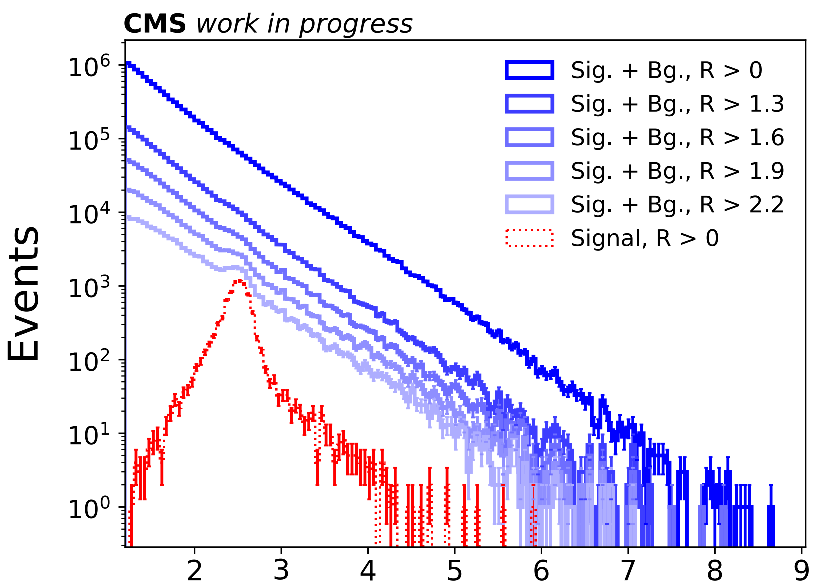
$$p(\boldsymbol{x}) = p(\boldsymbol{f}^{-1}(\boldsymbol{x})) \prod_i \left|\det\left(\frac{\partial \boldsymbol{f}_i^{-1}}{\partial \boldsymbol{x}}\right)\right| =$$

$$p(\boldsymbol{u}) \prod_i \left|\det\left(\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{u}}\right)\right|^{-1}$$

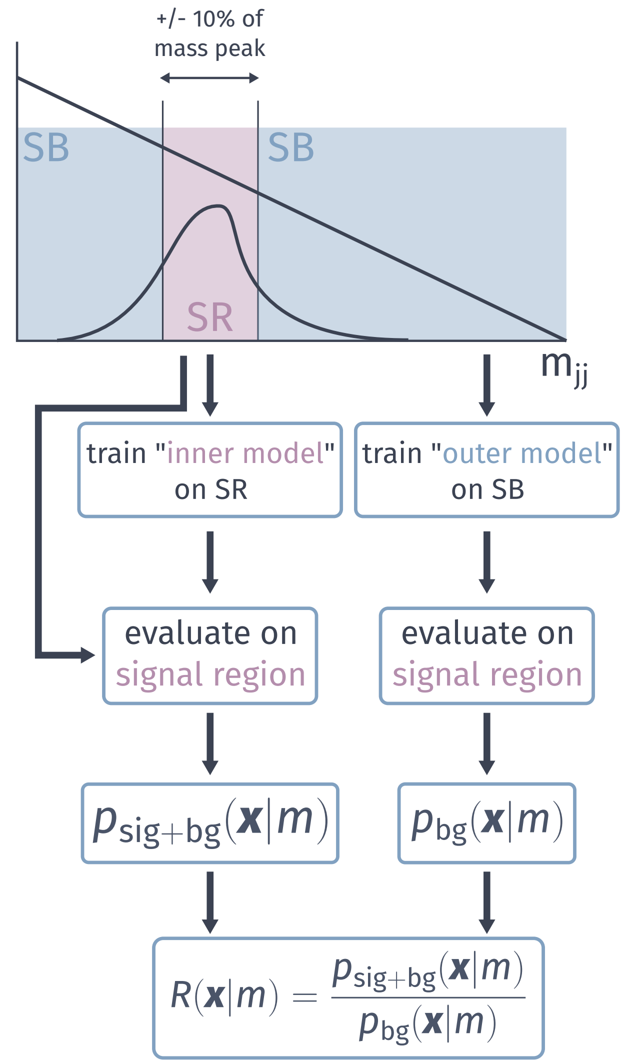


ANODE

- Use Masked Autoregressive Flow (1705.07057) to learn p (easy to invert NN with simple Jacobian)
- Compare extrapolated and in-region probability densities
- a(x) = R(x|m) = ratio of densities = $L_{D/B}$



Nachman, Shih, *Anomaly Detection with Density Estimation,* 2001.04990
Thanks to T. Loesche



How to build anomaly score?

- Anomaly score **a** should be high for anomalous (signal-like) and low for background-like events
- Some options:
  - a(x) = (Semi-) Supervised
  - a(x) = 1 / p(x|Background)
    (from simulation)
    (from data)
  - **a(x) = p(x|Signal) / p(x|Background)**

- Systematically look for differences between different phase space regions in data
- Show two examples:
  - Mixed sample training
  - **density estimation**
- Pros:
  - Relatively signal model independent
  - Powerful
- Cons:
  - Expensive to train
  - Some model assumptions needed for construction

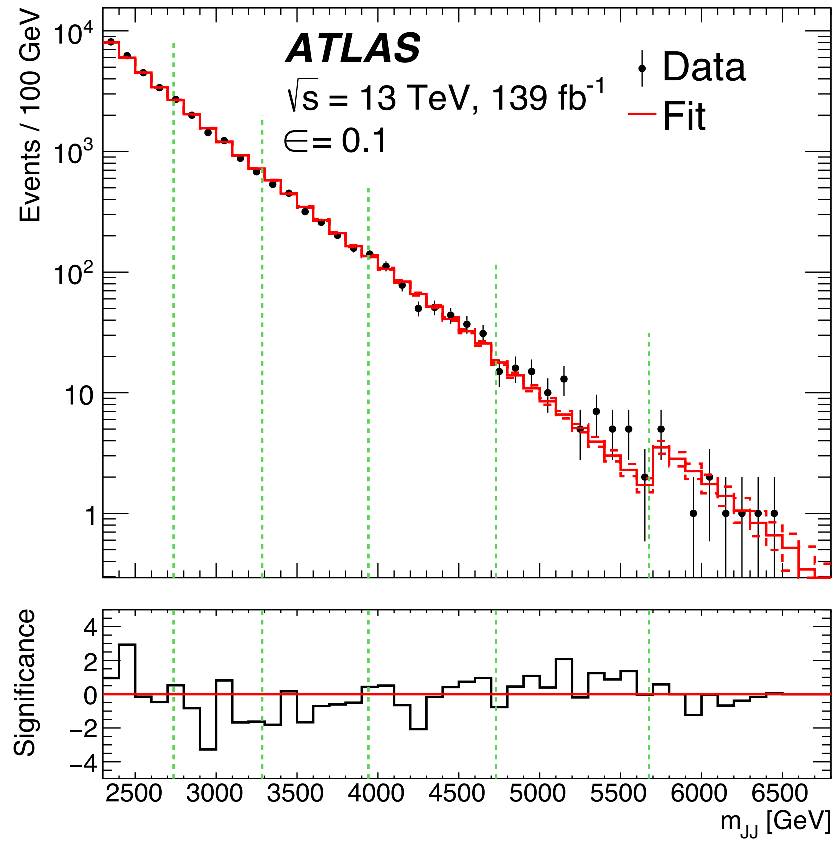
Moving on

- Many strategies exist to construct anomaly scores
  - a(x) = "p(x|Signal)” → Semi-**Supervised Cocktails**
  - a(x) = 1 / p(x|Background) → **General Search, Autoencoders**
  - a(x) = p(x|Signal) / p(x|Background) → **CWoLA, Density Estimation**
- **How can we use them in a search?**

MANY more ideas exist, see e.g. 2101.08320

Application: ATLAS Di-Jet Search

- ATLAS carried out a search following CWoLa approach
  - A → BC resonance search
    no assumption on masses of A,B,C
  - Resonance search in di-jet invariant mass using R=1.0 jets for B,C
  - Split spectrum into discrete signal regions
  - Use CWola method, cut on 10% and 1% most anomalous events
  - Fit spectrum from sidebands
  - Interpret results in W' model



ATLAS Collaboration, *Dijet resonance search with weak supervision using sqrt(s)=13 TeV pp collisions in the ATLAS detector*, 2005.02983

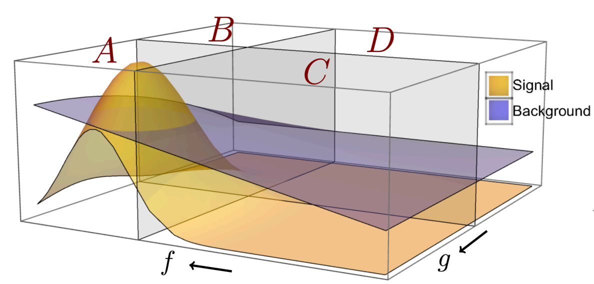
General

Simulation Driven

- Similar to classical analyses
- high signal model independence
- downside of data/simulation difference

Data Driven

- Straightforward idea to combine with bump hunt (see ATLAS example)
- Other data-driven techniques (ABCD?) should be possible as well, currently less explored
- Big advantage of data-only search:
  No systematic uncertainties
  (except background estimation from data)



How to interpret results?

Positive Result (some anomaly found)

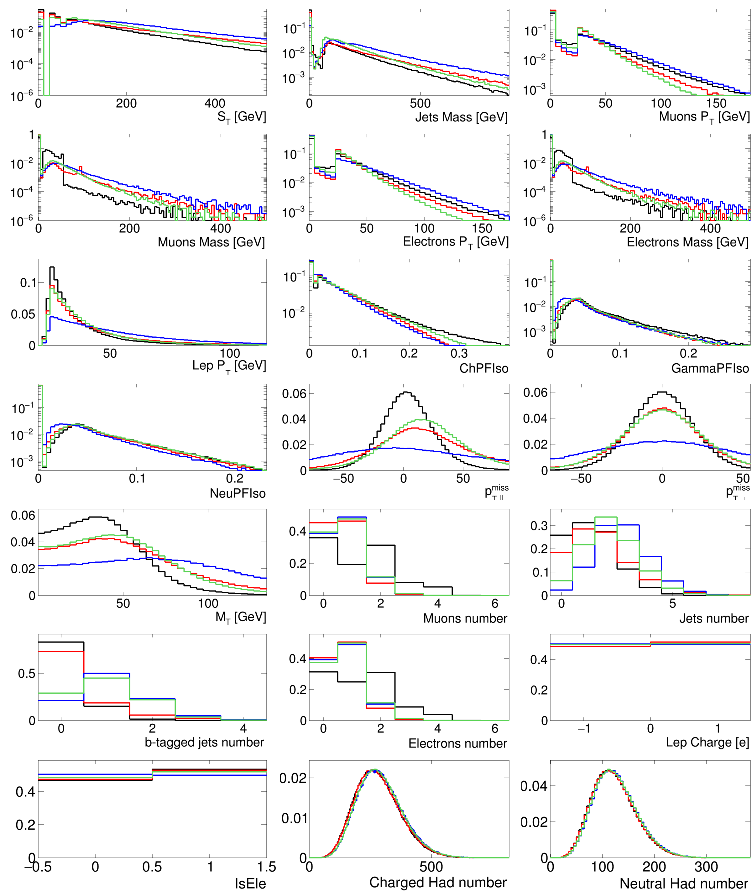
- Characterise what we found.
- Of course, can compare with different models and see which one fits. (Also test for detector effects, of course)
- Ideas to systematise this needed
- Publish events?

Negative Result (no anomaly found)

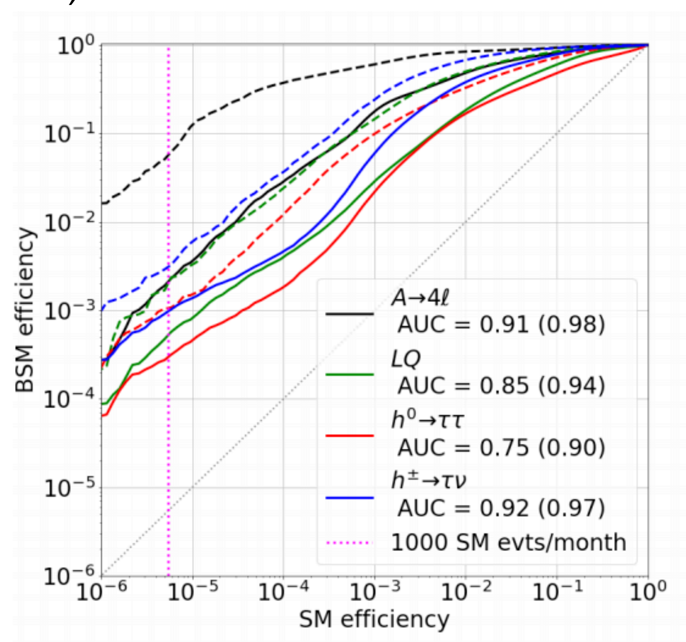
- Need to interpret resulting exlusion
- Of course, can run different models and test (systematic uncertainties enter here!)
- Again, strategy for interpretation needed. Publish anomaly score for recasting?

Aside: Trigger

- Consider anomaly detection for CMS/ATLAS triggers
- Strategy: (Variational) autoencoder trained on SM cockatail



21 high level observables as input



Cerri, Nguyen, Pierini, Spiropulu, Vlimant, *Variational Autoencoders for New Physics Mining at the Large Hadron Collider*, 1811.10276

Advertisment

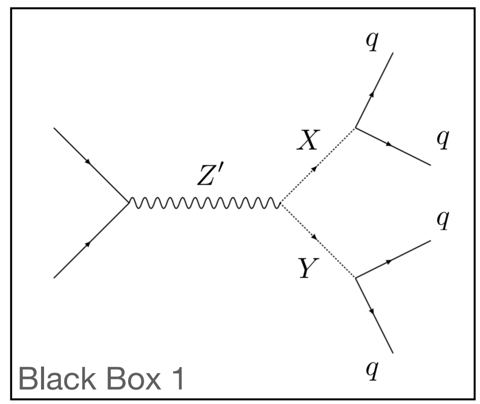
The LHC Olympics 2020

A Community Challenge for Anomaly Detection in High Energy Physics

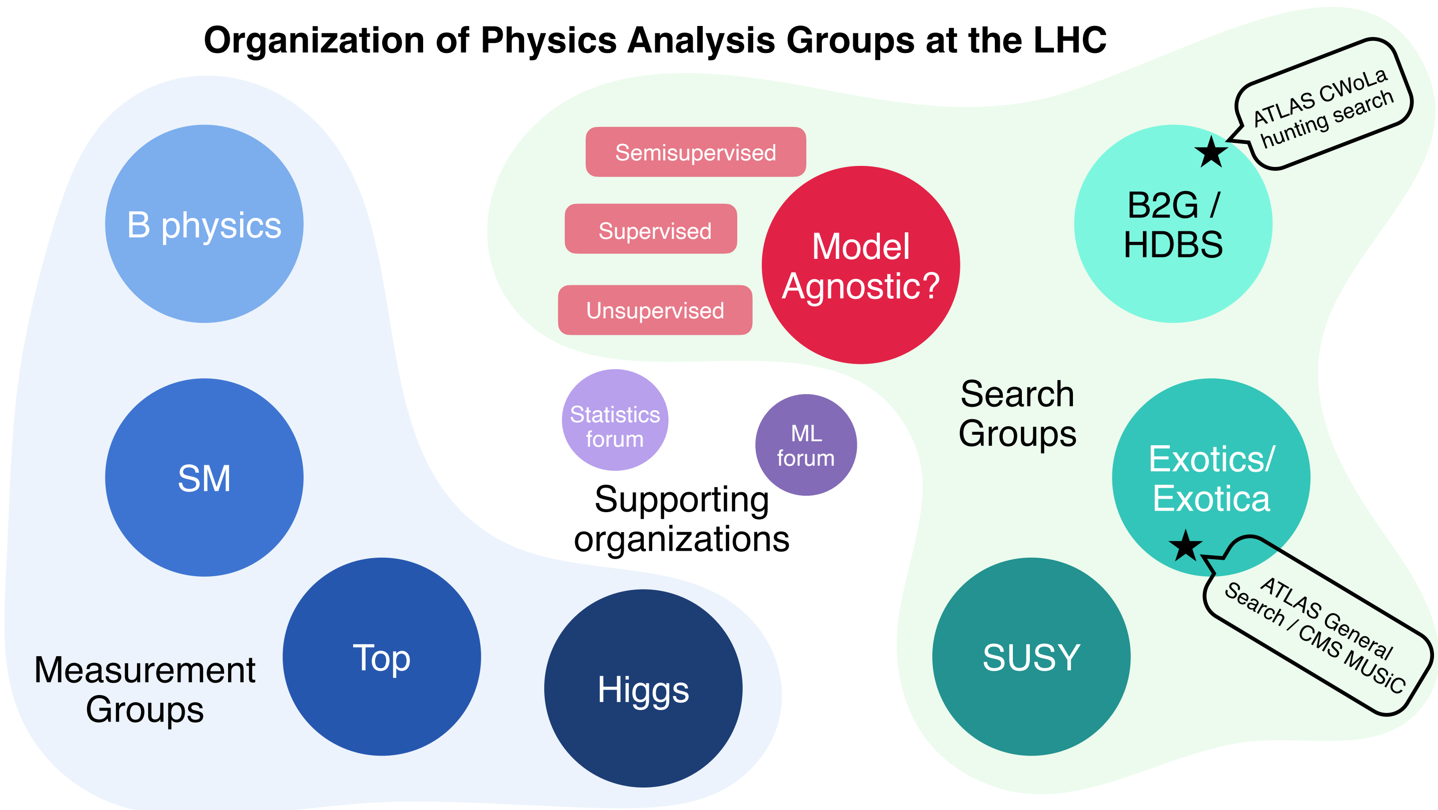
Gregor Kasieczka (ed),[1] Benjamin Nachman (ed),[2,3] David Shih (ed),[4] Oz Amram,[5] Anders Andreassen,[6] Kees Benkendorfer,[2,7] Blaz Bortolato,[8] Gustaaf Brooijmans,[9] Florencia Canelli,[10] Jack H. Collins,[11] Biwei Dai,[12] Felipe F. De Freitas,[13] Barry M. Dillon,[8,14] Ioan-Mihail Dinu,[5] Zhongtian Dong,[15] Julien Donini,[16] Javier Duarte,[17] D. A. Faroughy[10] Julia Gonski,[9] Philip Harris,[18] Alan Kahn,[9] Jernej F. Kamenik,[8,19] Charanjit K. Khosa,[20,30] Patrick Komiske,[21] Luc Le Pottier,[2,22] Pablo Martín-Ramiro,[2,23] Andrej Matevc,[8,19] Eric Metodiev,[21] Vinicius Mikuni,[10] Inês Ochoa,[24] Sang Eon Park,[18] Maurizio Pierini,[25] Dylan Rankin,[18] Veronica Sanz,[20,26] Nilai Sarda,[27] Uroš Seljak,[2,3,12] Aleks Smolkovic,[8] George Stein,[2,12] Cristina Mantilla Suarez,[5] Manuel Szewc,[28] Jesse Thaler,[21] Steven Tsan,[17] Silviu-Marian Udrescu,[18] Louis Vaslin,[16] Jean-Roch Vlimant,[29] Daniel Williams,[9] Mikaeel Yunus[18]

Kasieczka, Nachman, Shih (eds), et al, *The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics,* 2101.08320

- For more on anomaly detection see: https://indico.desy.de/e/anomaly2020
- Public datasets available: https://lhco2020.github.io/homepage/
- Community paper with ~20 methods



Advertisment



Kasieczka, Nachman, Shih (eds), et al, *The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics*, 2101.08320

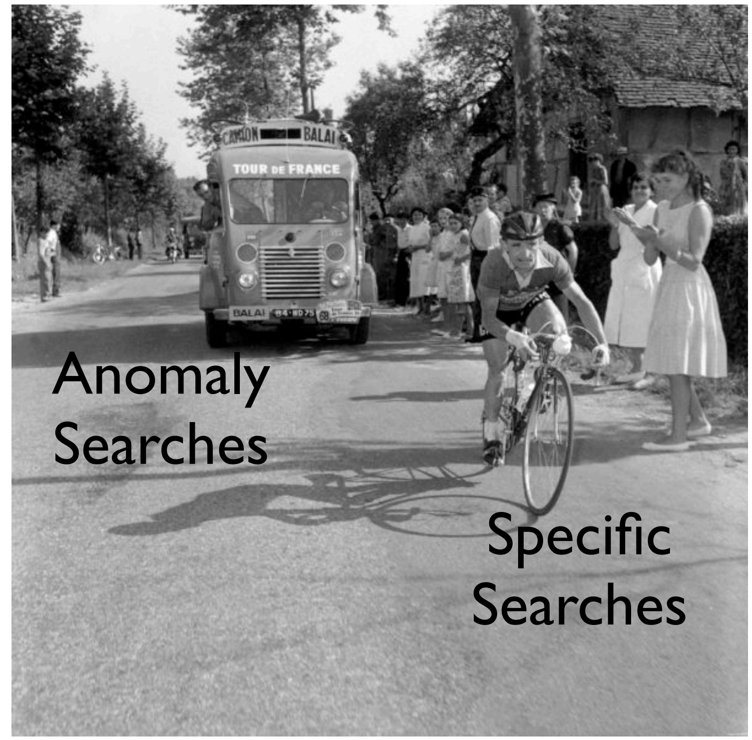
Closing

- Focused on new physics searches.
  Anomaly detection also considered for data quality monitoring, detector control, computing monitoring

- Improve power of anomaly detectors
- Extends to higher number of features
  - Beyond images / high-level observables

- How to properly encode normal physics / anomalous physics?
- Systematically understand sensitivity of different approaches
- Develop interpretation strategies

- Widely apply to experimental data



Thank you!

Backup

MADE/MAF

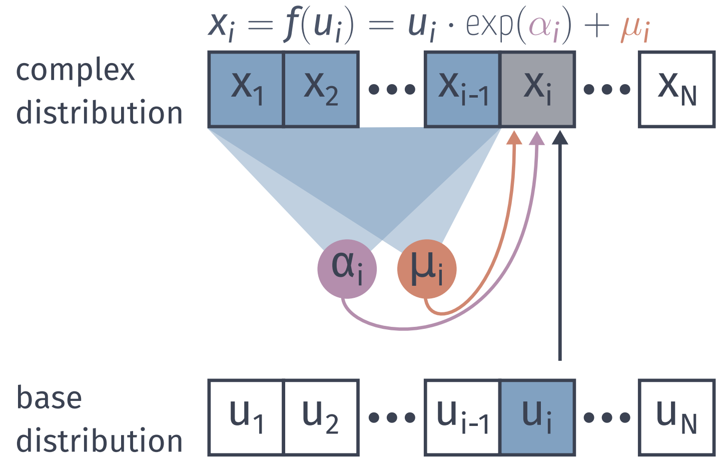
- Masked Autoregressive flow (1502.03509/1705.07057)
- Start with fully connected network, but drop connections so output a_j / mu_j are only connected to input x_1,...,x_j-1
- Autoregressive: no dependence of early features on late features
  - -> Jacobian is upper triangular matrix and easily invertible
- Combine multiple such blocks

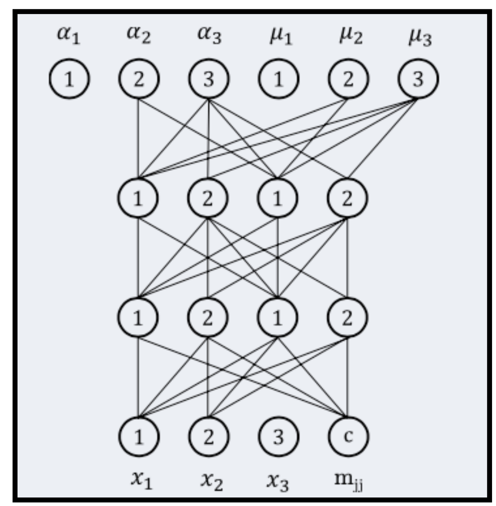
$$p(x) = \prod_i p(x_i|x_{1:i-1})$$

$$p(x_i|x_{1:i-1}) = \mathcal{N}(x_i|\mu_i, (\exp \alpha_i)^2)$$

$$\mu_i = f_{\mu_i}(x_{1:i-1})$$

$$\alpha_i = f_{\alpha_i}(x_{1:i-1})$$





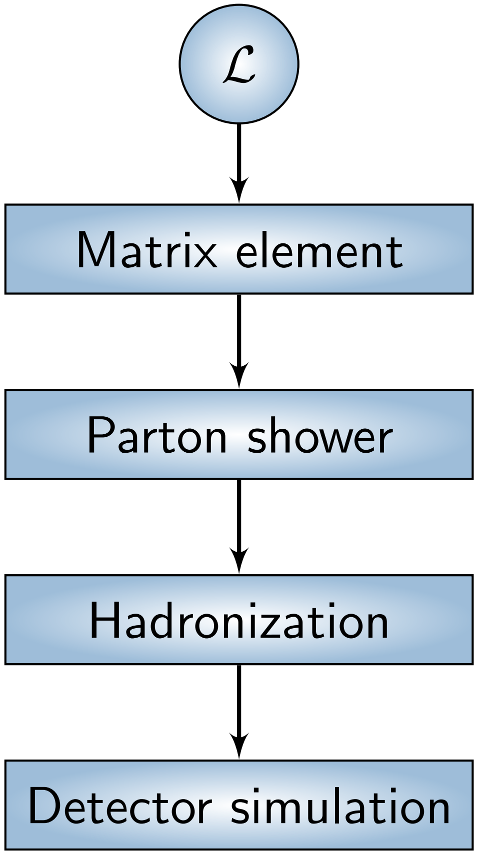
Unsupervised Learning for Fun and Precision

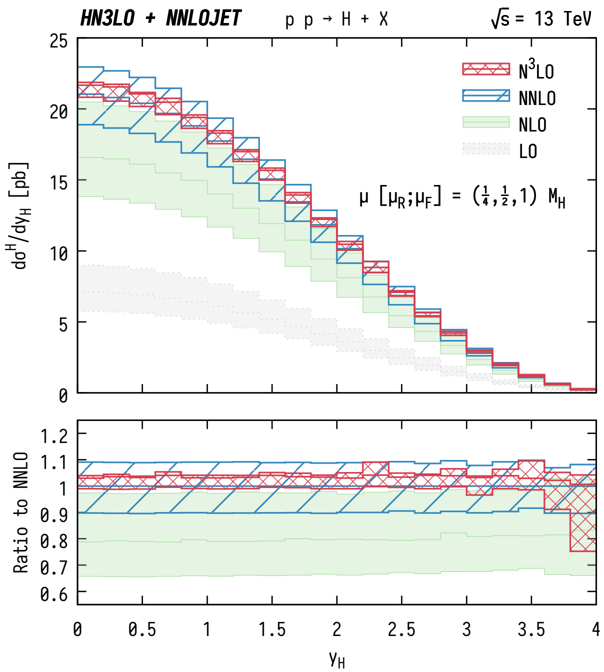
KITP - Precision21

Anja Butter & Gregor Kasieczka

ITP, Universität Heidelberg

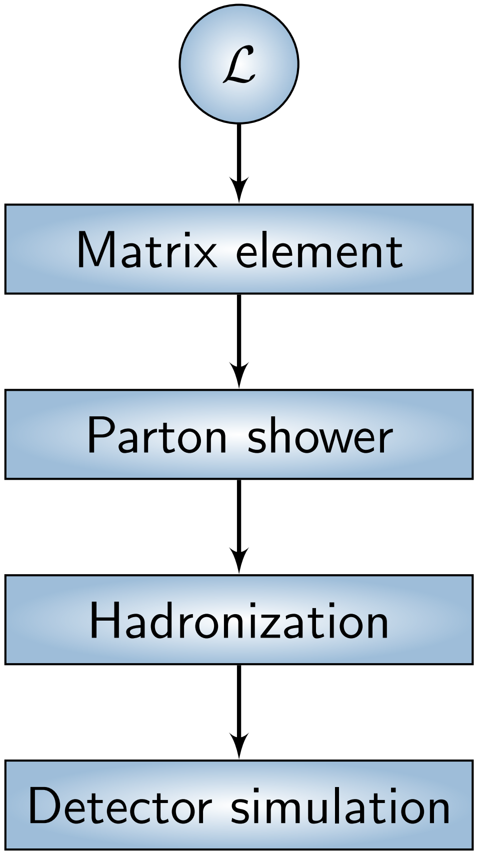
Precision simulations with limited resources

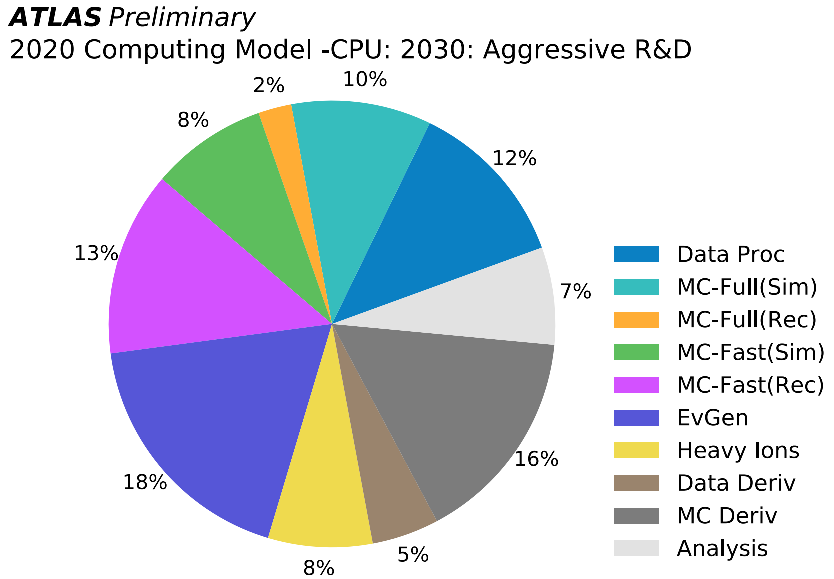




[1807.11501] Cieri, Chen, Gehrmann, Glover, Huss

Precision simulations with limited resources



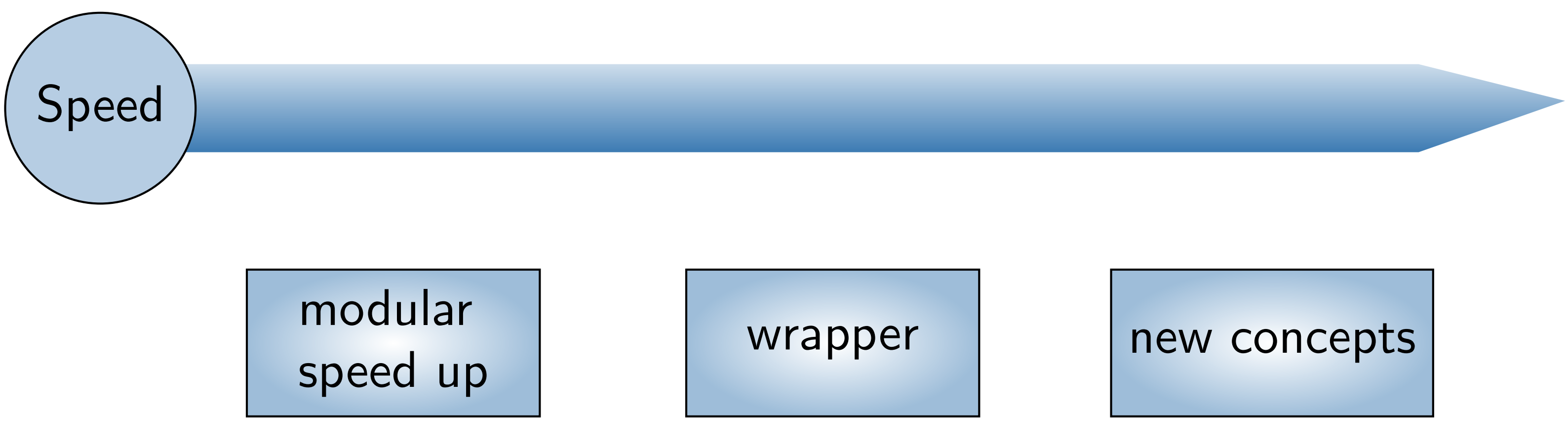


Speed = Precision

How can we boost MC simulations

- ML 2.0 Generative models
  - $\rightarrow$ Can we simulate new data?

Fast evaluation $\rightarrow$ more events = higher order $\rightarrow$ Precision



Boosting standard event generation...

1. Generate phase space points

2. Calculate event weight

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

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

Boosting standard event generation...

Matrix element

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

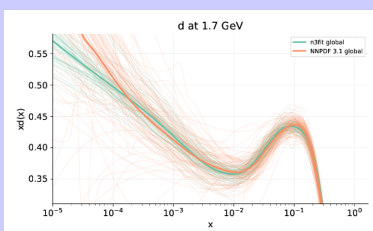
PDF

Phase space mapping

Boosting standard event generation...

Matrix element

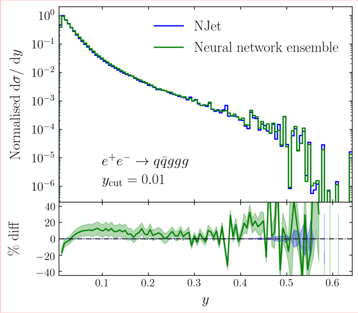
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \ldots p_n) \times J(p_i(r))^{-1}$$



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

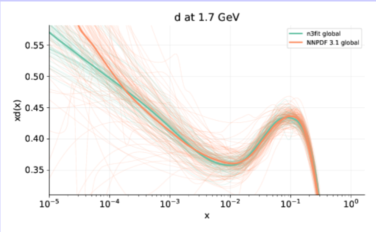
Phase space mapping

Boosting standard event generation...



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

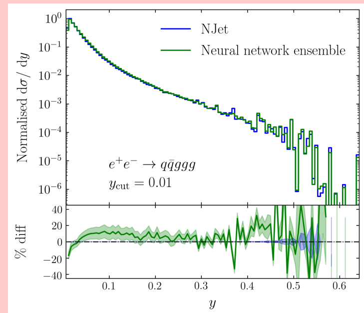
$$w_{event} = f(x_1, Q^2) f(x_2, Q^2) \times \mathcal{M}(x_1, x_2, p_1, \dots p_n) \times J(p_i(r))^{-1}$$



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

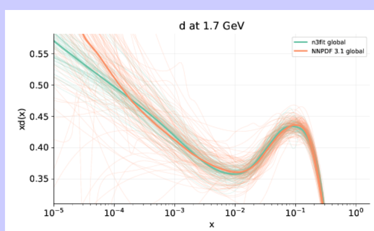
Phase space mapping

Boosting standard event generation...

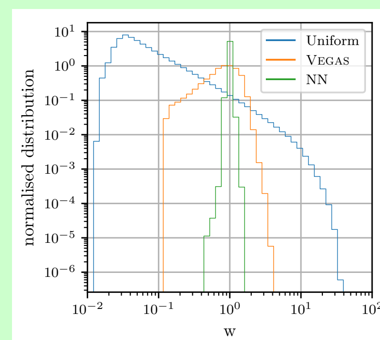


- Amplitude estimation
- S. Badger, J. Bullock [2002.07516]
- J. Bendavid [1707.00028]

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



- 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**

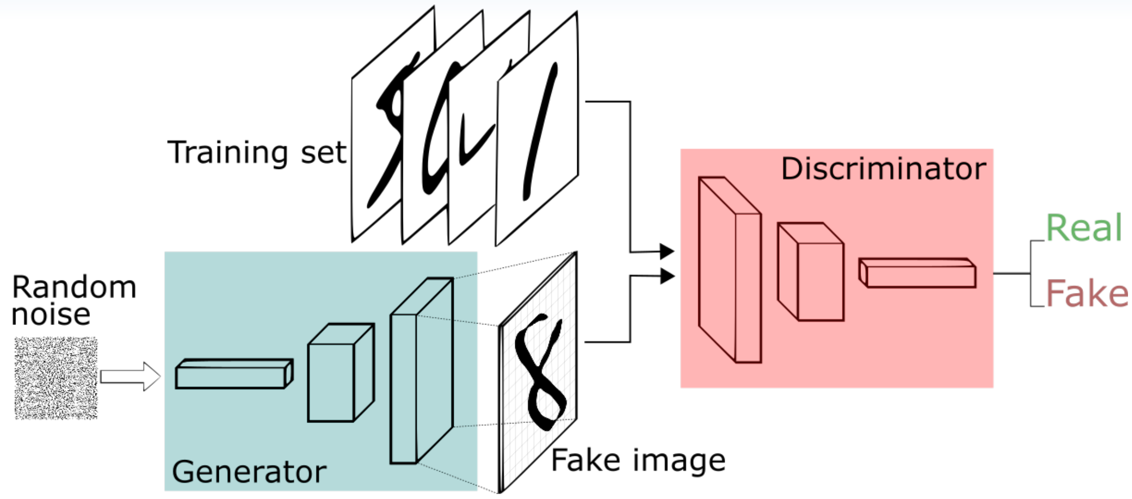
[2005.05334, 2102.12491] Buhmann et al. **VAE**

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

[2009.14017] Lu et al.

NO claim to completeness!

Generative Adversarial Networks



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

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

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

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

$\Rightarrow$ **Nash Equilibrium**

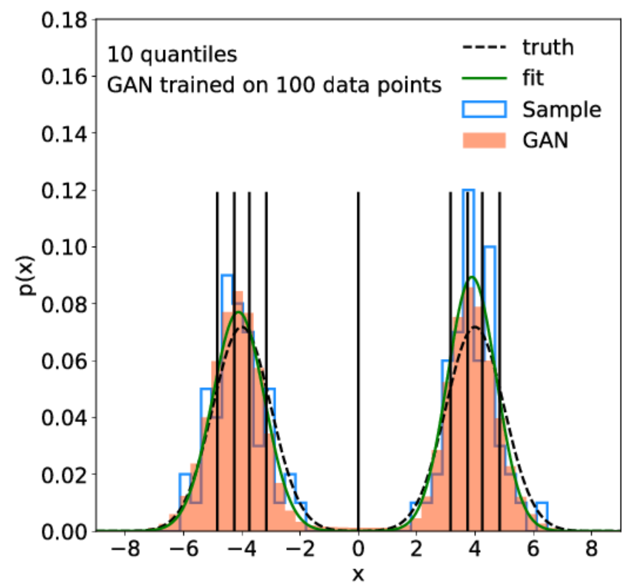
$\Rightarrow$ **New statistically independent samples**

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

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

Evaluation on quantiles:

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

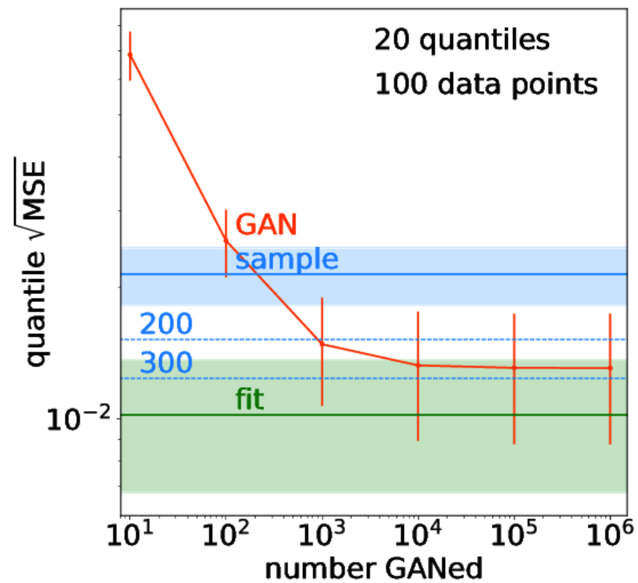


What is the statistical value of GANned events?[2008.06545]

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

Evaluation on quantiles:

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



→ Amplification factor 2.5

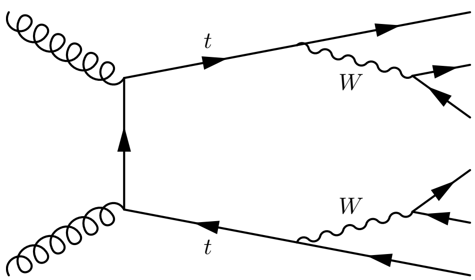
Sparser data → bigger amplification

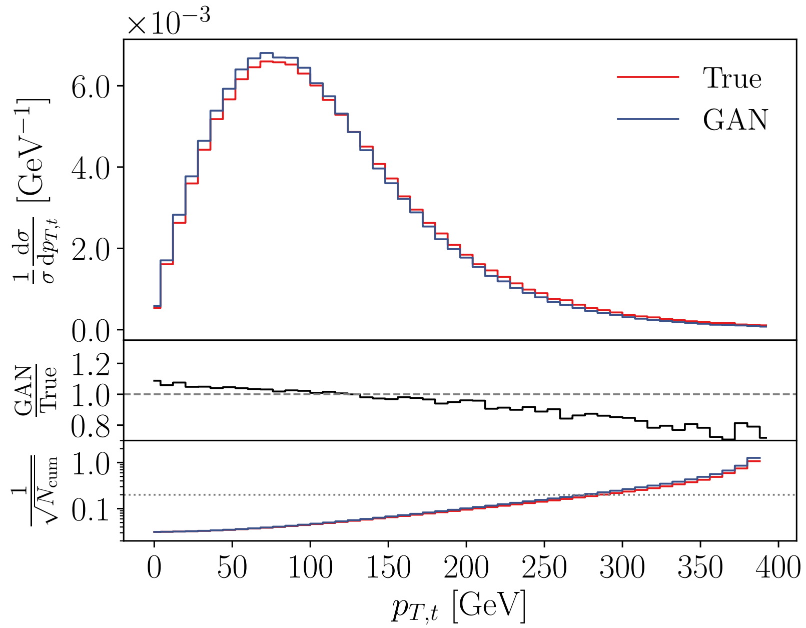
How to GAN LHC events [1907.03764]

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

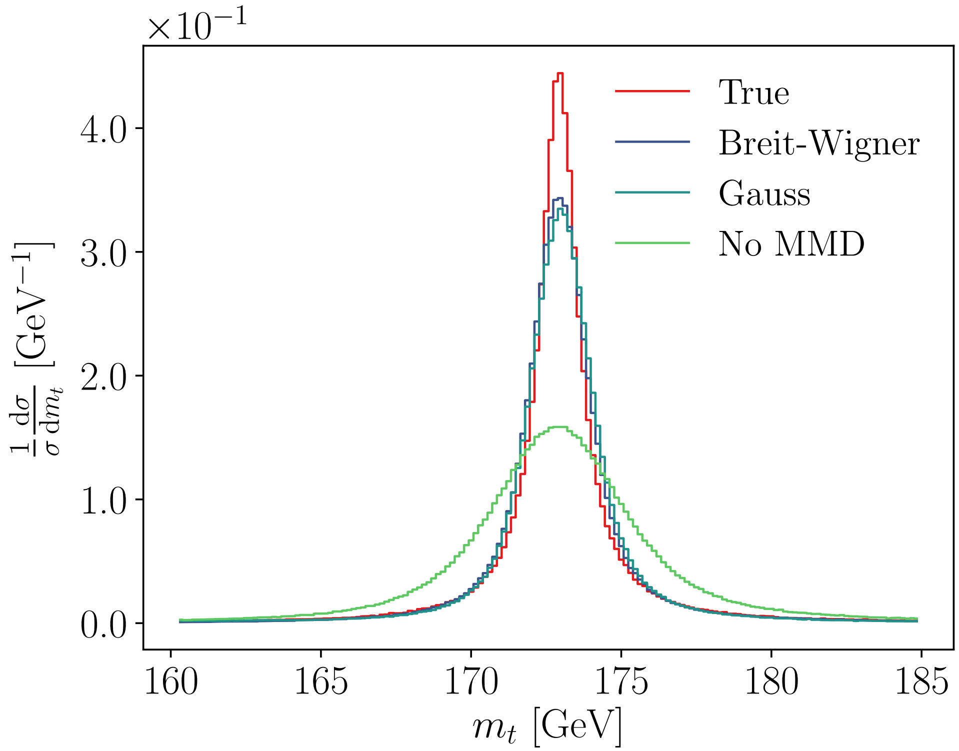
+ Flat observables ✓

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





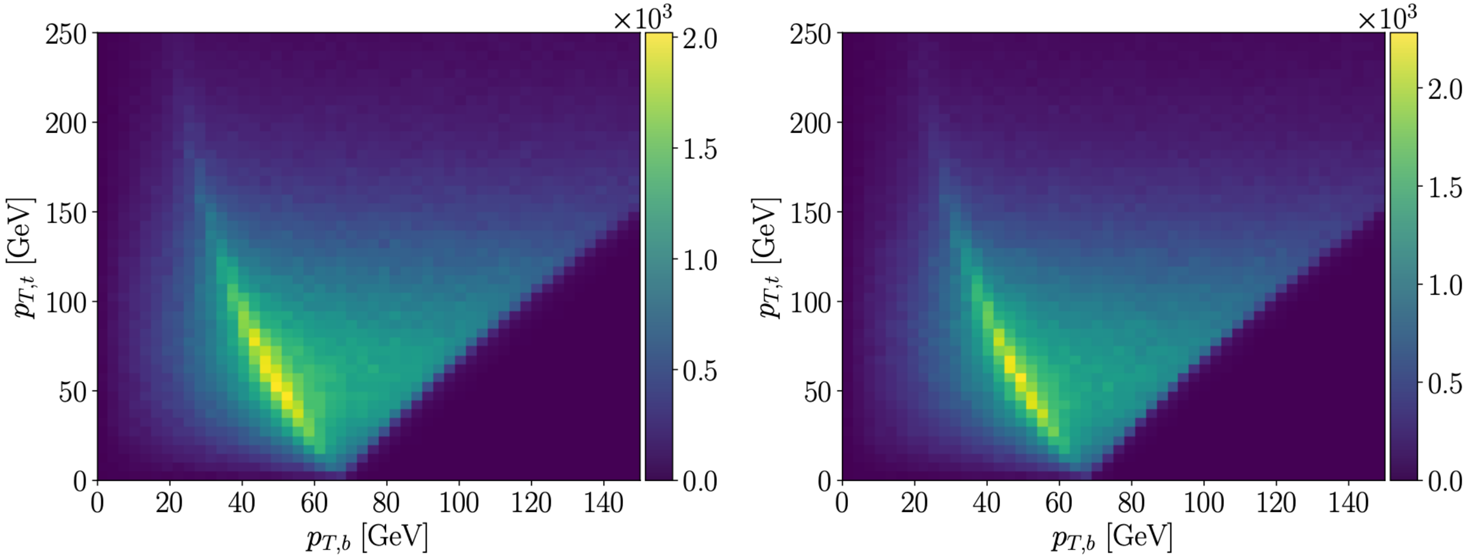
Special features

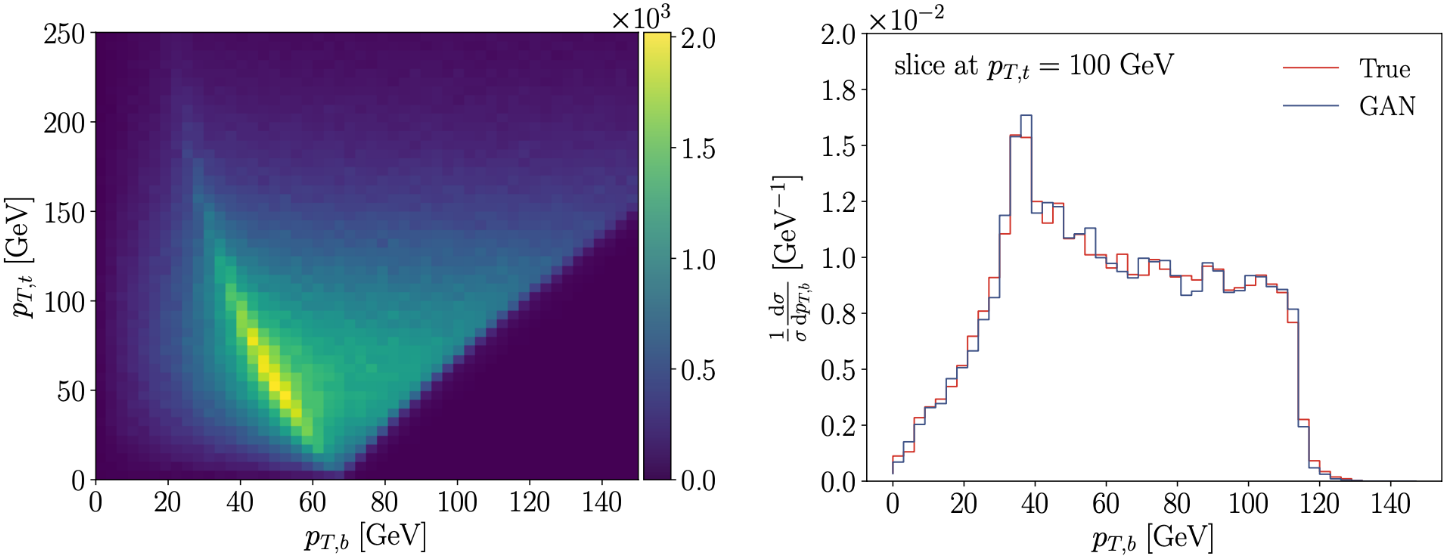


Solution: MMD kernel

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

Correlations





$\times 10^3$
$p_{T,t}$ [GeV]
$p_{T,b}$ [GeV]
$\times 10^{-2}$
slice at $p_{T,t} = 100$ GeV
True
GAN
$\frac{1}{\sigma}\frac{d\sigma}{dp_{T,b}}$ [GeV^{-1}]
$p_{T,b}$ [GeV]

✓

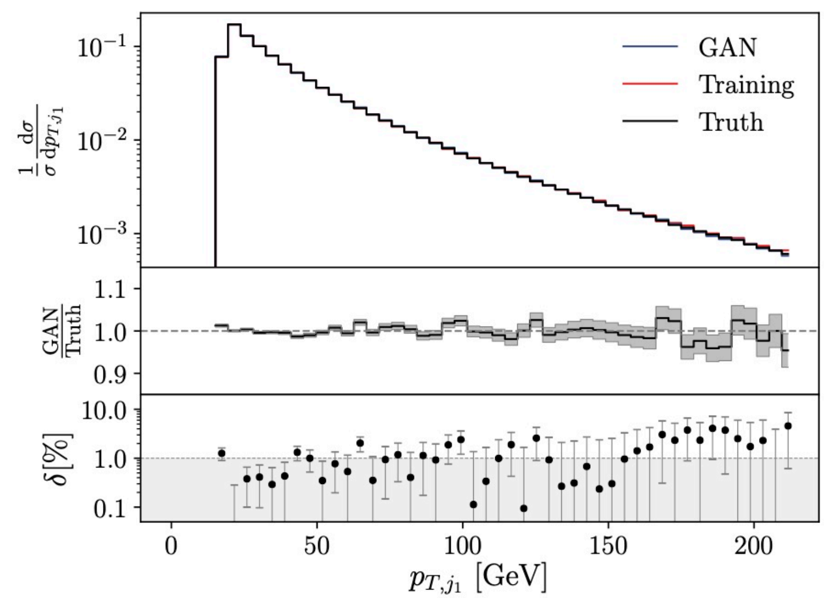
Reaching precision (preliminary)

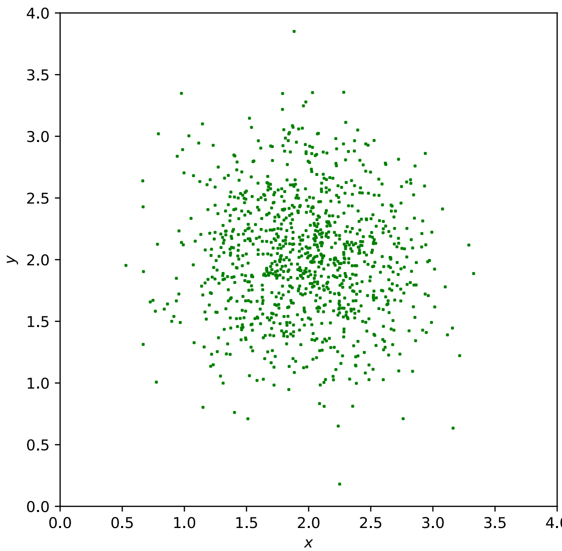
1. Representation p_T, η, ϕ
2. Momentum conservation
3. Resolve log p_T
4. Regularization: spectral norm
5. Batch information

$\rightarrow$ 1% precision ✓

Next step automization

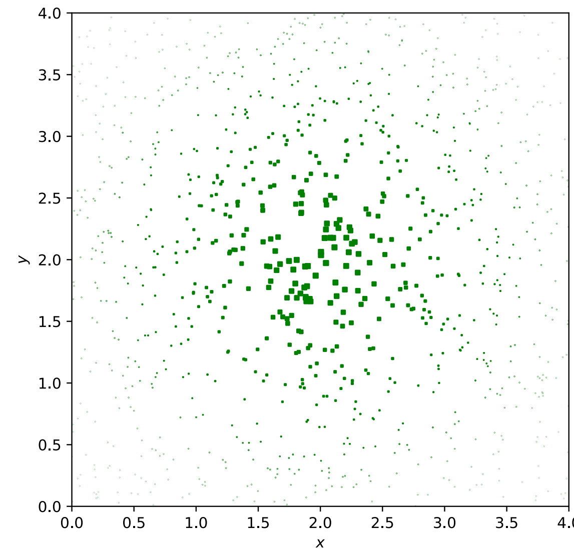
W + 2 jets





Information in space distribution
(what we want)

=



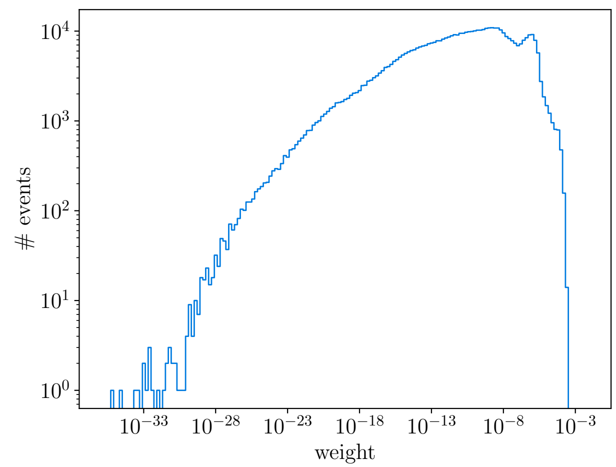
Information in weight
(what we have)

The unweighting bottleneck

- High-multiplicity / higher-order $\to$ unweighting efficiencies $< 1\%$

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

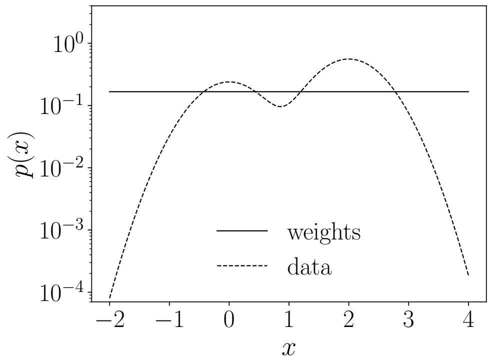
$pp \to \mu^+\mu^-$ with $m_{\mu\mu} > 50$ GeV



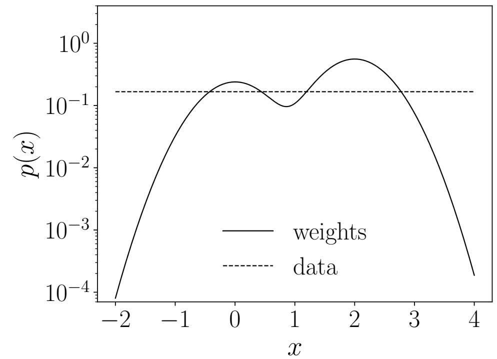
$\to$ unweighting efficieny 0.2%

Training on weighted events

Information contained in distribution or event weights

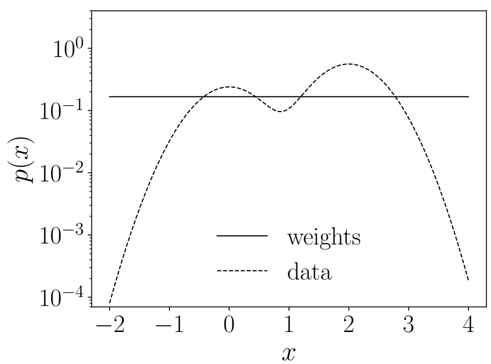


Train on
weighted events

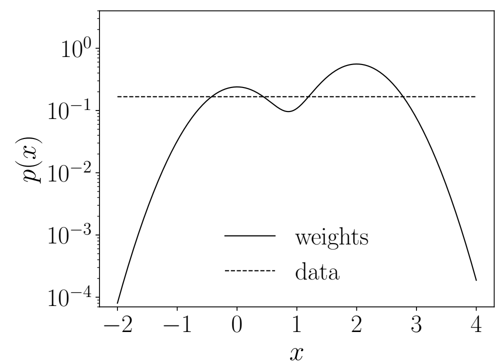


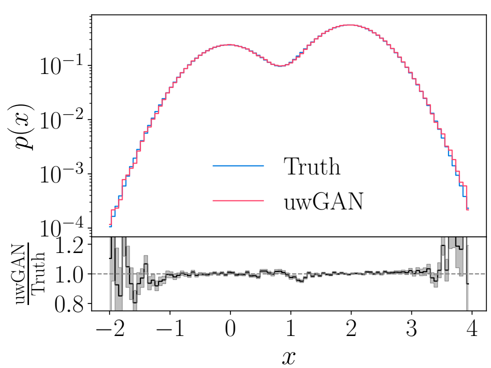
Training on weighted events

Information contained in distribution or event weights

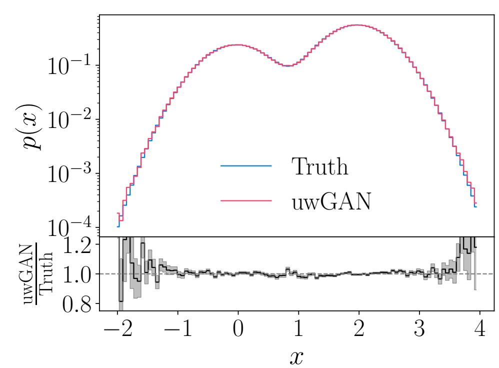


Train on weighted events





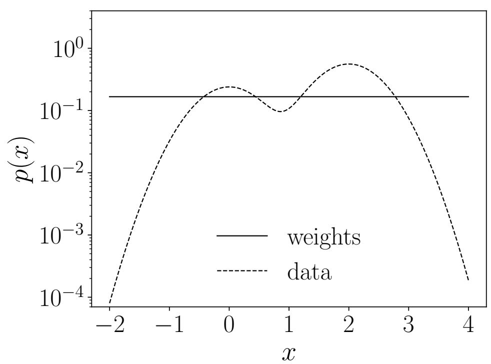
Generate unweighted events



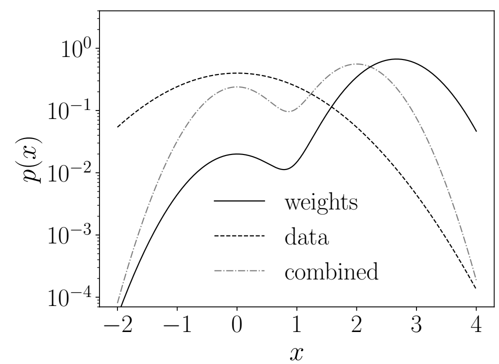
$$L_D = \big\langle -w \log D(x) \big\rangle_{x \sim P_{Truth}} + \big\langle -\log(1 - D(x)) \big\rangle_{x \sim P_{Gen}}$$

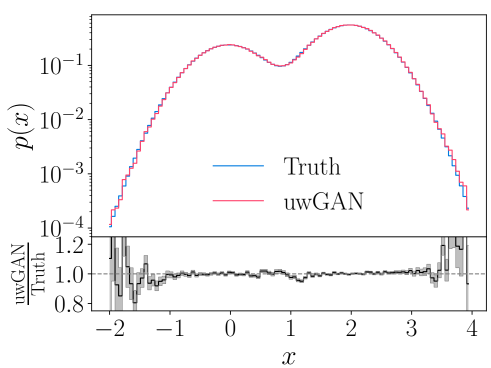
Training on weighted events

Information contained in distribution or event weights

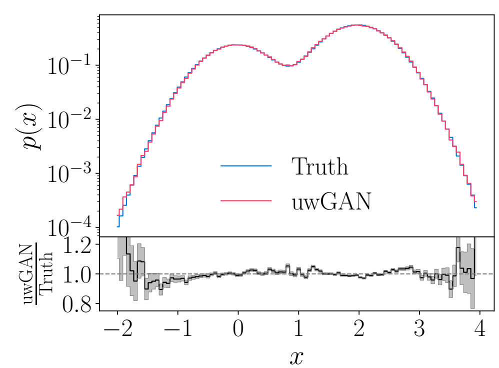


Train on
weighted events





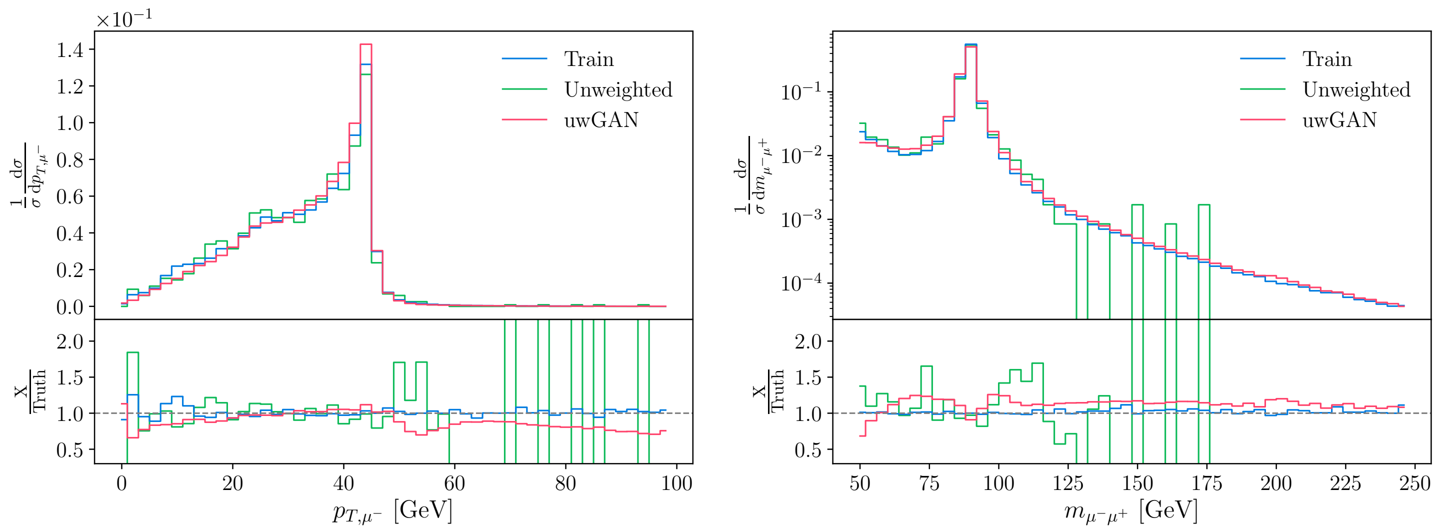
Generate
unweighted events



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

normalizing flow: B. Stienen, R. Verheyen [2011.13445]

uwGAN results

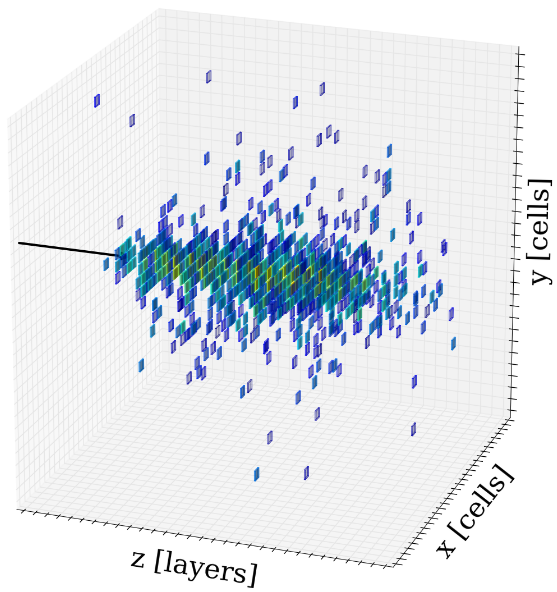


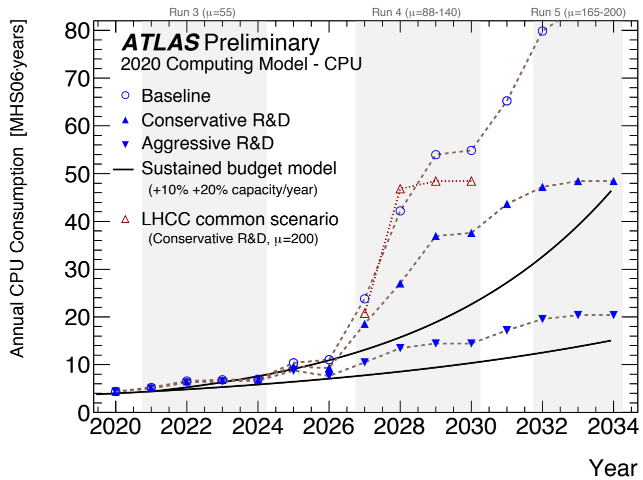
Populates high energy tails

Large amplification wrt. unweighted data!

Fast detector simulations

- Important R&D potential
  NN evaluation $\times$100-1000
  faster than GEANT4

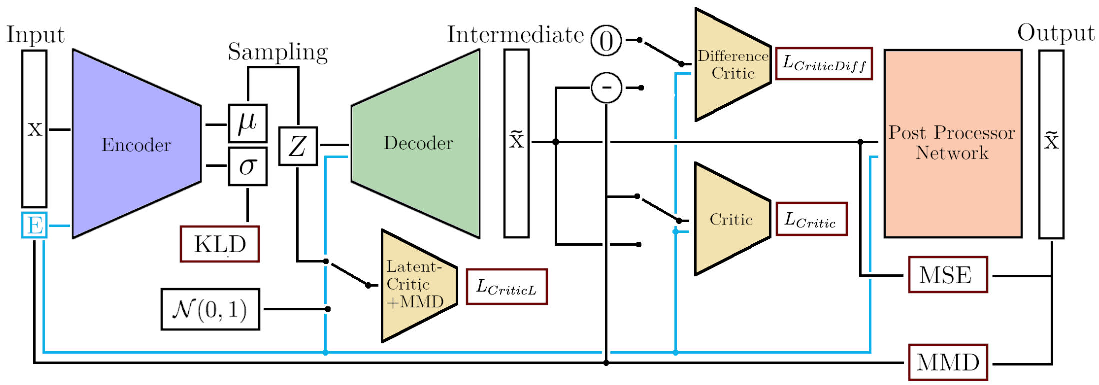




- Same underlying techniques
  [GAN, VAE, (NF)]
- Challenge:
  High-dimensional output
  $\leftarrow 30 \times 30 \times 30$

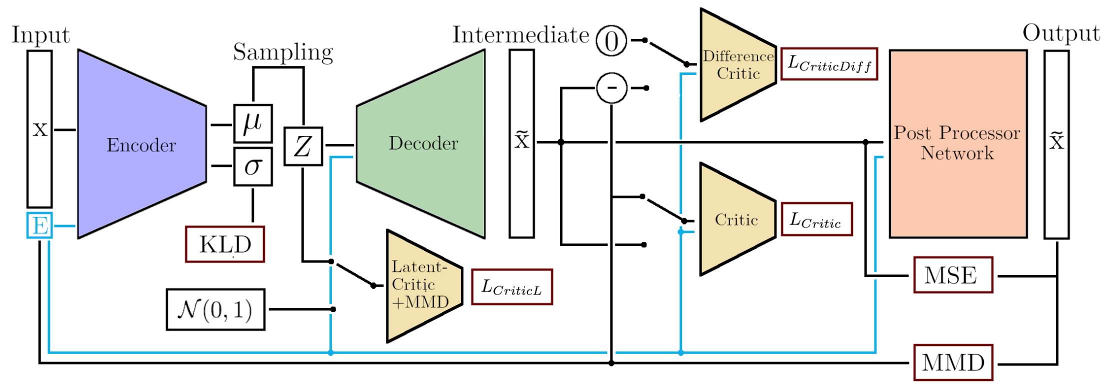
BIB-AE PP

Bounded-Information-Bottleneck autoencoder with post processing

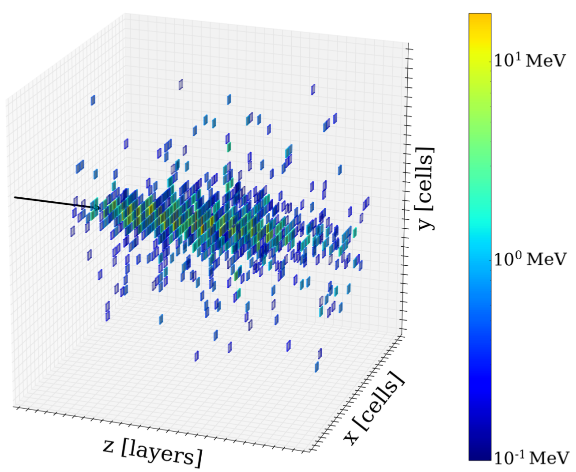


BIB-AE PP

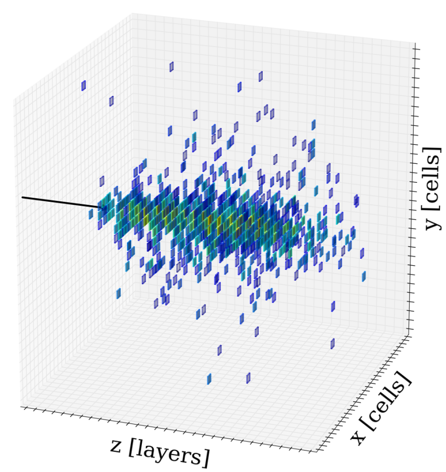
Bounded-Information-Bottleneck autoencoder with post processing



GEANT4

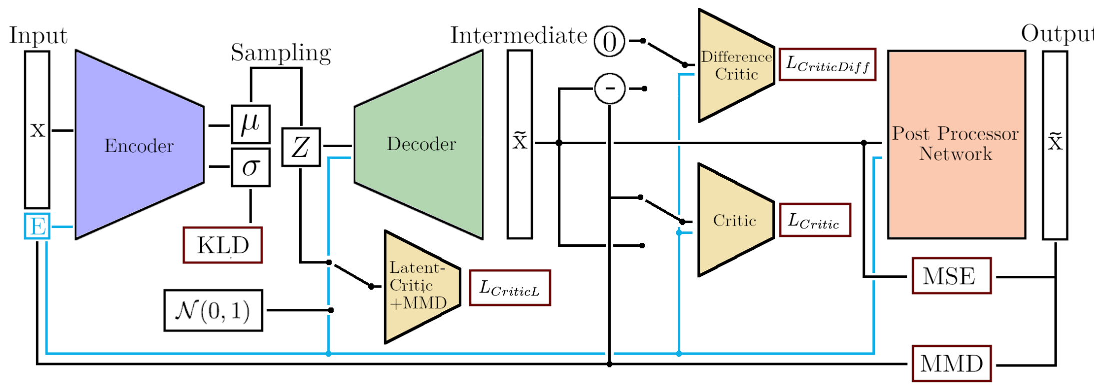


Simulation

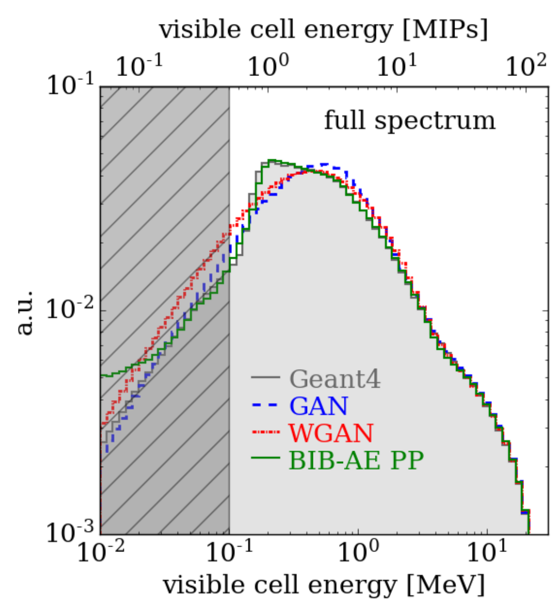


BIB-AE PP

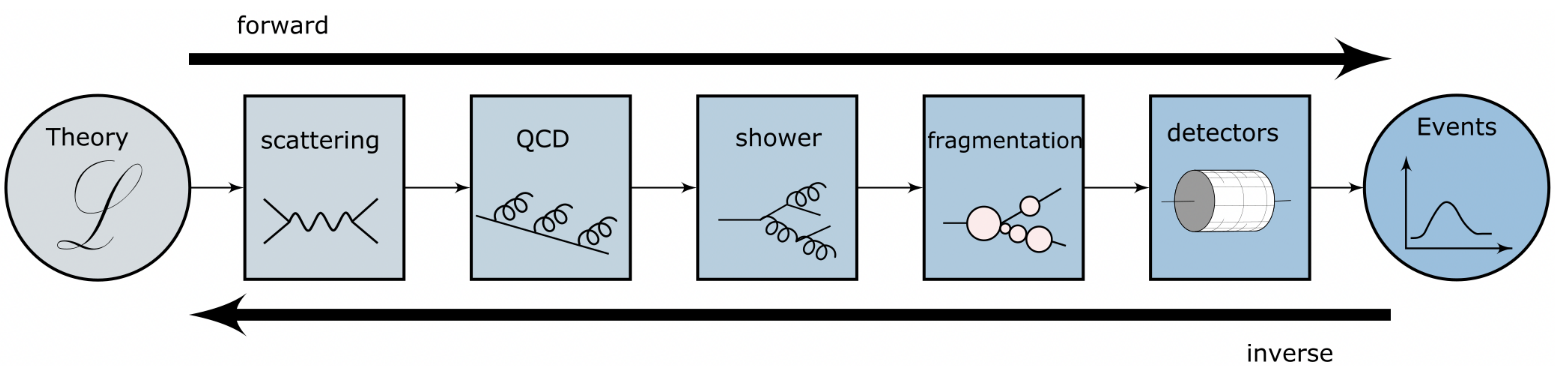
Bounded-Information-Bottleneck autoencoder with post processing



Post Processor Network adjusts energy to recover spectrum
$\rightarrow$ MIP bump



Can we invert the simulation chain?



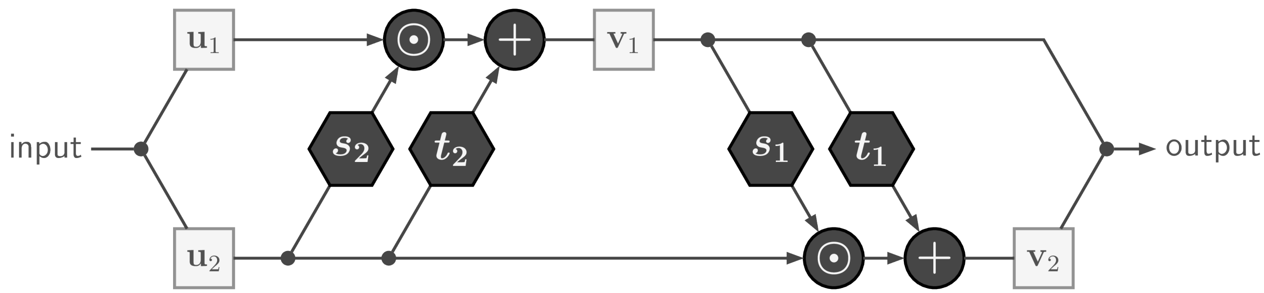
What we
want to know

What we
measure or simulate

wish list:
- ☐ multi-dimensional
- ☐ bin independent
- ☐ statistically well defined

Invertible networks

$$\left(x_{part}\right) \xleftrightarrow[\leftarrow \text{ unfolding:}\bar{g}]{\text{PYTHIA,DELPHES:}g\rightarrow} \left(x_{det}\right)$$



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

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

+ Bijective mapping

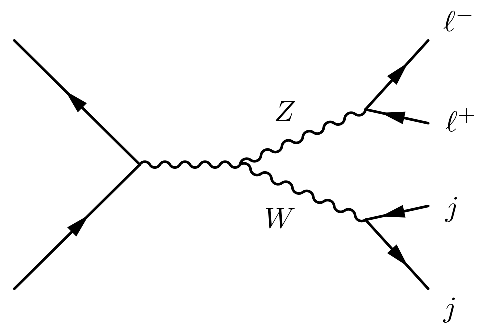
+ Tractable Jacobian

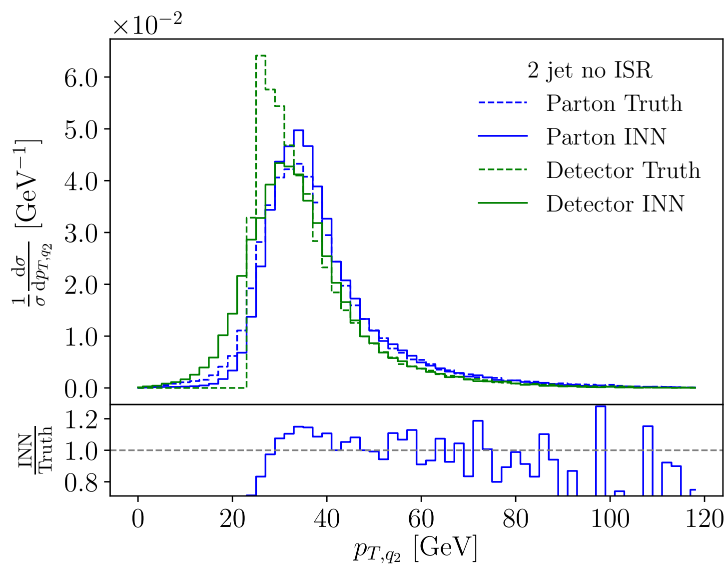
+ Fast evaluation in both directions

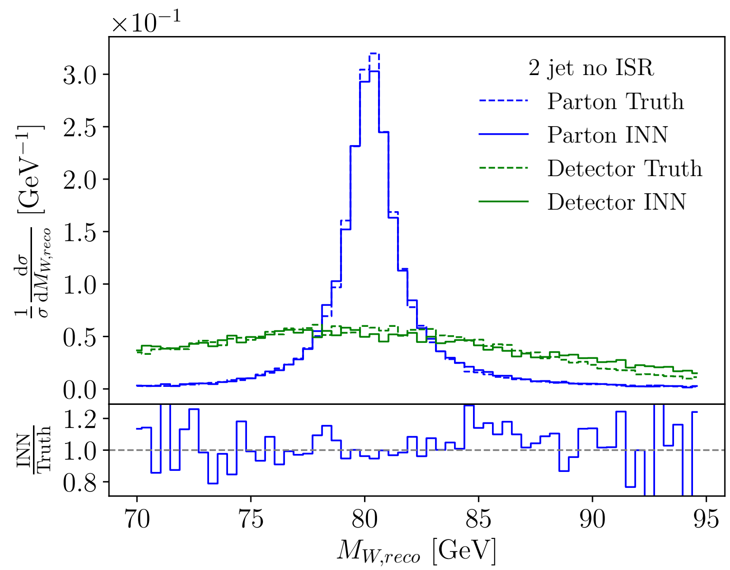
+ Arbitrary networks s and t

Inverting detector effects

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





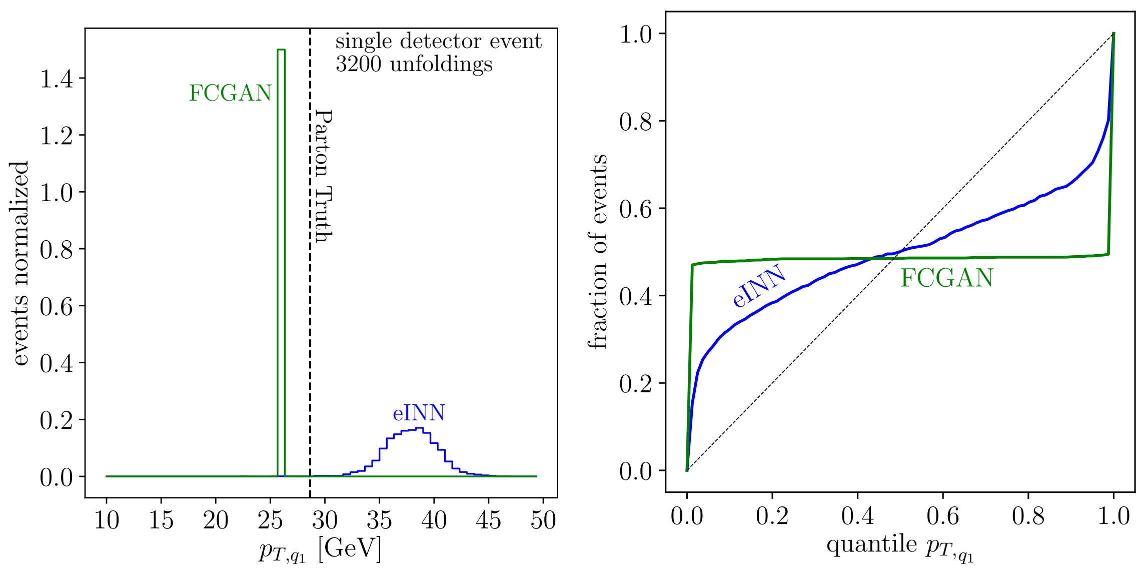


multi-dimensional ✓ bin independent ✓ statistically well defined ?

Including stochastical effects

$$\begin{pmatrix} x_p \\ r_p \end{pmatrix} \xleftrightarrow[\leftarrow \text{ unfolding:}\bar{g}]{\textsc{Pythia,Delphes}:g\rightarrow} \begin{pmatrix} x_d \\ r_d \end{pmatrix}$$

Sample r_d for fixed detector event
How often is Truth included in distribution quantile?

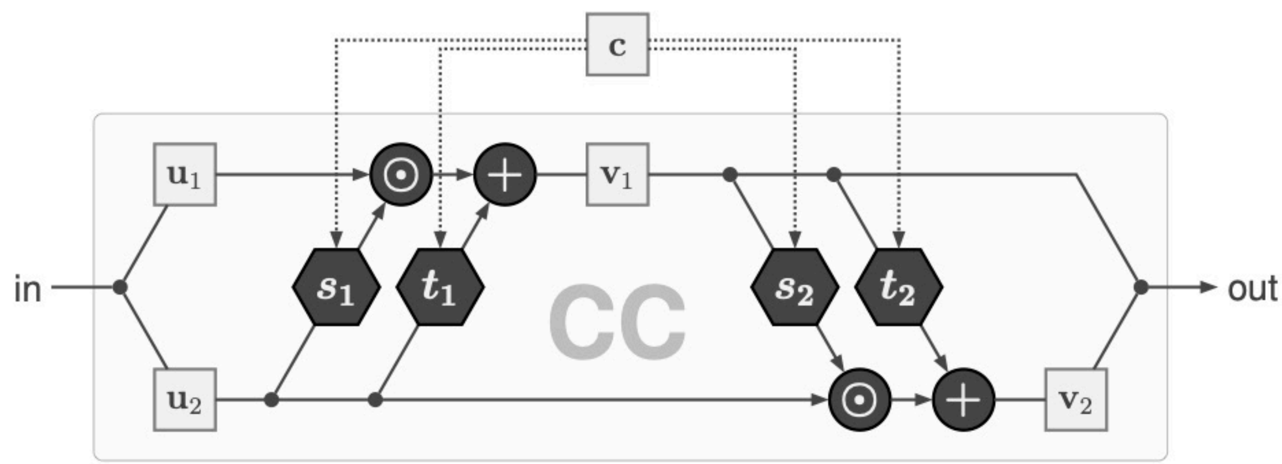


- Problem: arbitrary balance of many loss functions

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

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



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

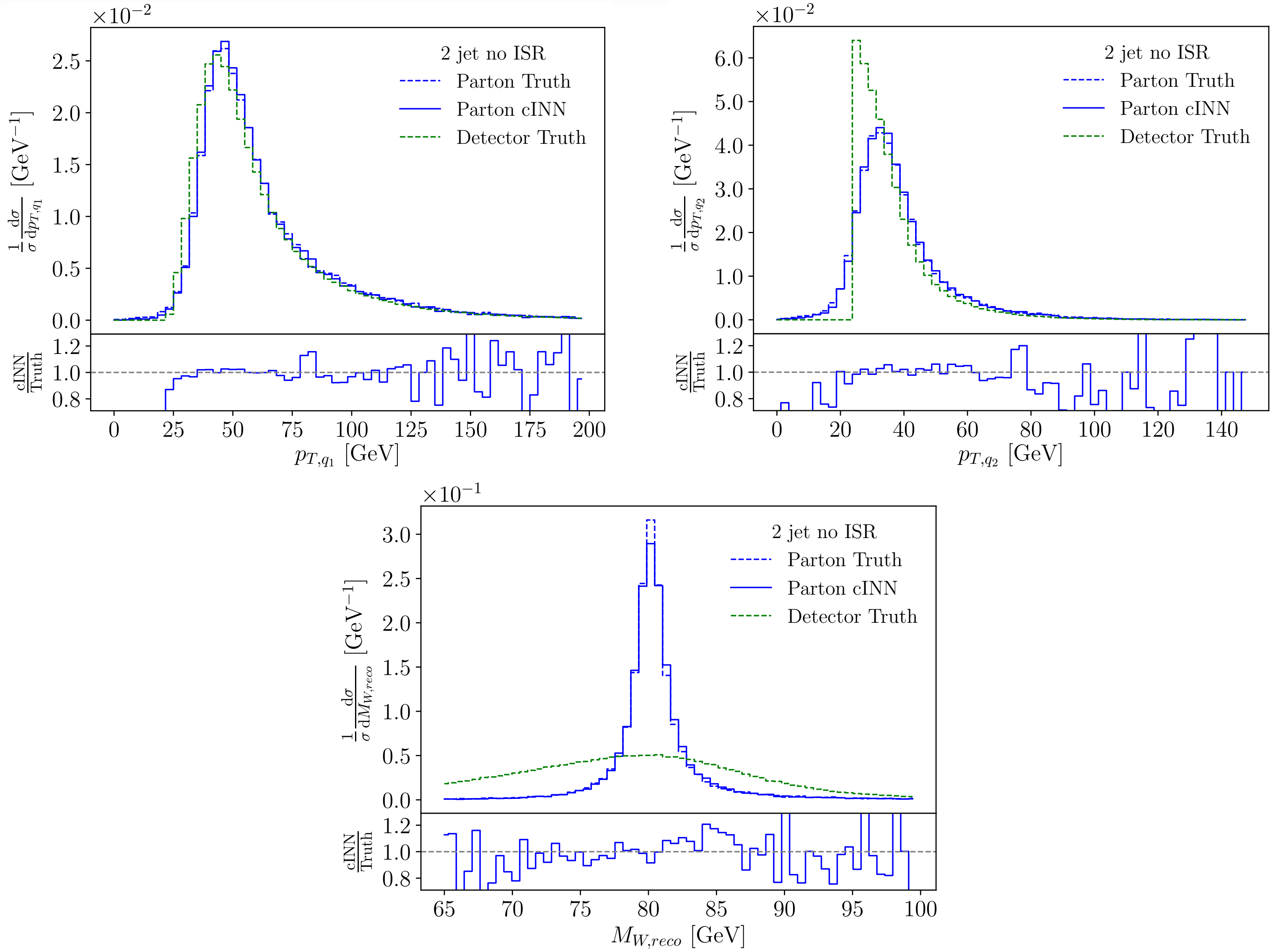
$$x_p \xleftrightarrow[\leftarrow \text{ unfolding: } \bar{g}(r, f(x_d))]{g(x_p, f(x_d)) \rightarrow} r$$

$\rightarrow$ Training: Maximize posterior over model parameters

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

$\rightarrow$ Jacobian of bijective mapping

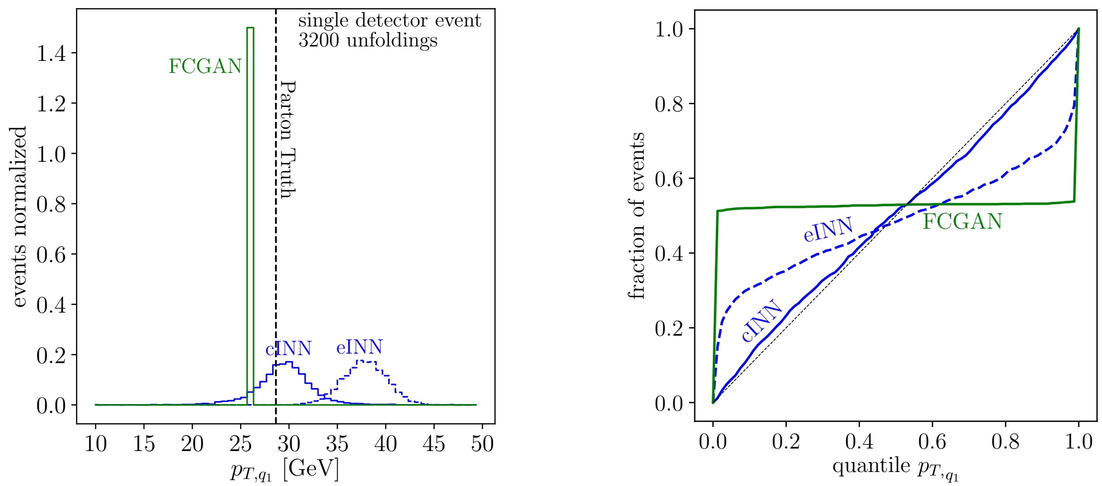
Cross check distributions



Condition INN on detector data [2006.06685]

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

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



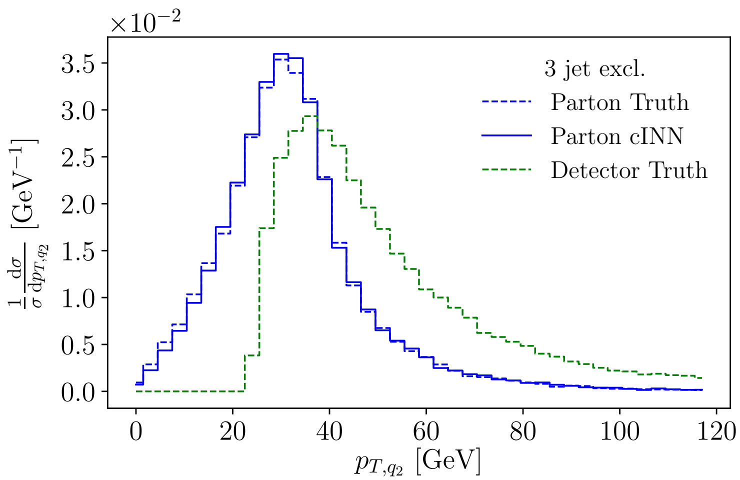
multi-dimensional ✓ bin independent ✓ statistically well defined ✓

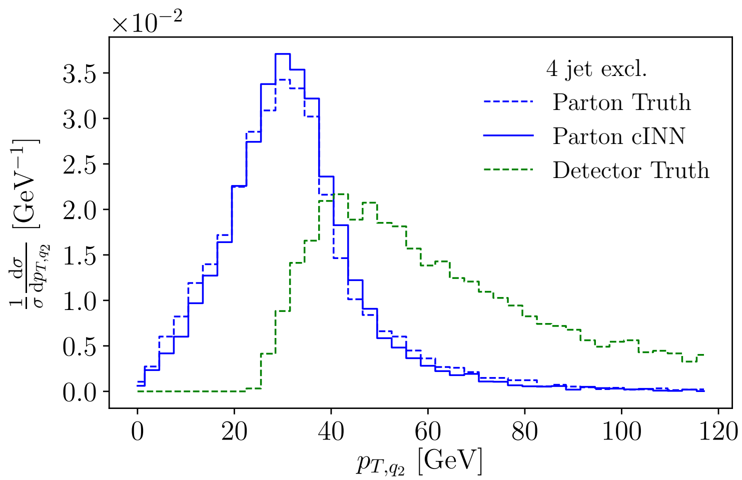
Inverting the full event I

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

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

- Train and test on exclusive channels

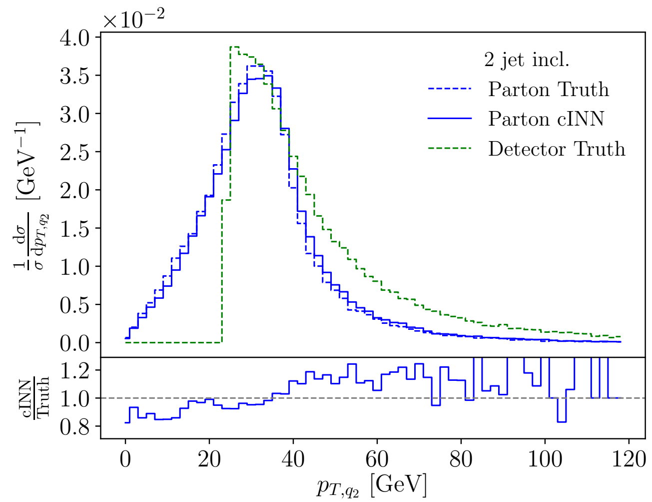




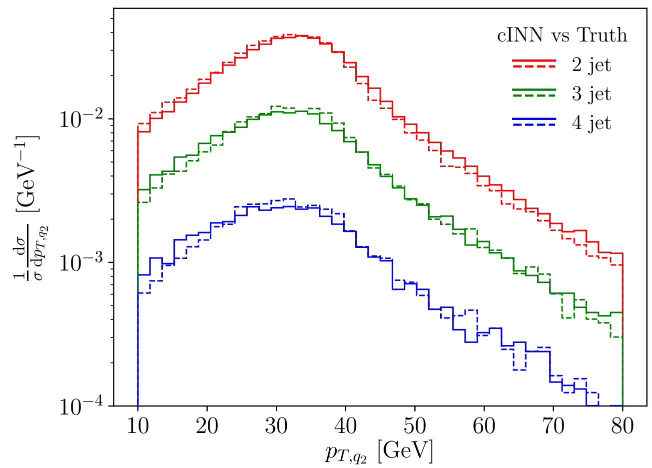
Inverting the full event II

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

Train on inclusive dataset



Evaluate
exclusive 2/3/4 jet channels



We can use ML ...

... to enable precision simulations in forward direction

... to turn weighted into unweighted events

... to invert the simulation chain statistically

... for fun and precision :)