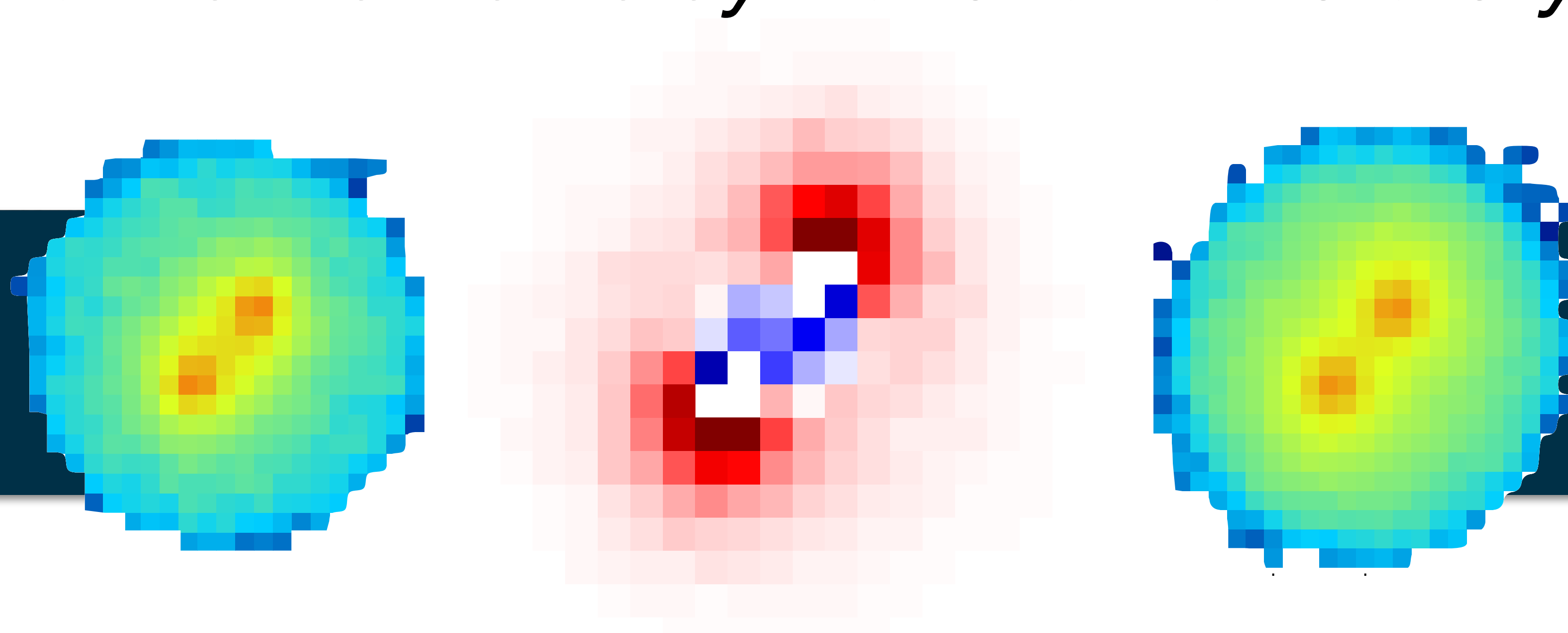
Modern Machine Learning

for Classification, Regression, and Generation in Jet Physics

Benjamin Nachman

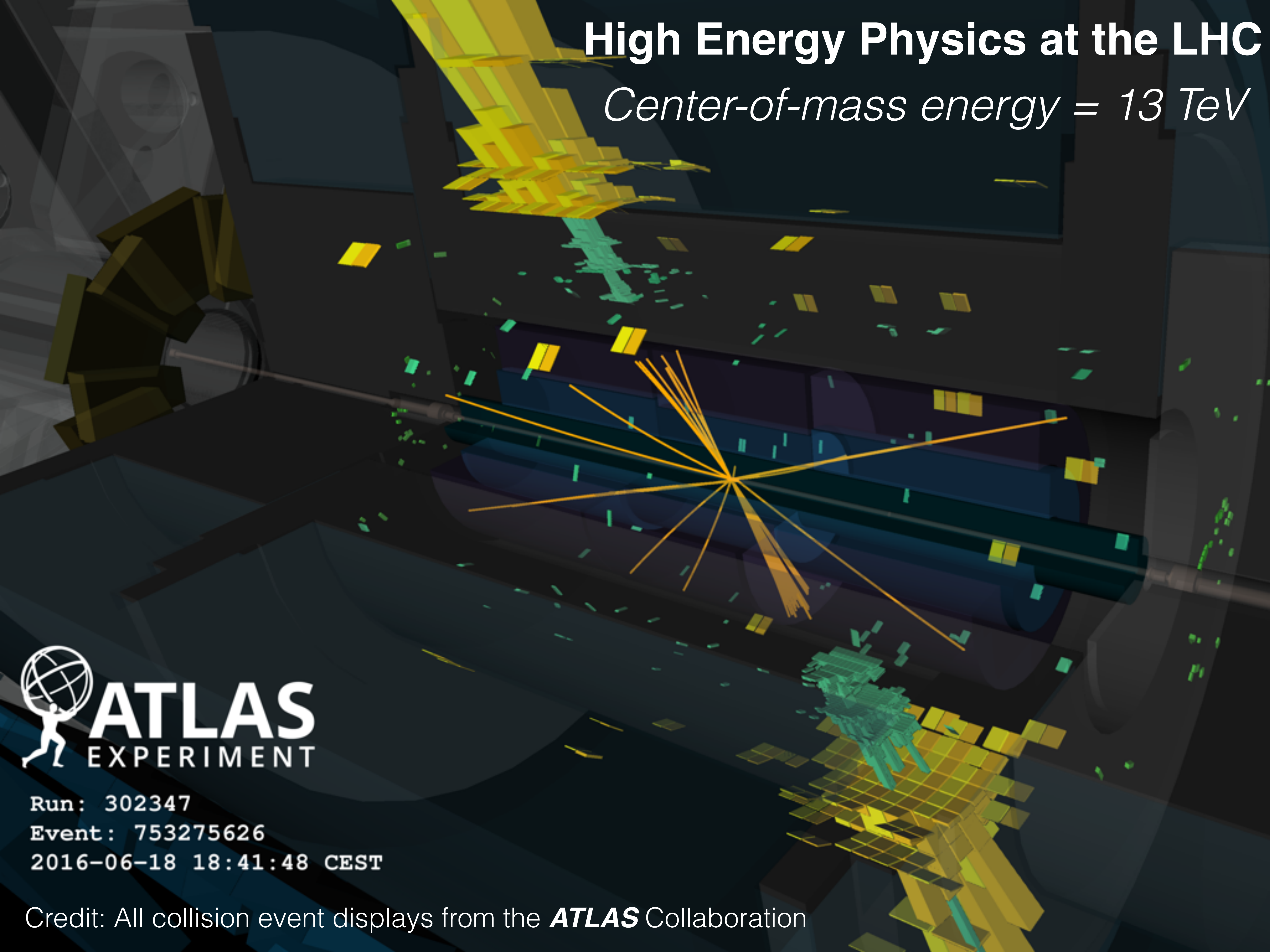
Lawrence Berkeley National Laboratory



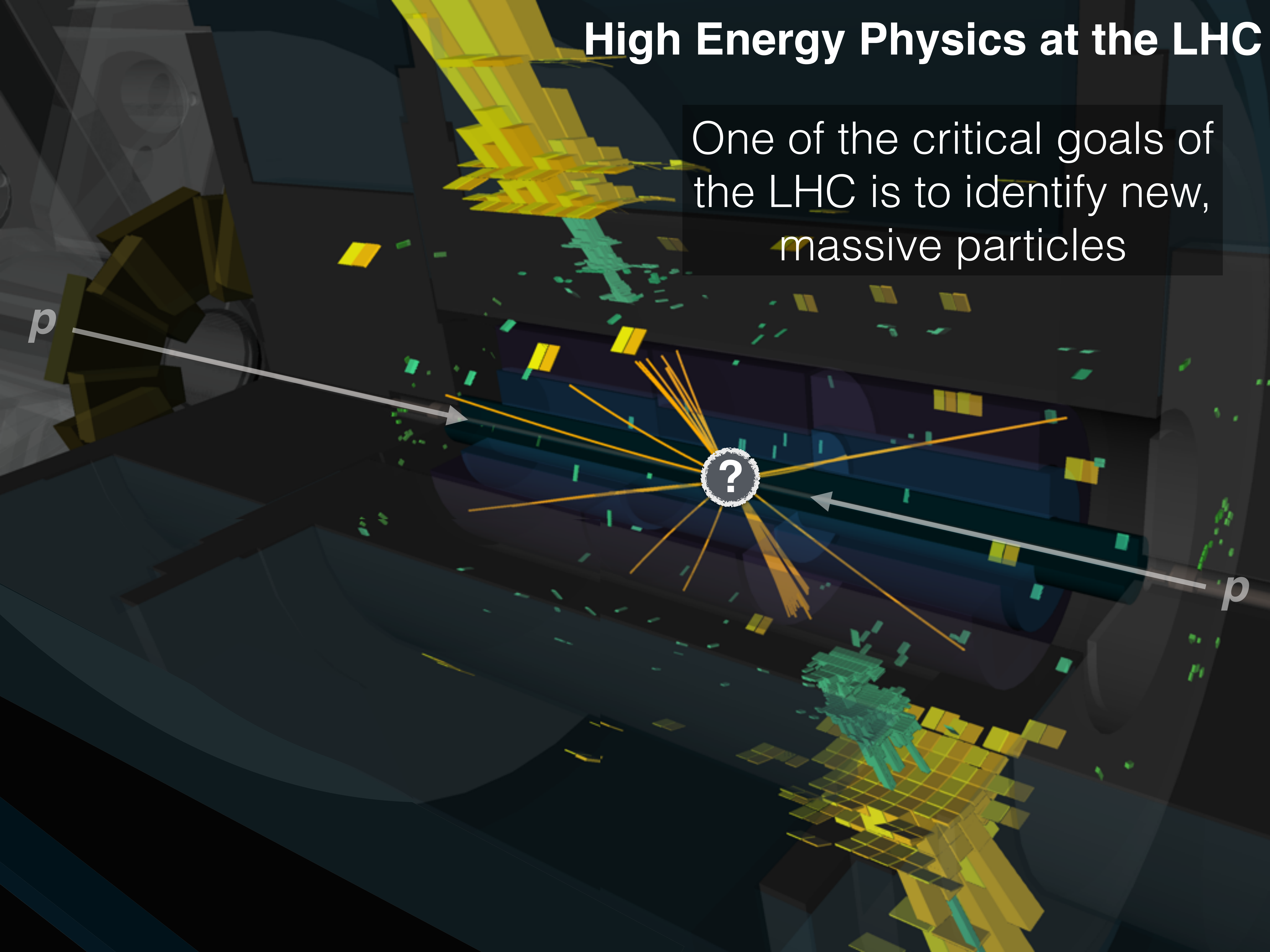
CERN Data Science Seminar, November 14, 2017

High Energy Physics at the LHC

Center-of-mass energy = 13 TeV

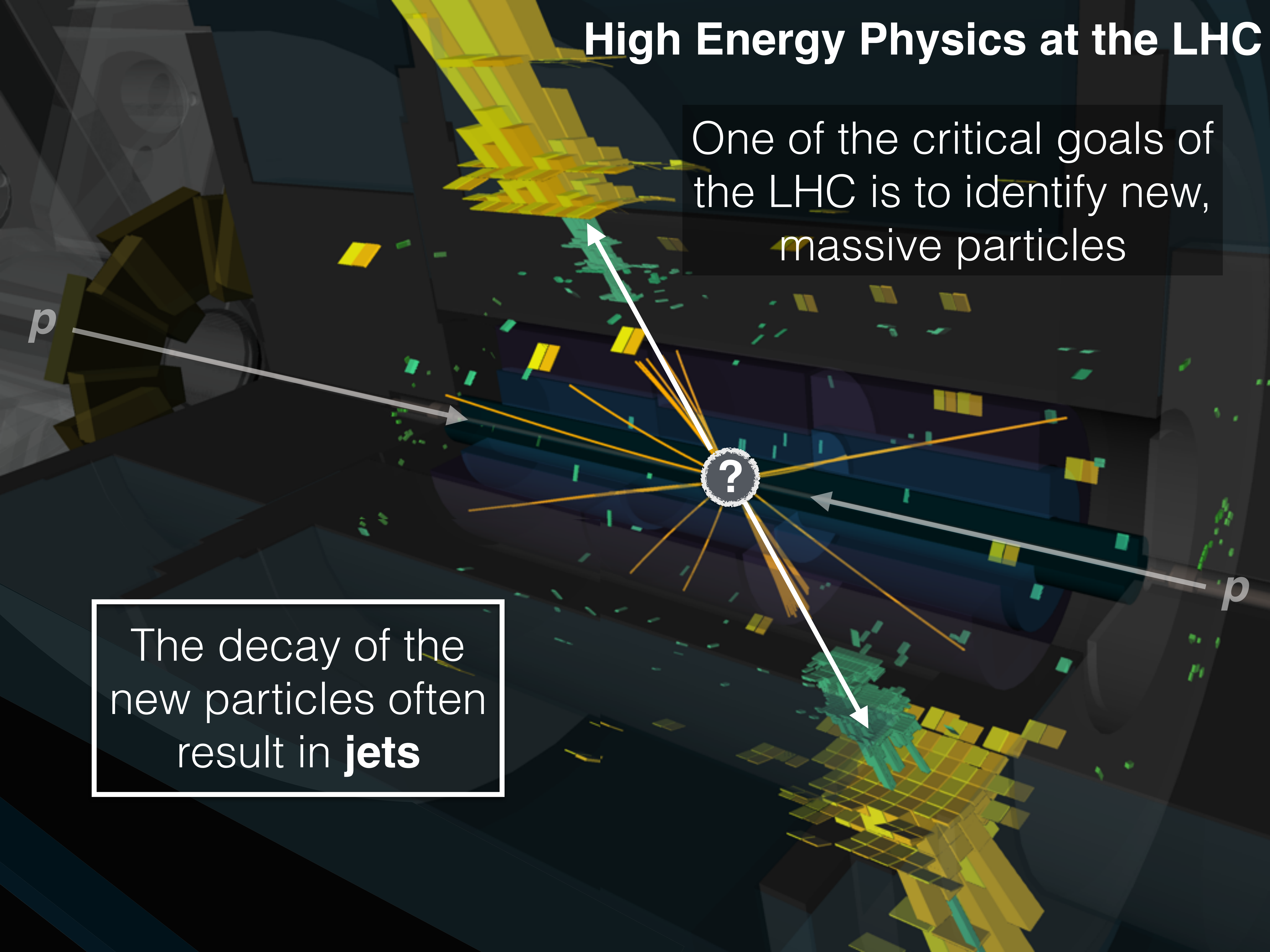


Credit: All collision event displays from the ***ATLAS*** Collaboration


High Energy Physics at the LHC
One of the critical goals of the LHC is to identify new, massive particles
p
?
p

High Energy Physics at the LHC

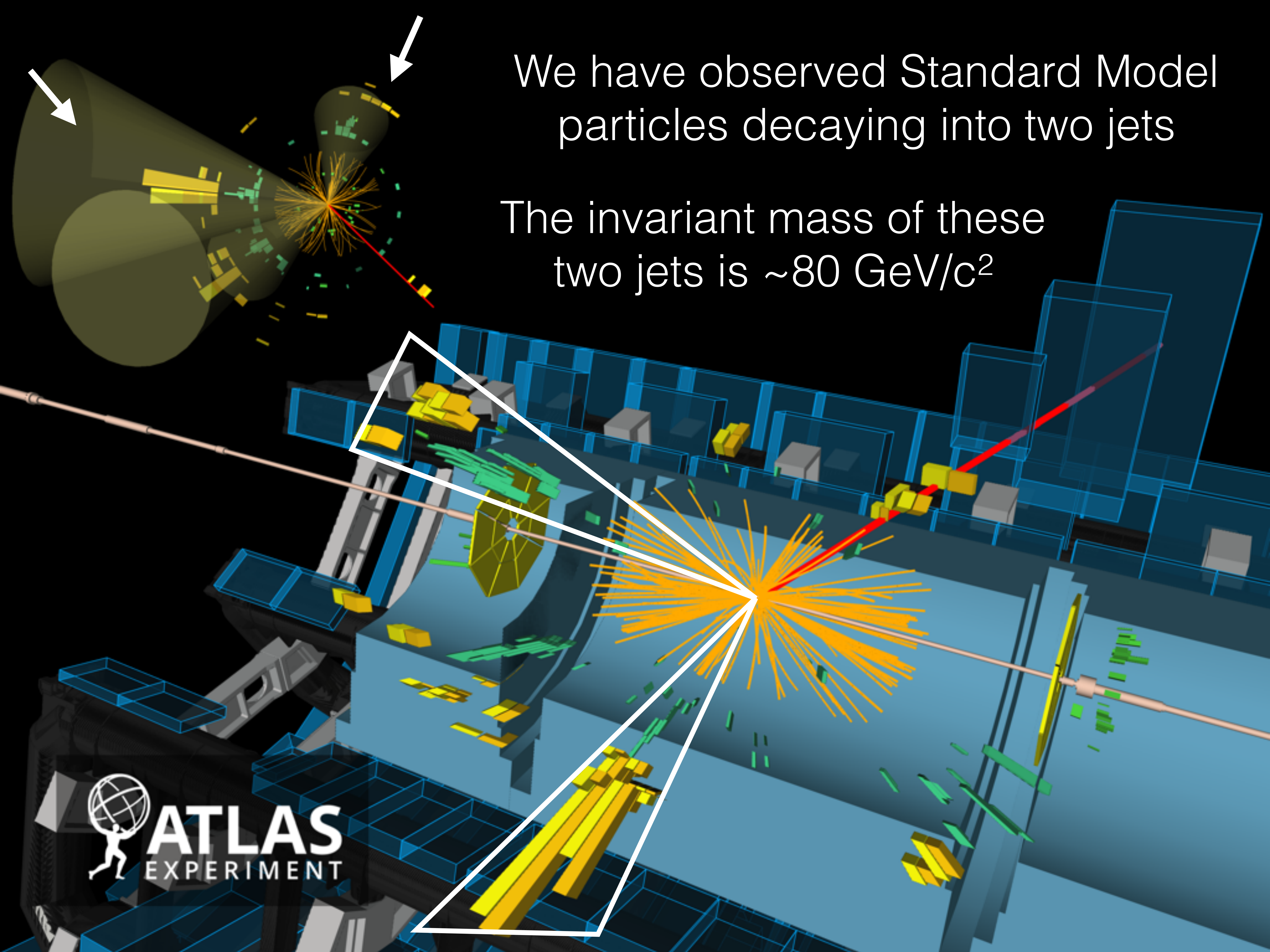
One of the critical goals of the LHC is to identify new, massive particles

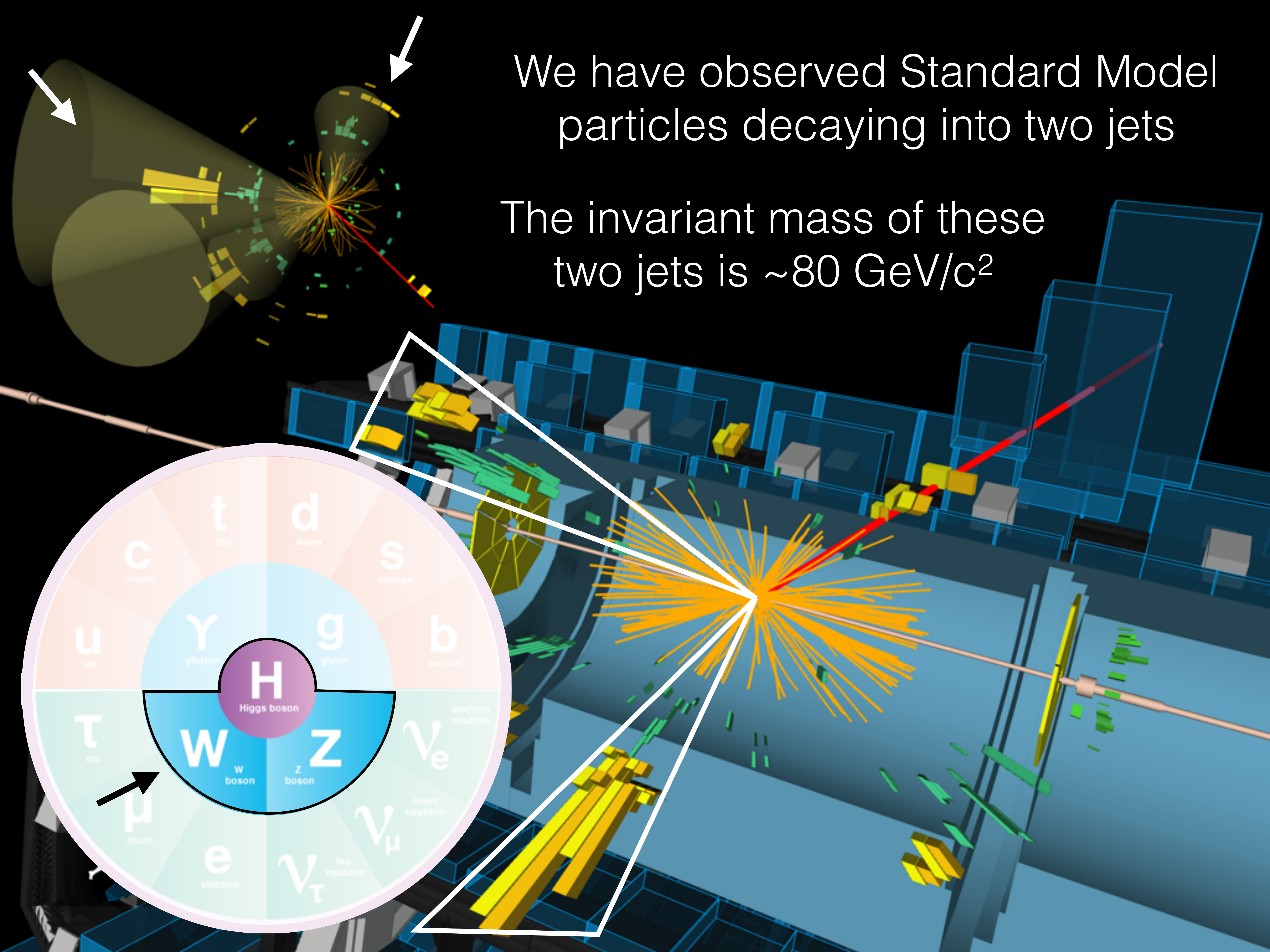


The decay of the new particles often result in **jets**

We have observed Standard Model particles decaying into two jets

The invariant mass of these two jets is ~80 GeV/c²

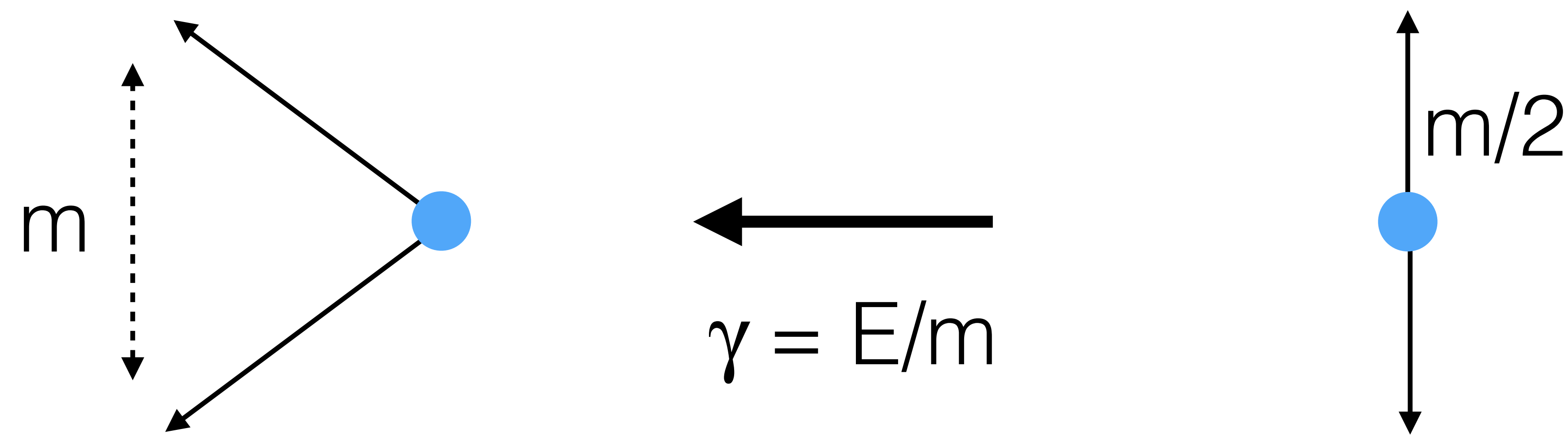


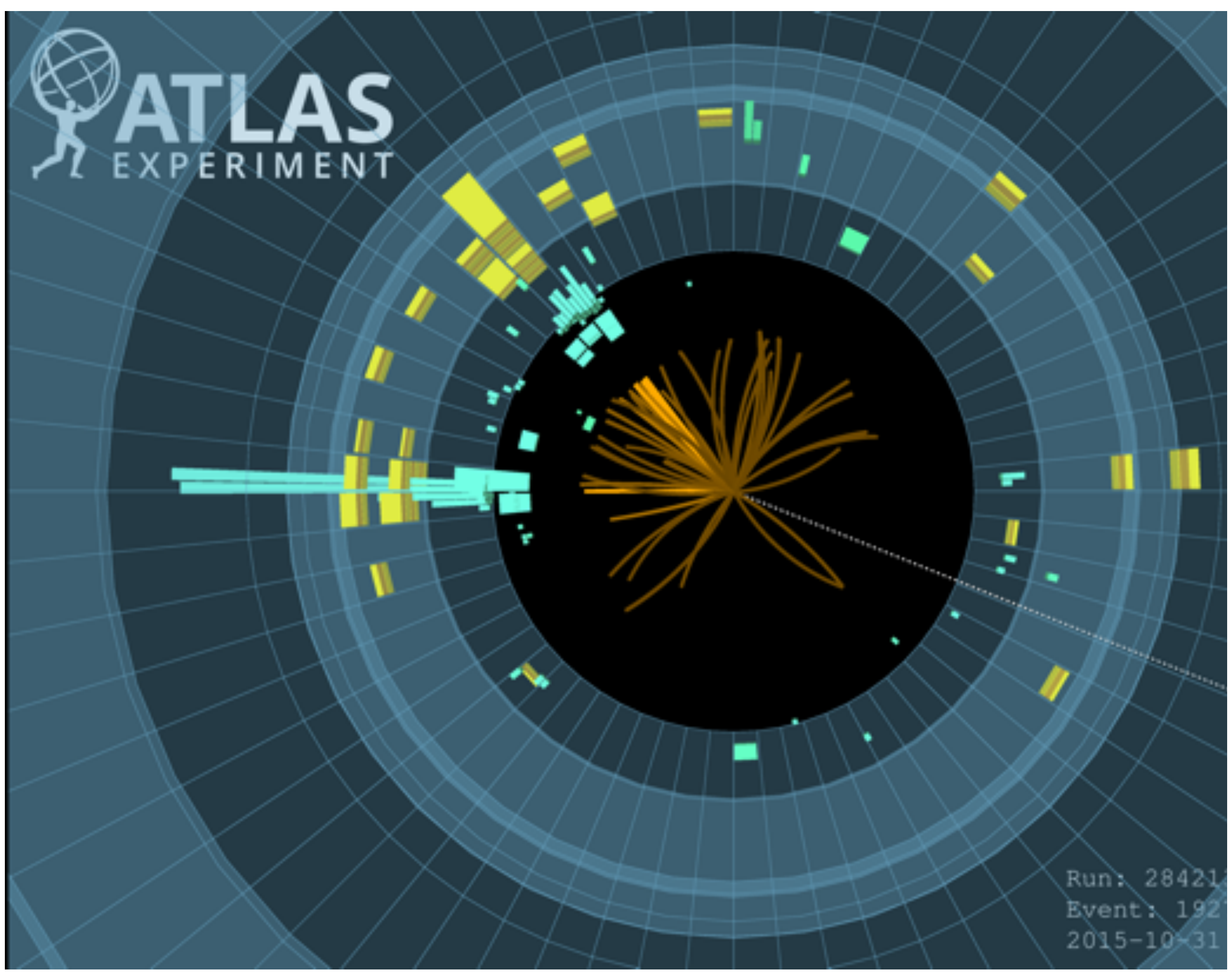


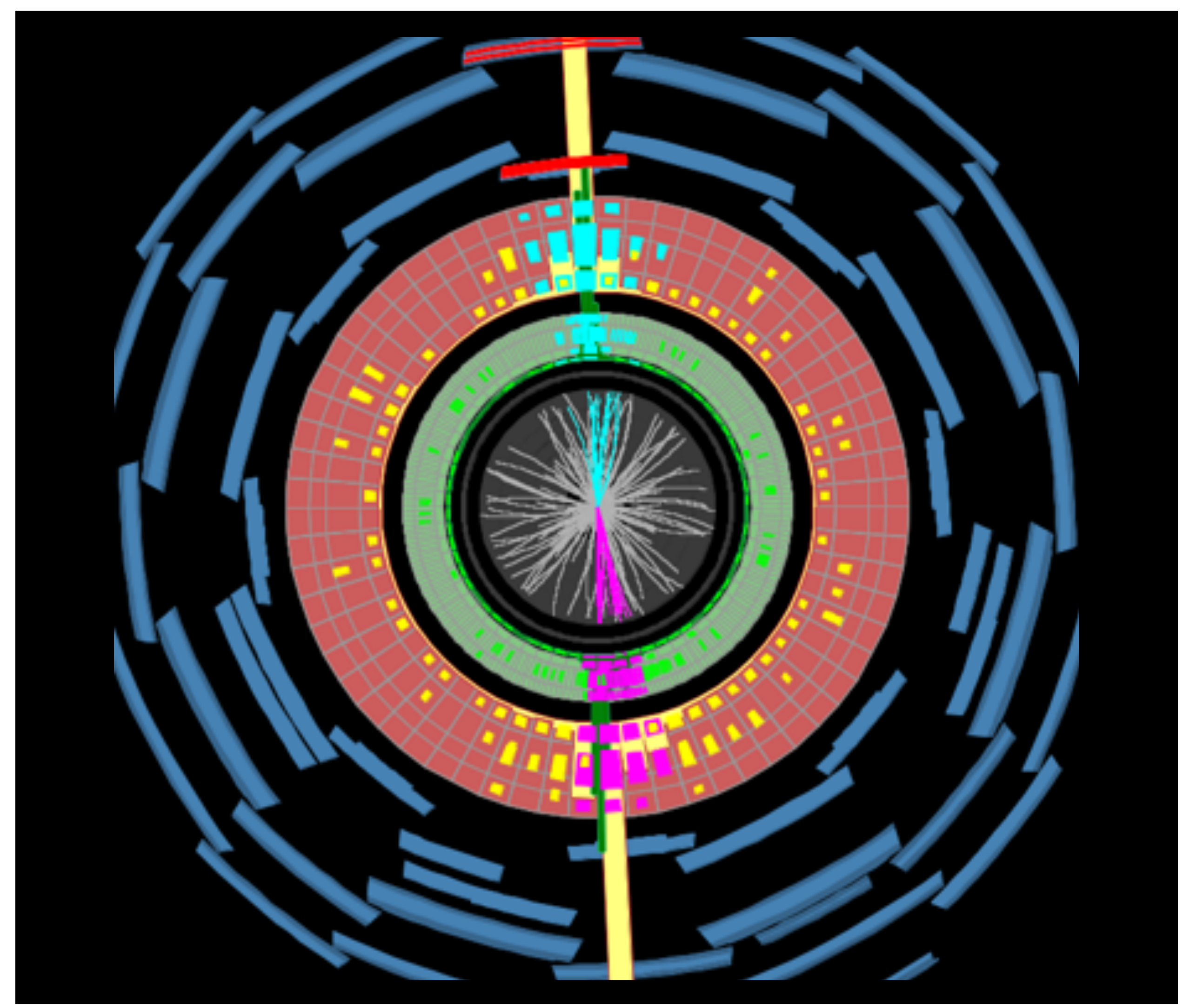
We have observed Standard Model particles decaying into two jets
The invariant mass of these two jets is ~80 GeV/c²
t
top
d
down
c
charm
s
strange
u
up
b
bottom
γ
photon
g
gluon
H
Higgs boson
W
w boson
Z
z boson
τ
tau
ν_e
electron neutrino
μ
muon
ν_μ
muon neutrino
e
electron
ν_τ
tau neutrino

What if you take one of those SM dijet resonances and Lorentz boost it?

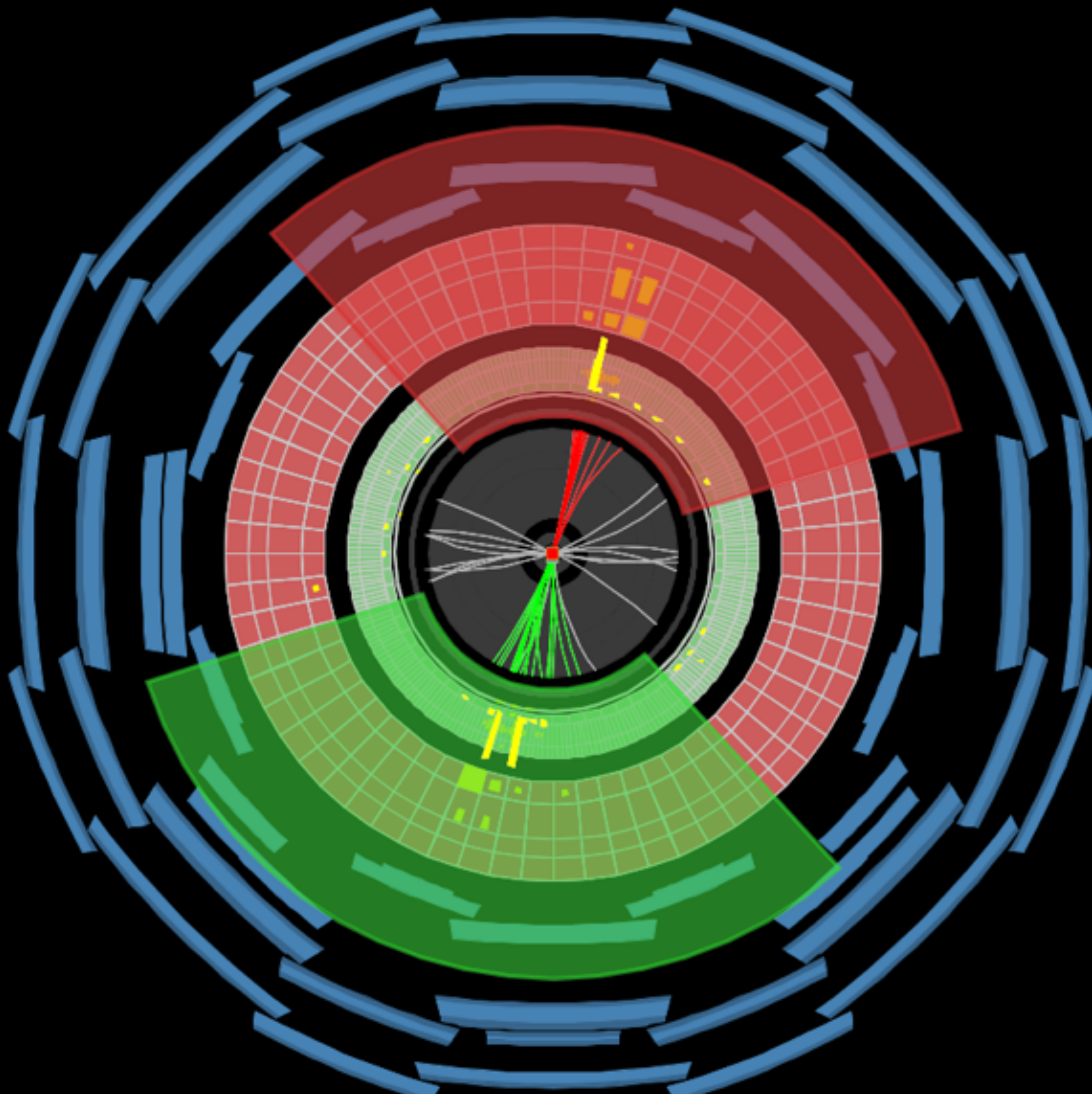
$\phi \sim 1/\gamma = m/E$

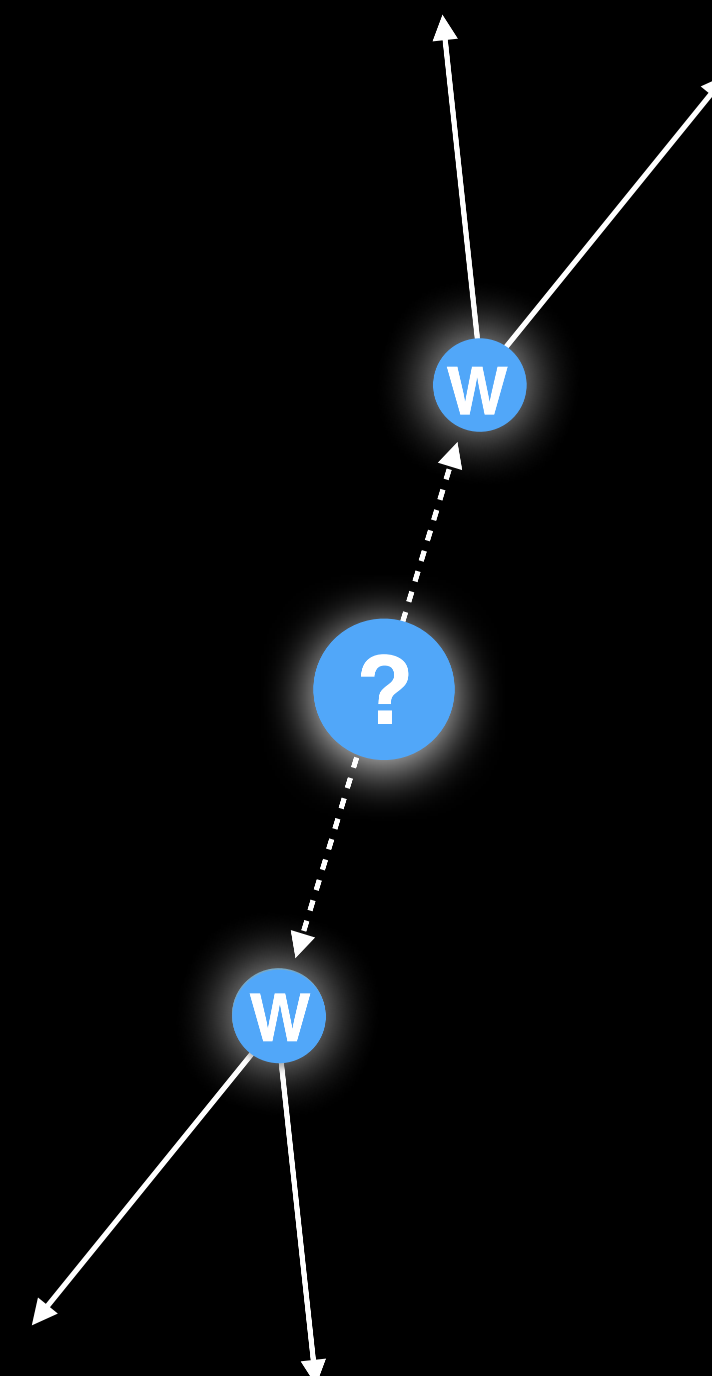






W bosons are naturally boosted if they result from the decay of something even heavier

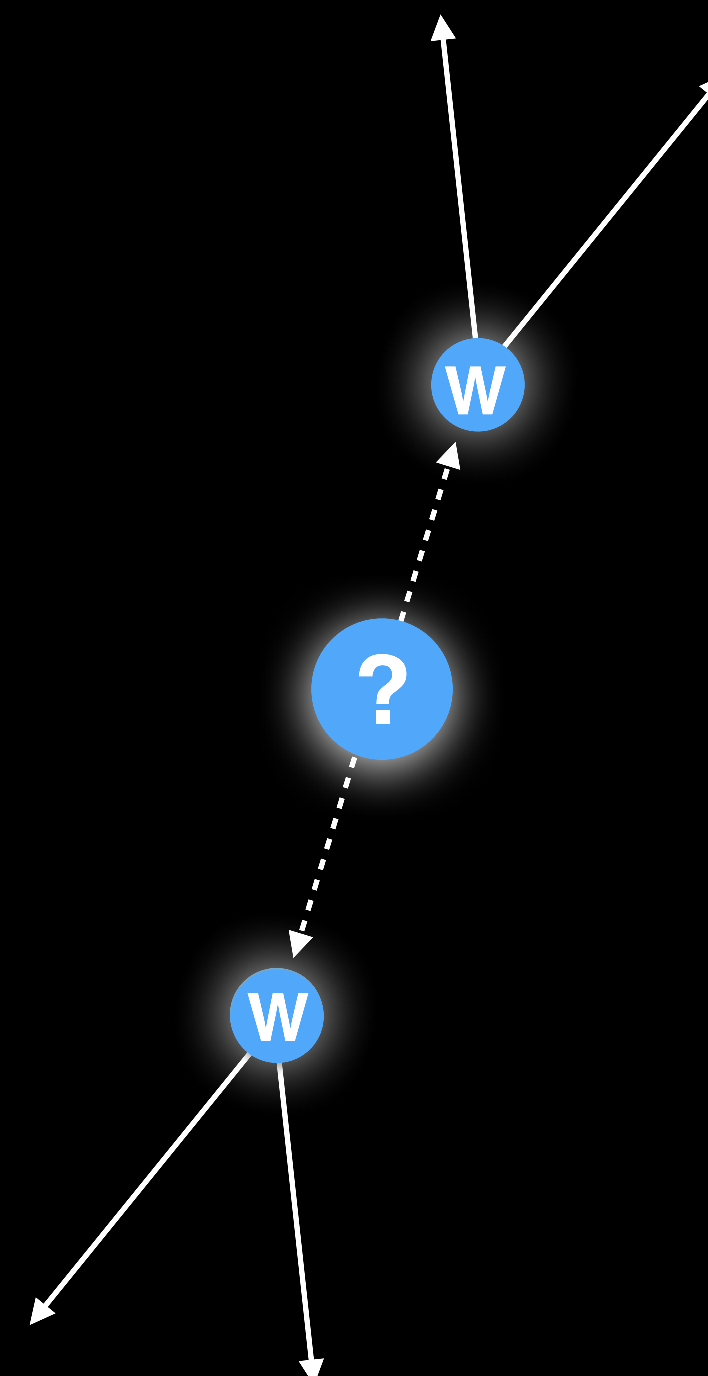


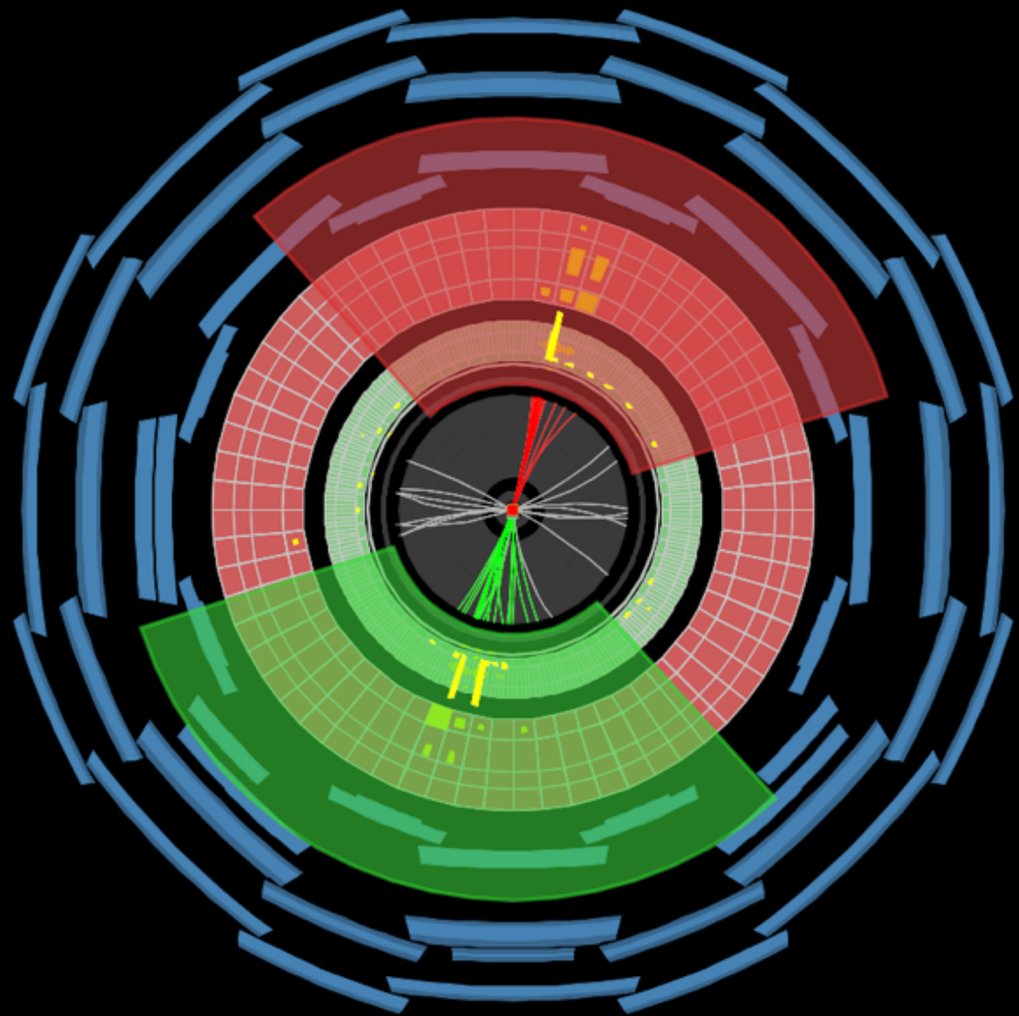


W bosons are naturally boosted if they result from the decay of something even heavier

Goal: Find W jets in an enormous sea of generic q/g jets

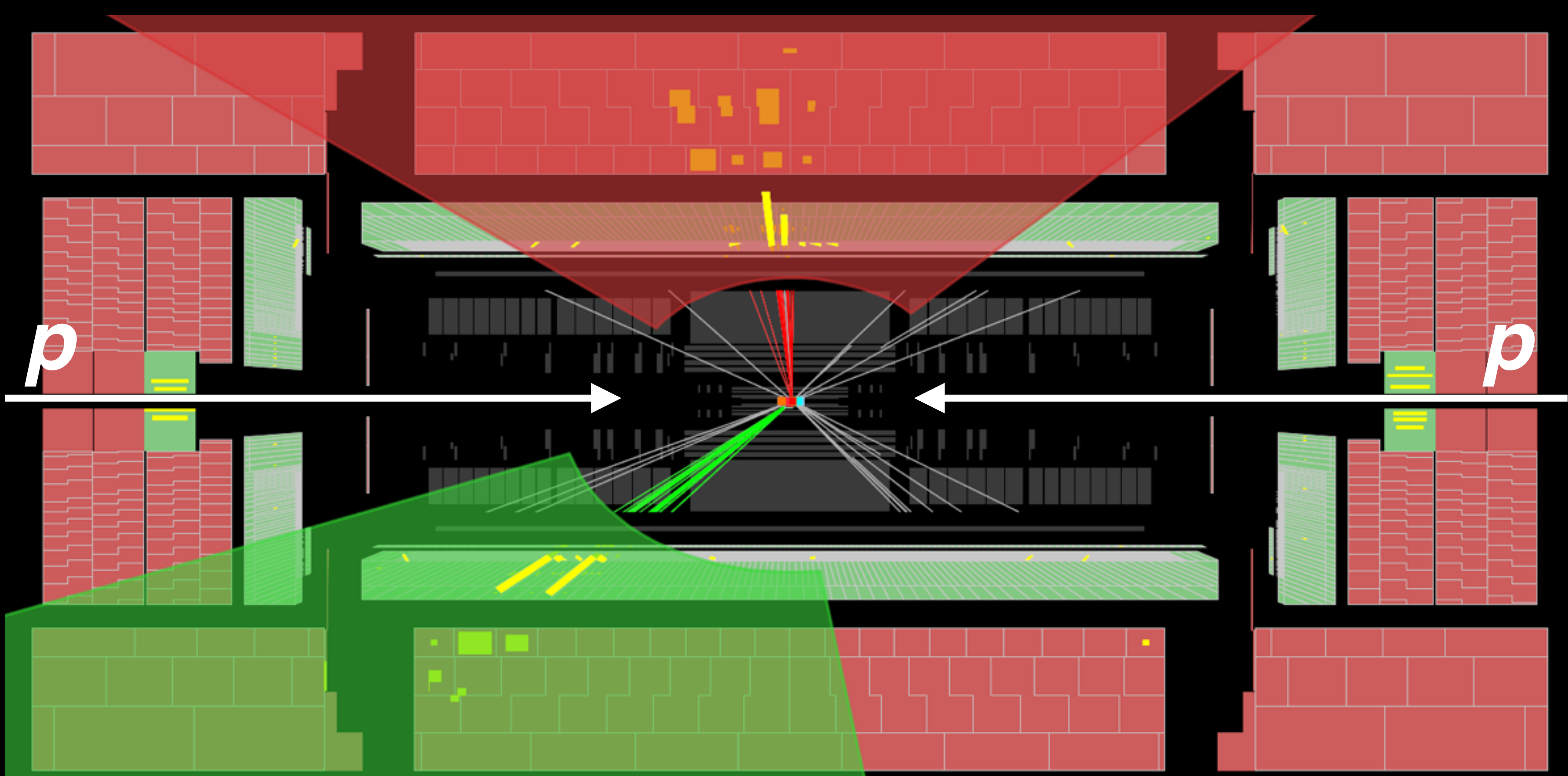
These jets have a non-trivial structure!

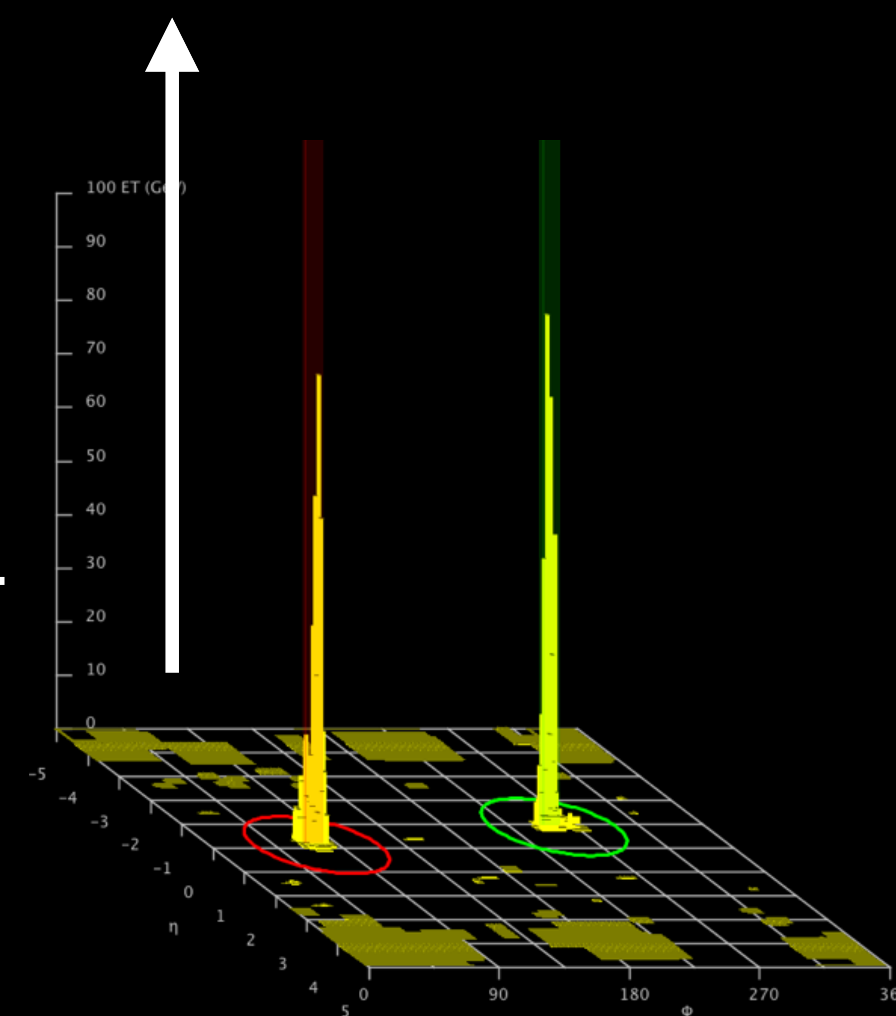




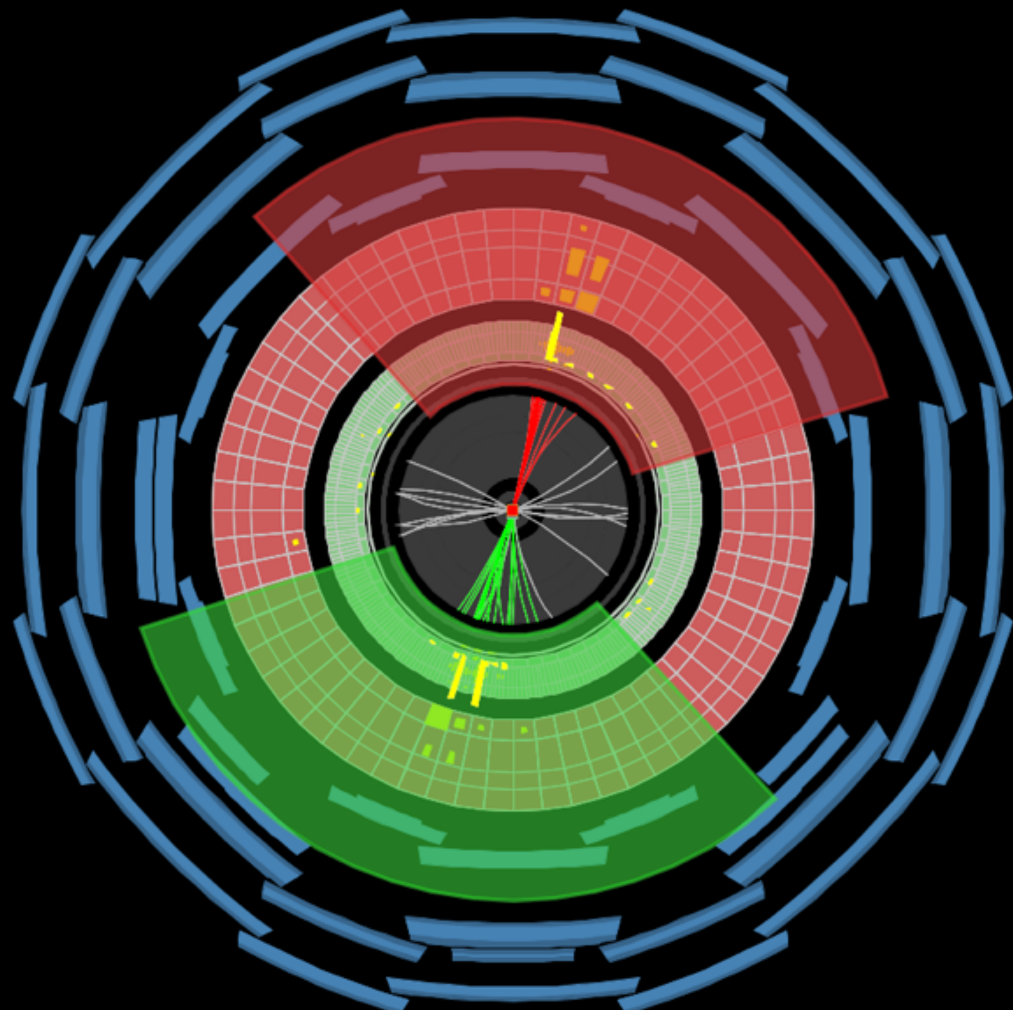
Searching for new particles decaying into boosted W bosons requires **looking at the radiation pattern inside jets**

momentum transverse to the beam (p_T)

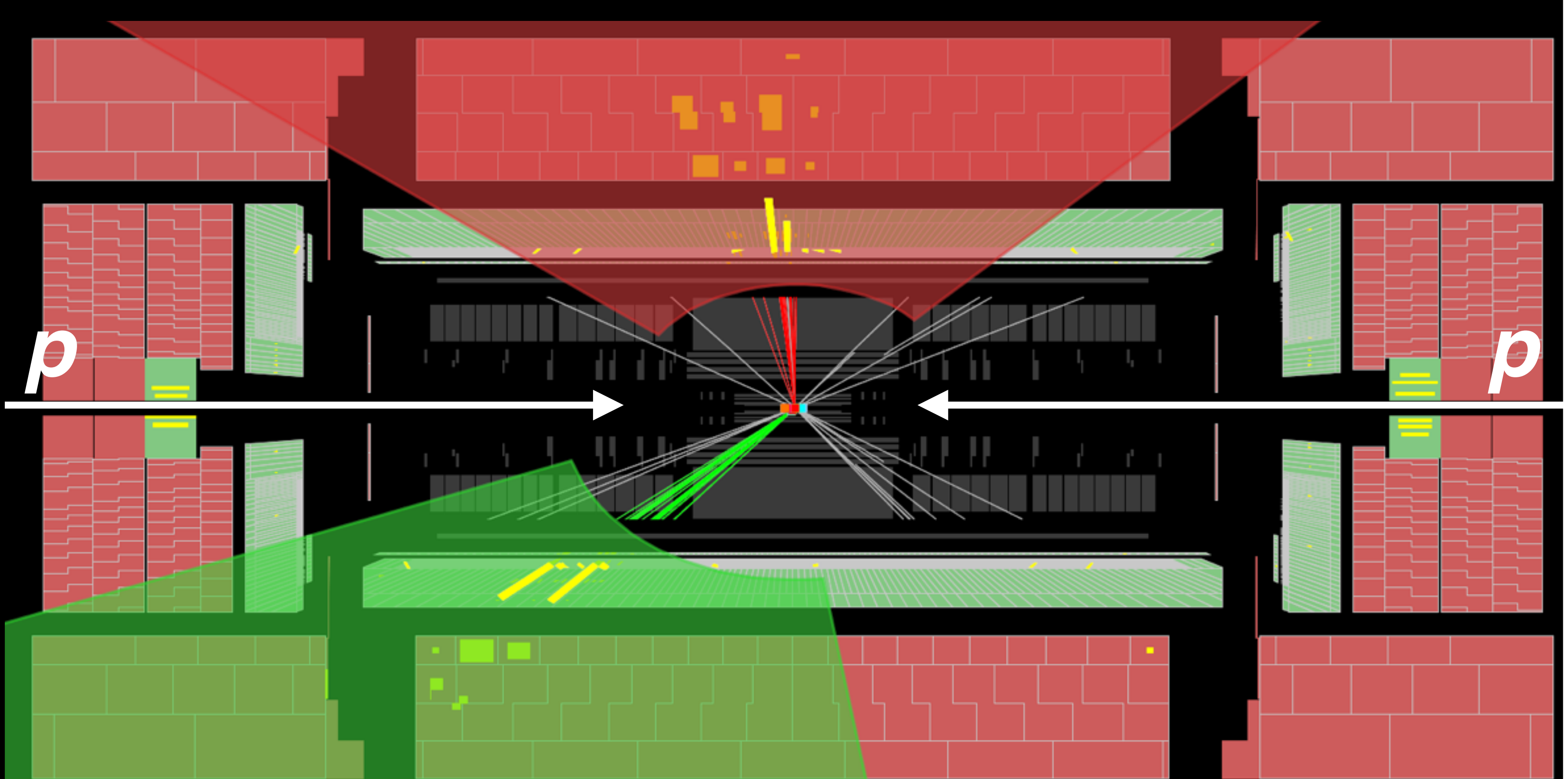


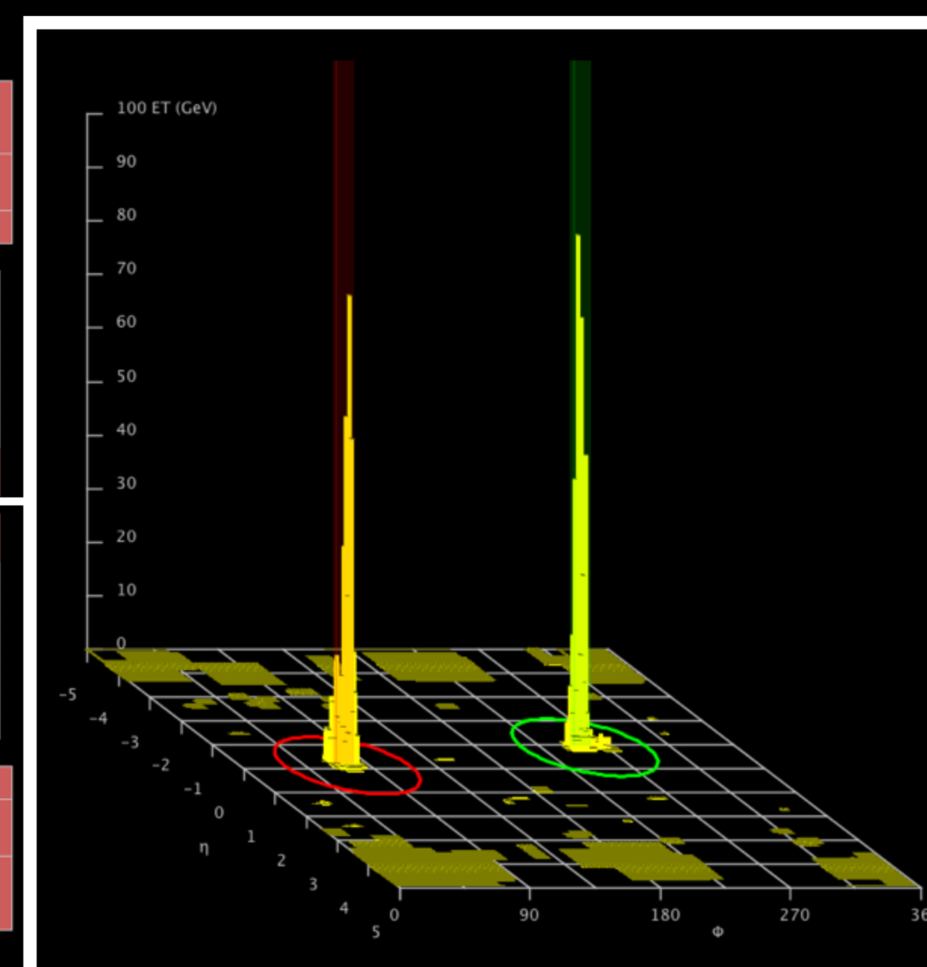


Up next: jet images



like a digital image!

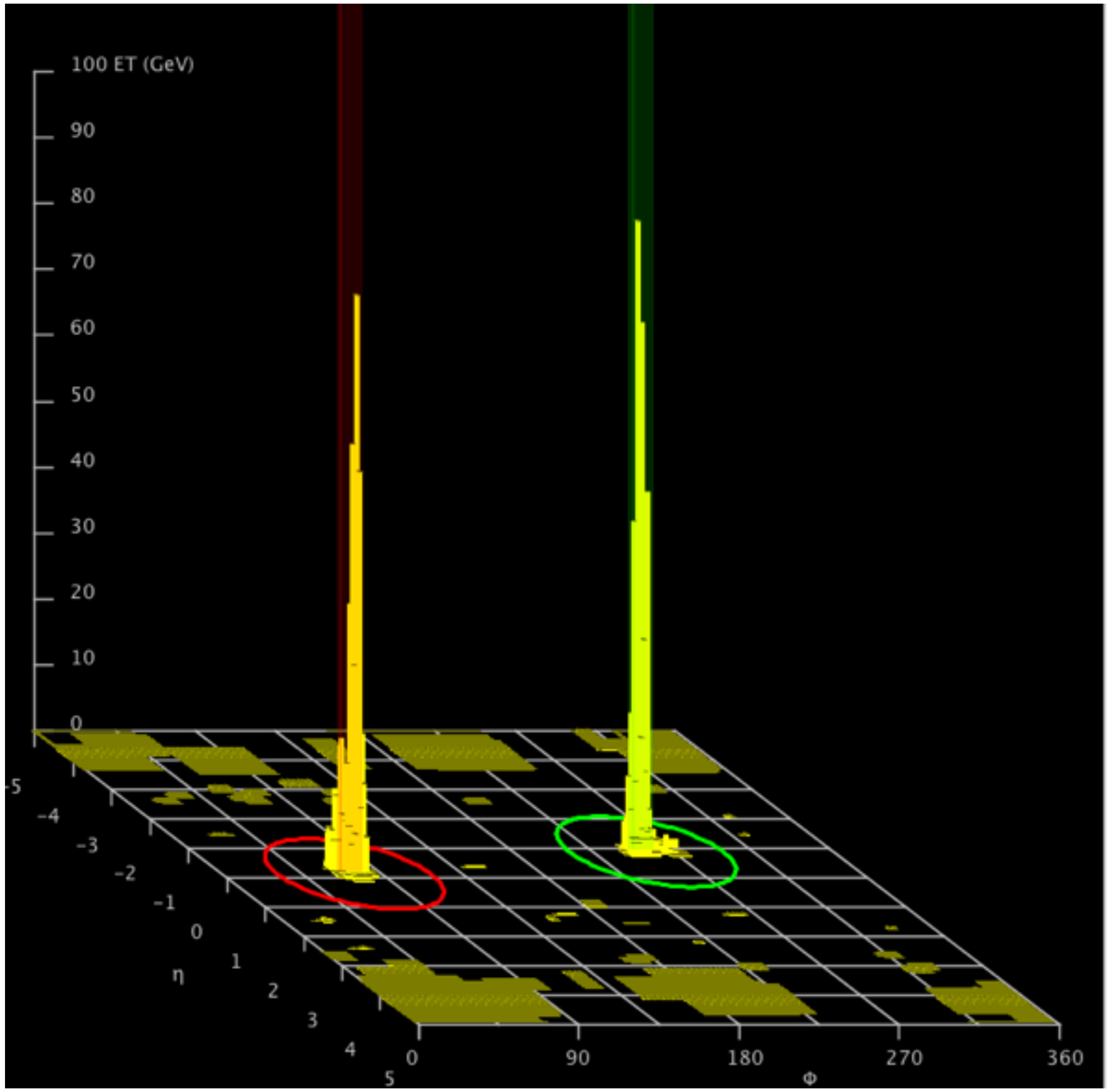




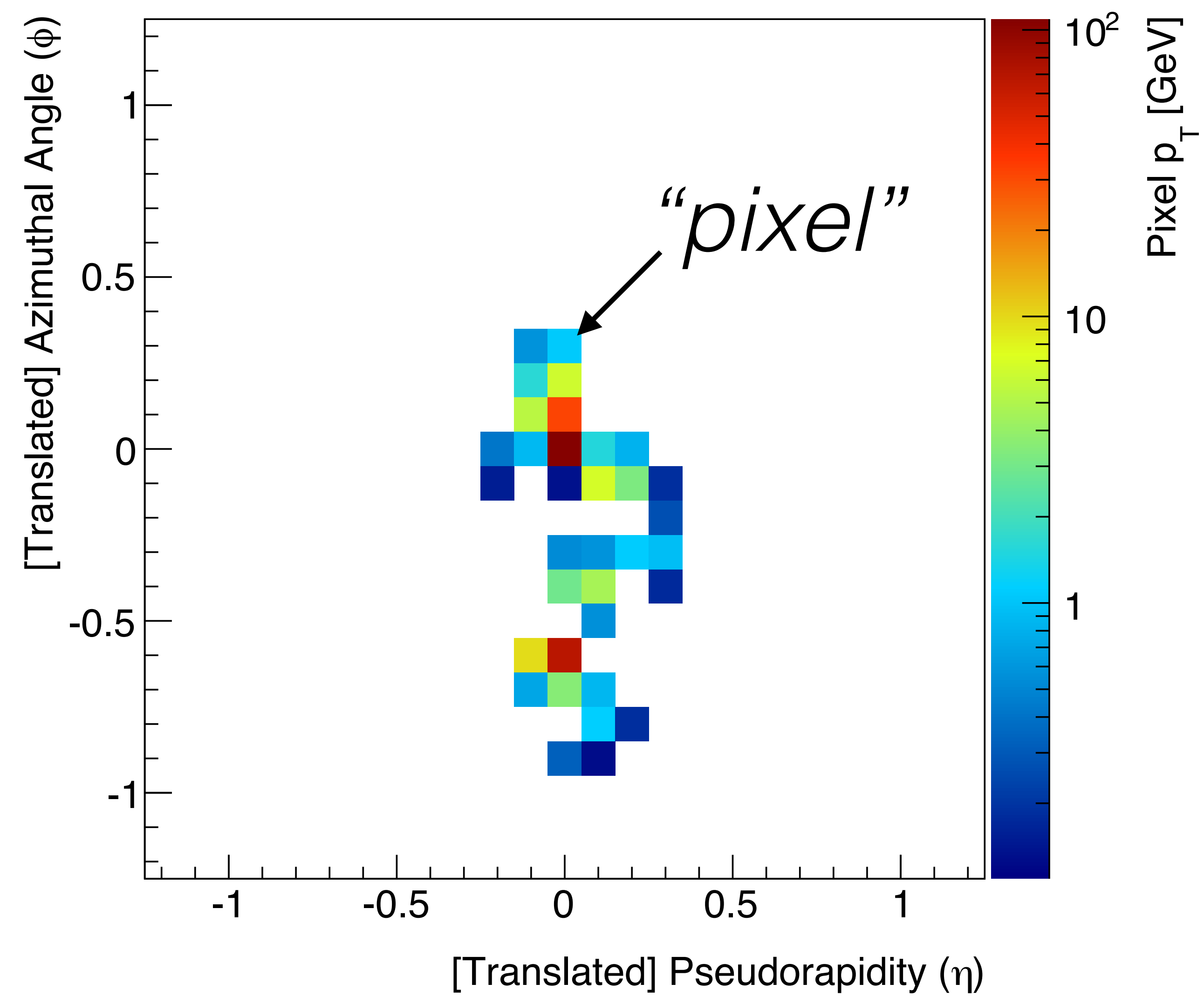
N.B. this is not the only way to represent a jet - more on that later

the Jet Image

J. Cogan et al. JHEP 02 (2015) 118



Boosted W



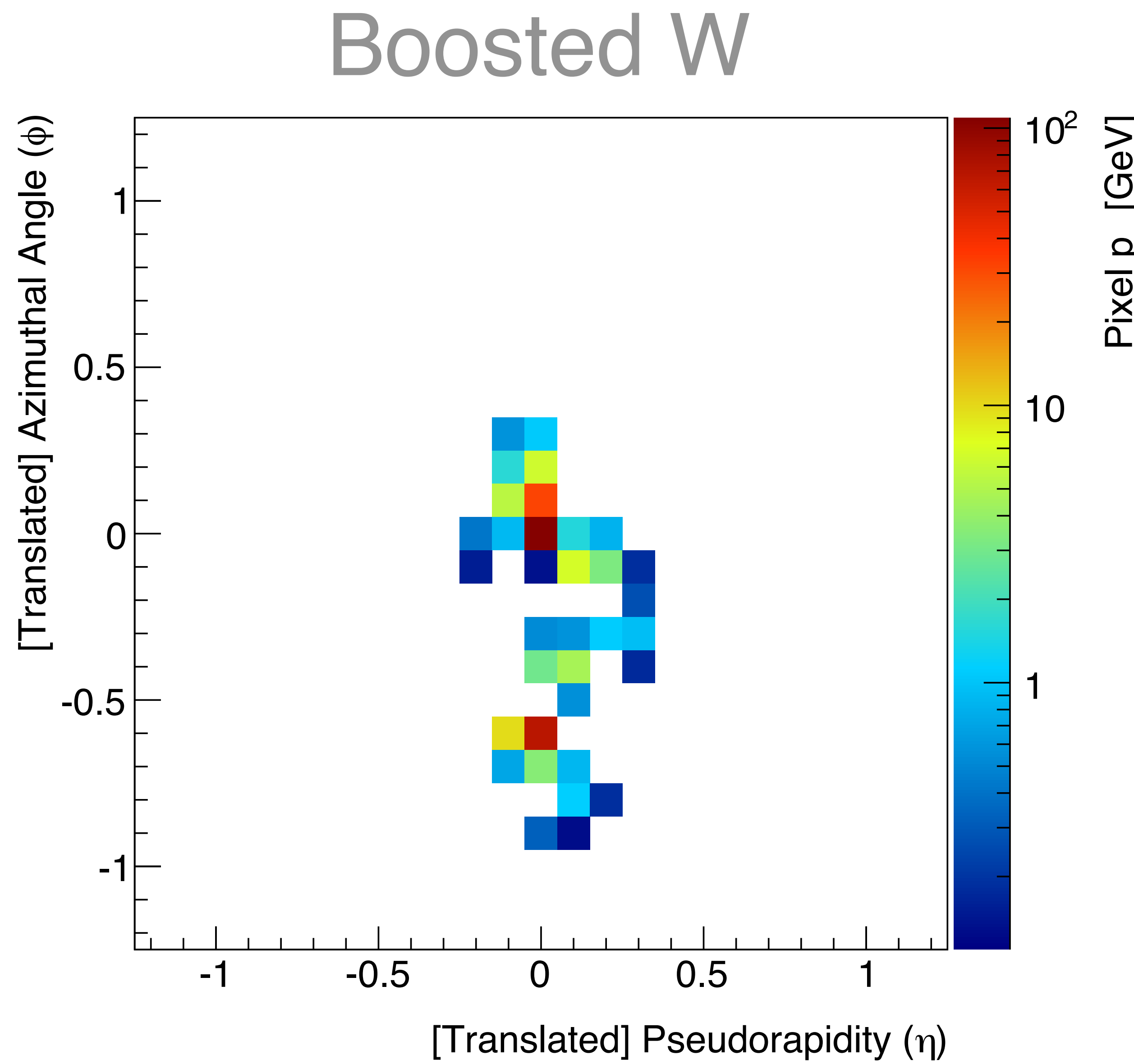
L. de Oliveira, et al., Comp. and Software for Big Science (2017) 1

nothing like a 'natural' image!

the Jet Image

J. Cogan et al. JHEP 02 (2015) 118

Credit: Peter G Trimming (Wikipedia)



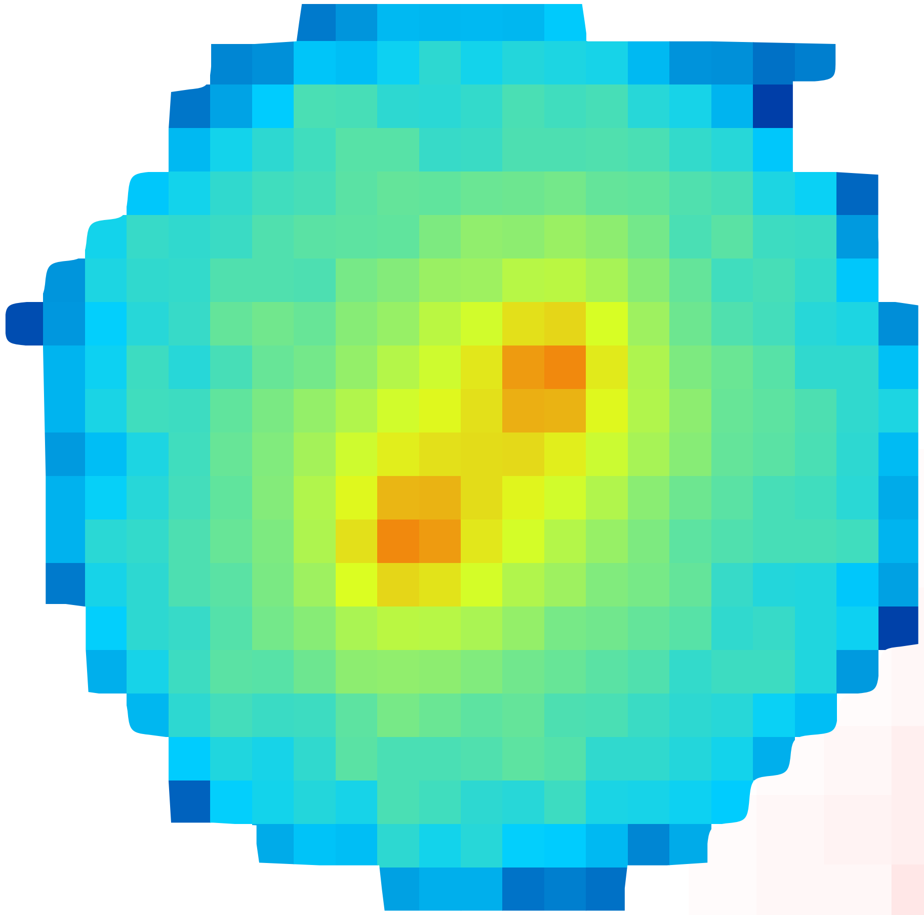
no smooth edges, clear features, low occupancy (number of hit pixels)

L. de Oliveira, et al., Comp. and Software for Big Science (2017) 1

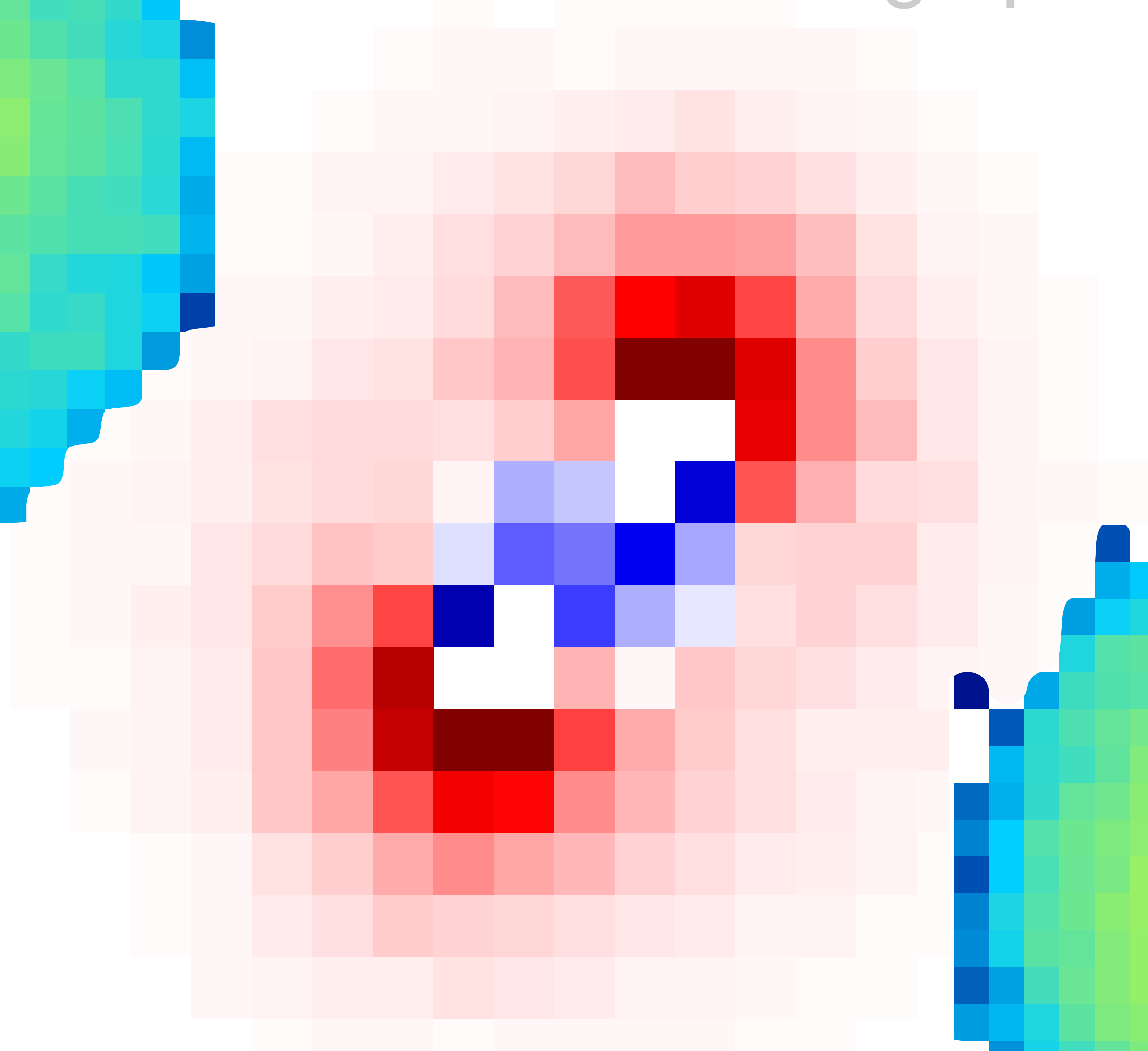
Why images?

Can directly visualize physics

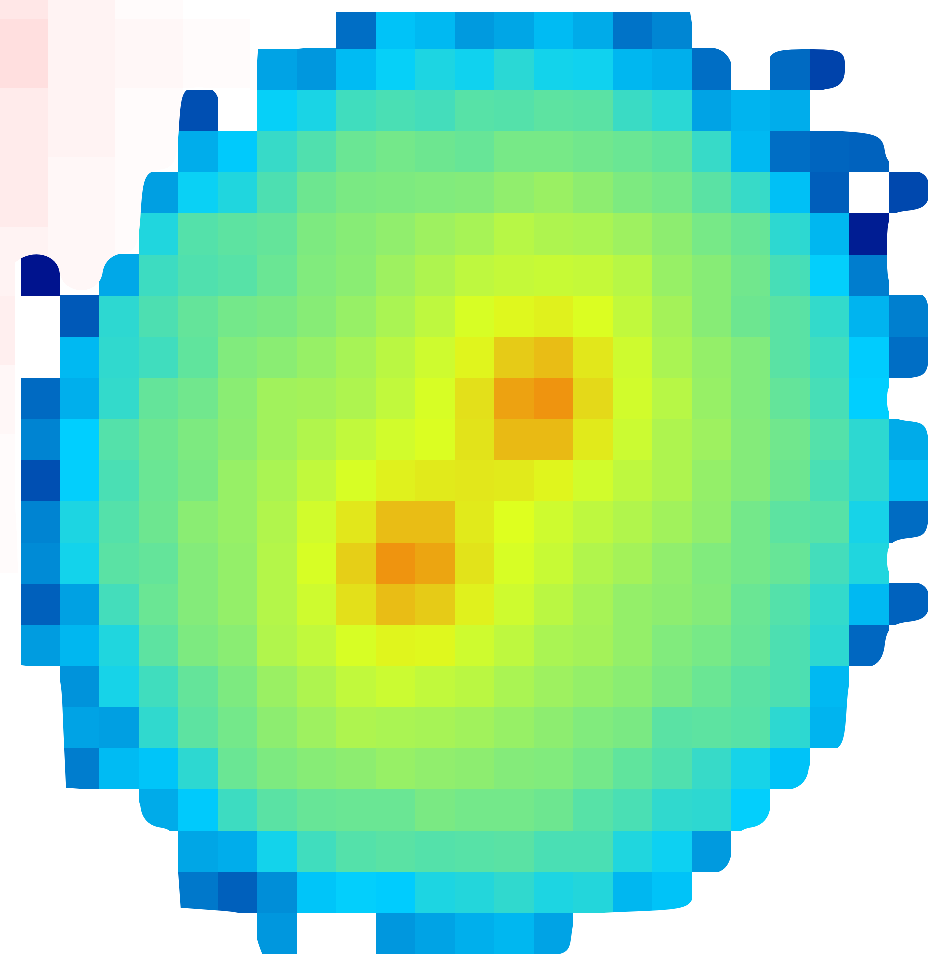
and we can benefit from the
extensive image processing literature



W ⇢ $q\bar{q}$



g ⇢ $q\bar{q}$

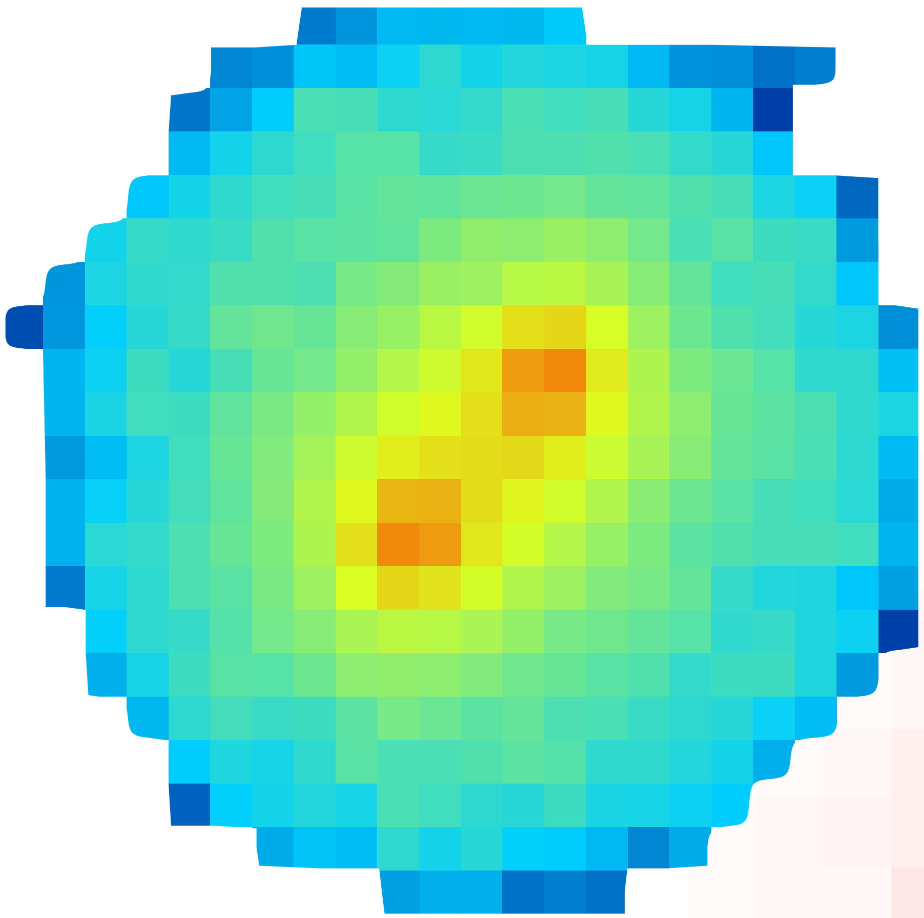


there is information encoded in the
physical distance between pixels

Why images?

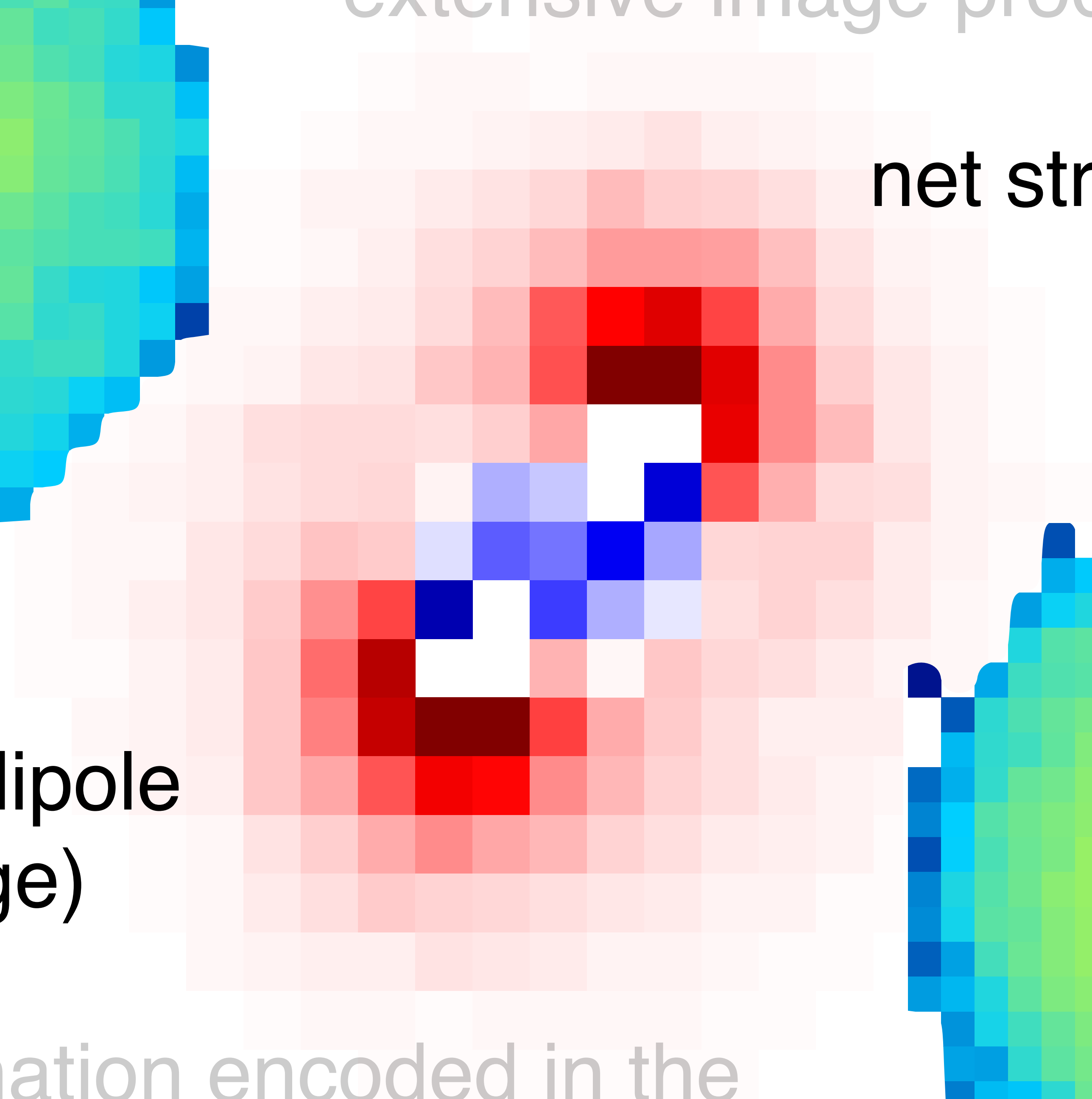
Can directly visualize physics

and we can benefit from the extensive image processing literature



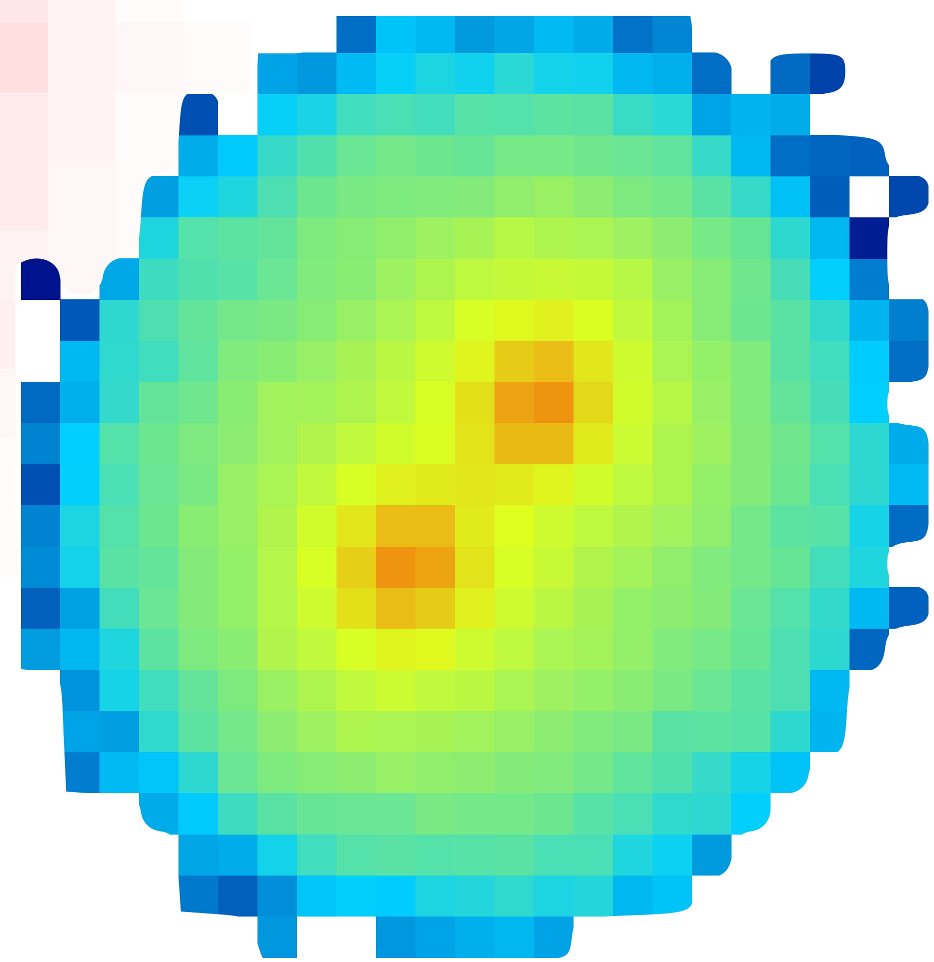
$\mathbf{W} \dashrightarrow q\bar{q}$

radiates like a dipole (no net charge)



net strong-force charge

$\mathbf{g} \dashrightarrow q\bar{q}$

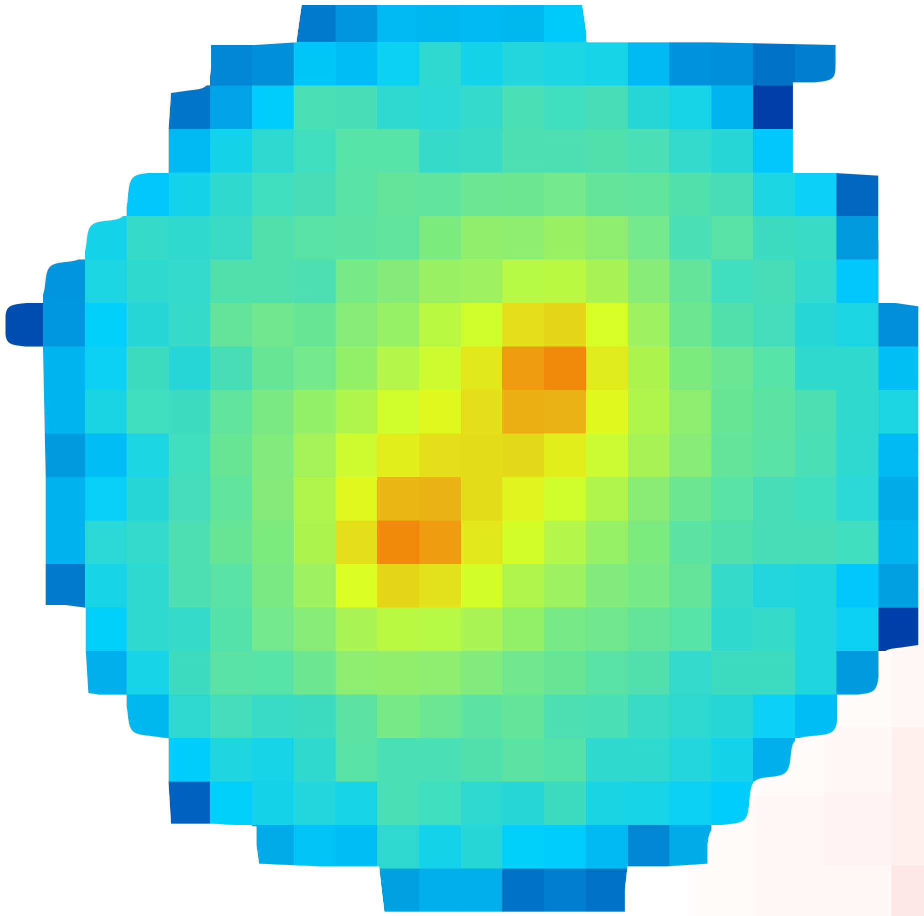


there is information encoded in the physical distance between pixels

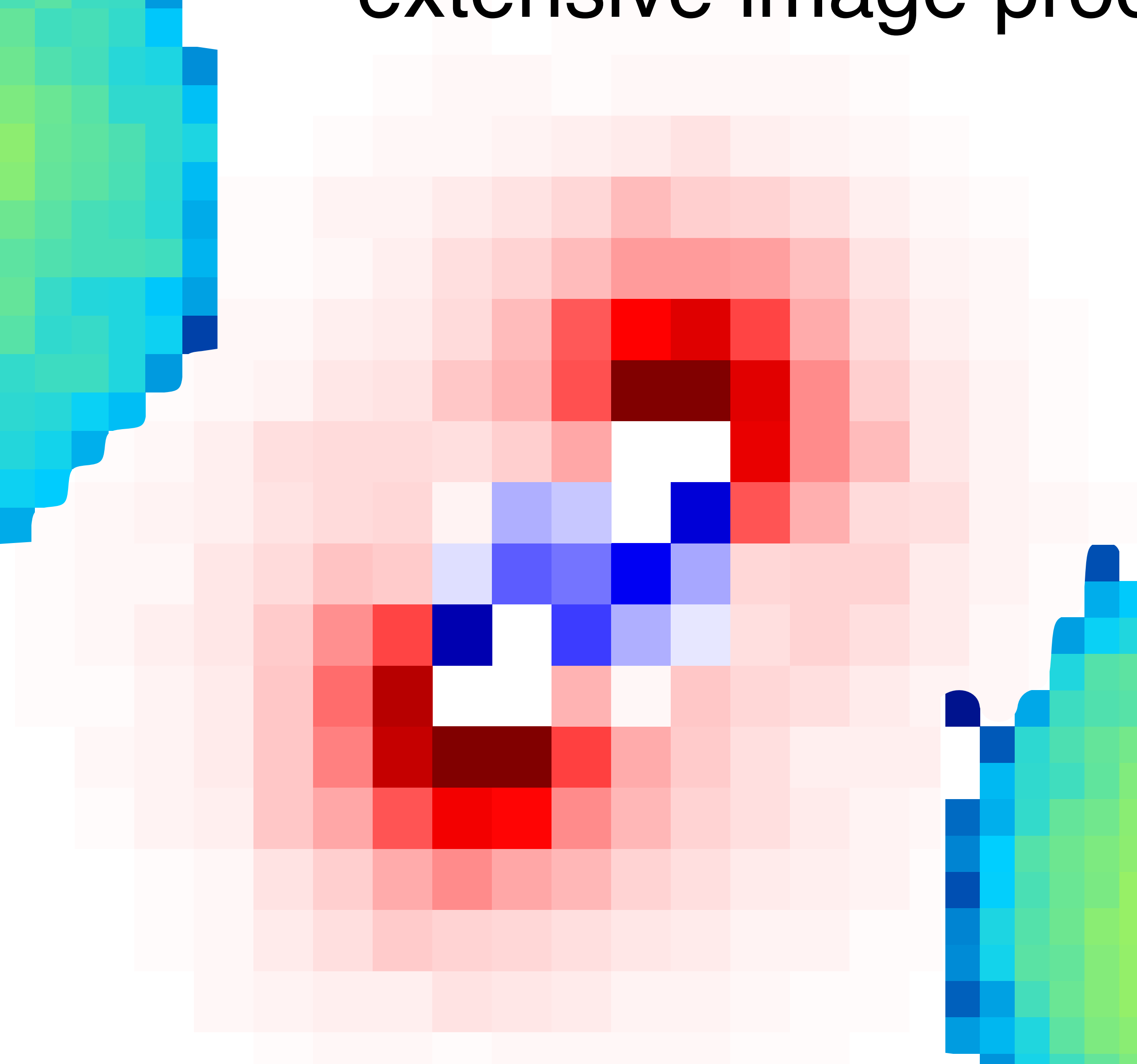
Why images?

Can directly visualize physics

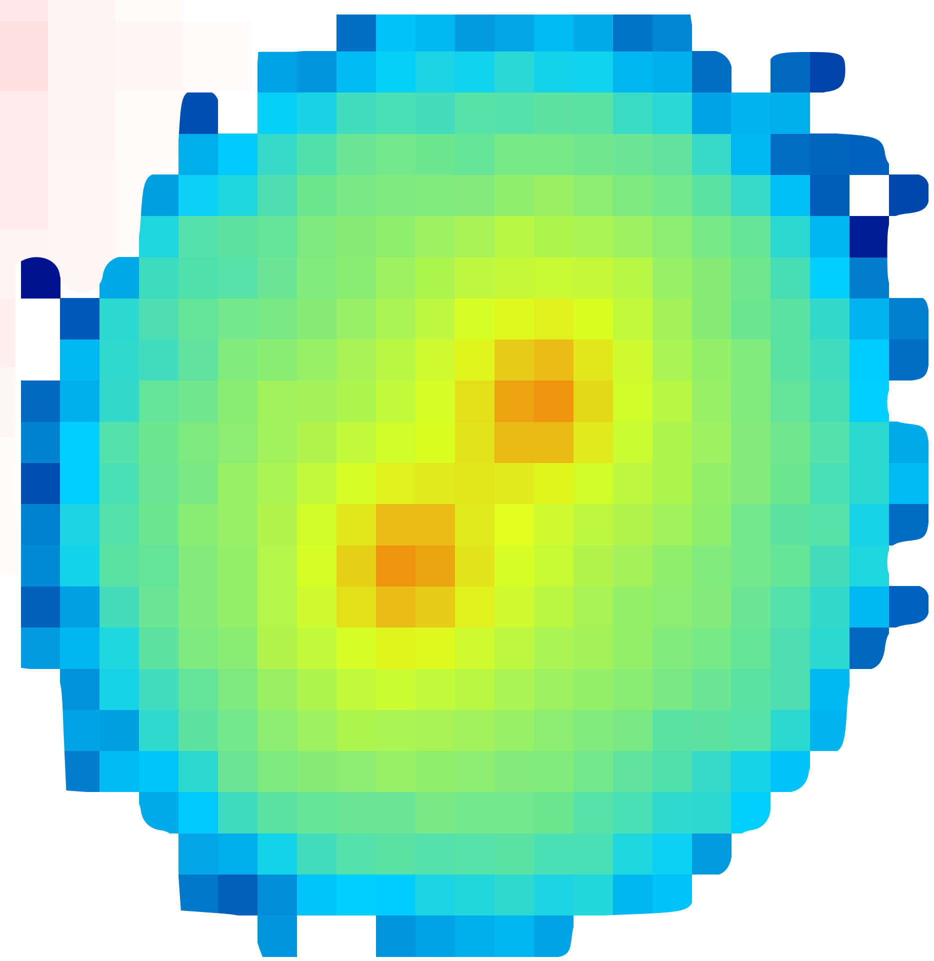
and we can benefit from the extensive image processing literature



$W \dashrightarrow q\bar{q}$



$g \dashrightarrow q\bar{q}$

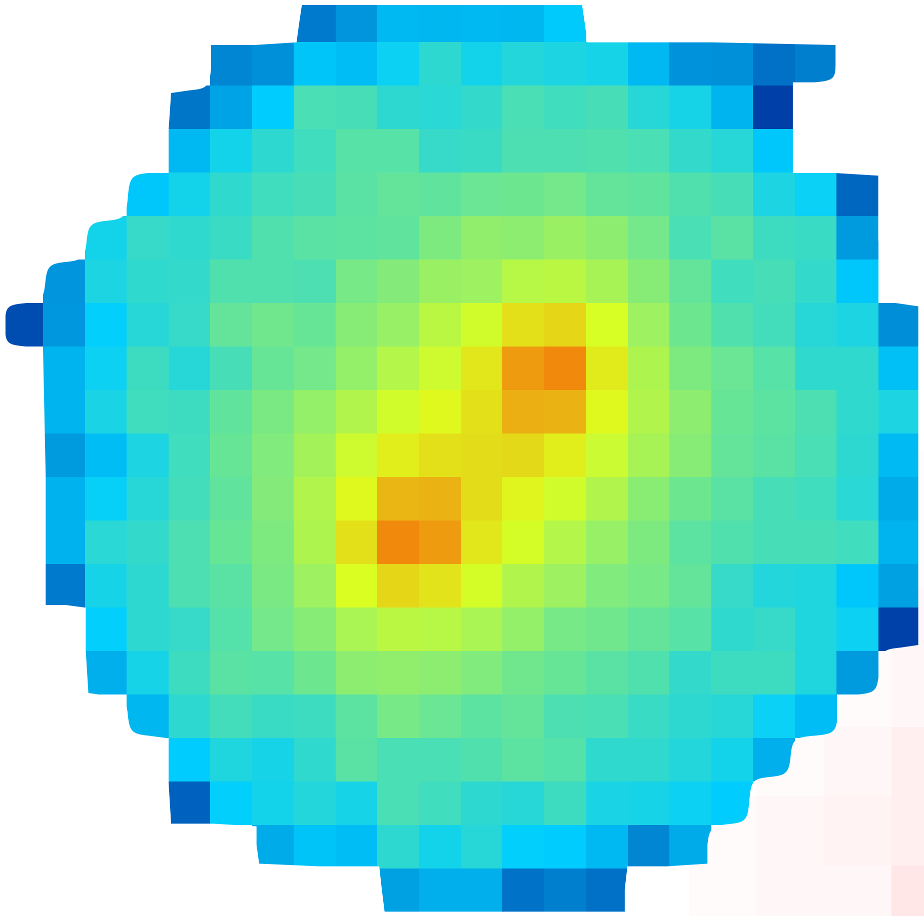


there is information encoded in the physical distance between pixels

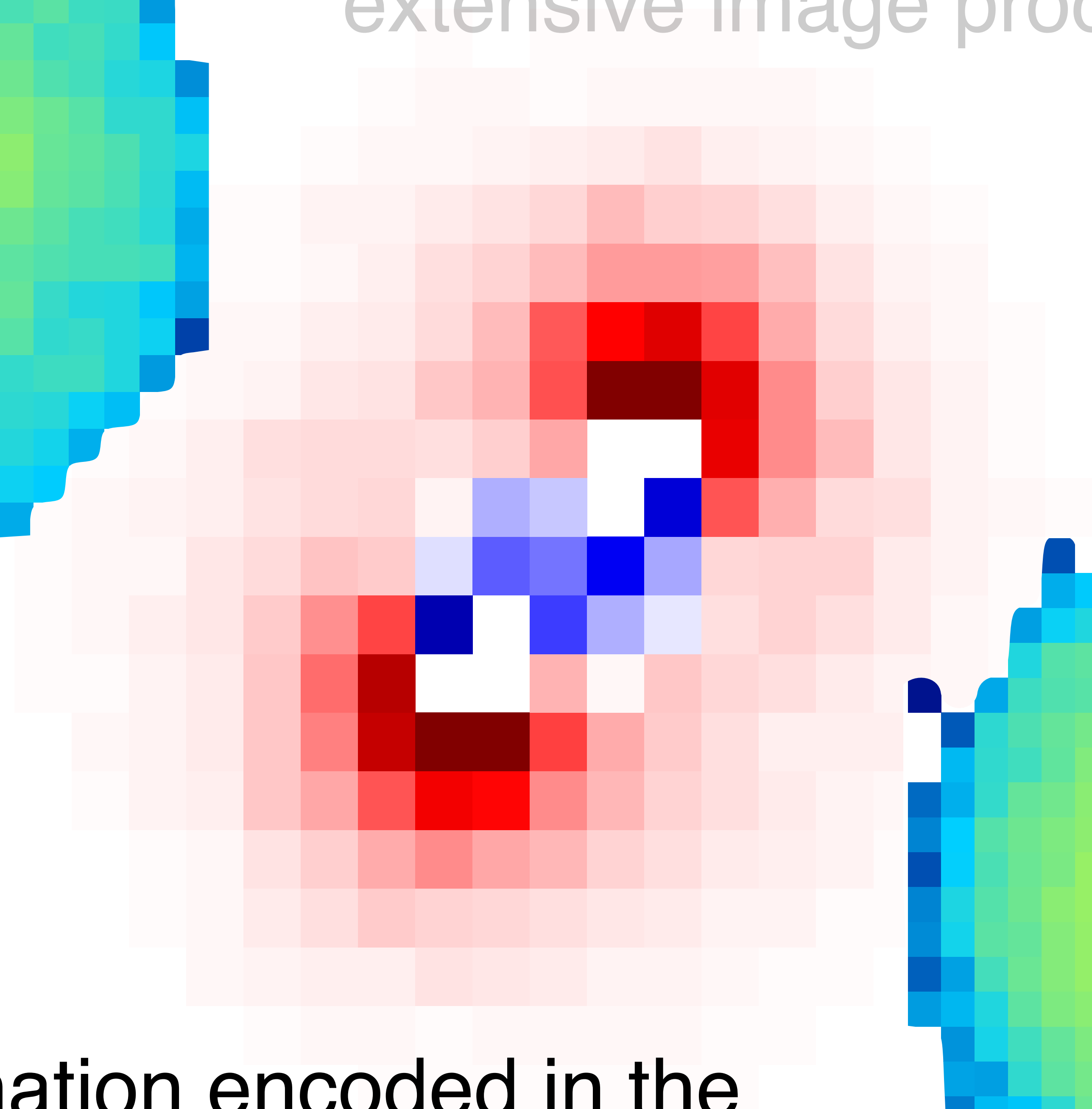
Why images?

Can directly visualize physics

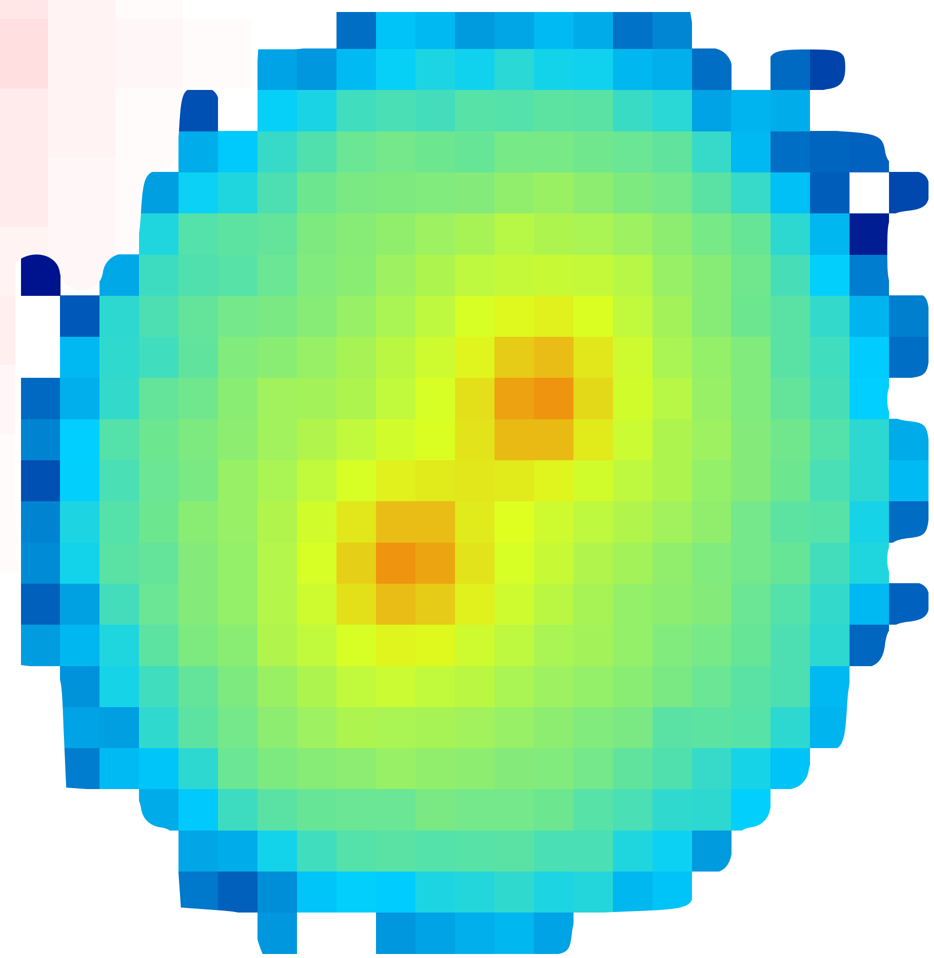
and we can benefit from the extensive image processing literature



W ⇢ $q\bar{q}$



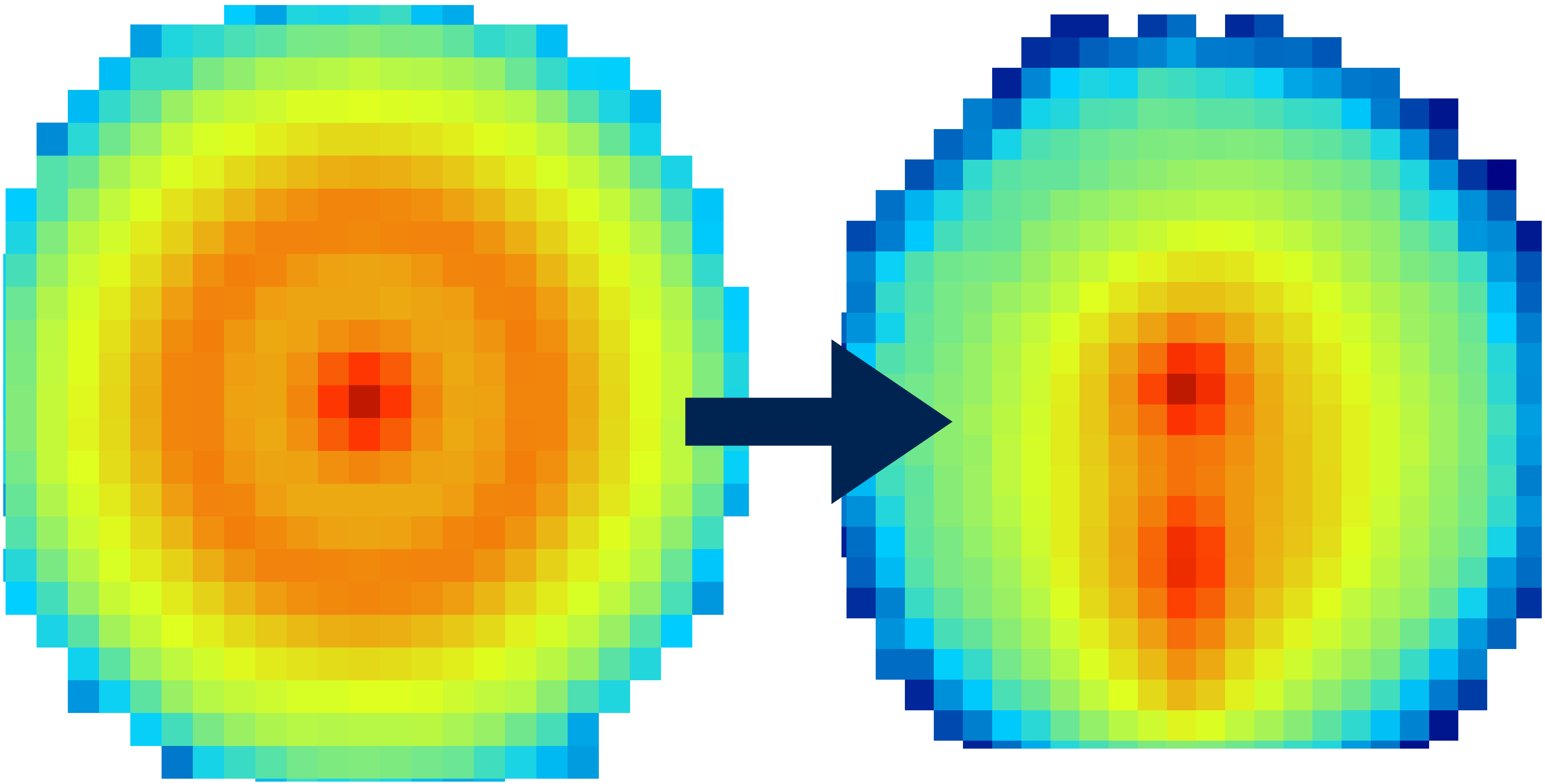
g ⇢ $q\bar{q}$



there is information encoded in the physical distance between pixels

Pre-processing & spacetime symmetries

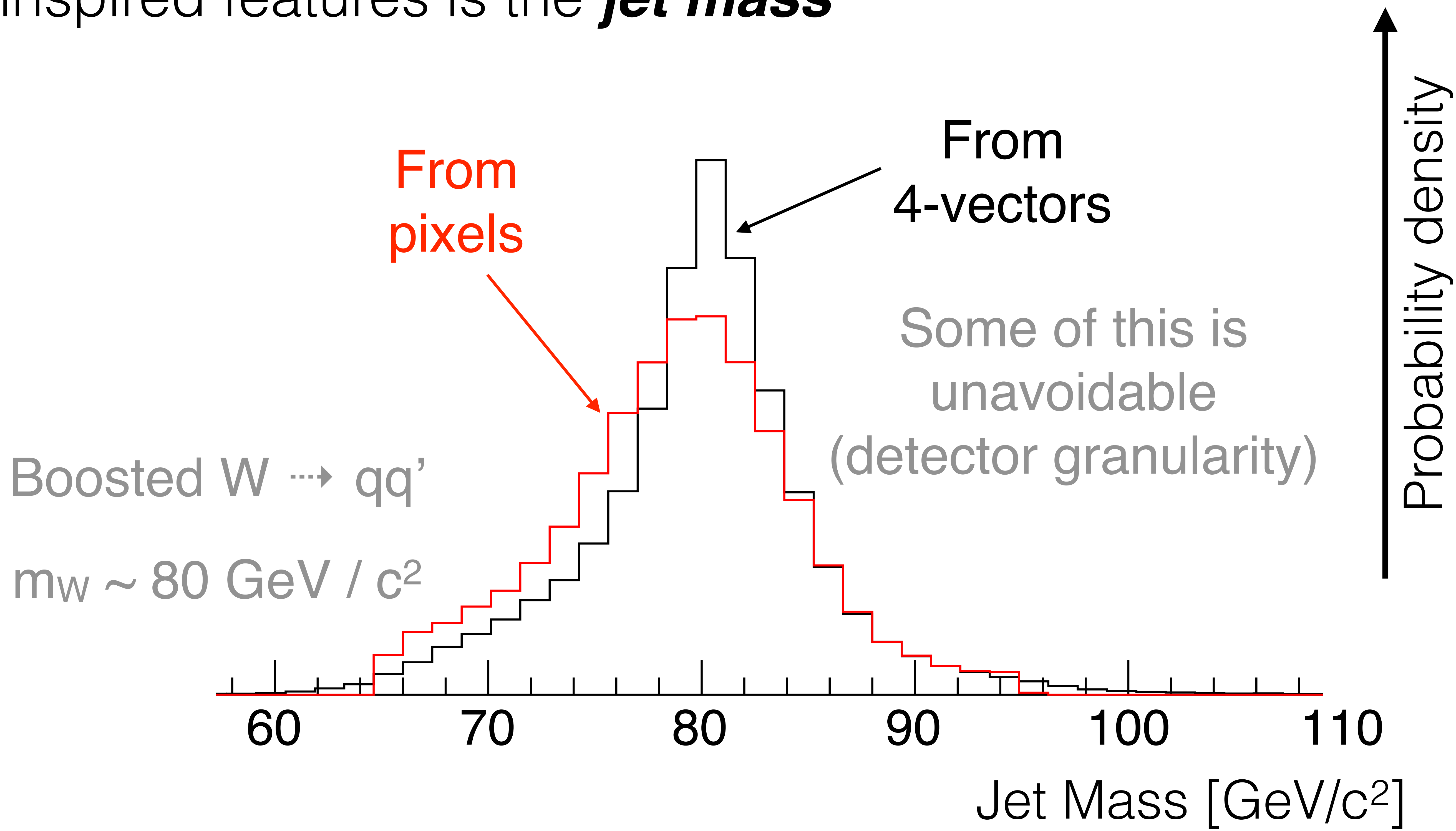
One of the first typical steps is pre-processing



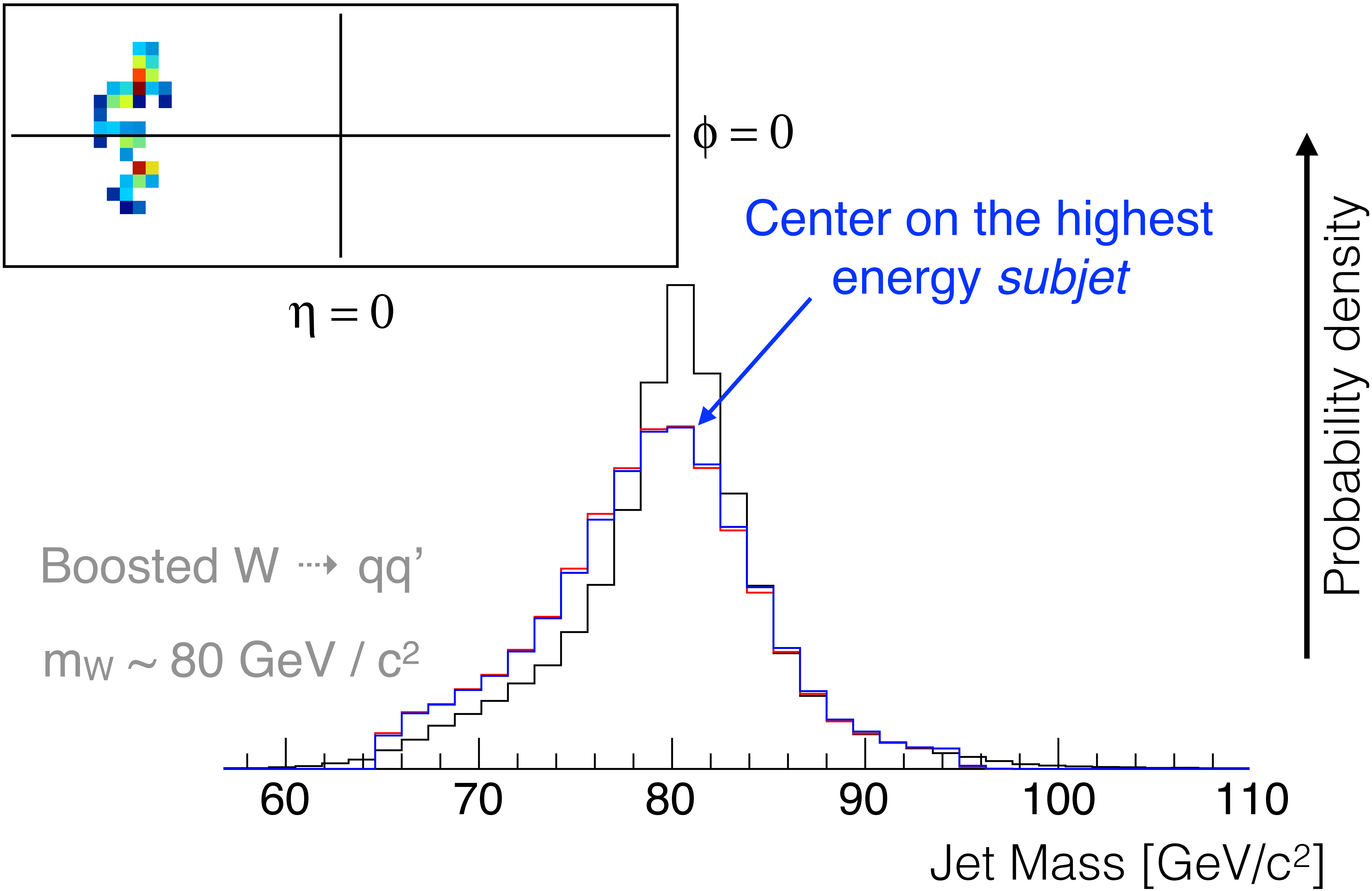
Can help to learn faster & smarter; but must be careful!

Pre-processing & spacetime symmetries

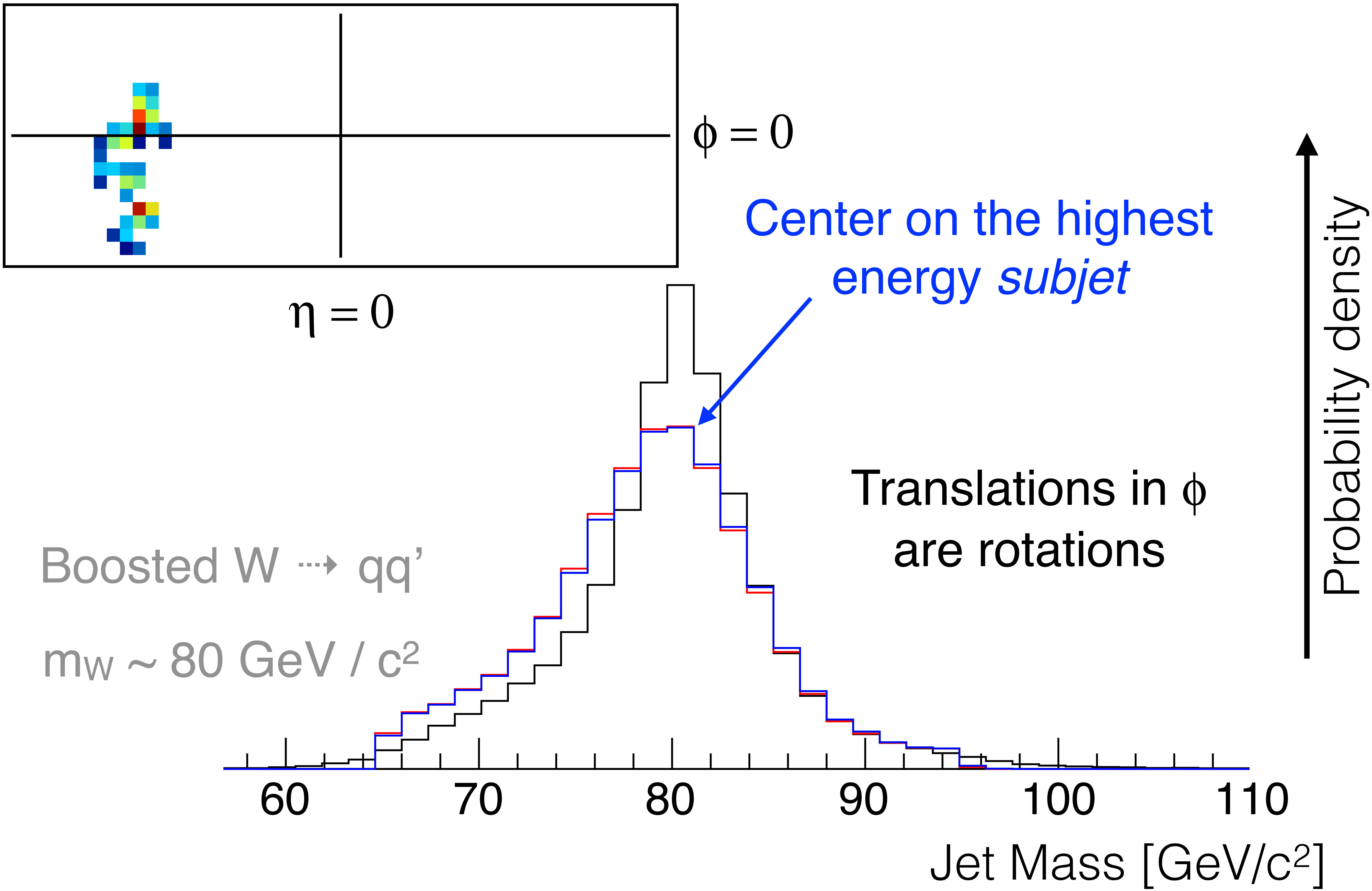
One of the most useful physics-inspired features is the ***jet mass***



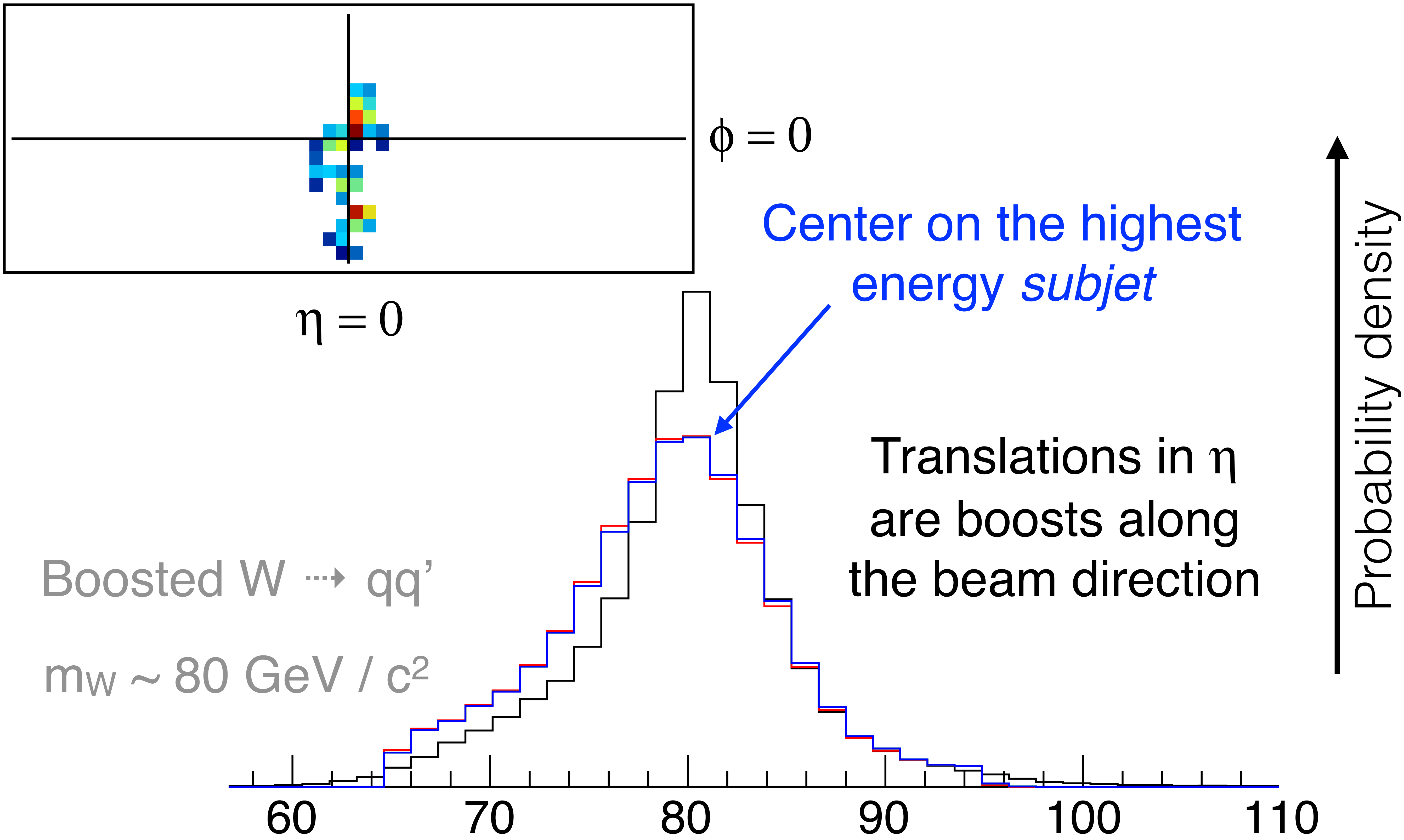
Pre-processing & spacetime symmetries



Pre-processing & spacetime symmetries



Pre-processing & spacetime symmetries



Boosted W ⇢ qq’

$m_W \sim 80$ GeV / c^2

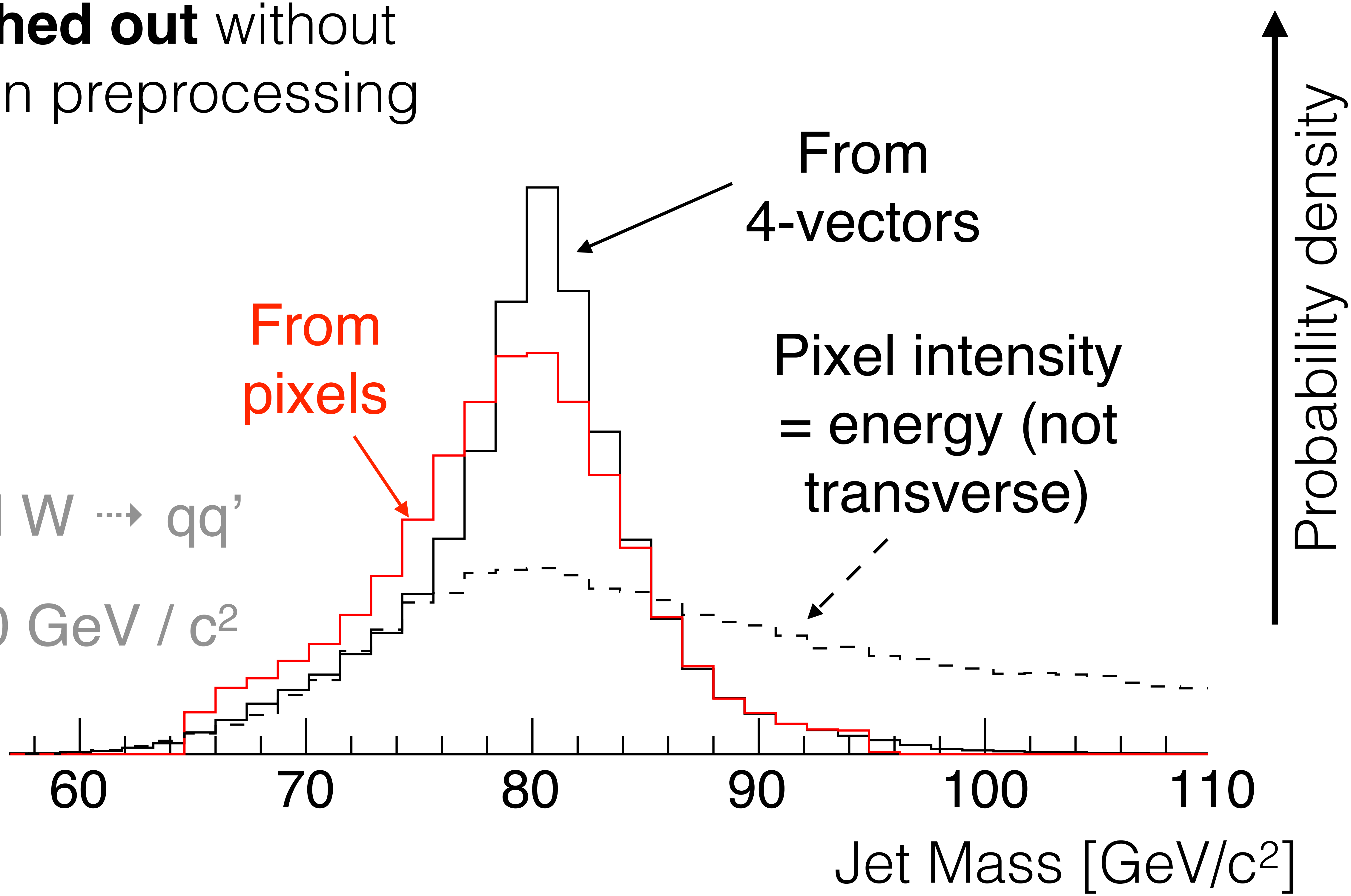
N.B. pixel intensity is p_T

Pre-processing & spacetime symmetries

Information can be **washed out** without care in preprocessing

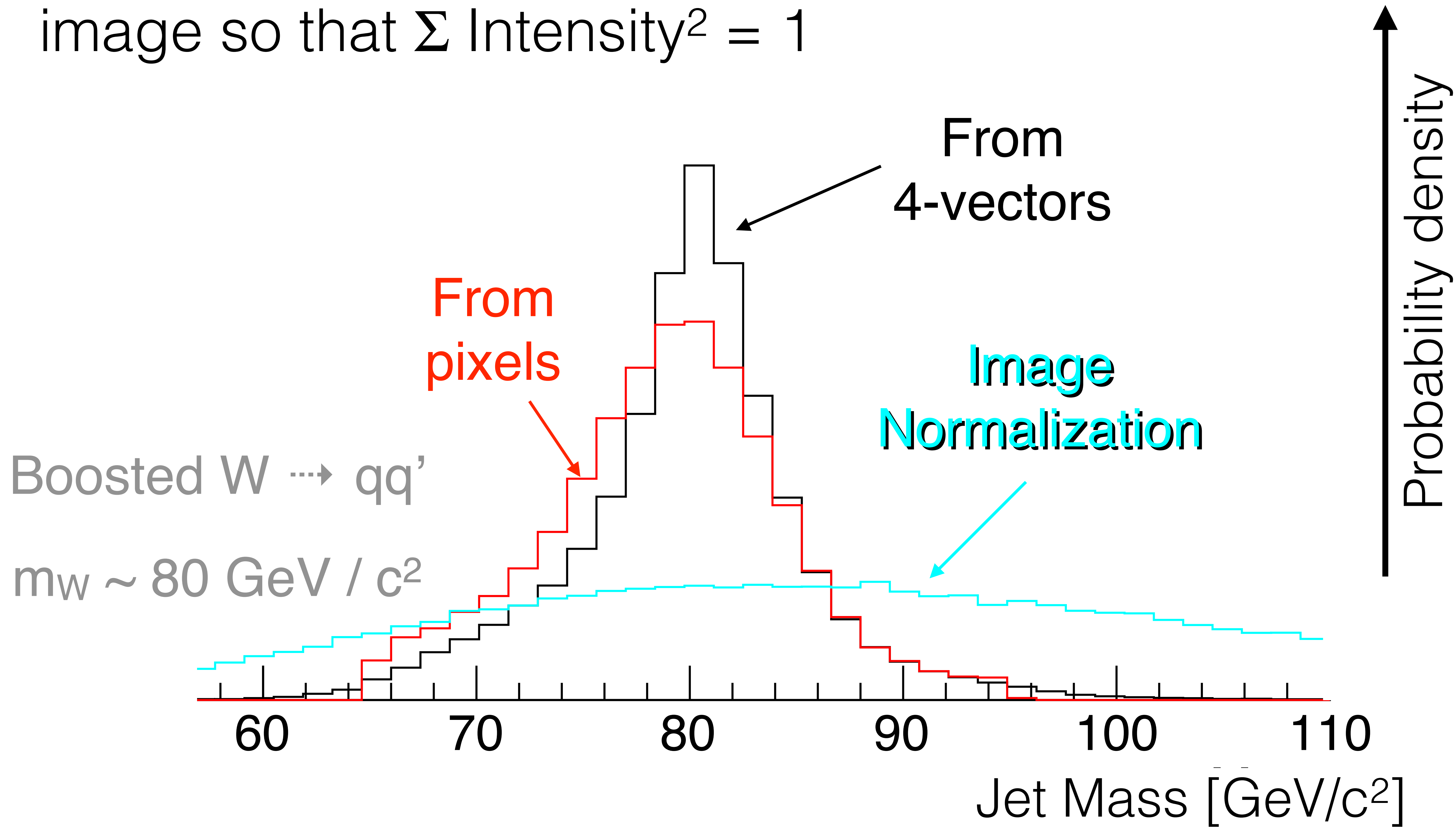
Boosted W ⇢ qq’

$m_W \sim 80$ GeV / c^2

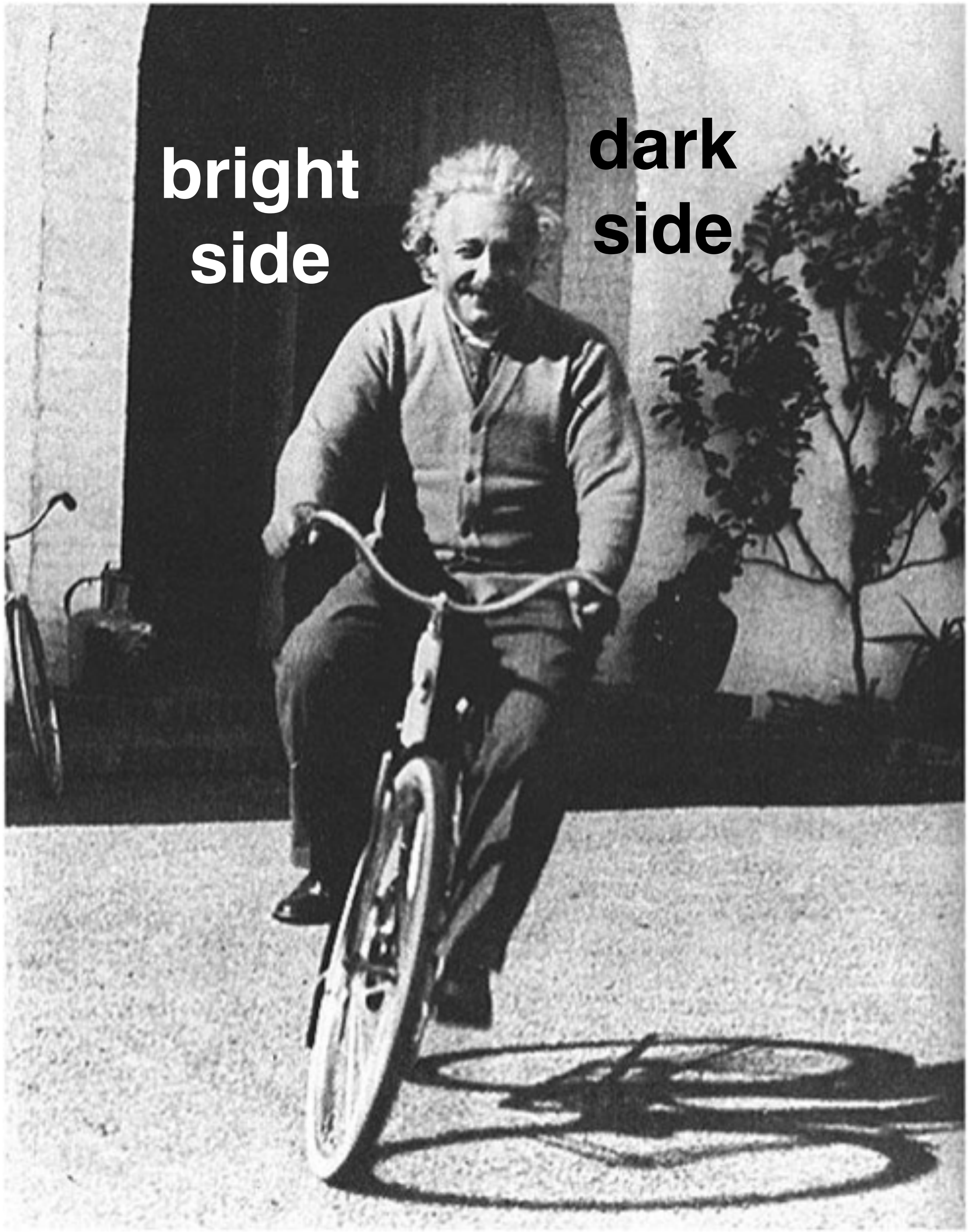


Pre-processing & spacetime symmetries

It is common to normalize each image so that Σ Intensity^2 = 1



Intuition via analogy *why normalization can hurt*



In both pictures, total intensity of Einstein's face is about the same.

However, his face's **image mass** is quite different!

Photos from: http://mentalfloss.com/article/49222/11-unserious-photos-albert-einstein

Intuition via analogy *why normalization can hurt*

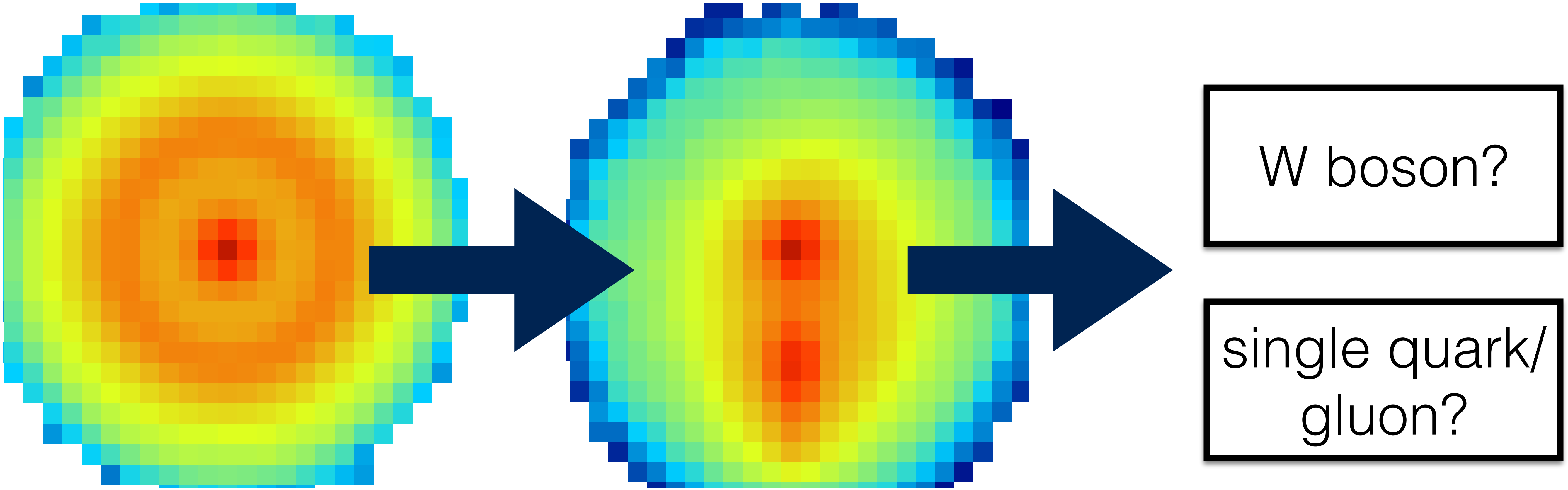
In standard computer vision, you likely don't want to be sensitive to this! …not the case for jet images!

In both pictures, total intensity of Einstein's face is about the same.

However, his face's **image mass** is quite different!

Photos from: http://mentalfloss.com/article/49222/11-unserious-photos-albert-einstein

Now, with a carefully processed image, we can ask: where did this jet come from?

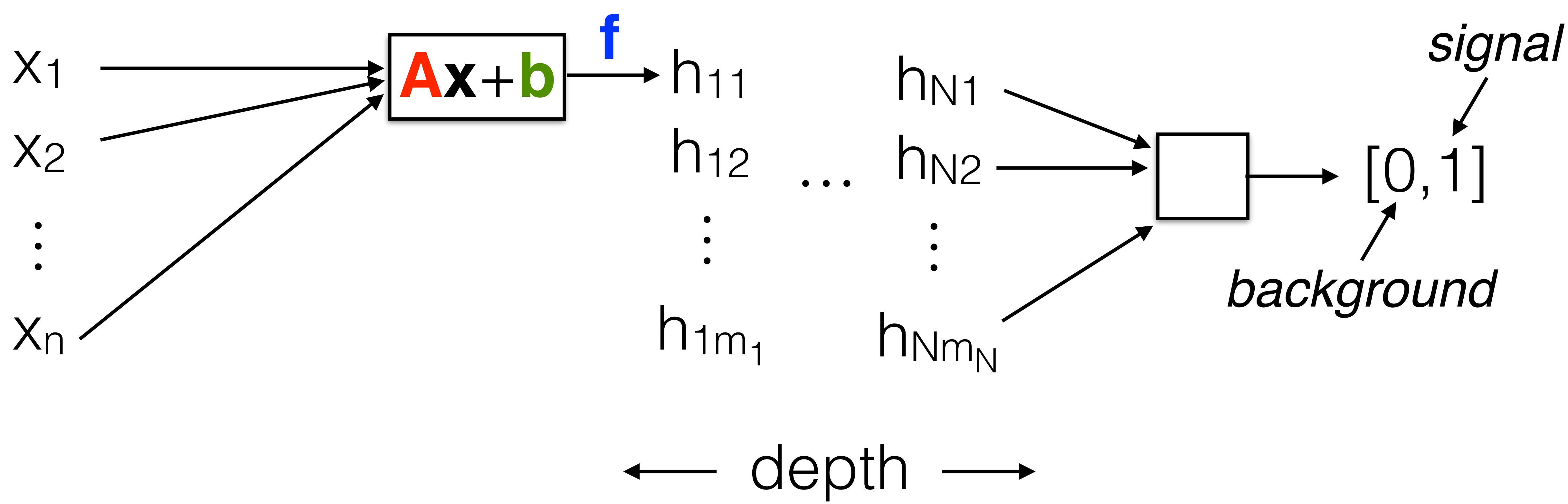


ultimate classification is achieved with modern machine learning using **all pixels as input**!

Modern Deep NN's for Classification

Neural Network: composition of functions **f**(**Ax**+**b**) for inputs **x** (features) matrix **A** (weights), bias **b**, non-linearity **f**.

N.B. I'm not mentioning biology - there may be a vague resemblance to parts of the brain, but that is not what modern NN's are about.



Modern Deep NN's for Classification

Neural Network: composition of functions $\mathbf{f}(\mathbf{A}\mathbf{x}+\mathbf{b})$ for inputs $\mathbf{x}$ (features) matrix $\mathbf{A}$ (weights), bias $\mathbf{b}$, non-linearity $\mathbf{f}$.

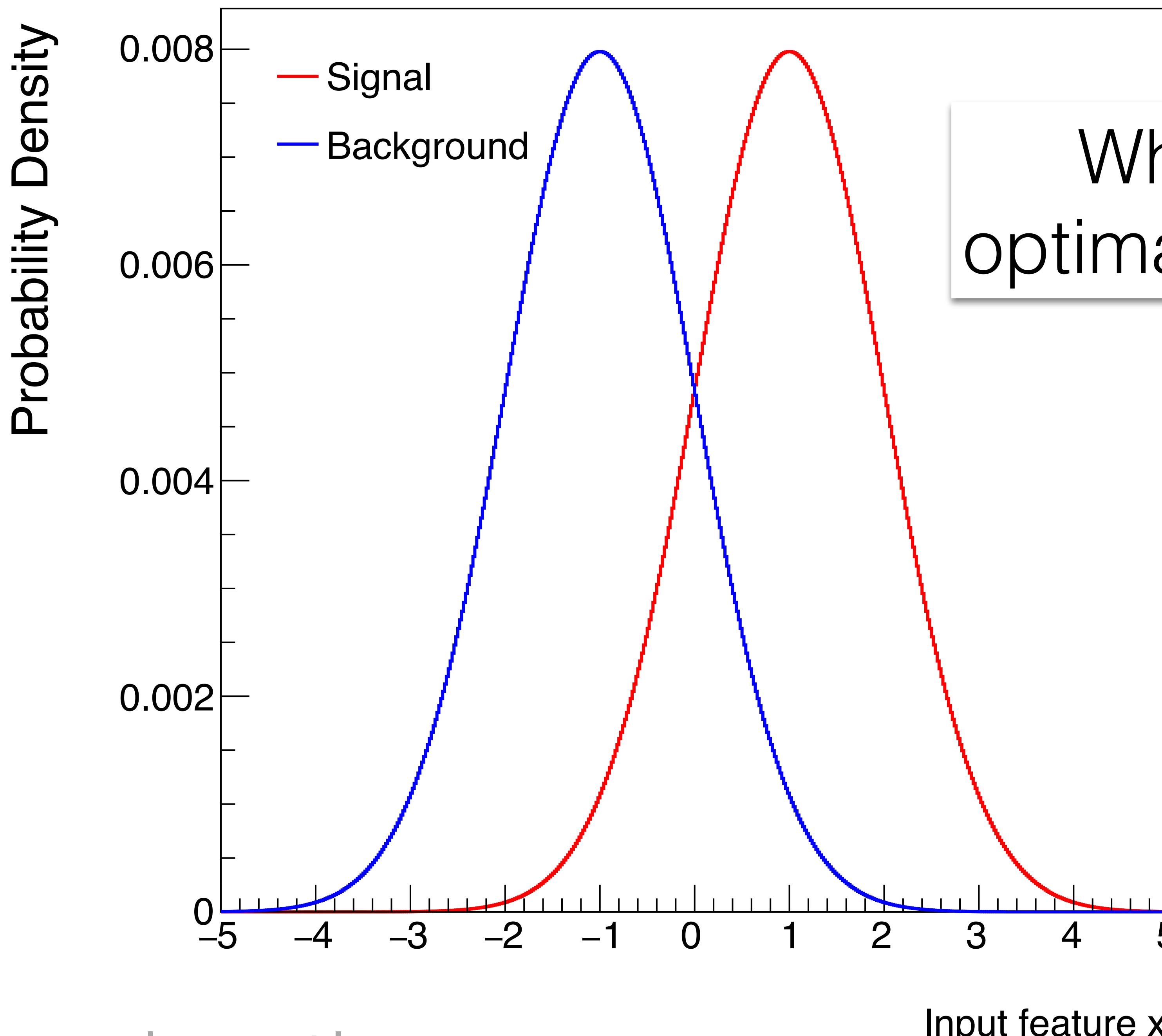
Fact: NN's can approximate "any" function.

Why useful?

For classification, there is an optimal function to learn: the likelihood ratio, $LL(x) = p_S(x) / p_B(x)$.

Getting into the machine's mind

Let's consider an important special case:
binary classification in 1D



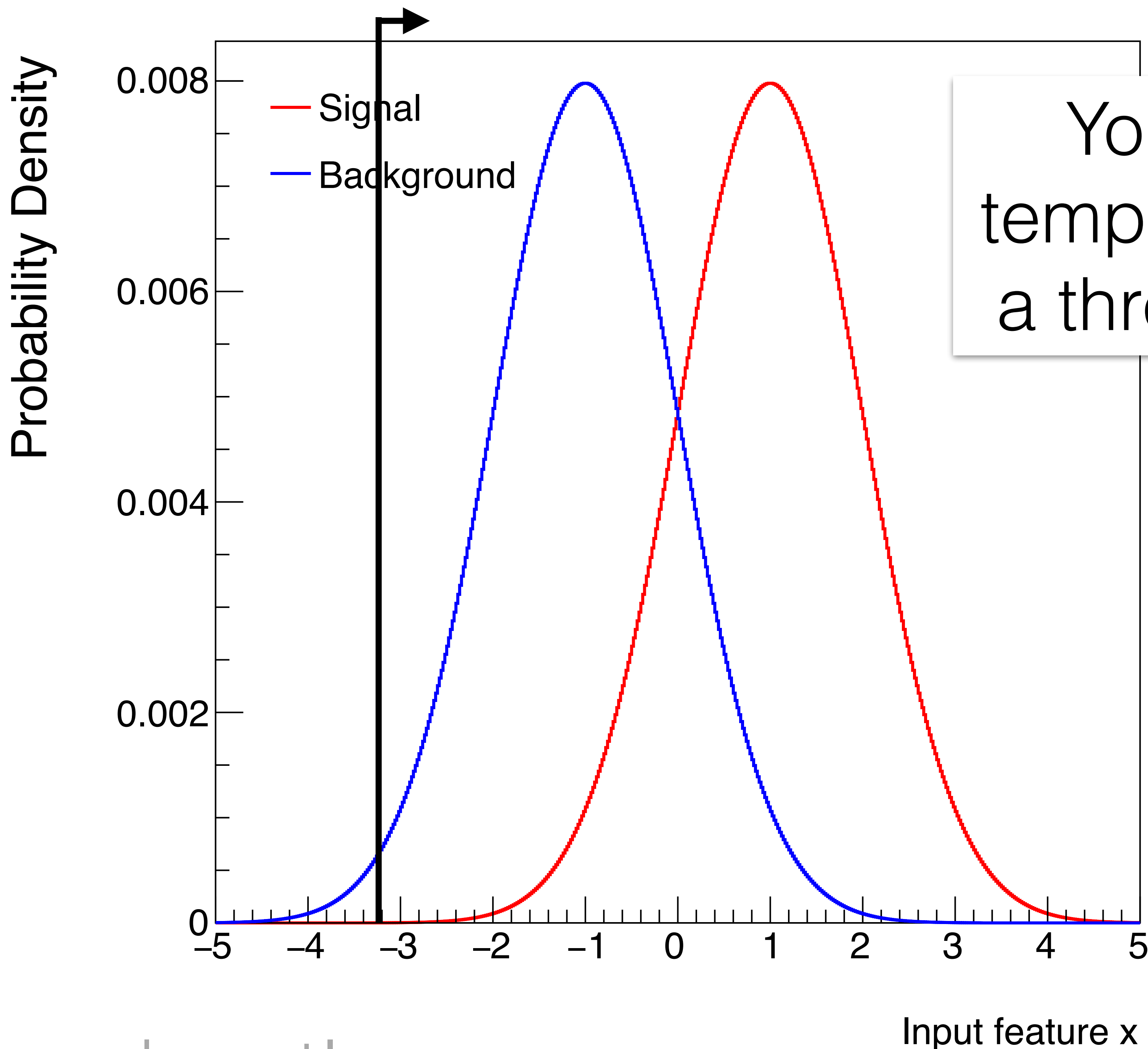
What is the optimal classifier?

could be e.g. the jet mass

⇨ Try this example out!

Getting into the machine's mind

Let's consider an important special case:
binary classification in 1D

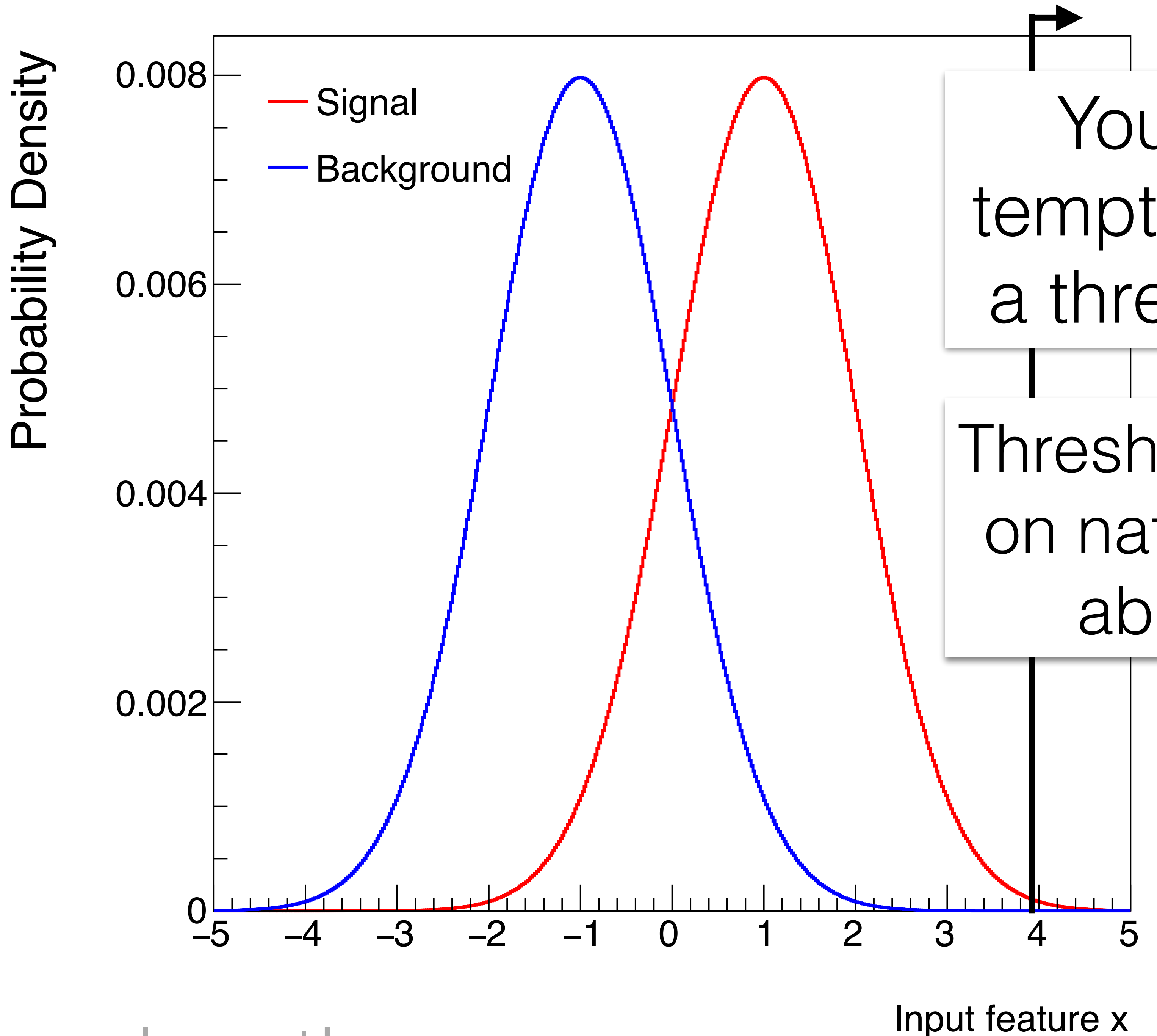


You may be tempted to place a threshold on x

⇨ Try this example out!

Getting into the machine's mind

Let's consider an important special case:
binary classification in 1D

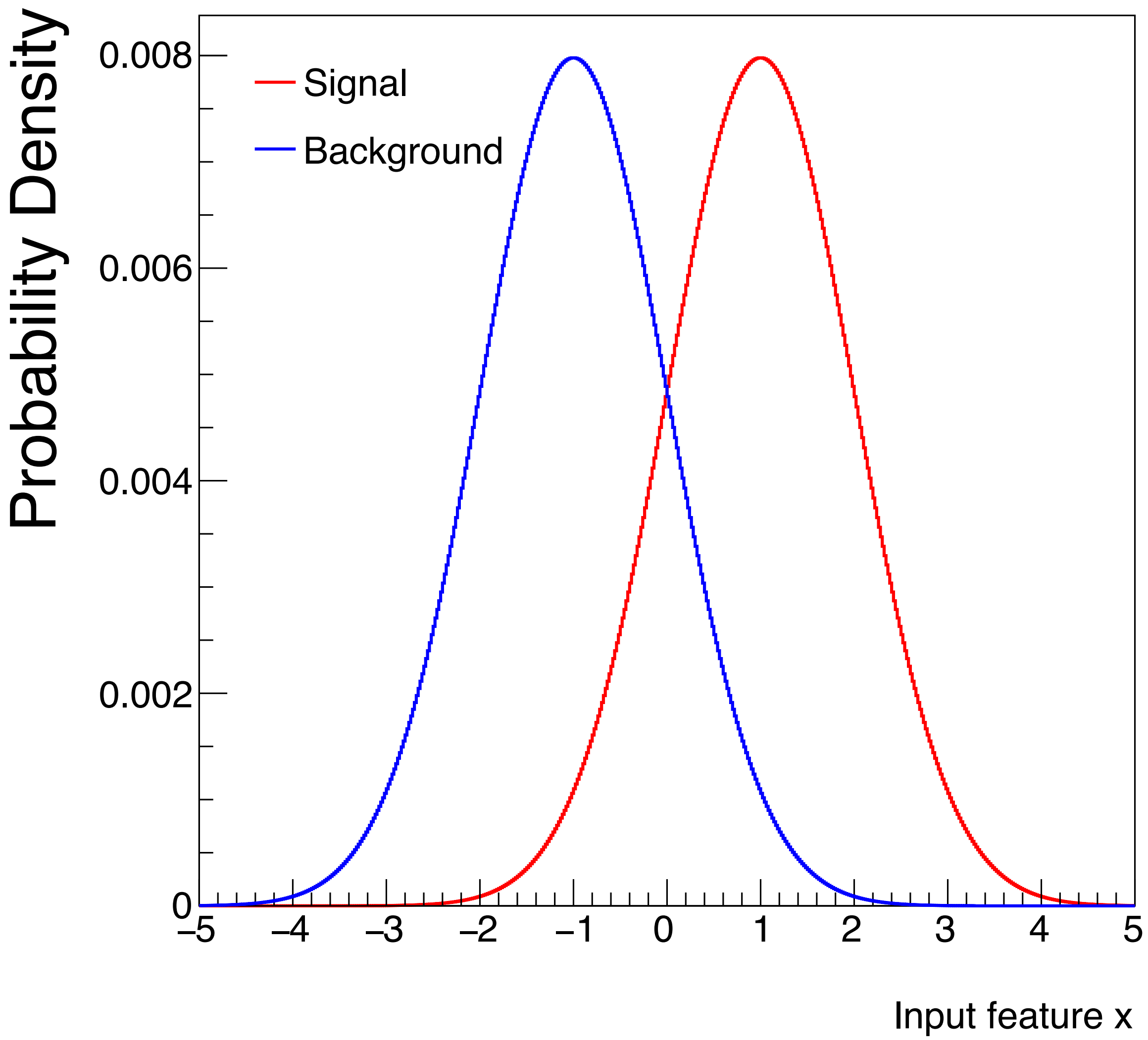


You may be tempted to place a threshold on x

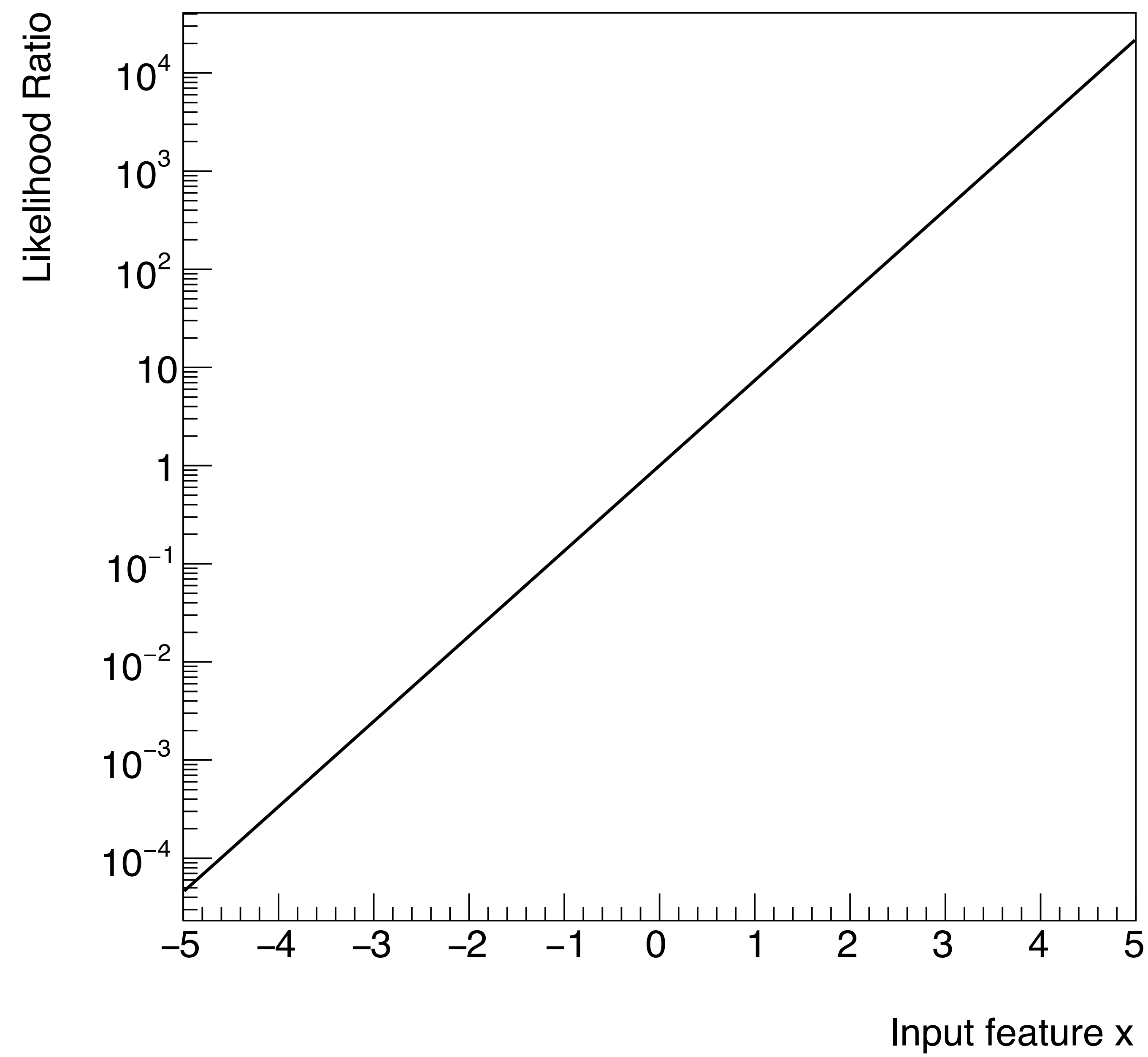
Threshold depends on natural relative abundance

⇨ Try this example out!

Getting into the machine's mind



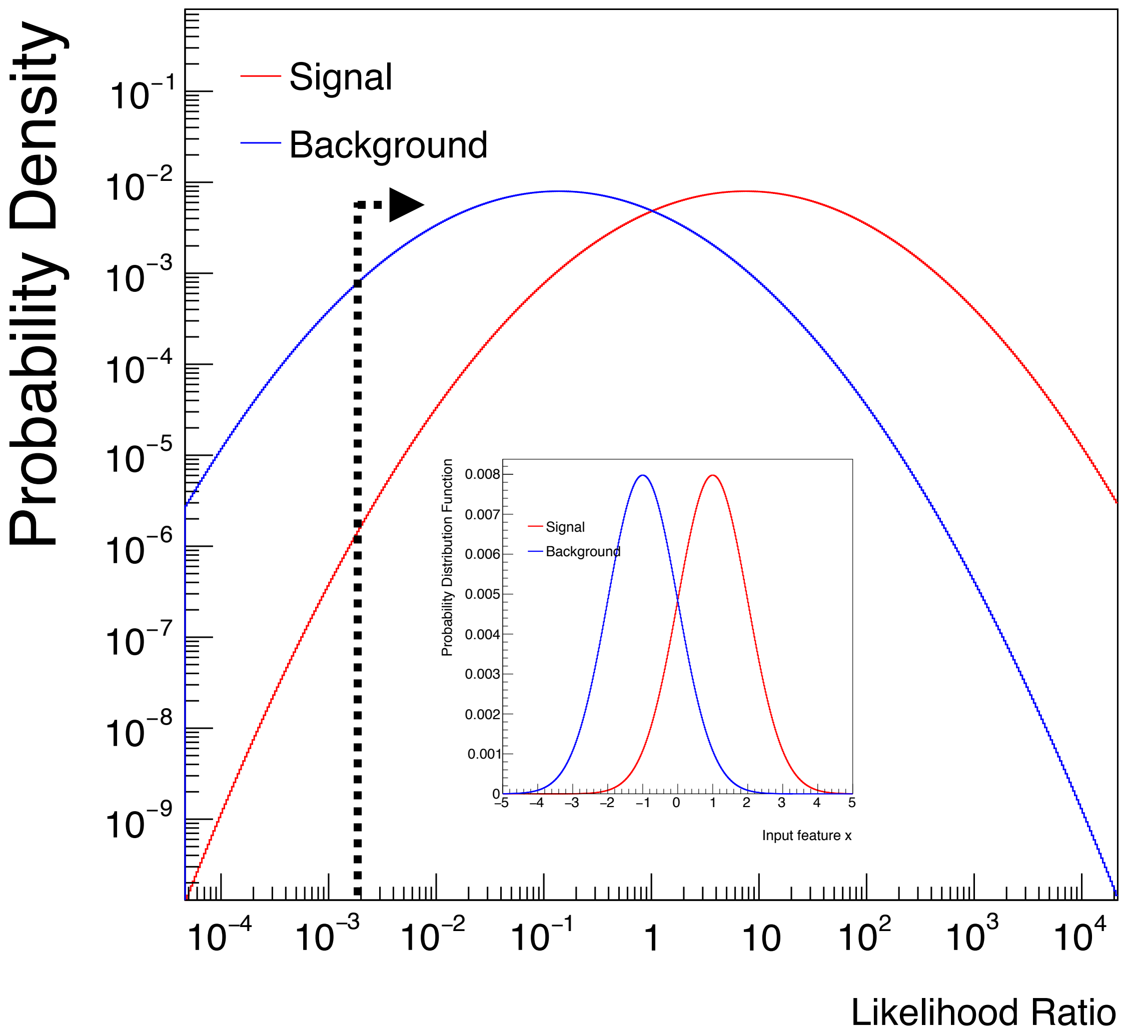
Is the simple threshold cut optimal?



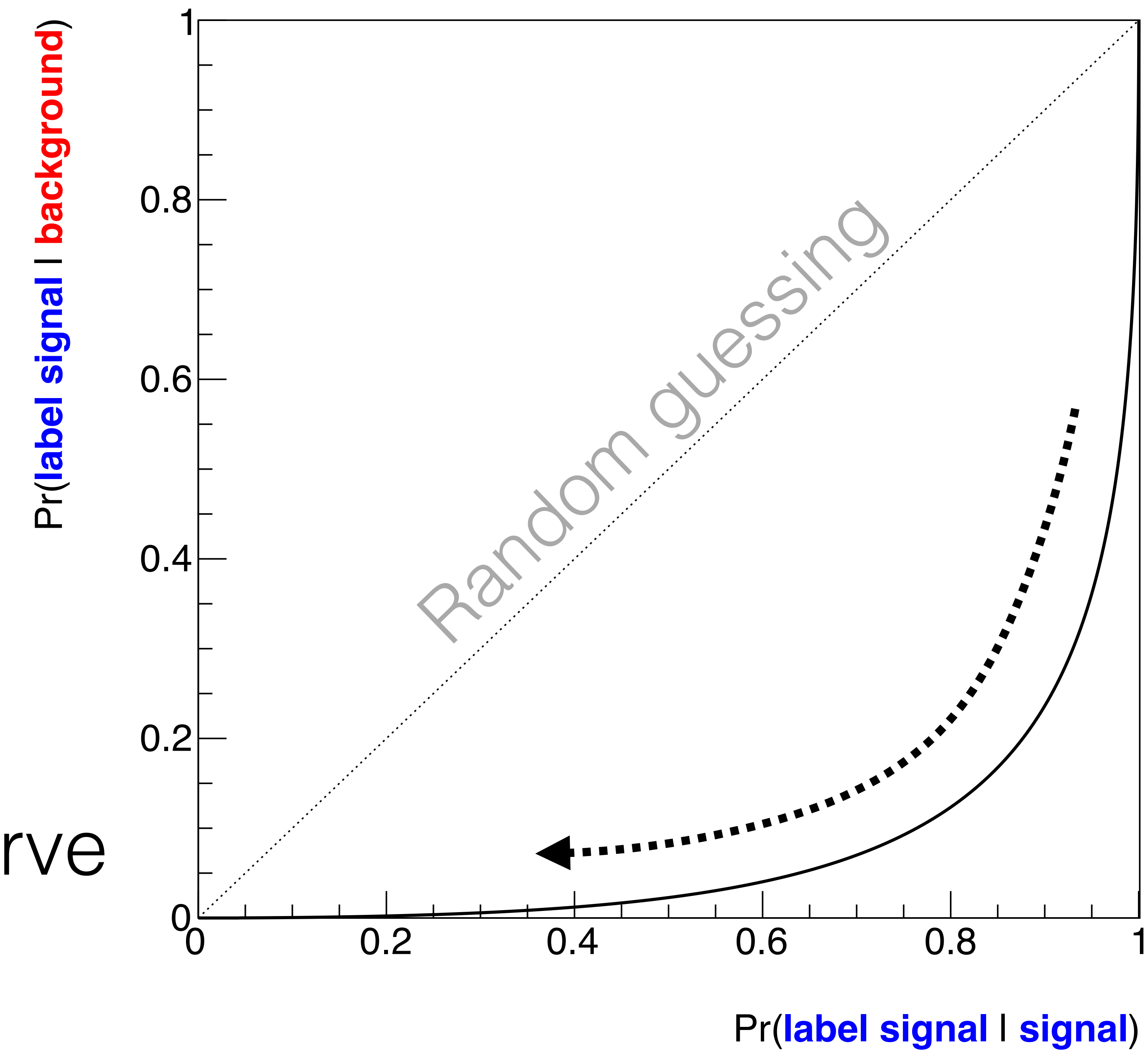
In this simple case, the log LL is proportional to x:
no need for non-linearities!

Threshold cut is optimal

Getting into the machine's mind



The optimal procedure is a threshold on the LL



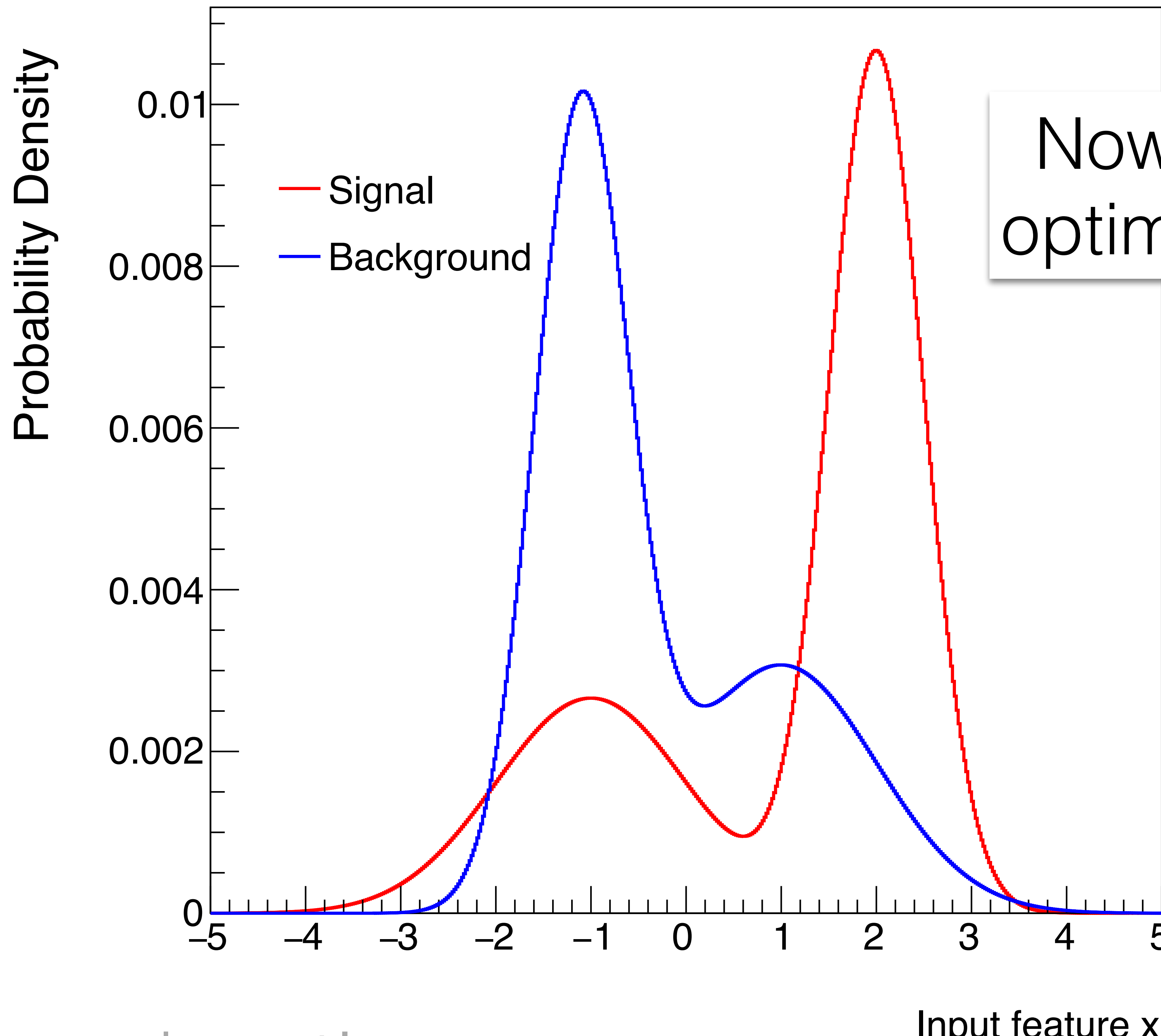
"Receiver Operating Characteristic" (**ROC**) Curve

⇨ Try this example out!

Getting into the machine's mind

What if the distribution of x is complicated?

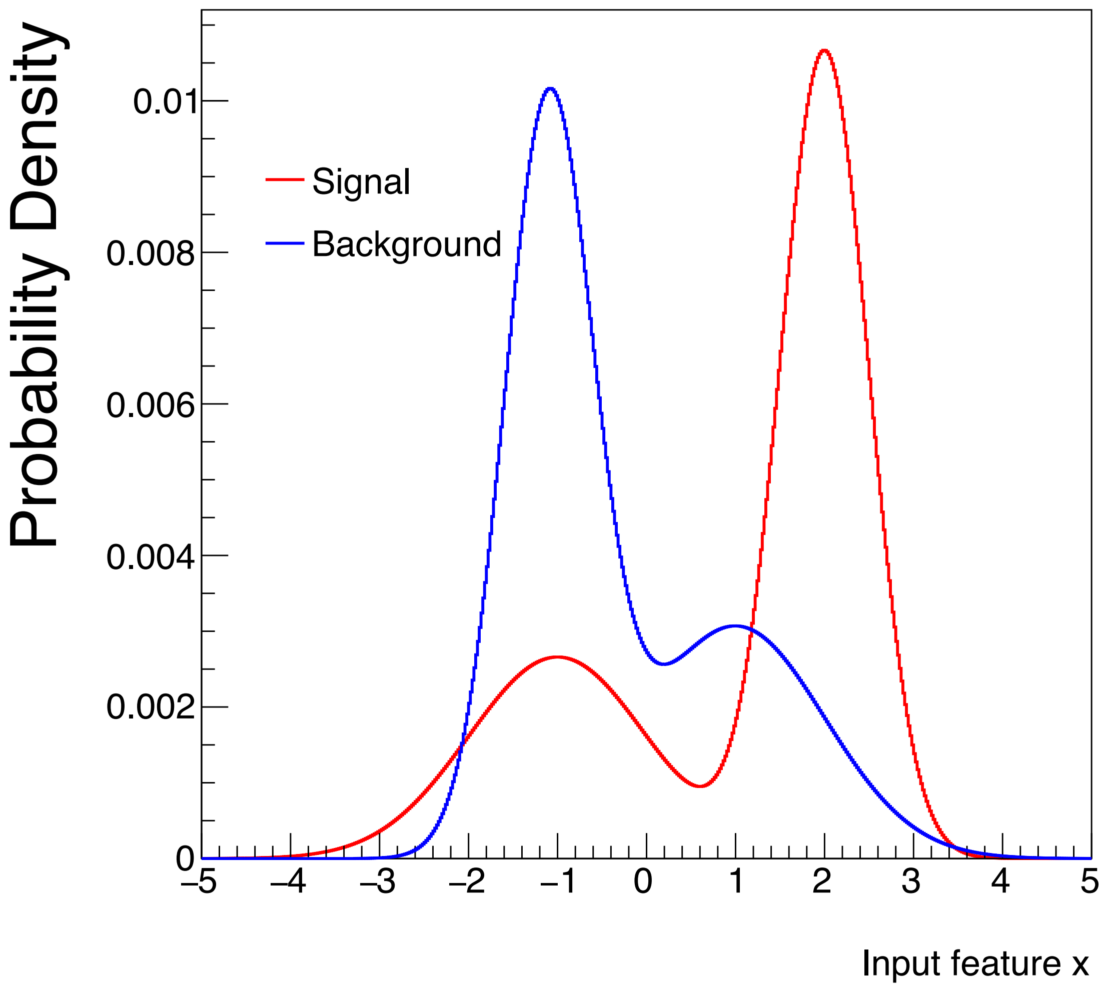
Real life is complicated!



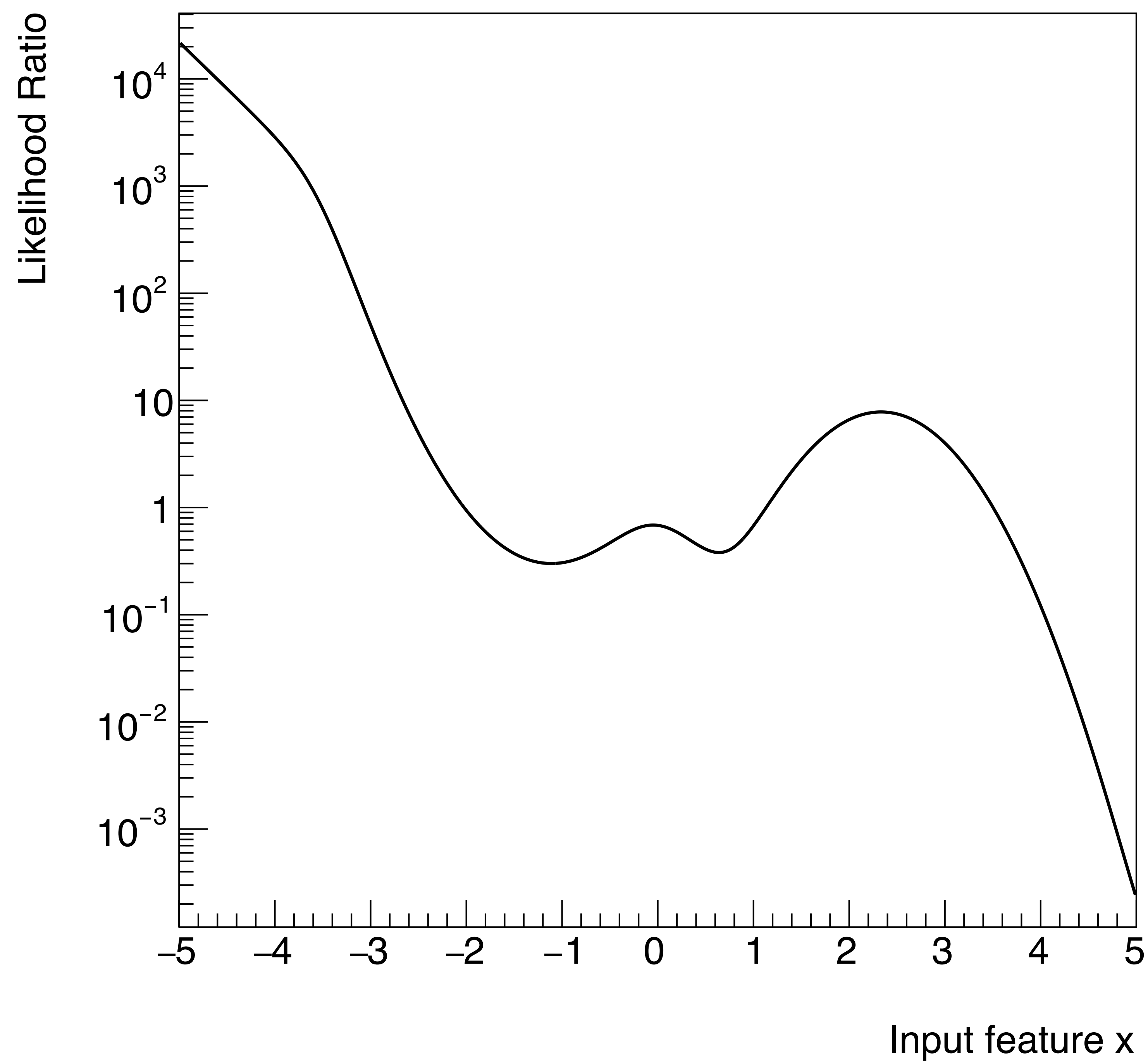
Now what is the optimal classifier?

⇨ Try this example out!

Getting into the machine's mind



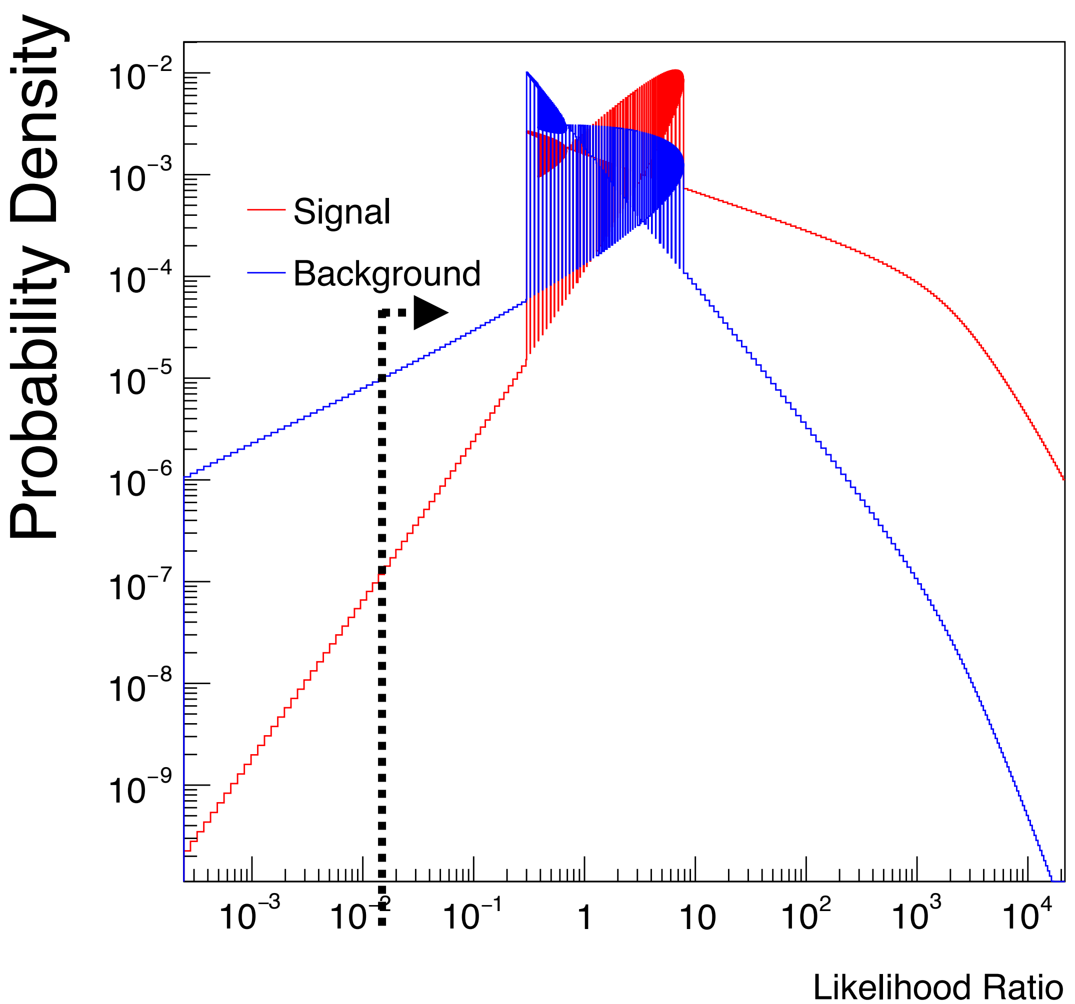
In this case, LL is highly non-linear function of x

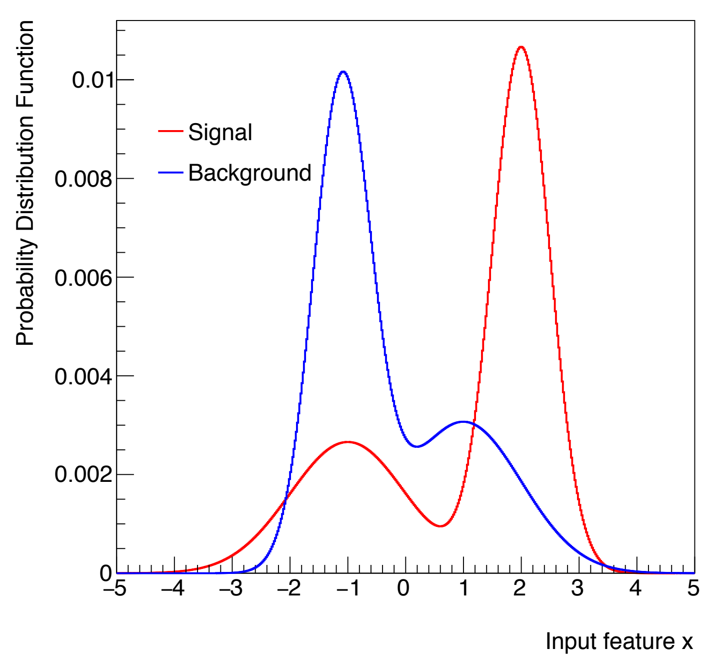


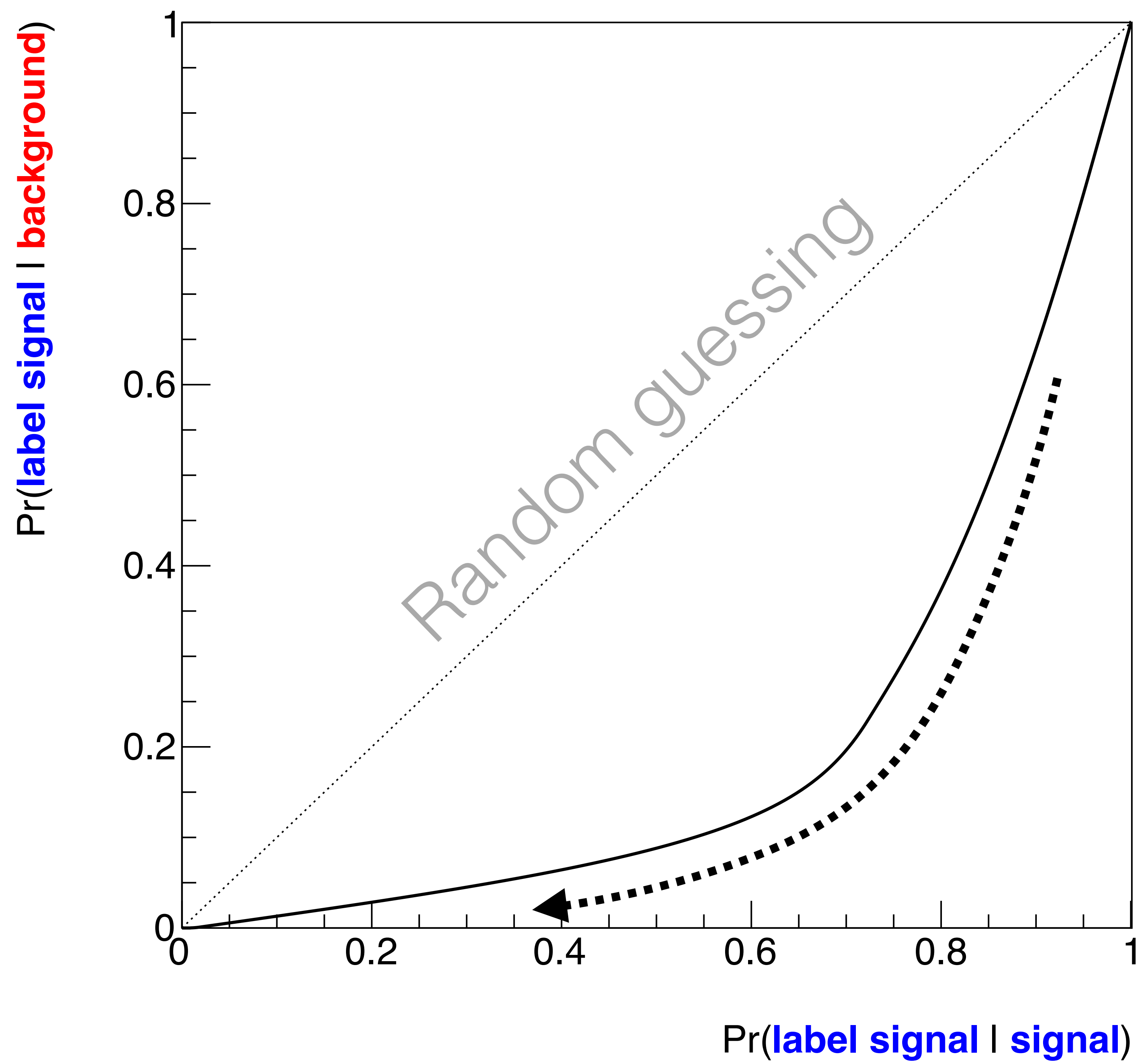
A threshold on x would be sub-optimal

⇨ Try this example out!

Getting into the machine’s mind





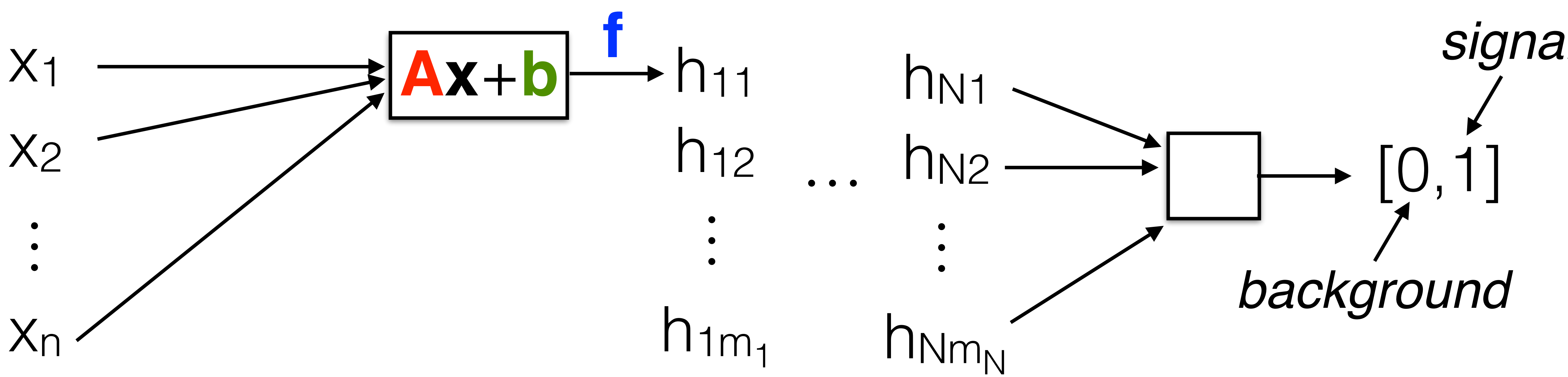


The curse of dimensionality

In principle, you can do the same thing in N > 1 dimensions. However, it very quickly gets out of hand!

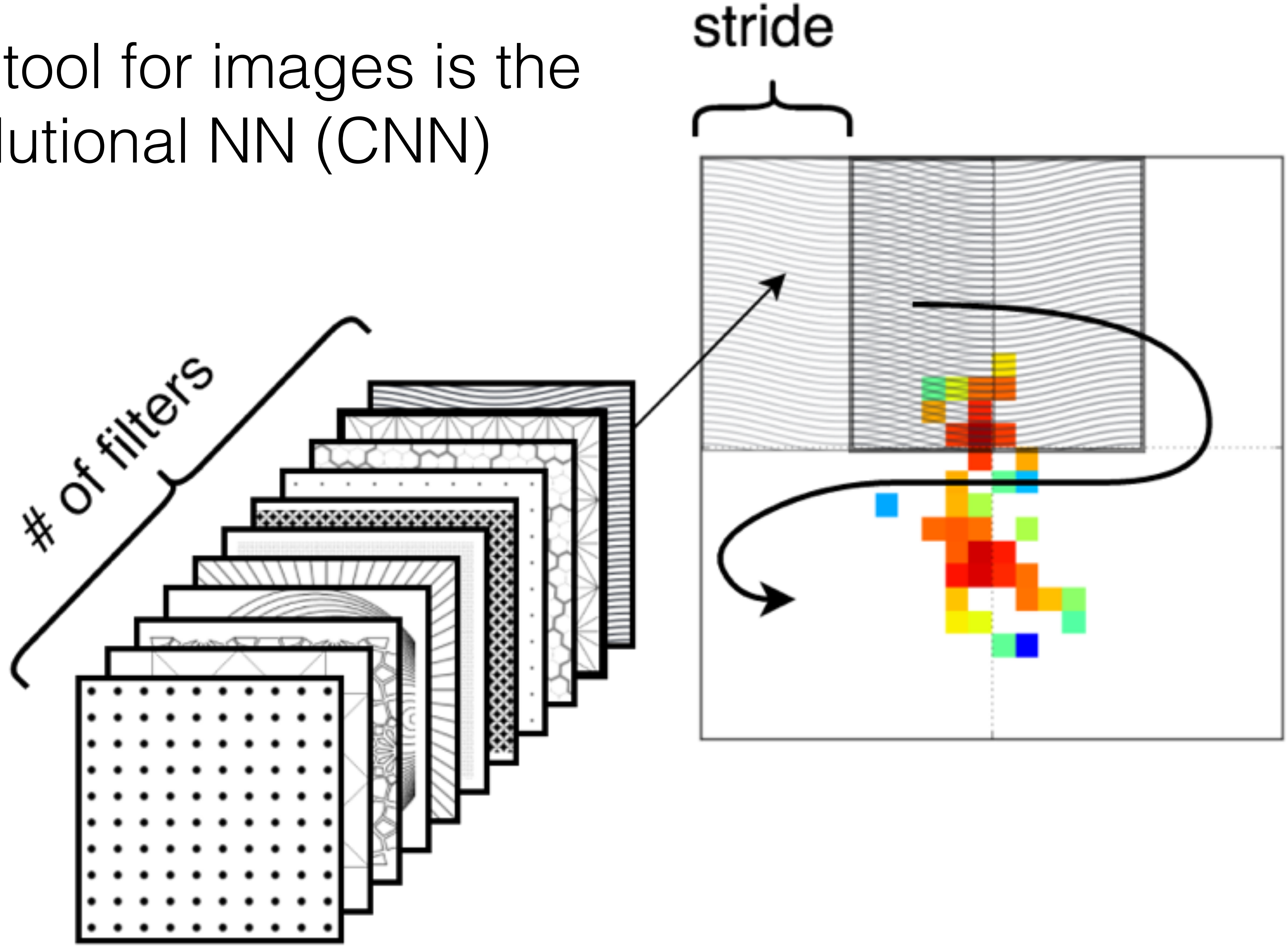
That is where NN's come in.

Image ~ 1000 dimensional



Let's see how we can use DNN's for jet image classification

Common tool for images is the convolutional NN (CNN)

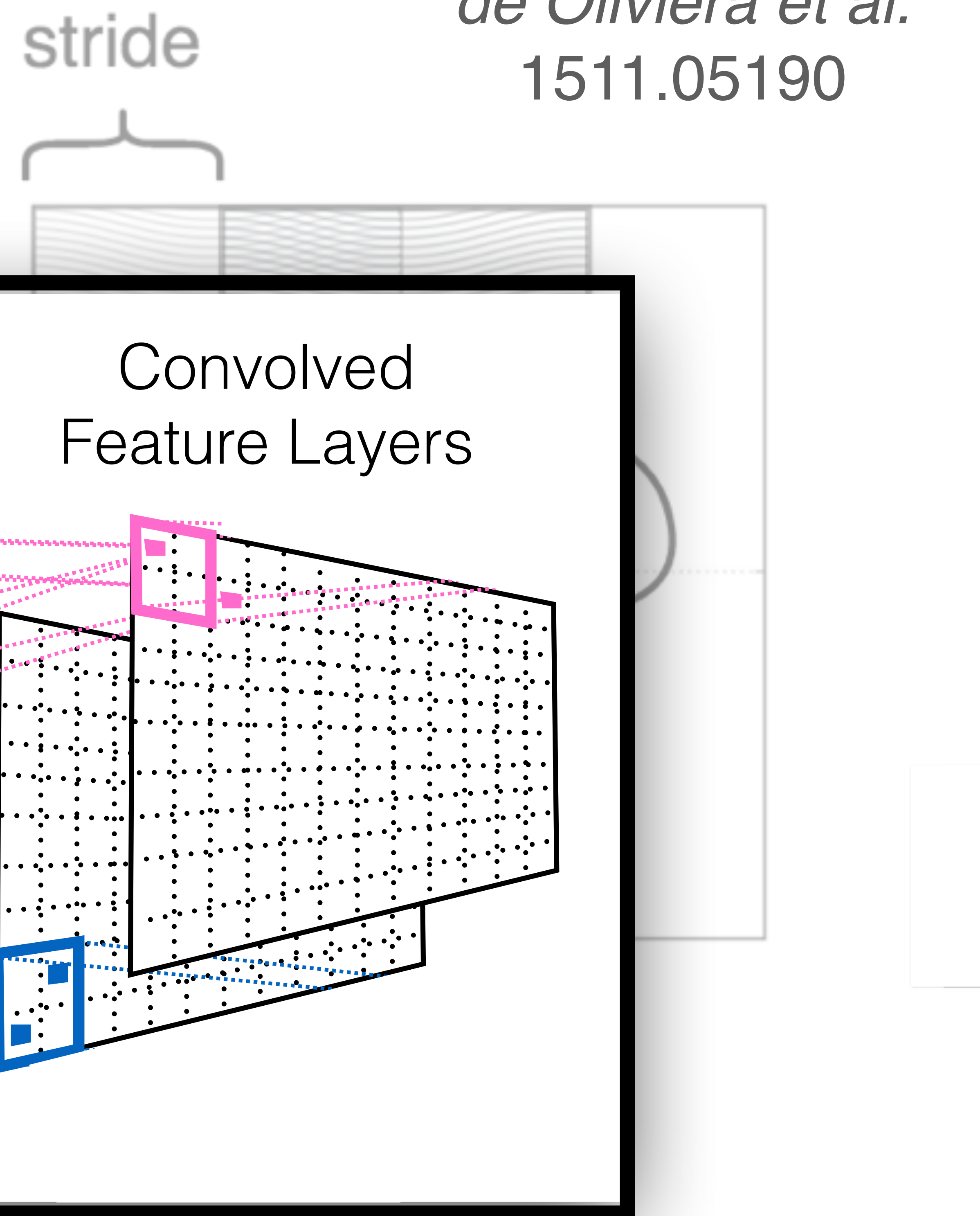


The filter is like the **A**, only the dimensionality is now the filter size (<< n) and not the image size (n).

L. de Oliveira, et al., Comp. and Software for Big Science (2017) 1

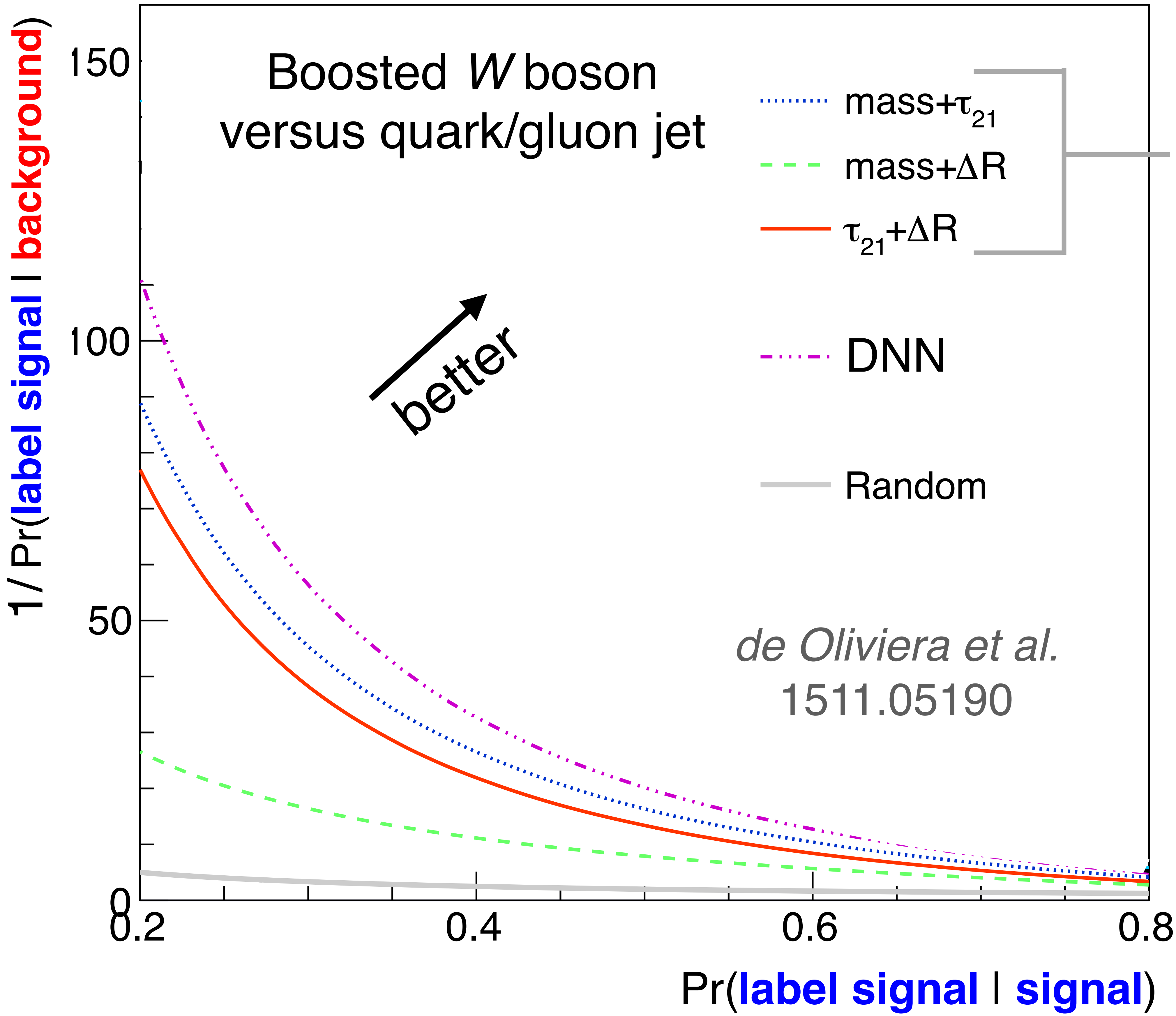
de Oliviera et al.
1511.05190

Common tool for images is the convolutional NN (CNN)



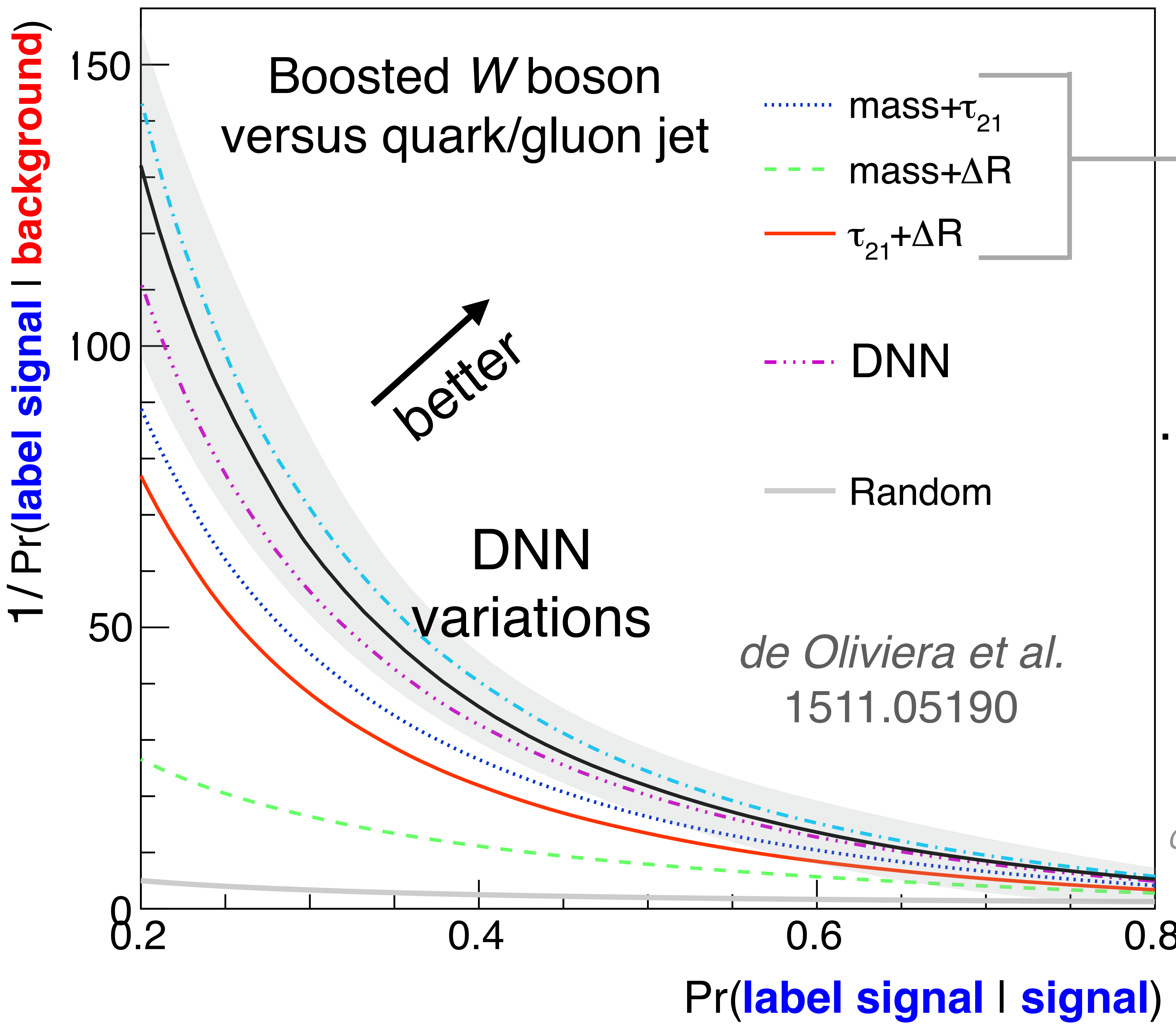
The filter is like the **A**, only the dimensionality is now the filter size (<< n) and not the image size (n).

Modern Deep NN's for Classification



mass, τ_{21}, ΔR are all simple functions of the image

Modern Deep NN's for Classification



mass, τ_{21}, ΔR are all simple functions of the image

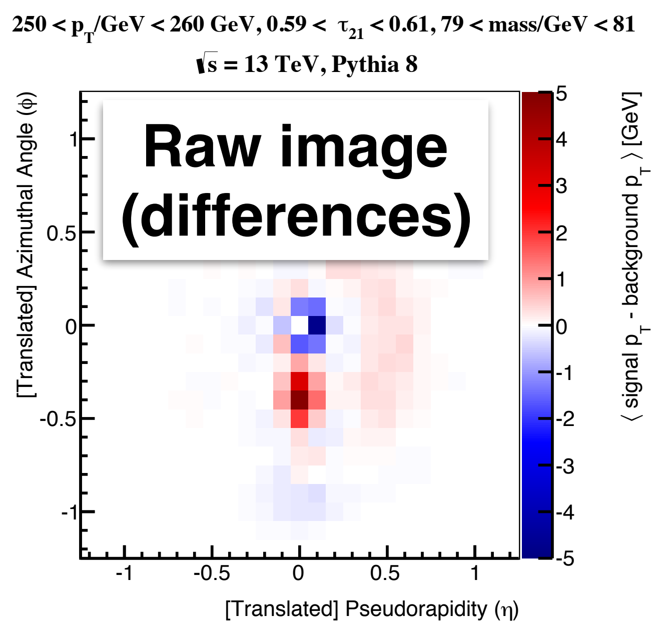
…what the DNN is learning is active R&D!

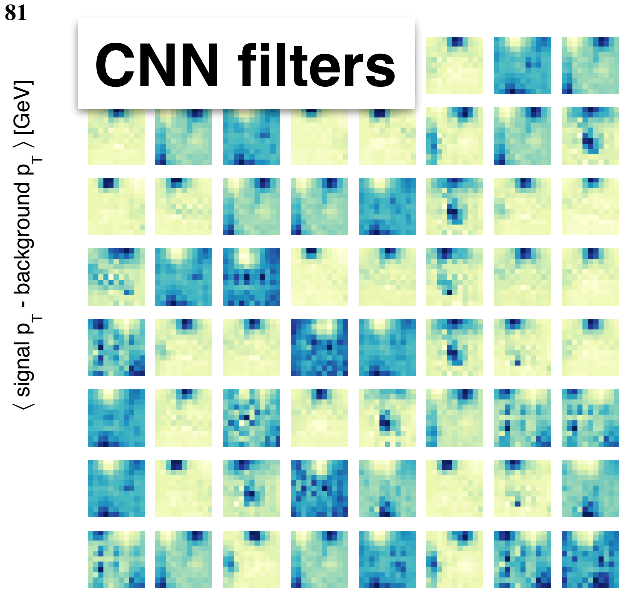
See also

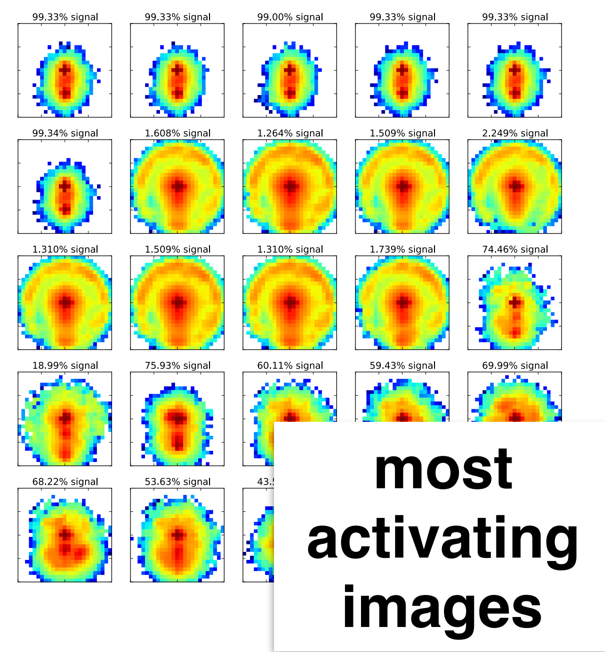
L. Almeida et al. 1501.05968
Baldi et al. 1603.09349
J. Barnard et al. 1609.00607
P. Komiske et al. 1612.01551
G. Kasieczka et al. 1701.08784
W. Bhimji et al. 1711.03573

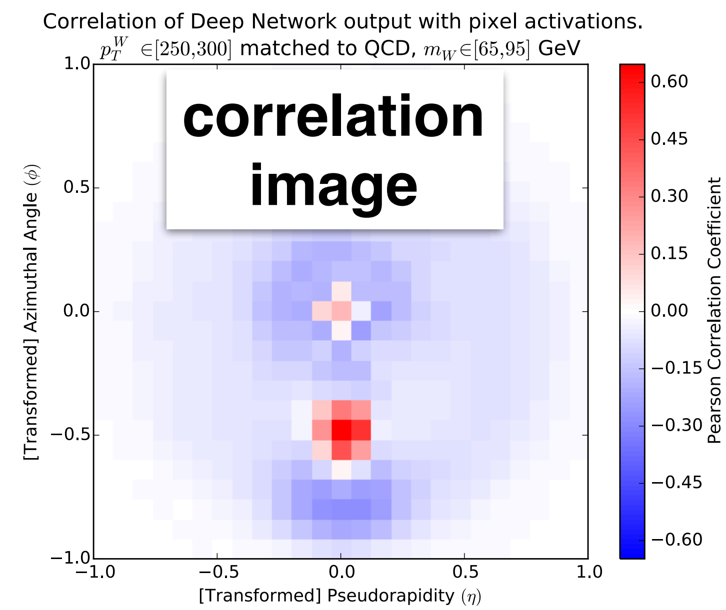
Learning about Learning

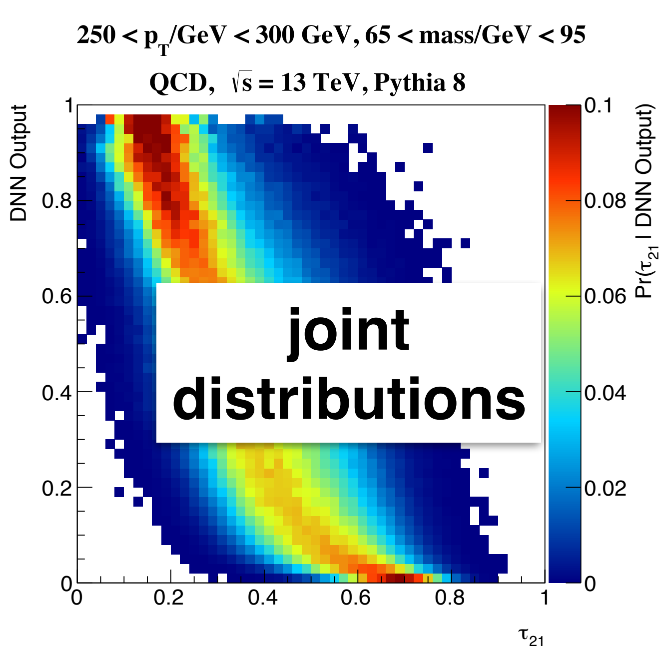
Opening the **box** is critical for improving robustness

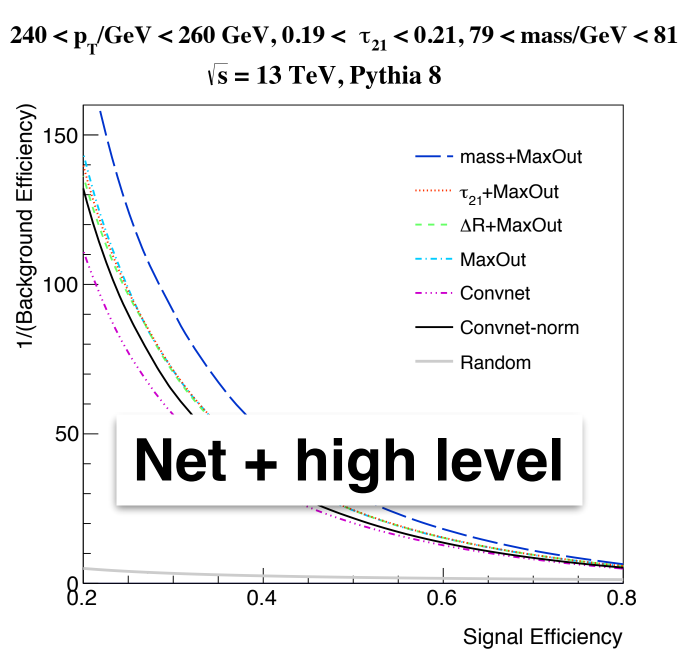


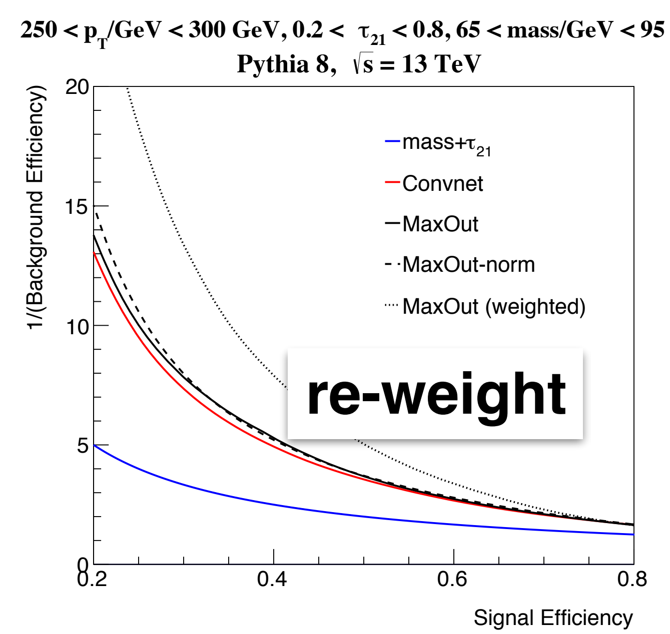


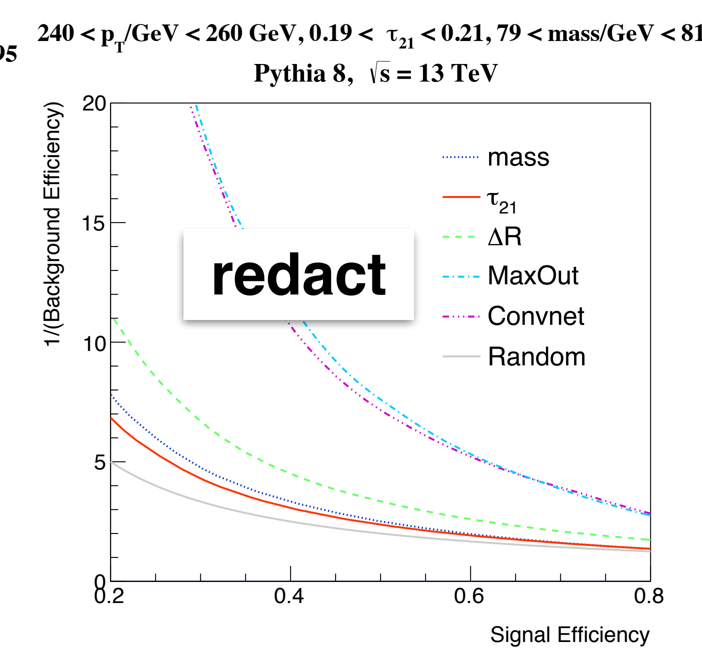






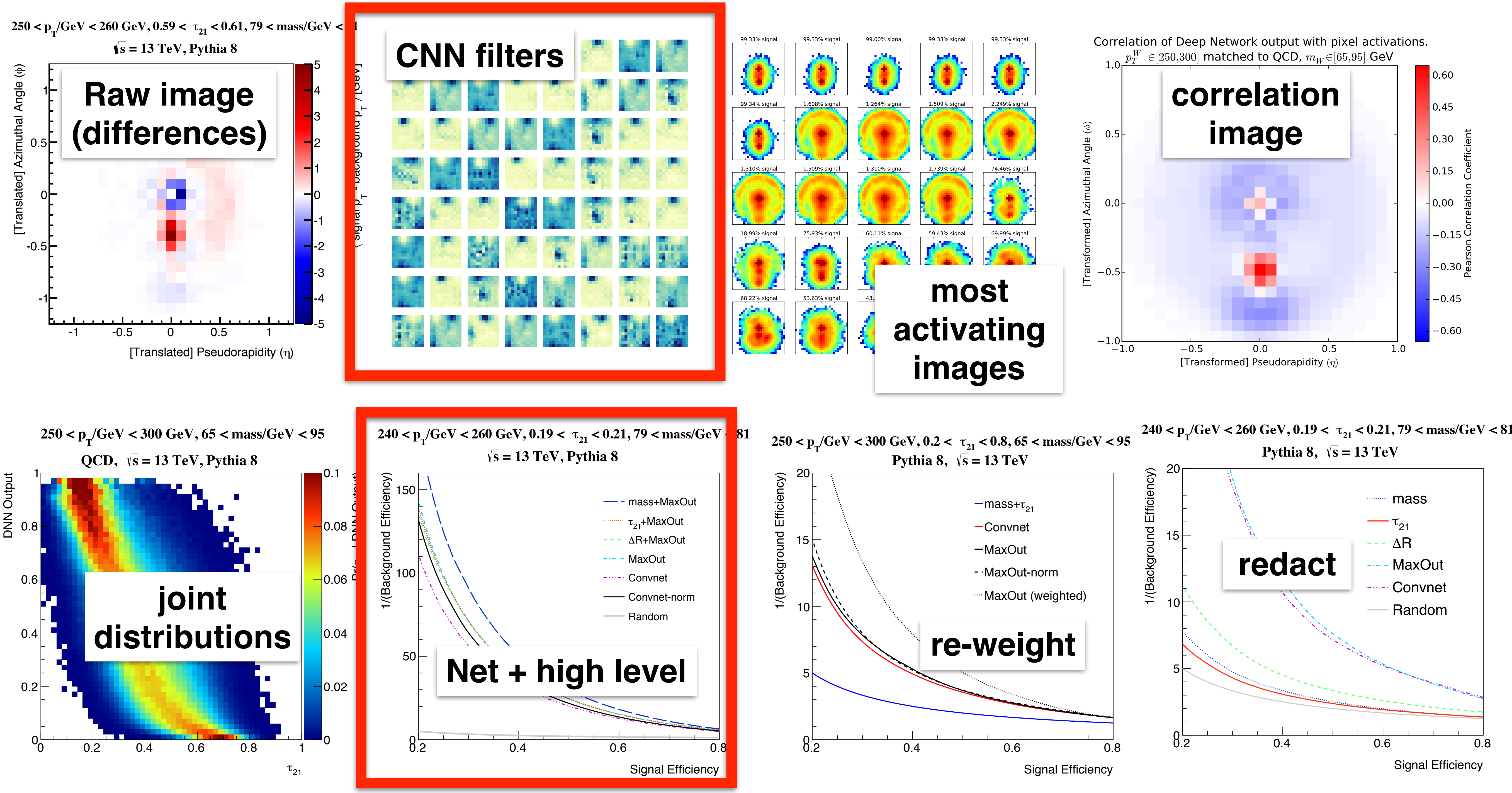






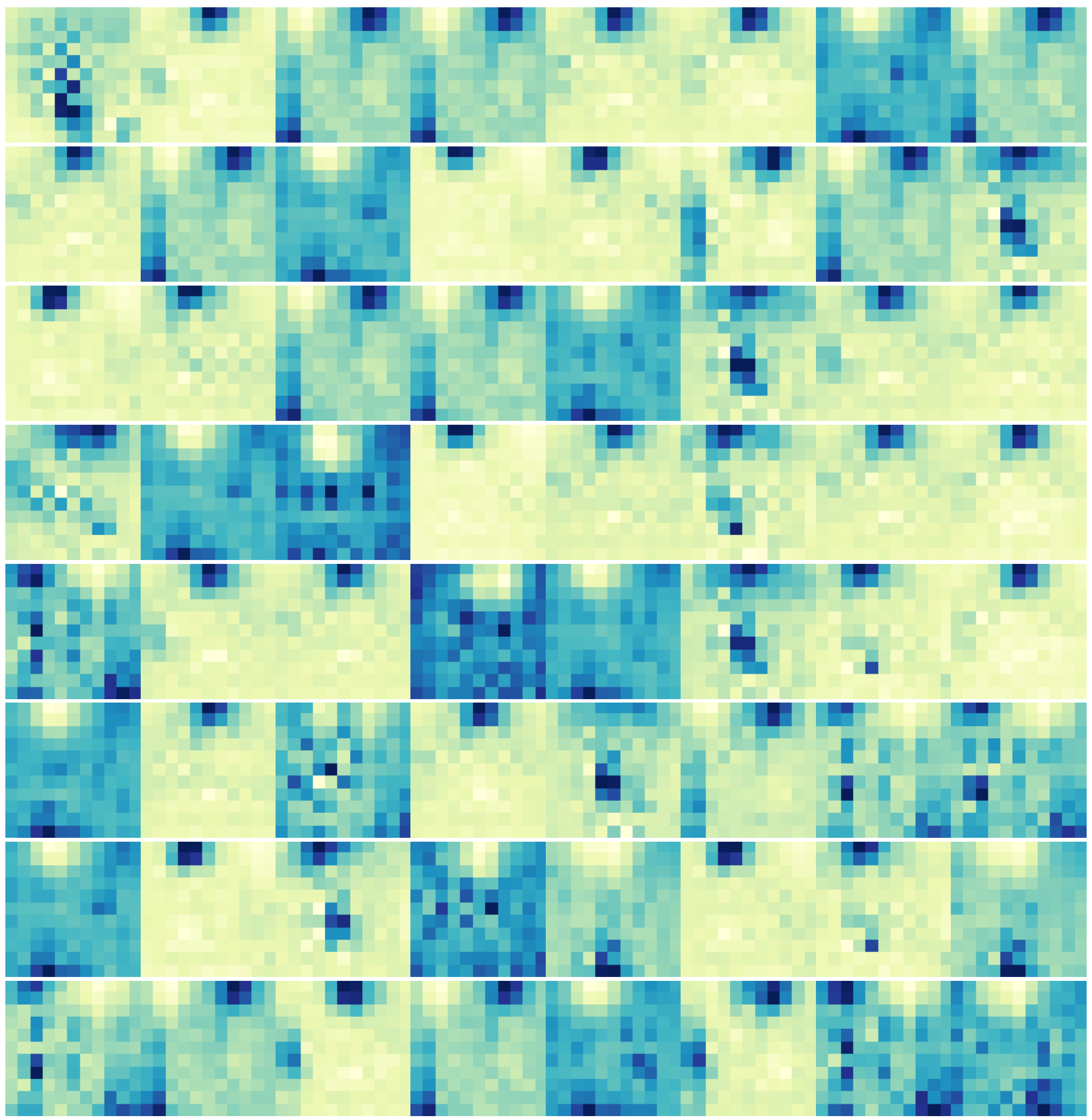
Learning about Learning

Opening the **box** is critical for improving robustness



Convolutional Filters

Filters are images! Can visualize ‘higher-level features’ learned by the network



Jet Images

learned edge detection

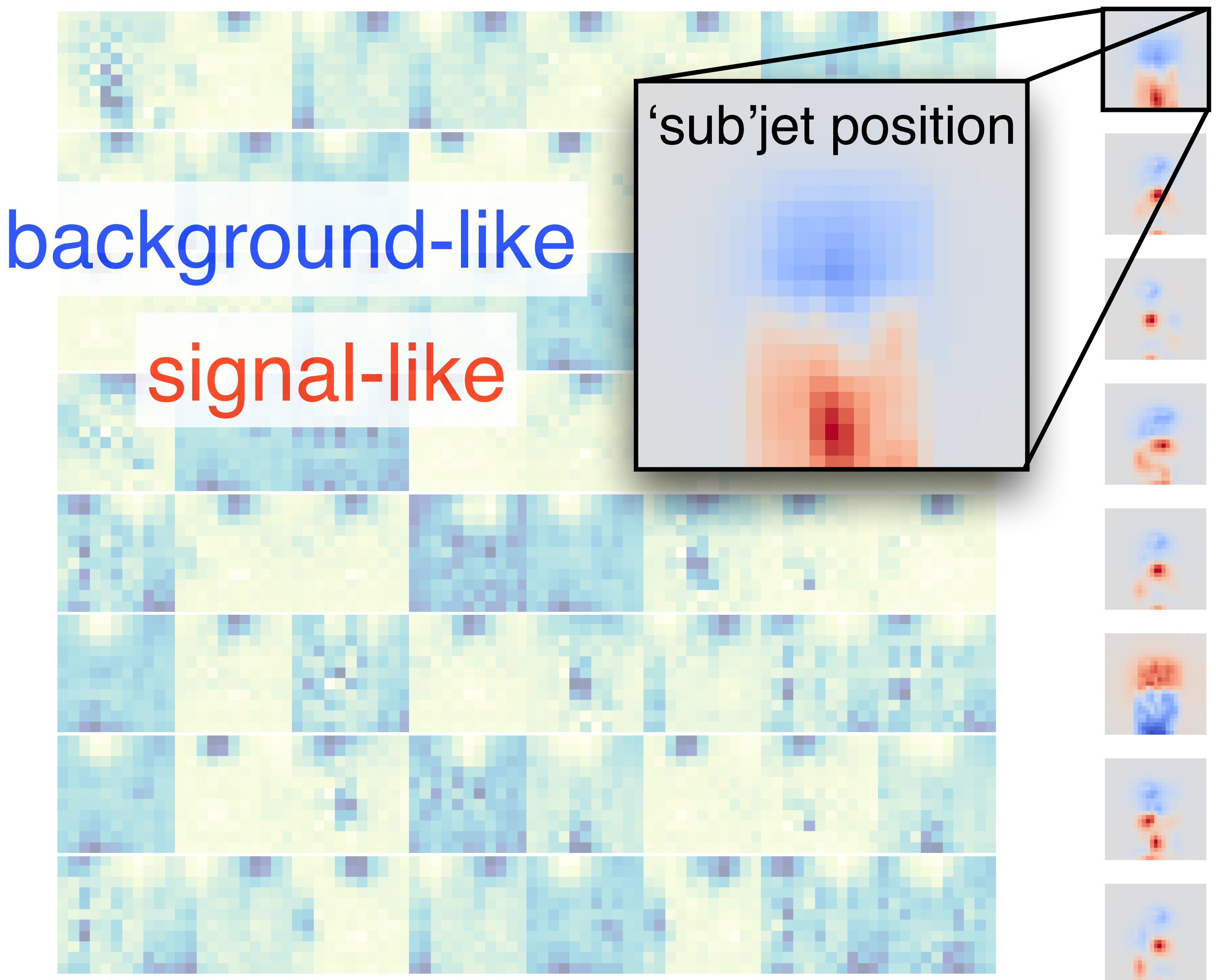
“Natural” Images

L. de Oliveira et al., JHEP 07 (2016) 069

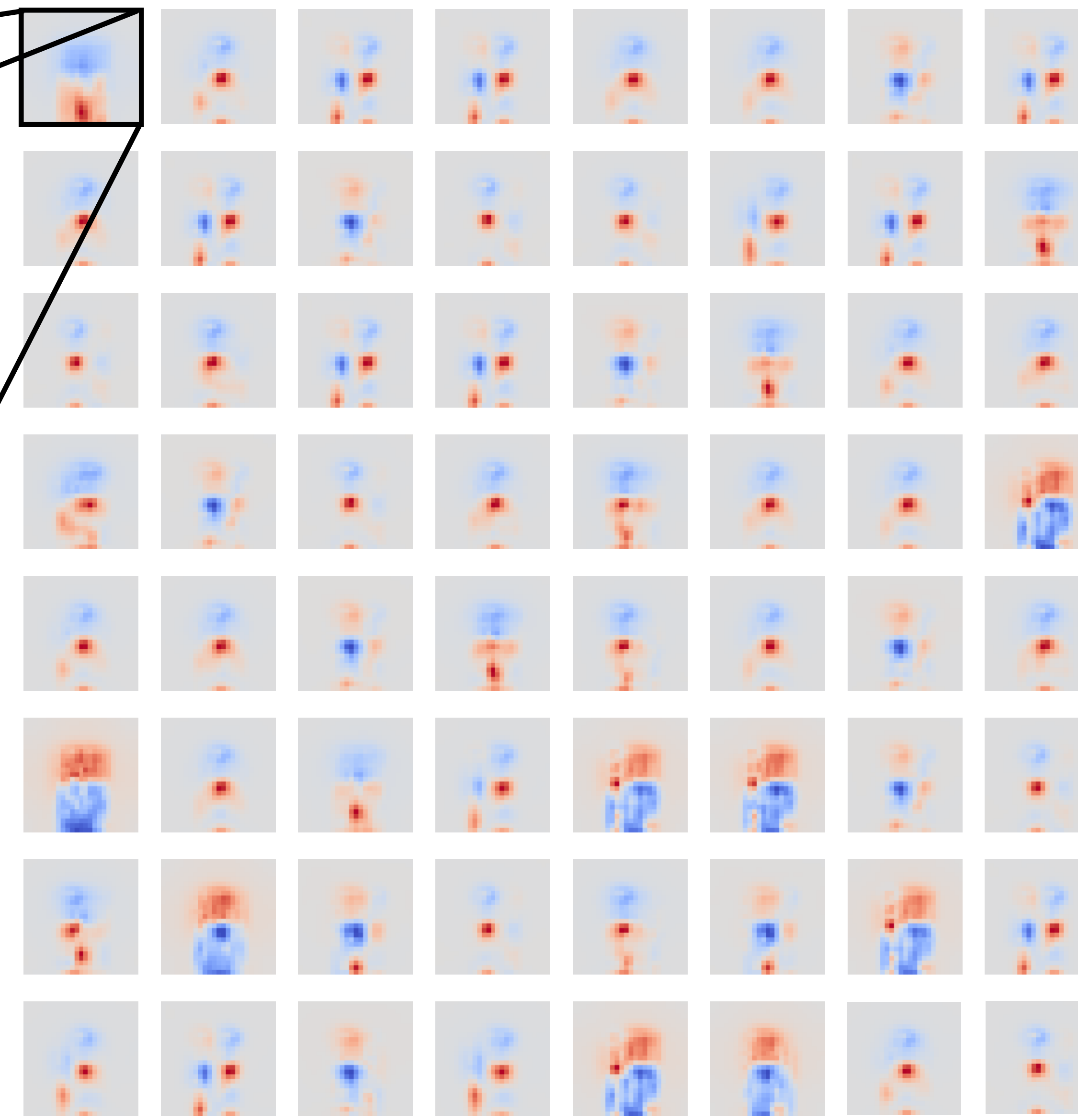
A. Krizhevsky et al. DNN for ImageNet

Convolutional Filters

Filters are images! Can visualize 'higher-level features' learned by the network



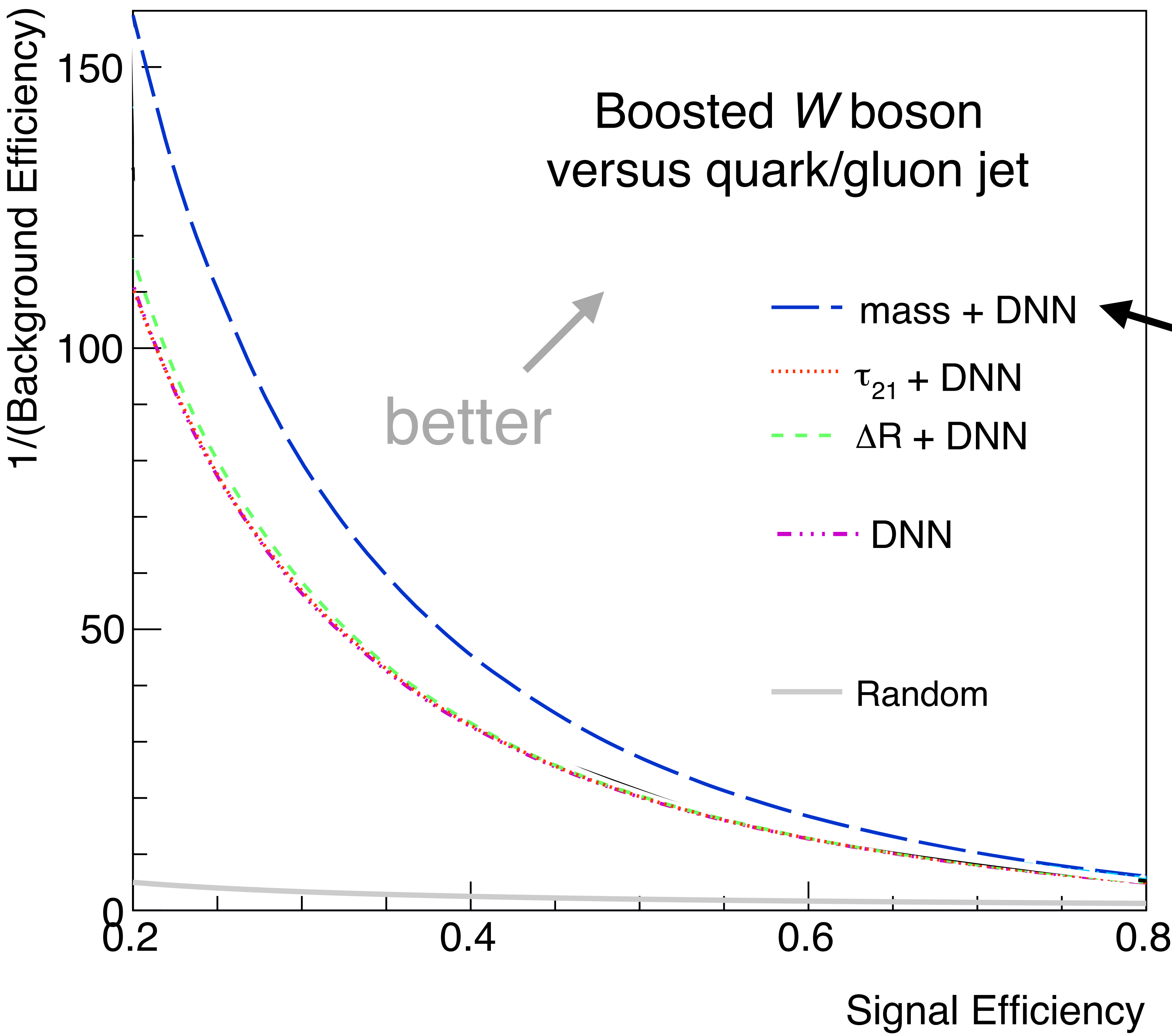
Jet Images Layer 1 Filters



Filters convolved with signal - background

L. de Oliveira et al., JHEP 07 (2016) 069

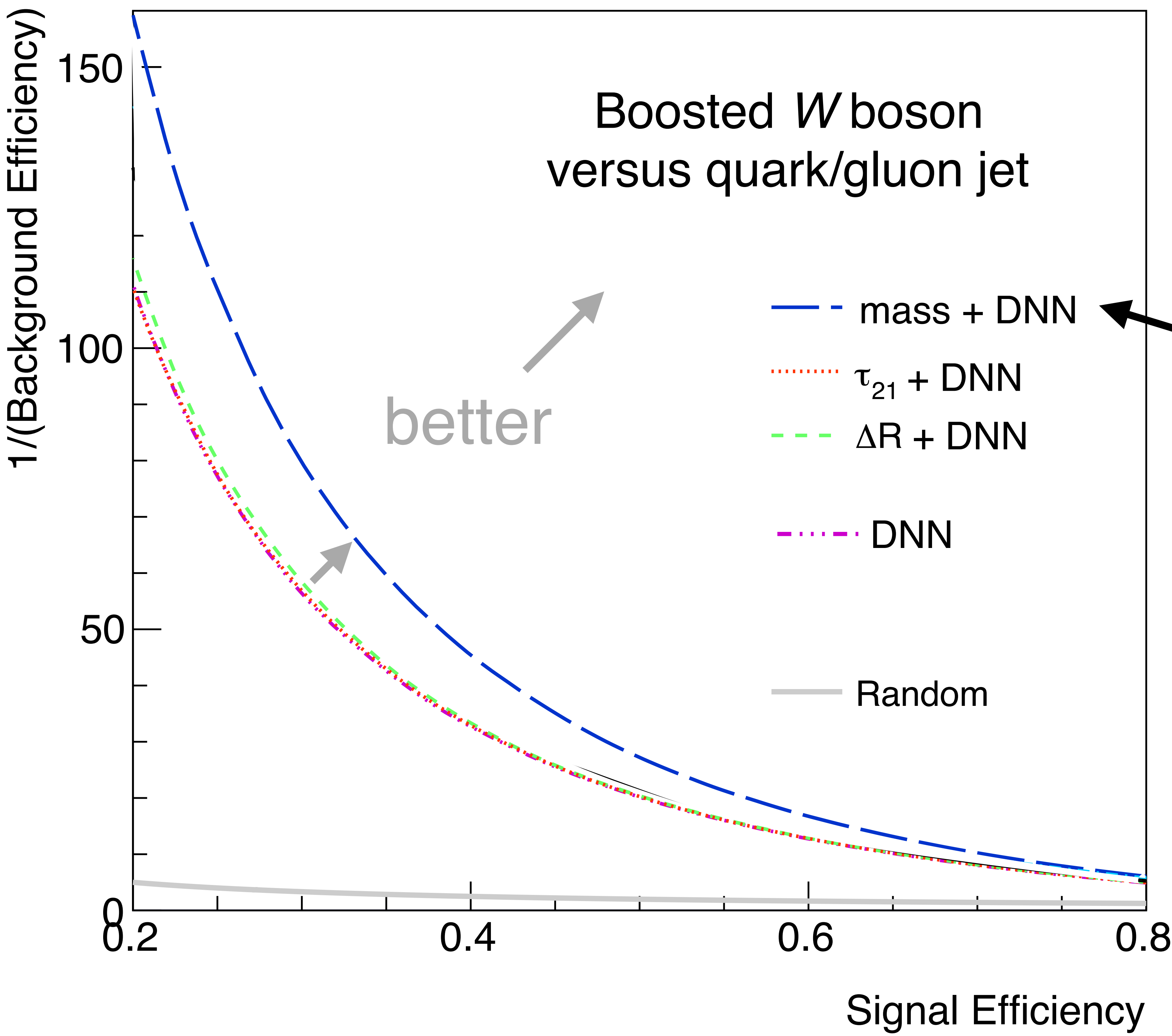
Is there more to learn that we know about?



Idea: explicitly combine NN with a known feature

Has learned image mass?

Is there more to learn that we know about?

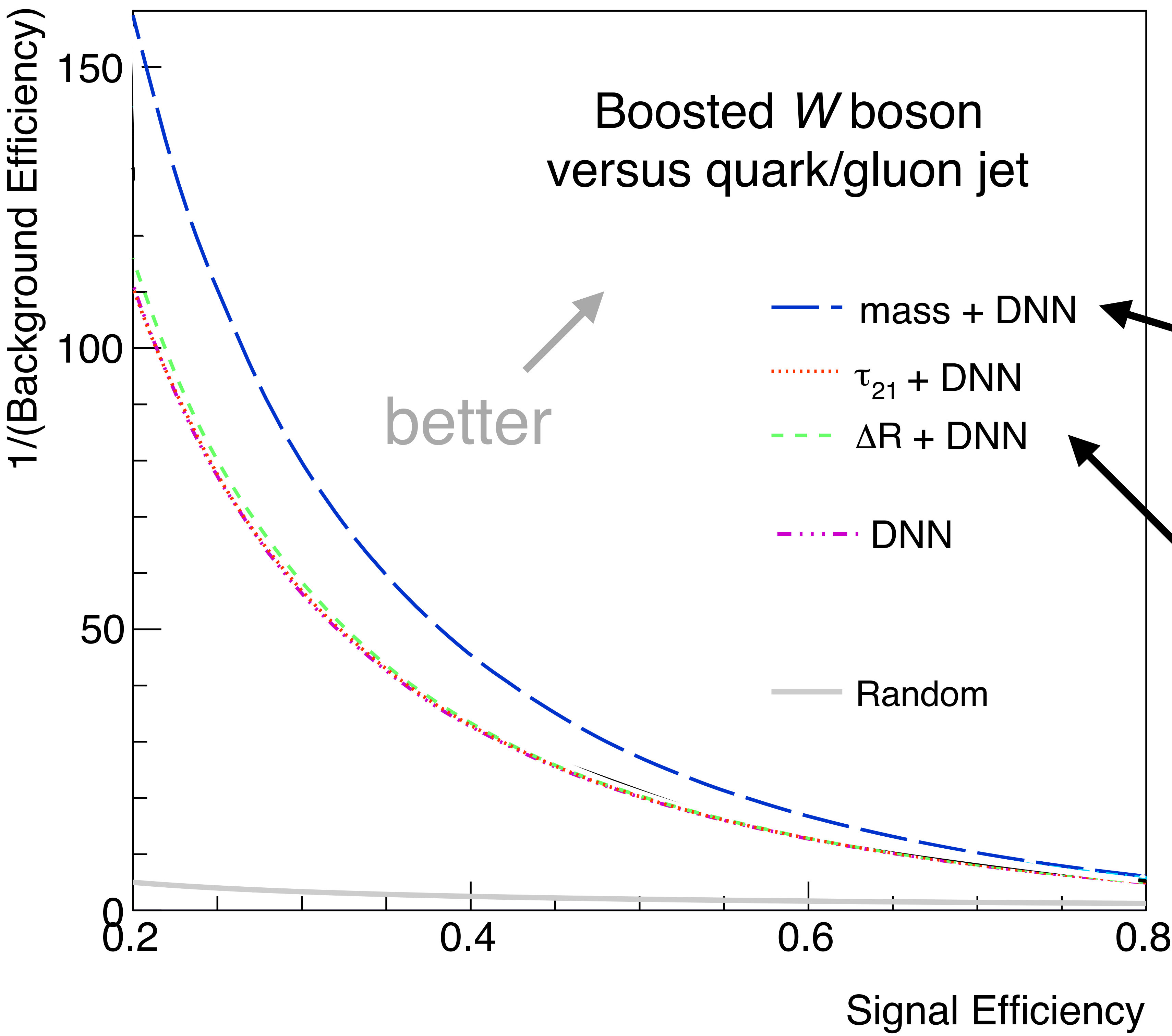


Idea: explicitly combine NN with a known feature

Has learned image mass?

not fully!

Is there more to learn that we know about?



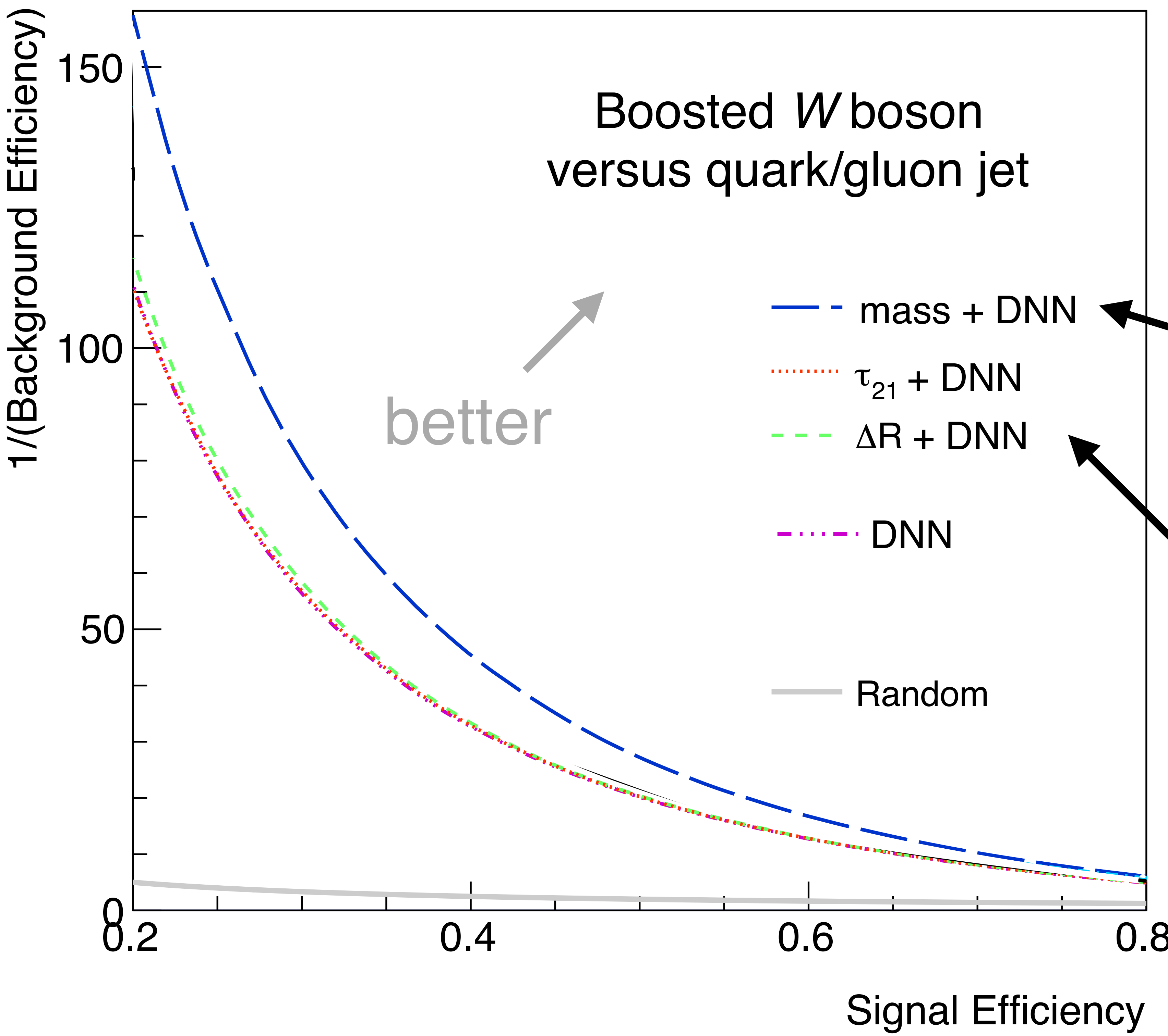
Idea: explicitly combine NN with a known feature

Has learned image mass?

not fully!

What about the distance between subjets?

Is there more to learn that we know about?



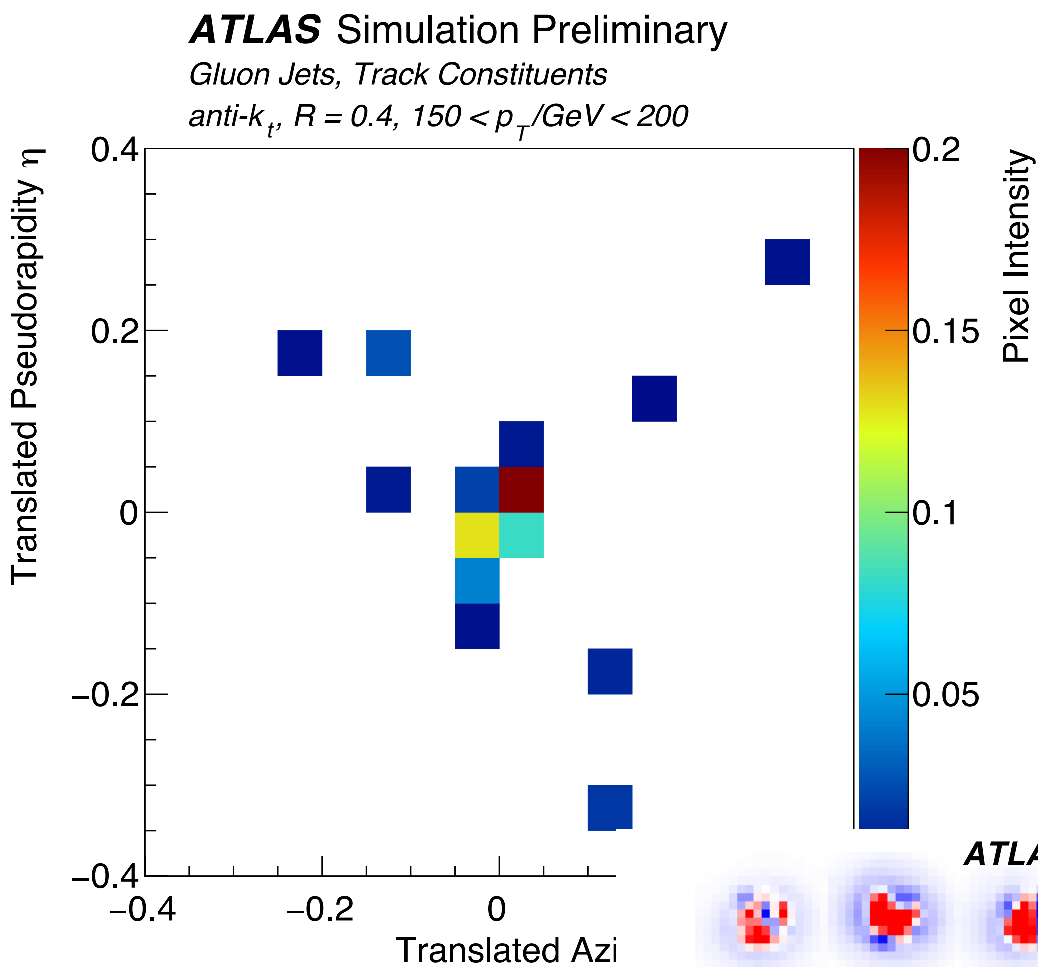
Idea: explicitly combine NN with a known feature

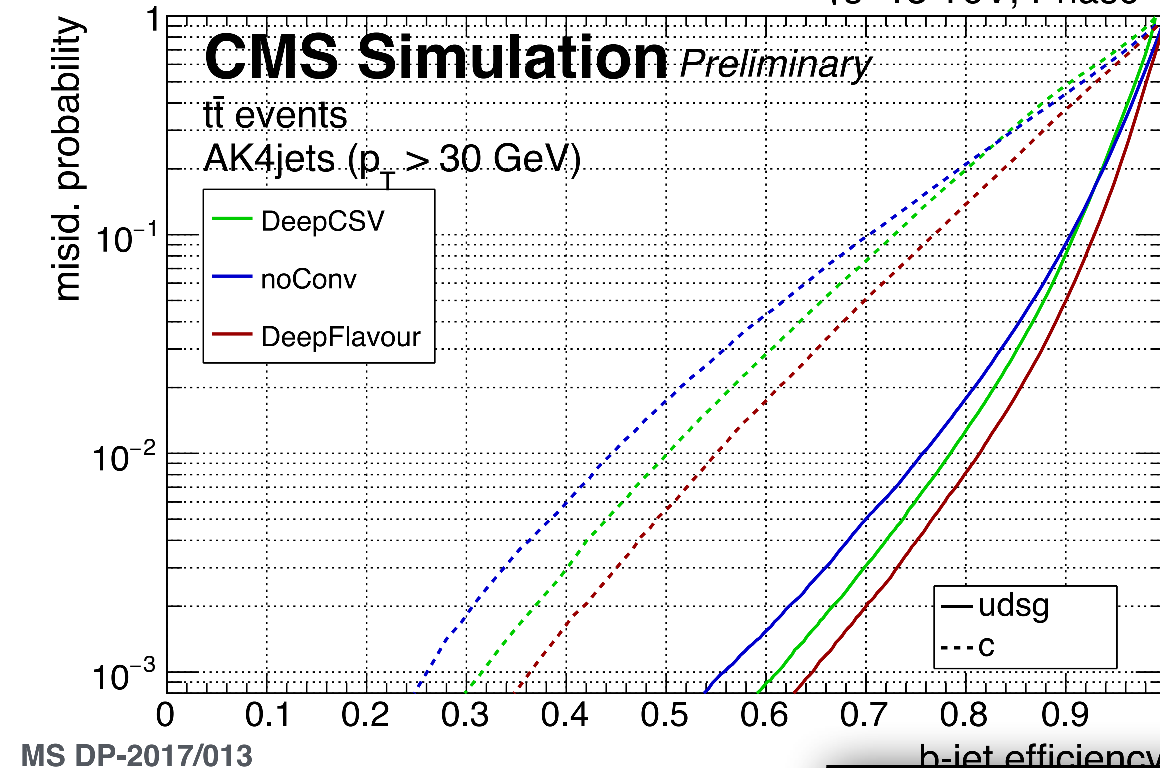
Has learned image mass?

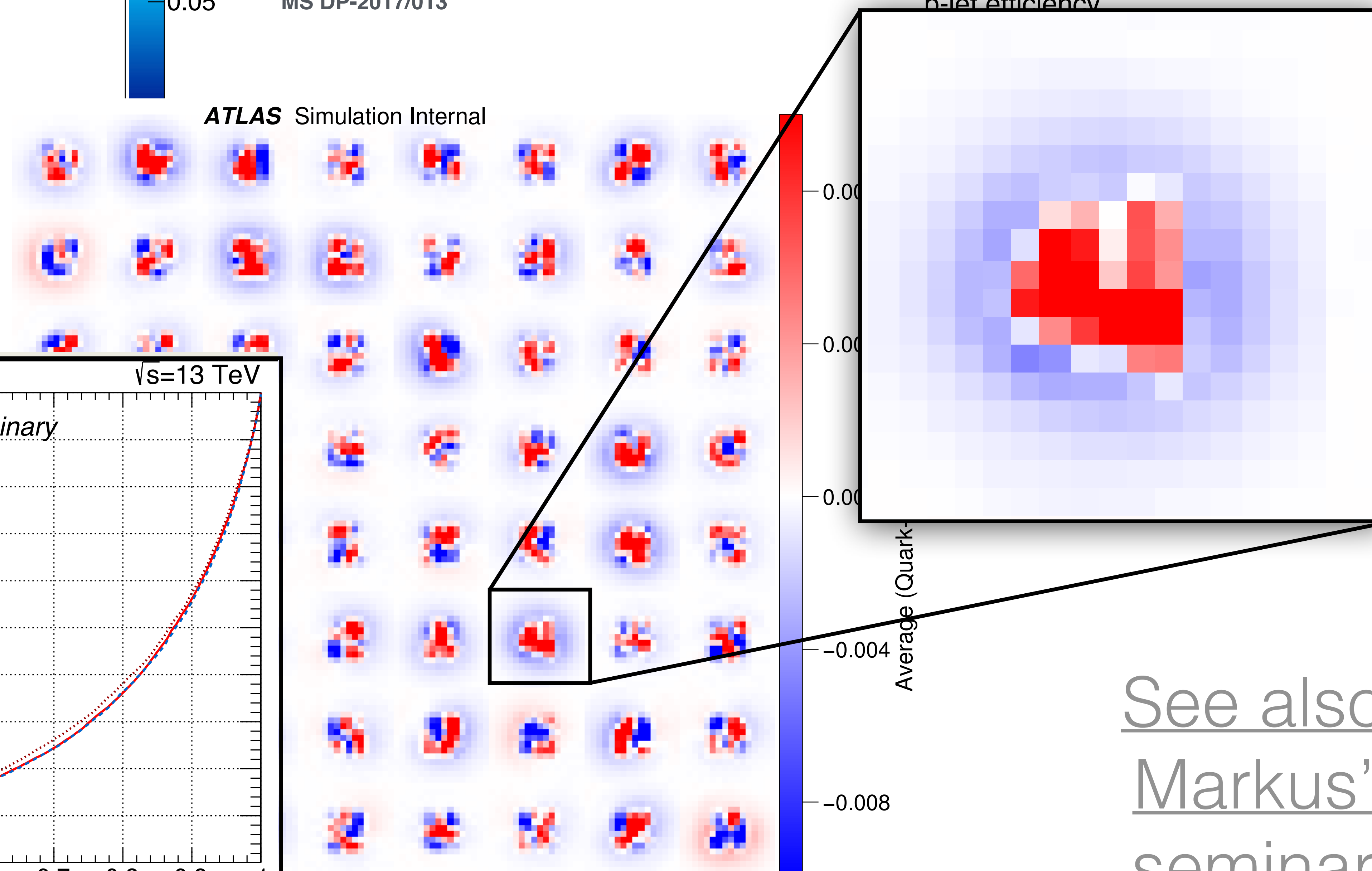
not fully!

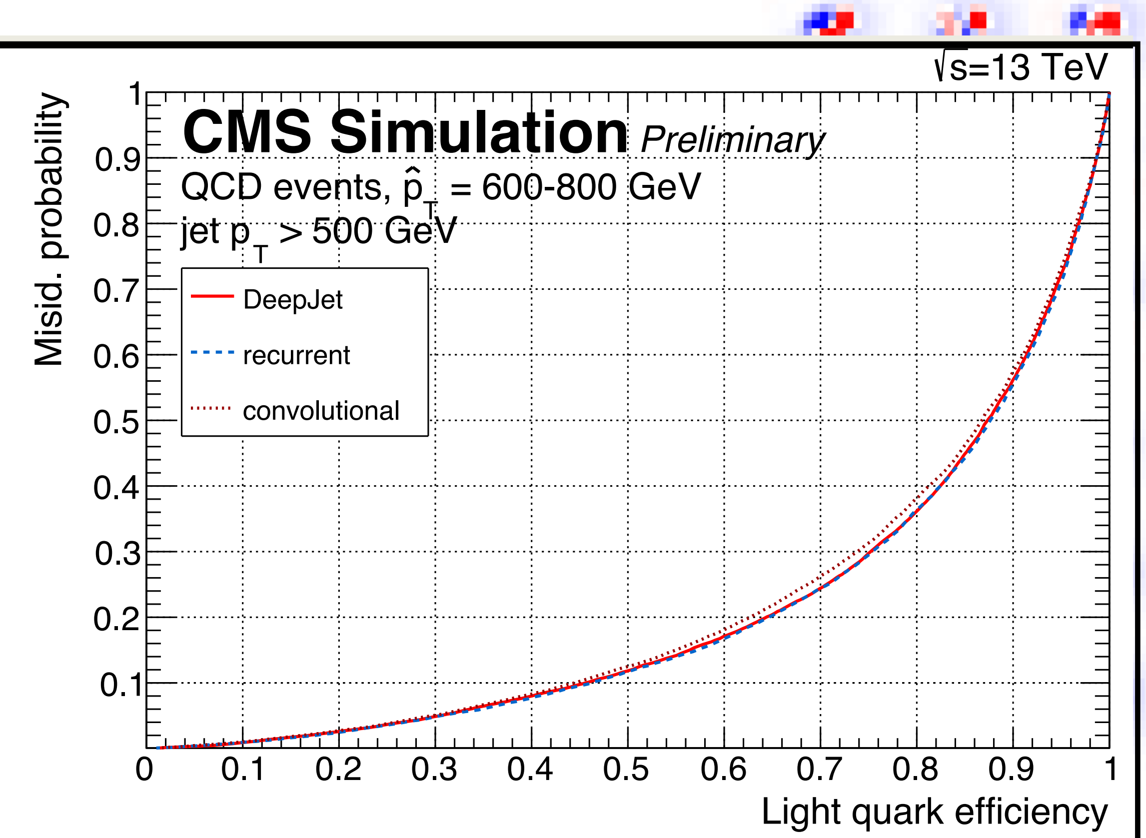
What about the distance between subjets?

yes!





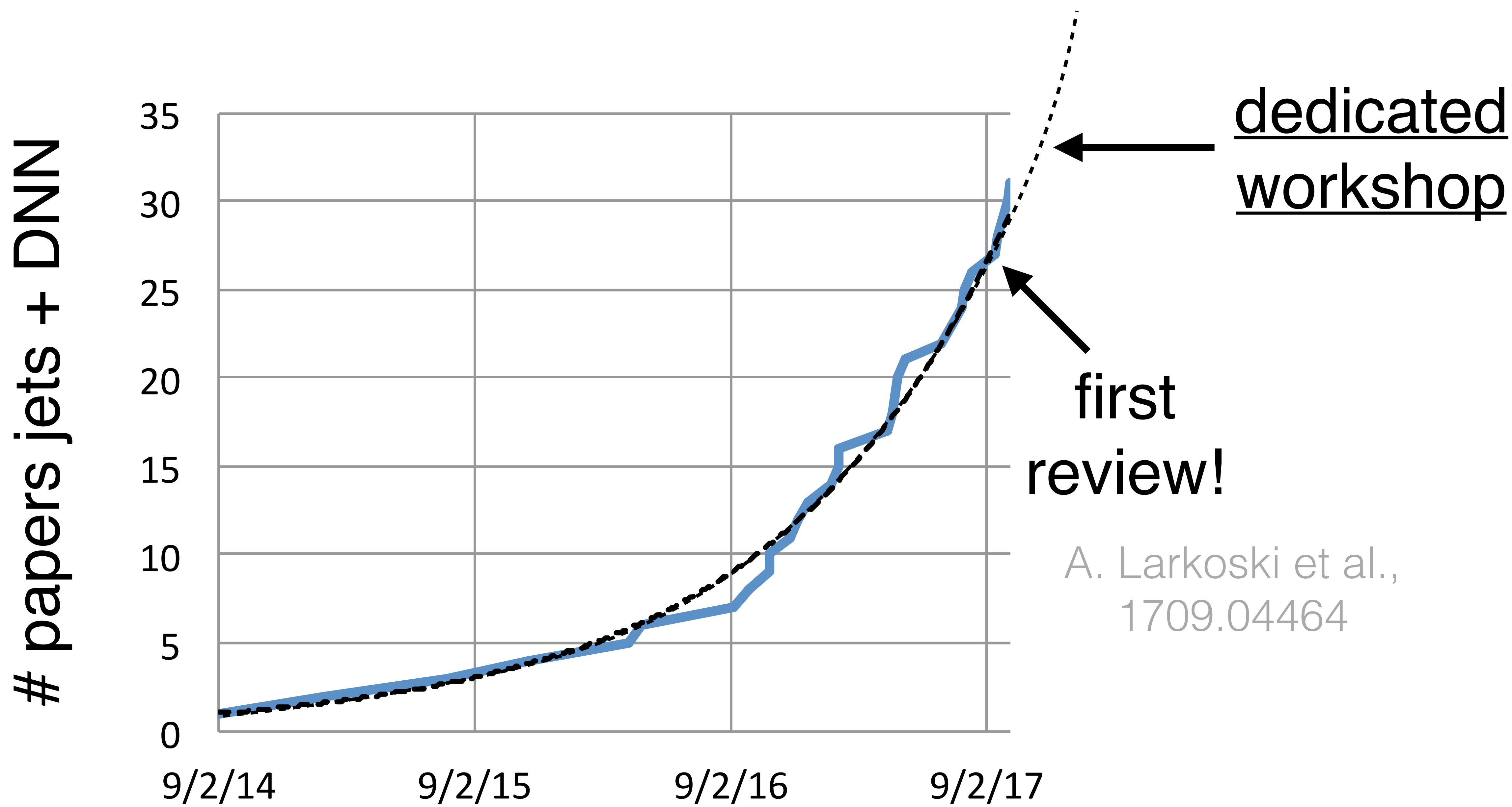




See also Markus' seminar

Exciting New Directions

So far only scratches the surface
…this is a very active field of research!

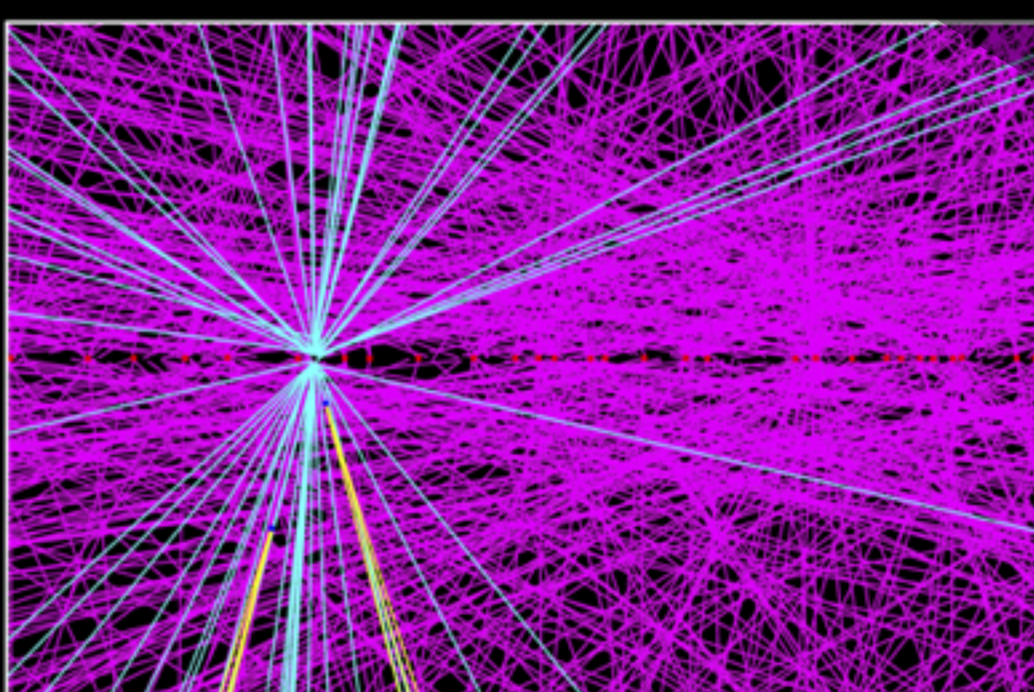


A. Larkoski et al., 1709.04464

Exciting New Directions I: Removing Noise

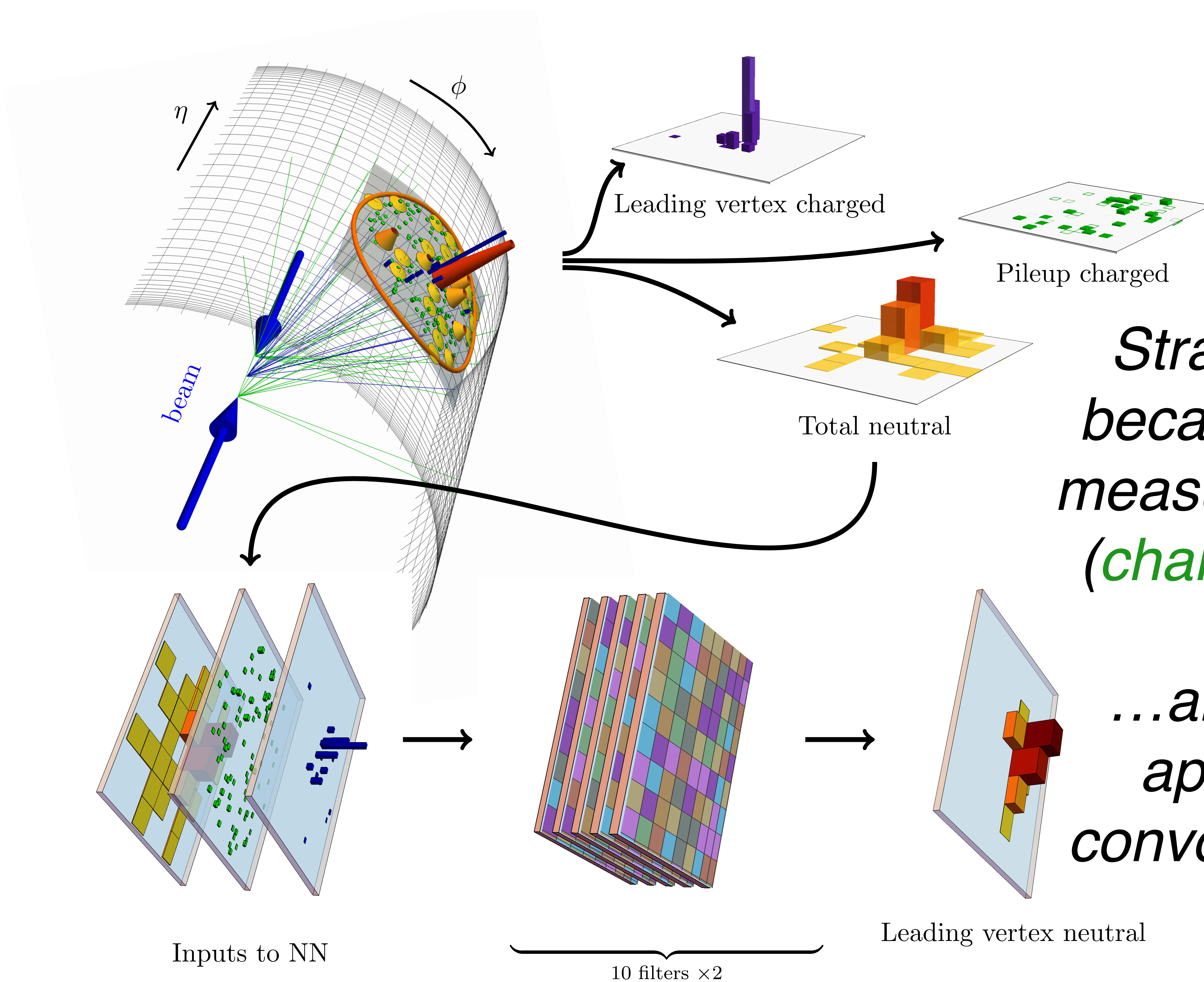
pp collisions at the LHC don't happen one at a time!

the extra collisions are called **pileup** and add soft radiation on top of our jets



this is akin to image de-noising - we can use ML for that!

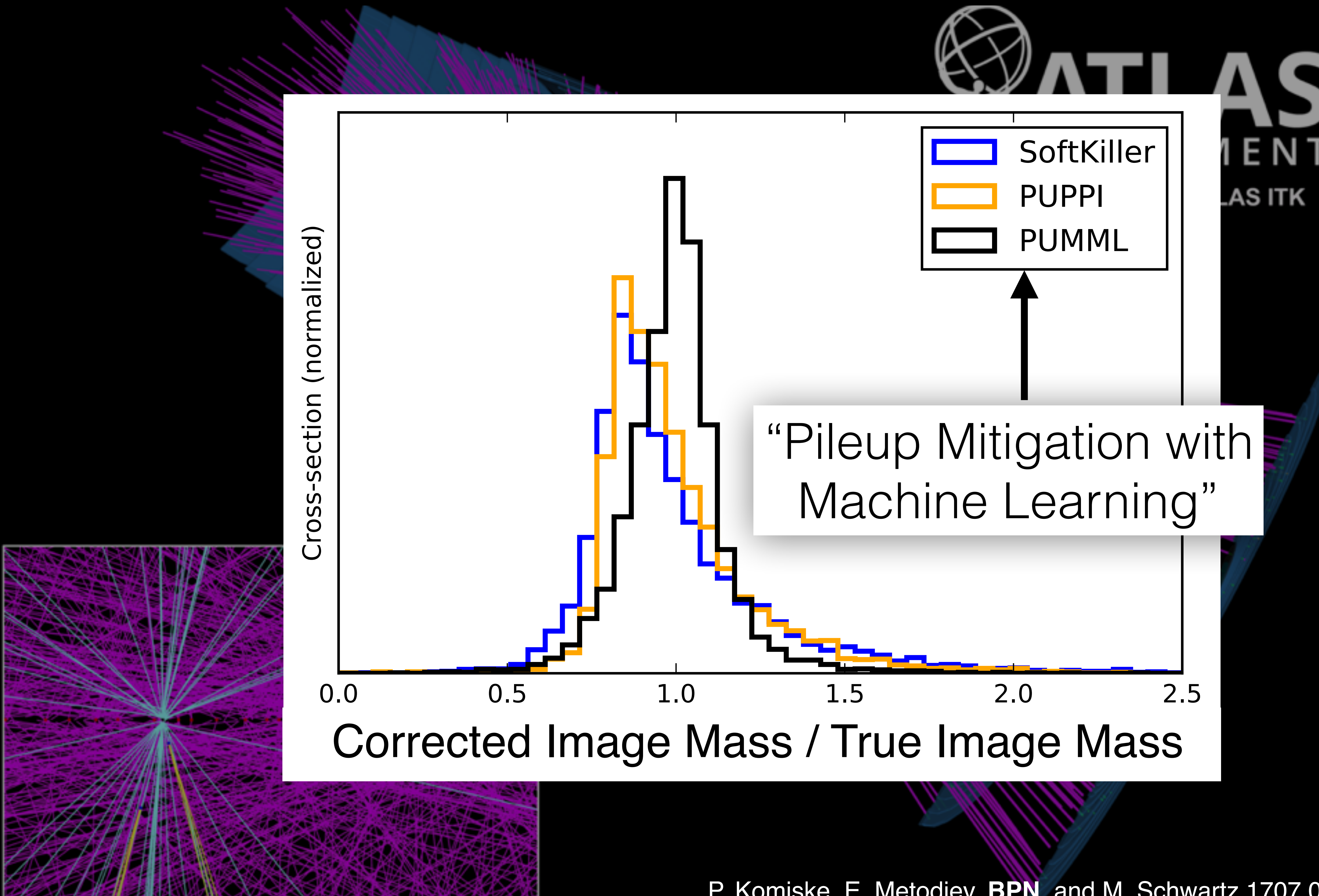
Exciting New Directions I: Removing Noise



Strange noise because we can measure ~2/3 of it (charged pileup)

…also a natural application of convolutional NNs!

Exciting New Directions I: Removing Noise

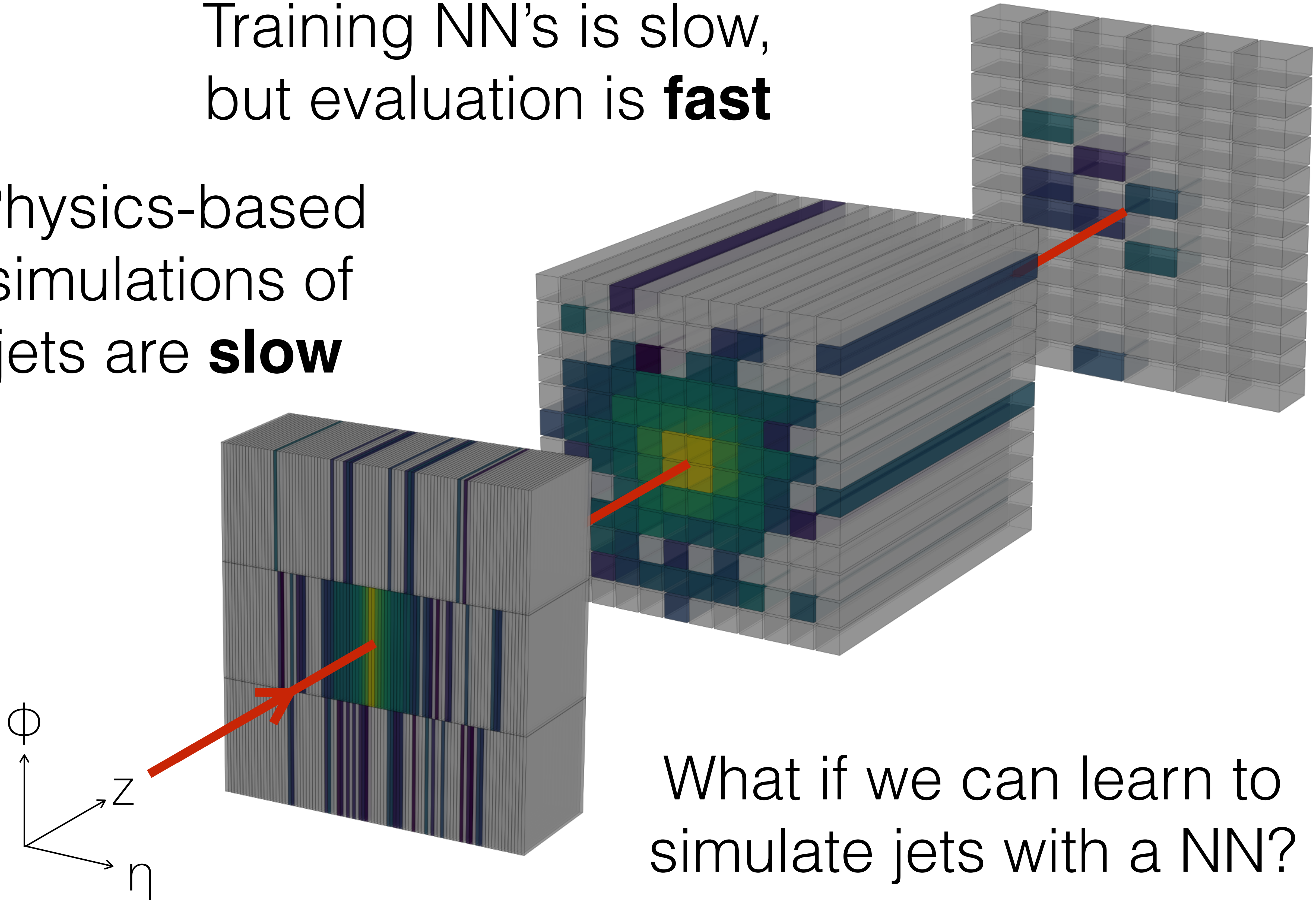


P. Komiske, E. Metodiev, **BPN**, and M. Schwartz,1707.08600

Exciting New Directions II: Simulation NN

Training NN's is slow, but evaluation is **fast**

Physics-based simulations of jets are **slow**



What if we can learn to simulate jets with a NN?

Exciting New Directions II: Simulation NN

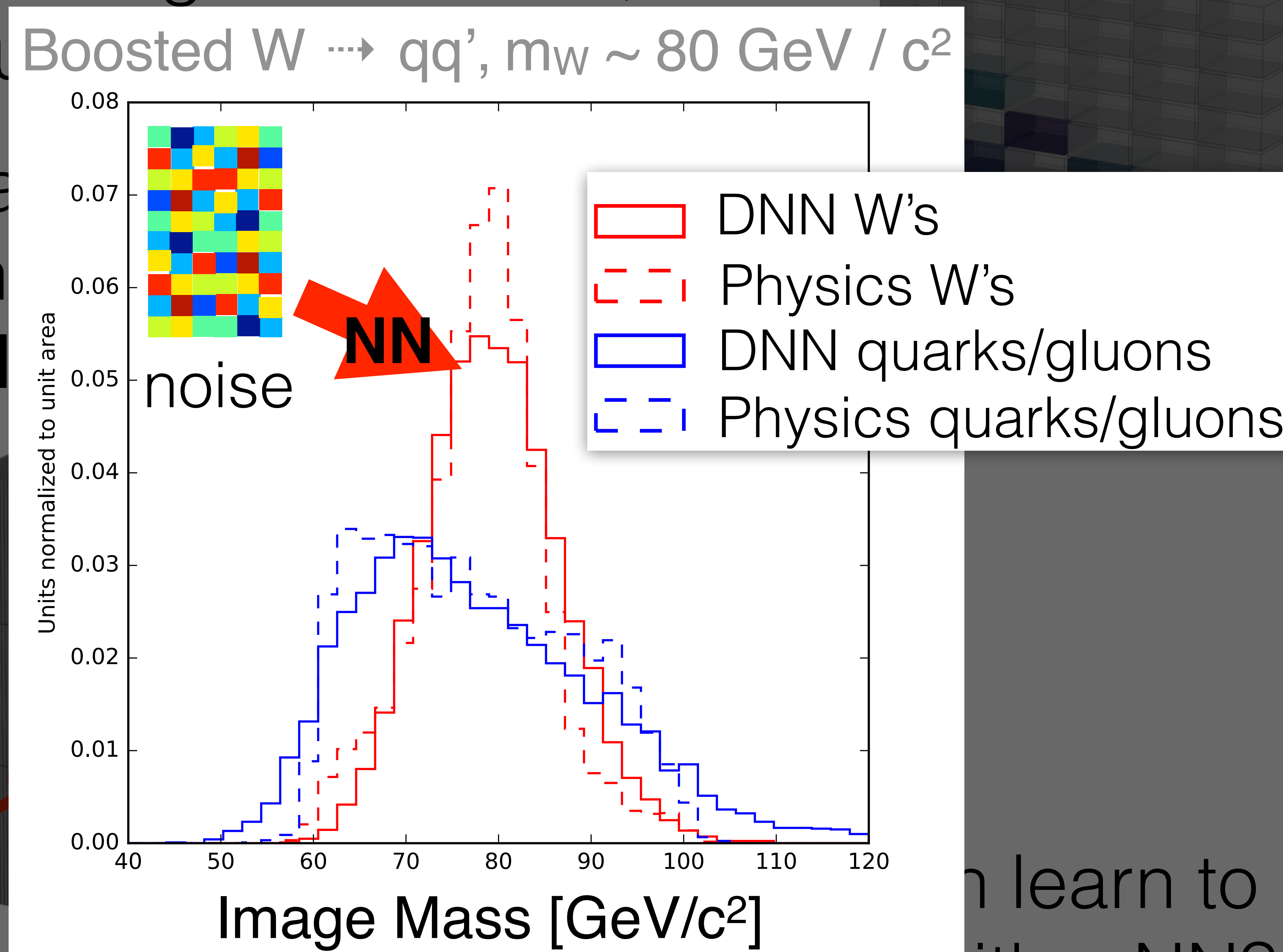
Training NN's is slow,

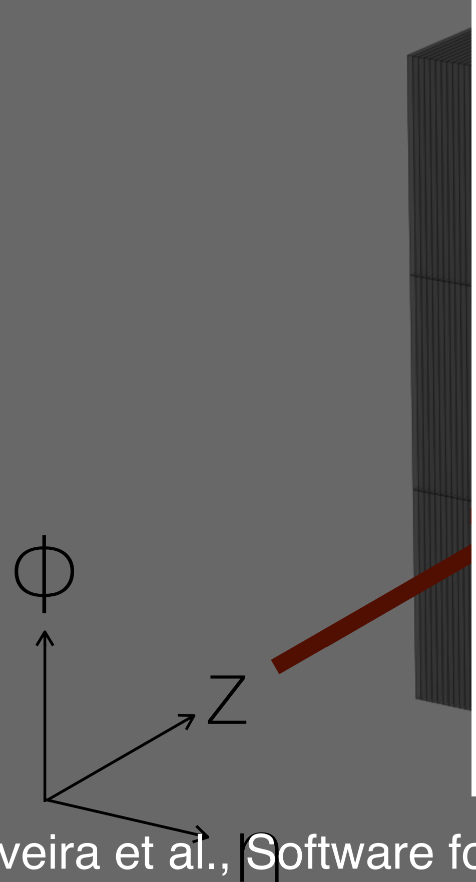
bu

Physics-ba

simulation

jets are **sl**





n learn to

simulate jets with a NN?

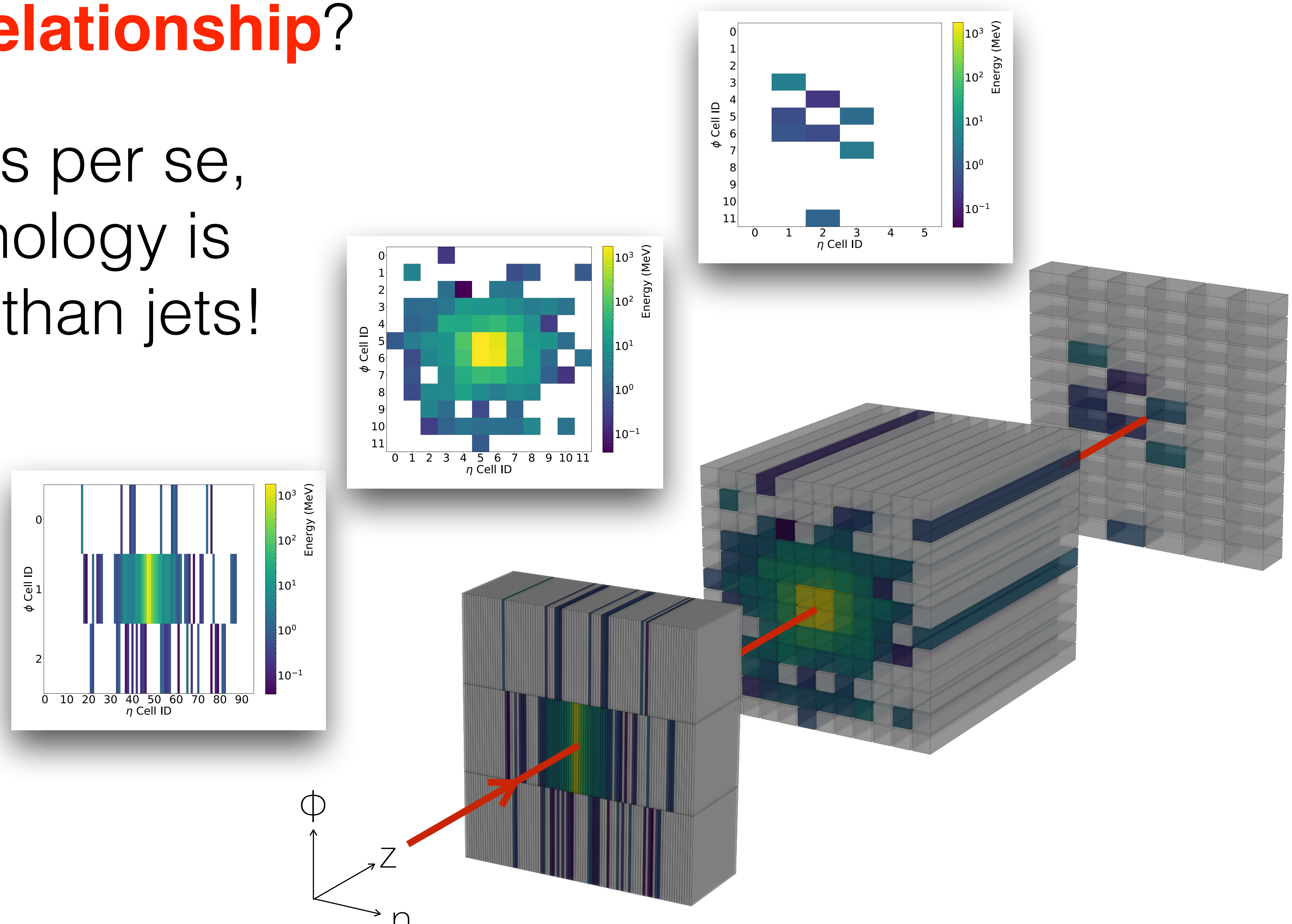
L. de Oliveira et al., Software for Big Science (2017) 1

M. Paganini et al.,1705.02355

+ More Layers for Generation

What about **multiple layers** with **non-uniform granularity** and a **causal relationship**?

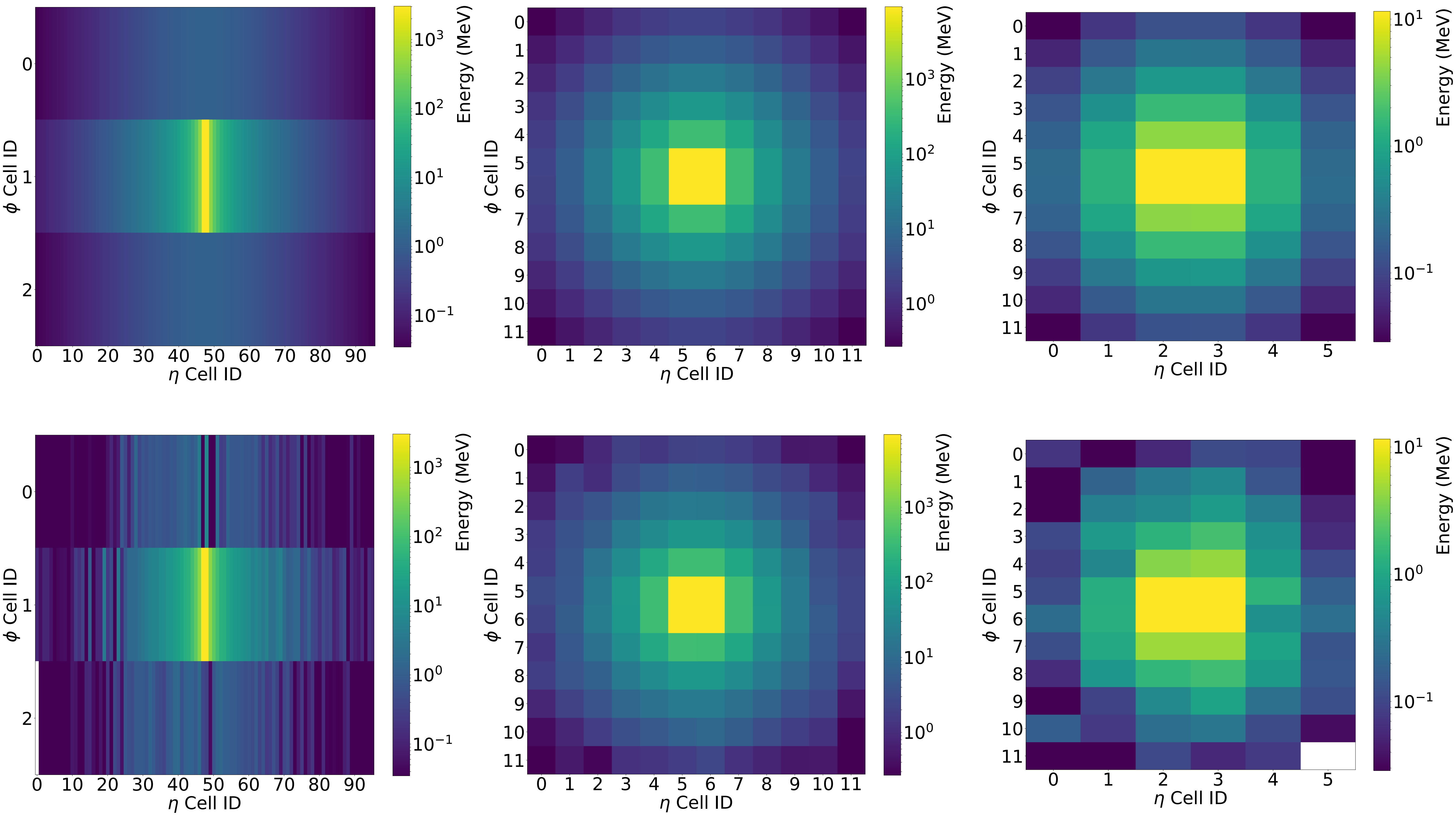
Not jet images per se, but the technology is more general than jets!



M. Paganini et al., 1705.02355

Average Images

Geant4



CaloGAN

M. Paganini et al., 1705.02355

Timing

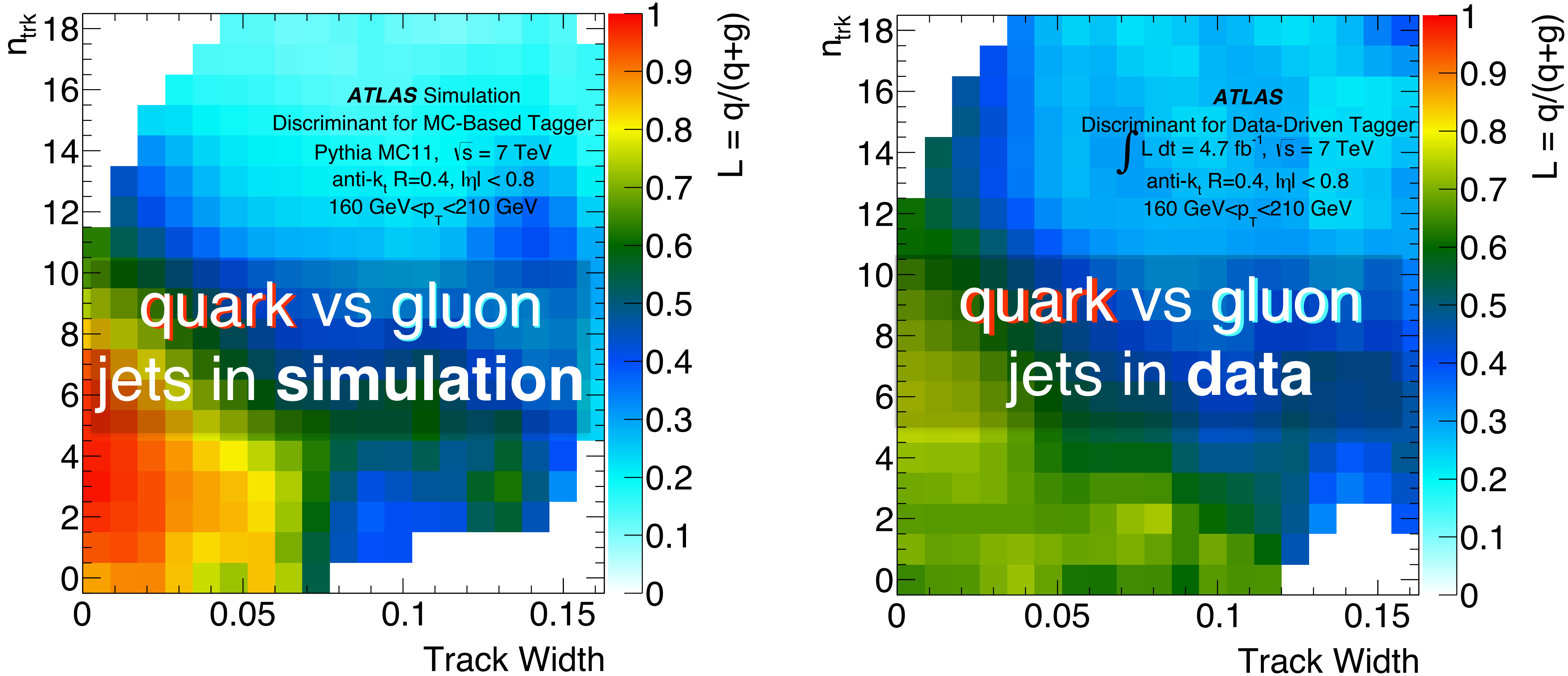
M. Paganini et al., 1705.02355

Generation Method	Hardware	Batch Size	milliseconds/shower
GEANT4	CPU	N/A	1772 ←
CALOGAN	CPU	1	13.1
		10	5.11
		128	2.19
		1024	2.03
	GPU	1	14.5
		4	3.68
		128	0.021
		512	0.014
		1024	0.012 ←

See also S. Vallecorsa et al. (GeantV), C. Guthrie et al. (NYU), W. Wei et al. (LCD dataset group), D. Salamani et al. (Geneva), D. Rousseau et al. (Orsay), L. de Oliveira et al. (Berkeley)

Where next III: Learning directly from data

For supervised learning, we depend on labels
labels usually come from simulation

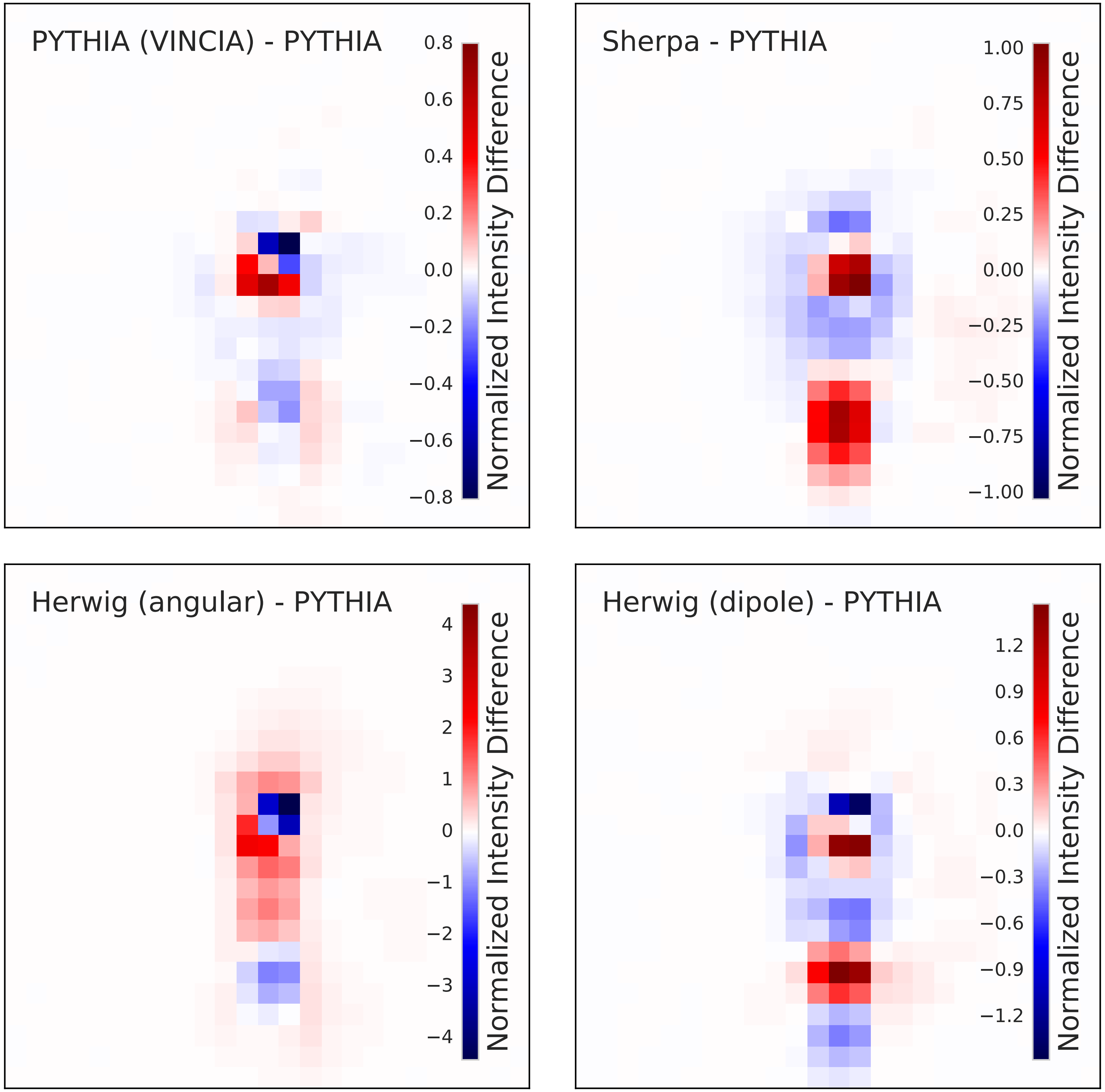


What if data and simulation are very different?

…your classifier will be sub-optimal

Where next III: Learning directly from data

Boosted *W* boson jets



N.B. not all of these have been tuned to the same data

J. Barnard et al.
Phys. Rev. D 95, 014018 (2017)

DNN classifiers can **exploit** subtle features

subtle features are **hard to model** !

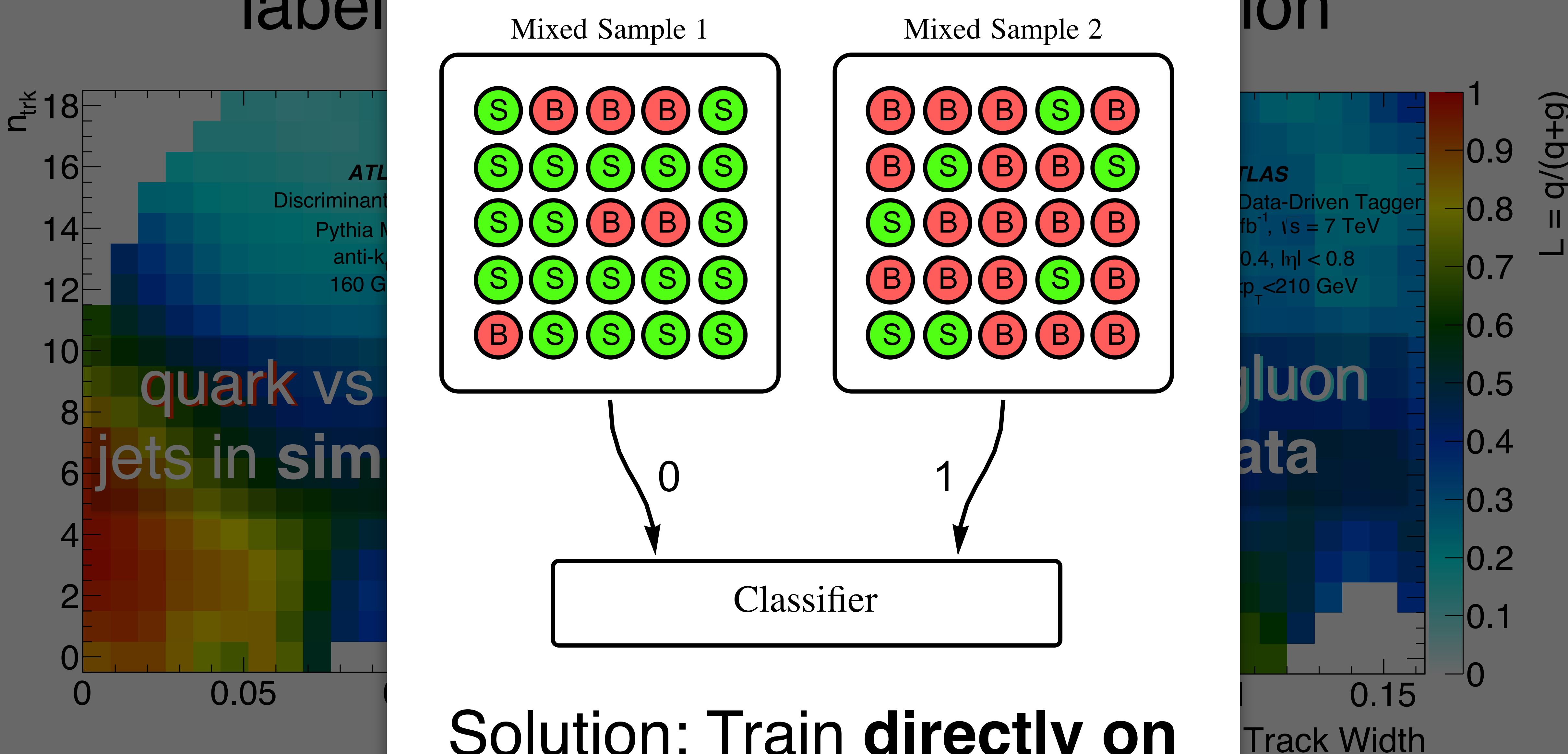
we need to be careful about which models we use - **only data is correct**

For a mixed approach, see G. Louppe et al.

Where next III: Learning directly from data

For supervised learning, we depend on labels

label ion



Solution: Train **directly on data** using mixed samples

What if da ifferent?

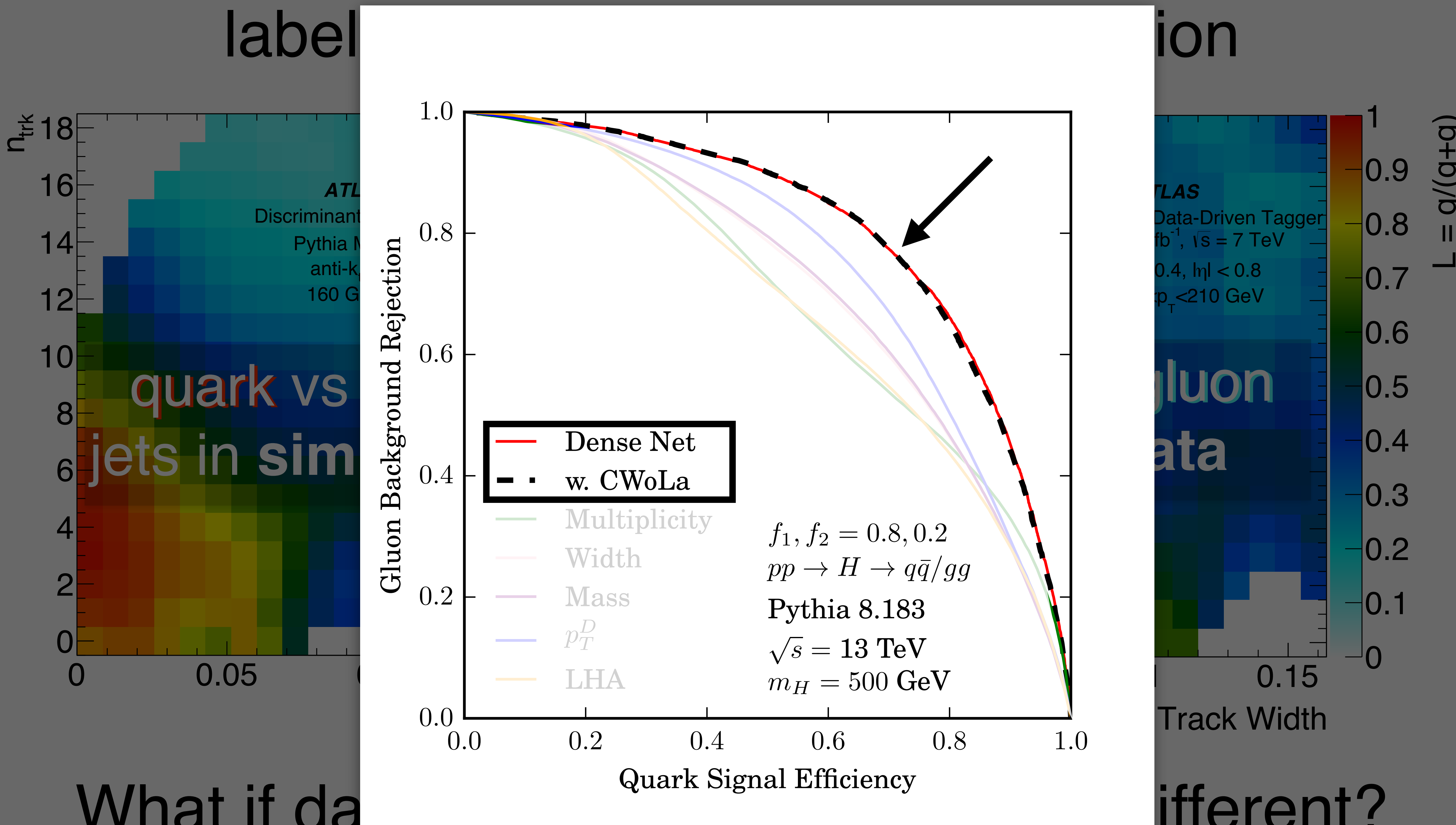
…your classifier will be sub-optimal

L. Dery et al., JHEP 05 (2017) 145
E. Metodiev et al., JHEP 10 (2017) 174

Where next III: Learning directly from data

For supervised learning, we depend on labels

label   ion

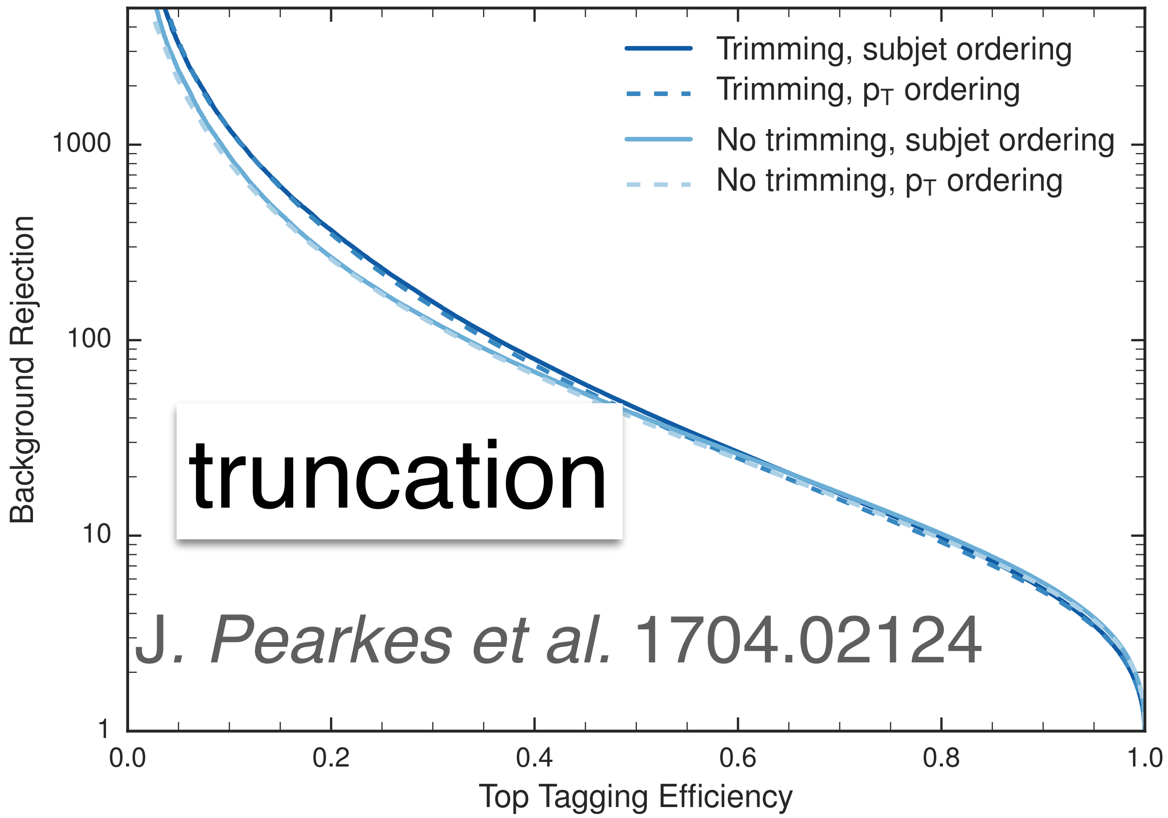


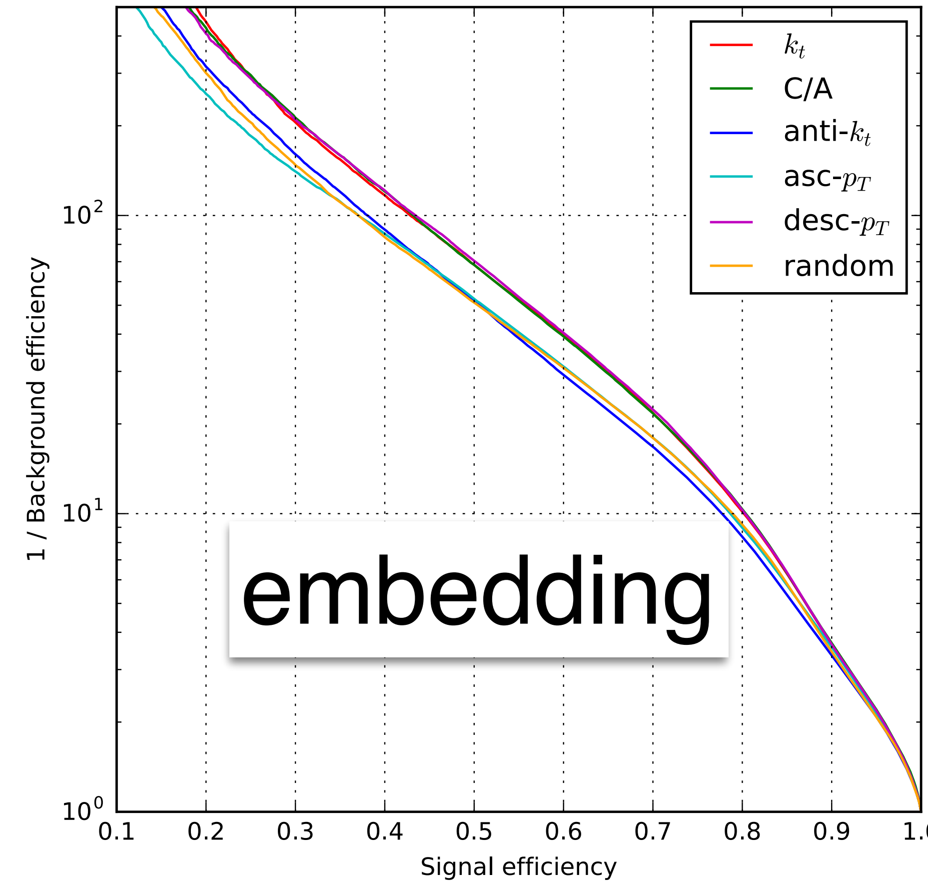
What if da   ifferent?

…your classifier will be sub-optimal

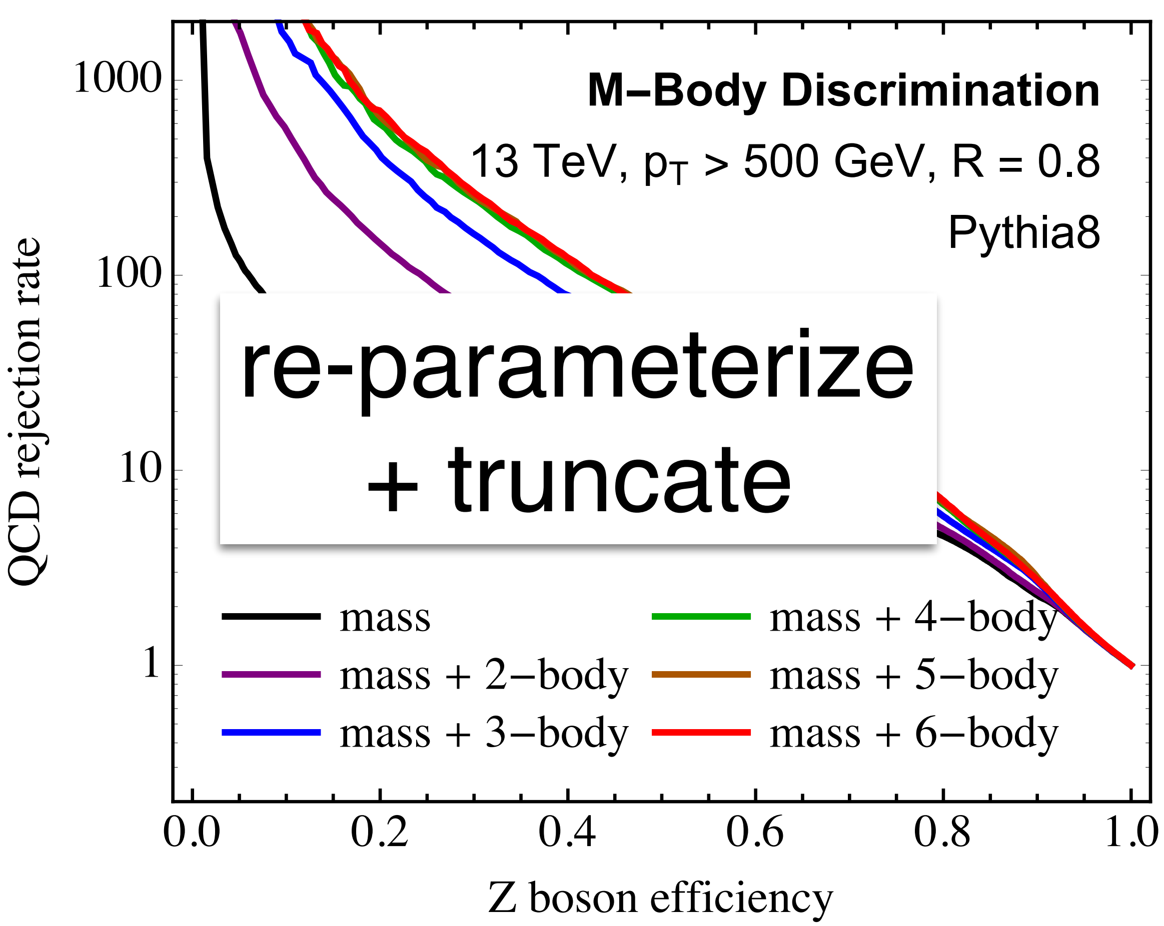
E. Metodiev et al., JHEP 10 (2017) 174

Beyond Images





G. Louppe et al. 1702.00748



K. Datta et al. 1704.08249

A. Butter *et al.* 1707.08966 (truncate + augment/embed)

K. Datta et al. 1710.01305 (re-param)

T. Cheng 1711.02633 (RNN)

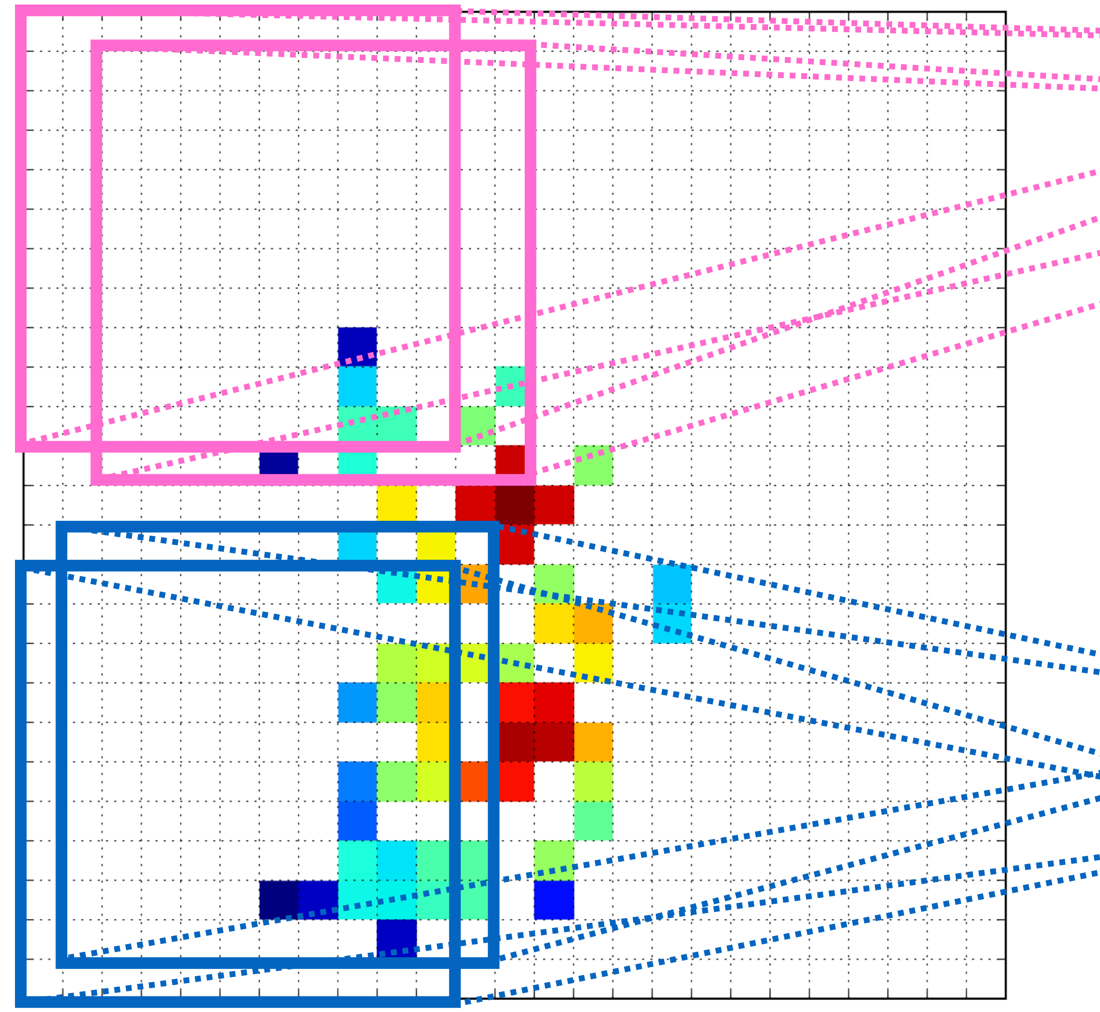
J. A. Aguilar-Saavedra et al. 1709.01087 (re-param)

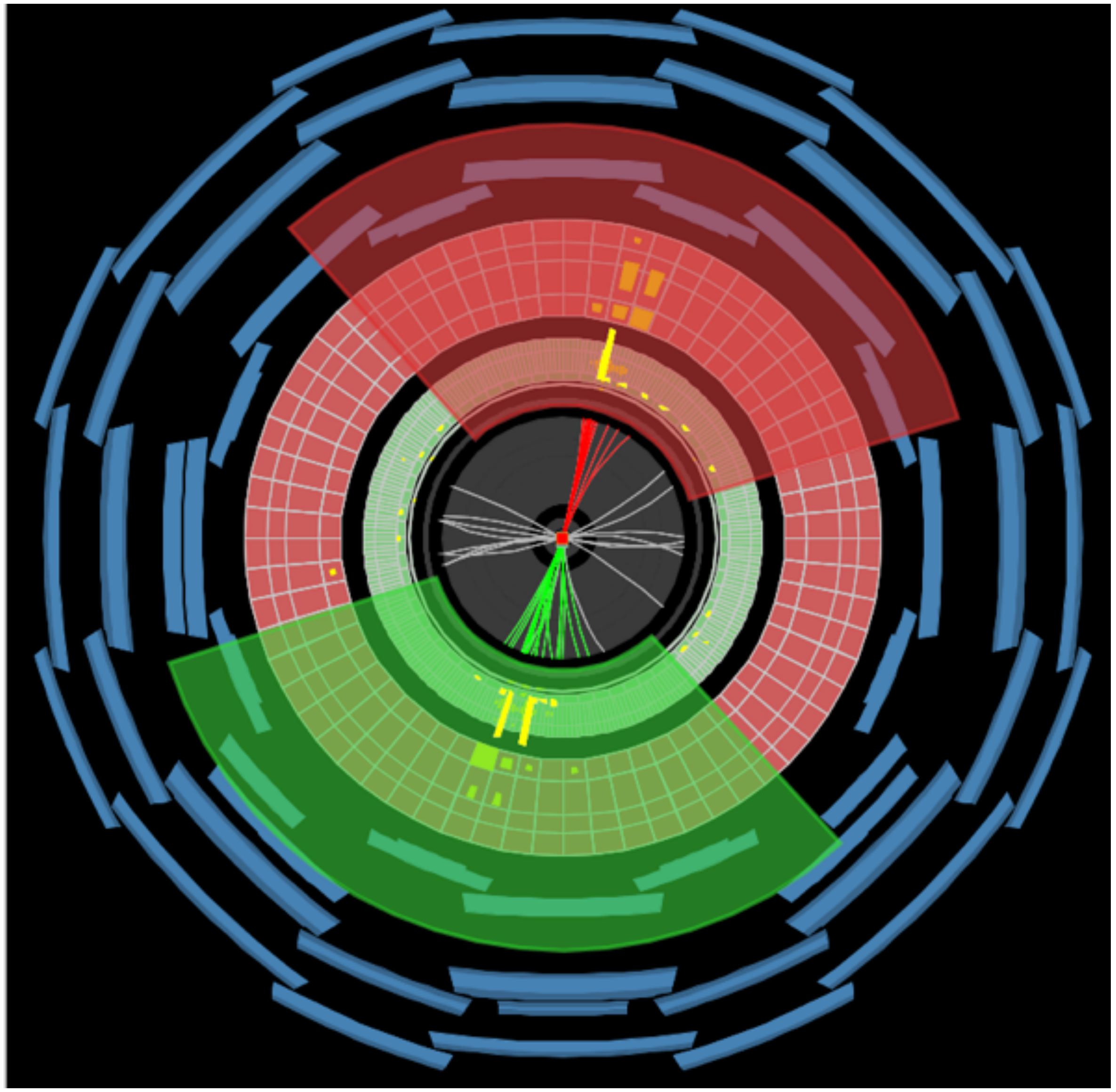
+ flavor tagging (see backup)

+ many more results at the <u>dedicated workshop</u> next month!

Conclusions and outlook

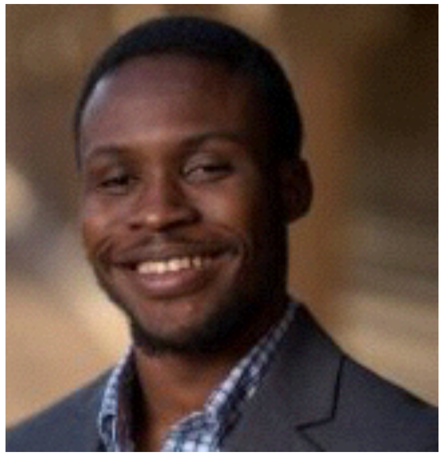
(Jet) image-based NN classification, regression, and generation are powerful tools for fully exploiting the physics program at the LHC





The key to robustness is to study what is being learned; this may even help us to learn something new about nature!

Collaborators


Lucio
Dery
Stanford


Michela
Paganini
Yale


Eric
Metodiev
MIT


Patrick
Komiske
MIT


Zihao
Jiang
Stanford

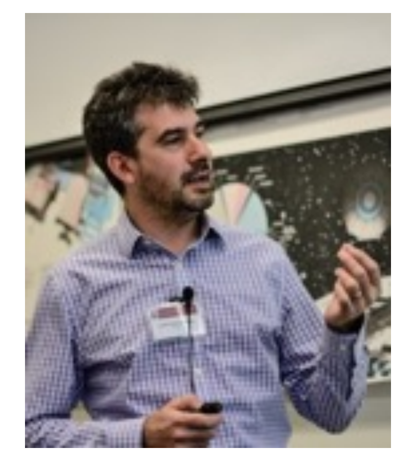

Francesco
Rubbo
SLAC

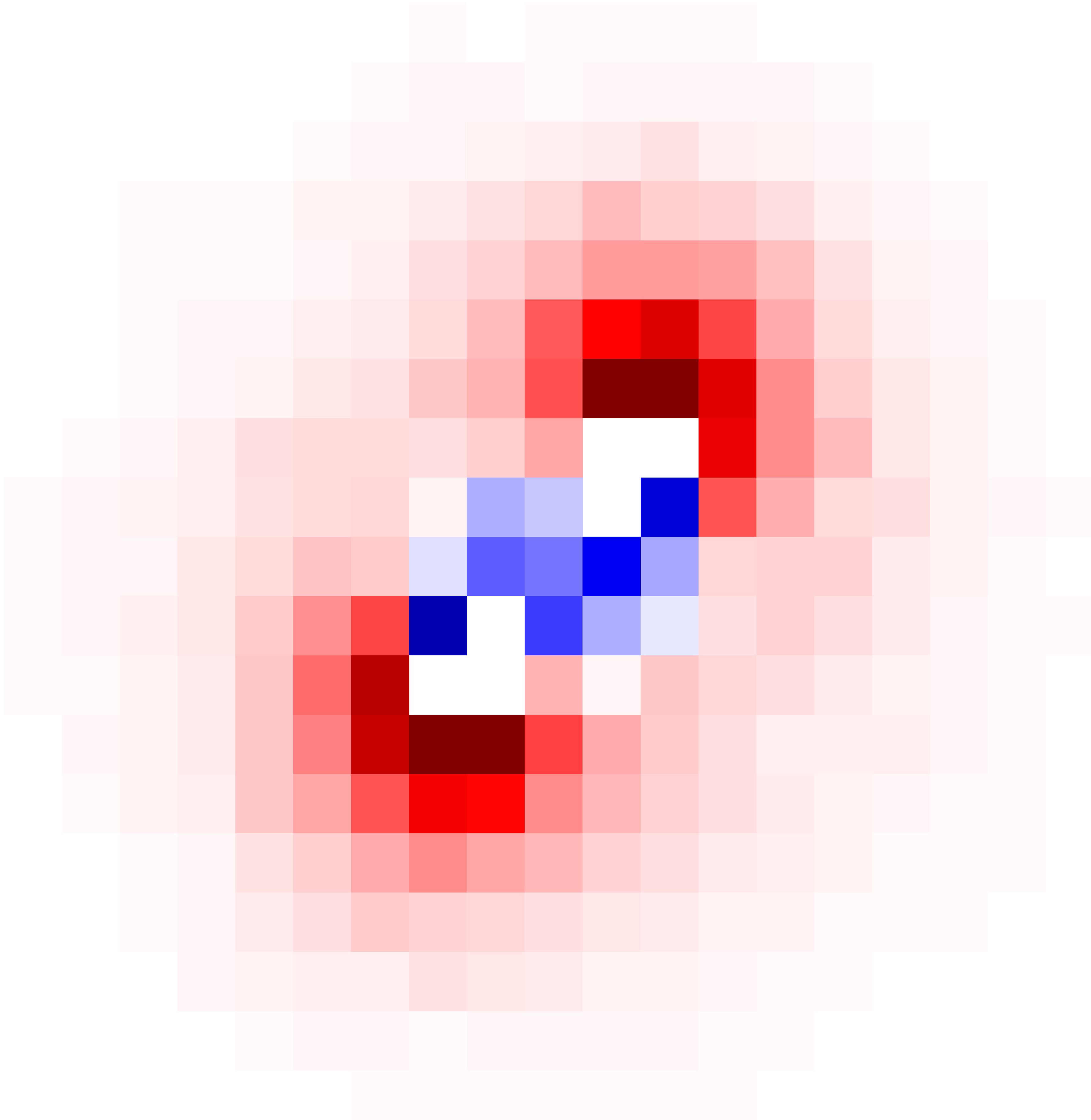

Luke
de Oliveira
VAI tech.


Michael
Kagan
SLAC


Jesse
Thaler
MIT


Matt
Schwartz
Harvard

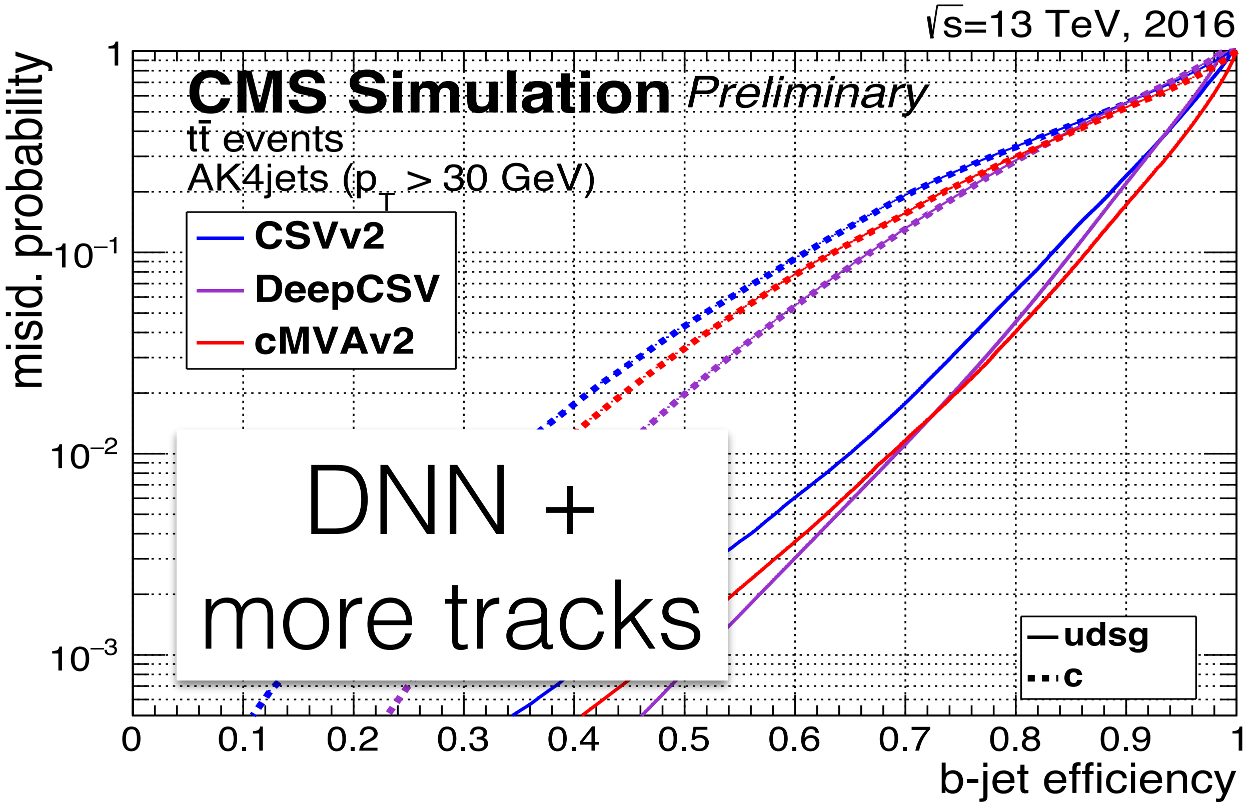

Ariel
Schwartzman
SLAC



Fin.

Backup

b-tagging in **CMS** and **ATLAS**

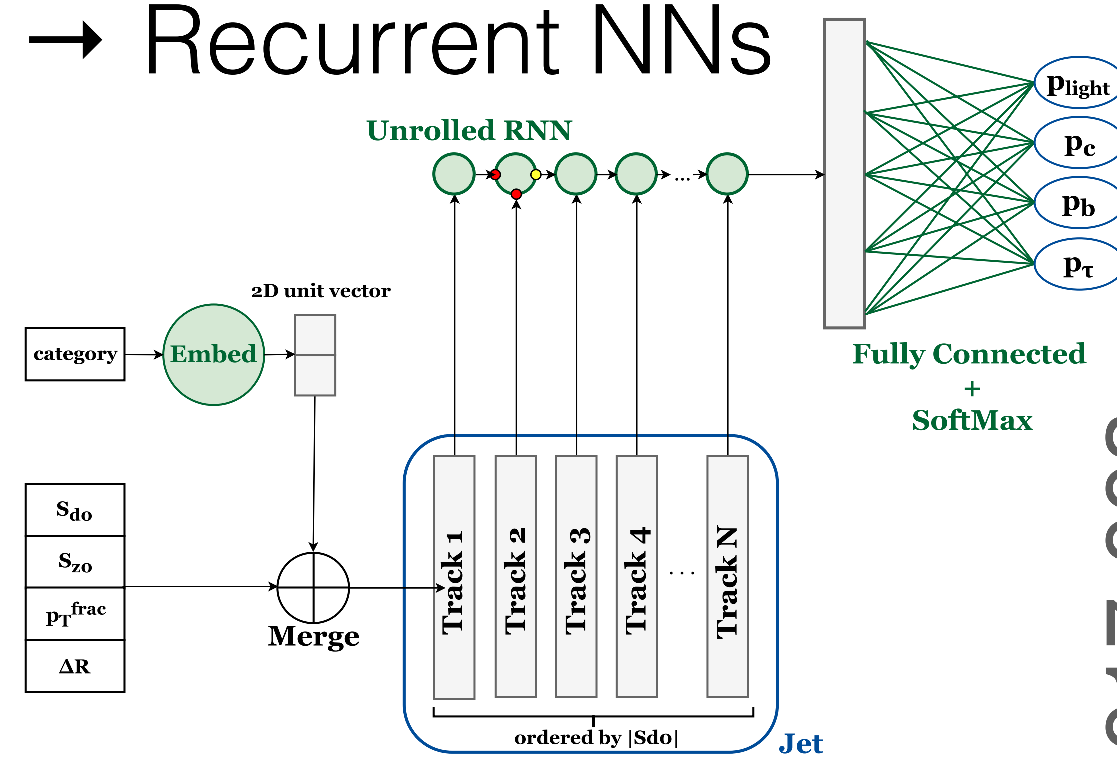


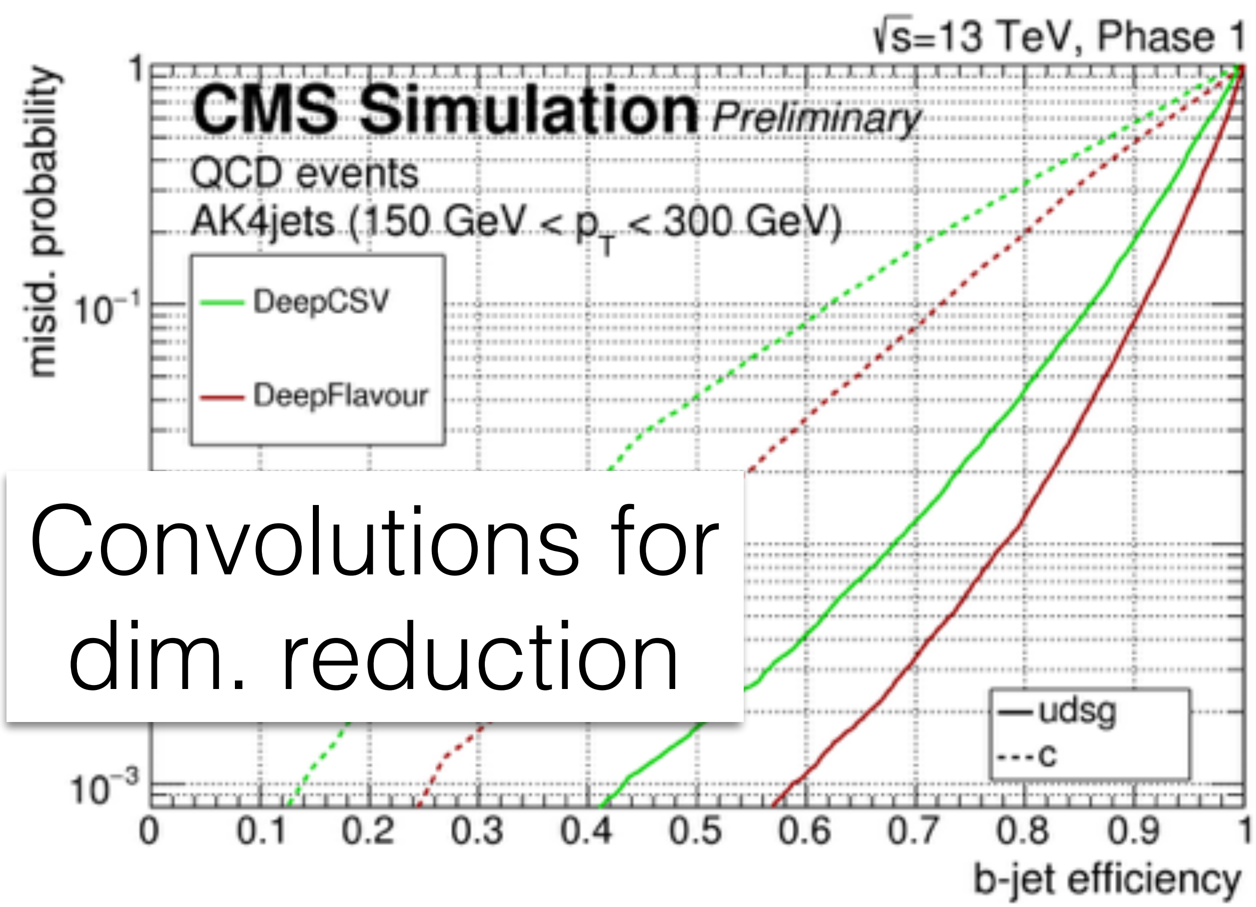
DNN + more tracks

CMS-DP-2017-005

Sequence of tracks

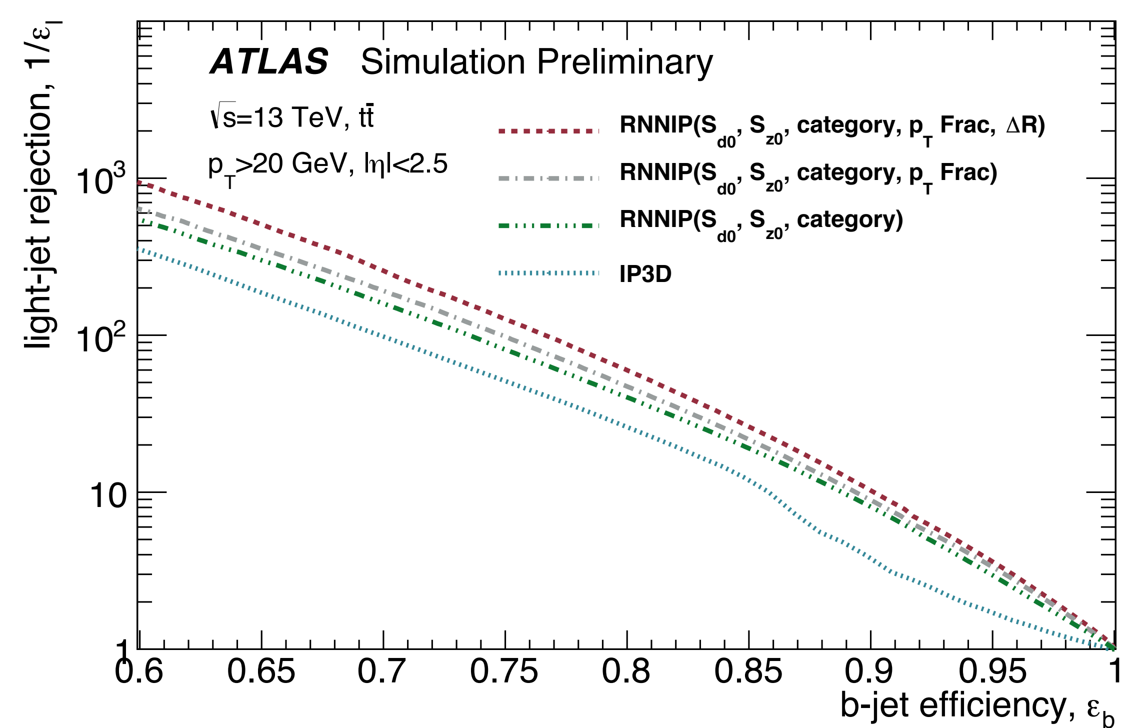
→ Recurrent NNs





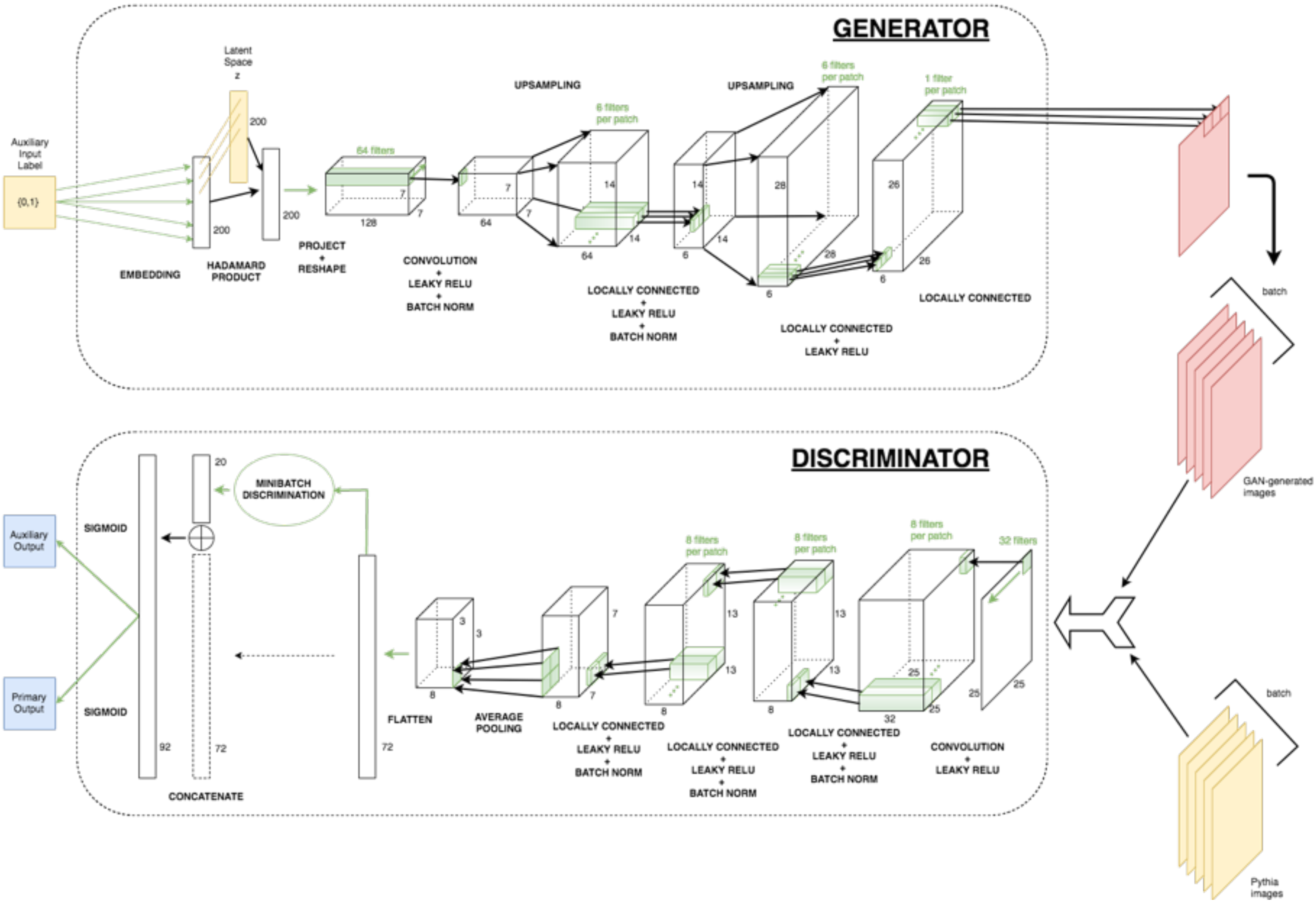
Convolutions for dim. reduction

CMS-DP-2017-013

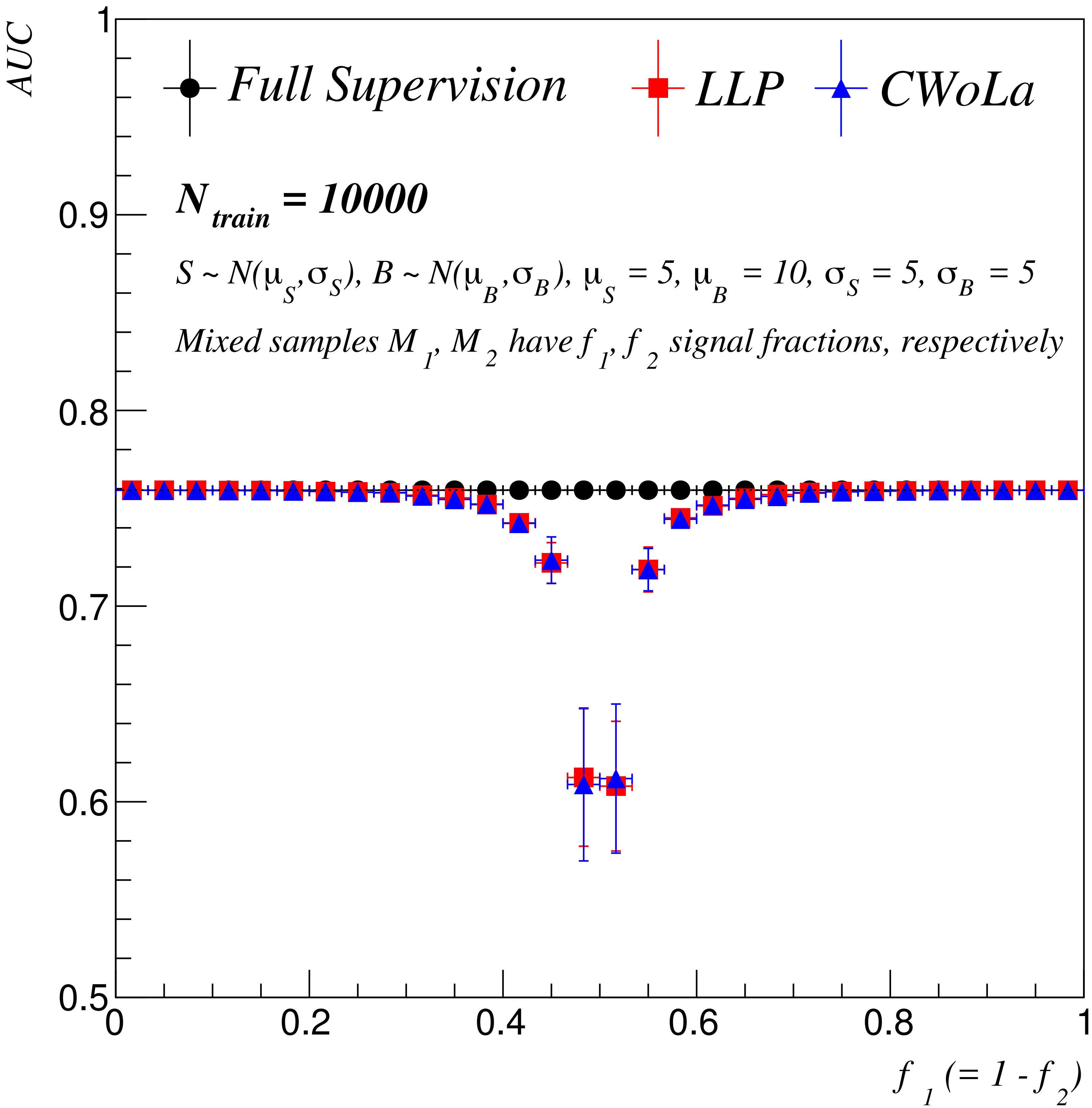


ATL-PHYS-PUB-2017-003

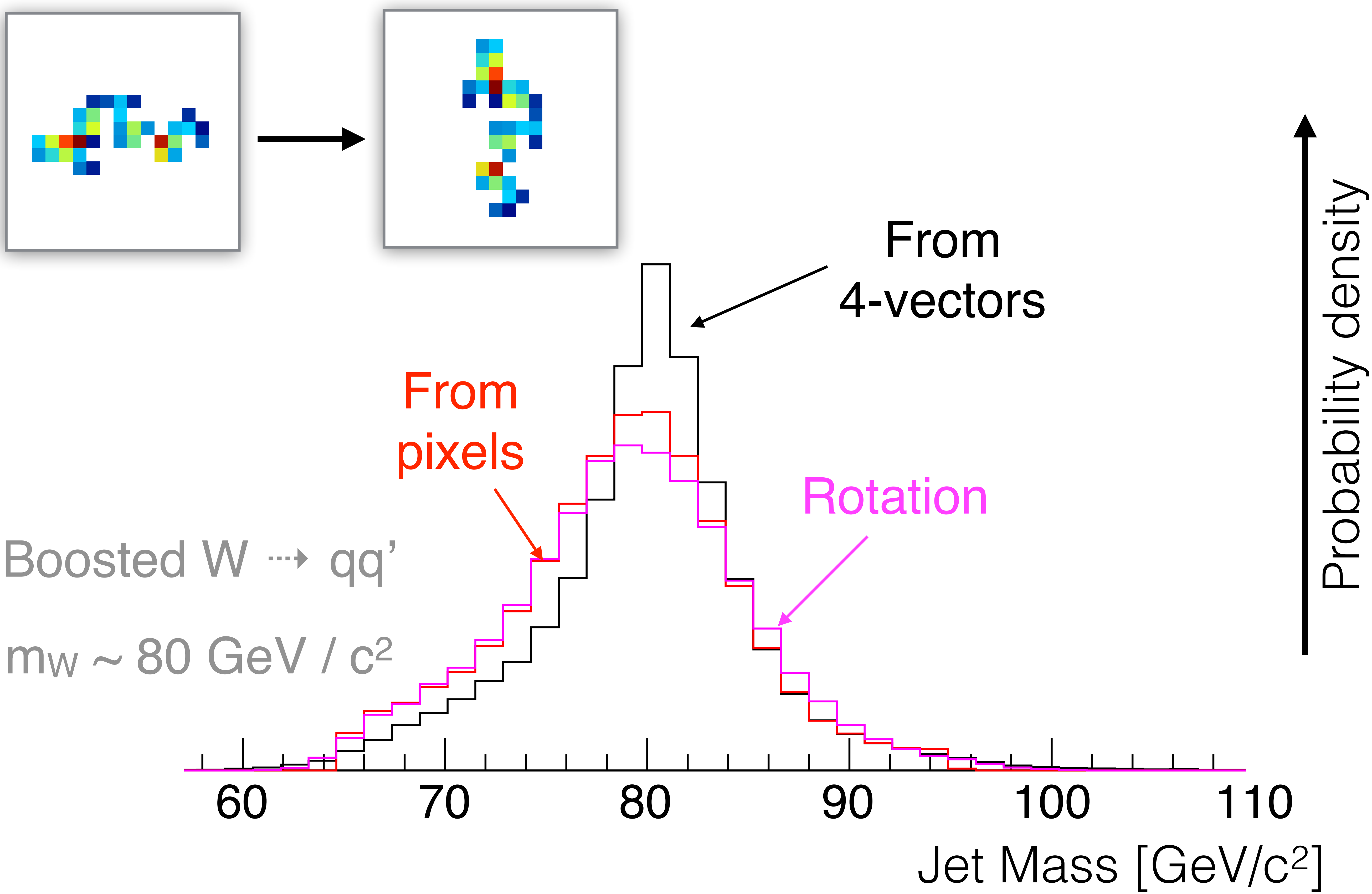
Locally Aware GAN (LAGAN)



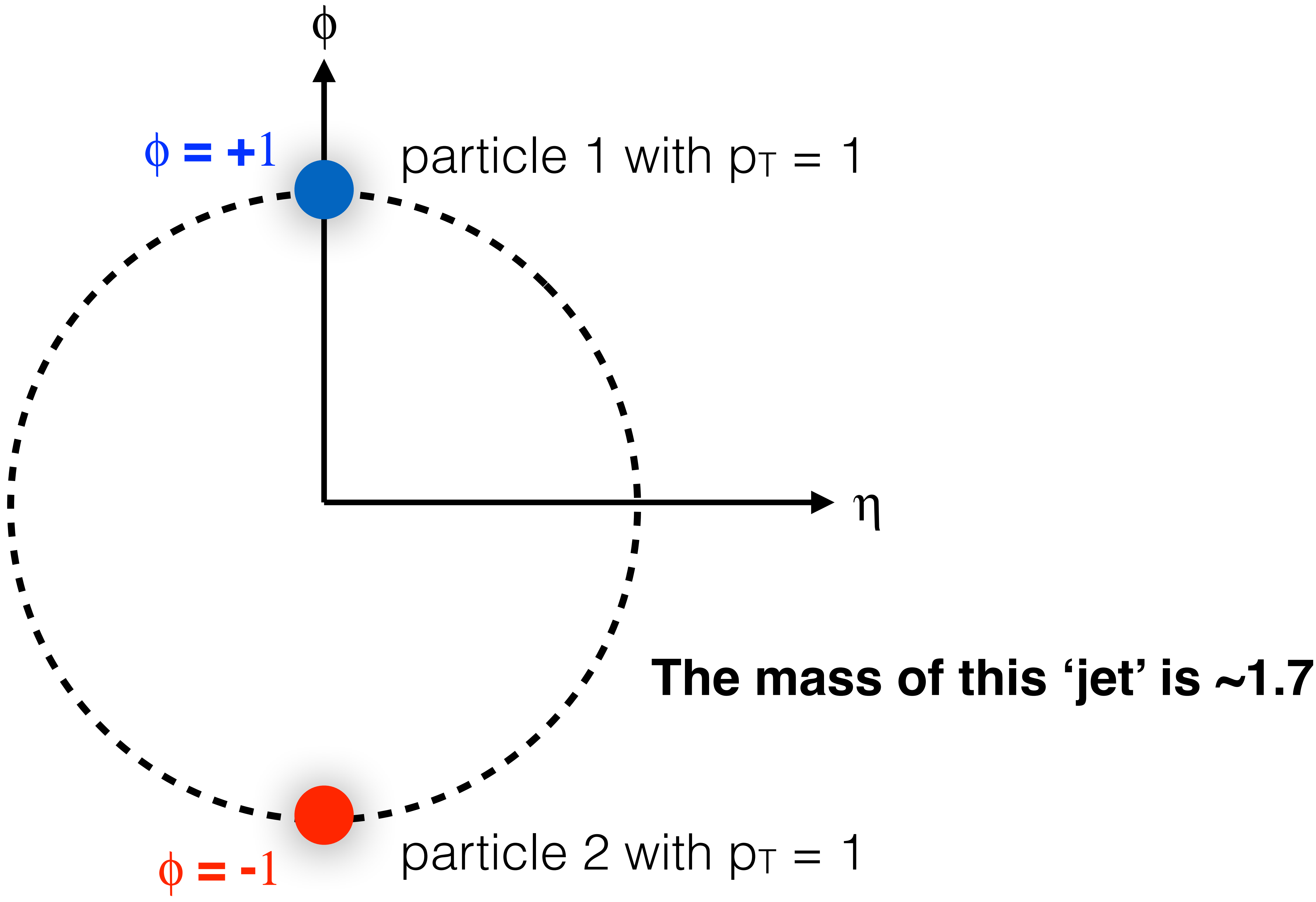
Learning when you know (almost) nothing



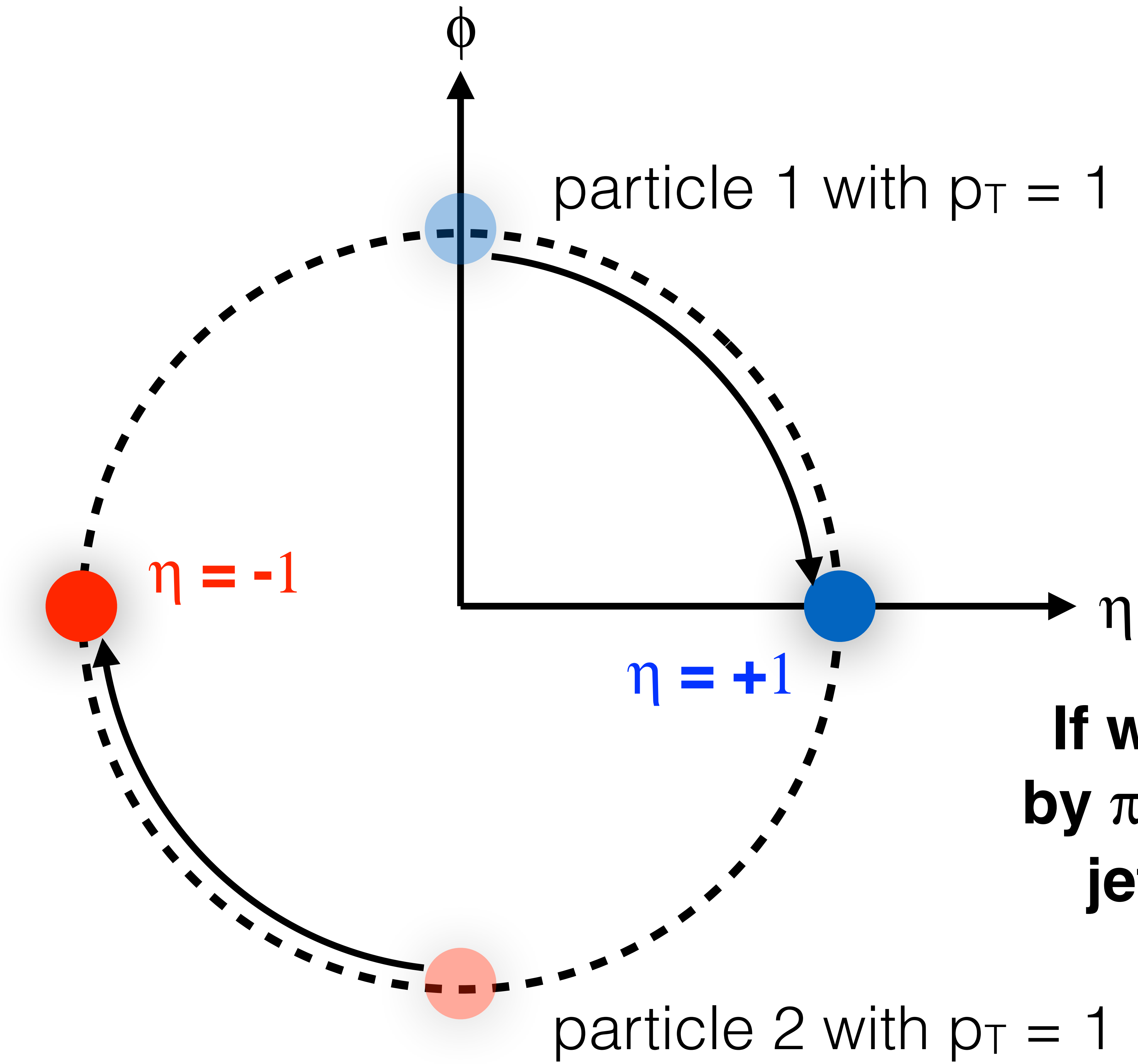
Pre-processing & spacetime symmetries



Pre-processing & spacetime symmetries



Pre-processing & spacetime symmetries

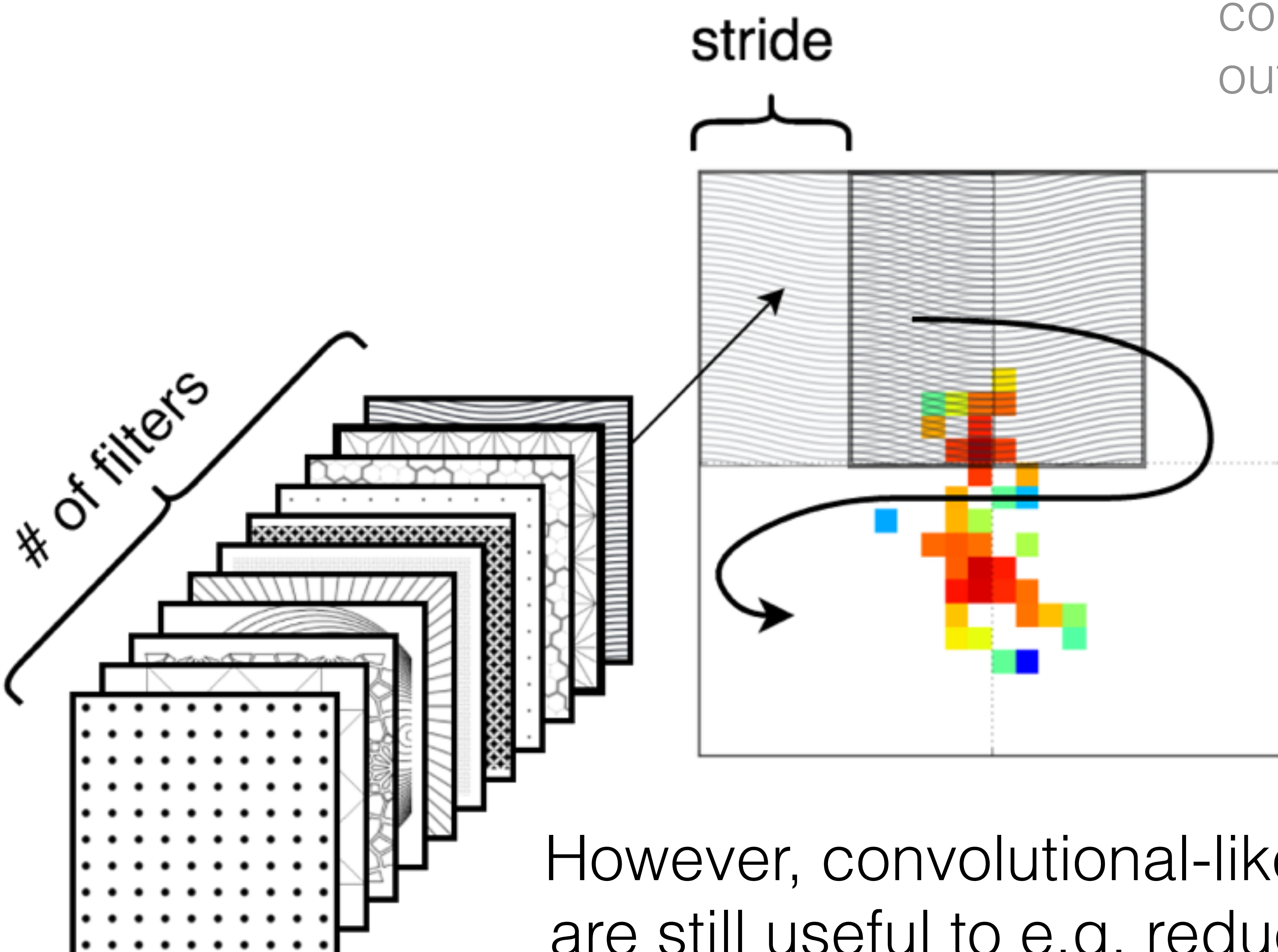


If we rotate the jet by $\pi/2$, then the new jet mass is ~2.4

Locally Connected Layers

Due to the structure of the problem, we do not have translation invariance.

Classification studies found fully connected networks outperformed CNNs

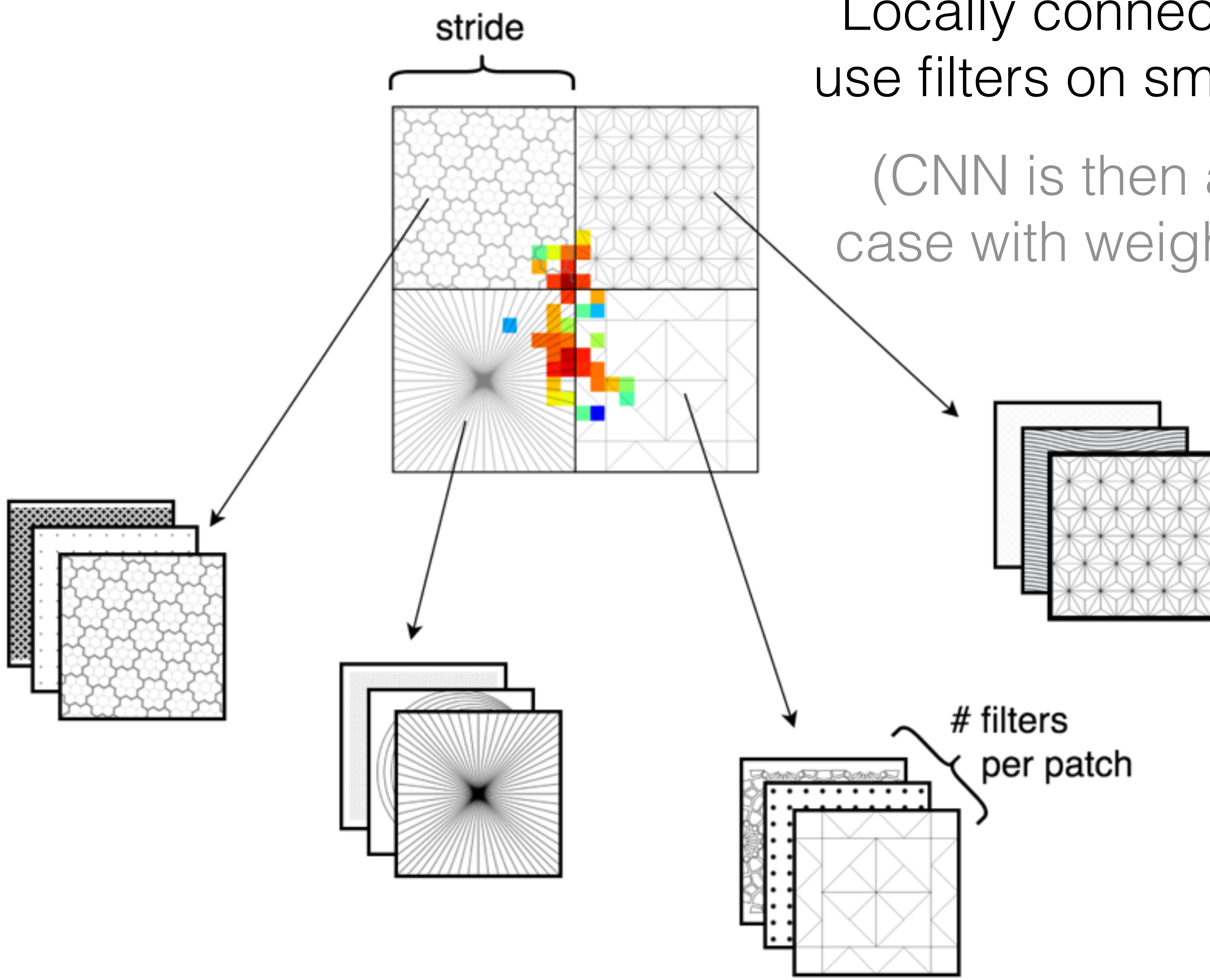


However, convolutional-like architectures are still useful to e.g. reduce parameters

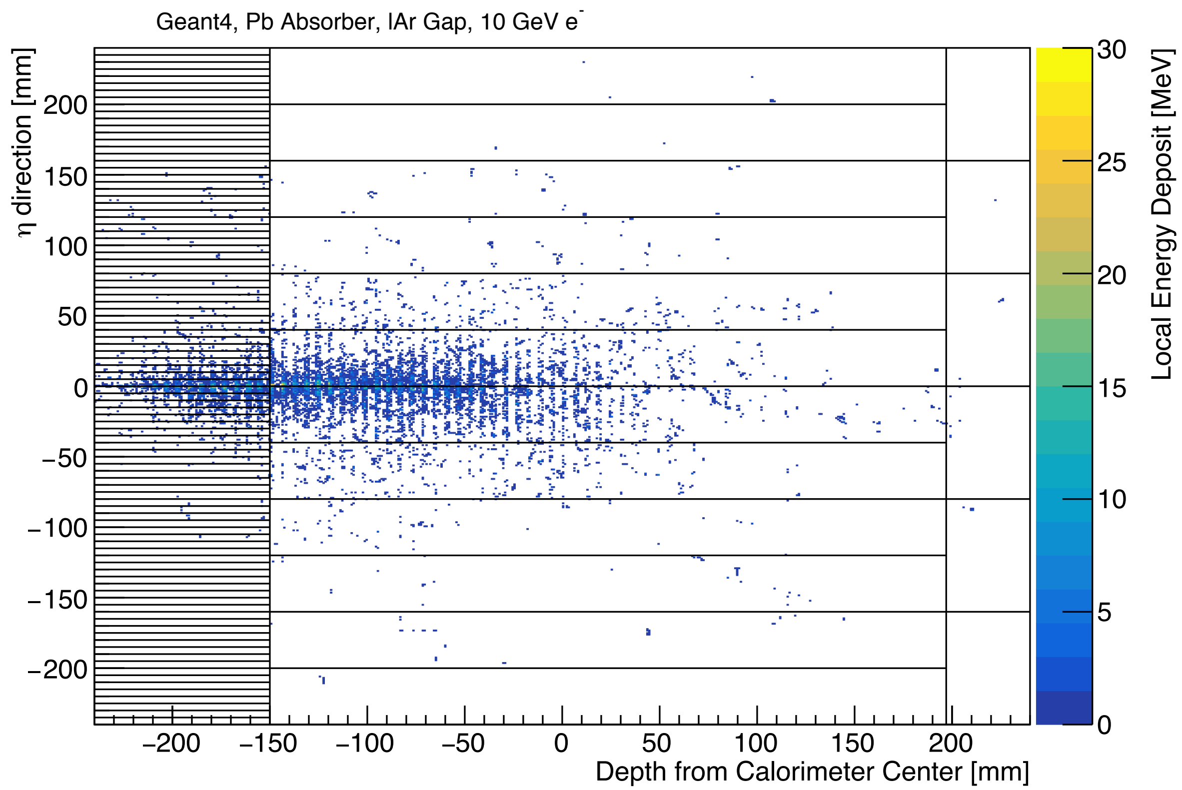
Locally Connected Layers

Locally connected layers use filters on small patches

(CNN is then a special case with weight sharing)



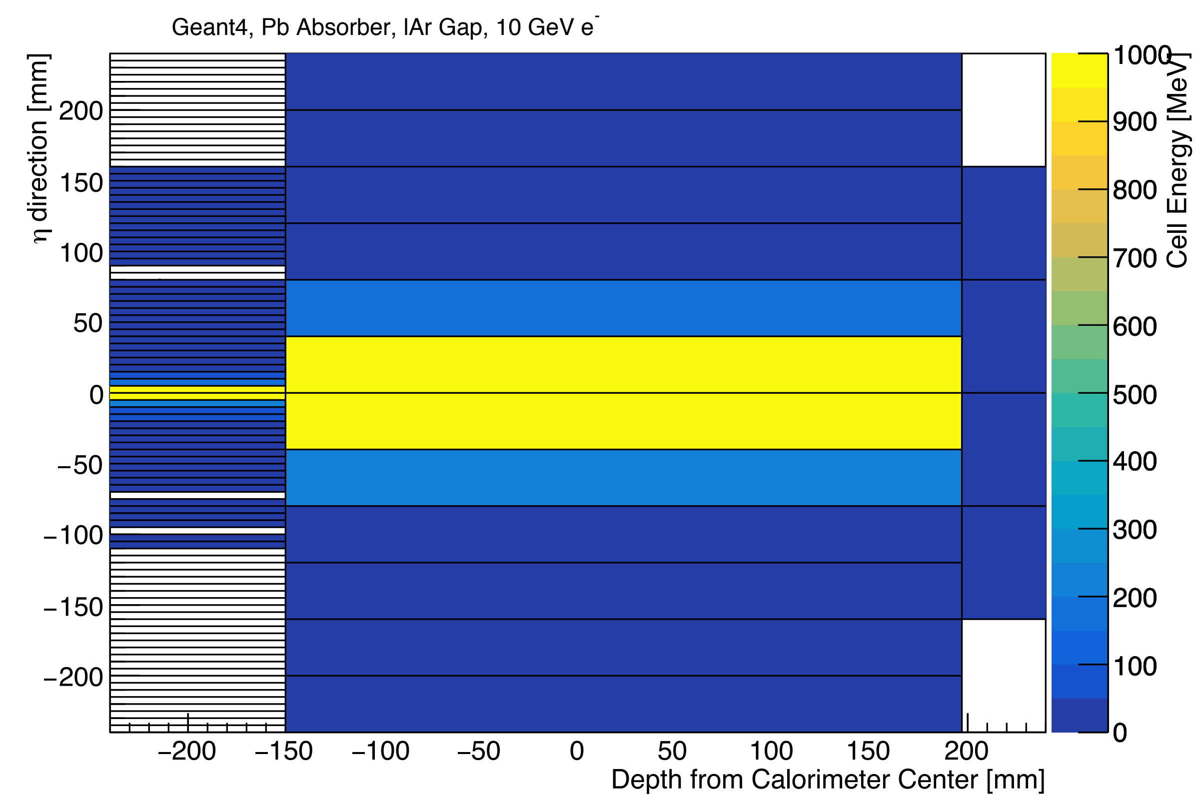
Calorimeter Simulation



We take as our model a 3-layer LAr calorimeter, inspired by the ATLAS barrel EM calorimeter

A single event may have $O(10^3)$ of particles showering in the calorimeter - too cumbersome to do all at once (now)

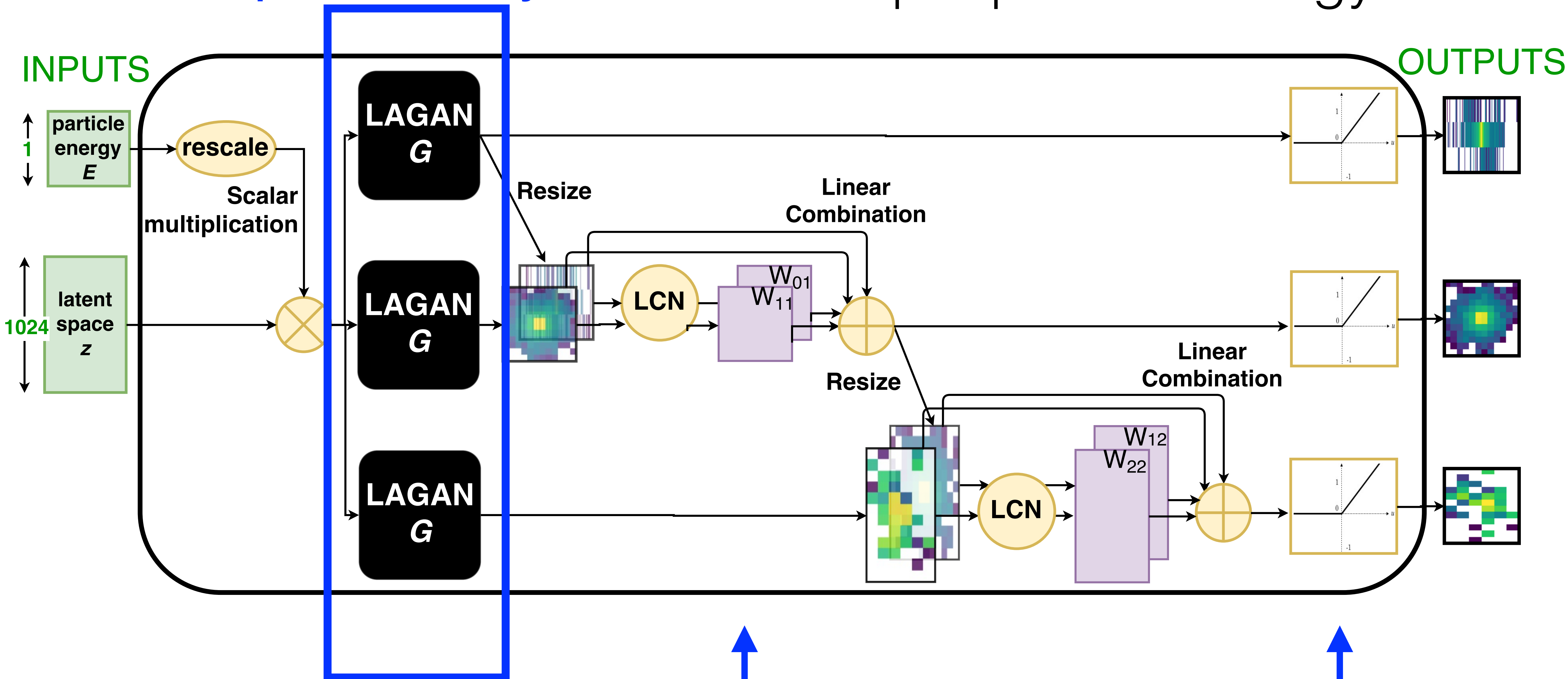
We exploit factorization of energy depositions



Generator Network for CaloGAN

One 'jet image' per calo layer

One network per particle type; input particle energy

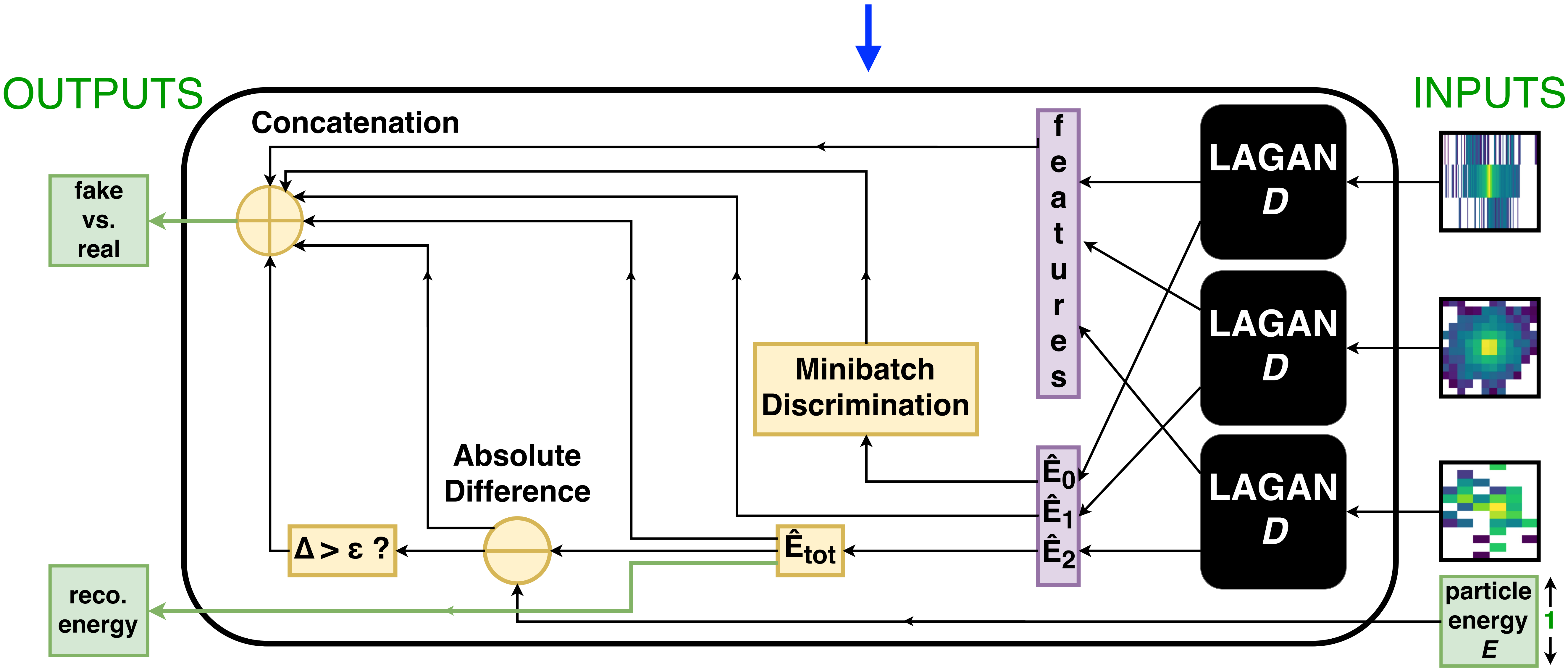


use layer i as input to layer i+1

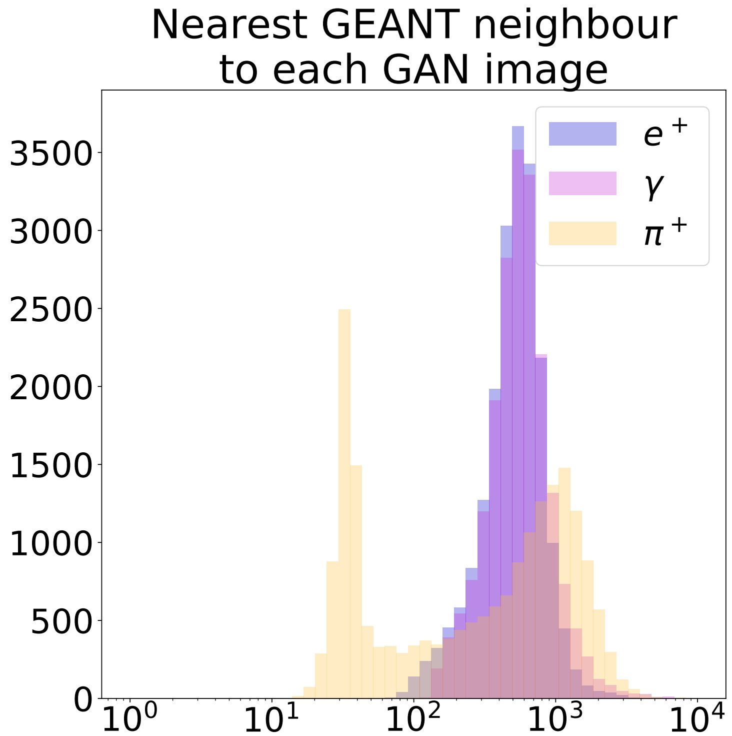
ReLU to encourage sparsity

Discriminator Network for CaloGAN

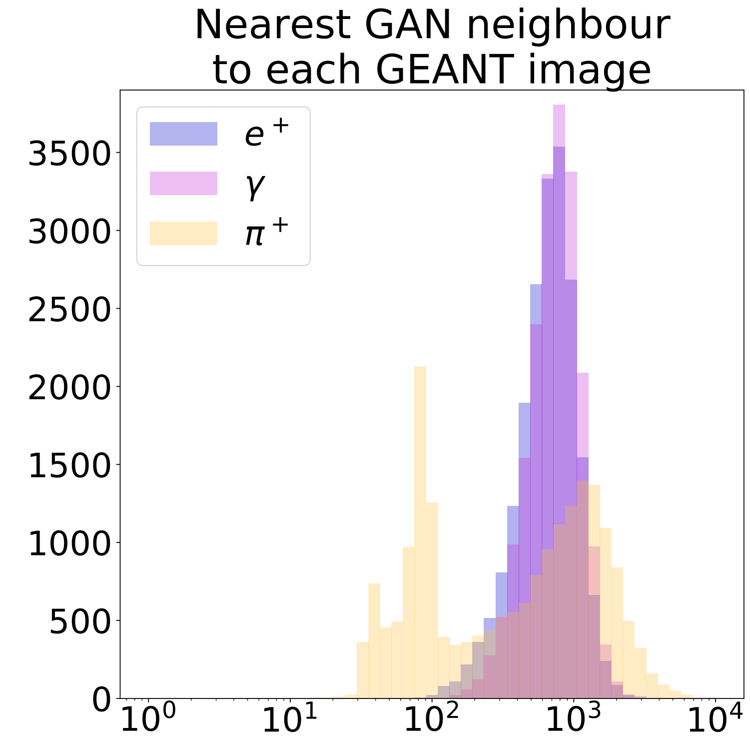
help avoid
'mode collapse'



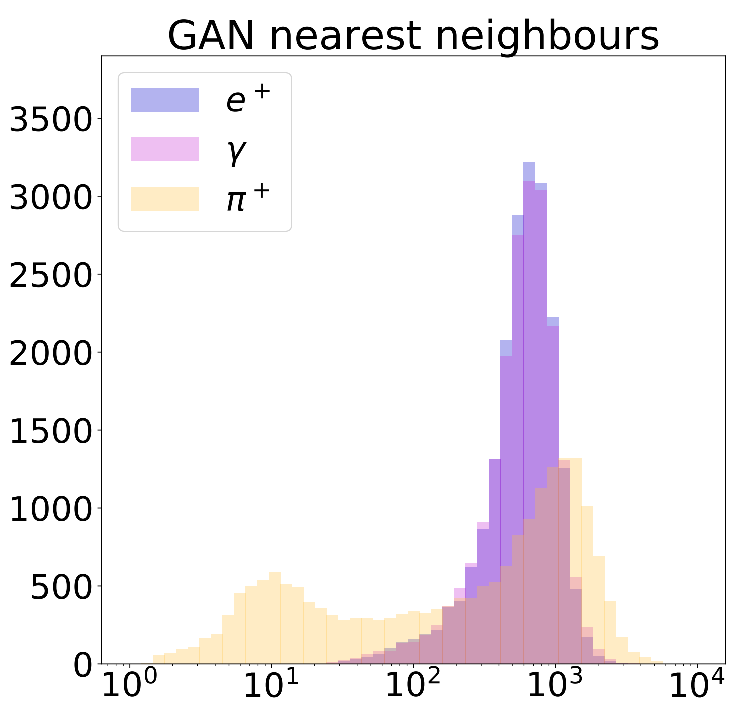
"Overtraining"



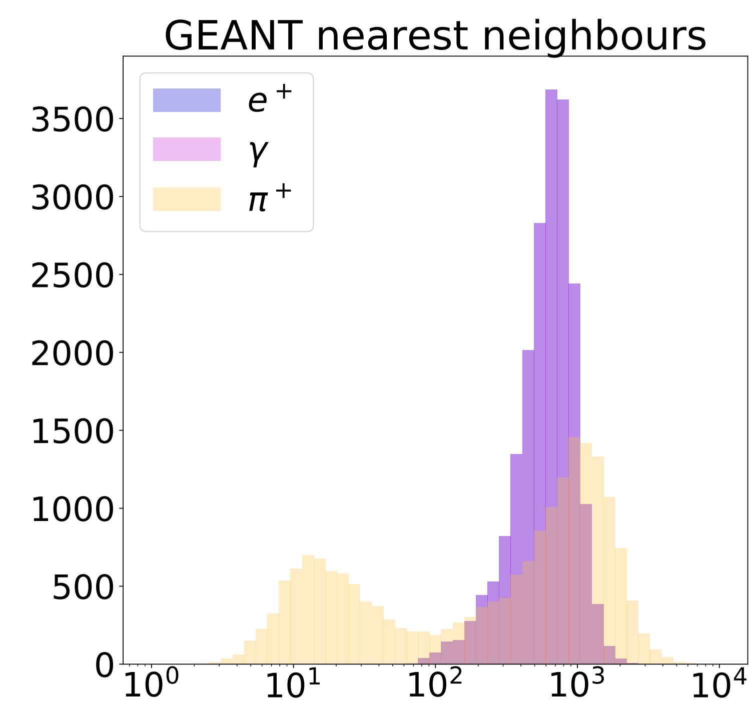
not
memorizing



A key challenge in training GANs is the diversity of generated images. This does not seem to be a problem for CaloGAN.



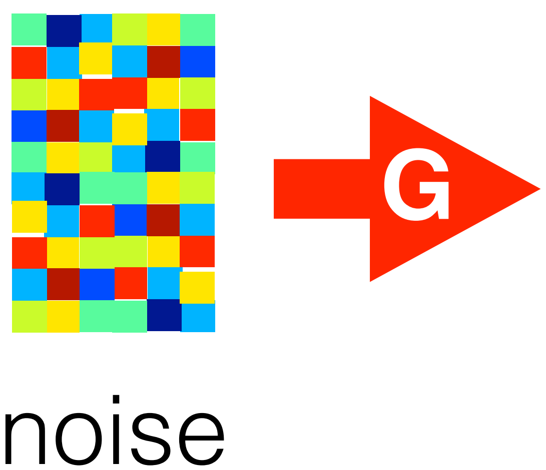
no mode
collapse

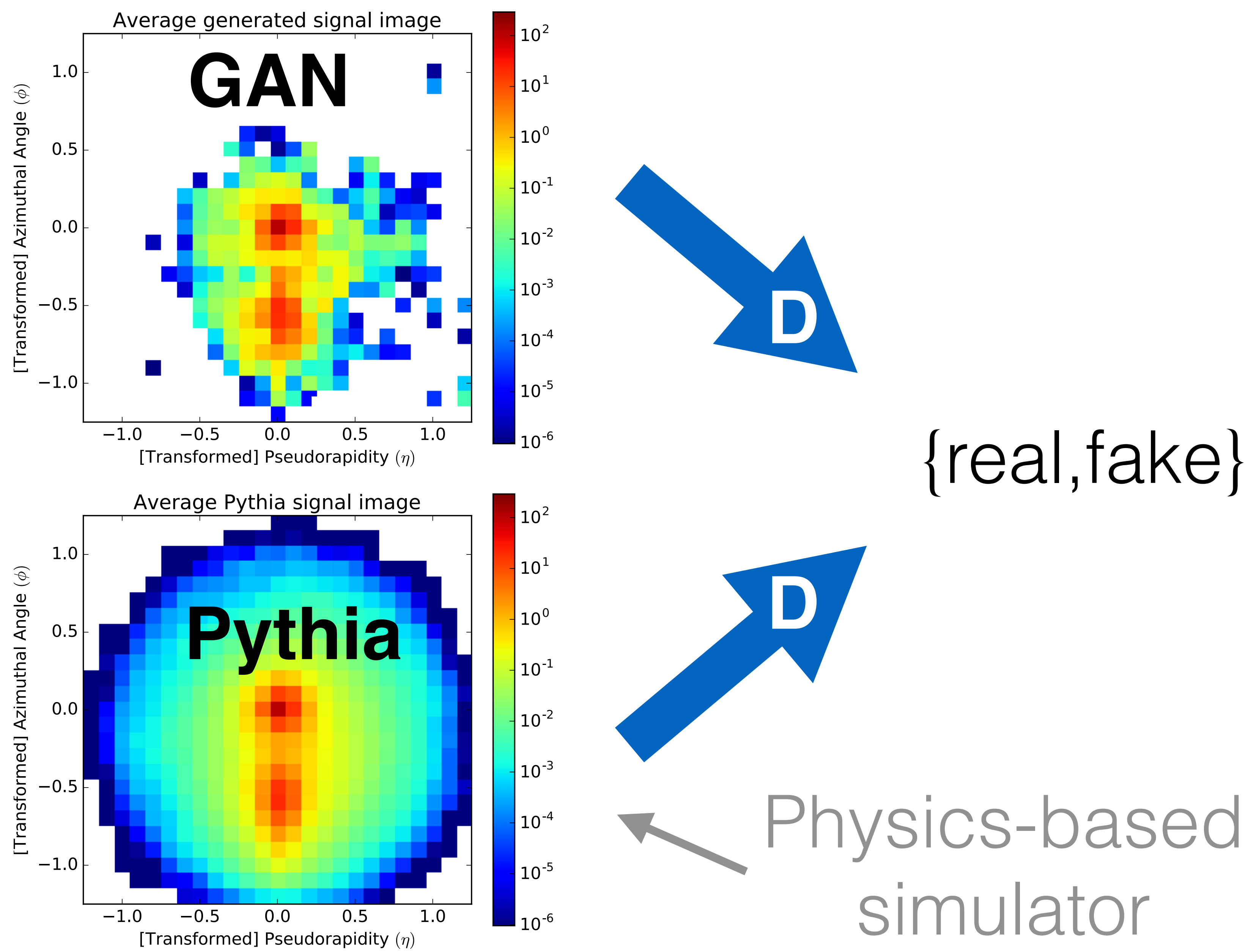


And now: Modern Deep NN's for Generation

M. Paganini, L. de Oliveira, and ***BPN*** 1705.05927, 1705.02355

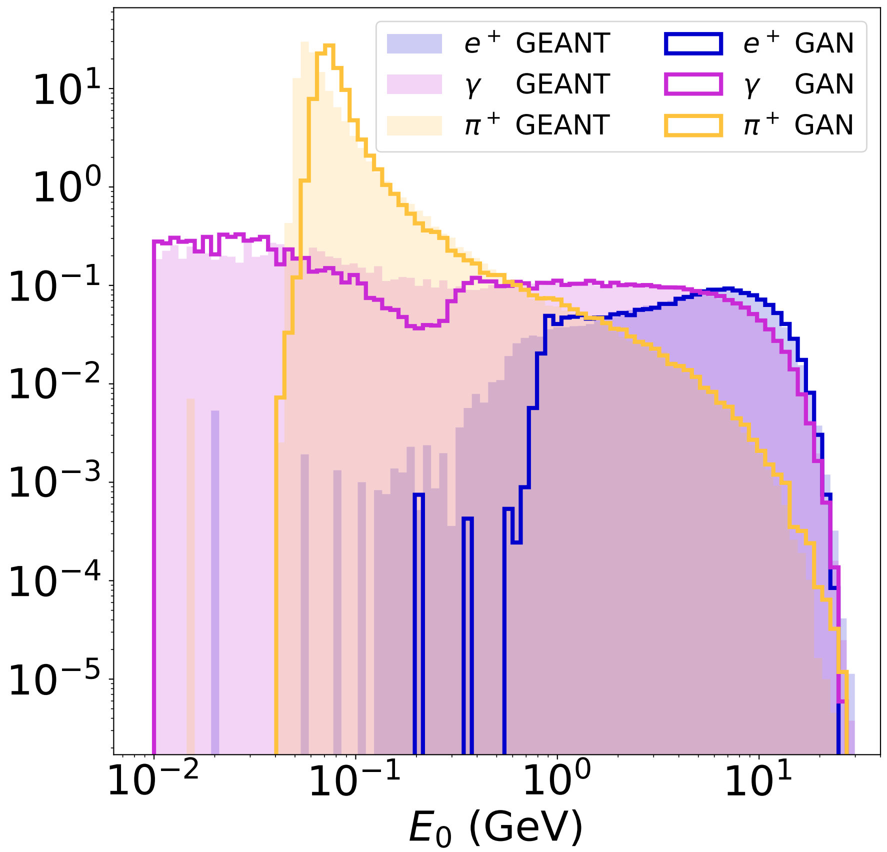
Generative Adversarial Networks (GAN):
A two-network game where one ***maps noise to images*** *and one* ***classifies images as fake or real****.*



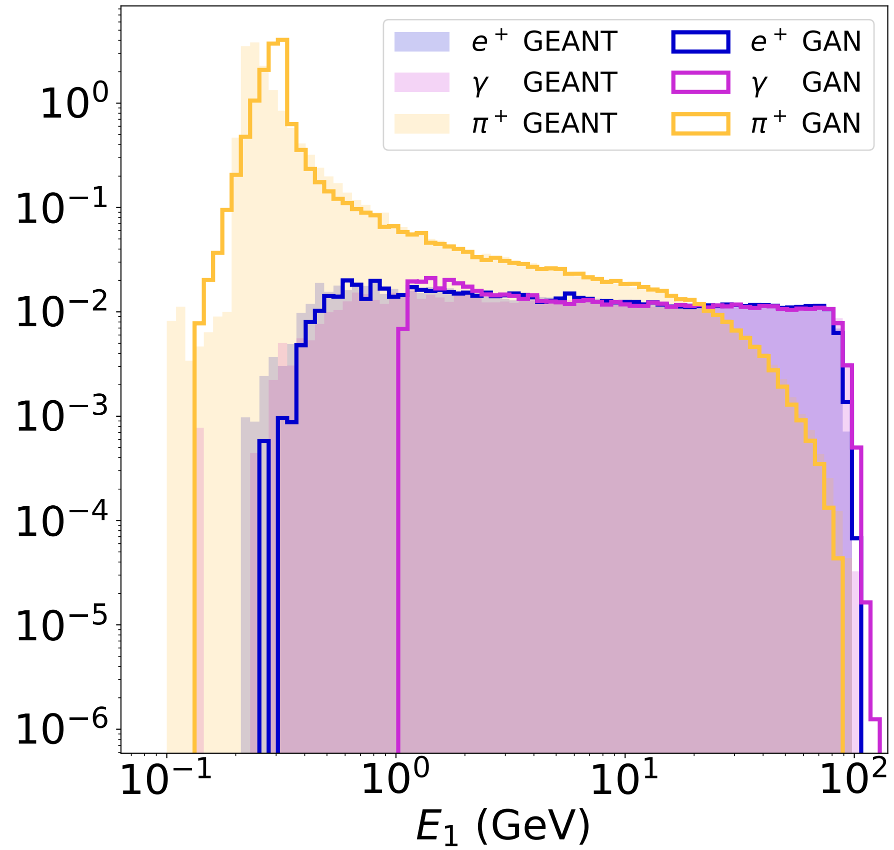


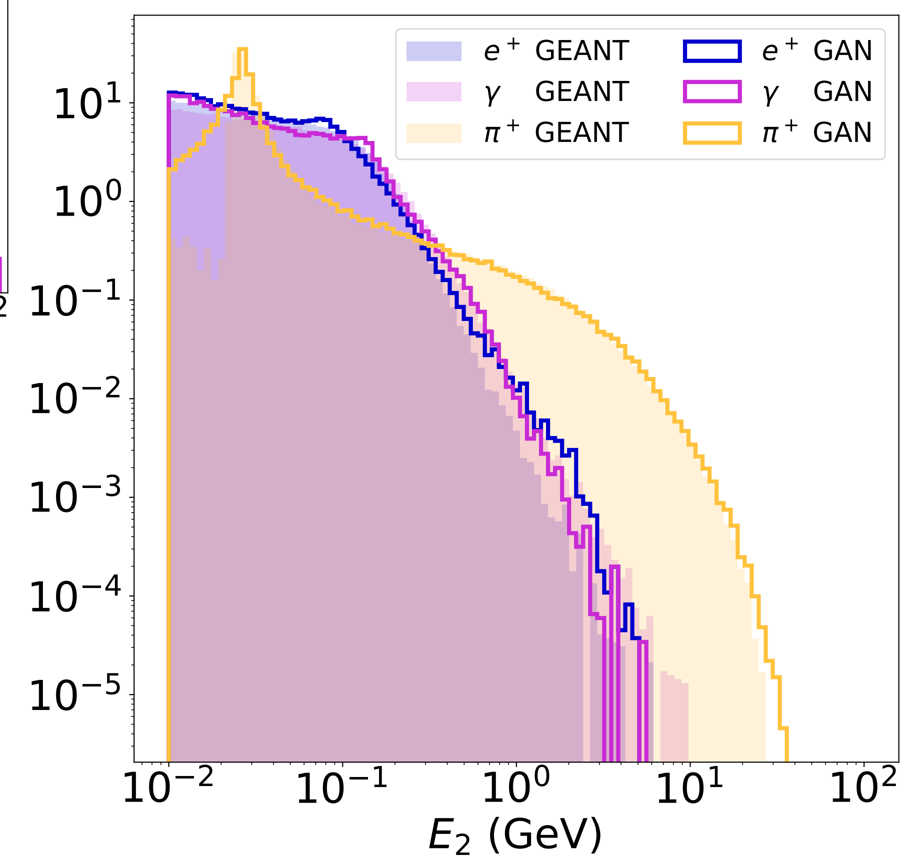
When **D** is maximally confused, **G** will be a good generator

Energy per layer



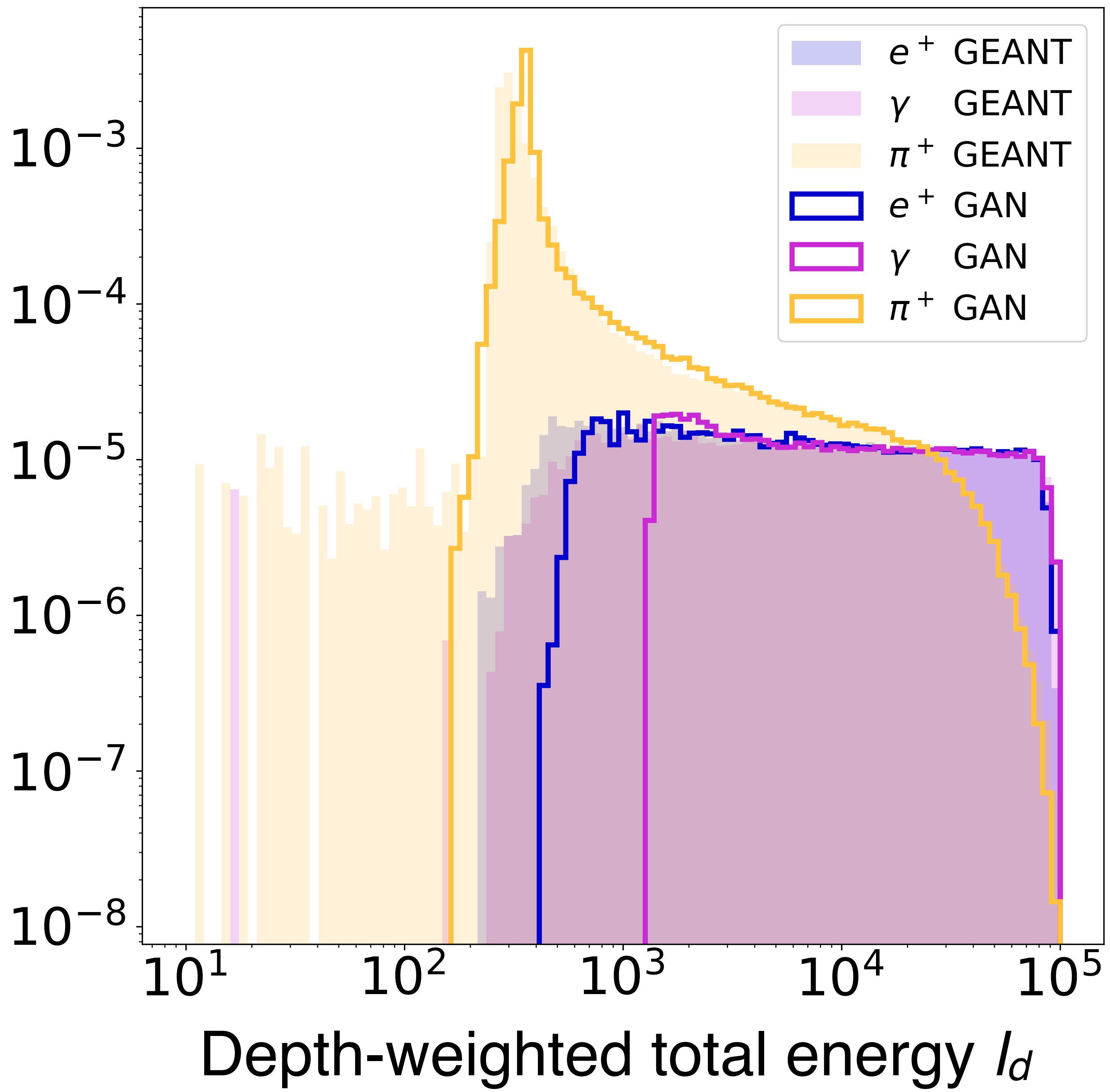
Pions deposit much less energy in the first layers; leave the calorimeter with significant energy



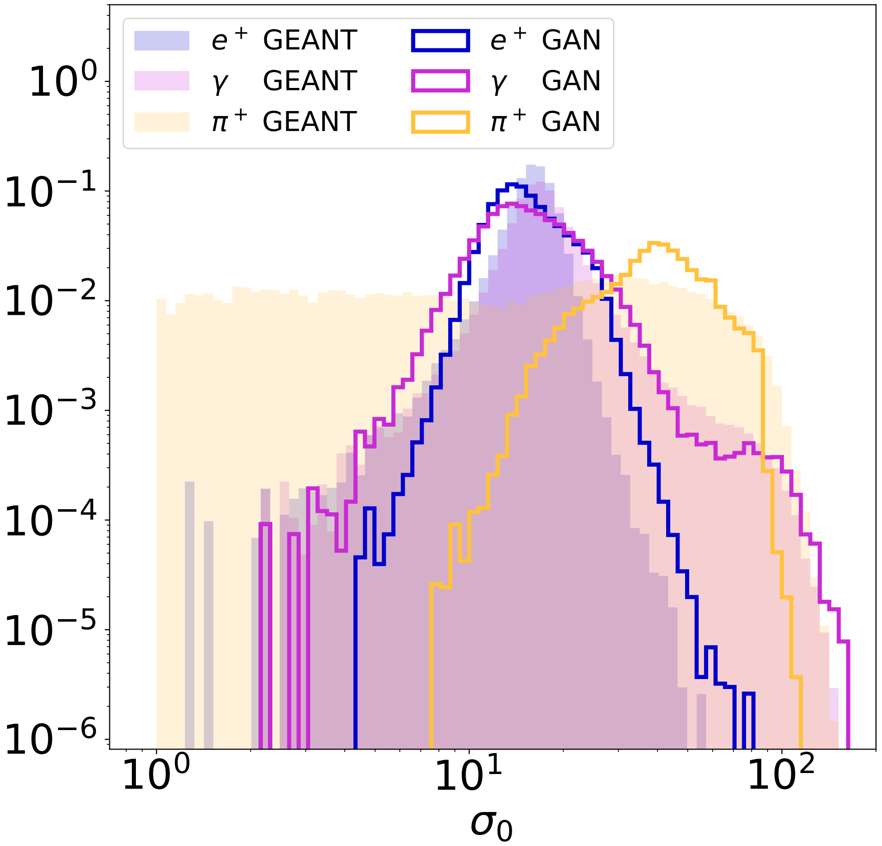


N.B. can always add these (and others) explicitly to the training

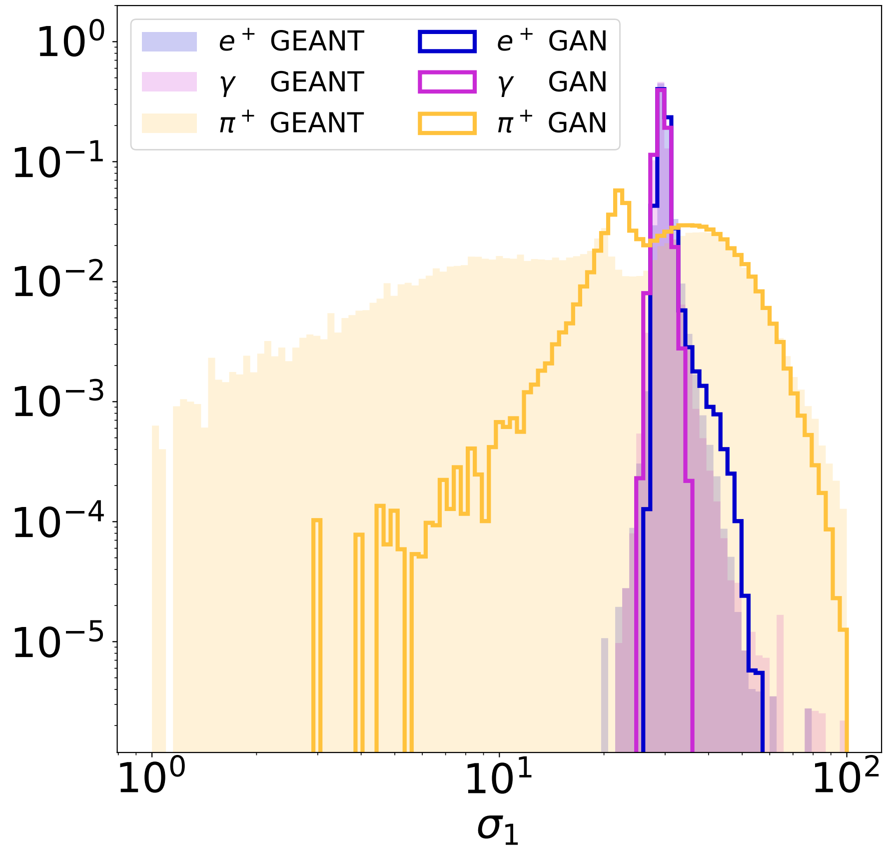
Depth of the shower

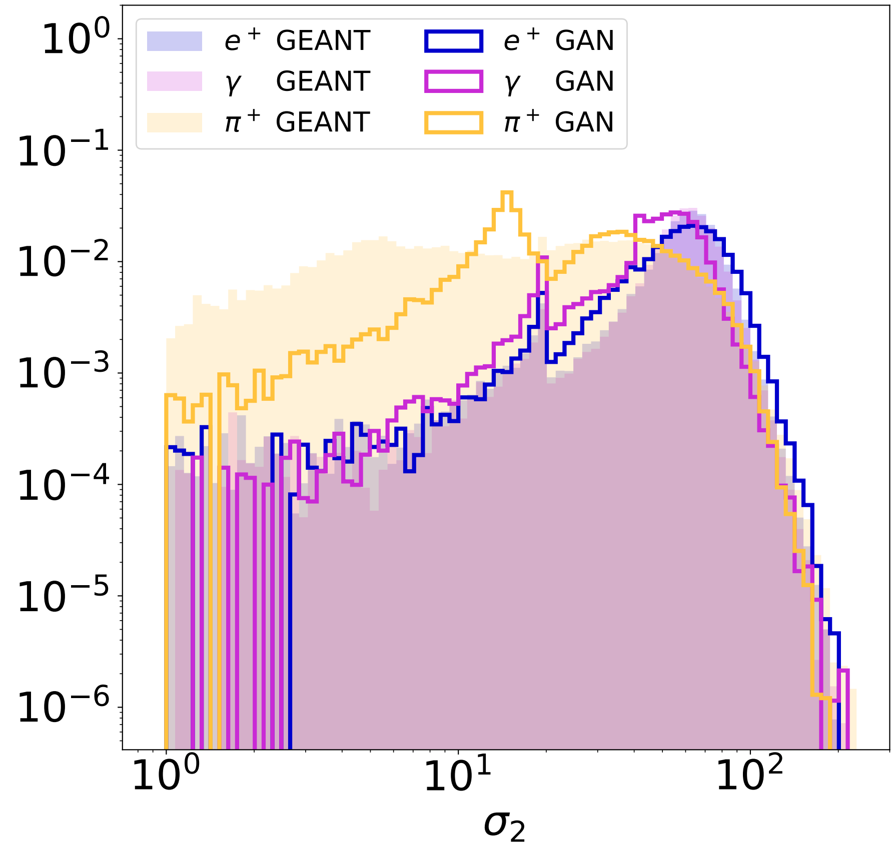


Lateral spread

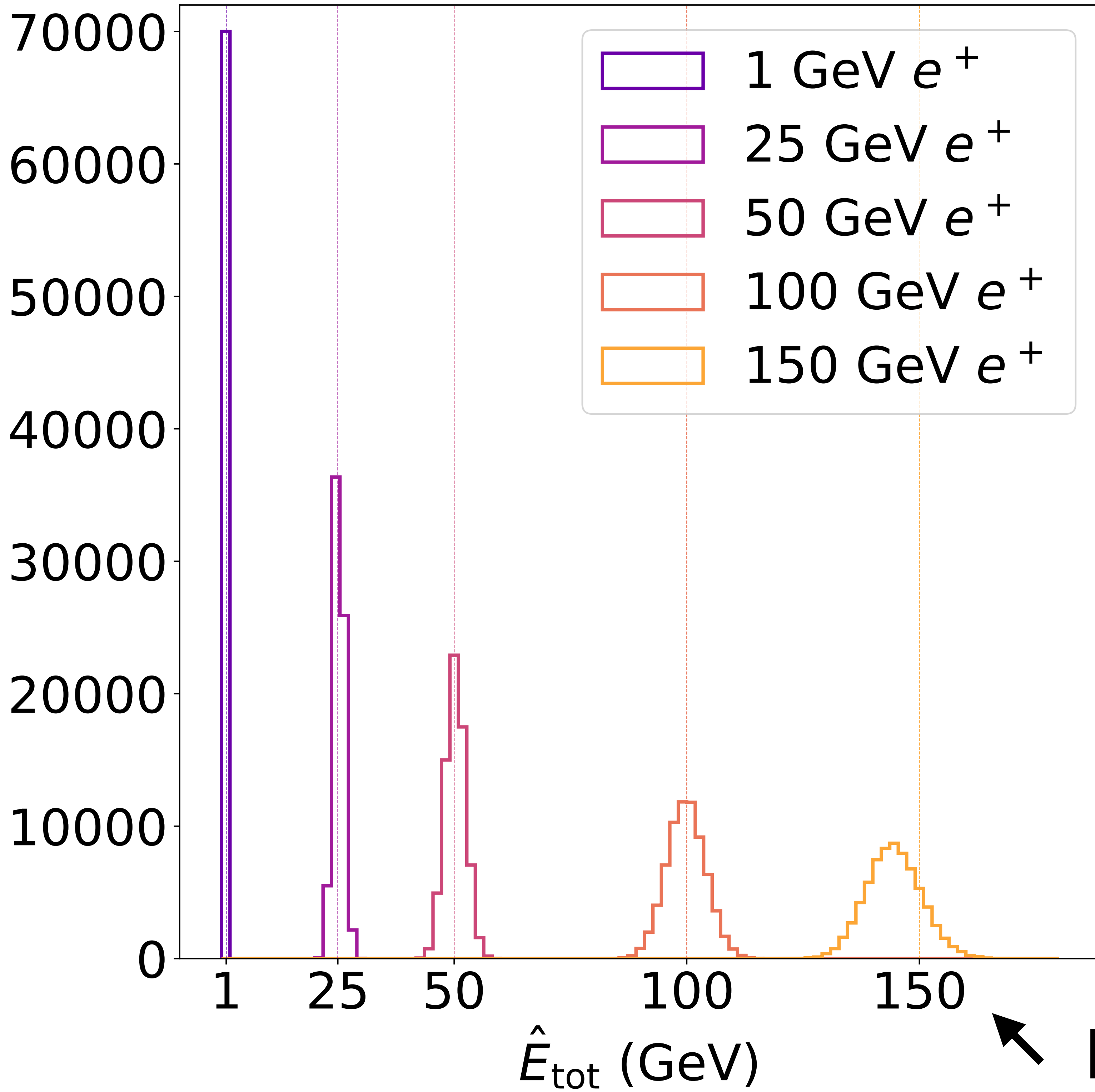


The much larger variation in the pion showers is a challenge for the network.



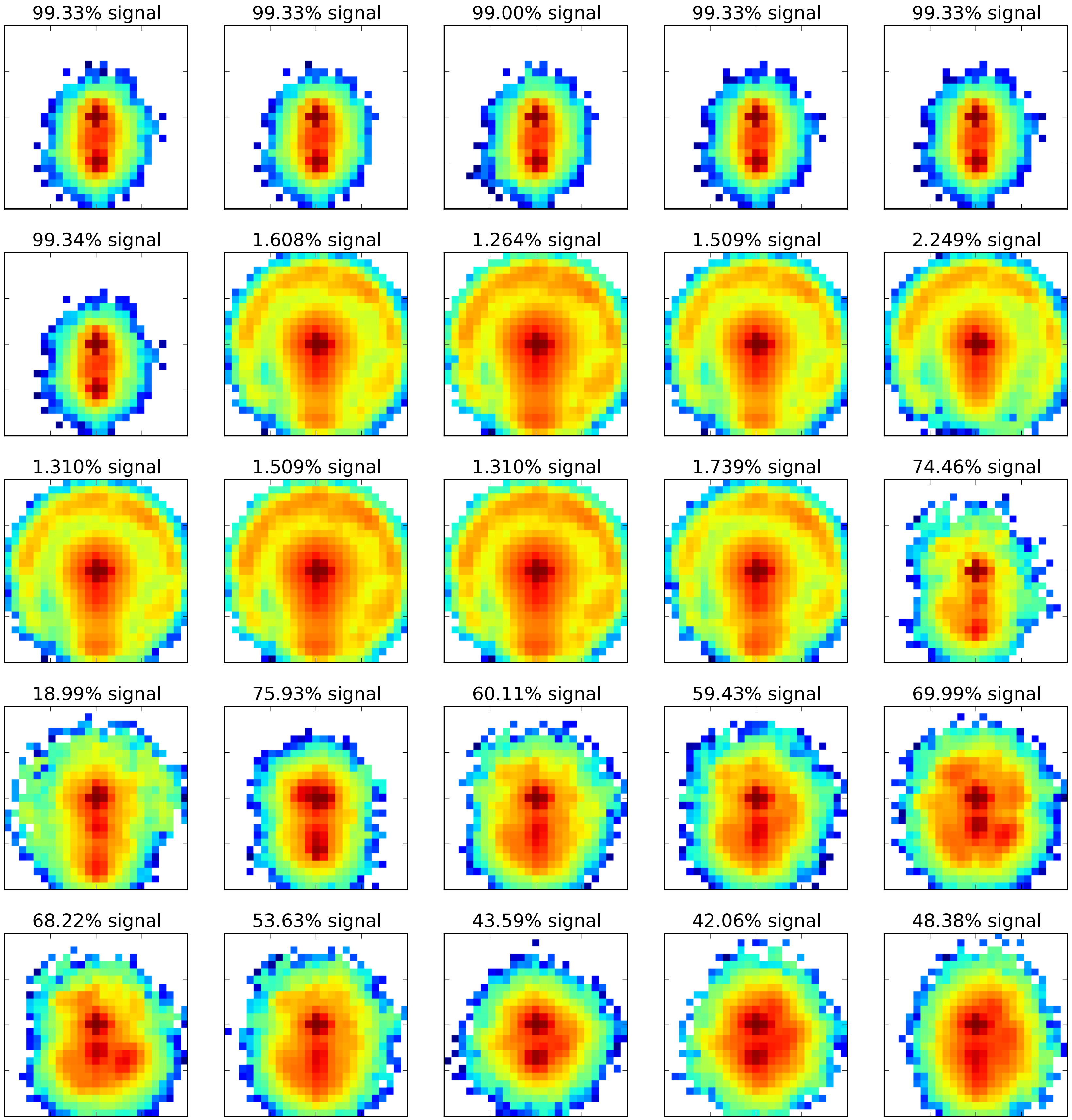


These moments and others are useful for classification; we have also tested this as a metric (NN on 3D images)



Beyond our training sample!

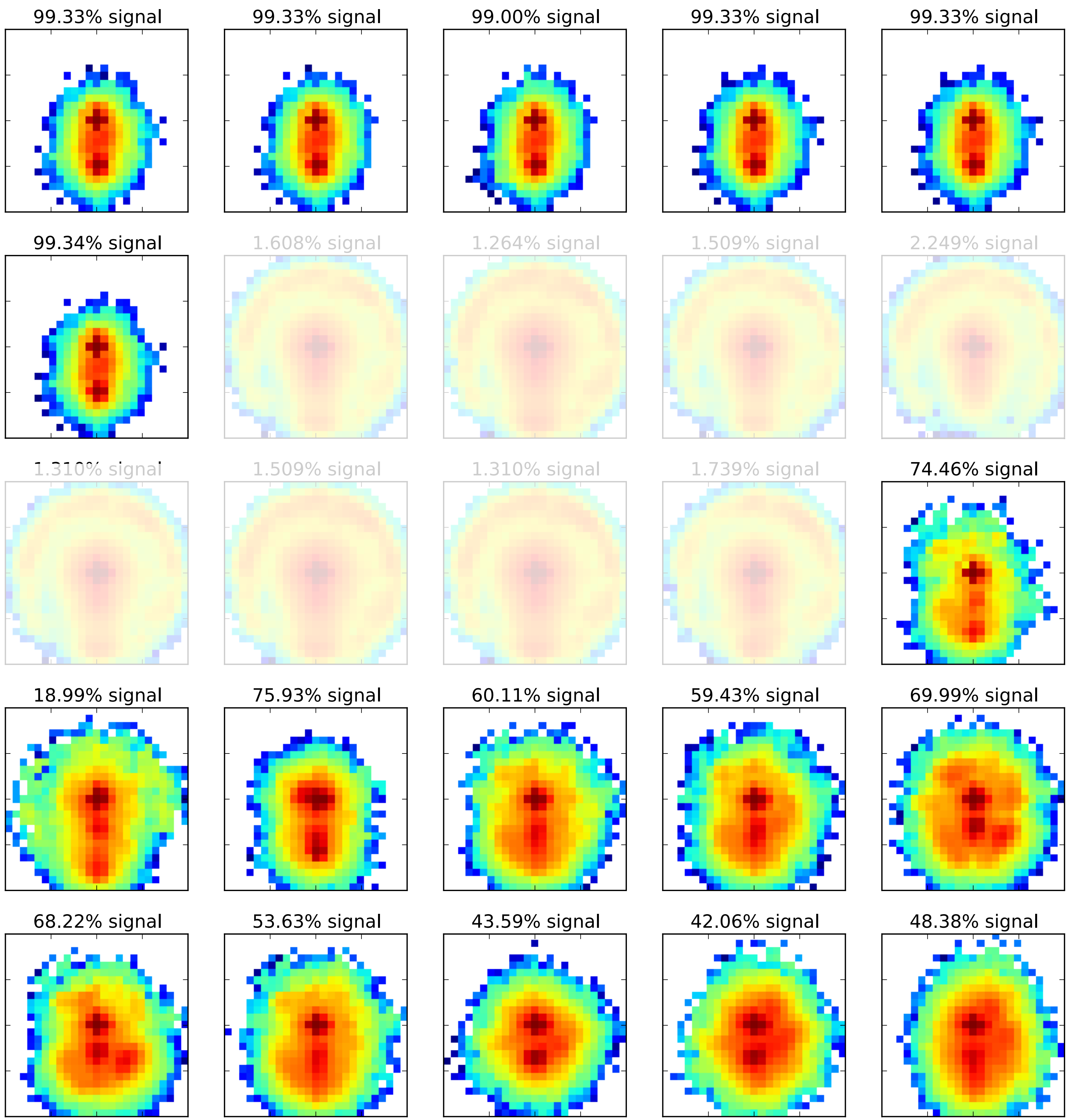
Most activating images



Take a node in the NN and ask which input images activate it the most

Some nodes learn about subjets and some learn about peripheral radiation

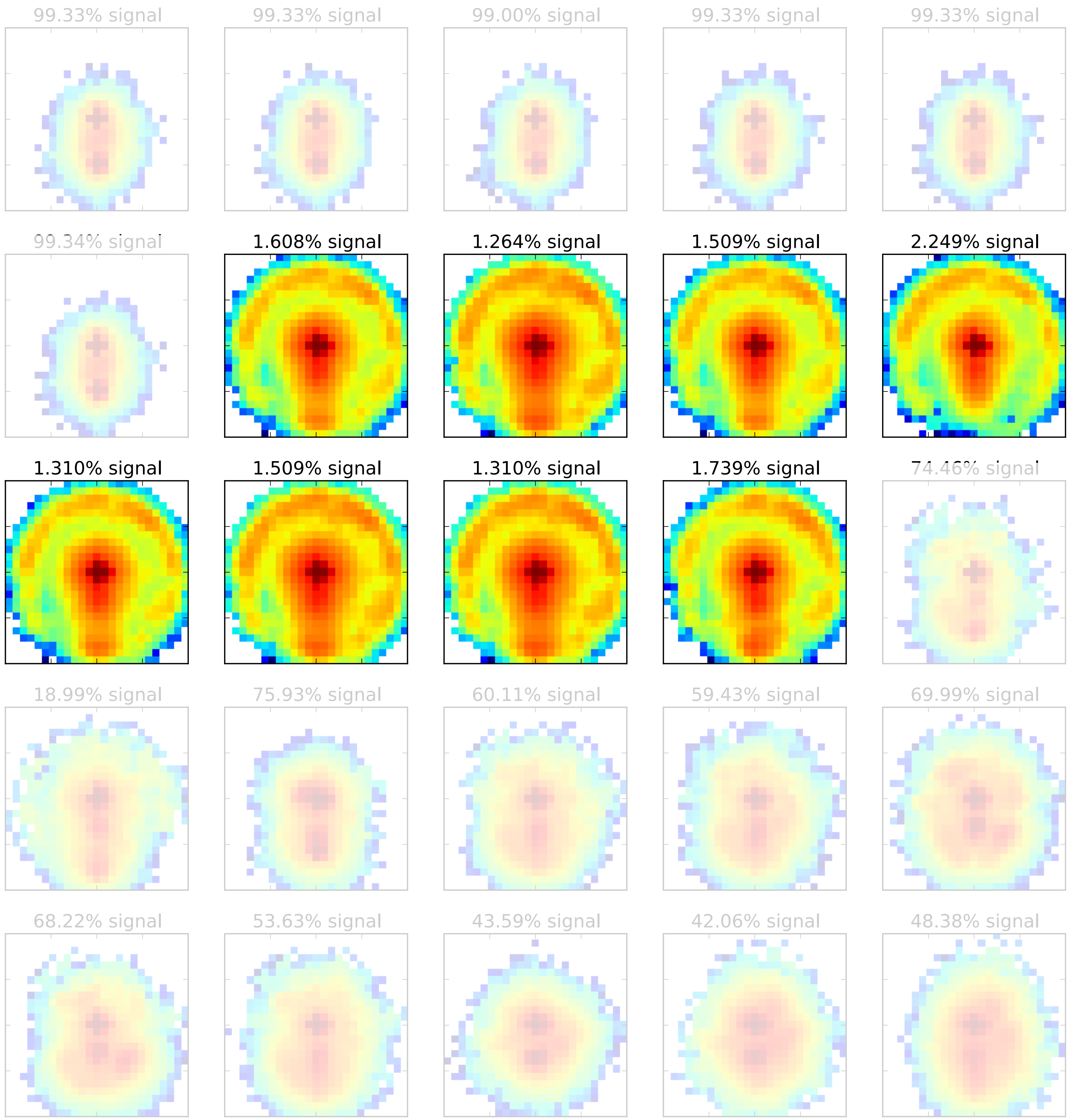
Most activating images



Take a node in the NN and ask which input images activate it the most

Some nodes learn about subjets and some learn about peripheral radiation

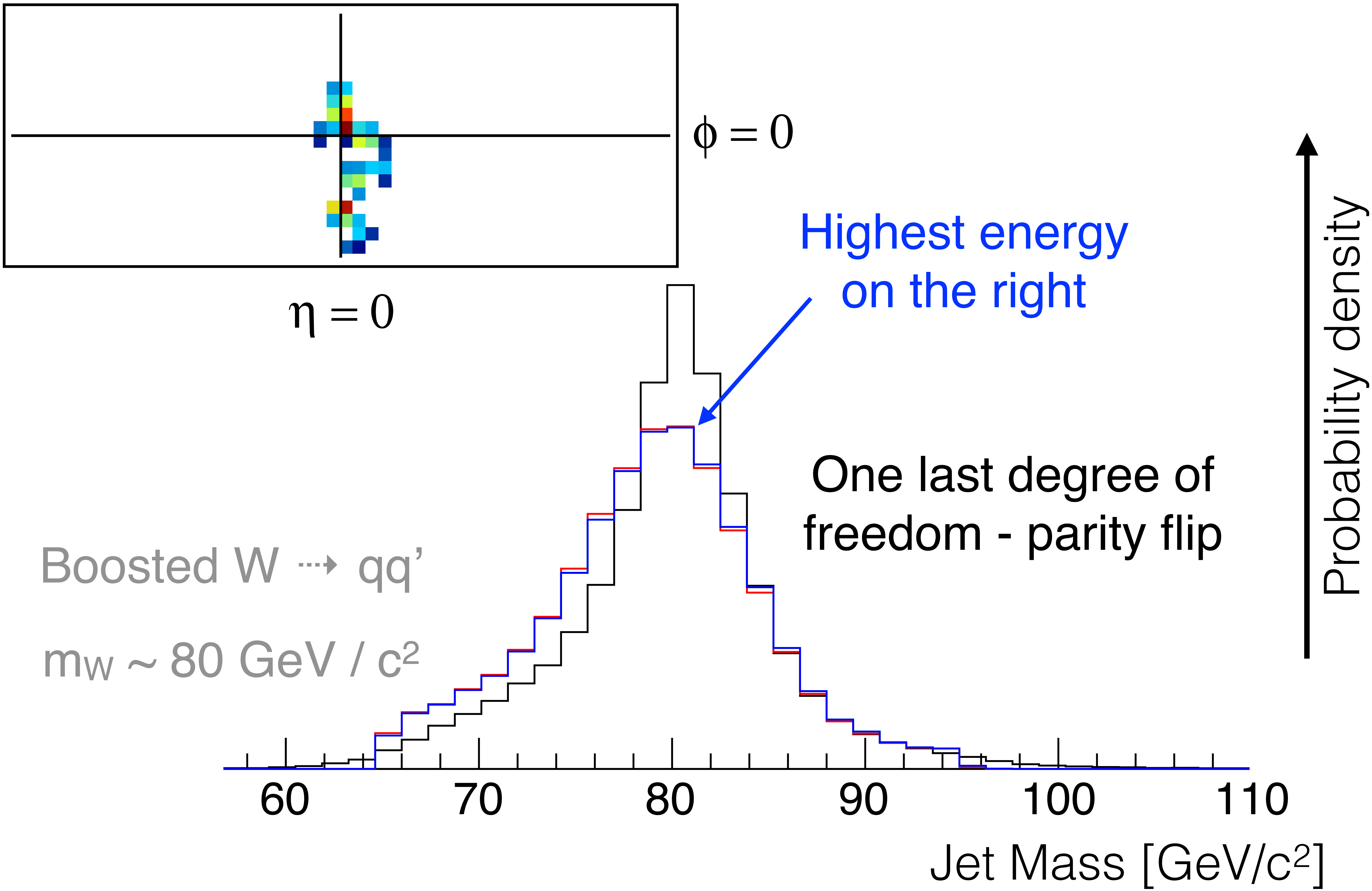
Most activating images



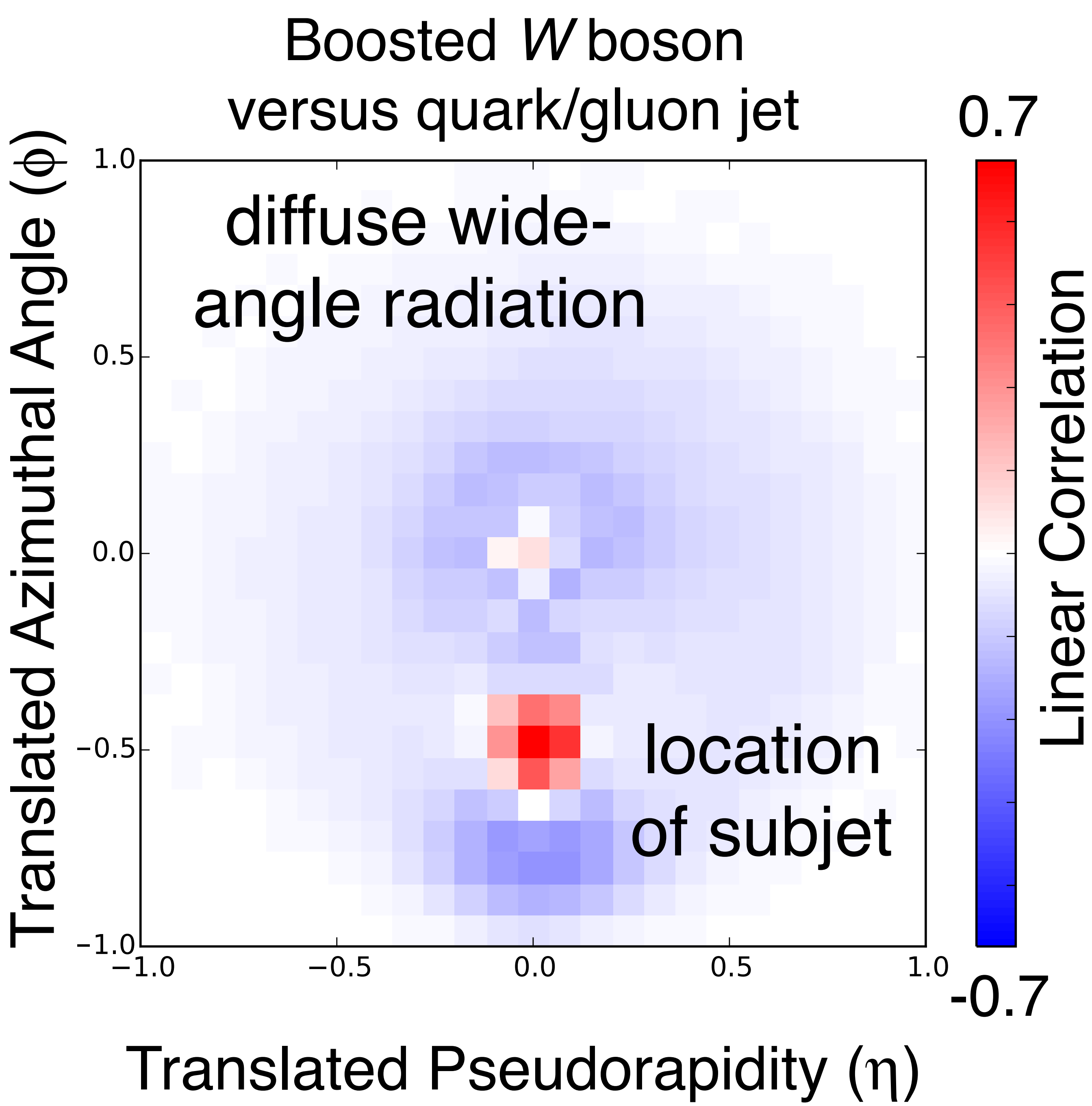
Take a node in the NN and ask which input images activate it the most

Some nodes learn about subjets and some learn about peripheral radiation

Pre-processing & spacetime symmetries



Correlation between input and output



Red = network is more activated (more signal-like)