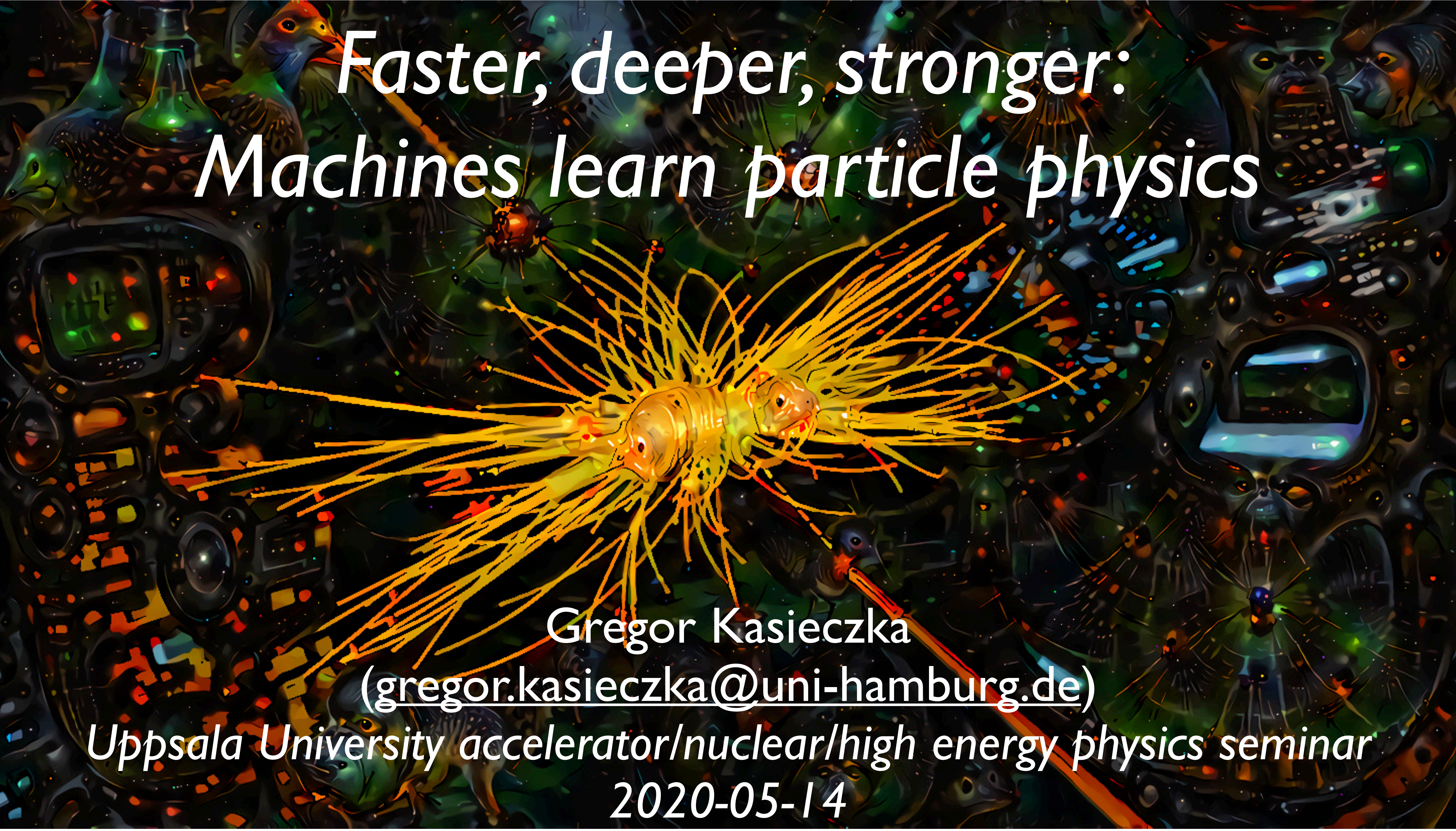


Faster, deeper, stronger: Machines learn particle physics

Gregor Kasieczka
(gregor.kasieczka@uni-hamburg.de)

Uppsala University accelerator/nuclear/high energy physics seminar

2020-05-14

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

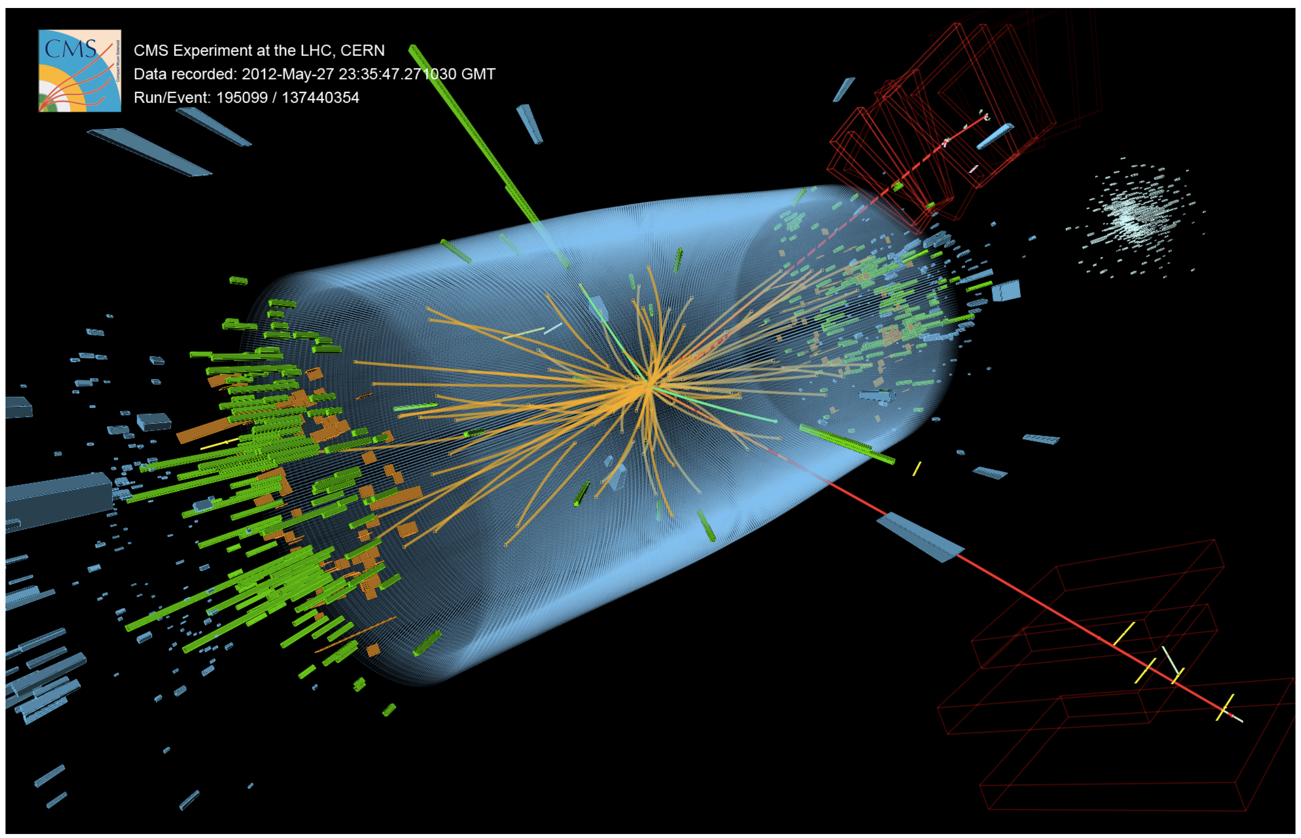
Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

PIER
Partnership of Universität Hamburg and DESY

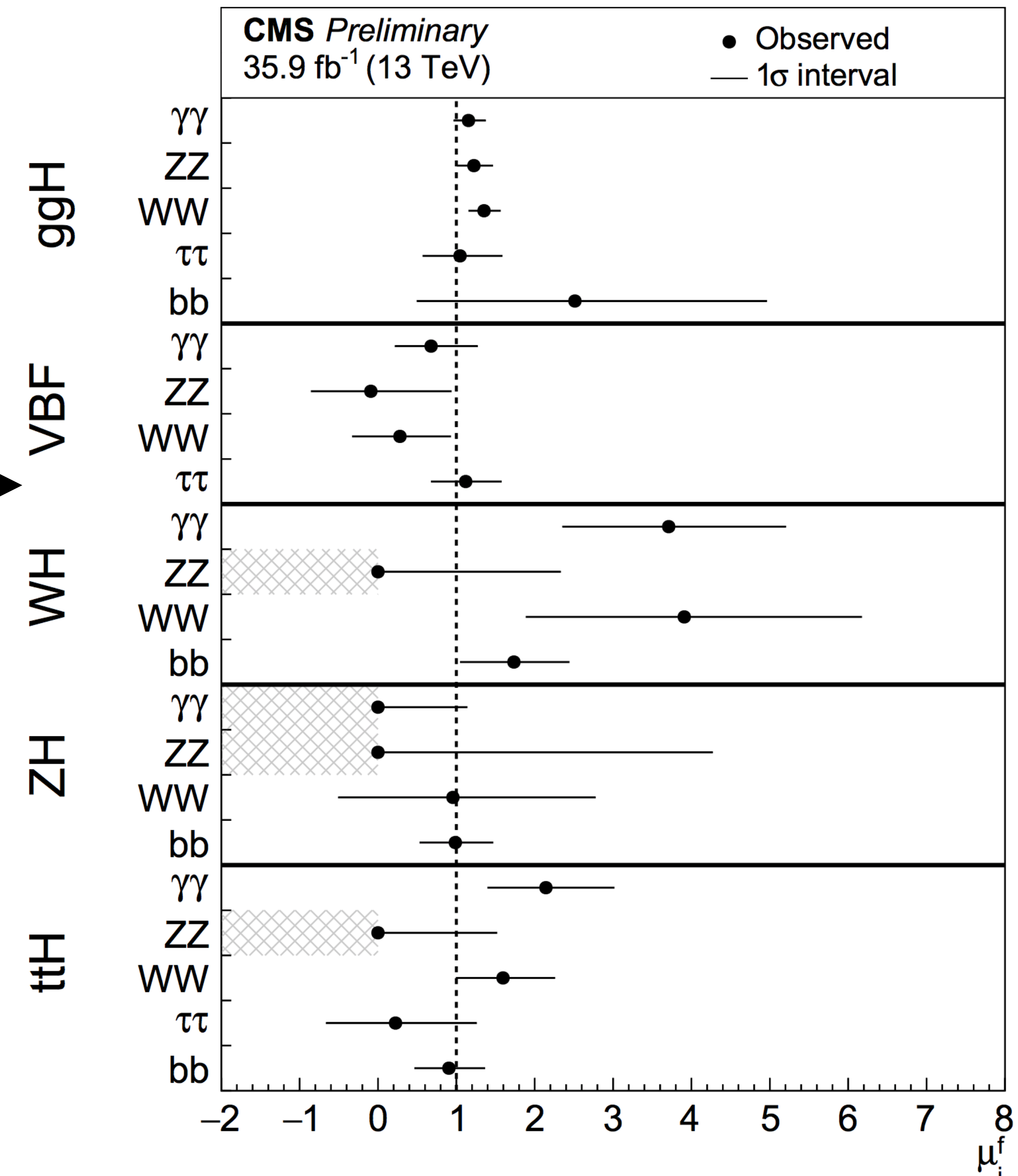
Bundesministerium für Bildung und Forschung

Higgs Boson: Discovery to Precision…

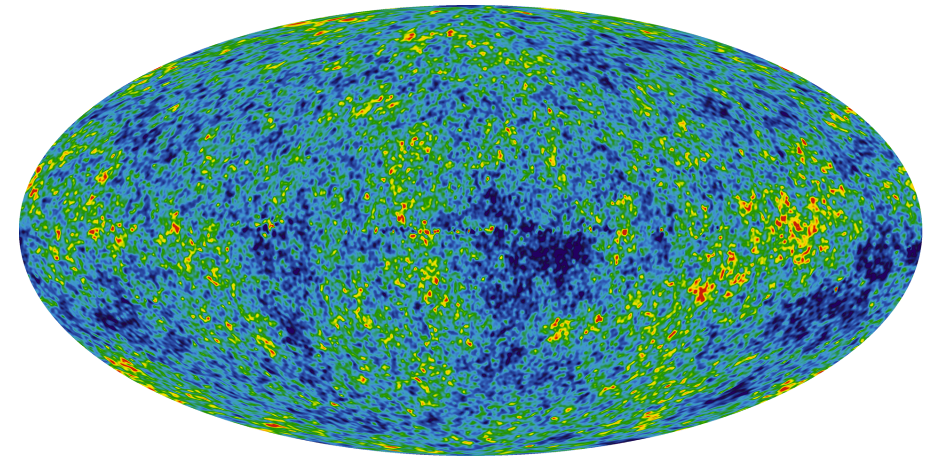


2012: Discovery of the Higgs boson

Now →

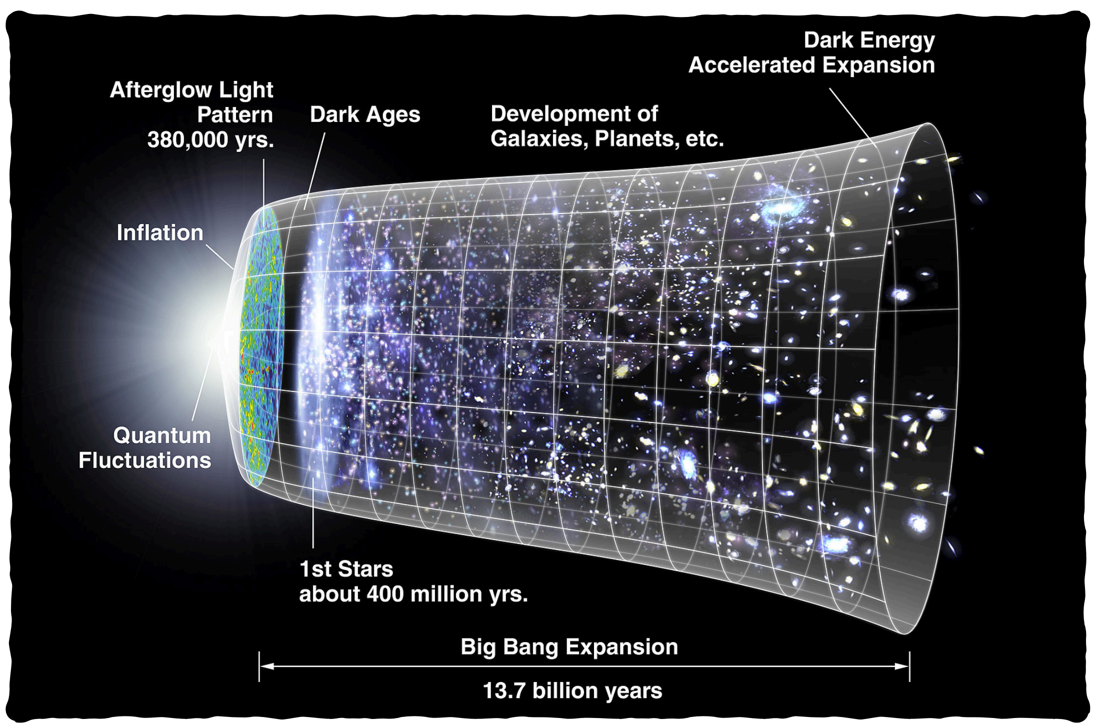


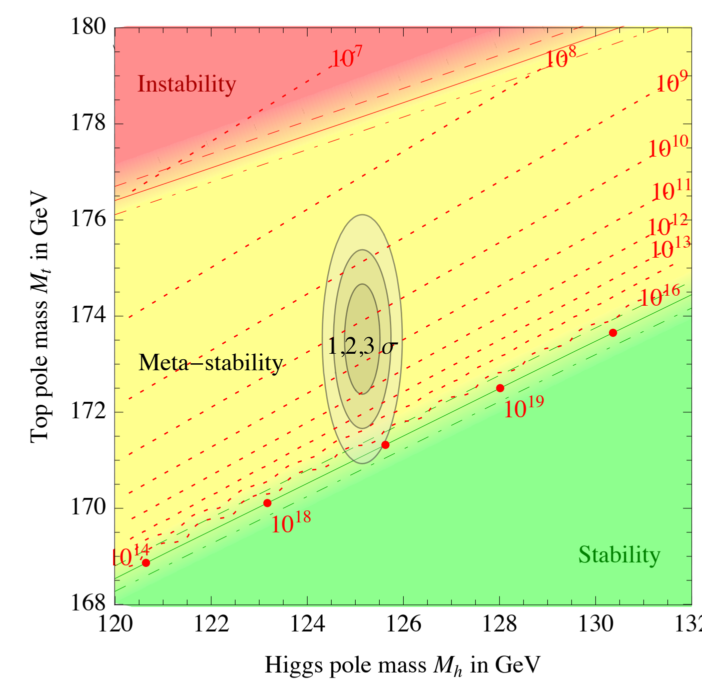
Why are neutrinos massive?



What are the origins of the LHCb flavour anomaly?

What is the nature of dark matter & dark energy?



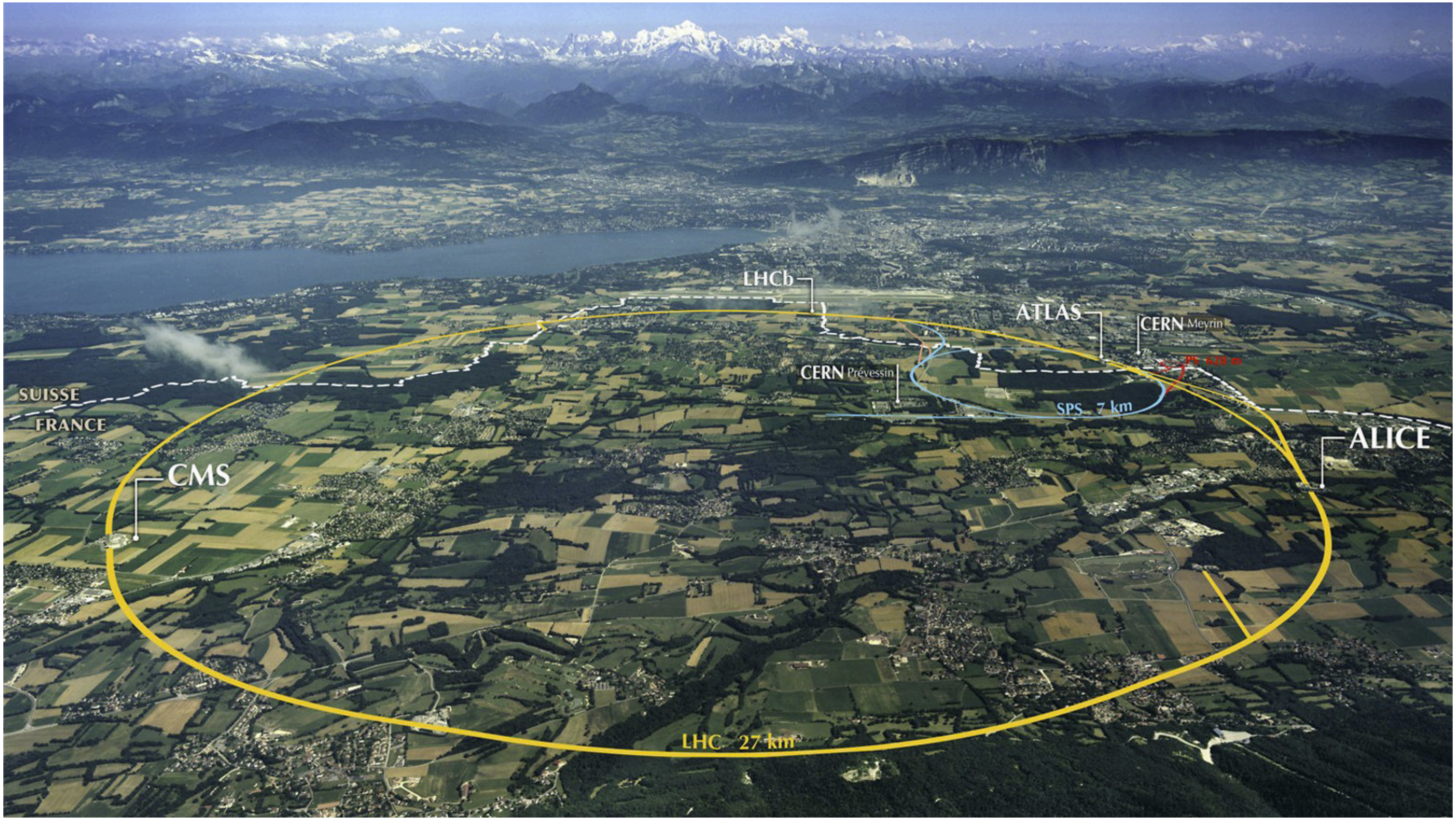


Is the electroweak vacuum stable?

Why is there more matter than anti-matter?

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

What are the details of cosmic inflation?

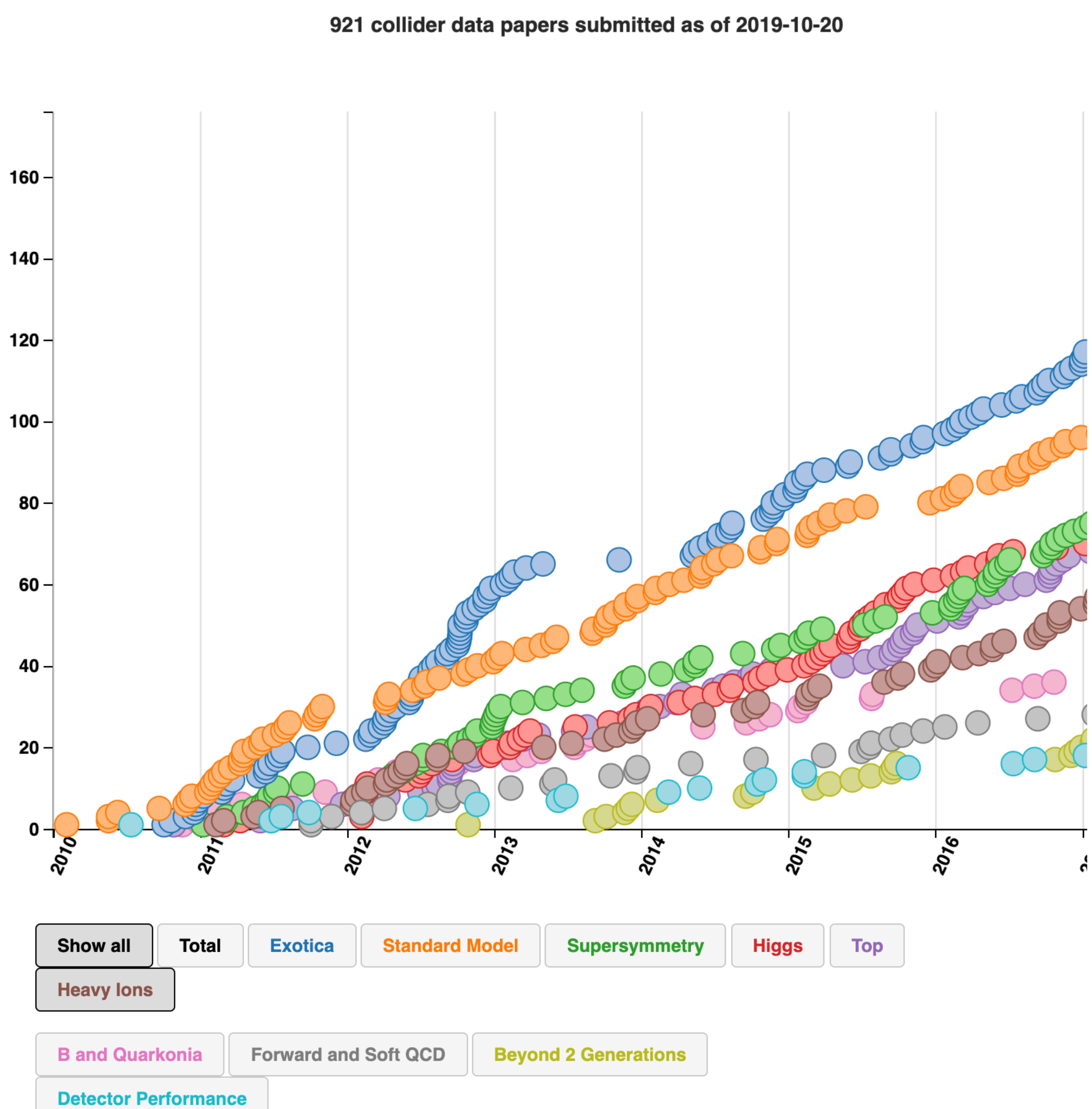


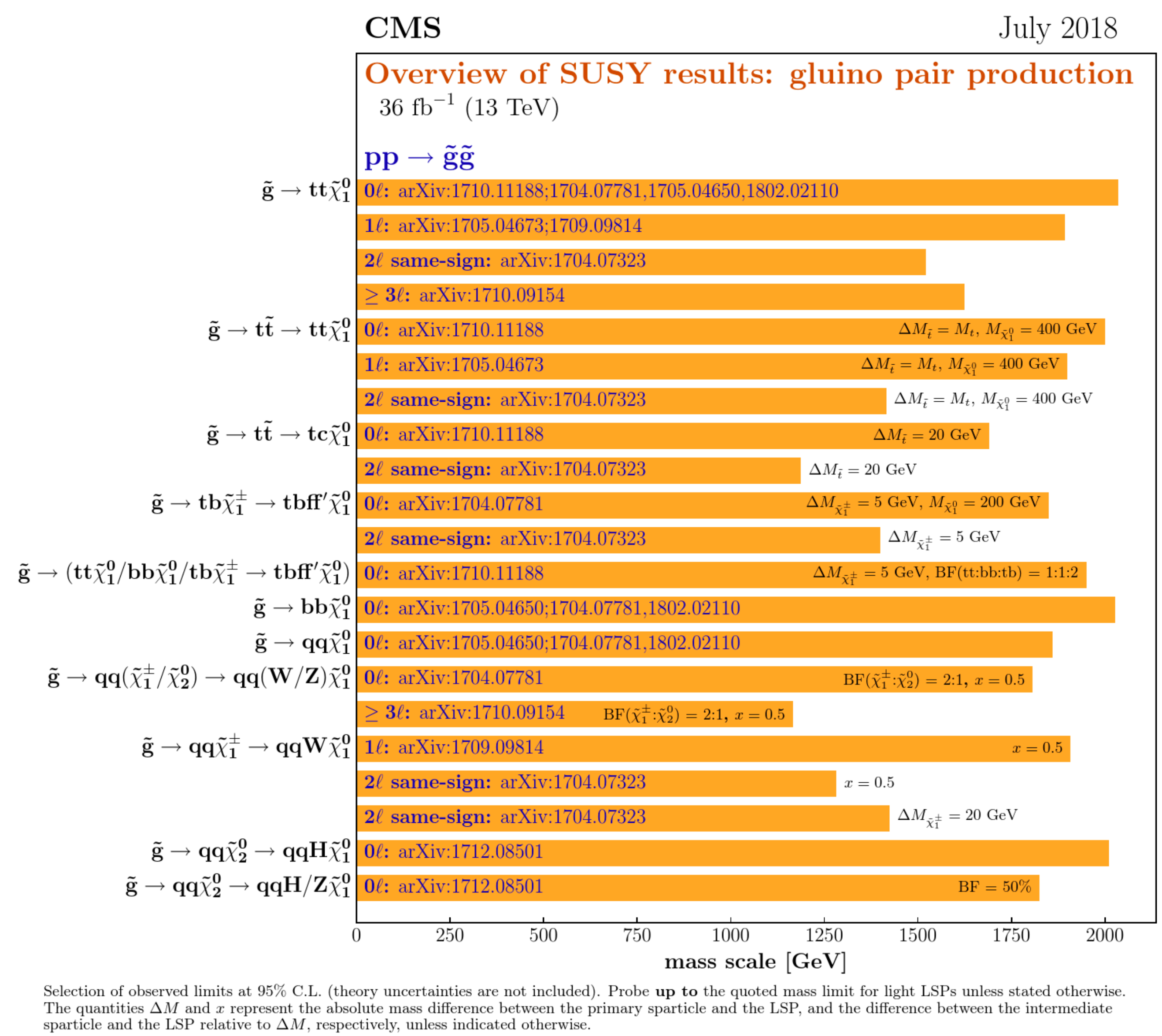
- *LHC: 27 km circumference*
- *Collide protons with a centre-of-mass energy of 13 TeV (99.999999% of speed of light)*
- *40 Million collisions/second in ATLAS/CMS*
- *~25 Petabyte collision data/year / experiment*

- *Planned High Luminosity Upgrade (HL-LHC)*
- *Higher data rate, higher pile-up*
  - *Big data challenges ahead!*
- *Begin operation O(2026)*

Many results…

…but no new physics so far



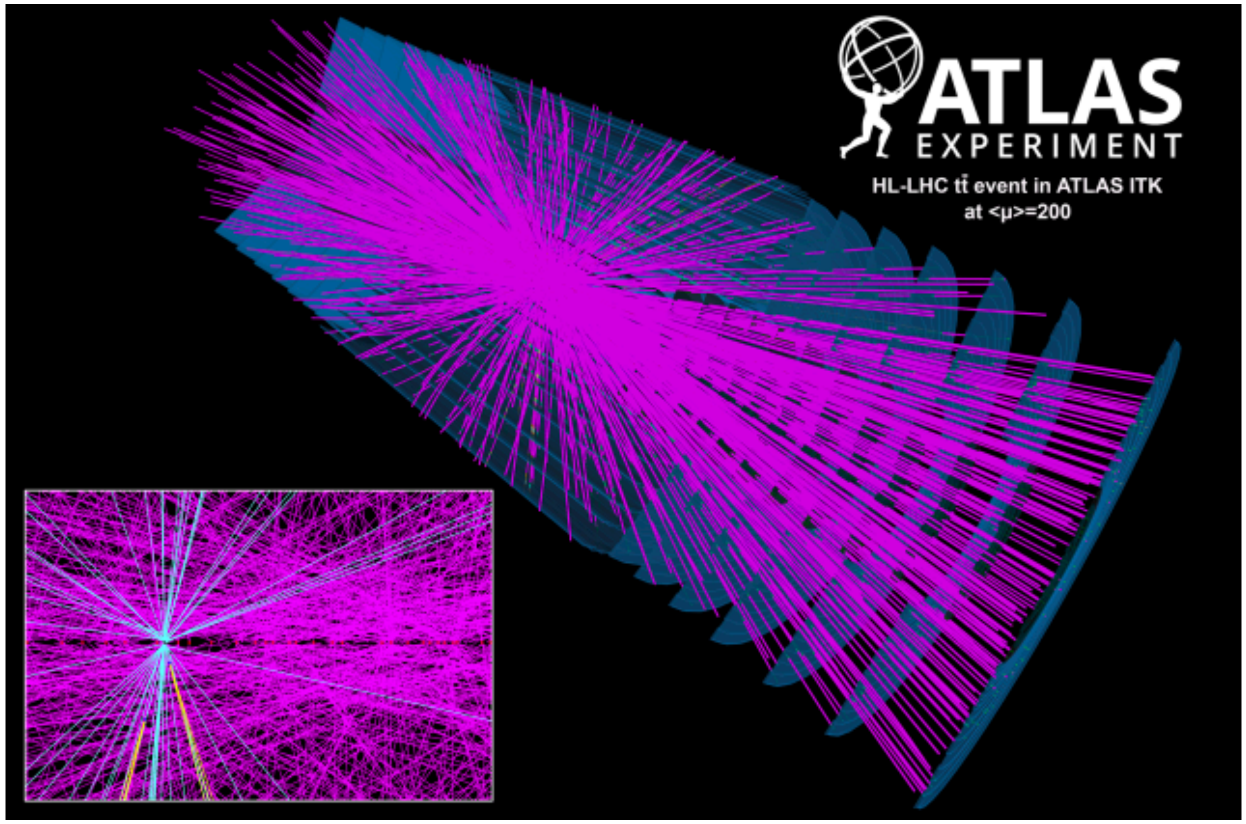


http://cms.web.cern.ch/org/physics-papers-timeline

What next?

- Precision measurements and searches for new physics need
  - better tools to identify known particles and processes
  - higher accuracy and speed
- Finding unknown signatures needs
  - new ways of analysing data
- Future data taking with higher collision rates needs:
  - faster reconstruction and triggering
  - faster simulation and event generation

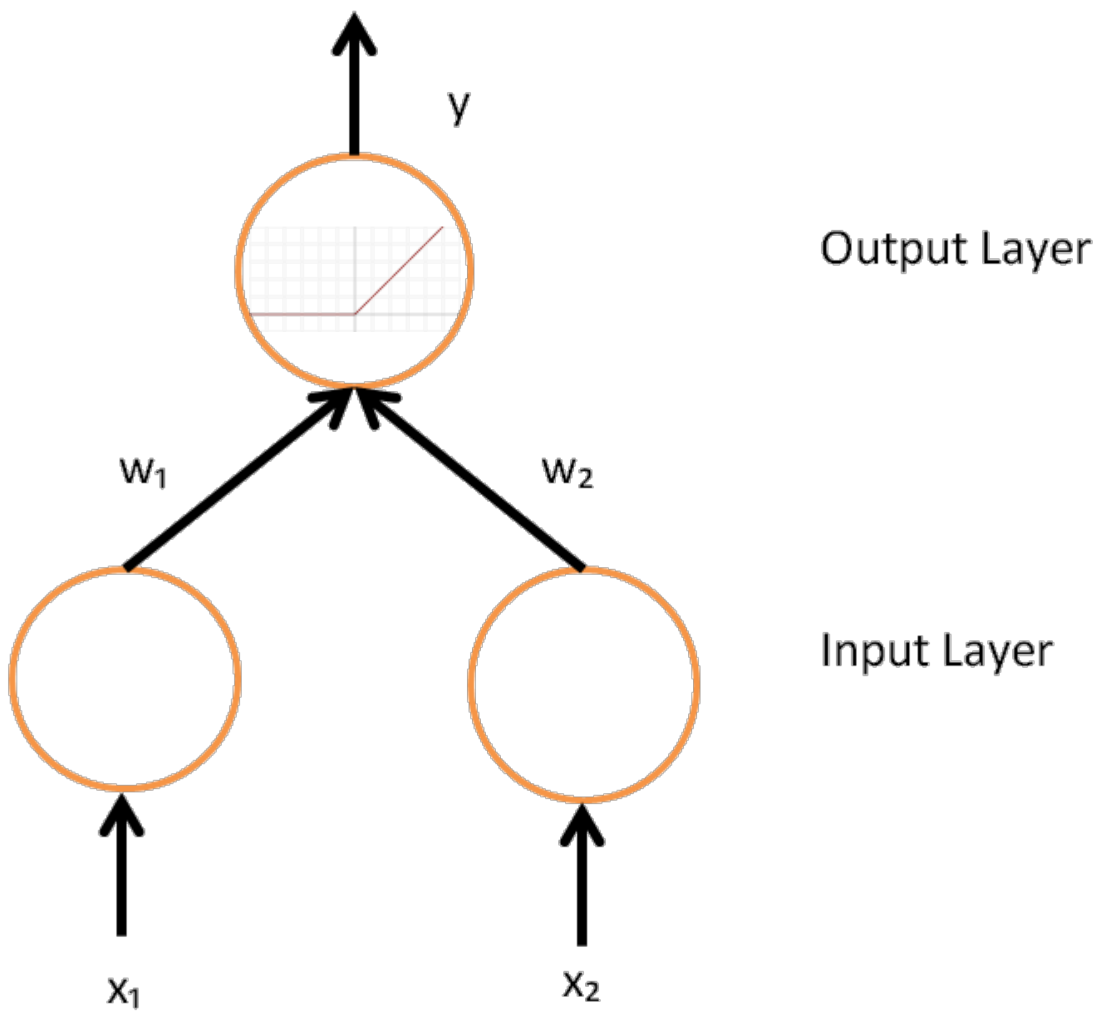
(a) promising answer: **Deep Learning**



Prelude

What is deep learning?

Basics of Neural Networks



- *Backpropagation + Gradient descent*
  - Pass input (x_1, x_2) to neural networks
  - Calculate output *y* and (problem specific) loss function *L*
  - Find gradient of loss function with respect to weights
  - Use gradient to find new weights

Supervised

Attempt to learn some target: classification or regression tasks

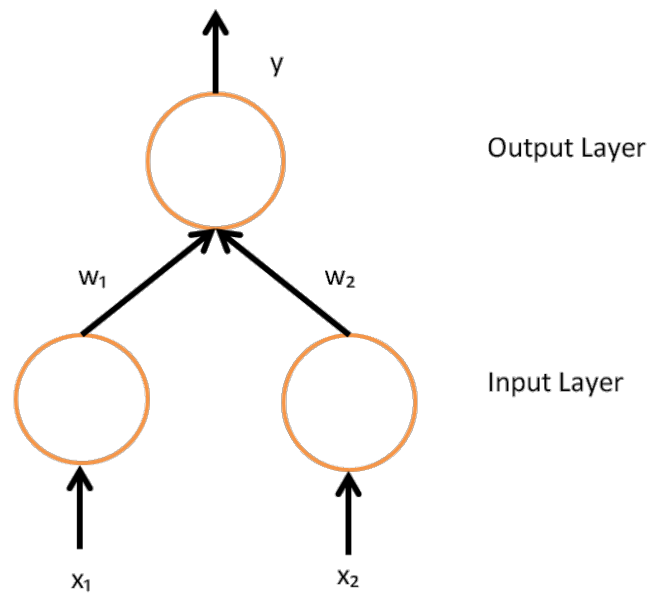
Need to have a dataset with known targets (typically from MC simulation)

Unsupervised

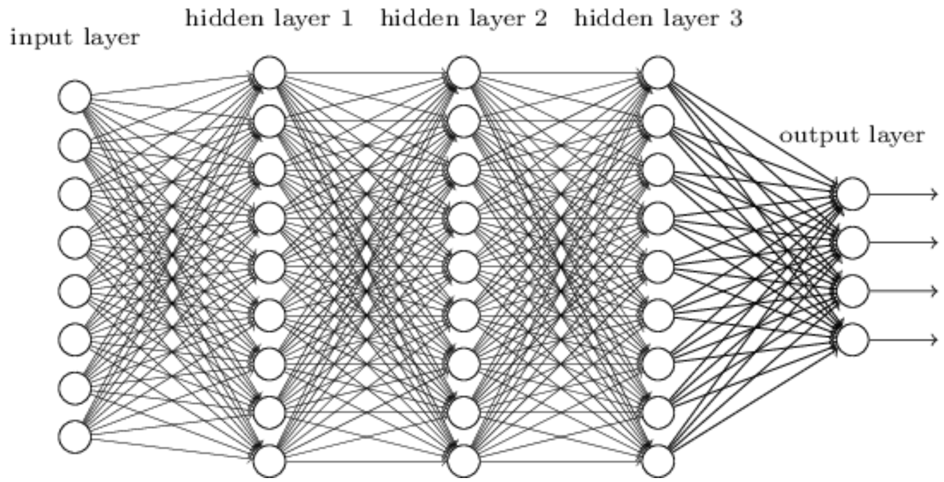
No target, learn the probability distribution

Generative models and anomaly detection

Complexity



2 weights



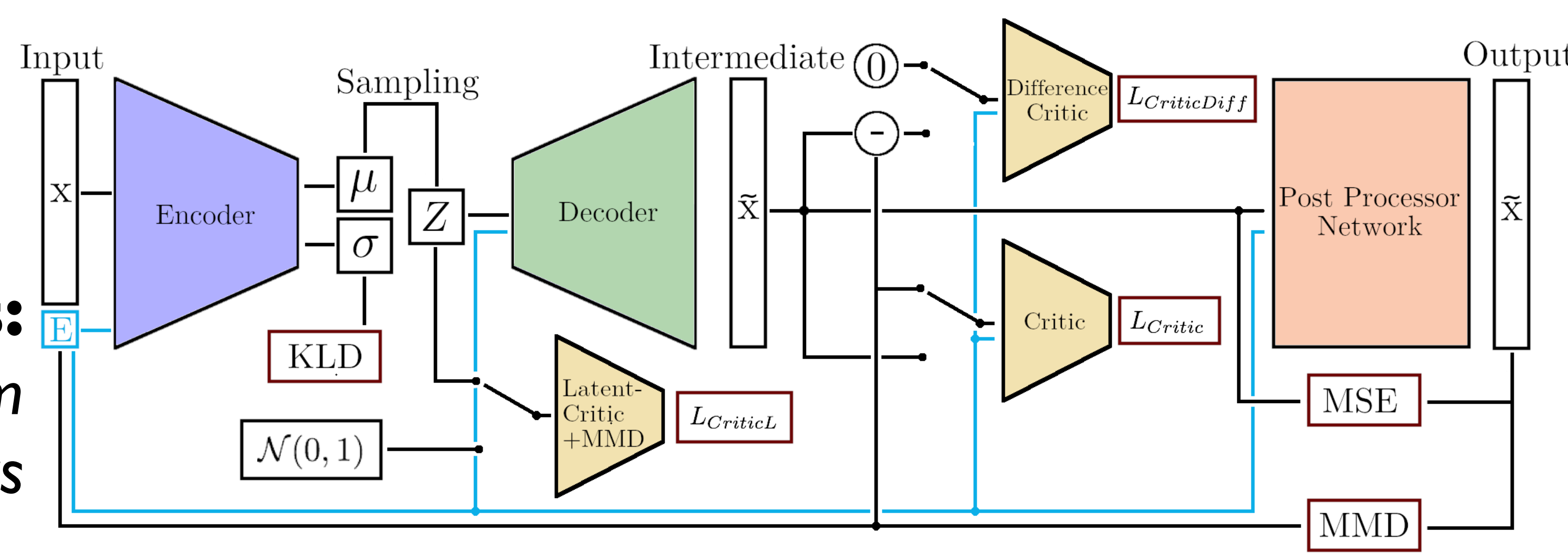
300 weights

25 million weights:
2016 state of the art for image classification

stage	output	ResNet-50	ResNeXt-50 (32×4d)
conv1	112×112	7×7, 64, stride 2	7×7, 64, stride 2
conv2	56×56	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128, C{=}32 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3	28×28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256, C{=}32 \\ 1\times1, 512 \end{bmatrix}\times4$
conv4	14×14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512, C{=}32 \\ 1\times1, 1024 \end{bmatrix}\times6$
conv5	7×7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 1024 \\ 3\times3, 1024, C{=}32 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	global average pool 1000-d fc, softmax	global average pool 1000-d fc, softmax
# params.		**25.5**$\times10^6$	**25.0**$\times10^6$
FLOPs		**4.1**$\times10^9$	**4.2**$\times10^9$

Deep Learning:
Complex network + low level inputs

71 million weights:
generative network from physics



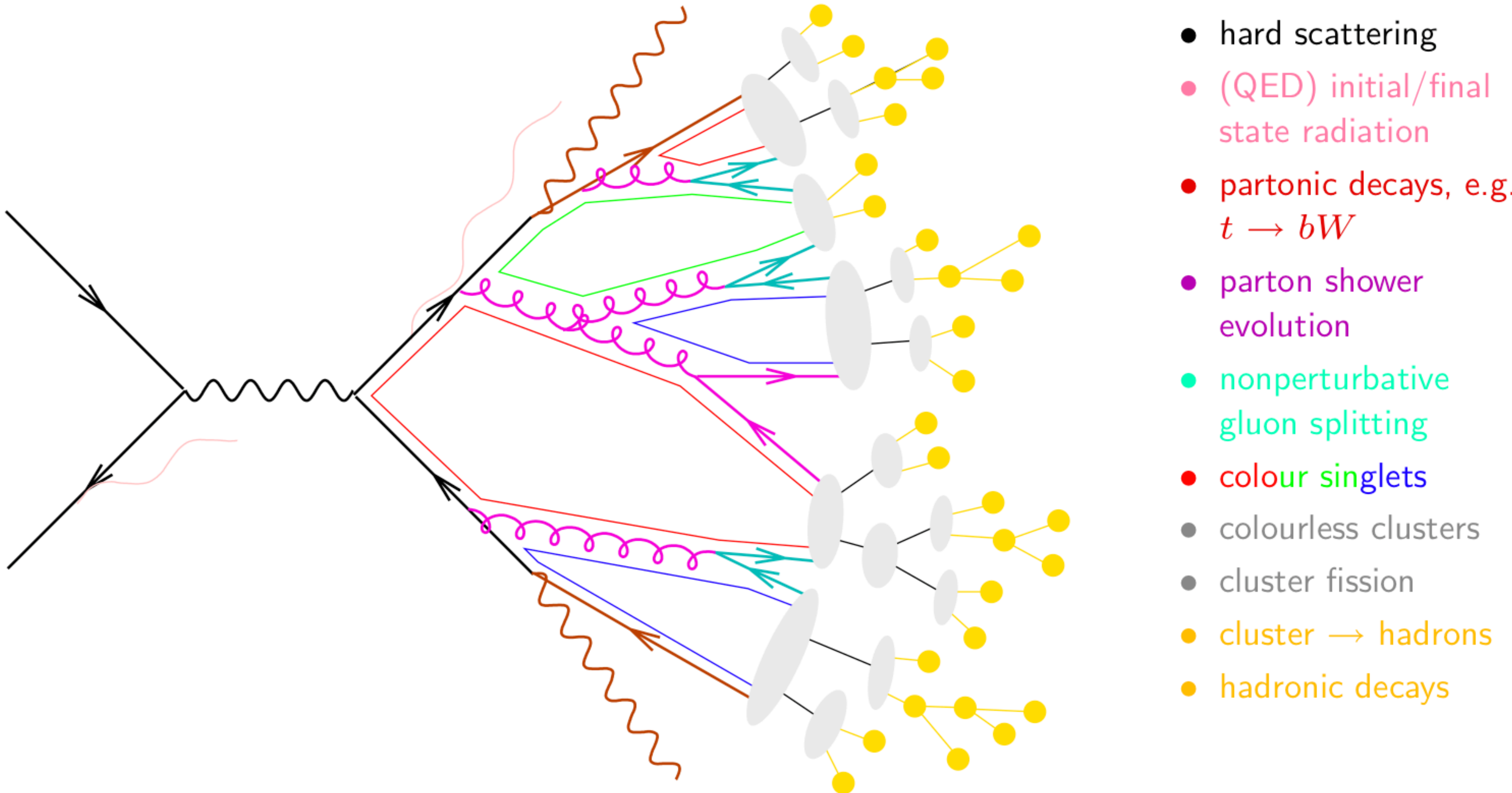
Menu

Supervised Classification:
Heavy Resonance Tagging
Event Classification

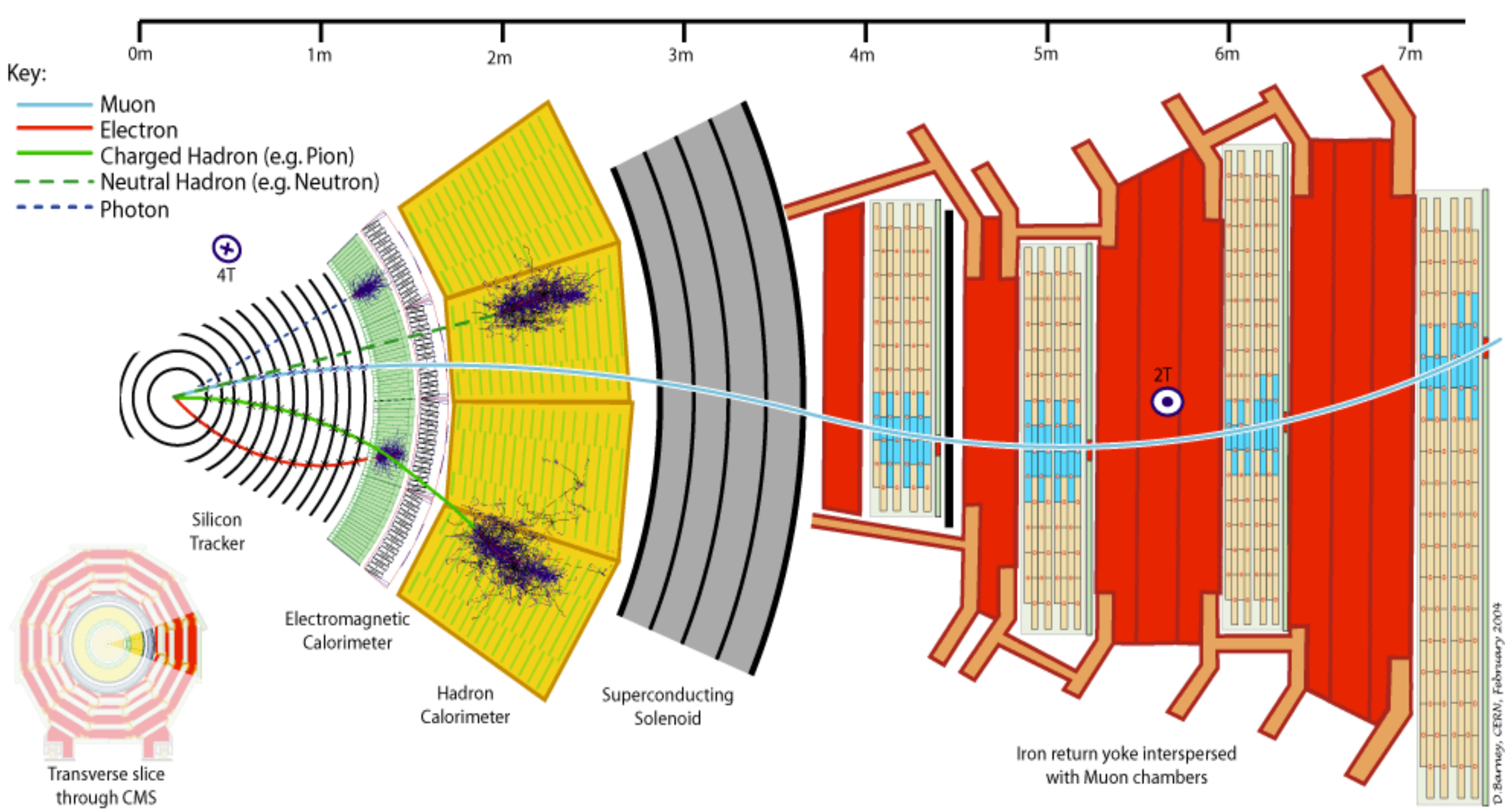
Generative Models:
Fast calorimeter simulation

Learning from Data:
Anomaly Detection

Build better tools to identify known particles



+



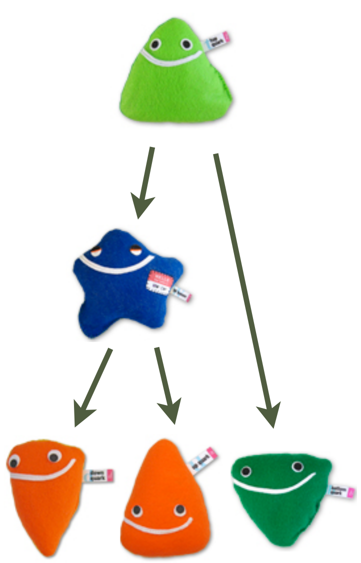
We want to infer underlying physics from measurements in the detector.

How can deep neural networks assist us?

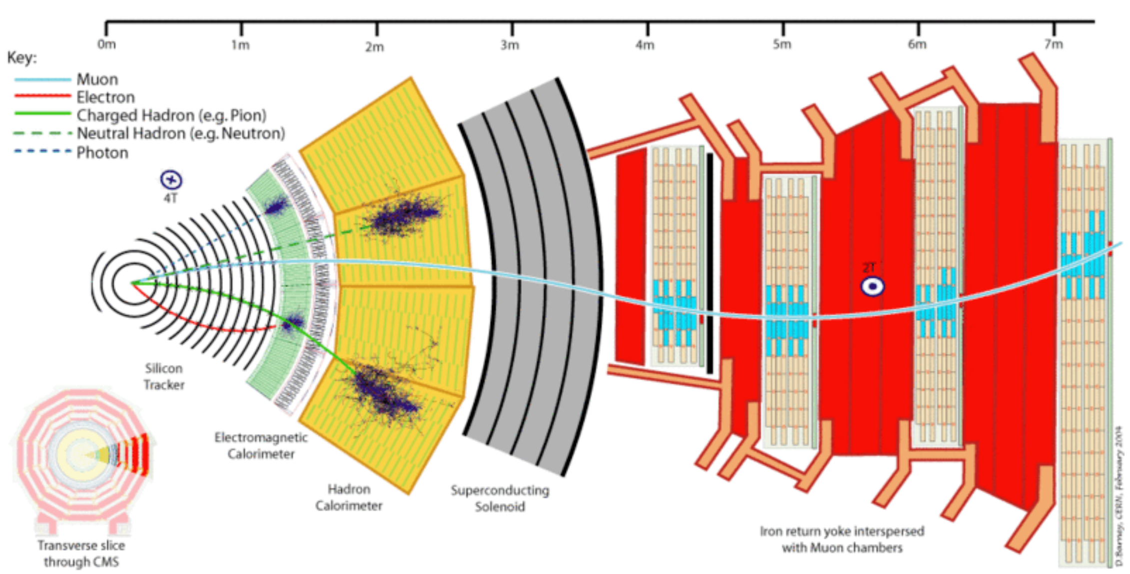
Heavy Resonance Tagging

- **Goal**: Distinguish decay products of heavy resonance (top quark, W/Z boson, Higgs boson) from other particles (light quark/gluon jets)
- Needed for searches and measurements
- Achieve by looking at substructure of jets in the detector

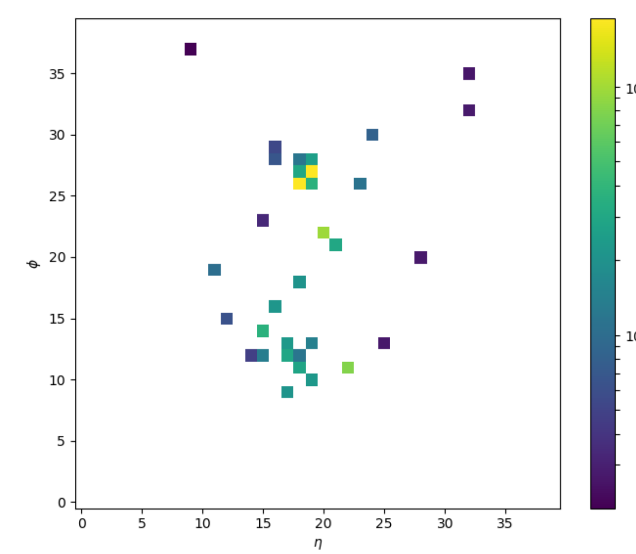
Top Quark *(Simulated) Detector* *10 Images*



+



=

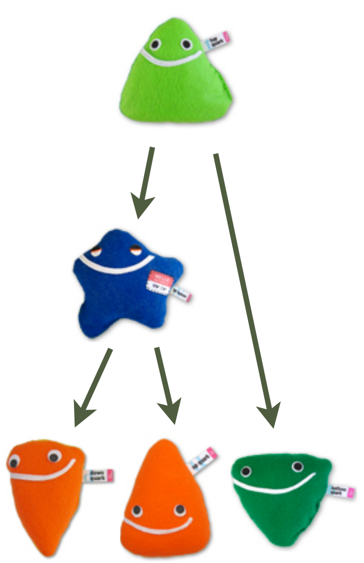


(ignoring parton shower, hadronisation,...)

Heavy Resonance Tagging

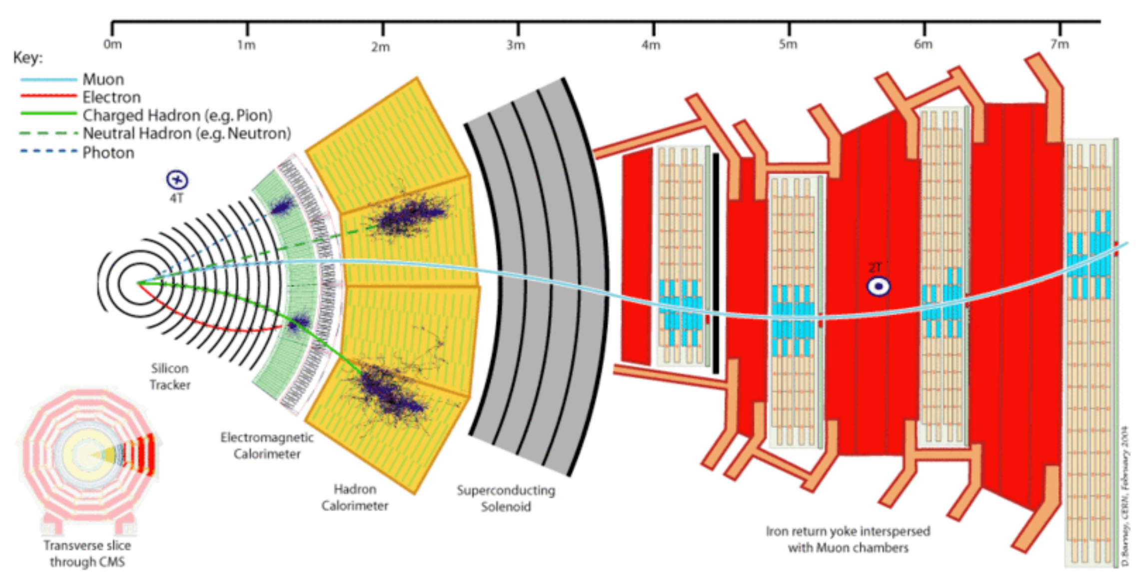
- **Goal**: Distinguish decay products of heavy resonance (top quark, W/Z boson, Higgs boson) from other particles (light quark/gluon jets)
- Needed for searches and measurements
- Achieve by looking at substructure of jets in the detector

Top Quark



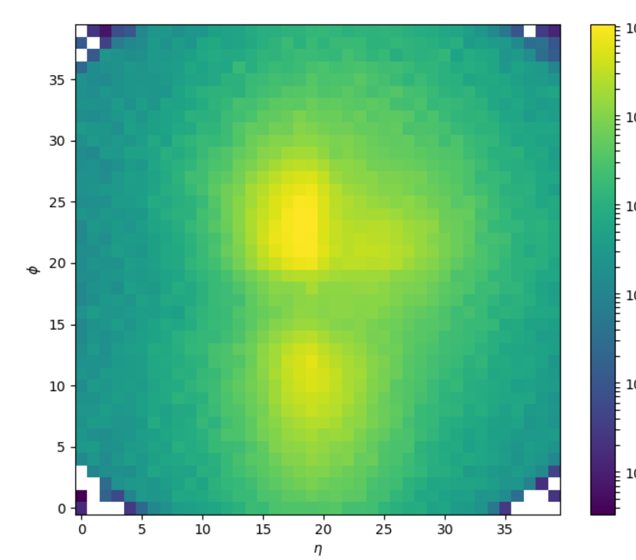
+

(Simulated) Detector



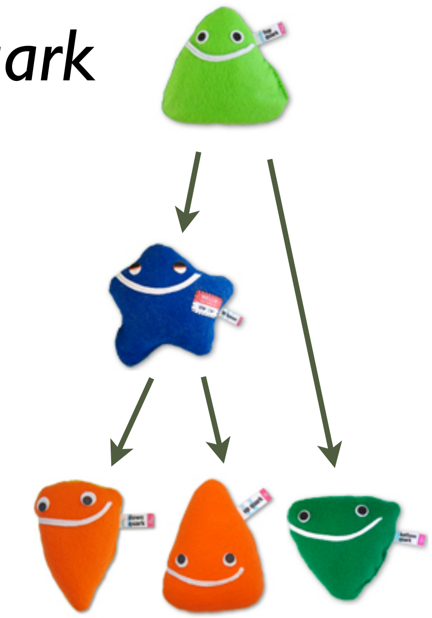
=

10000 Images

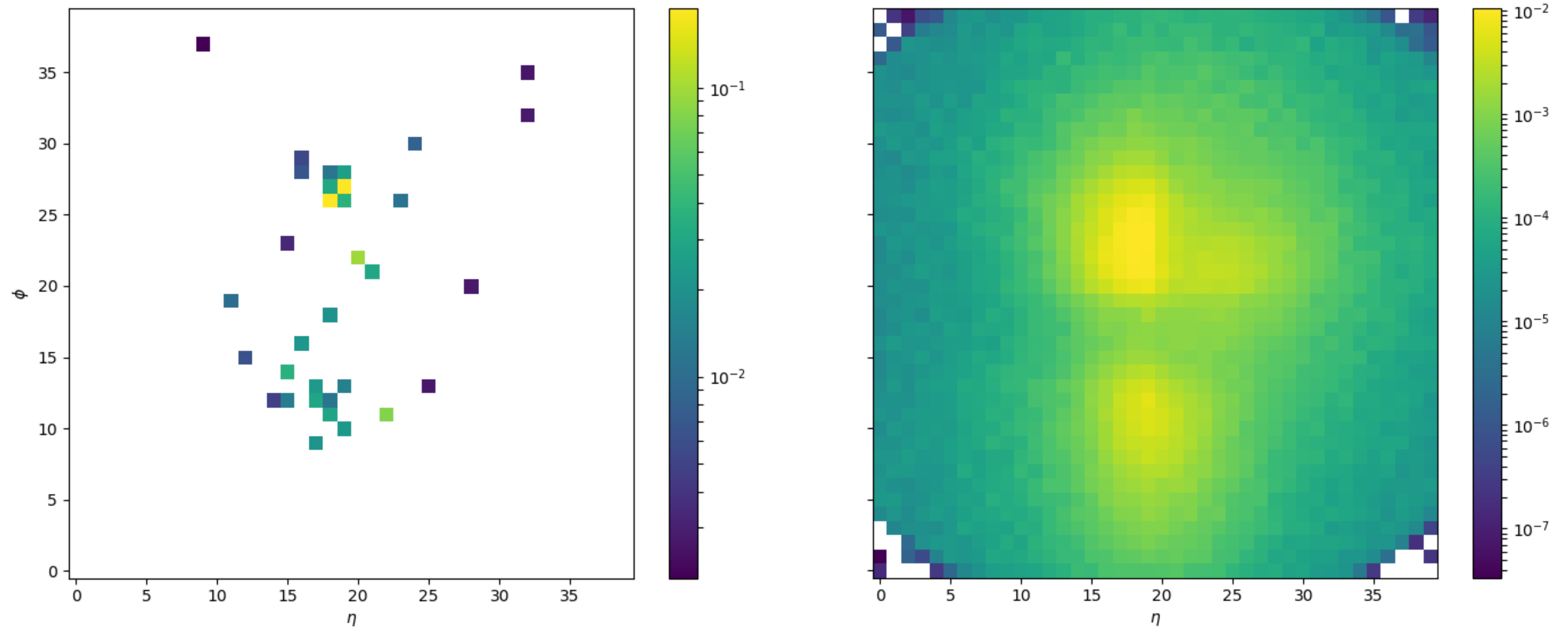


(ignoring parton shower, hadronisation,...)

Top Quark Jet

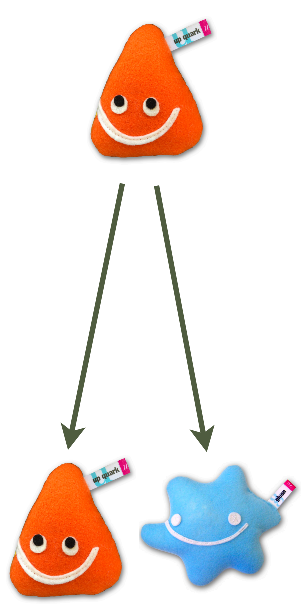


=

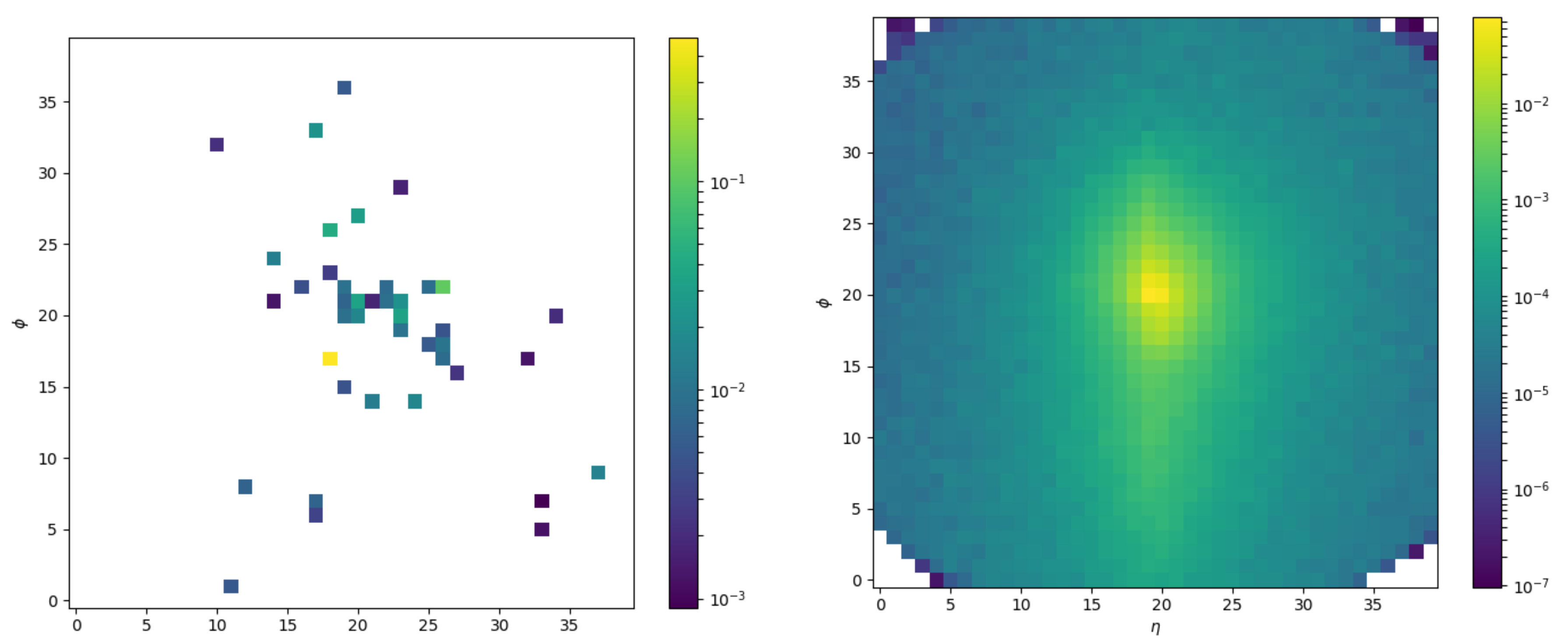


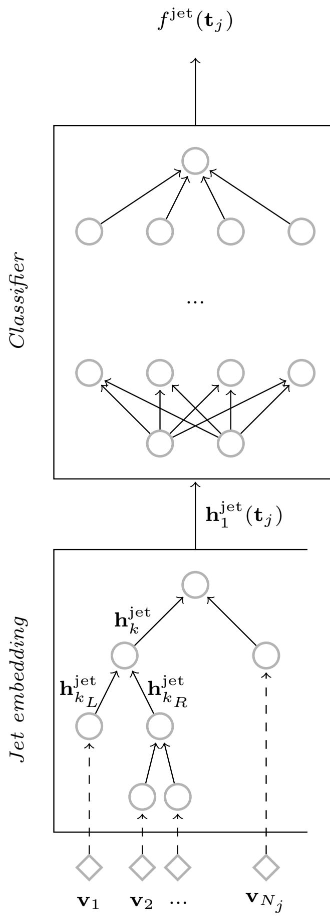
- *Binary classification task*
- *Fully supervised learning (using simulation)*
- *40x40 Pixels, E_T*
- *Perfectly suited for deep learning algorithms*

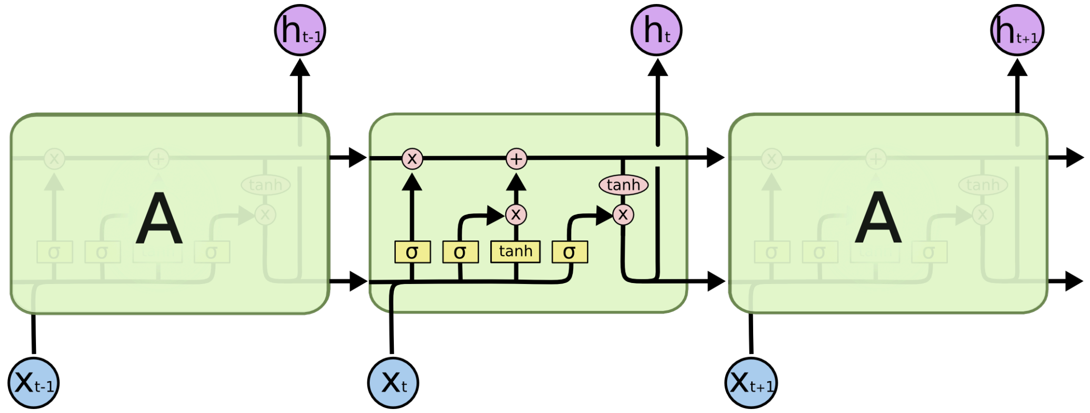
QCD Jet

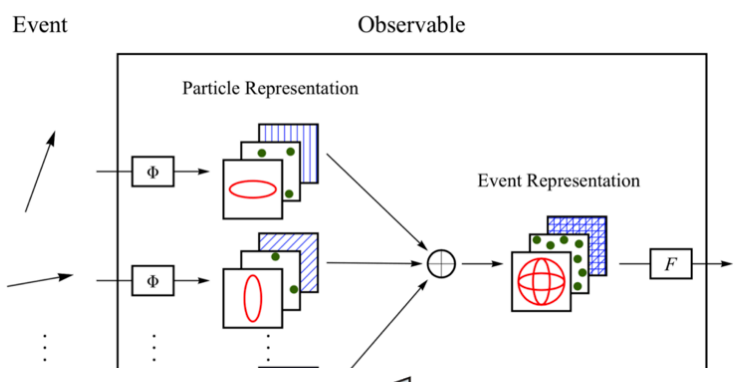


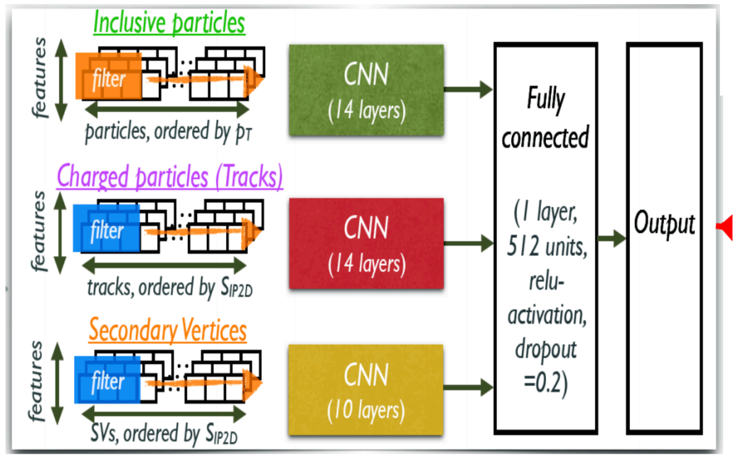
=

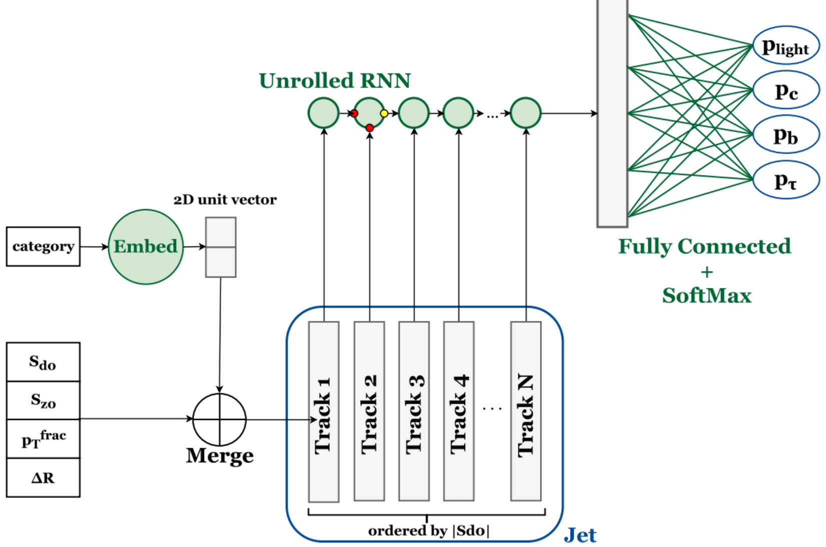


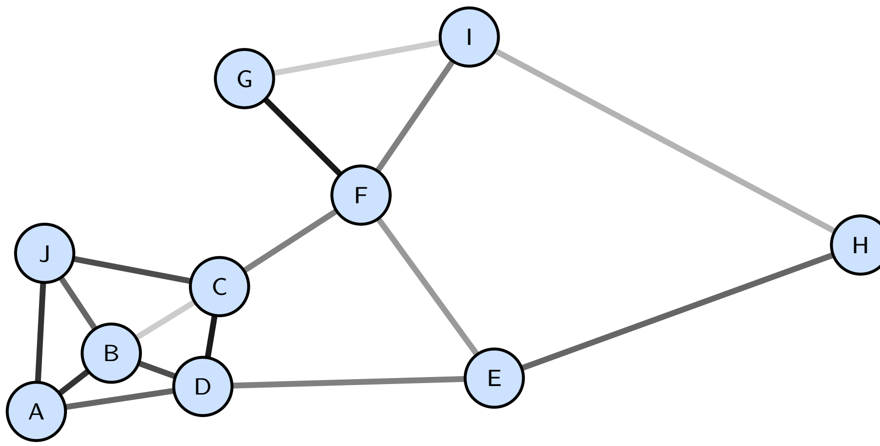


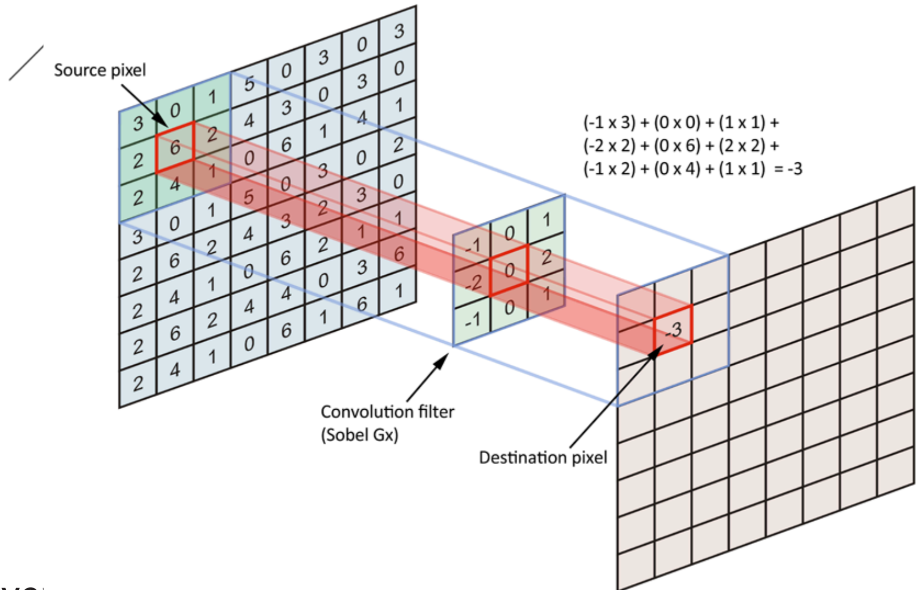


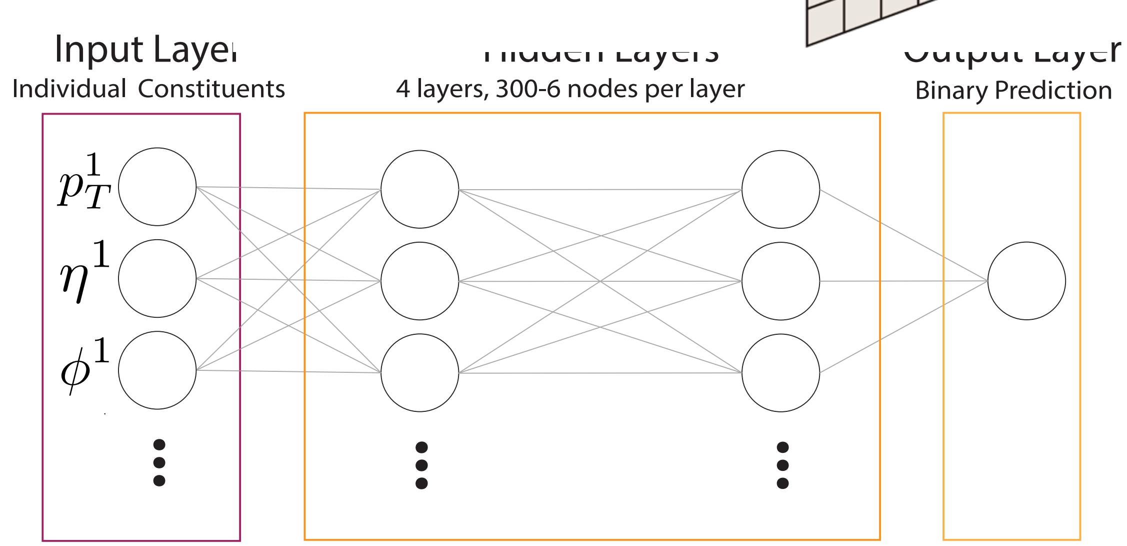












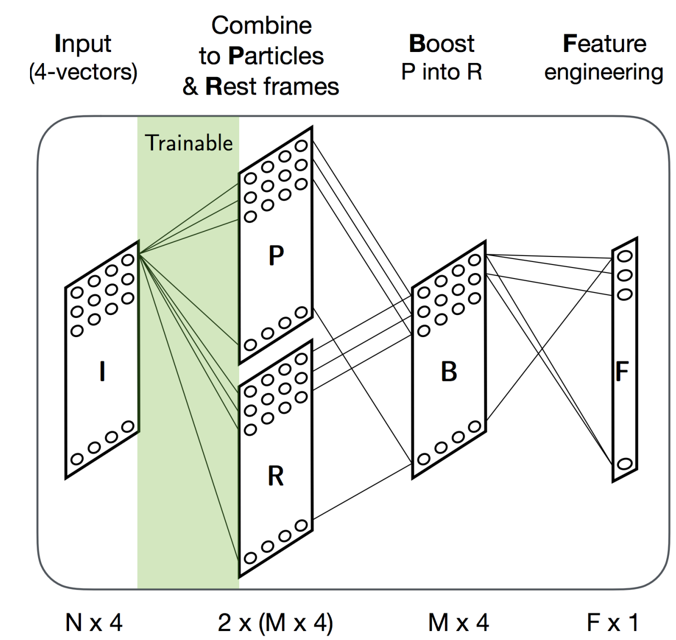
$$k_{\mu,i} = \begin{pmatrix} E_0 & E_1 & \cdots & E_N \\ p_{x,0} & p_{x,1} & \cdots & p_{x,N} \\ p_{y,0} & p_{y,1} & \cdots & p_{y,N} \\ p_{z,0} & p_{z,1} & \cdots & p_{z,N} \end{pmatrix}$$

Combination Layer (**CoLa**): create linear combinations:

$$k_{\mu,i} \xrightarrow{\text{CoLa}} \widetilde{k}_{\mu,j} = k_{\mu,i}\, C_{ij}$$

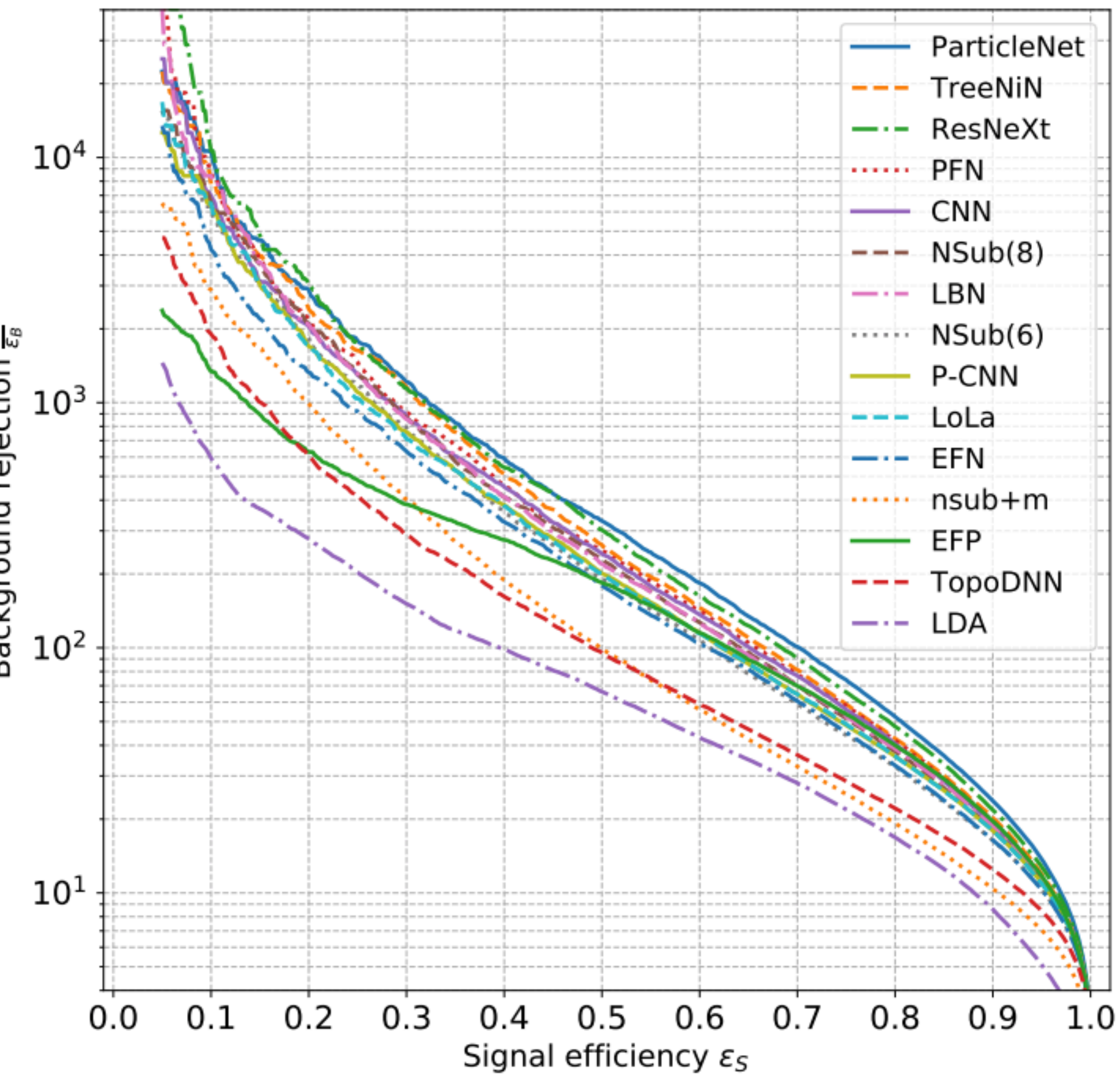
Lorentz Layer (**LoLa**): Use resulting matrix to extract physics features.
Main assumption is the Minkowski metric

$$\widetilde{k}_{\mu,i} \rightarrow \sum_j (\widetilde{k}_i - \widetilde{k}_j)_\mu (\widetilde{k}_i - \widetilde{k}_j)_\nu \eta^{\mu\nu} B_{ij}$$



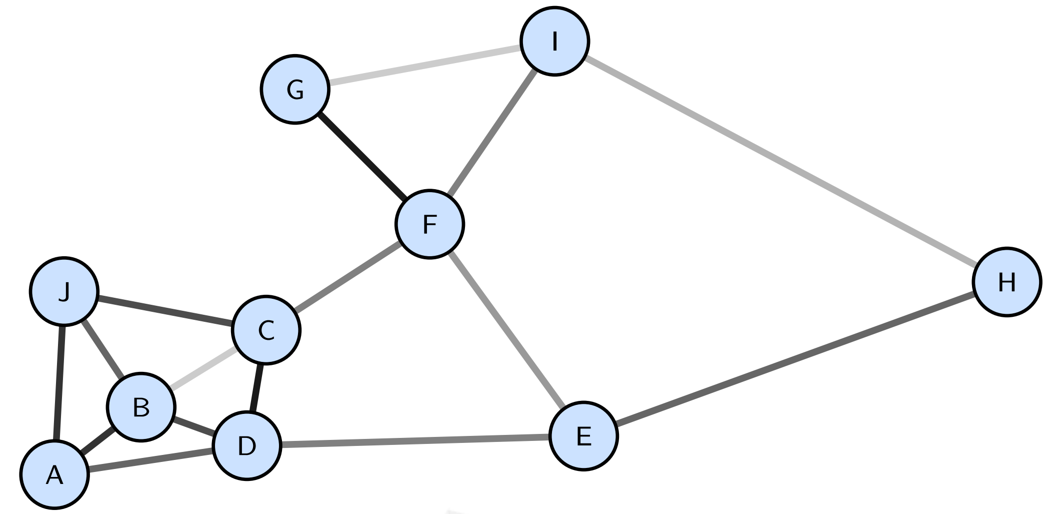
Architecture Overview

Community performance comparison (toy <u>dataset public</u>):
1902.09914 (GK, Plehn, et al)

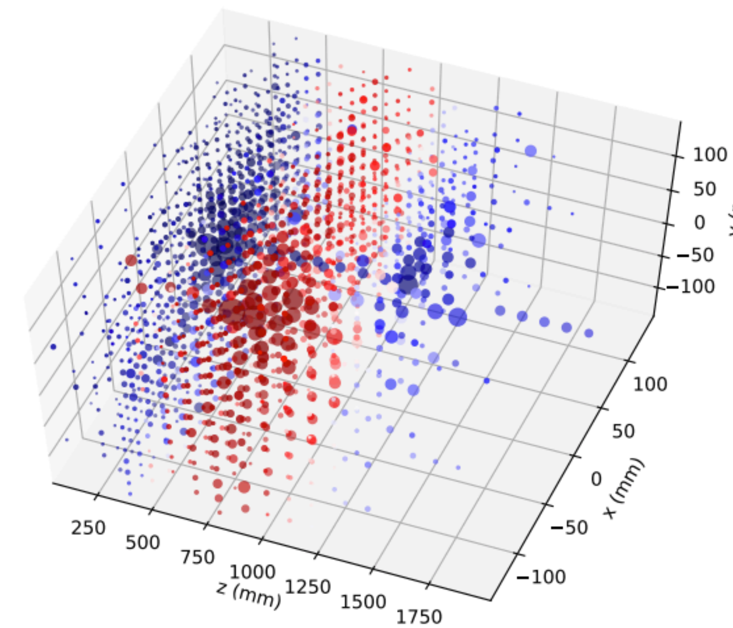


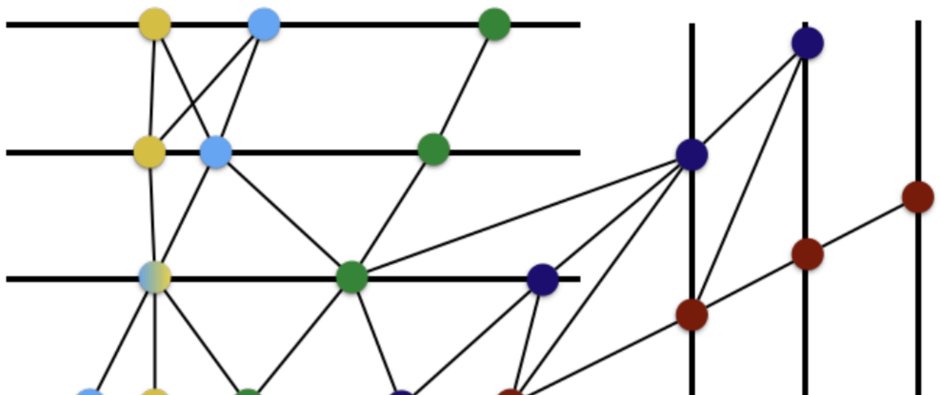
- 1.2M simulated top quark and background events
- Great test-bed to compare different data representations
  - (and, of course, useful for new physics searches, top/Higgs measurements)
- Still surprising gains in performance
  - Although it needs to be seen how well these translate to data
- (Also developments in flavour tagging, not covered here)
- Best performance from graph networks

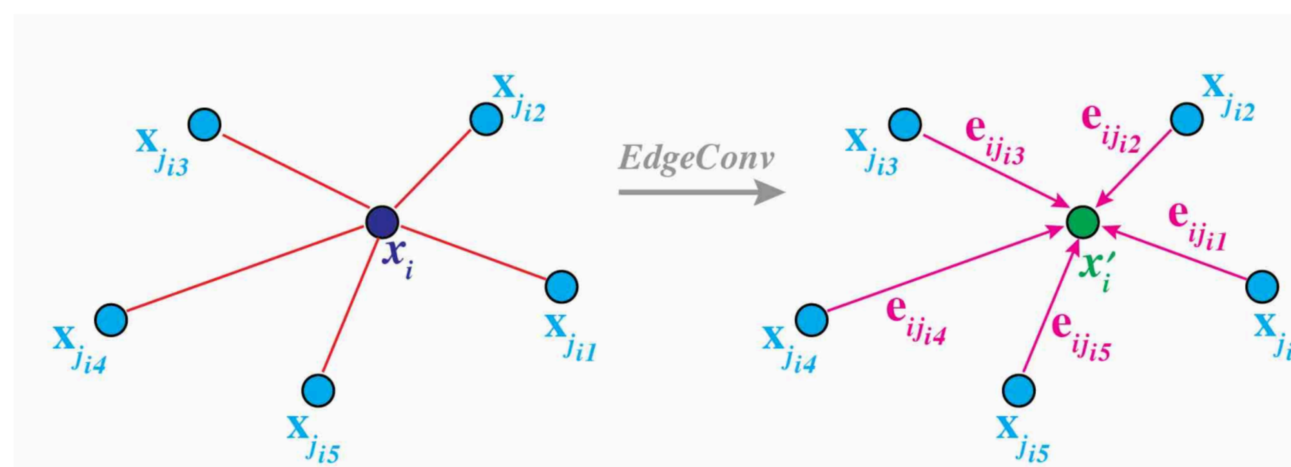
ParticleNet = Graphs



- Images are a convenient representation, but do not capture real structure of our measurements
- Alternative: Graphs
  - Vertex: Particle
  - Edge: Distance (for example in eta-phi space)
- Active development of graphs on CS side, but already HEP applications:
  - Particle Net (best performing top tagger in community study based on EdgeConv)
  - Calorimeter Clustering (1902.07987)
  - Tracking (1810.06111)

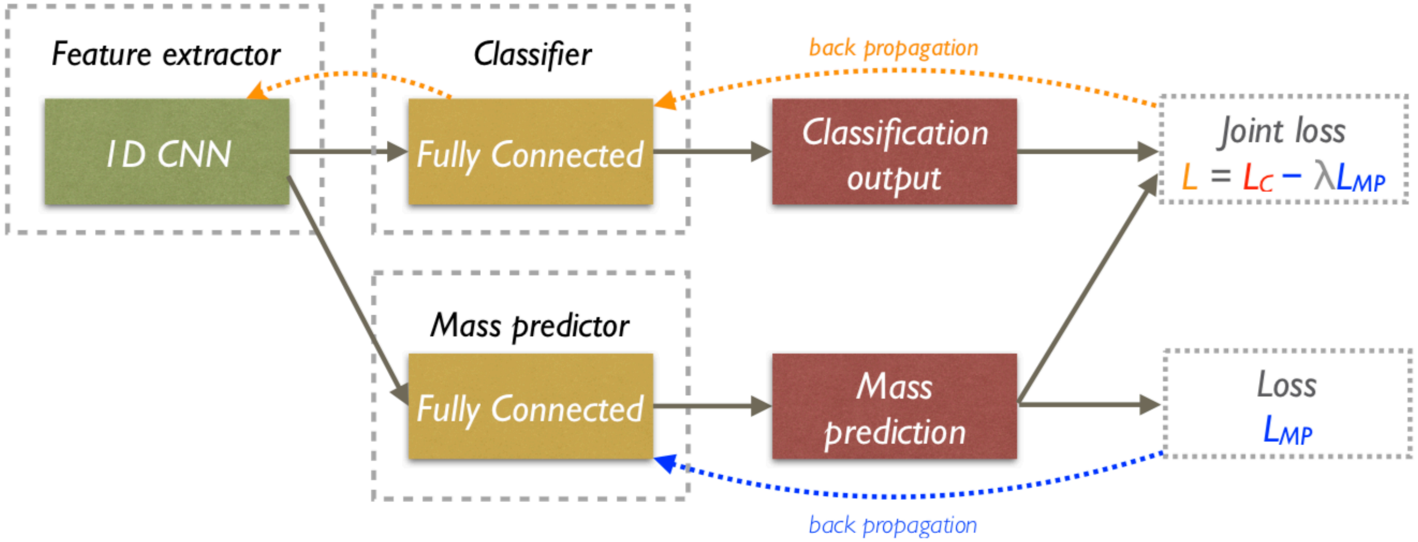




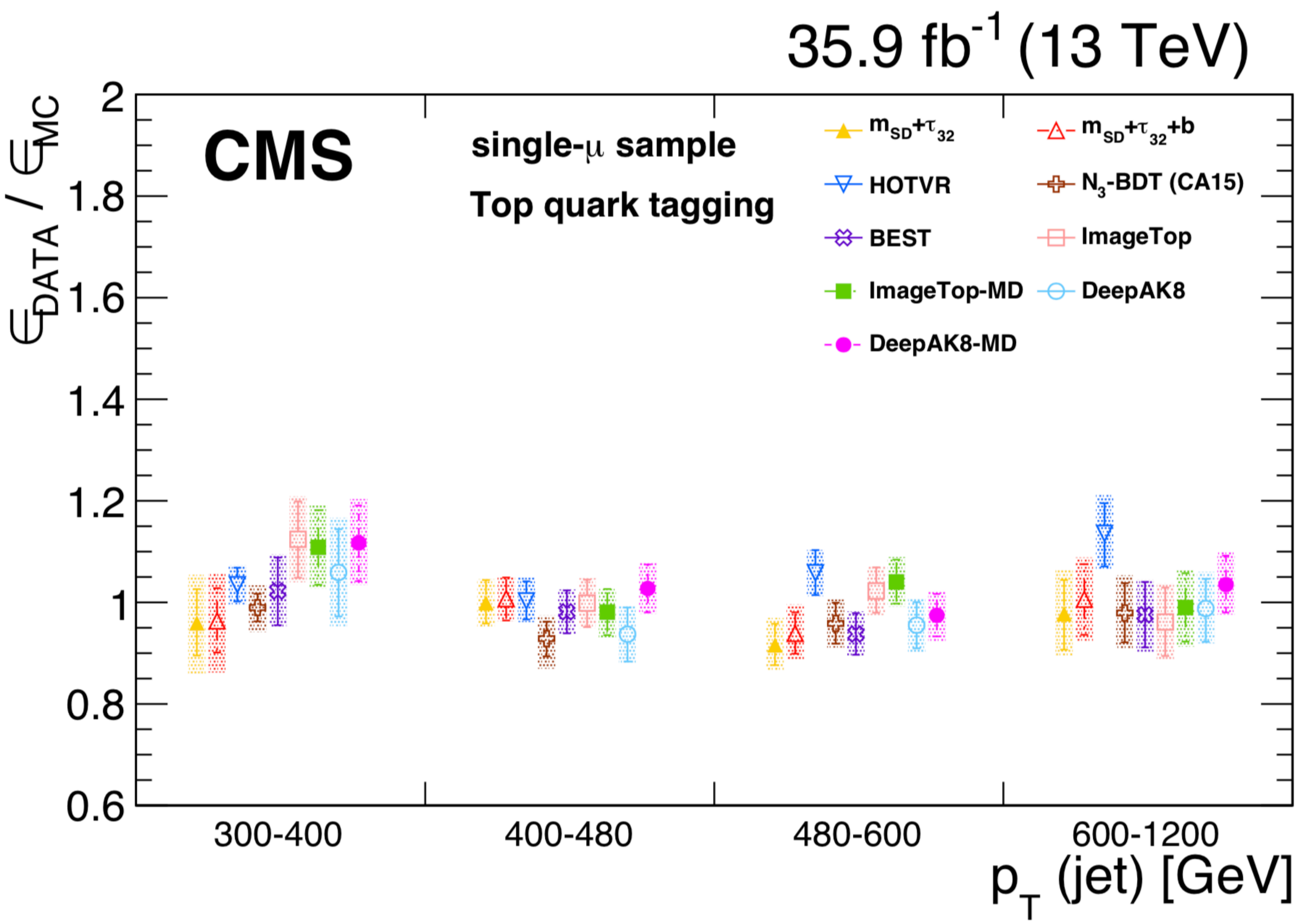


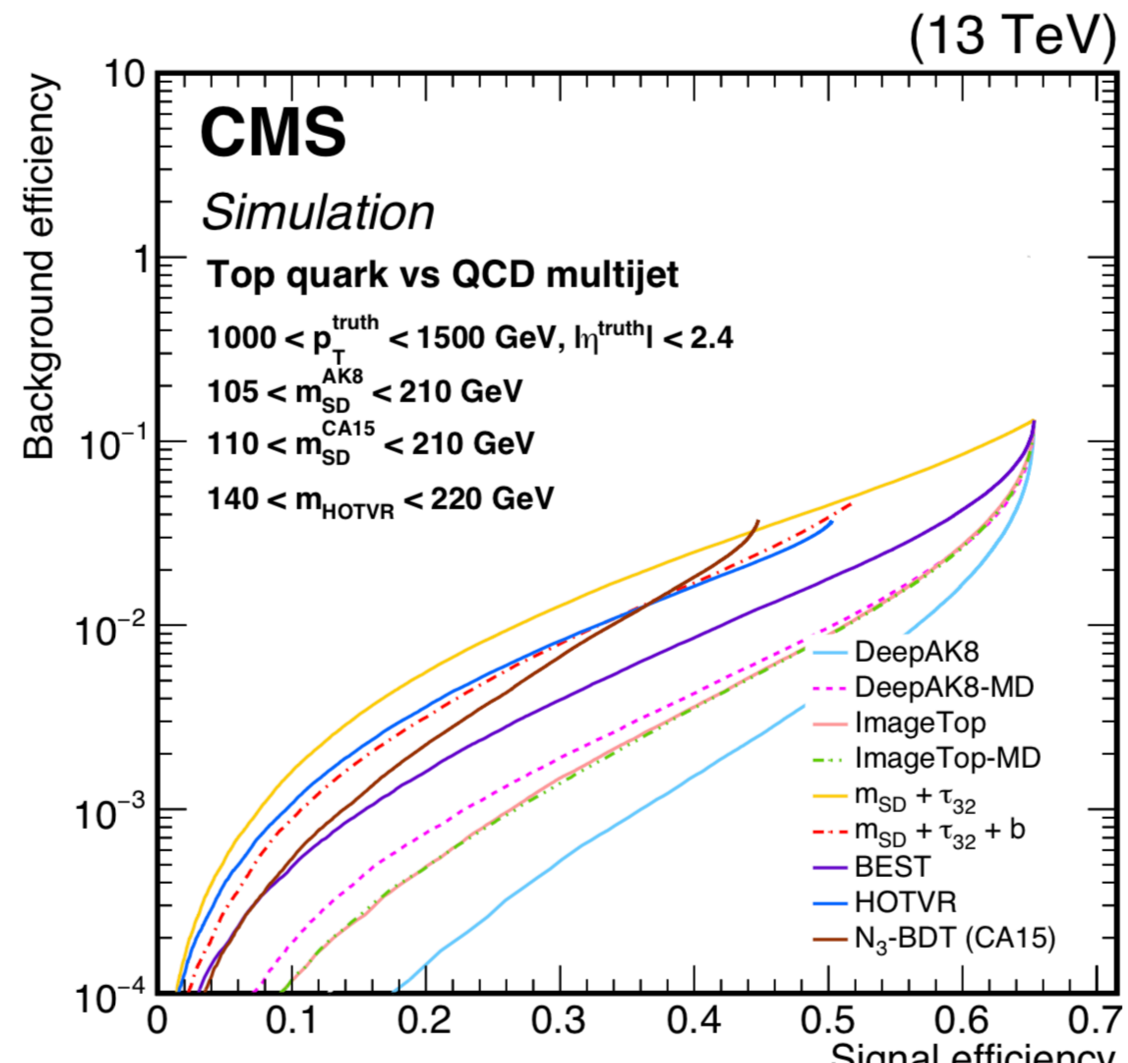
1902.08570

Heavy Resonance Tagging in CMS



Performance on realistic simulation and data in CMS: JME-18-002

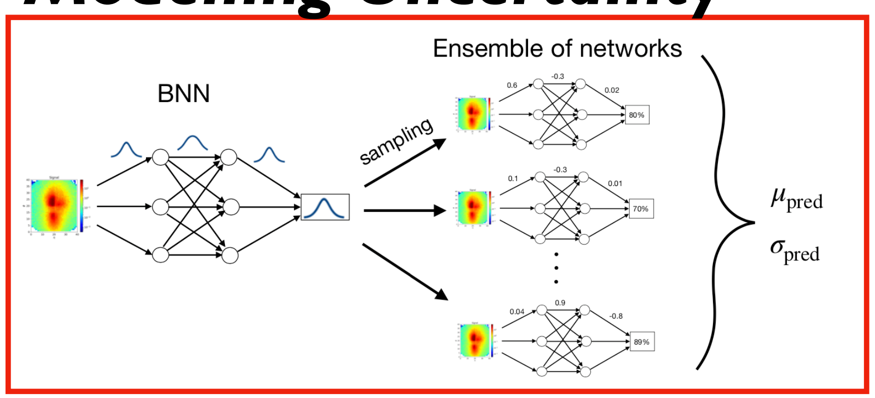




Learning uncertainty

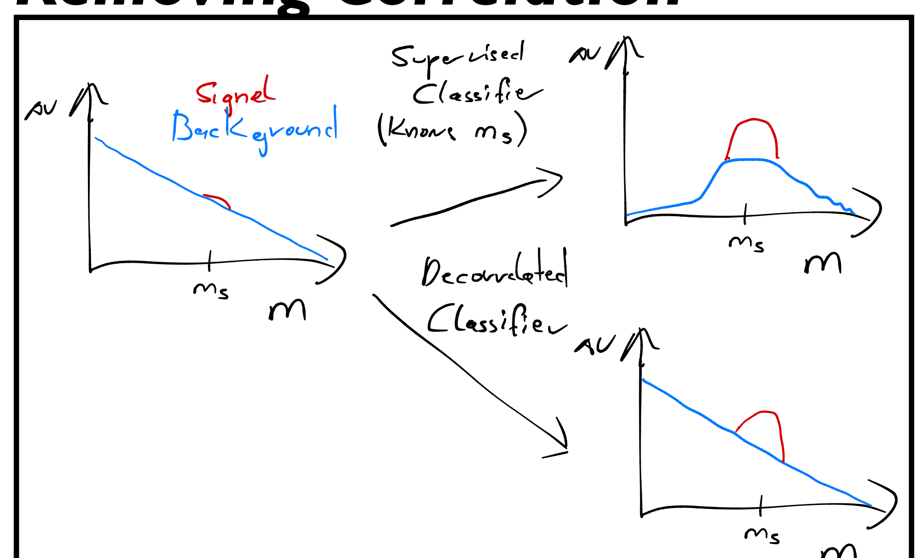
- Stable algorithmic predictions and well-understood uncertainties are key requirements for quantitative science
- Deal with systematic differences between training data (*simulation*) and testing data (*experiment*)

Modelling Uncertainty



1904.10004 & 2003.11099

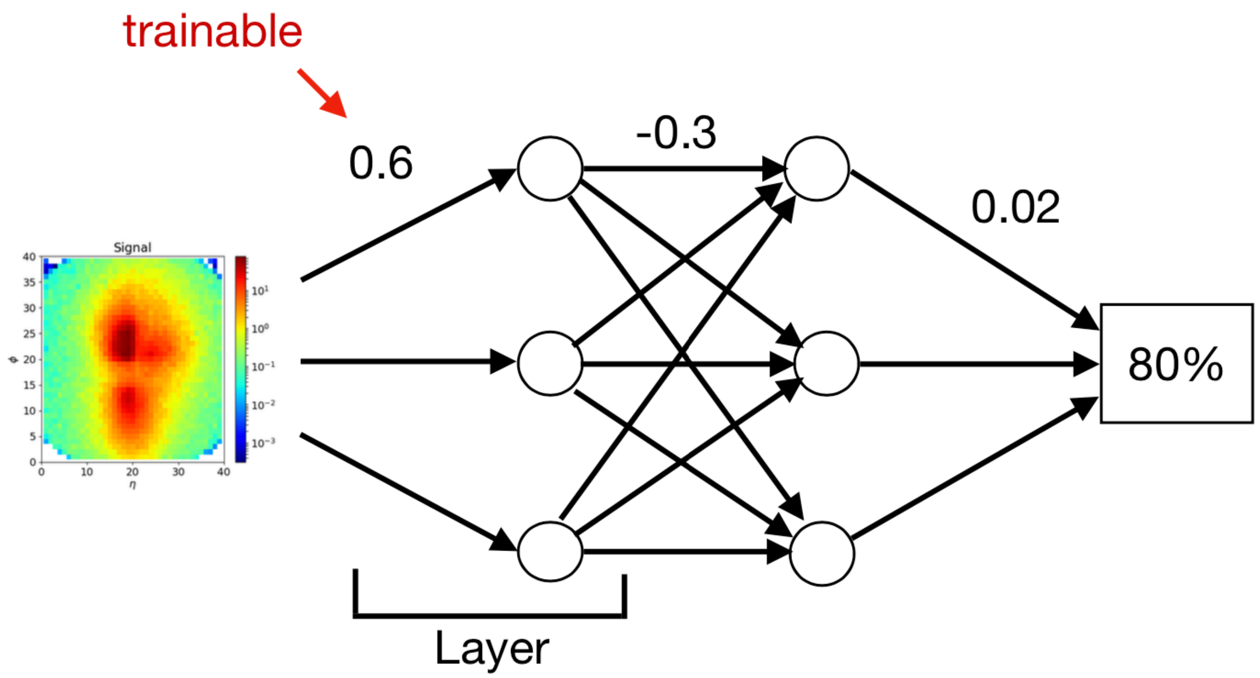
Removing Correlation



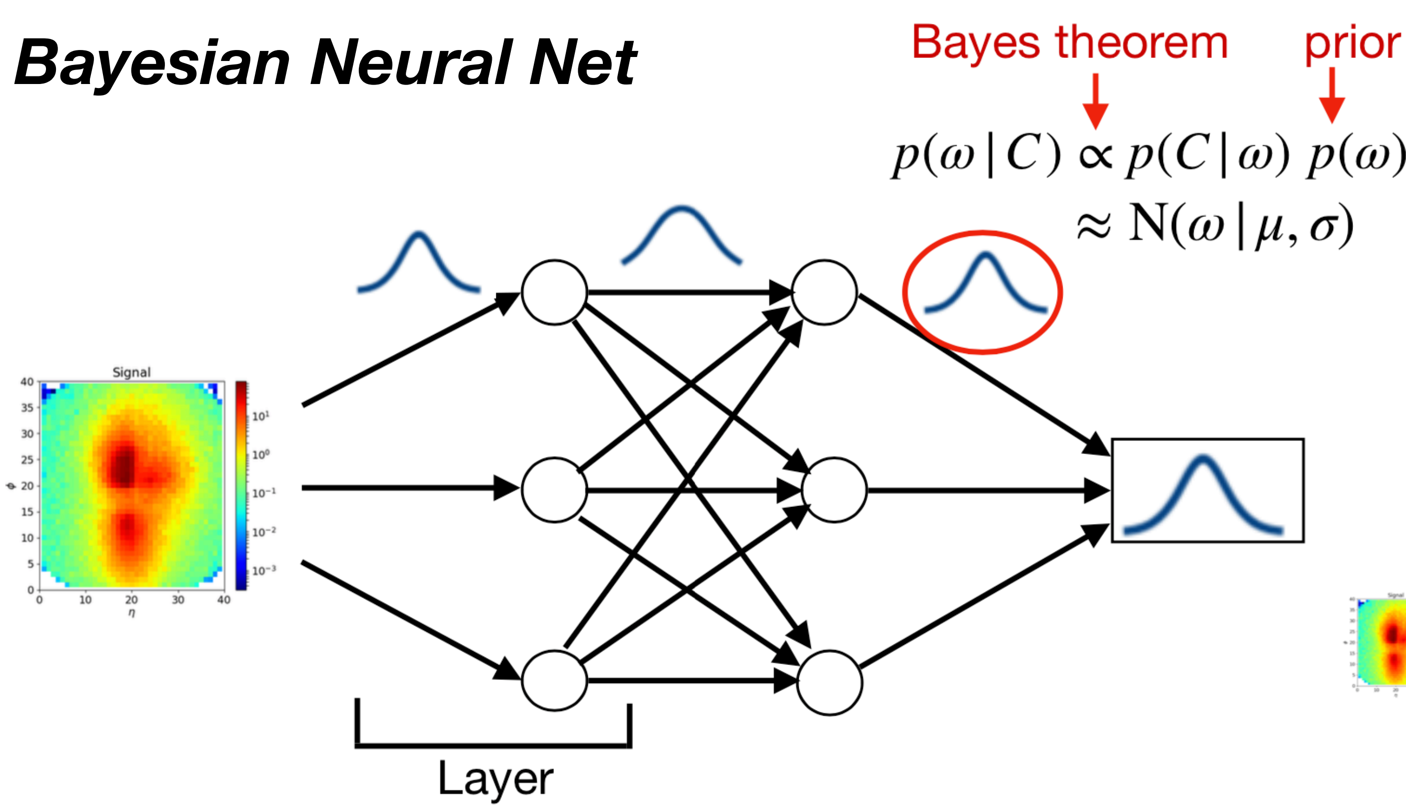
2001.05310

Quantifying Uncertainty

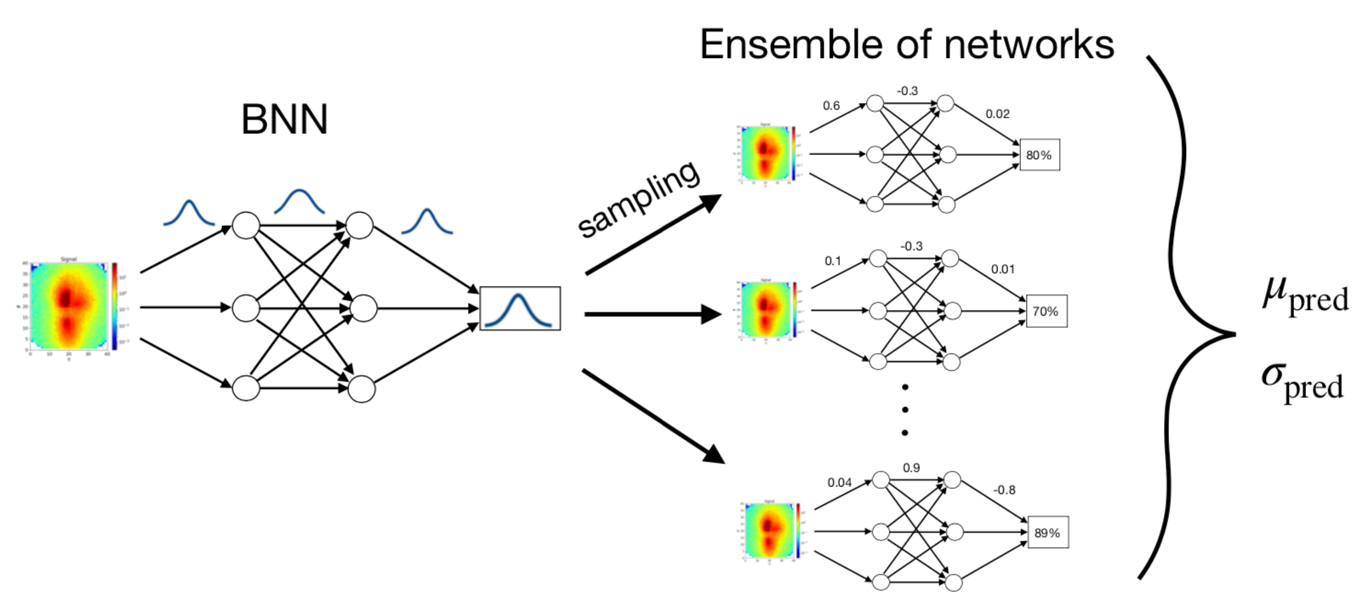
Standard (Deterministic) Neural Net



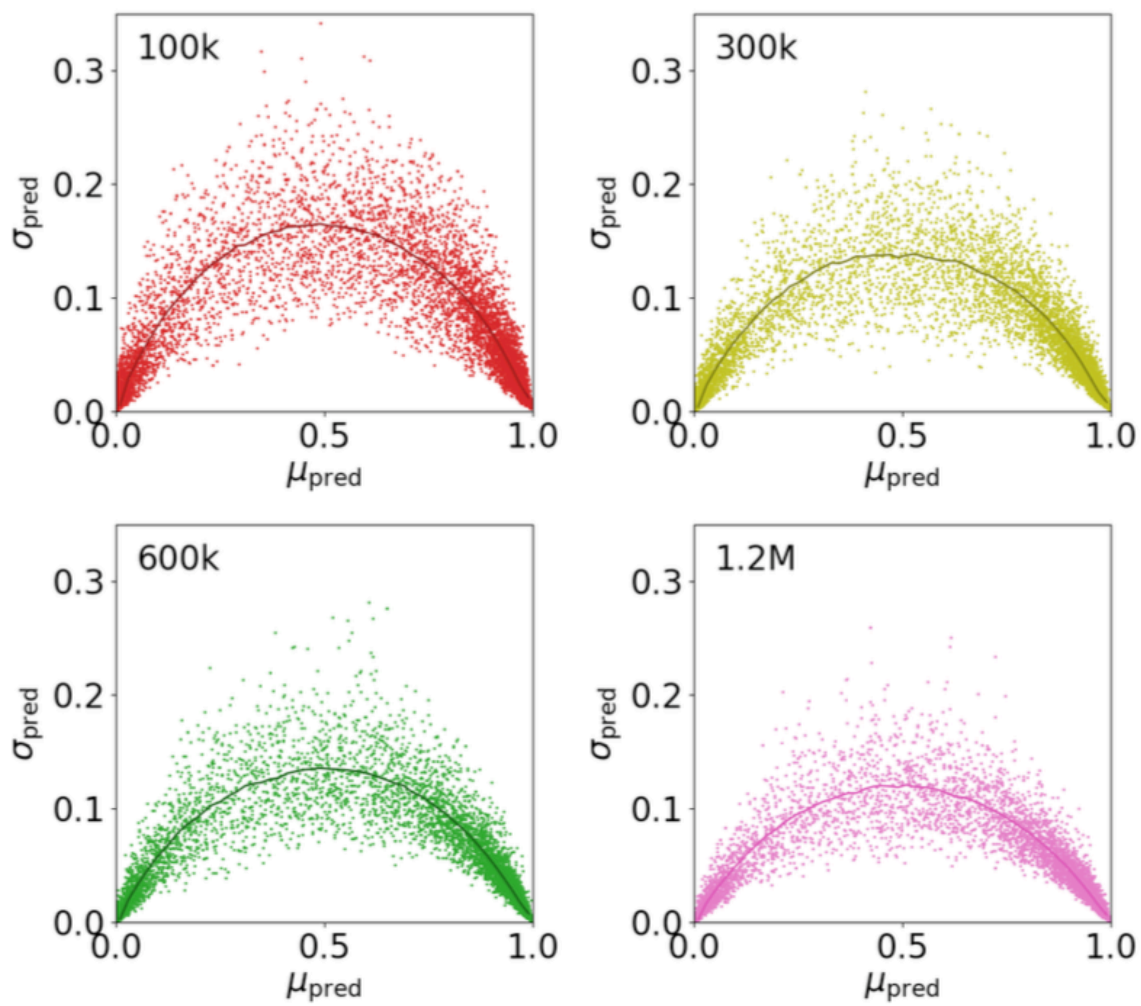
Bayesian Neural Net



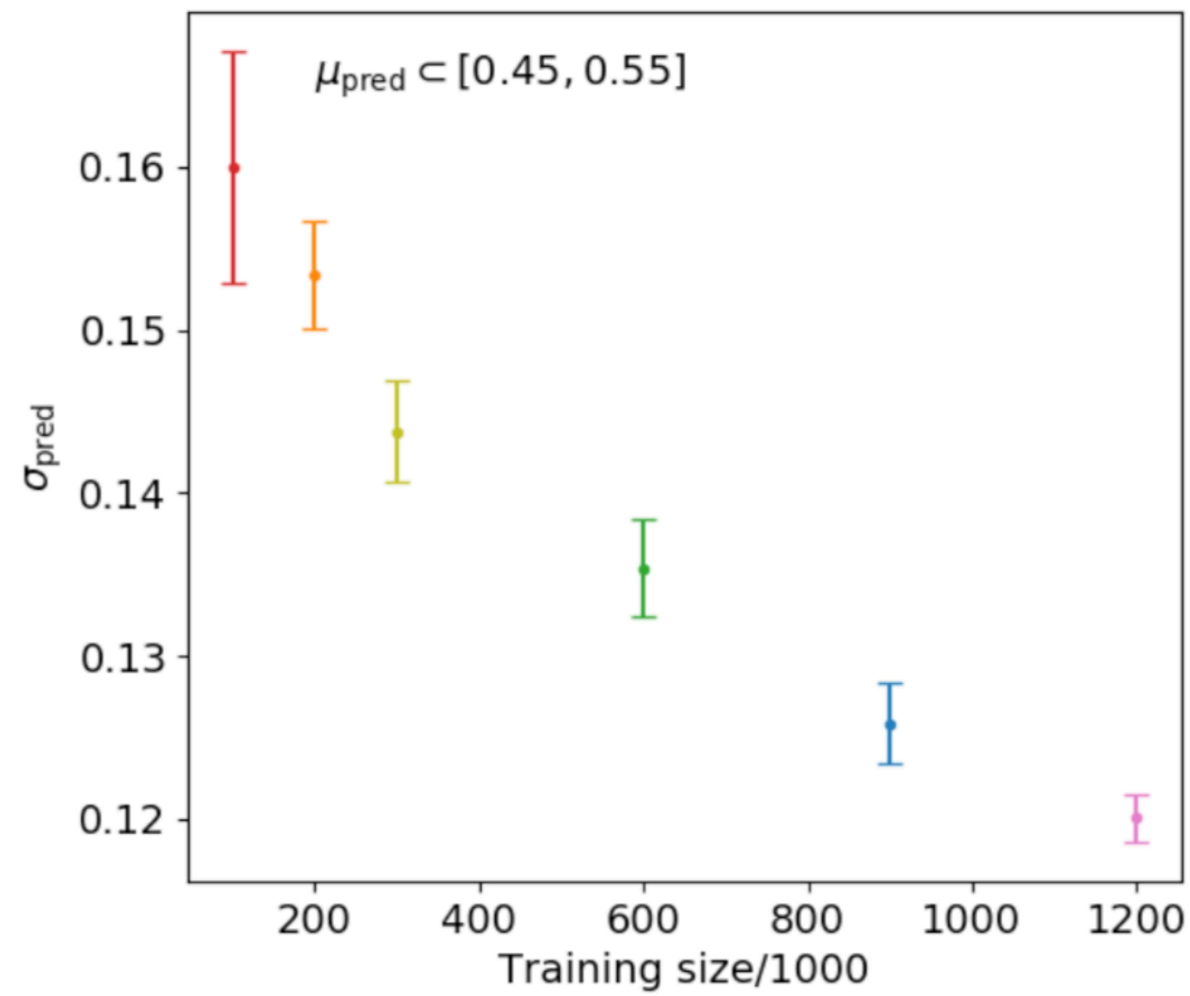
- Provide per-prediction uncertainty on neural network output: Bayesian networks
- Weights replaced by probability distributions
- Prediction vis MC sampling



Statistical Uncertainty



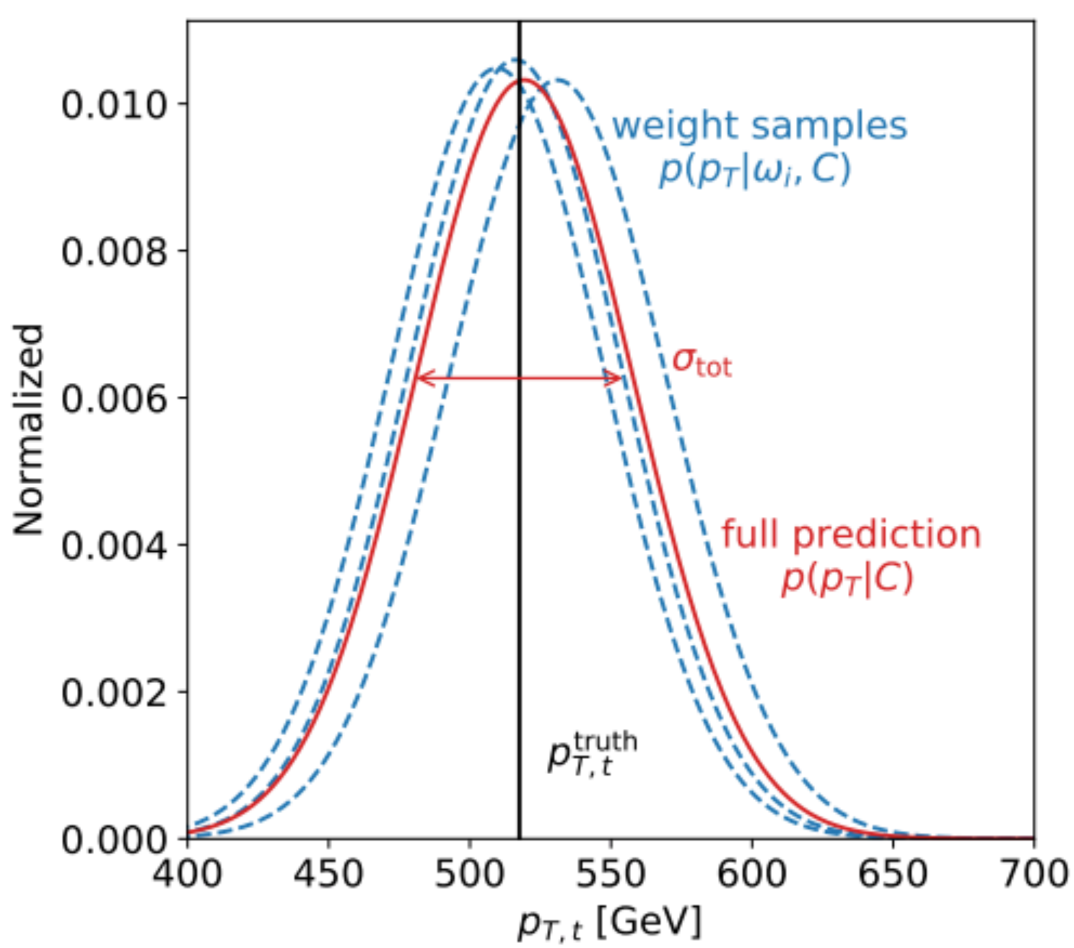
Classification sigmoid correlates mean and standard deviation



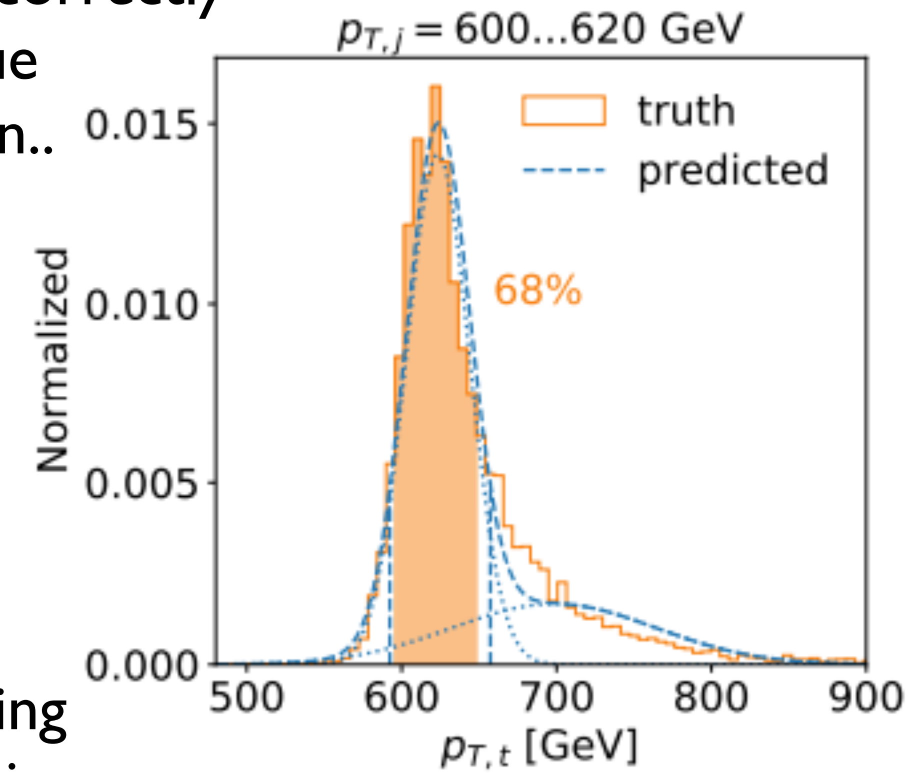
BNN captures effect of finite training data

Systematic Uncertainty

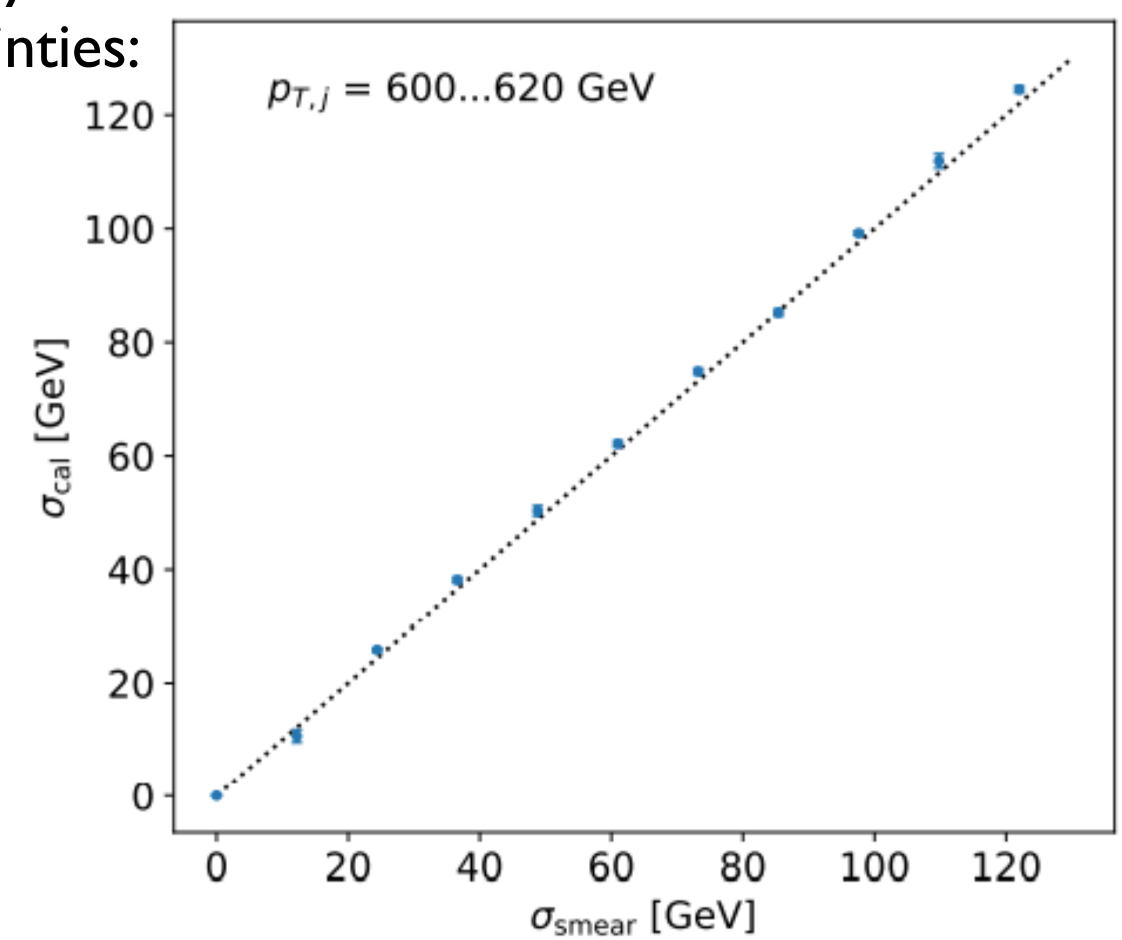
- Look at regression task *(predict top quark momentum from measured calorimeter energies)*
- Each sampling predicts one Gauss or multi-Gauss distribution
- Average to get network prediction



Learns to correctly predict true distribution..

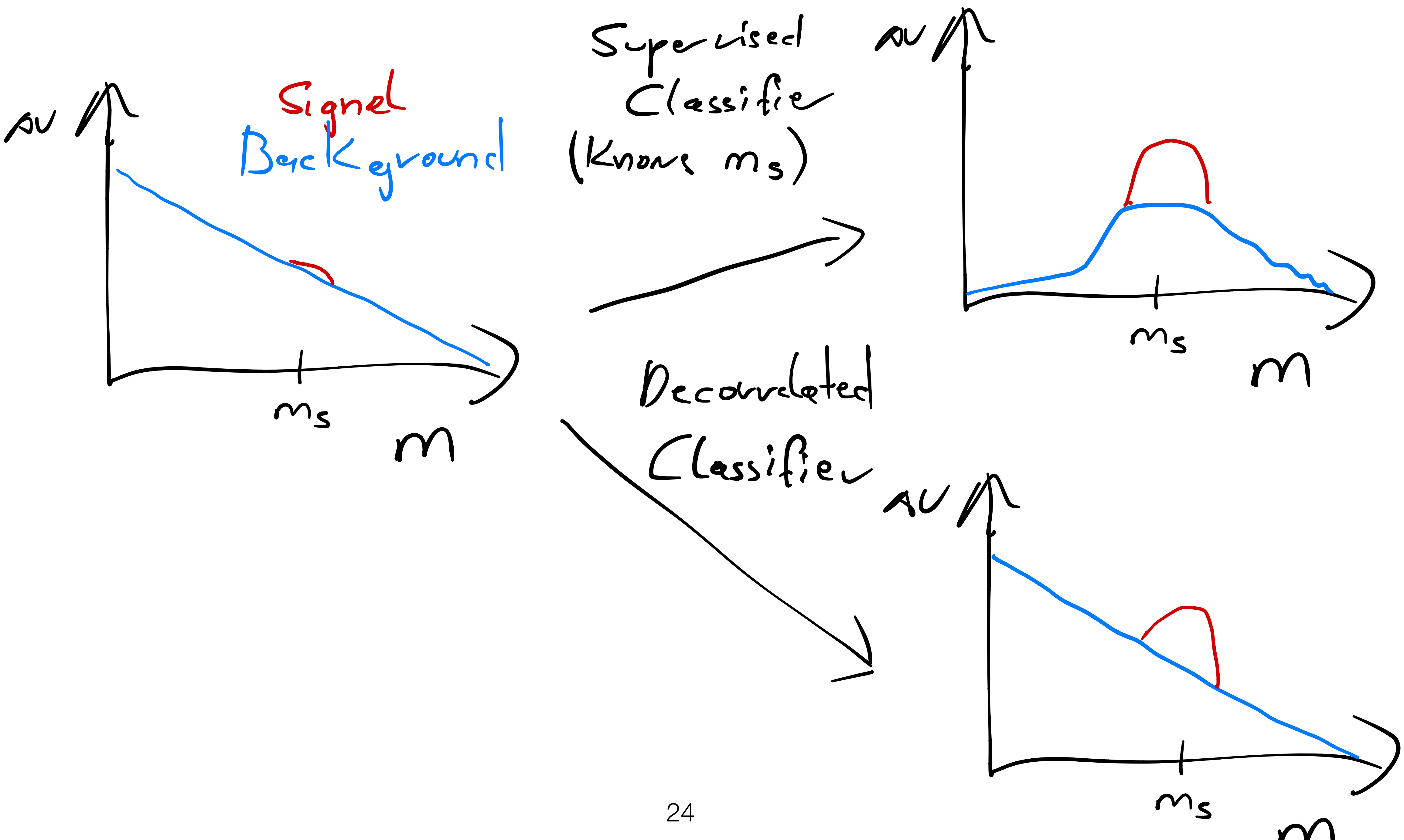


Including smearing due to systematic uncertainties:



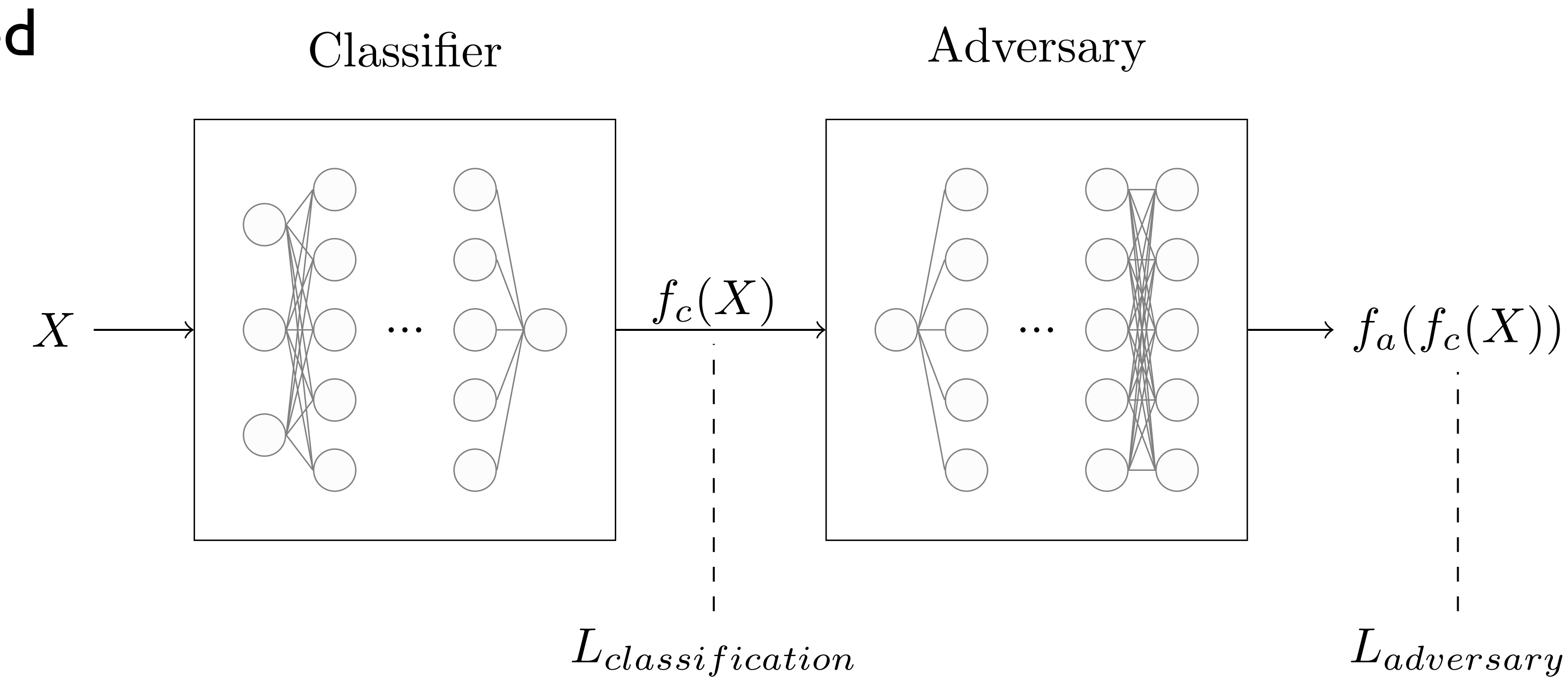
Decorrelation

- Reduce impact of other variables on analysis result
- Either remove correlation of classifier output with a systematic uncertainty or another variable

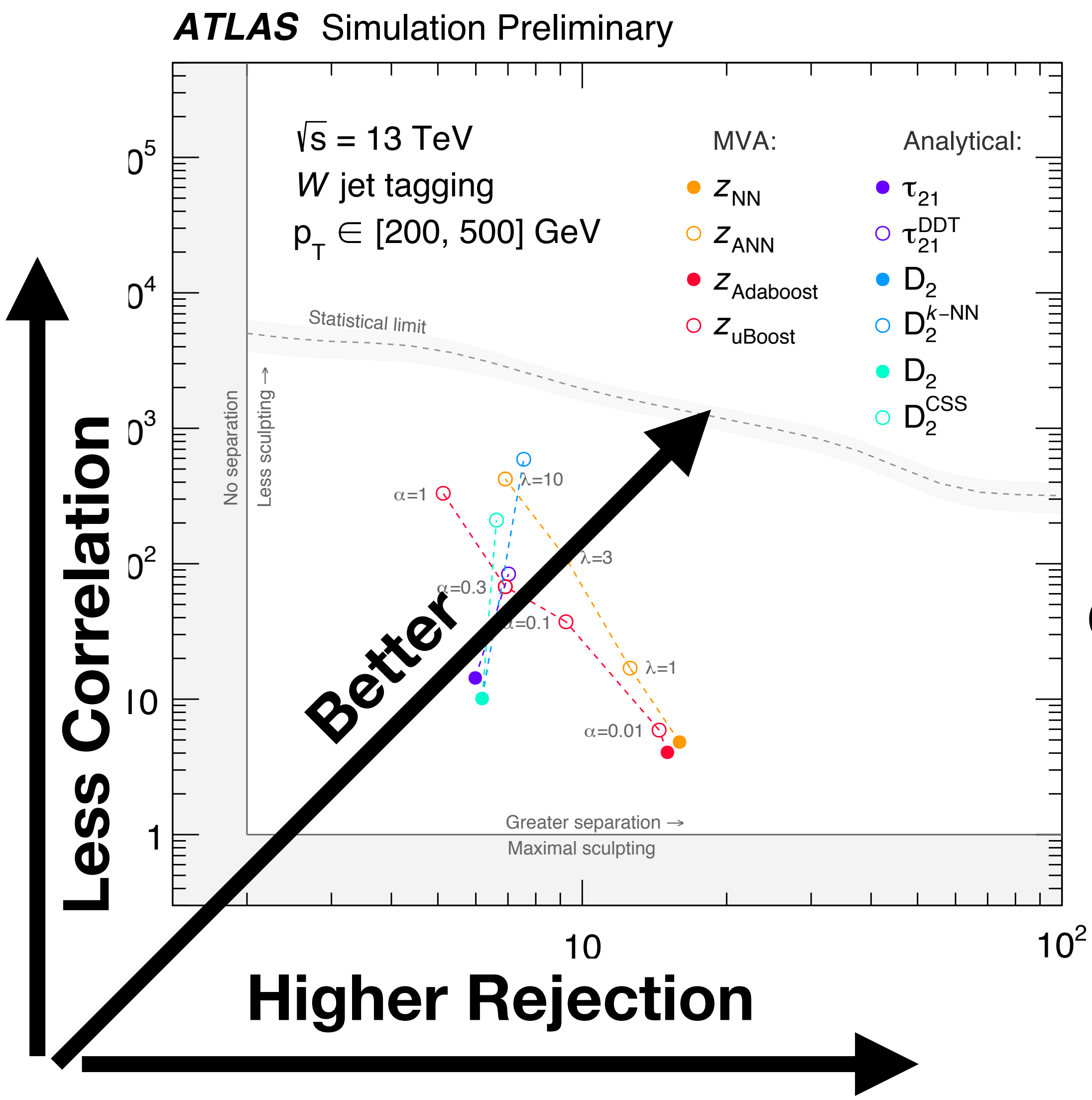


Approaches

- **Obscurity**:
  - Do not give mass [will be using this as stand-in for any variable we want to decor relate agains] as input
  - Simple, does not work
- **Data planing** (1709.10106, 1908.08959):
  - Reweight input distributions to be flat
  - Simple, limited power
- Designing Decorrelated Taggers - **DDT** (1603.00027):
  - Linearly transform output to be stable for one working point by subtracting for each bin
- Add **KL/JS divergence** to loss
  - Promising idea, but only works for one working point. Binning needed.
- Use complex **adversarial ML** (1611.01046, 1703.03507)
  - Powerful, hard to tune
  - Basic idea: If adversary can infer mass from classifier output, the output is not decorrelated

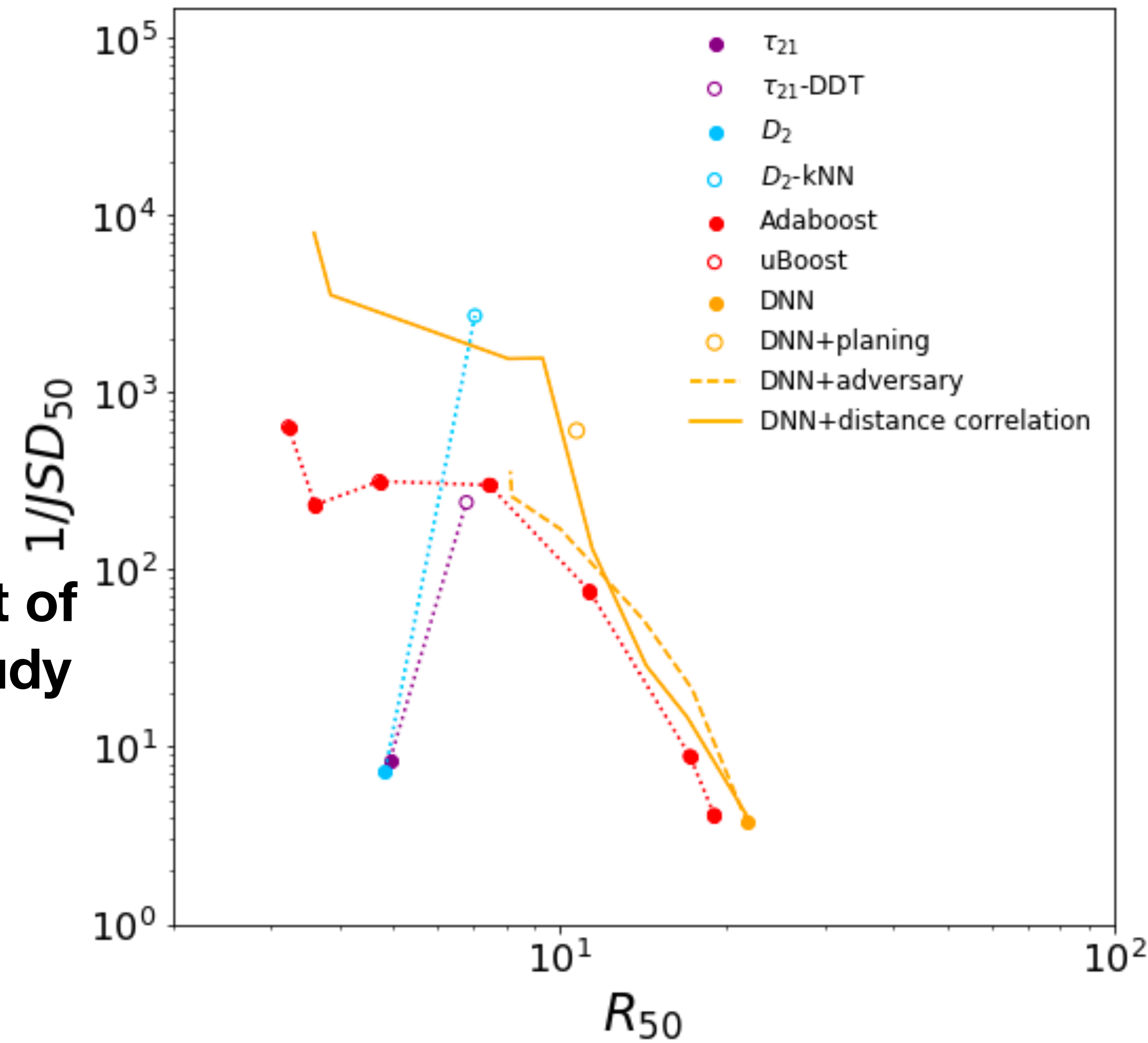


Comparison



ATL-PHYS-PUB-2018-014

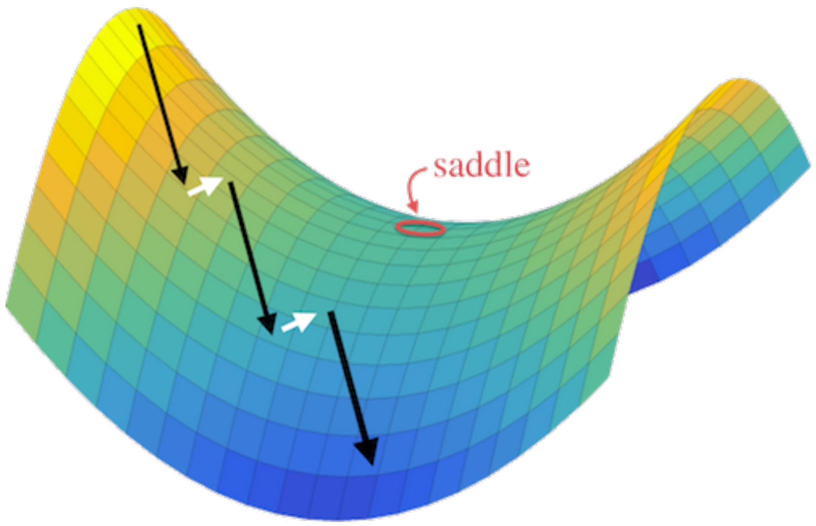
Our recast of ATLAS study



Problem

- Adversarial training is inherently unstable (hard to set up and sensitive to hyper parameter changes)
  - Looking for a saddle point

$$\min_{\theta_{\mathrm{clf}}} \max_{\theta_{\mathrm{adv}}} L_{\mathrm{clf}}(y(\theta_{\mathrm{clf}})) - \lambda L_{\mathrm{adv}}(y(\theta_{\mathrm{clf}}), m; \theta_{\mathrm{adv}})$$



- Find a regulariser term that fulfils the same goal but allows simple training to convergence

$$\min_{\theta_{\mathrm{clf}}} L_{\mathrm{clf}}(y(\theta_{\mathrm{clf}})) + \lambda C_{\mathrm{reg}}(y(\theta_{\mathrm{clf}}), m)$$

- Use distance correlation

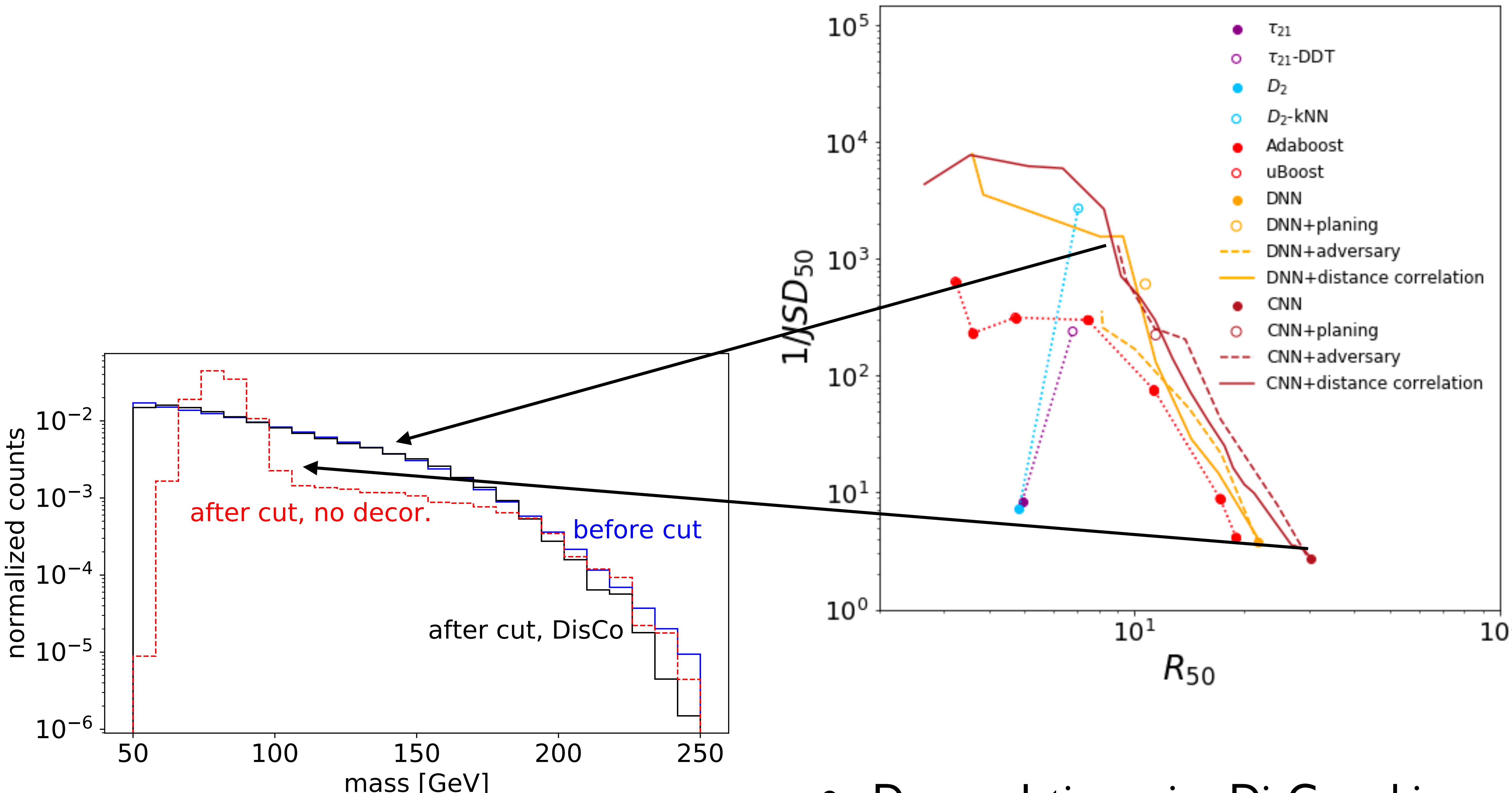
$$x_{jk} = |X_j - X_k|$$ Distances of all examples in batch for classifier output

$$y_{jk} = |Y_j - Y_k|$$ ... for variable to decorrelate

$$\hat{x}_{jk} = x_{jk} - \overline{x}_{j\cdot} - \overline{x}_{\cdot k} + \overline{x}_{\cdot\cdot}$$
$$\hat{y}_{jk} = y_{jk} - \overline{y}_{j\cdot} - \overline{y}_{\cdot k} + \overline{y}_{\cdot\cdot}$$ Center distributions

$$\mathrm{dCov}^2 = \frac{1}{n} \sum_j \sum_k \hat{x}_{jk} \hat{y}_{jk}$$ And calculate average product per batch

Comparison



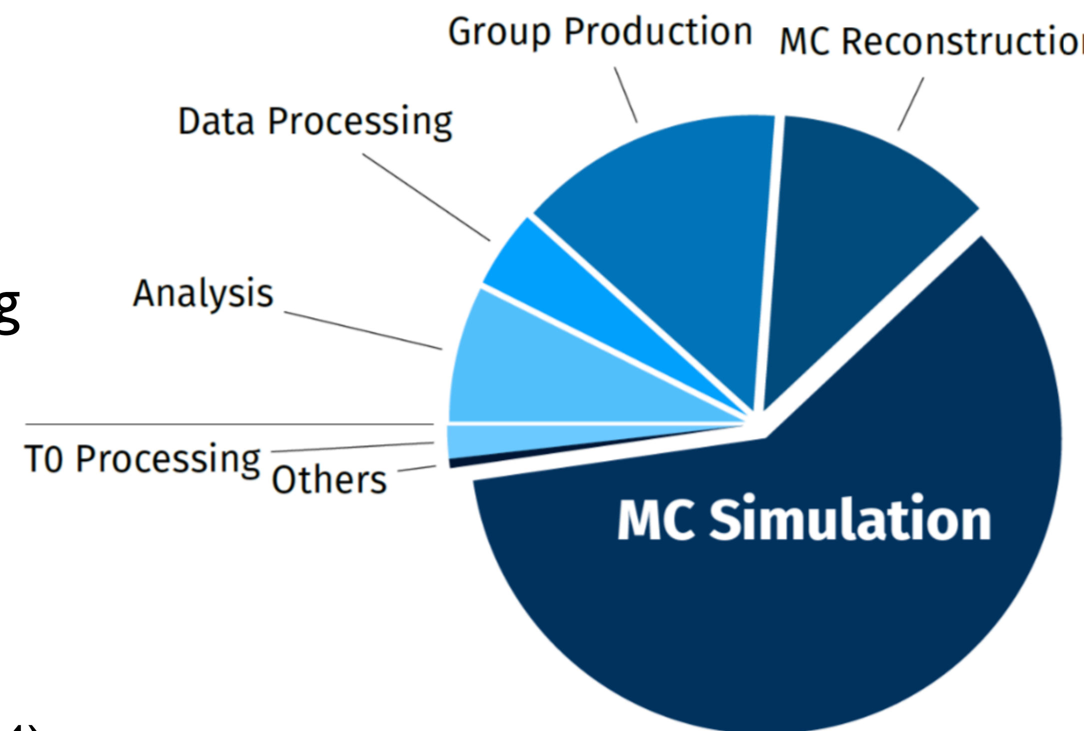
- Decorrelation using DisCo achieves same performance as adversarial method, easier to train

Fast Simulation

Fast Simulation / Generation

Problem: Spend large fraction of our computing resources on event generation and detector simulation

Potential Solution: Use generative machine learning models to accelerate simulation (1705.02355, 1807.01954, 1912.06794, 2005.05334)



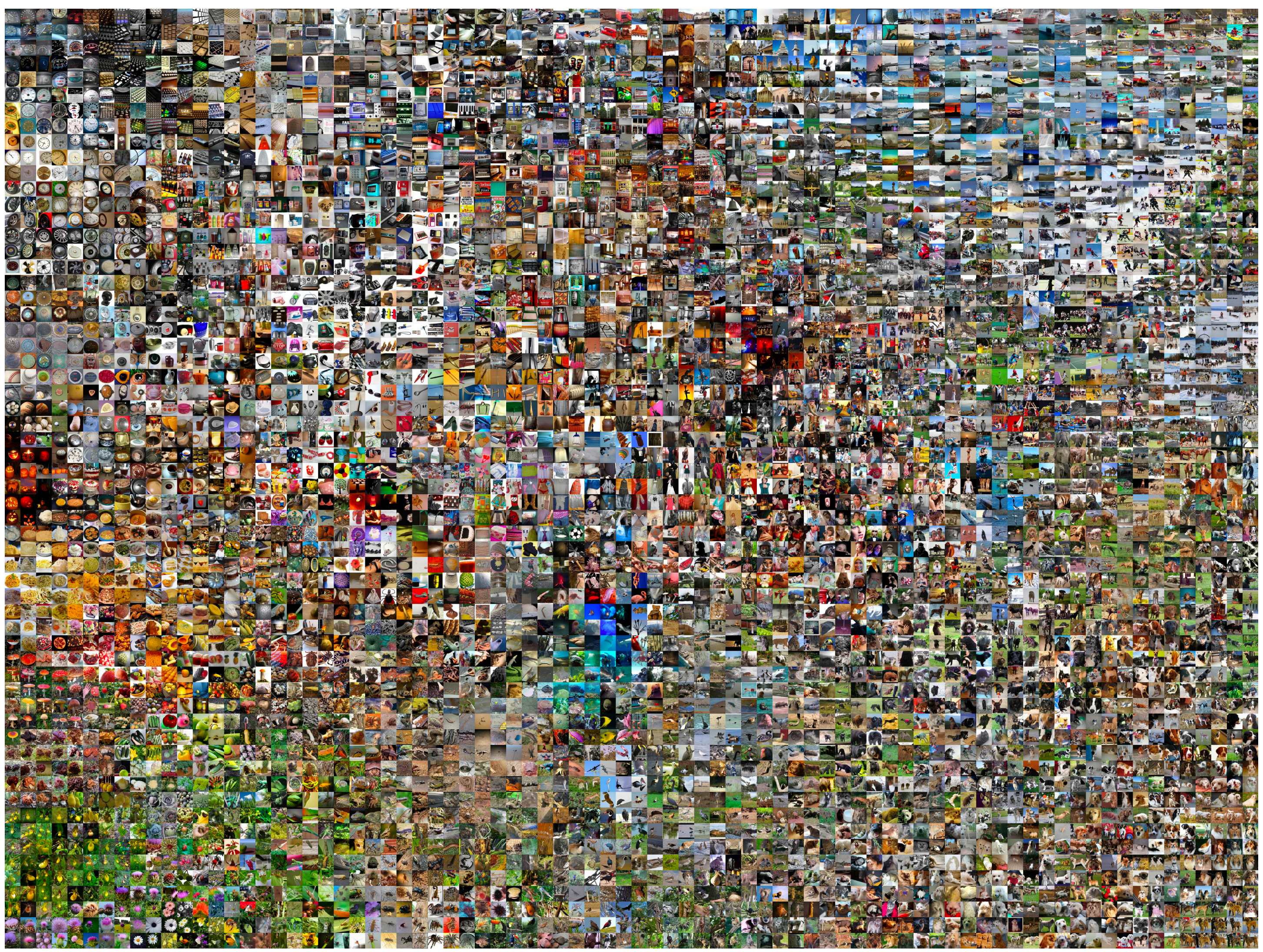
Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed

Erik Buhmann · Sascha Diefenbacher · Engin Eren · Frank Gaede · Gregor Kasieczka · Anatolii Korol · Katja Krüger

(out since yesterday!)

Generators

We **have**:
many images
(or collision events,
or detector readouts, …)



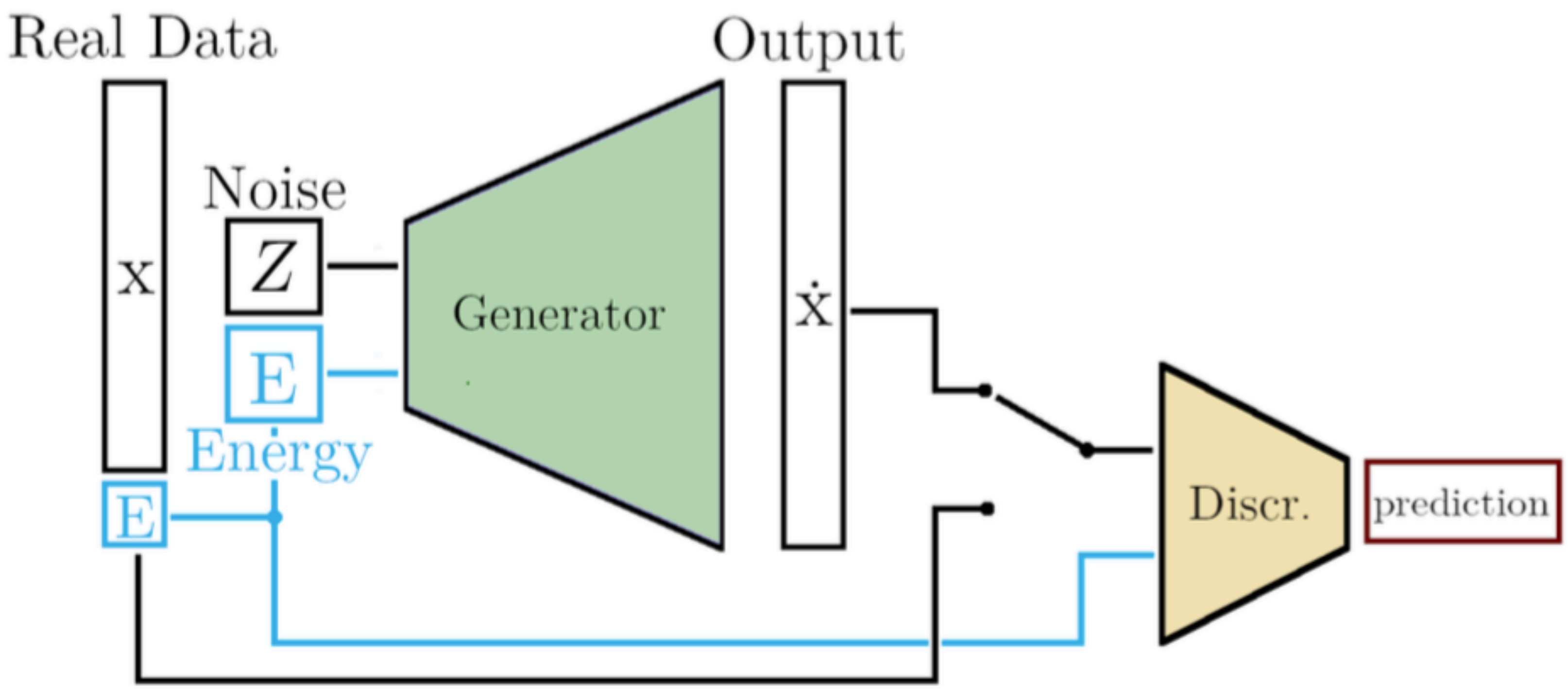
wired.com

We **want**: more images.

(Specifically: New examples that are similar to the examples, but not exact copies)

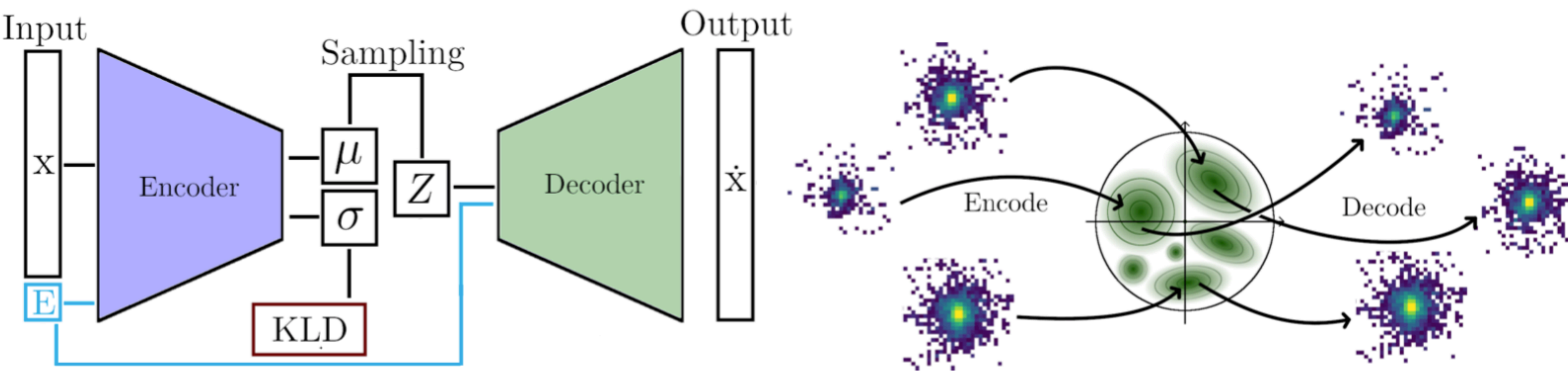
How to encode in neural net?

GAN



- **G**enerative **A**dversarial **N**etwork
- Generator generates new fake images from noise
- Second network (discriminator) learns to distinguish fake from real images
- Training via mutual feedback

VAE



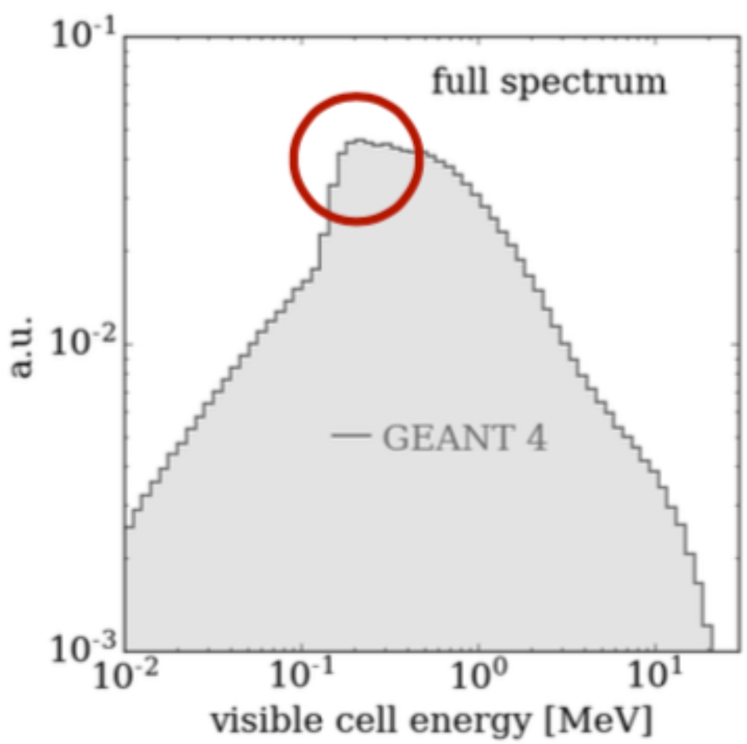
- **V**ariational **A**uto**e**ncoder
- Encode examples into latent space of network
- Sample from latent space to produce new examples

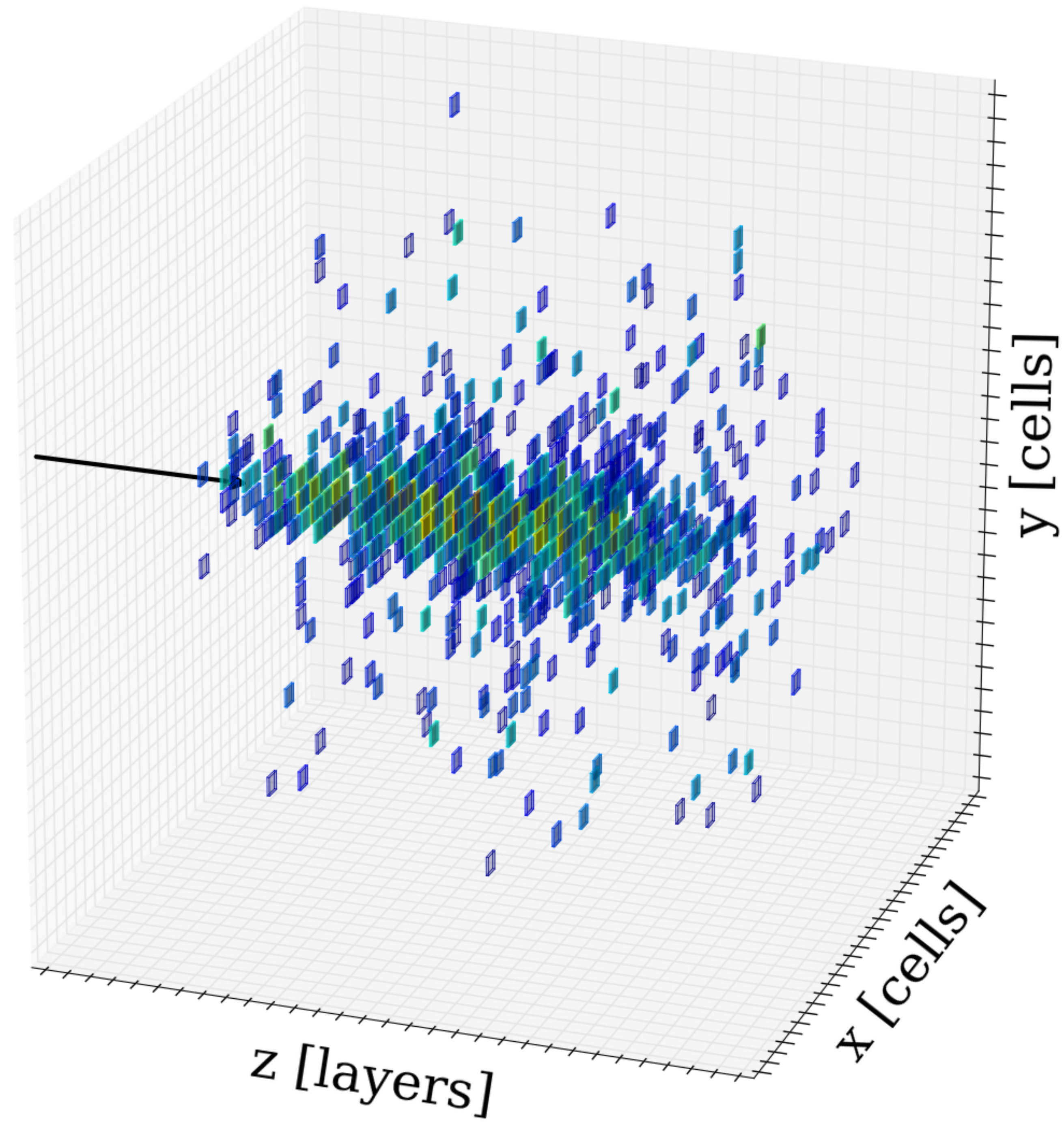
https://thispersondoesnotexist.com/

Concrete Problem

Describe photon showers in high granularity calorimeter prototype

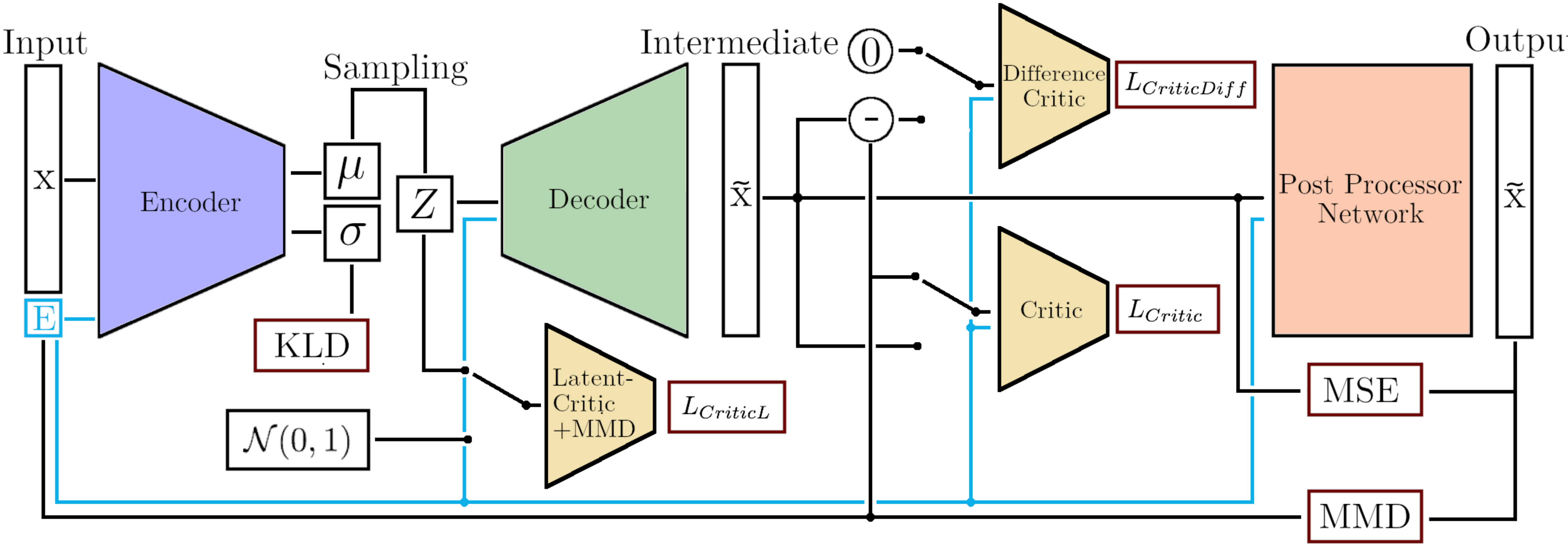
- 30x30x30 cells (Si-W)
- Photon energies from 10 to 100 GeV
- Use 950k examples (uniform in energy) created with GEANT4 to train





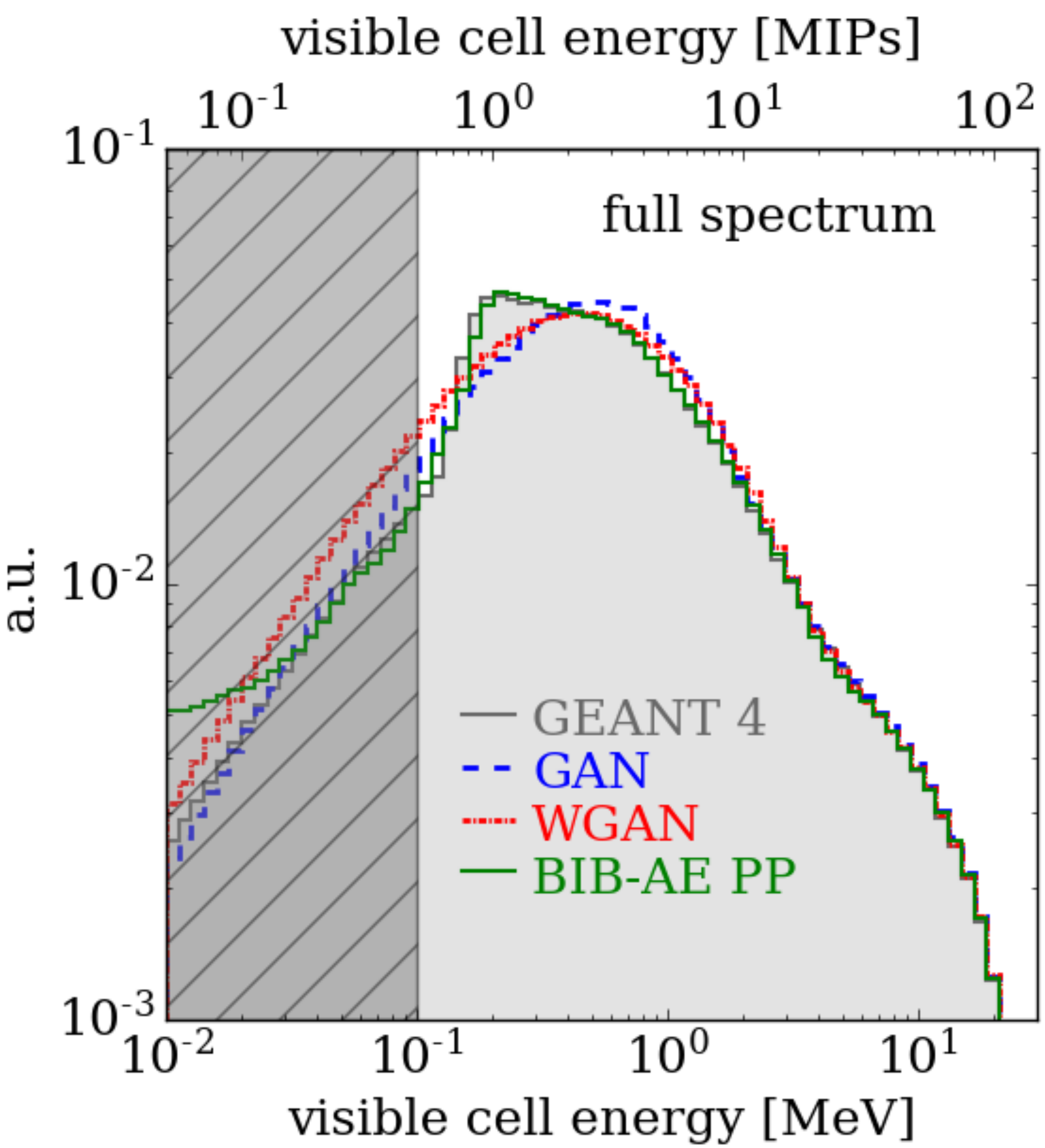
- Not only model individual images but also **differential distributions**

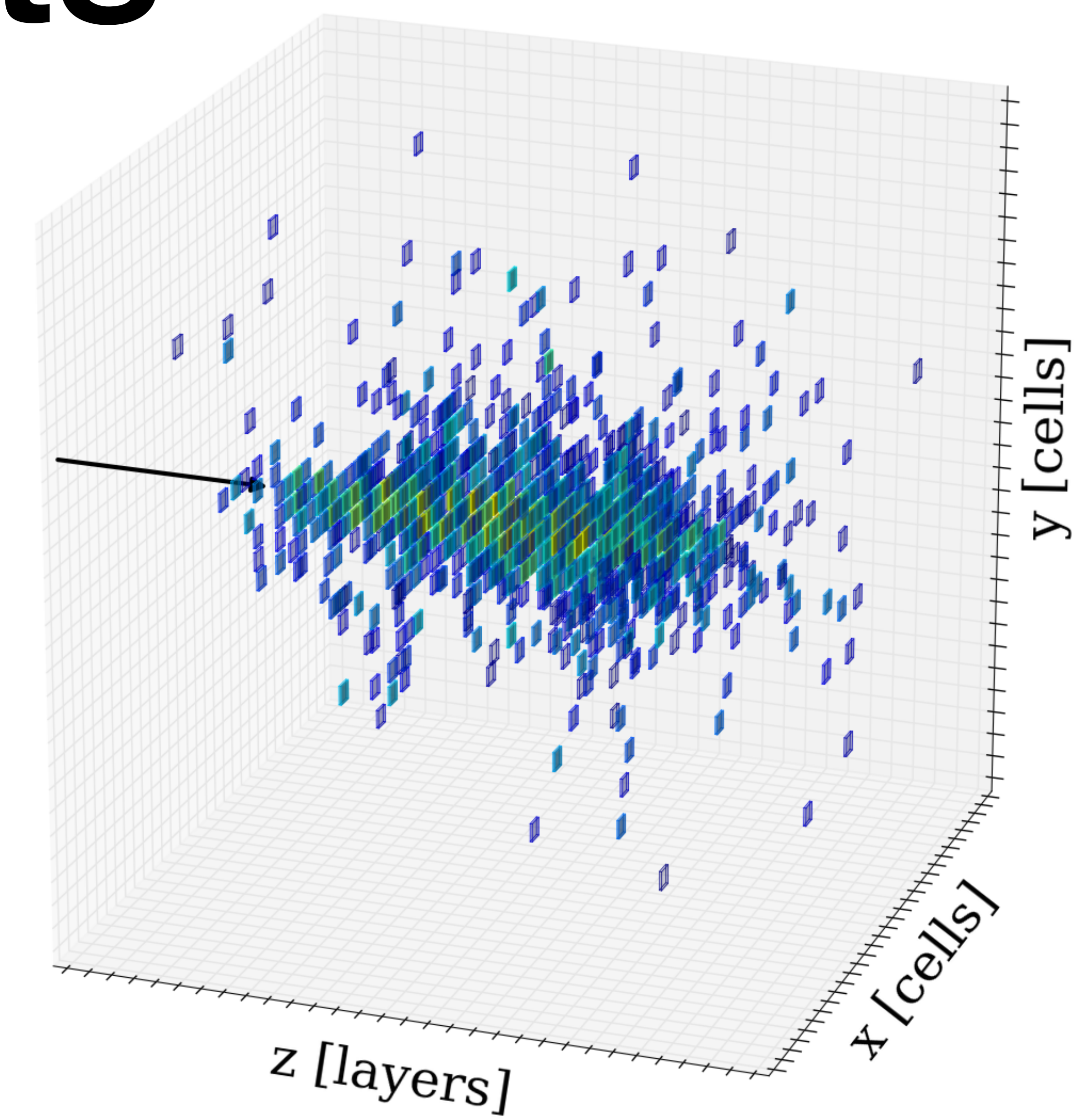
Architecture



- BIB-AE (based on 1912.00830) with added post-processing
- Unifies features of GAN and VAE
- 71M trainable parameters

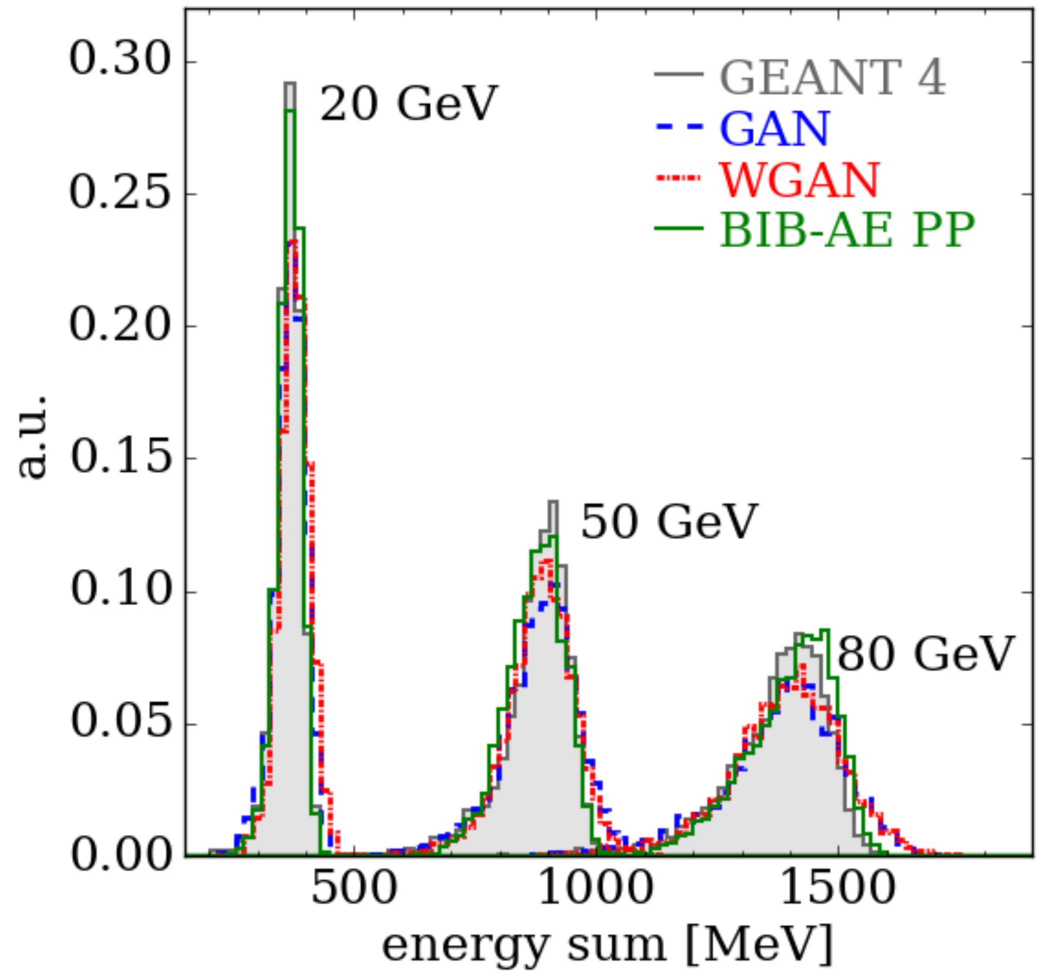
Results



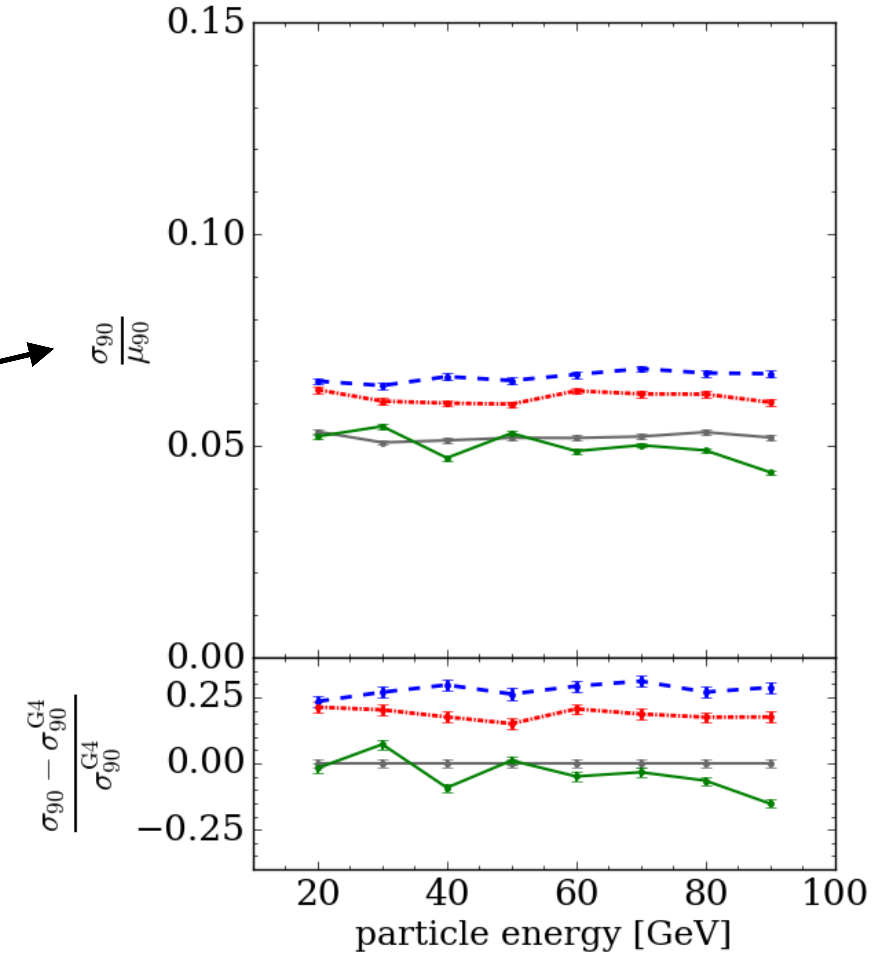


Correctly describe distribution down to 0.2 MIPs

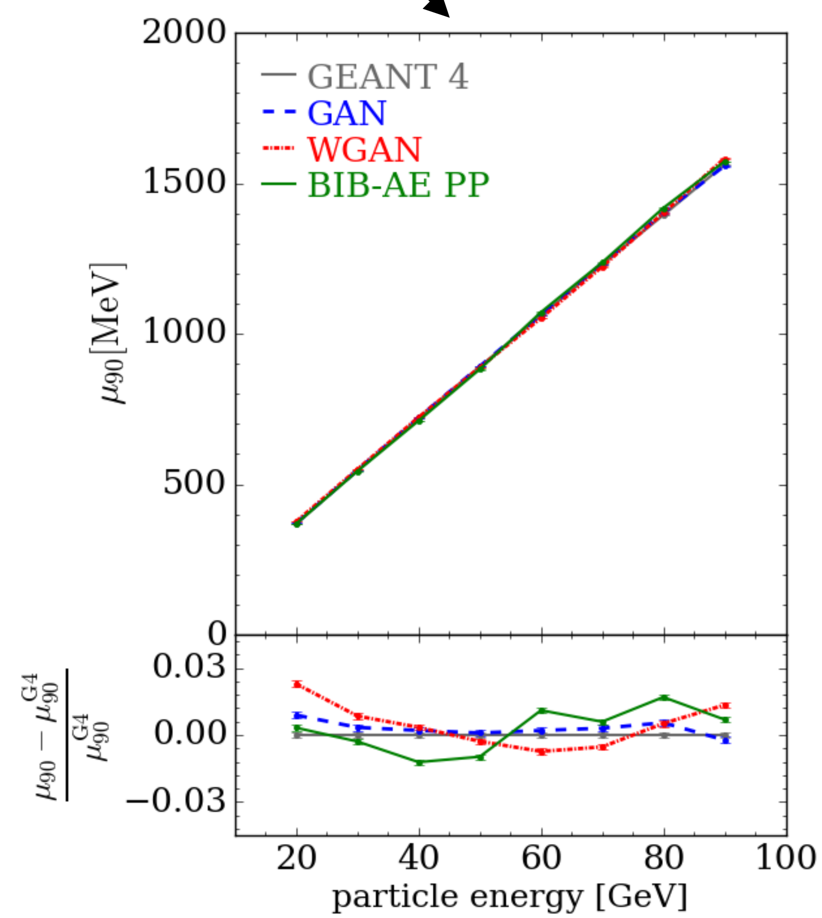
Results



Width



Mean



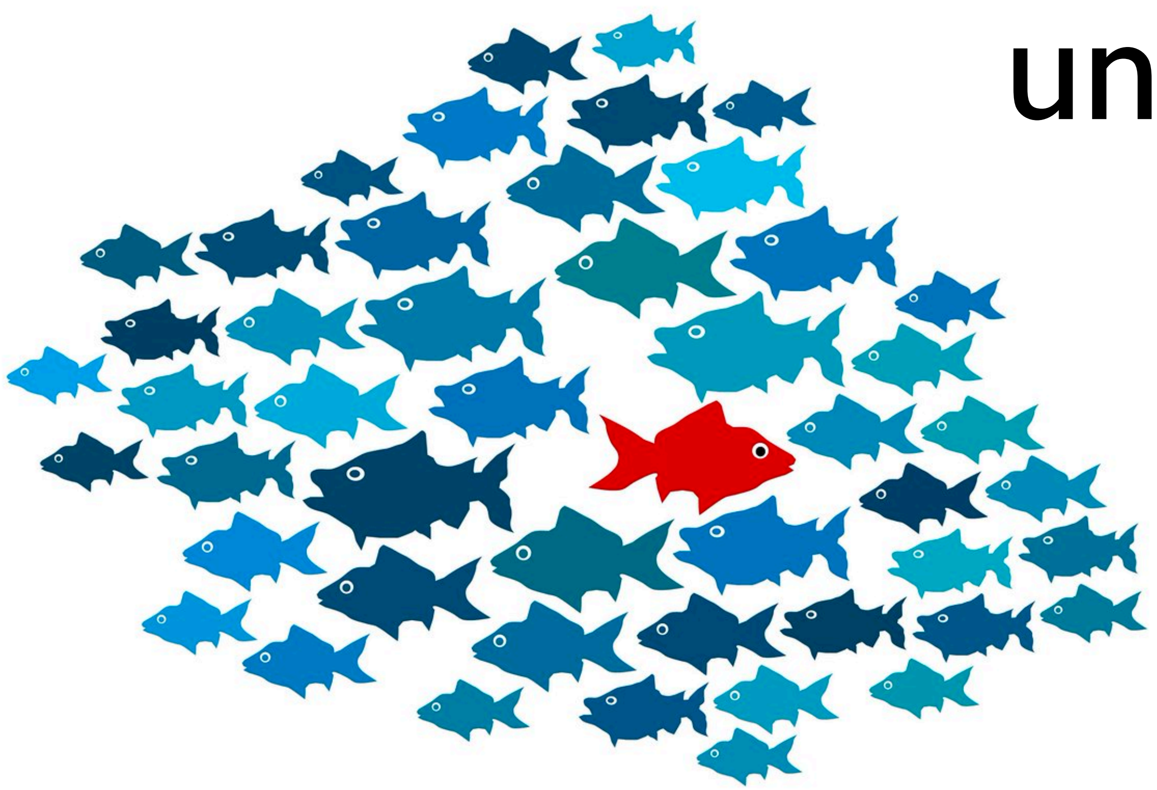
Total energy is also well described.

Next challenge:
Also perfect correlations, more varied inputs

Anomaly Searches

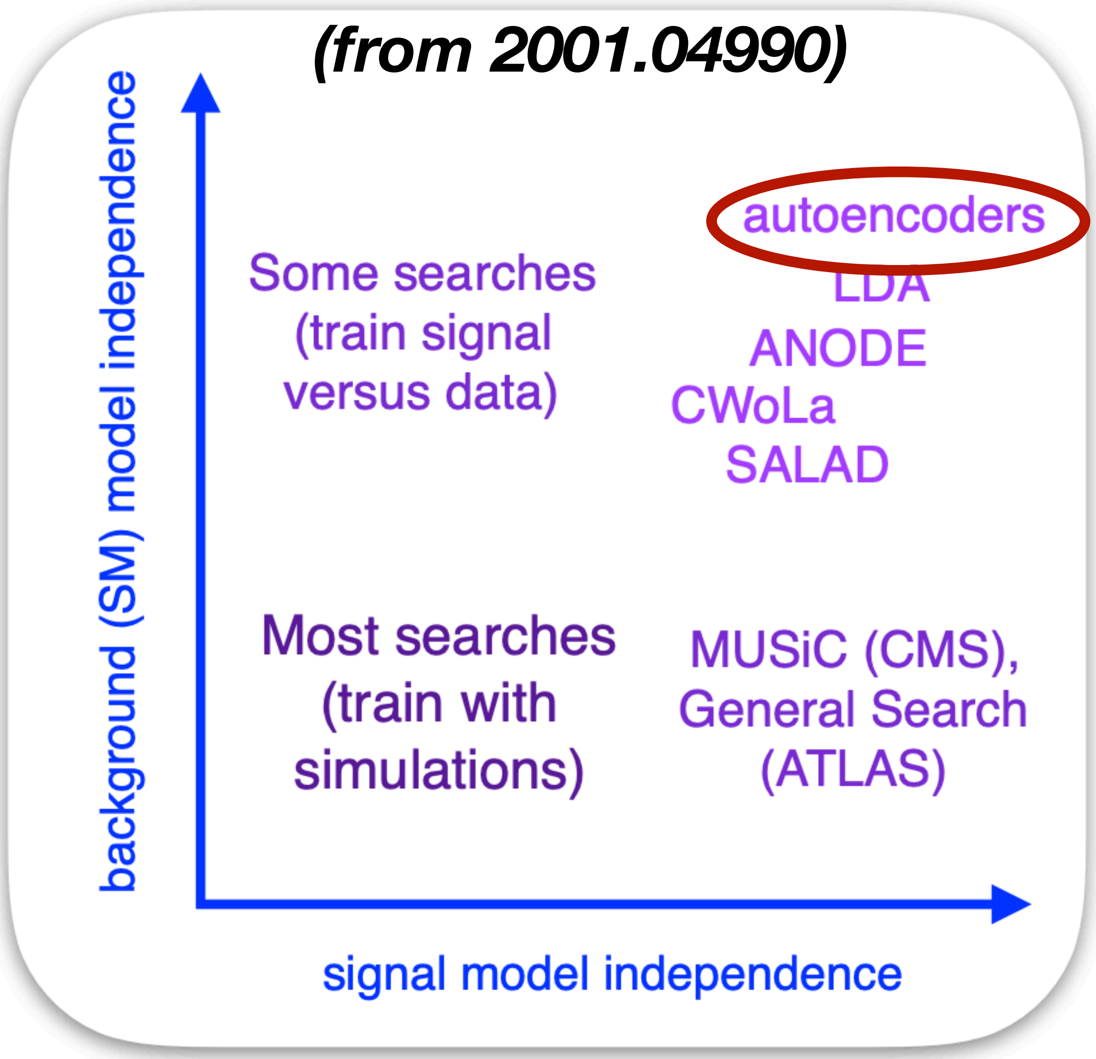
Can we look for new physics, without knowing what to look for?

Can we avoid systematic uncertainties in searches?

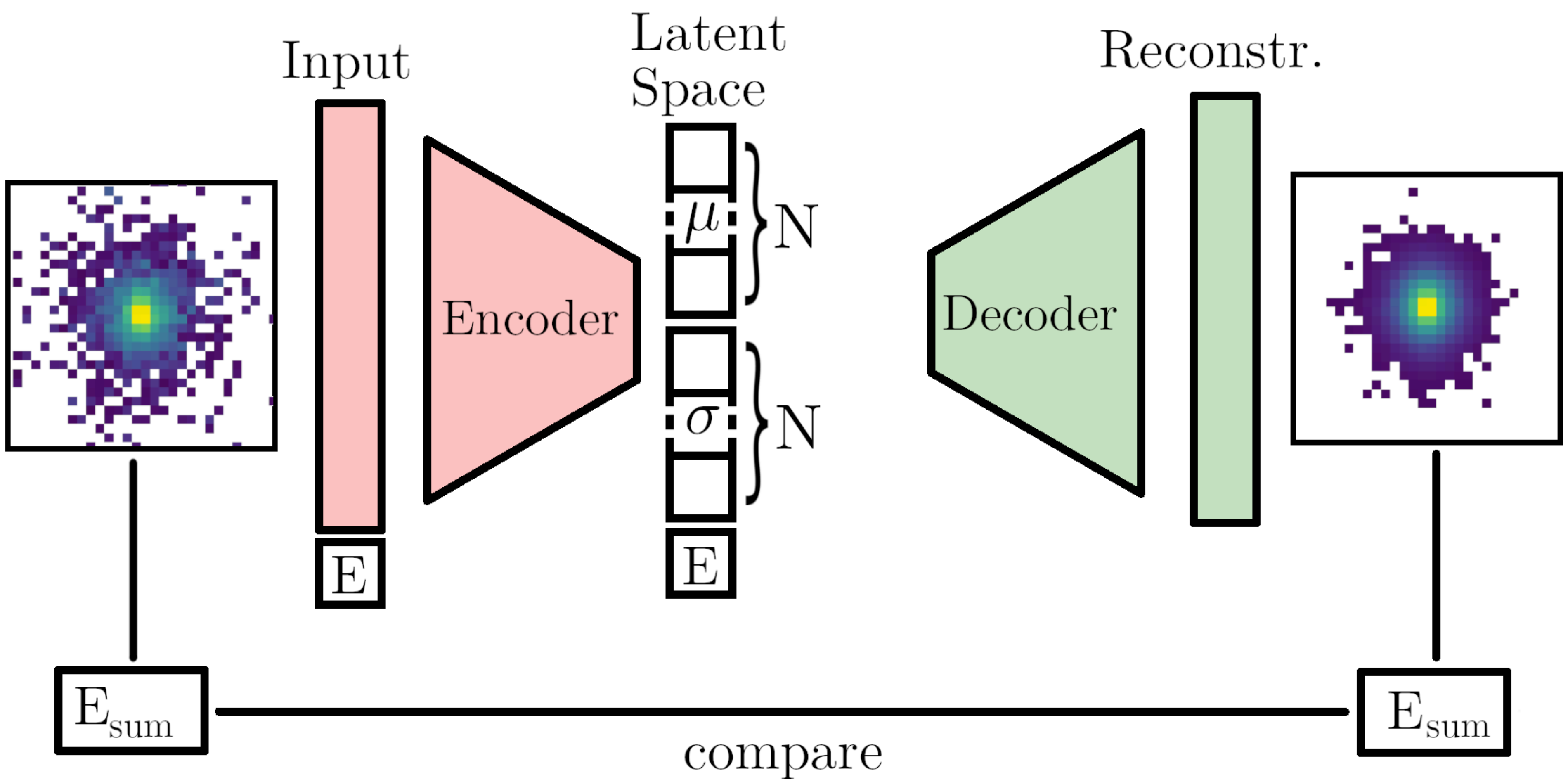


Anomaly detection ideas

- Developing field of different approaches:



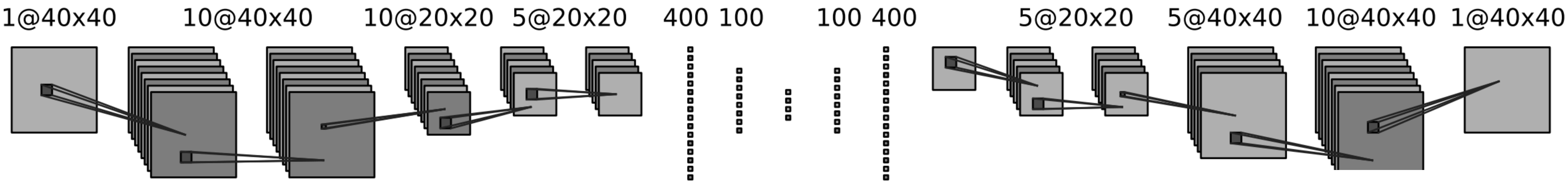
(Variational) Autoencoder

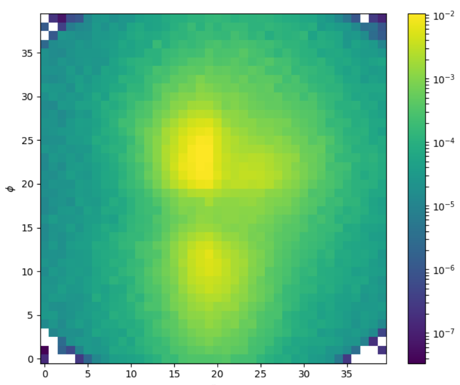


- *Latent space/bottleneck* with compressed representation
  - Dimension reduction
  - Denoising
  - Generation
  - And **anomaly finding**

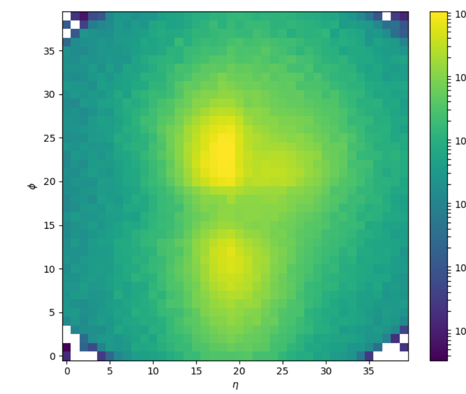
kvfrans
deeplearningbook.org

Autoencoder for Physics

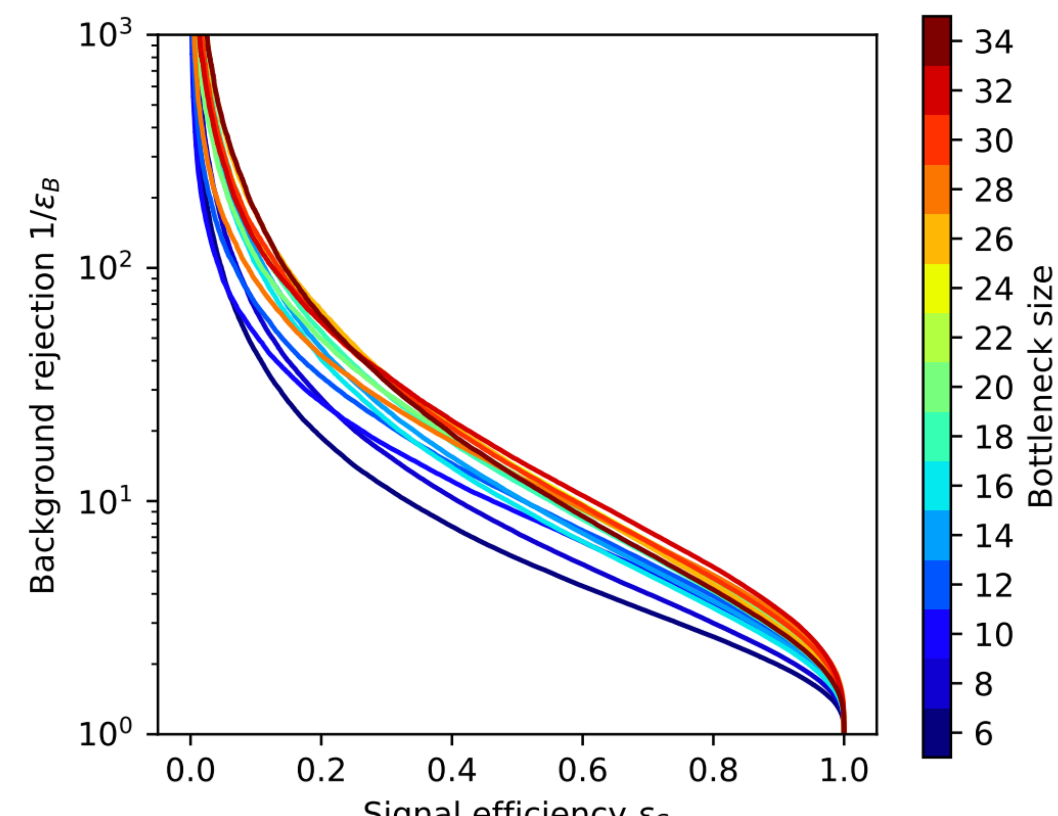




$$L_{\text{auto}} = \sum_{1600 \text{ pixels}} \left(k_T^{\text{norm,in}} - k_T^{\text{auto}} \right)^2$$



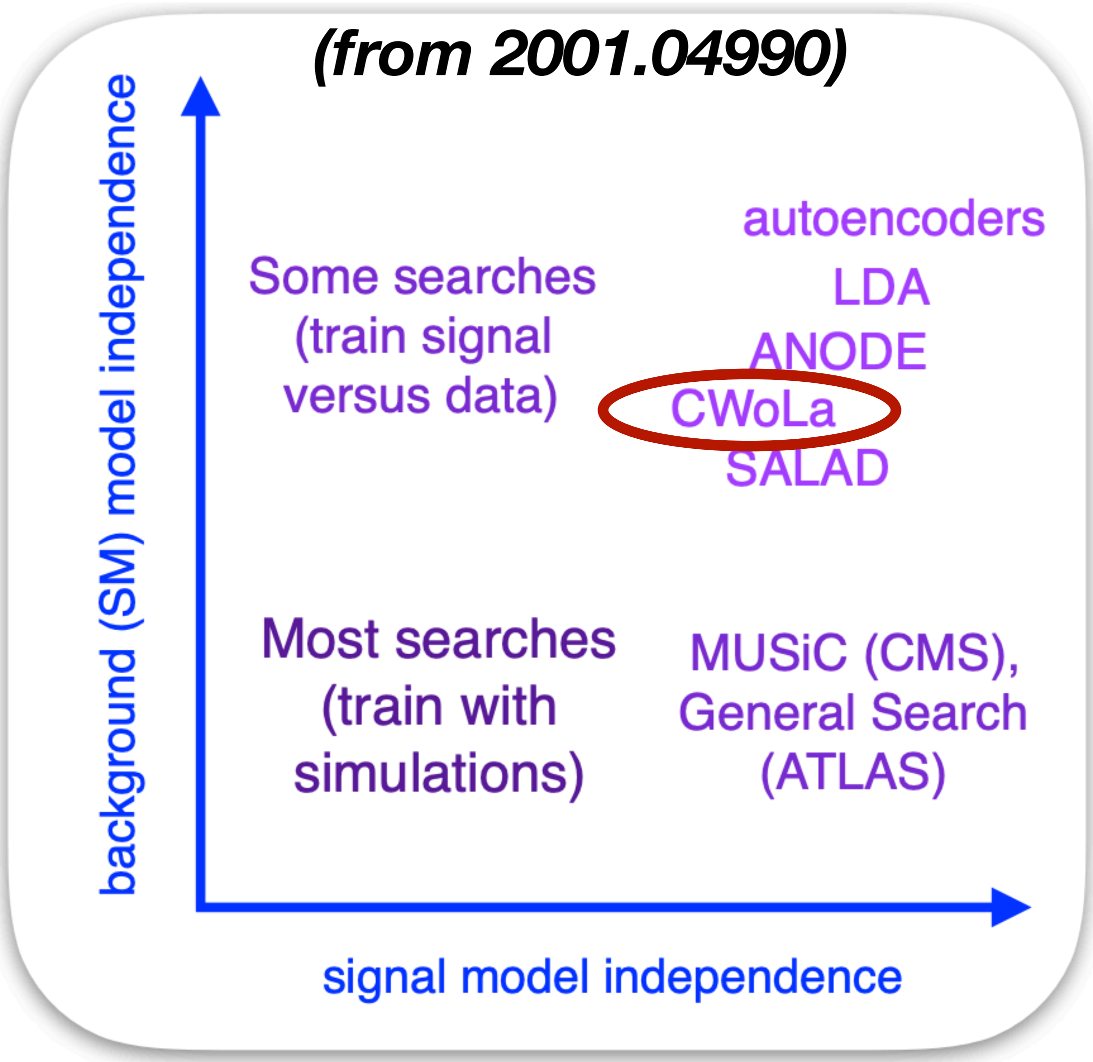
- Can we find new physics without knowing what to look for?
- Train on pure QCD light quark/gluon jets and apply to top tagging
- Top quarks/ new physics identified as anomaly

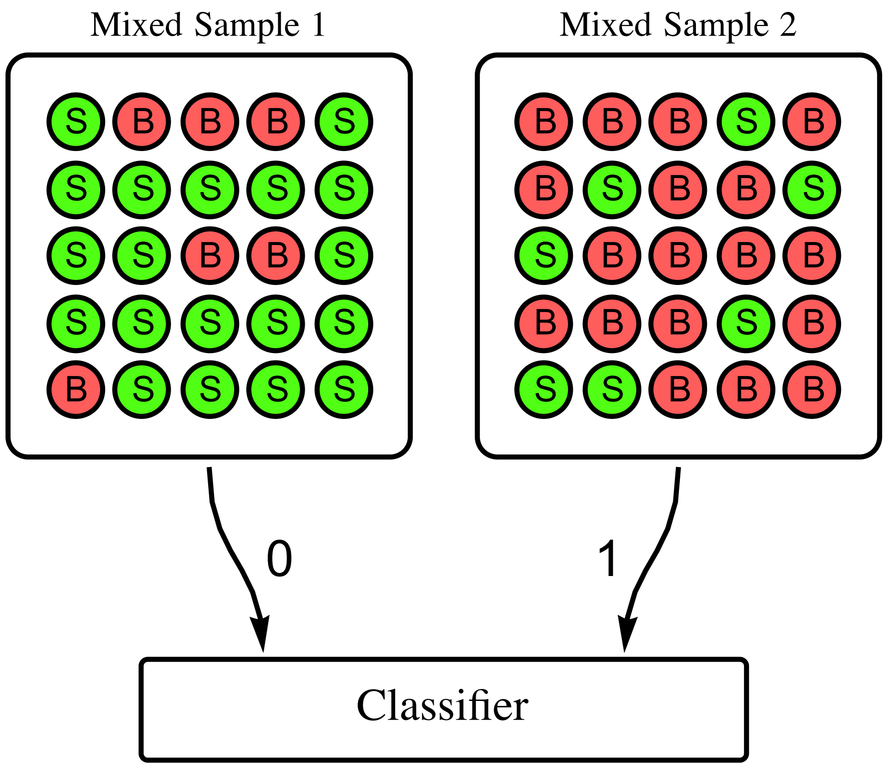


QCD or What?
T Heimel, GK, T Plehn, JM Thompson, 1808.08979
Searching for New Physics with Deep Autoencoders
M Farina, Y Nakai, D Shih, 1808.08992

Anomaly detection ideas

- Developing field of different approaches



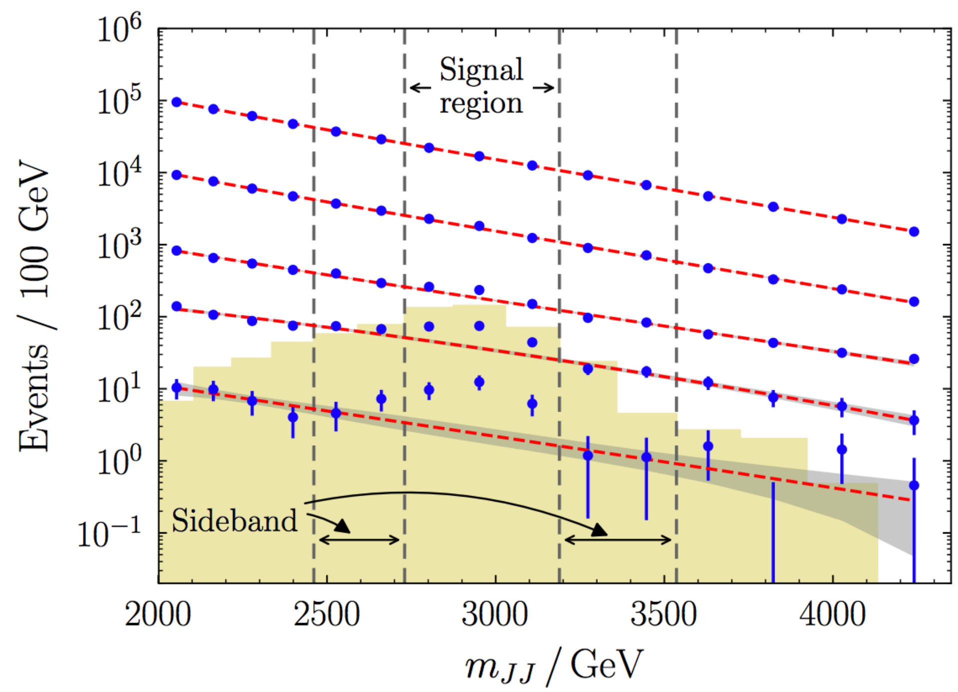


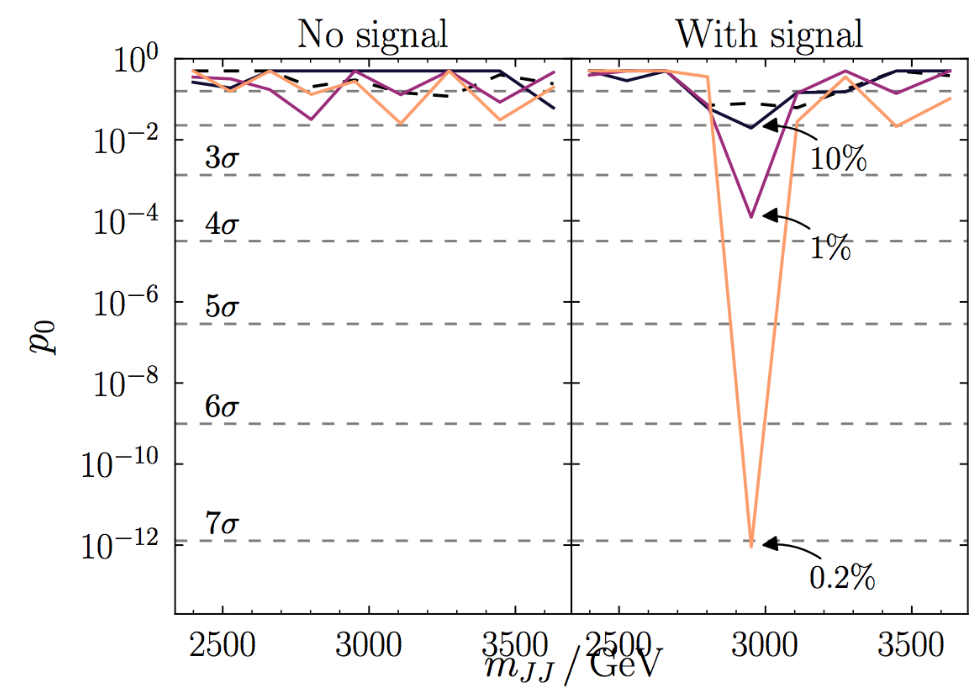
CWola Hunting

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

Distinguishing mixed samples is equivalent to signal/background classification!

First result by ATLAS!!! 2005.02983



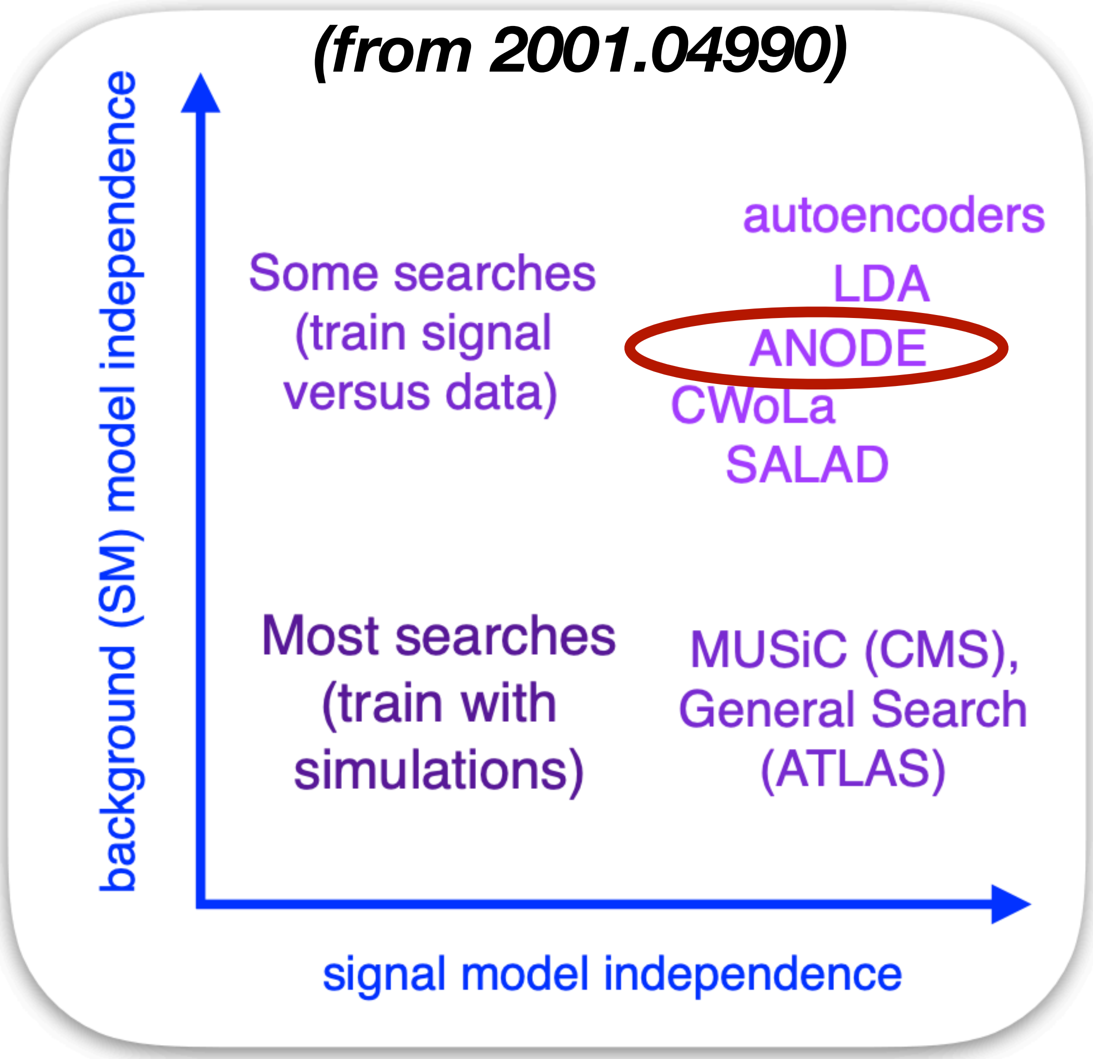


- Assume signal is resonant in one variable
- Define signal region and sidebands
- Train classifier and look for excess

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

Anomaly detection ideas

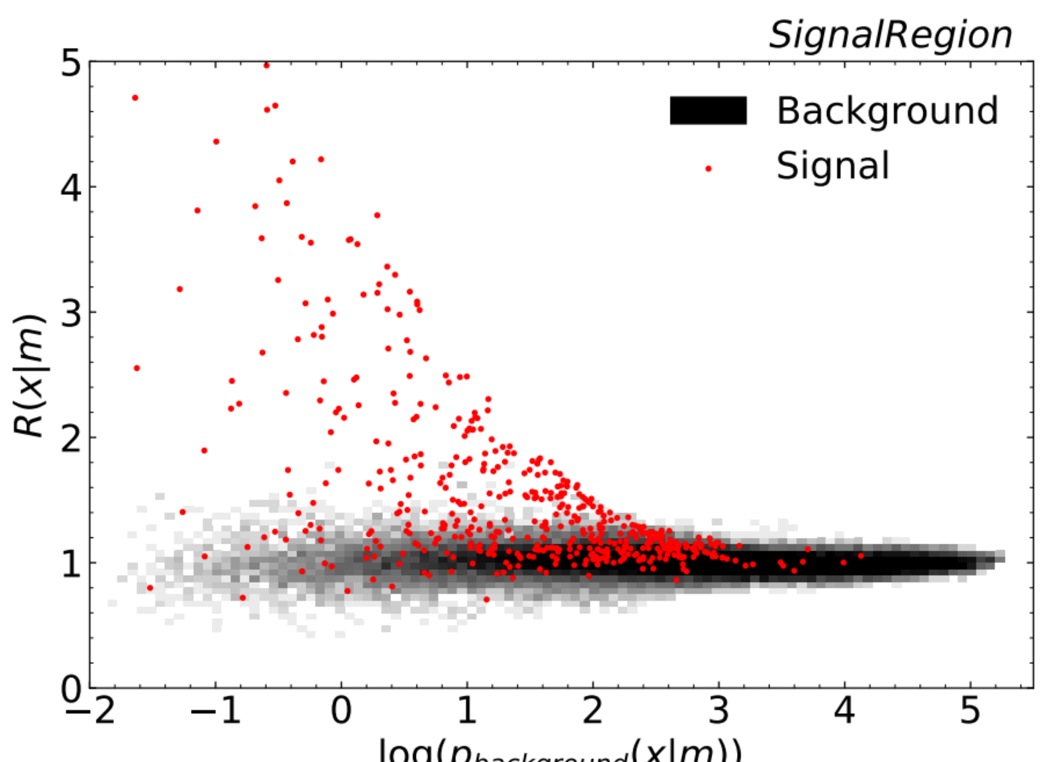
- Developing field of different approaches

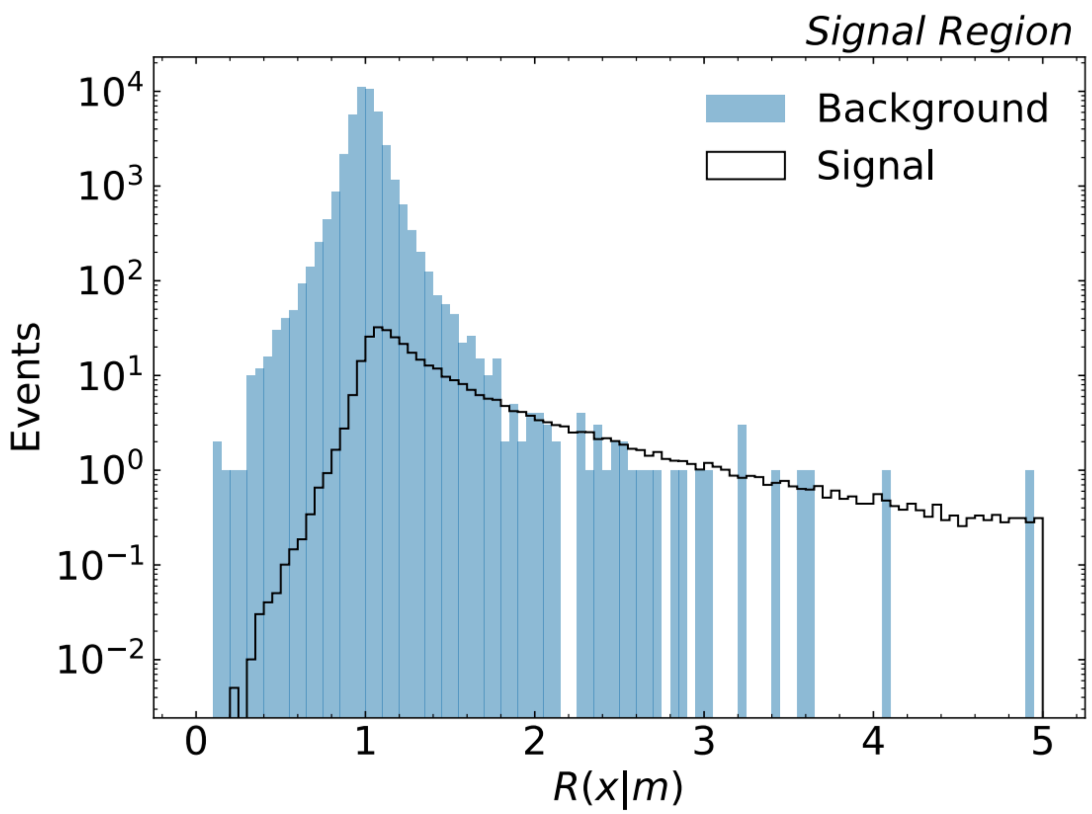


ANODE: ANOmaly detection with Density Estimation

An anomaly is a local over density of events

- Build density estimator in sideband region P_{SB}
- Extrapolation to signal region gives background estimate P_{SB} -> P_{BG}
- Build density estimator in signal region P_{SR}
- Likelihood ratio $R = P_{SR}/P_{BG}$
- *Density estimation via MAF (1705.07057) (Masked Autoregressive Flow)*

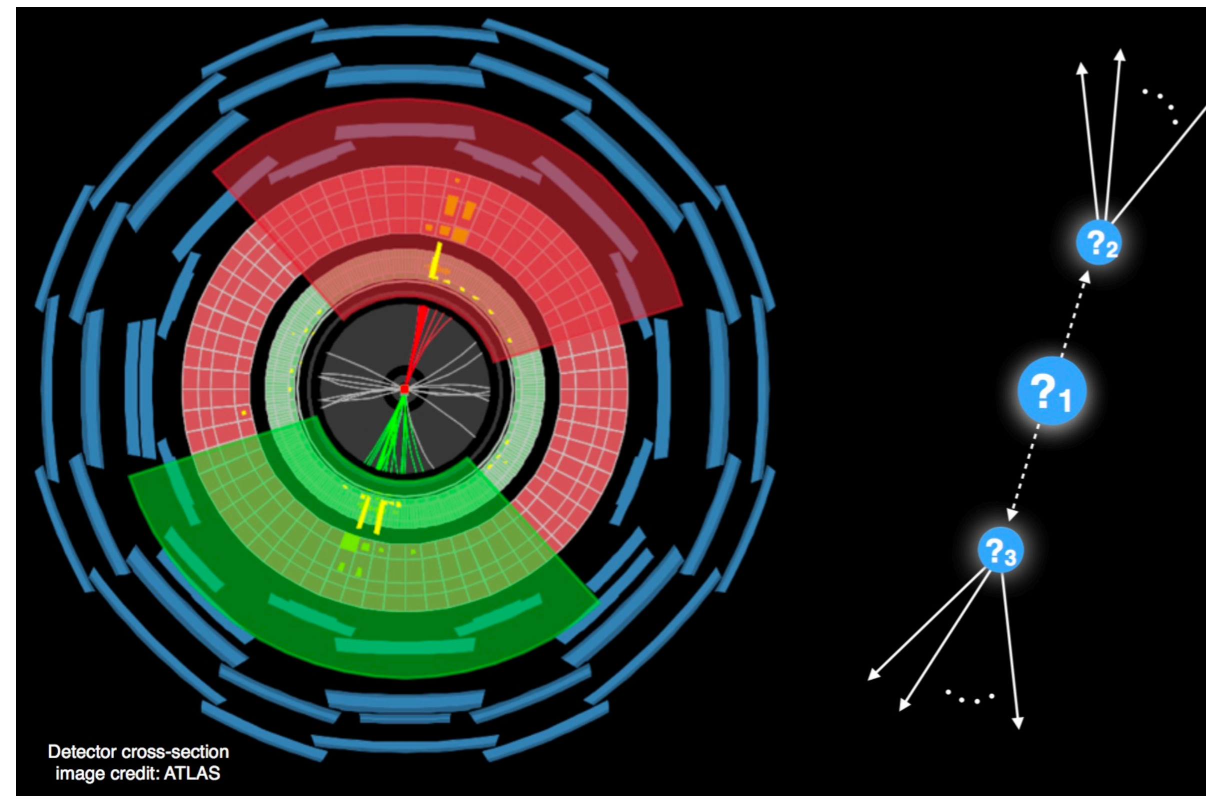




Anomaly Detection with Density Estimation, B Nachman, D Shhih 2001.04990

LHC Olympics 2020

- No evidence for new physics found at the LHC
- Could it be that we are looking in the wrong places?
- Recent new ideas for anomaly detection
- Data challenge to spur development of new approaches and understand the trade-offs between different approaches
    - -> LHC Olympics 2020

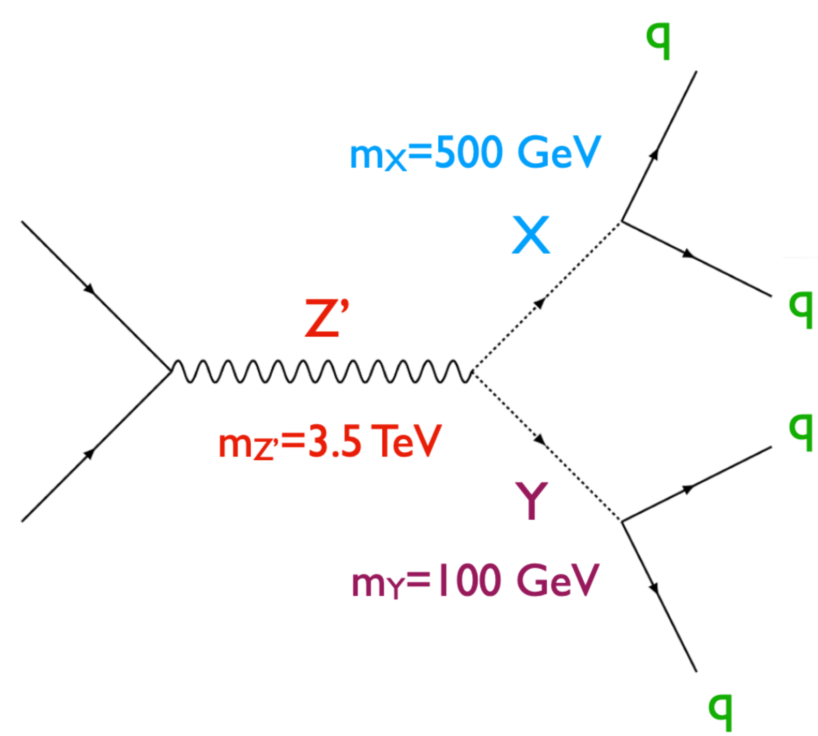


Organisers: Ben Nachmann, David Shih & GK

Timeline & Dataset

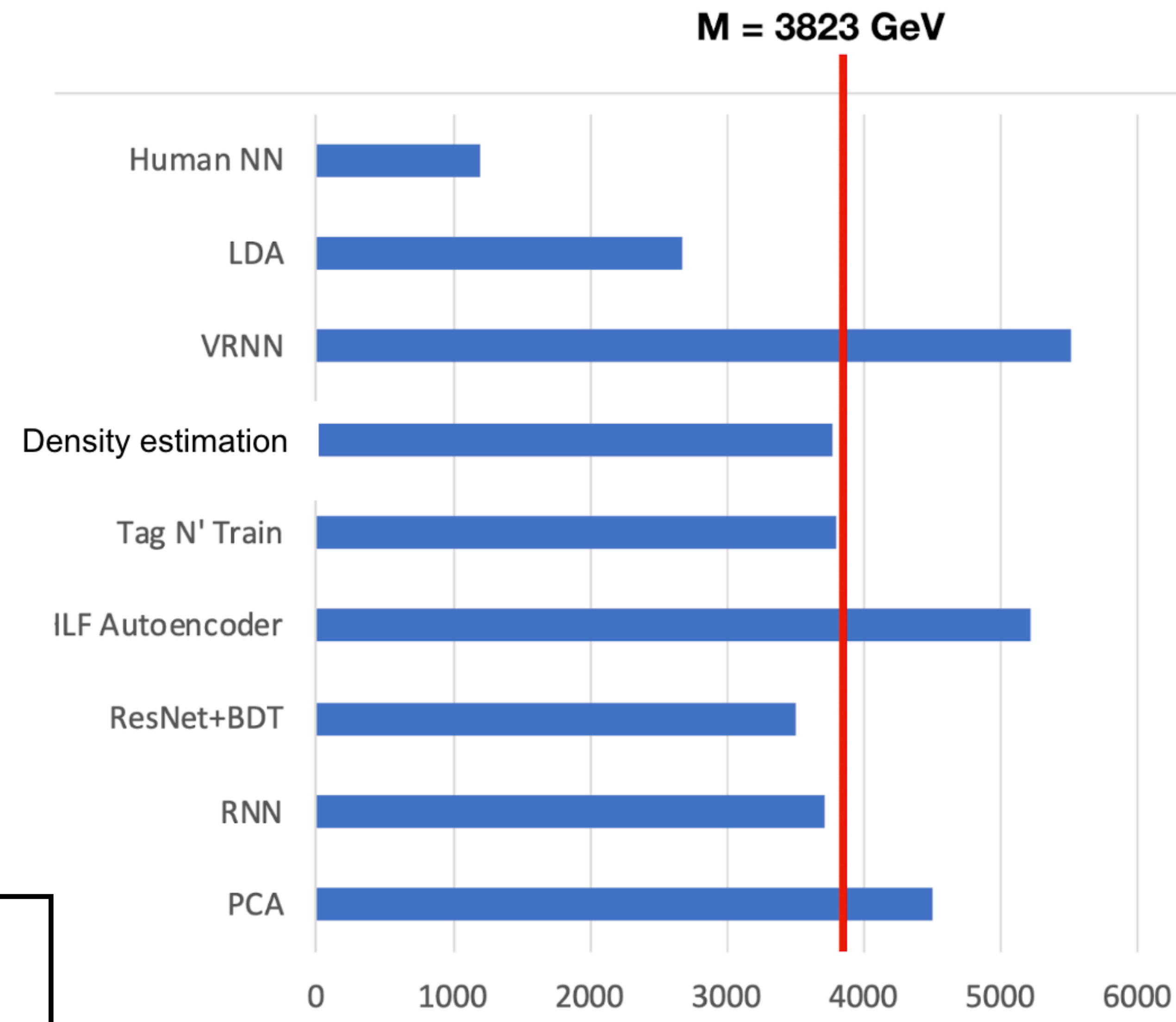
- April 2019:
  Released R&D dataset
- Novemeber/December 2019:
  Released 3 black box datasets
- January 2020:
  Submission of black-box results by participating groups
- January 16, 2020:
  Unveiling of black box 1 at ML4Jets
- *July 2020:*
  Online Anomaly mini-workshop
  (likely Week of July 13th 2020)

- R&D Dataset:
  - Pythia 8 / Delphes
  - 1M QCD events
    100k labelled signal
  - 4-vectors of reconstructed particles
  - No particle ID, charge, …
  - Single R=1 jet trigger pT>1.2 TeV



Opening black box 1

- Total number of 10 submissions, including:
  - Autoencoders
  - CWoLa hunting
  - PCA outlier detection
  - LSTM / RNN
  - CNN+BDT
  - density estimation
  - biological neural network



Resonance Mass

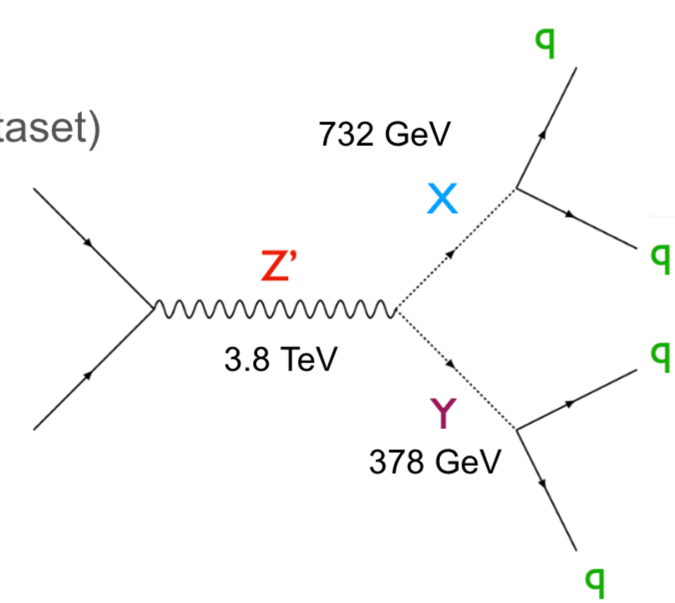
Signal: 834 events

Z'->XY; X,Y->qq
(same topology as R&D dataset)

mZ' = 3823 GeV

mX = 732 GeV

mY = 378 GeV



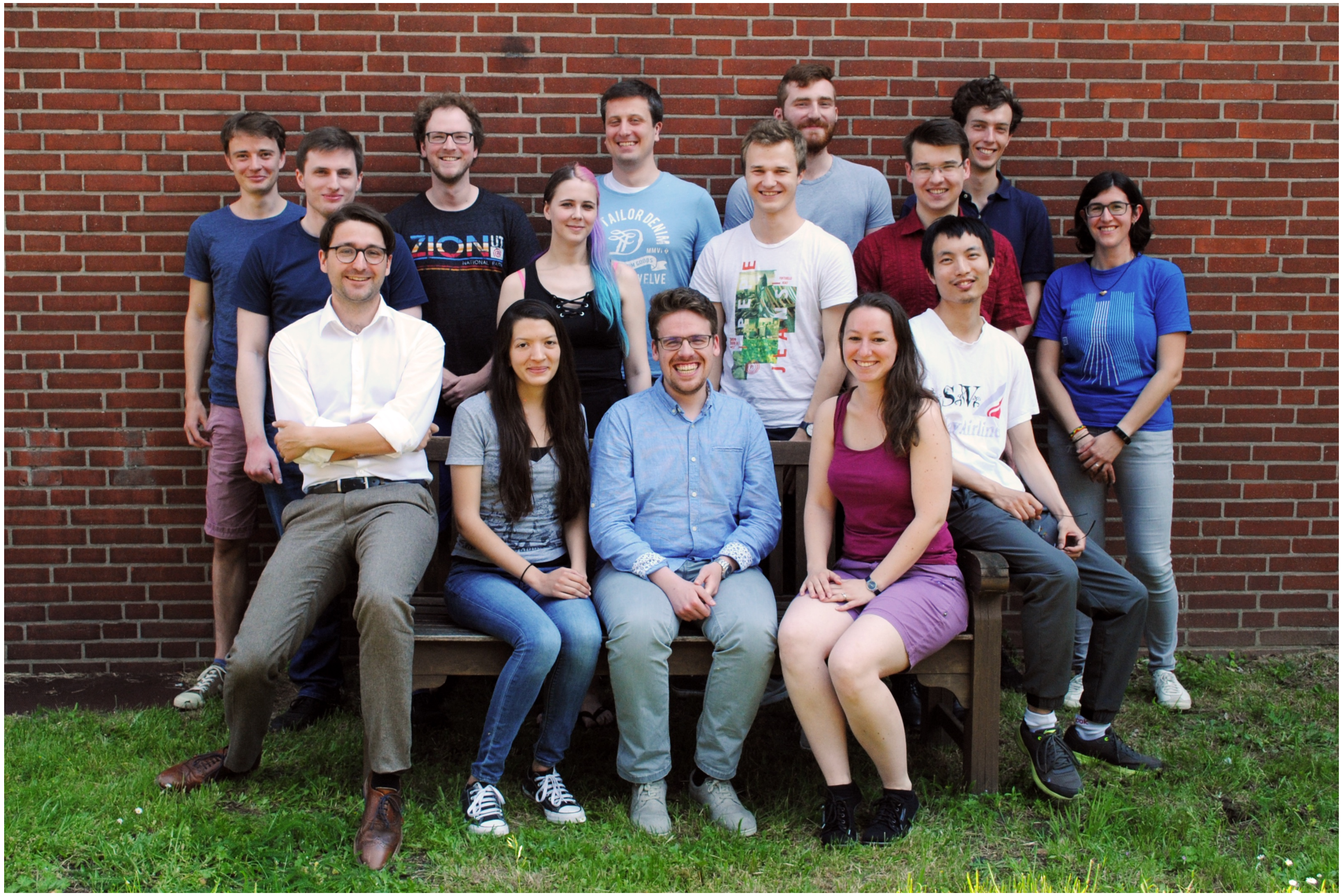
Closing

Conclusions

- Deep Learning for particle physics is rapidly developing solutions to a wide range of problems
  - **Object and Event classification**
  - **Anomaly detection**
  - **Robustness** and uncertainties
  - Fast reconstruction and simulation
- If you are excited now:
  - Consider joining the LHC Olympics: -> LHC Olympics 2020
- Automate parts of our ML tasks?

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Emmy Noether-Programm
Deutsche Forschungsgemeinschaft
DFG



Thank you!

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

PIER
Partnership of
Universität Hamburg and DESY

Bundesministerium
für Bildung
und Forschung