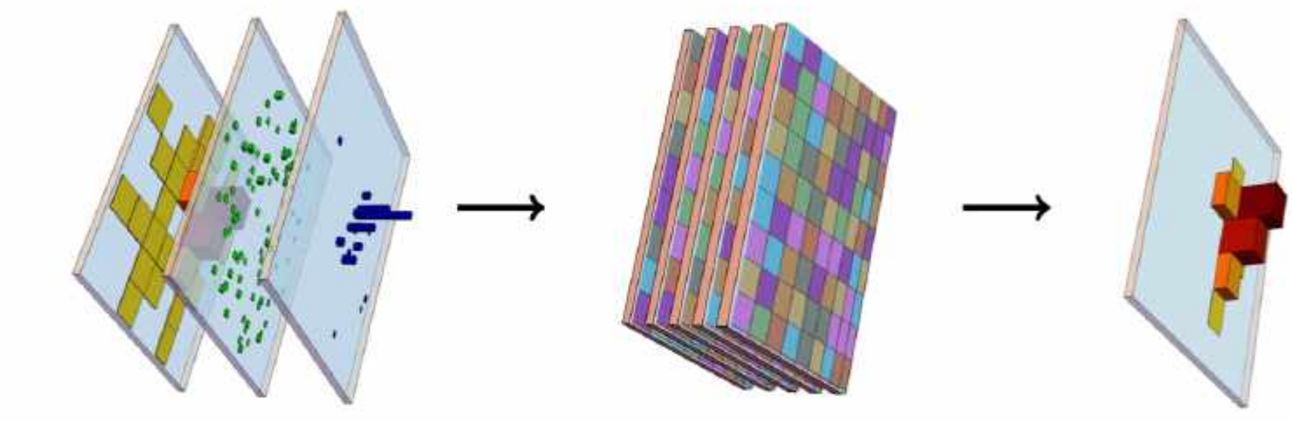
Modern Machine Learning and the Large Hadron Collider

KIAS QQUC School on "AI in High Energy Physics"

July 24, 2020

Matthew Schwartz

Harvard University



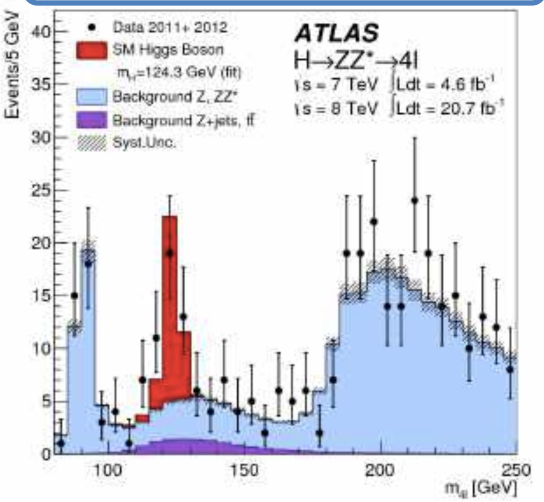
Outline

- Introduction to jet physics
- Traditional approach vs ML approach
- ML applications in collider physics
  - Quarks vs. Gluons
  - Top tagging
  - Pileup removal
  - Unsupervised learning
  - Unfolding
  - Weak supervision
  - Decorrelation

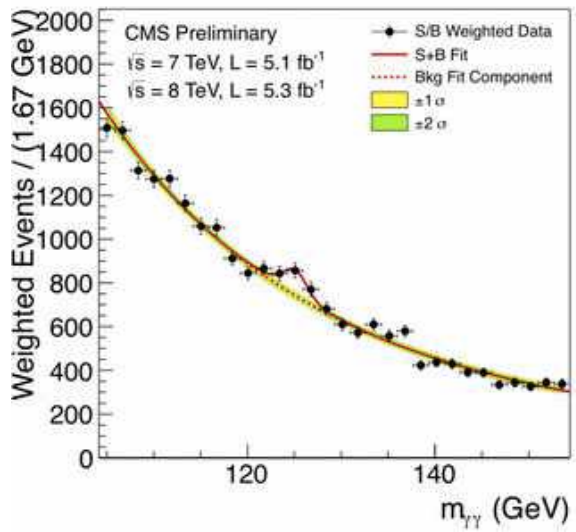
INTRODUCTION TO JETS

Higgs discovered 2012

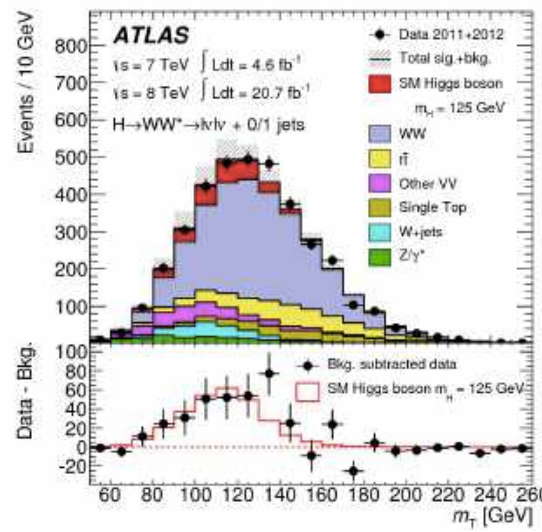
$h \to e^- e^+ \mu^- \mu^+$, 0.01%



$h \to \gamma\gamma$, 0.1%



$h \to e^- \mu^+ \bar{\nu} \nu$, 1%



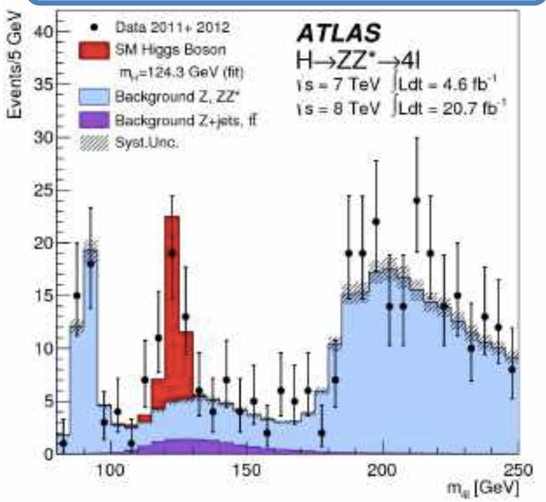
4 leptons = Golden Channel
Clear bump
Large signal/background
0.01% of Higgs decays

Photons
Not as clear bump
Small signal/background
0.2% of Higgs decays

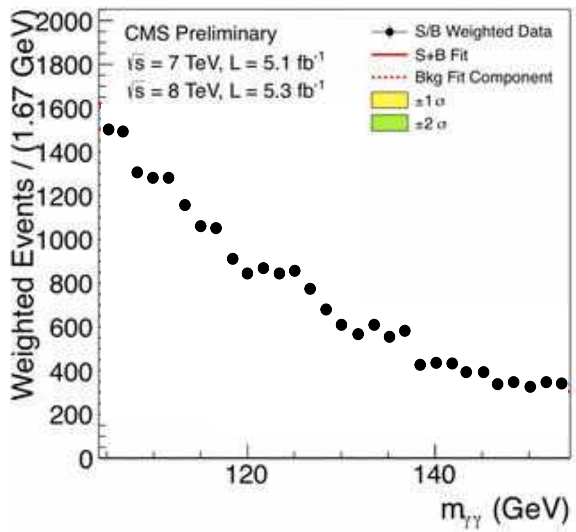
WW
No bump
Need backgrounds
0.8% of Higgs decays

Higgs discovered 2012

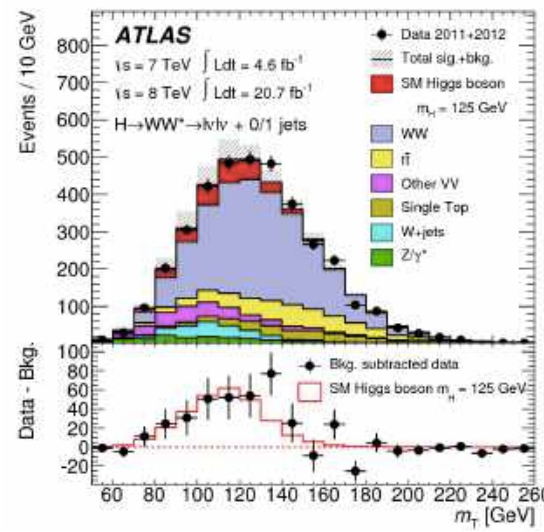
$h \to e^- e^+ \mu^- \mu^+$, 0.01%



$h \to \gamma\gamma$, 0.1%



$h \to e^- \mu^+ \bar{\nu} \nu$, 1%



4 leptons = Golden Channel
Clear bump
Large signal/background
0.01% of Higgs decays

Photons
Not as clear bump
Small signal/background
0.2% of Higgs decays

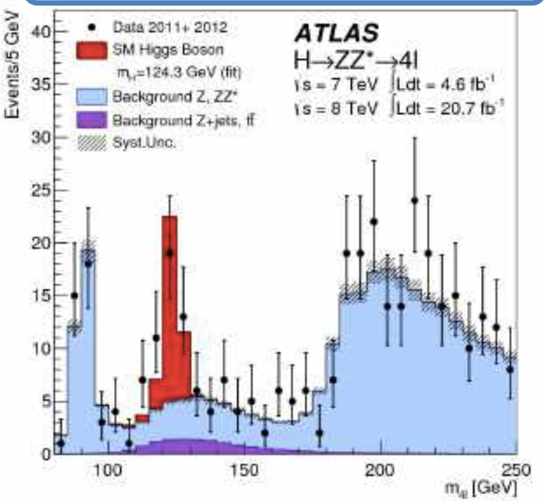
WW
No bump
Need backgrounds
0.8% of Higgs decays

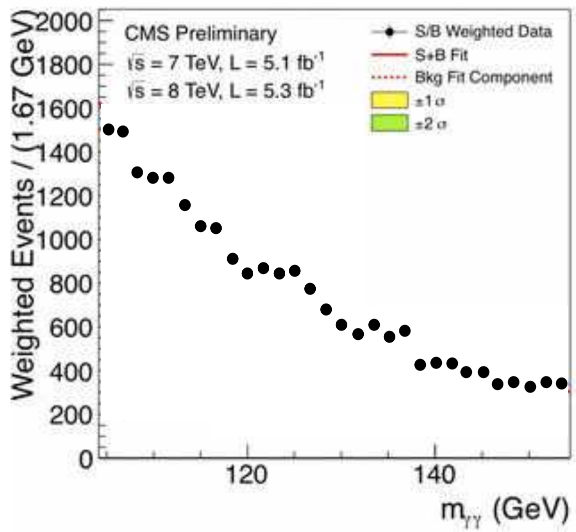
Higgs discovered 2012

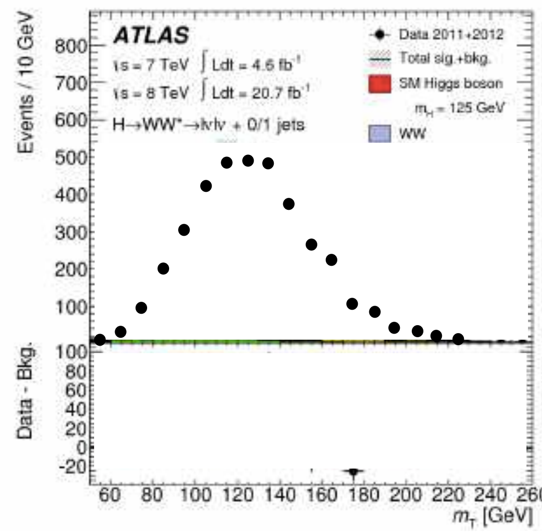
$h \to e^- e^+ \mu^- \mu^+$, 0.01%

$h \to \gamma\gamma$, 0.1%

$h \to e^- \mu^+ \bar{\nu}\nu$, 1%





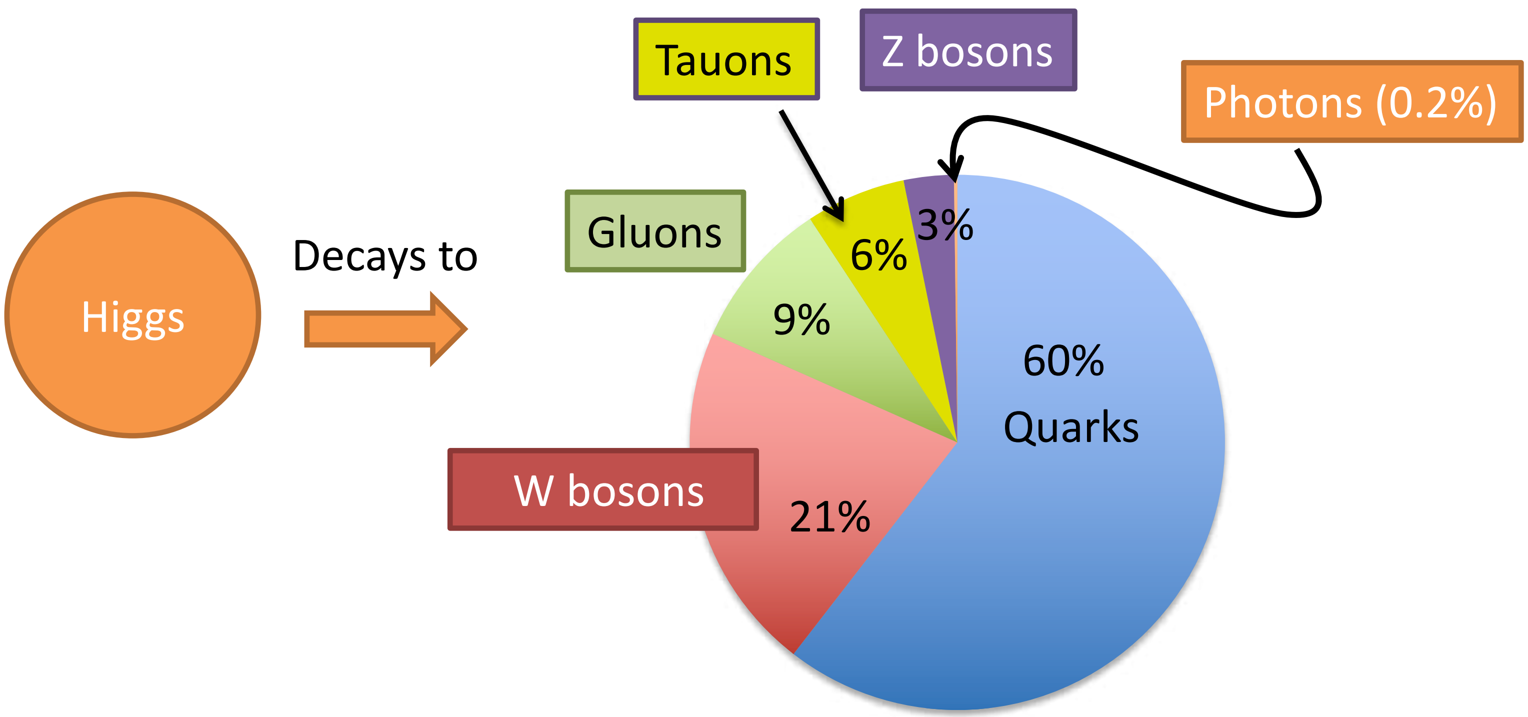


4 leptons = Golden Channel
Clear bump
Large signal/background
0.01% of Higgs decays

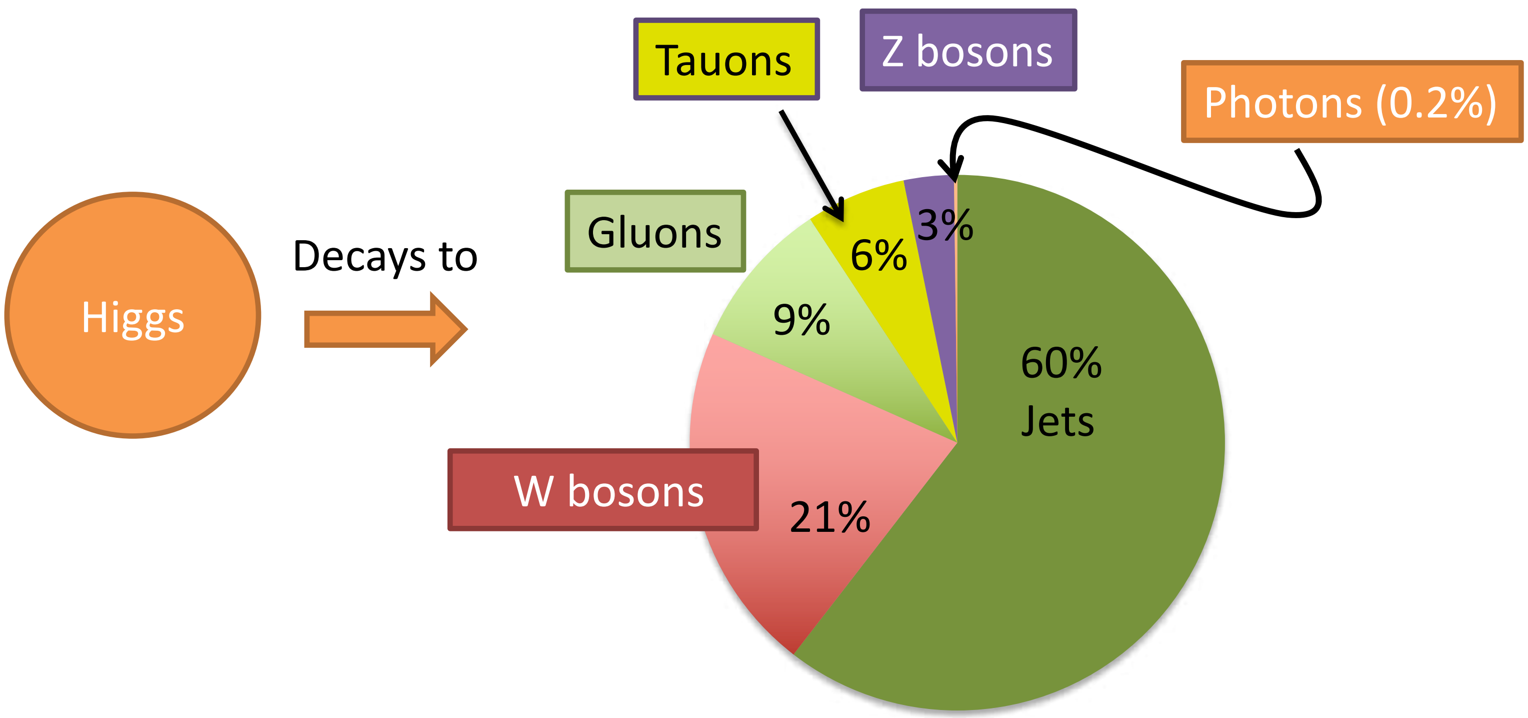
Photons
Not as clear bump
Small signal/background
0.2% of Higgs decays

WW
No bump
Need backgrounds
0.8% of Higgs decays

Higgs boson decay modes

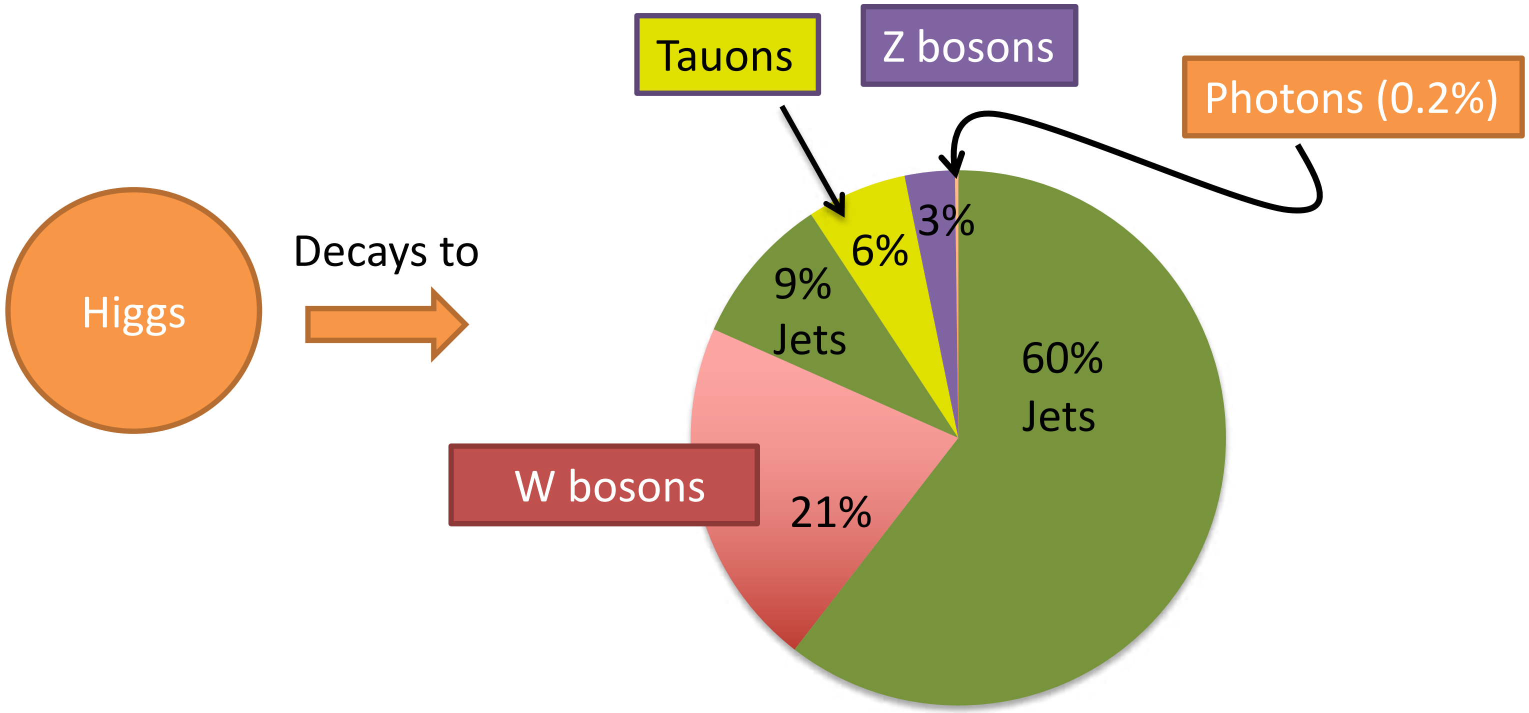


Higgs boson decay modes

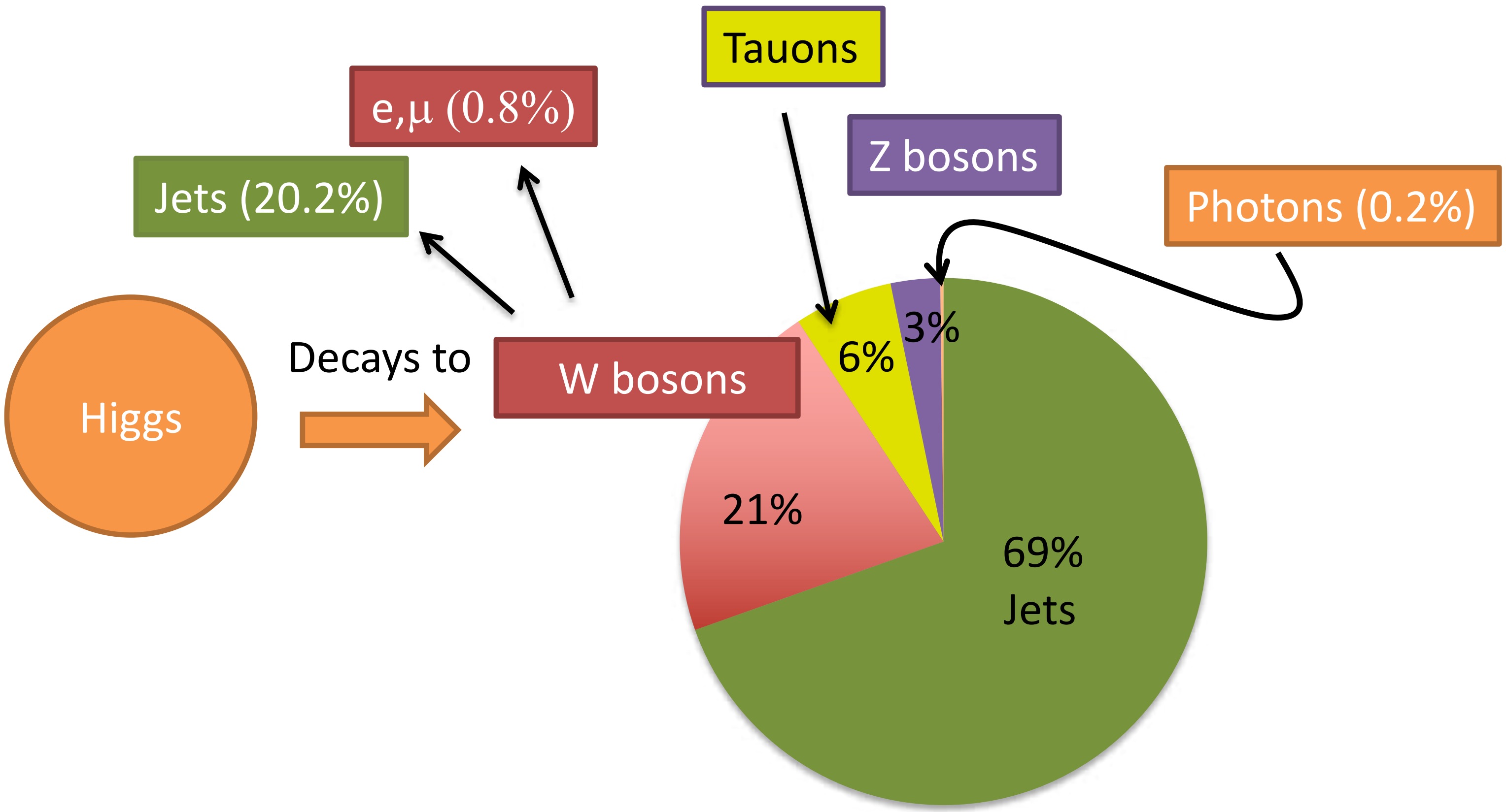


Matthew Schwartz

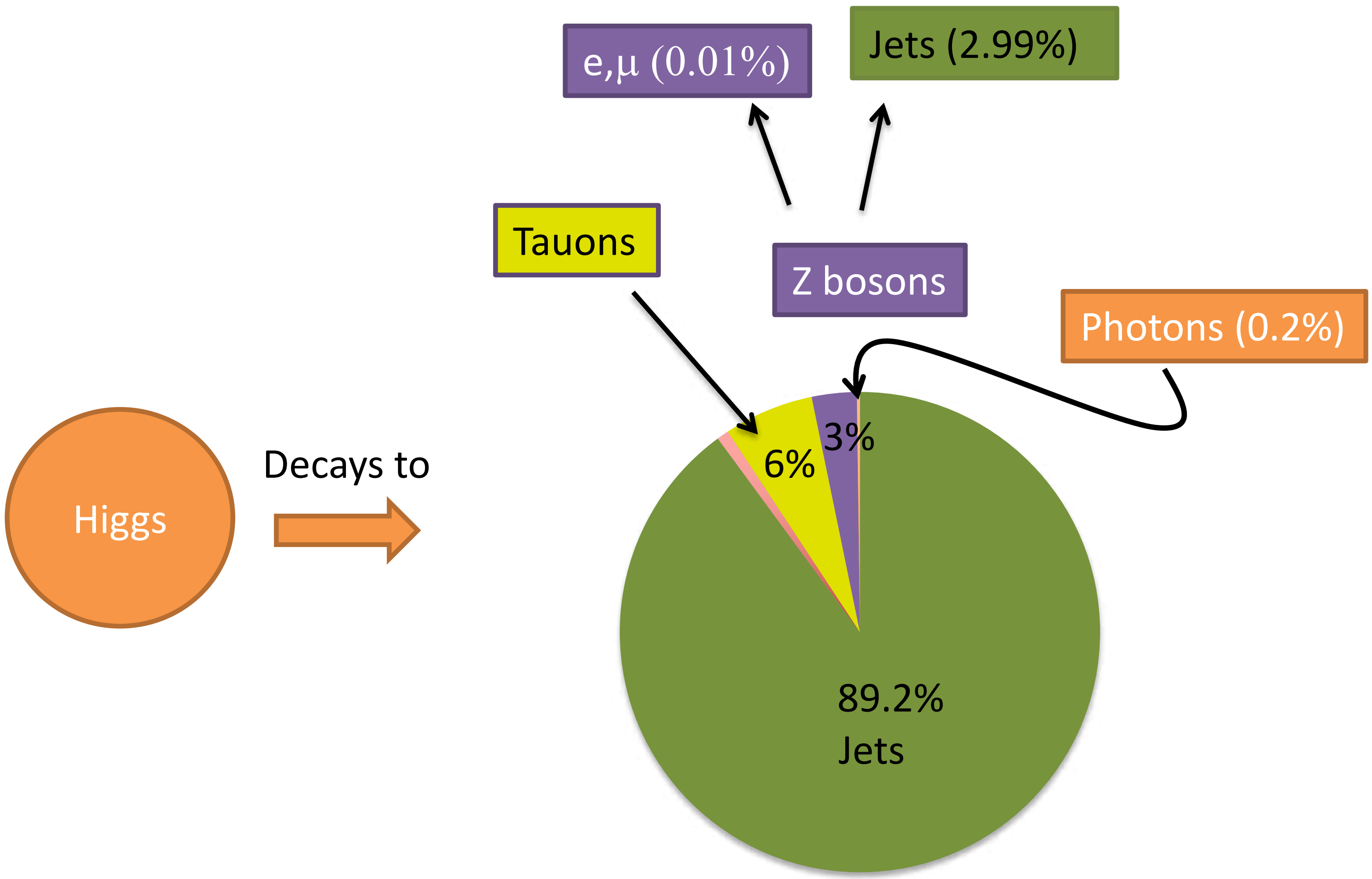
Higgs boson decay modes



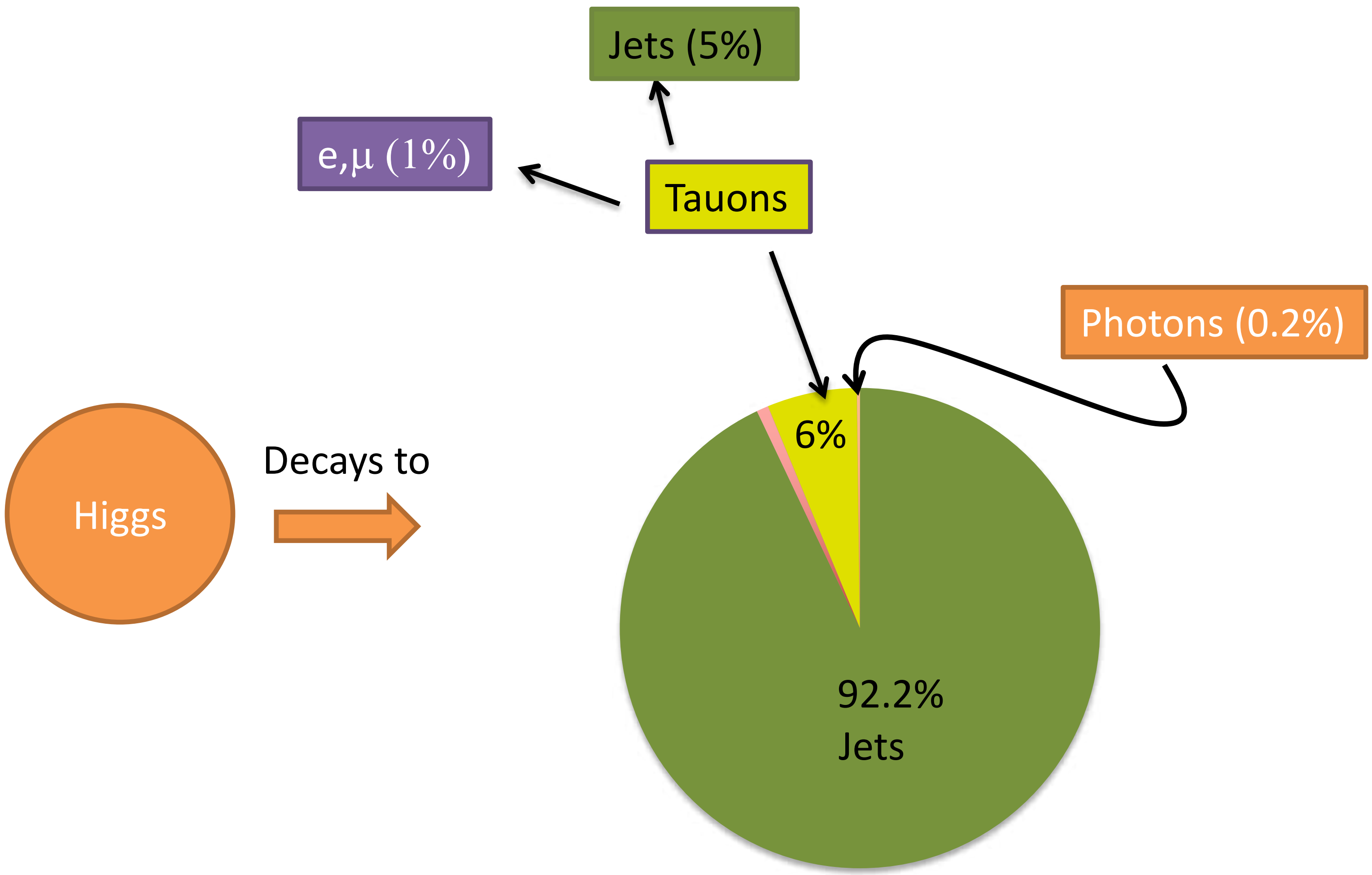
Higgs boson decay modes



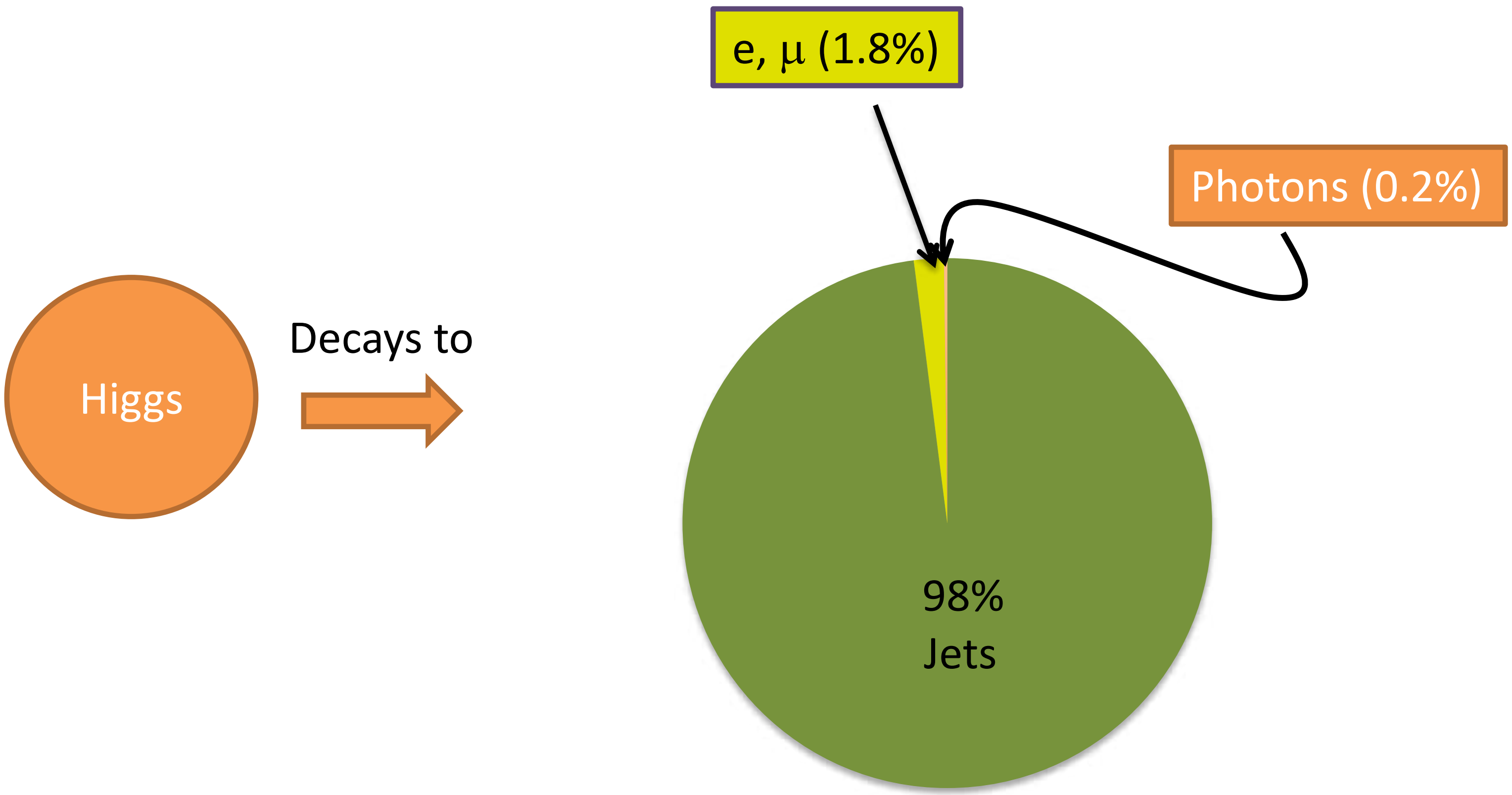
Higgs boson decay modes



Higgs boson decay modes

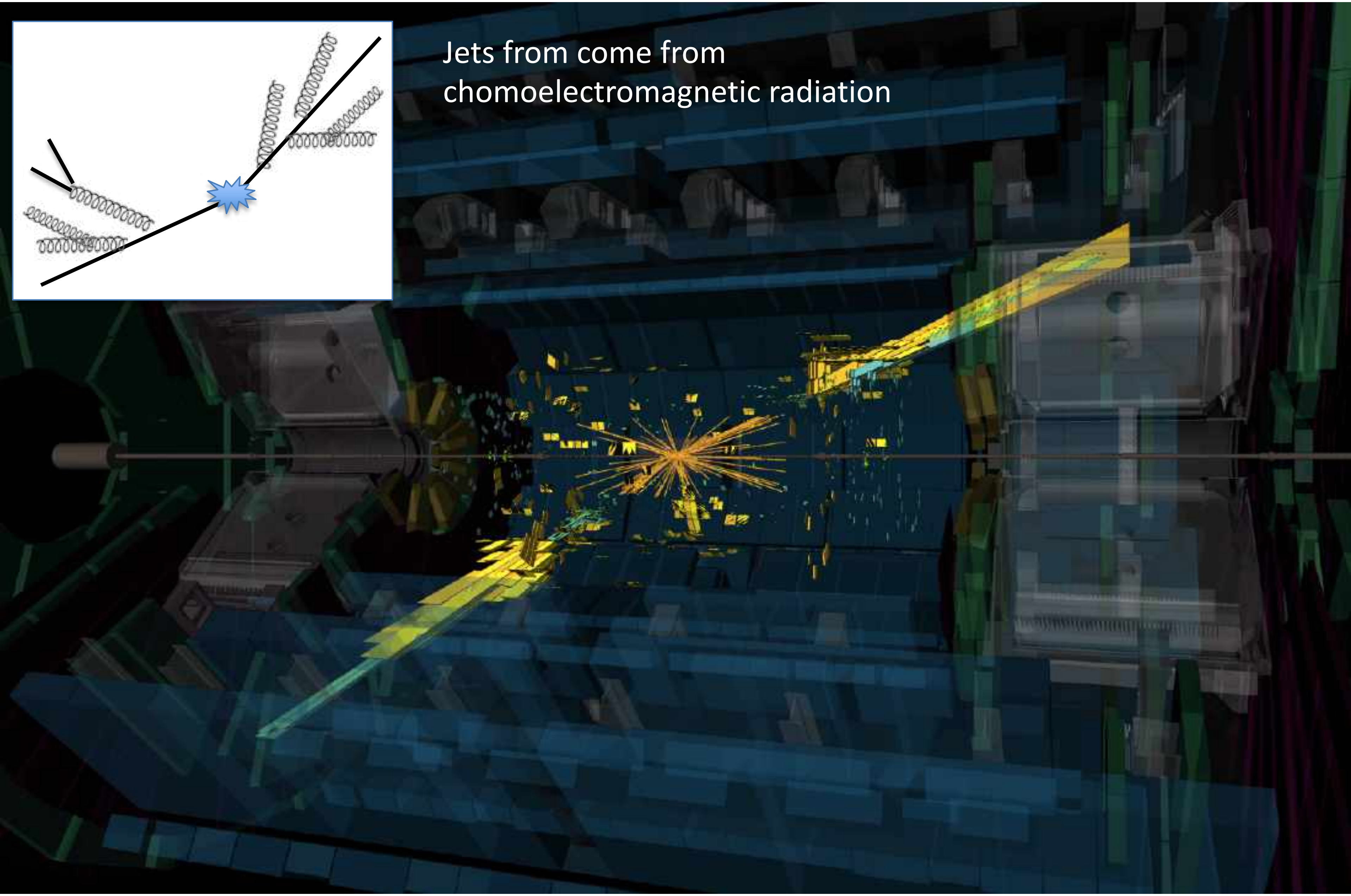


Higgs boson decay modes

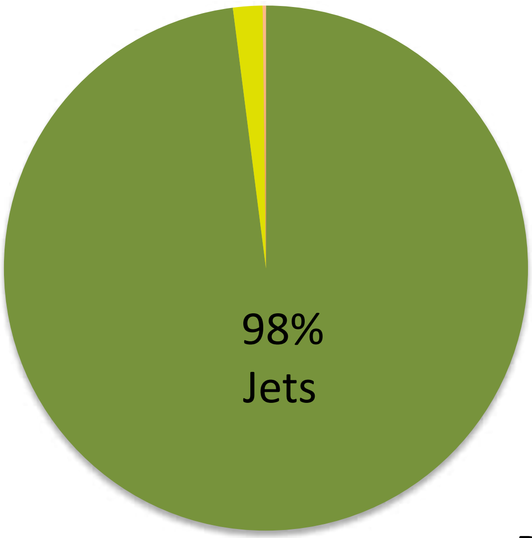


- Higgs discovery involved just these special decays

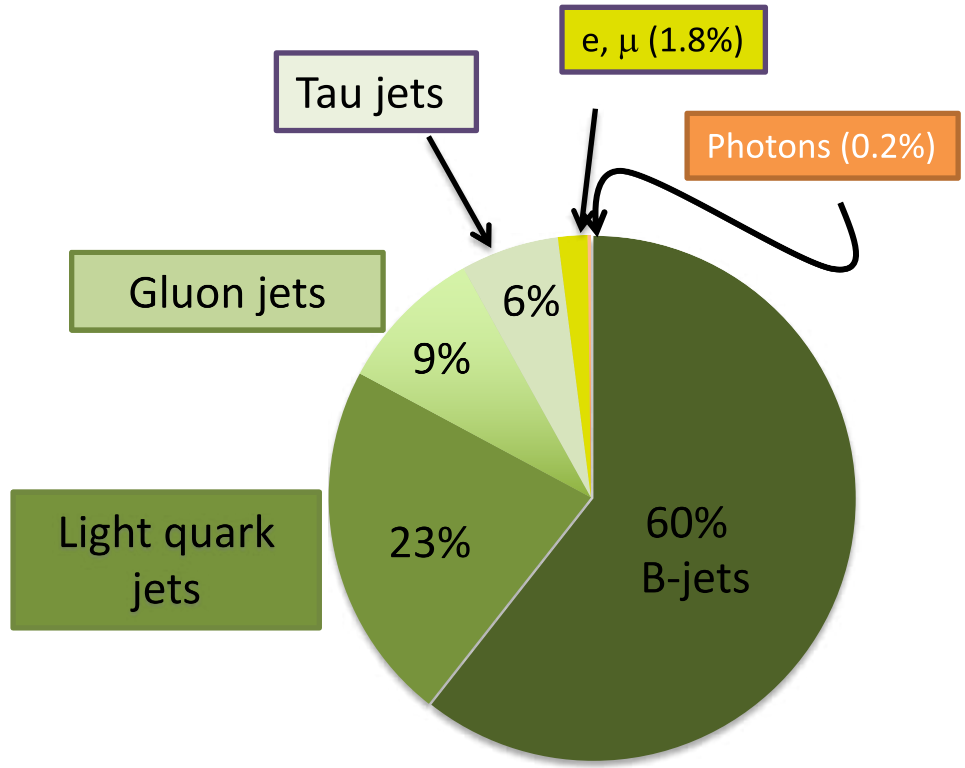
Jets from come from
chomoelectromagnetic radiation



Not all jets are created equal



=



Background is

- 80 billion gluon jets
- 10 billion light quark jets

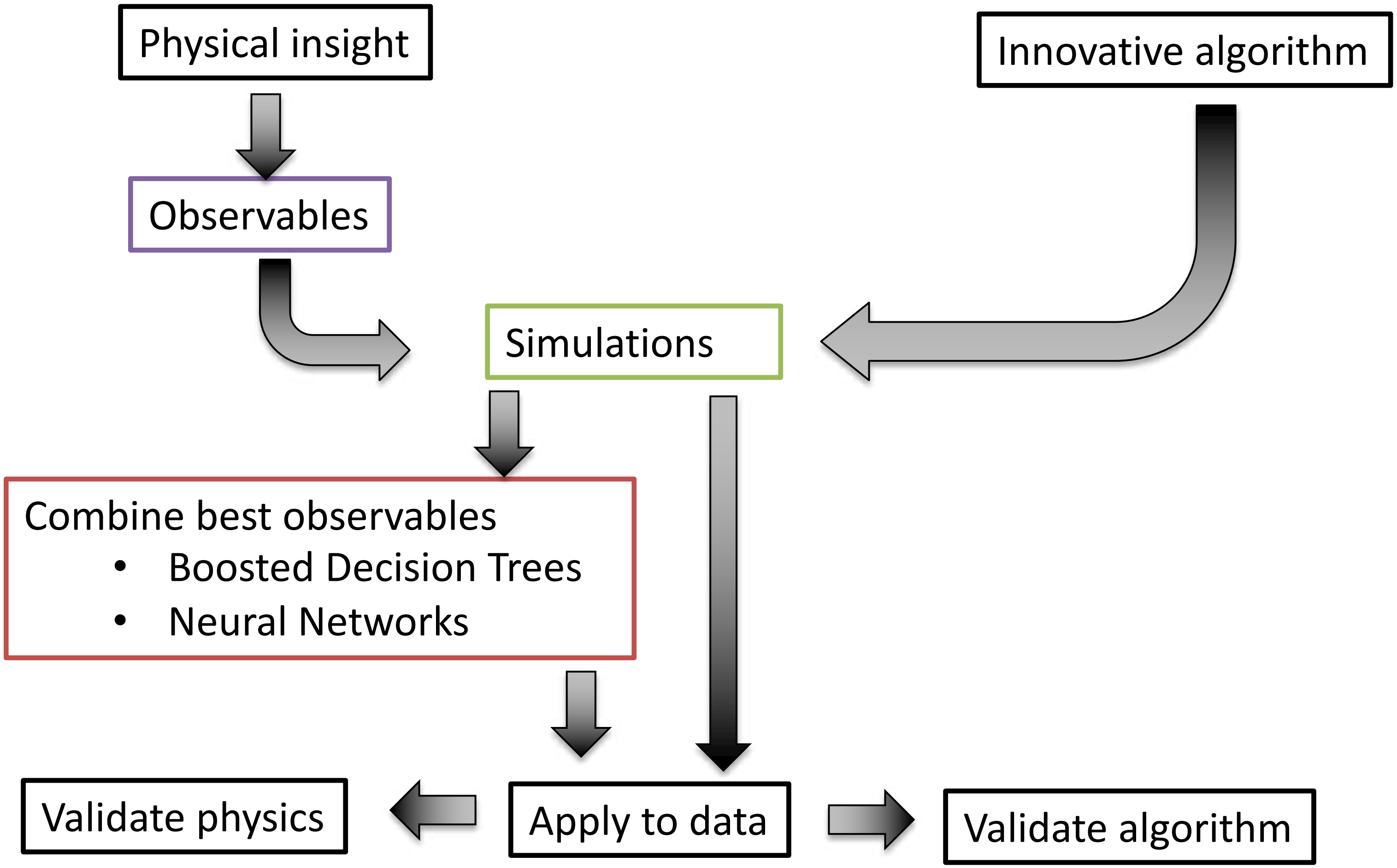
How can we tell all these different jet types apart?

MODERN MACHINE LEARNING

Modern Machine Learning for Particle Physics

Traditional approach

Modern machine learning



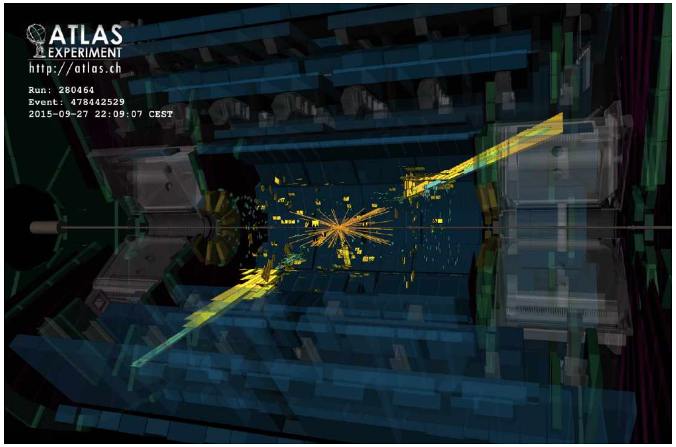
Game is about Hammers and Nails:



- Convolutional Neural Networks
- Recurrent Neural Networks
- Variational Auto-encorders
- Latent Dirichlet Allocation
- Reinforcement learning
- Point cloud networks
- Cluster networks

- Top tagging
- W tagging
- Quark/gluon discrimination
- Pileup removal
- b/c/s-tagging
- Jet-energy scale calibration
- Missing energy measurement
- Jets in heavy ion collisions

Physics domain is *distributions*



Is this event

- two quark jets
- two gluon jets
- two Higgs bosons?

- Any individual event has no “truth” identity
- All that exists are the probability distributions for different truths

$$dP_q(x) = \frac{d^n\sigma_q}{\underbrace{dp_1 \cdots dp_n}} \qquad dP_g(x) = \frac{d^n\sigma_g}{dp_1 \cdots dp_n}$$

1000-dimensional phase space
(from 10^8 dimensional measurements!)

goal is to test/measure

- data is a combination

$$dP_{\text{data}}(x) = \alpha_q dP_q + \alpha_g dP_g + \cdots$$

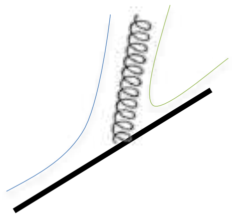
Matthew Schwartz

QUARKS VS GLUONS

Quark jets vs gluon jets: theory

Probability of quark radiating:

Color charge = 3

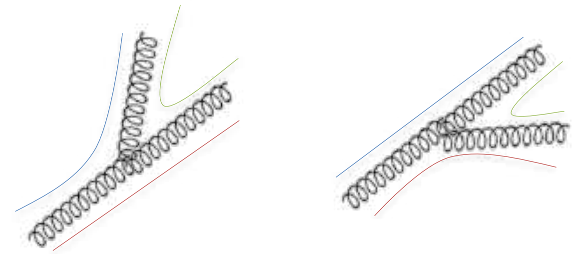


$$dP_q \sim 3 \times \frac{\alpha_s}{4\pi}(\cdots)$$

3 colors of quark

Probability of gluon radiating:

Color charge = 8

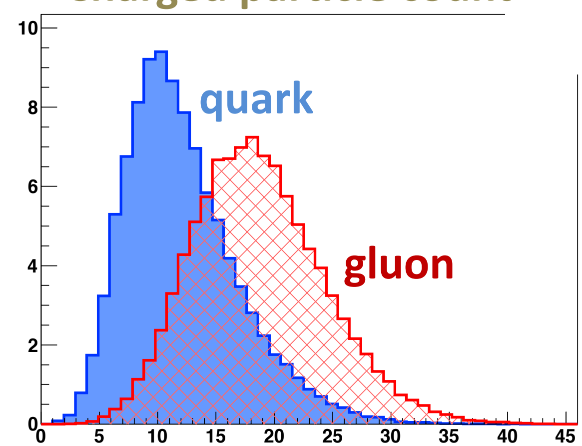


$$dP_g \sim 8 \times \frac{\alpha_s}{4\pi}(\cdots)$$

8 colors of gluon

- Gluons around twice as likely to radiate than
  - Gluon jets are fatter, on average
  - Gluon jets are more massive, on average
  - Gluon jets have more particles, on avera
  - …

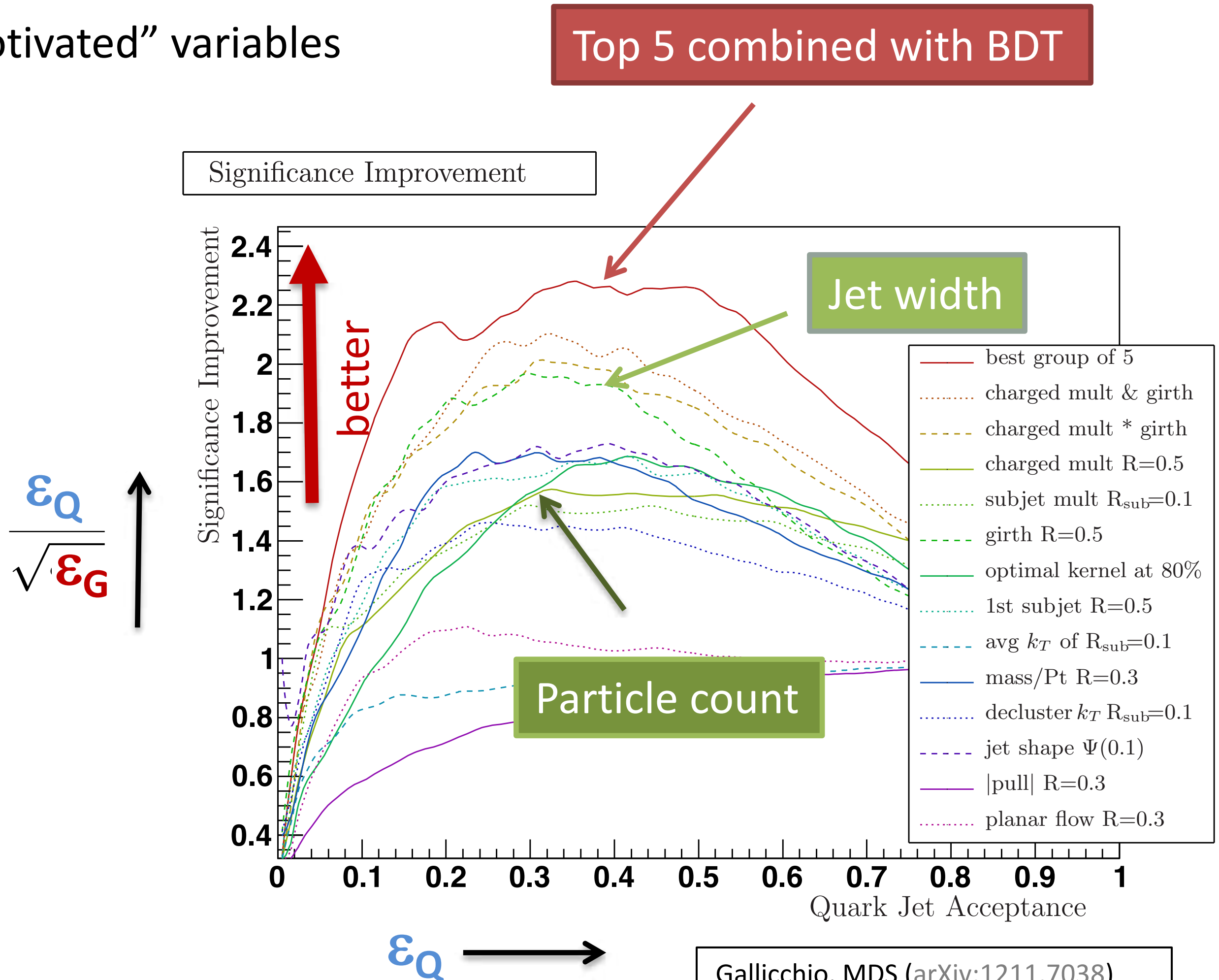
Charged particle count



Traditional approach

Consider lots of "motivated" variables

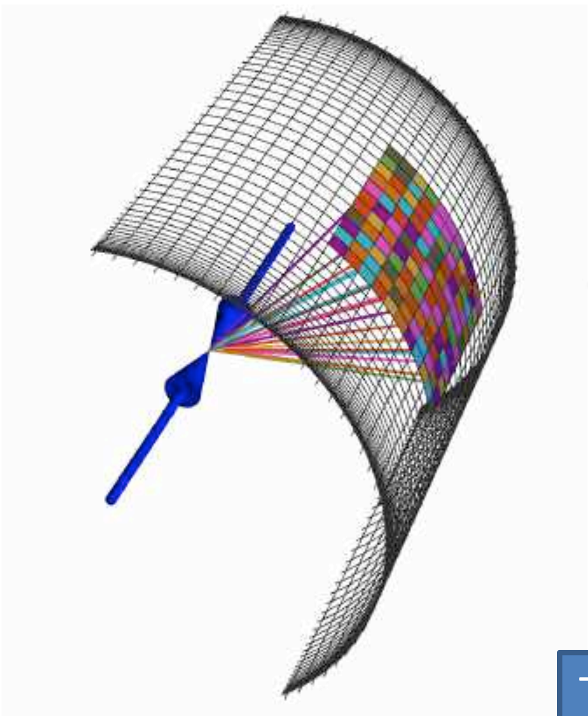
- Jet width
- # of particles
- # of subjets
- Jet "shape"
- Jet mass
- …

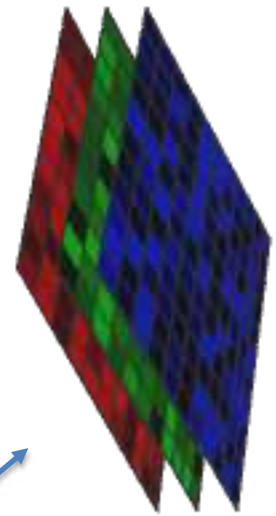


Gallicchio, MDS (arXiv:1211.7038)

Convolutional Neural Networks for quark/gluon jet discrimination

Komiske, Metodiev, MDS (arXiv:1612.01551)



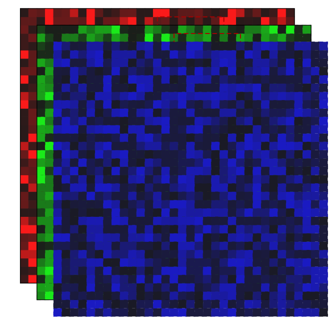


Three input layers

preprocess

- Center
- Crop
- Normalize
- Zero
- Standardize

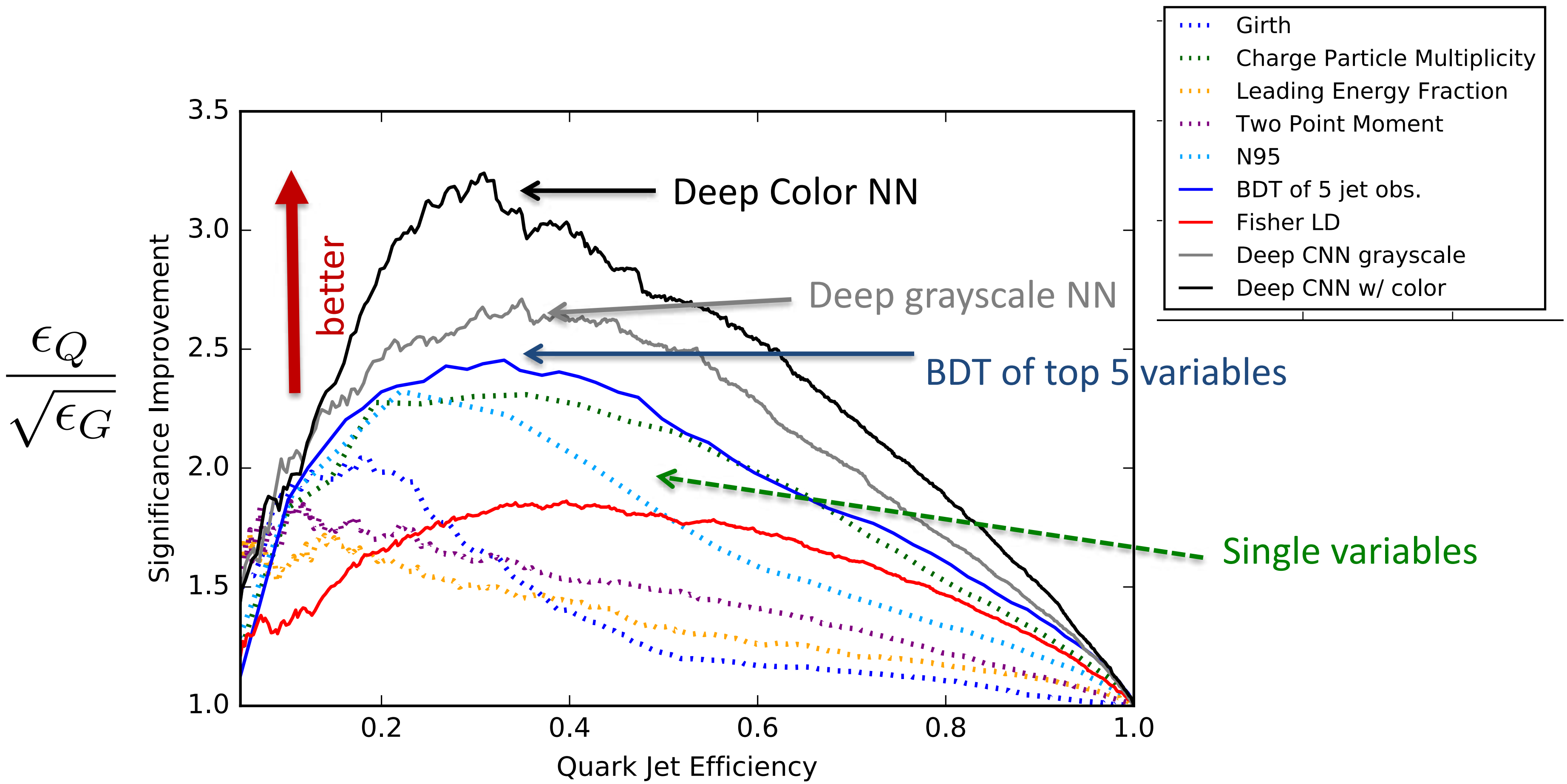
NN inputs



- Red = energy of charged particles
- Green = energy of neutral particles
- Blue = number of charged particles

Quark/Gluon CNN results

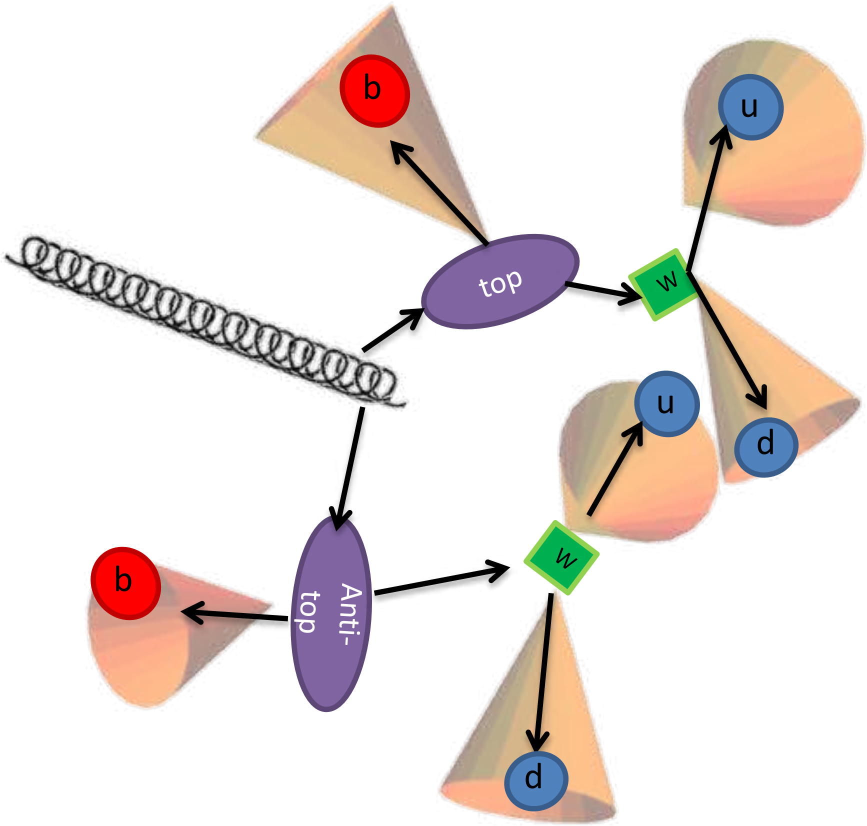
Komiske, Metodiev, MDS (arXiv:1612.01551)



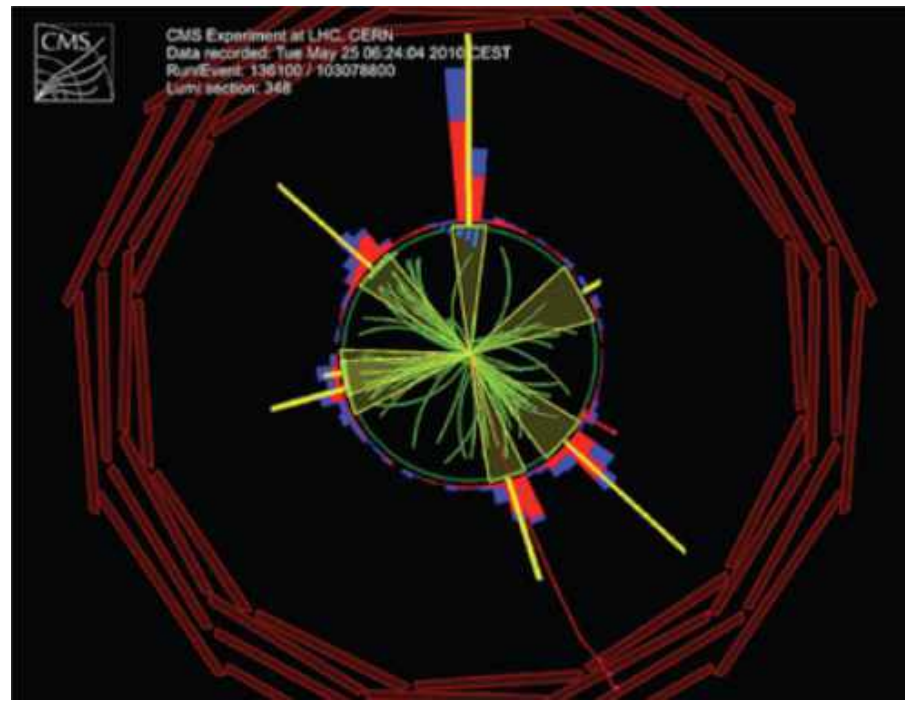
TOP TAGGING

Top-tagging

Hypothetical new heavy particles often decay to top quarks:



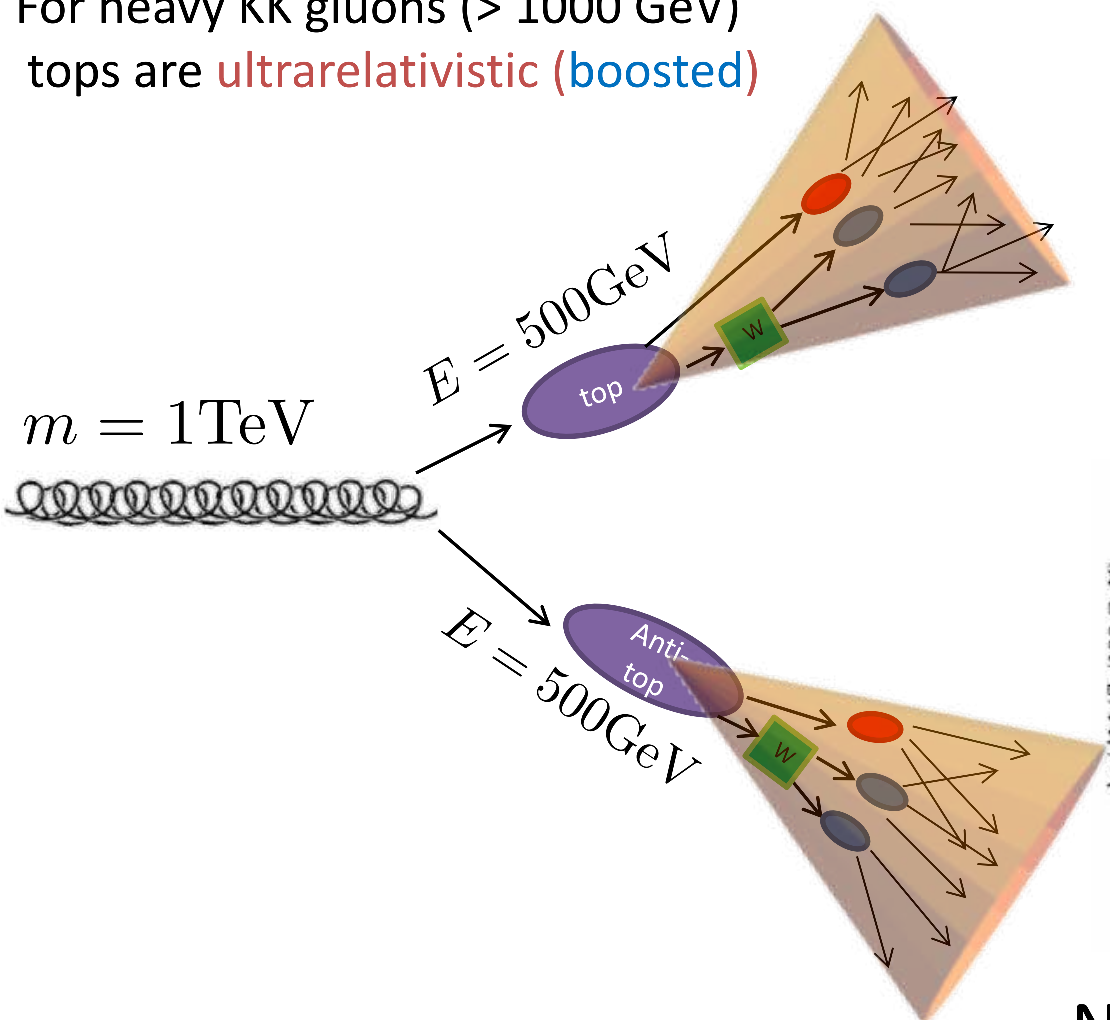
Looks like 6 Jets

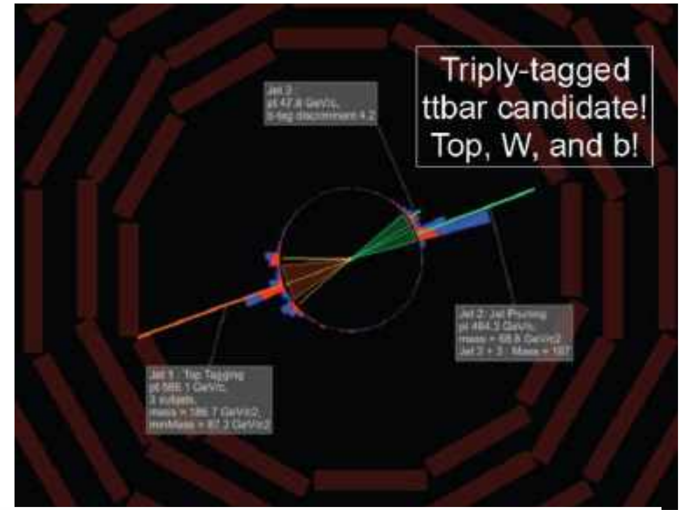


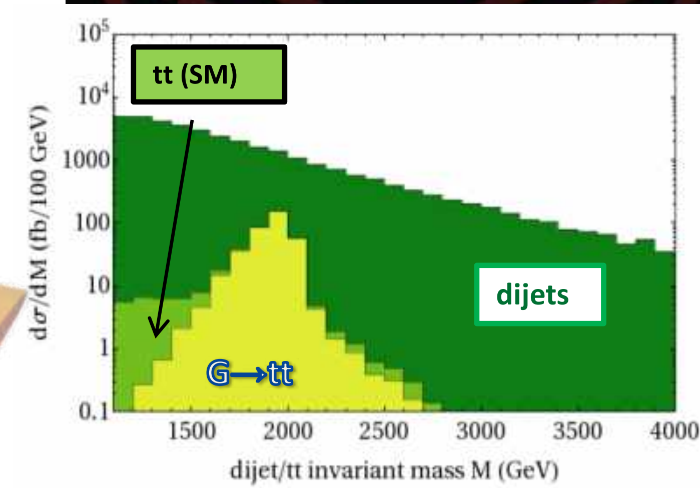
Matthew Schwartz

Tops are often *boosted*

For heavy KK gluons (> 1000 GeV)
tops are ultrarelativistic (boosted)



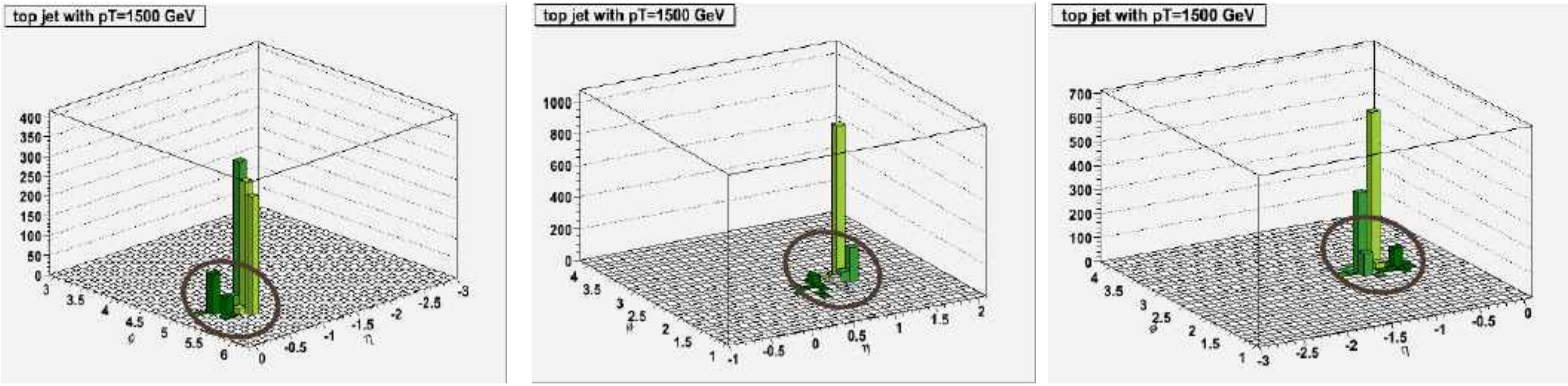




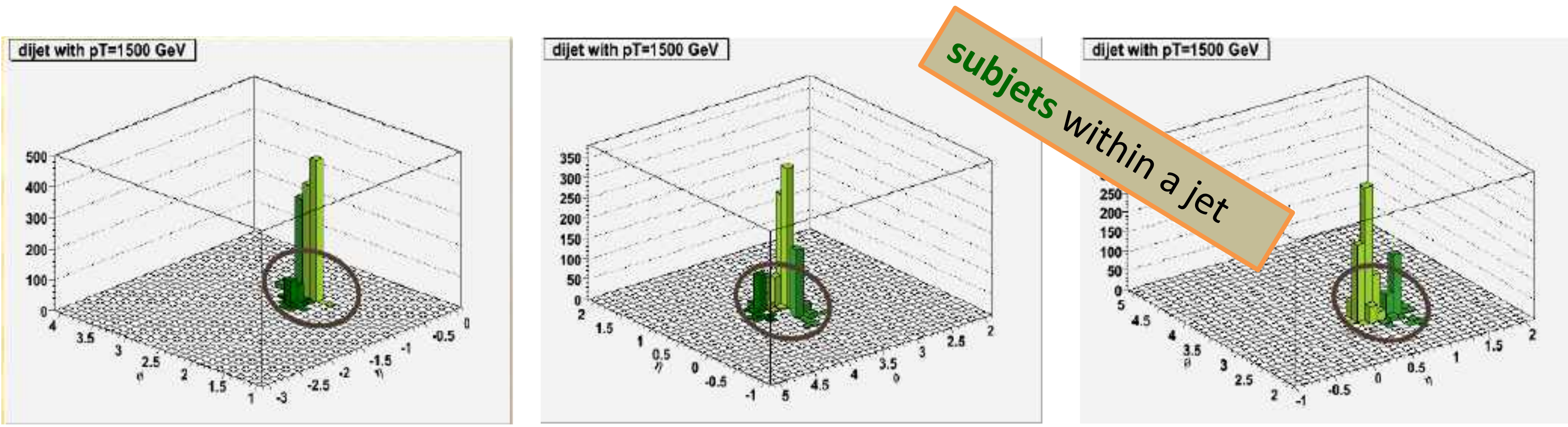
Now it looks like 2 jets!

Typical top jets

Large boost (P_T = 1500 GeV)



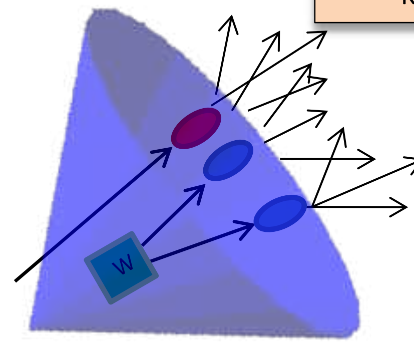
Typical background jets



Matthew Schwartz

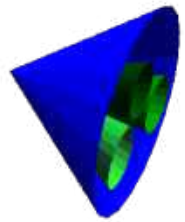
Hopkins top-tagger
Kaplan et al. arXiv:0806.0848

Top-tagging

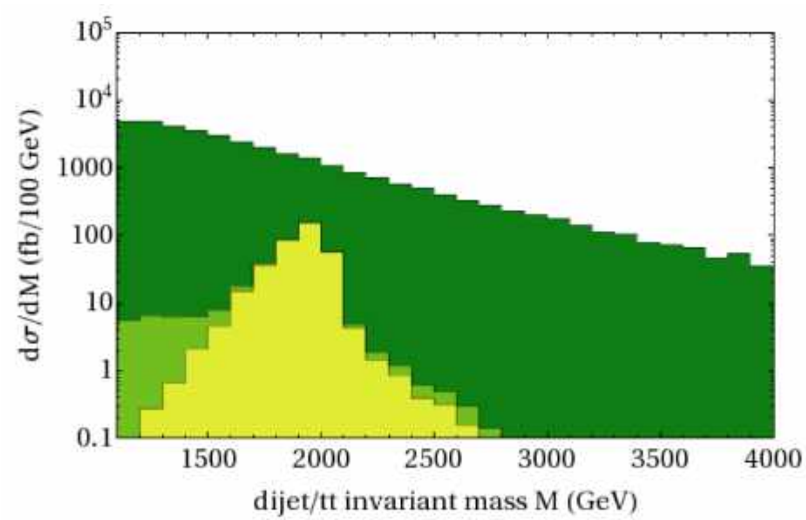


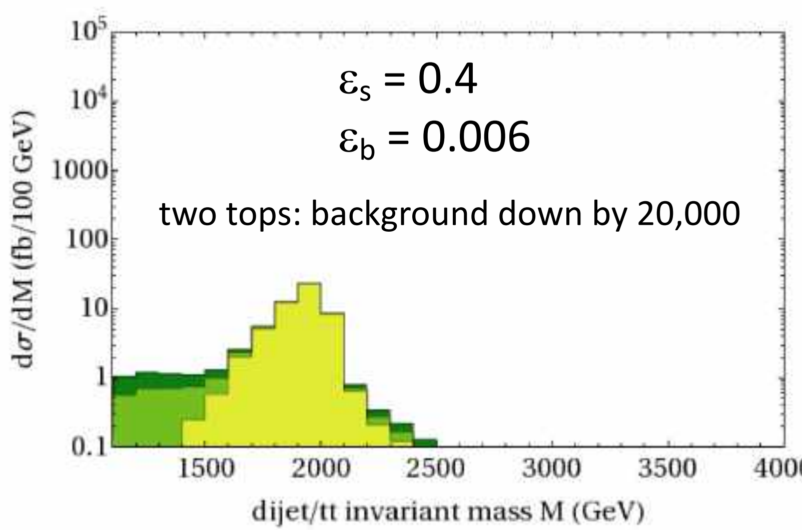
1. Look for big jets (R = 1.2)

2. with subjets within the jet



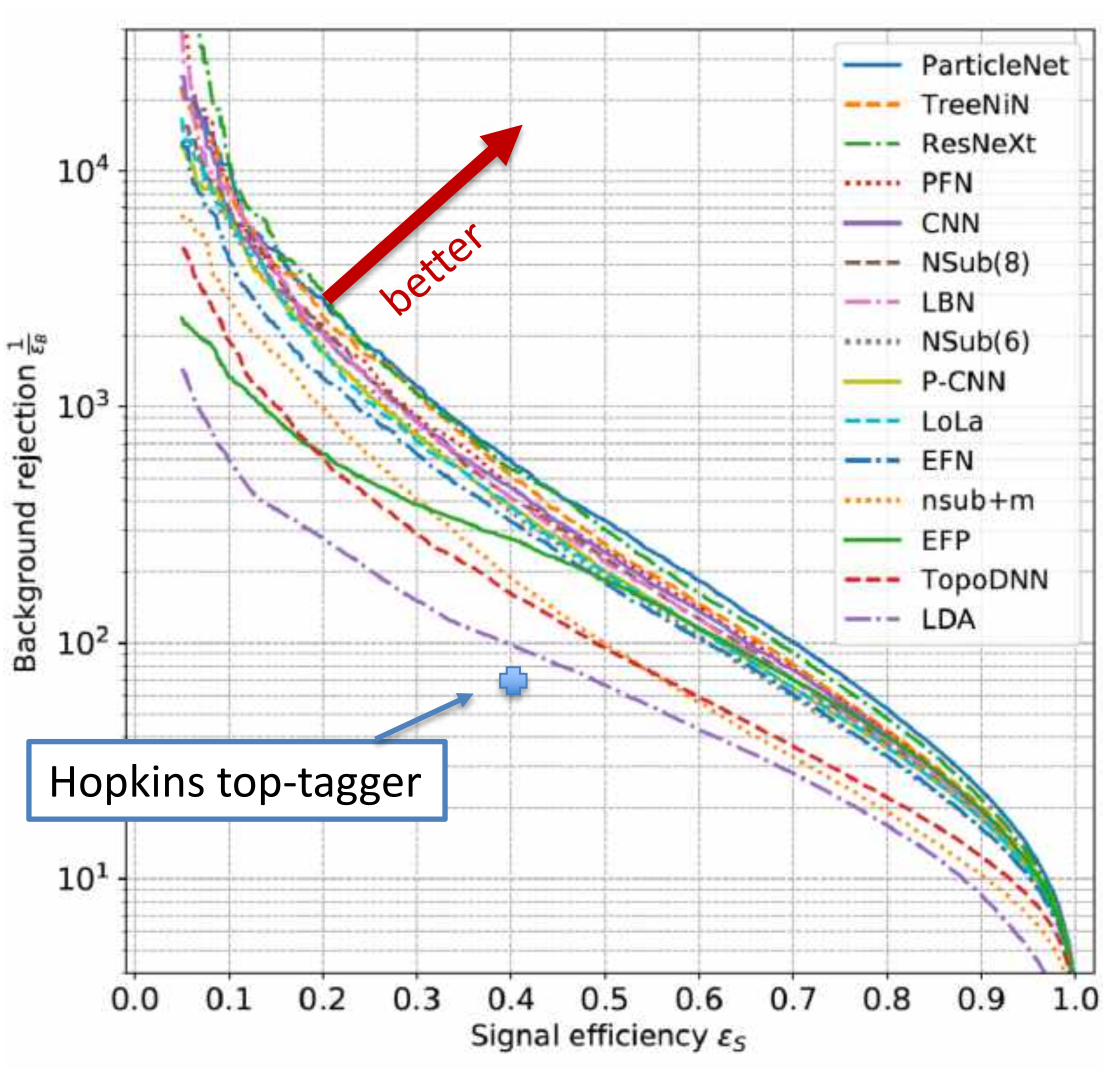
3. Analyze the subjets
   - W mass peak, top mass peak, and helicity angle





Many ML hammers hit the top-tagging nail

arXiv:1902.09914



- Factor of 10 better background rejection than traditional taggers

Apples-to-apples top-tagging comparison

arXiv:1902.09914

		AUC	Acc	$1/\epsilon_B$ ($\epsilon_S = 0.3$) single	mean	median	#Param
image based	CNN [16]	0.981	0.930	914±14	995±15	975±18	610k
	ResNeXt [30]	0.984	0.936	1122±47	1270±28	1286±31	1.46M
4-vector based	TopoDNN [18]	0.972	0.916	295±5	382± 5	378 ± 8	59k
	Multi-body N-subjettiness 6 [24]	0.979	0.922	792±18	798±12	808±13	57k
	Multi-body N-subjettiness 8 [24]	0.981	0.929	867±15	918±20	926±18	58k
	TreeNiN [43]	0.982	0.933	1025±11	1202±23	1188±24	34k
	P-CNN	0.980	0.930	732±24	845±13	834±14	348k
	ParticleNet [47]	0.985	0.938	1298±46	1412±45	1393±41	498k
theory inspired	LBN [19]	0.981	0.931	836±17	859±67	966±20	705k
	LoLa [22]	0.980	0.929	722±17	768±11	765±11	127k
	Energy Flow Polynomials [21]	0.980	0.932	384			1k
	Energy Flow Network [23]	0.979	0.927	633±31	729±13	726±11	82k
	Particle Flow Network [23]	0.982	0.932	891±18	1063±21	1052±29	82k
particle cloud	GoaT	0.985	0.939	1368±140		1549±208	35k
	ParticleNet-Lite	0.984	0.937	1262±49			26k
	ParticleNet	0.986	0.940	1615±93			366k

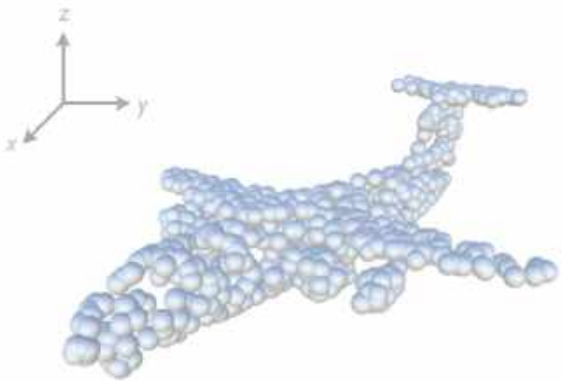
- Uses same samples
- 800k training, 200k test
  - So good, limited by sample size
  - GoaT: 200k/1368 = 146 bg events survive

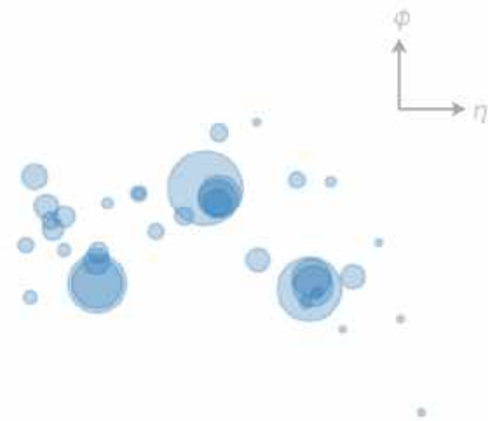
Particle net uses point cloud approach

Gouskos and Qu: (arXiv:1902.08570)

- Respects permutation symmetry

(other approaches include energy flow polynomials, arXiv:1712.07124)





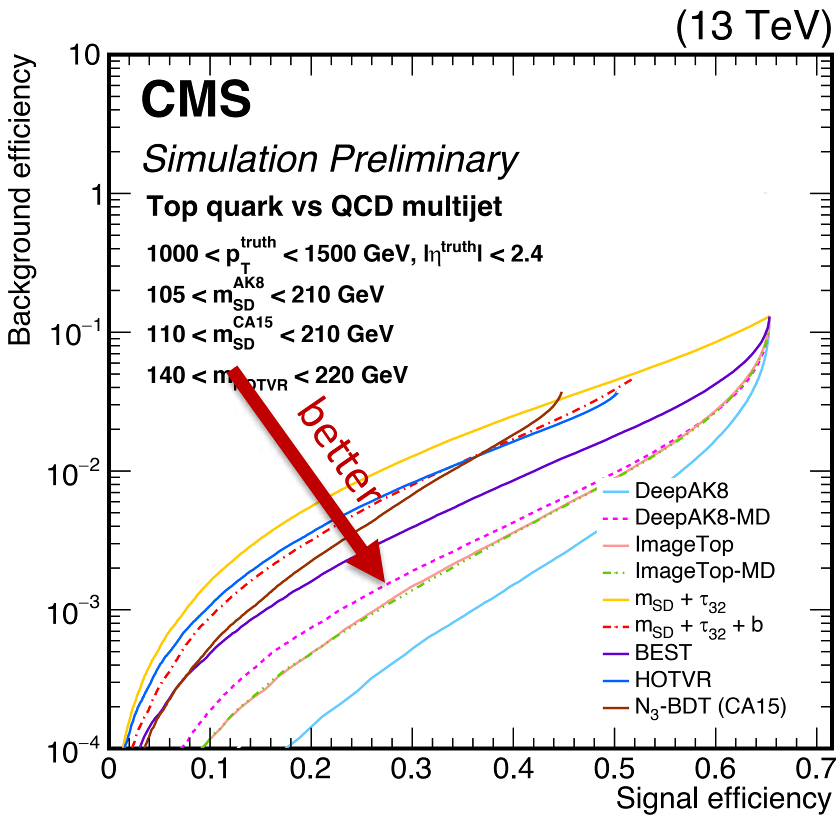
- Point cloud
  - points are intrinsically *unordered*
  - primary information:
    - 3D coordinates in the xyz space

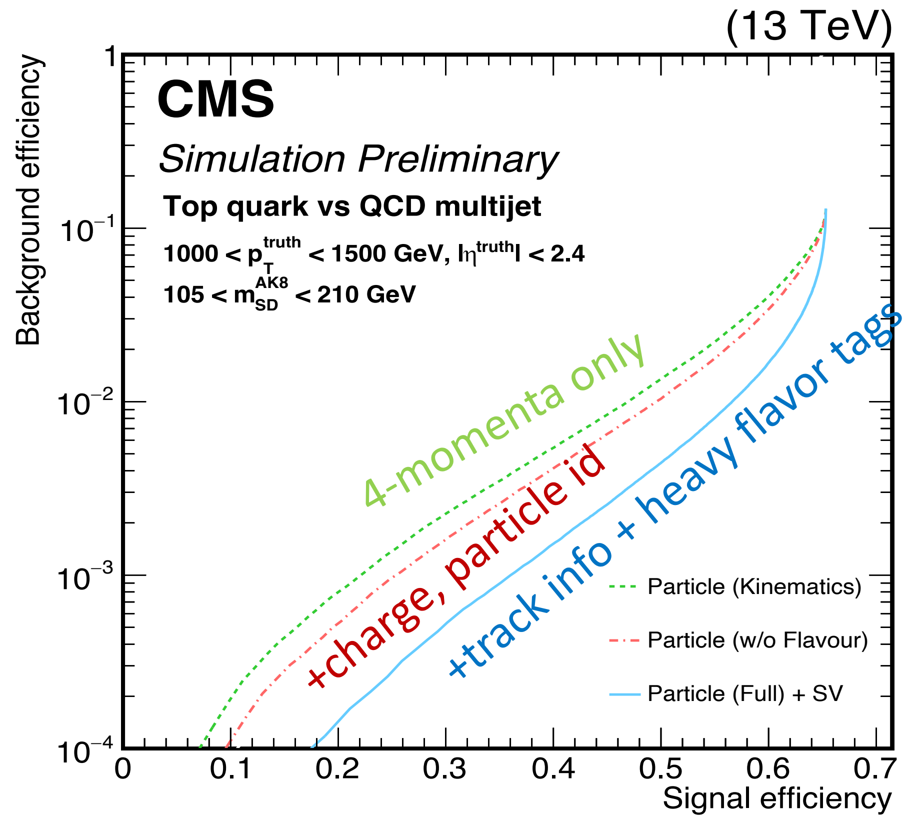
- Particle cloud
  - particles are intrinsically *unordered*
  - primary information:
    - 2D coordinates in the η-ϕ space
  - *but also many additional features!*
    - energy/momenta
    - charge/particle ID
    - track quality/impact parameters/etc.

- uses EdgeConv
- angular distance metric
- k-nearest neighbors

CMS top-tagging: DeepAK8 ML algorithm

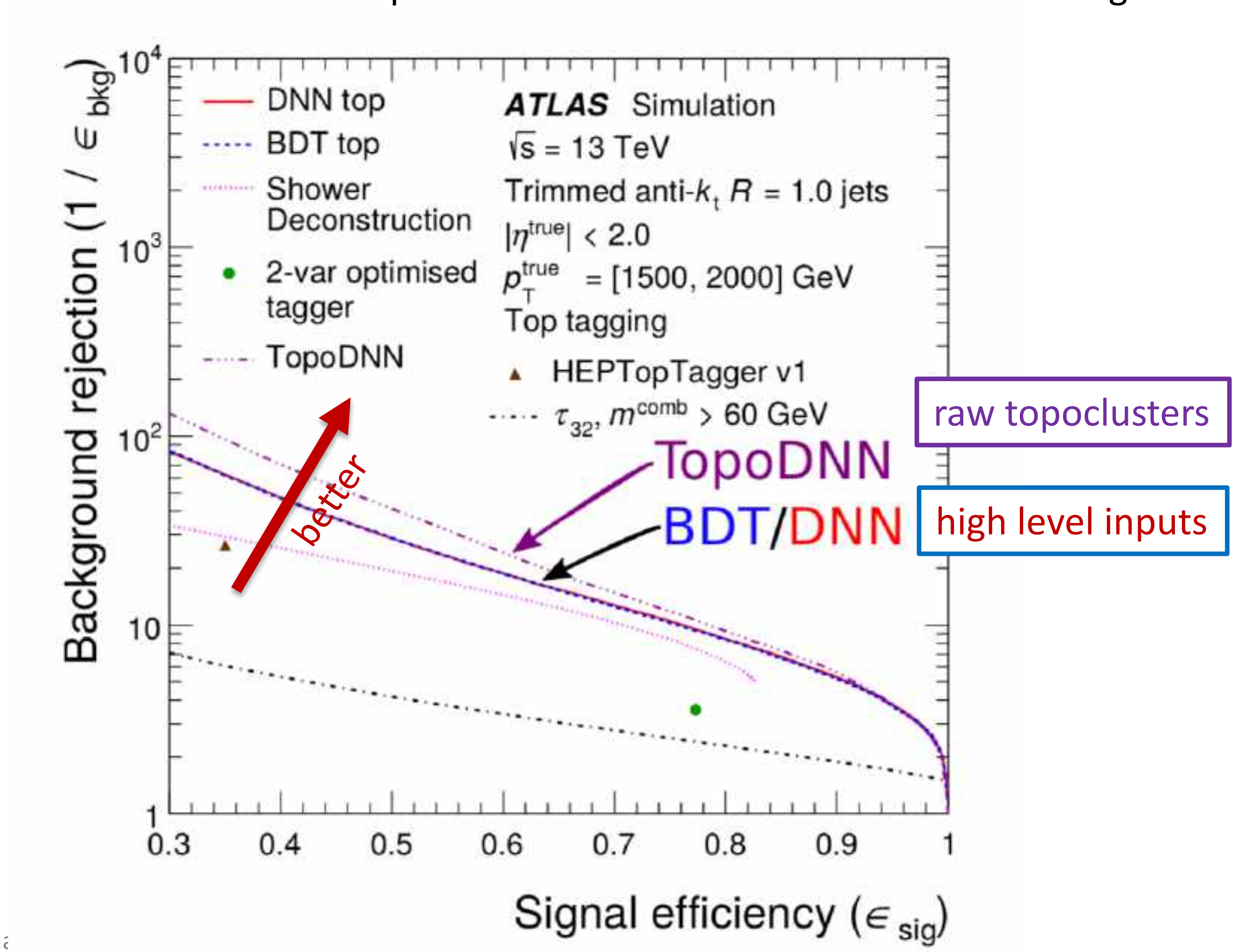
- Inputs particles/tracks to ResNet archictecture





ATLAS top-tagging : topoDNN

- inputs topoclusters to deep neural network
- also finds better performance with modern machine learning

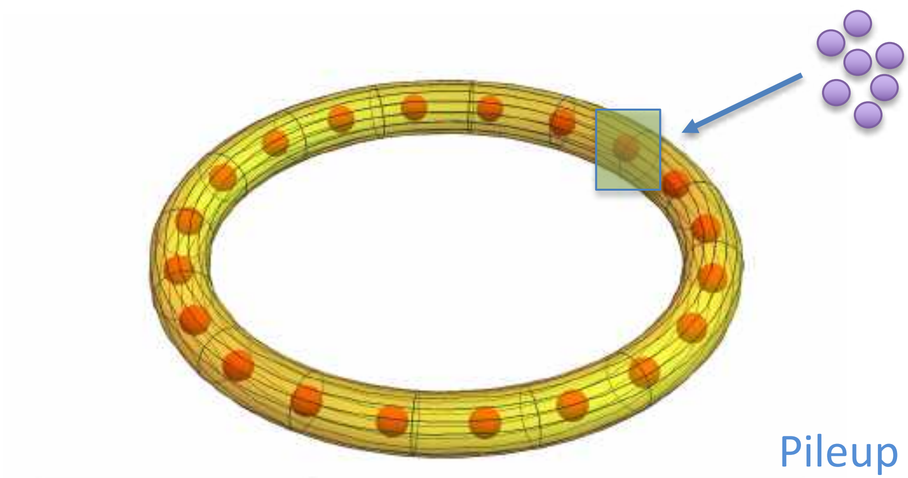


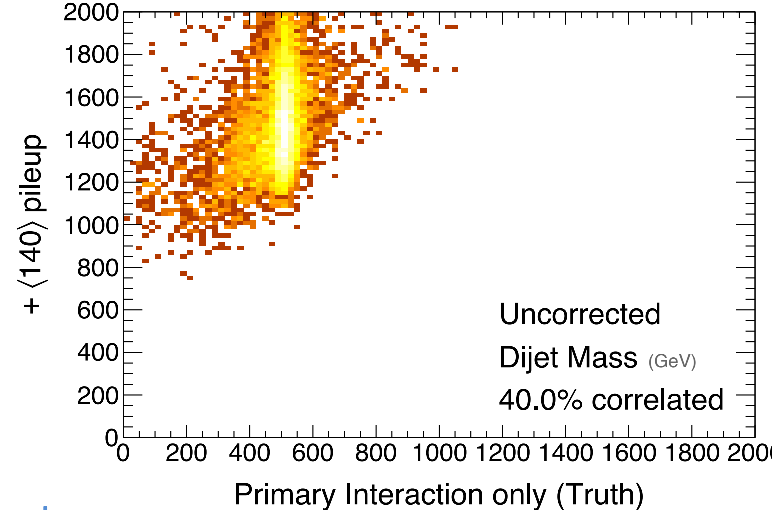
Ma

PILEUP REMOVAL

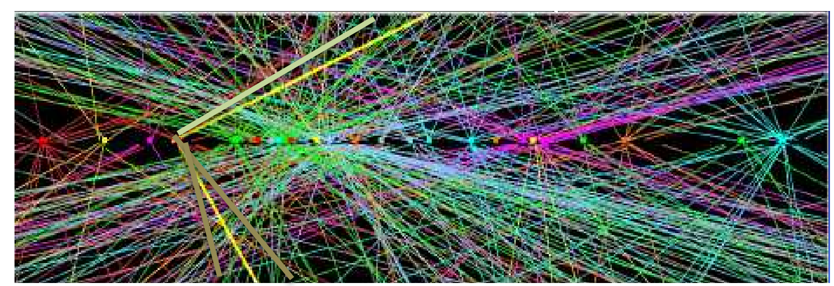
Pileup

- LHC collides protons in bunches
  - 10^{11} protons/bunch
  - Up to 200 collisions per bunch crossing





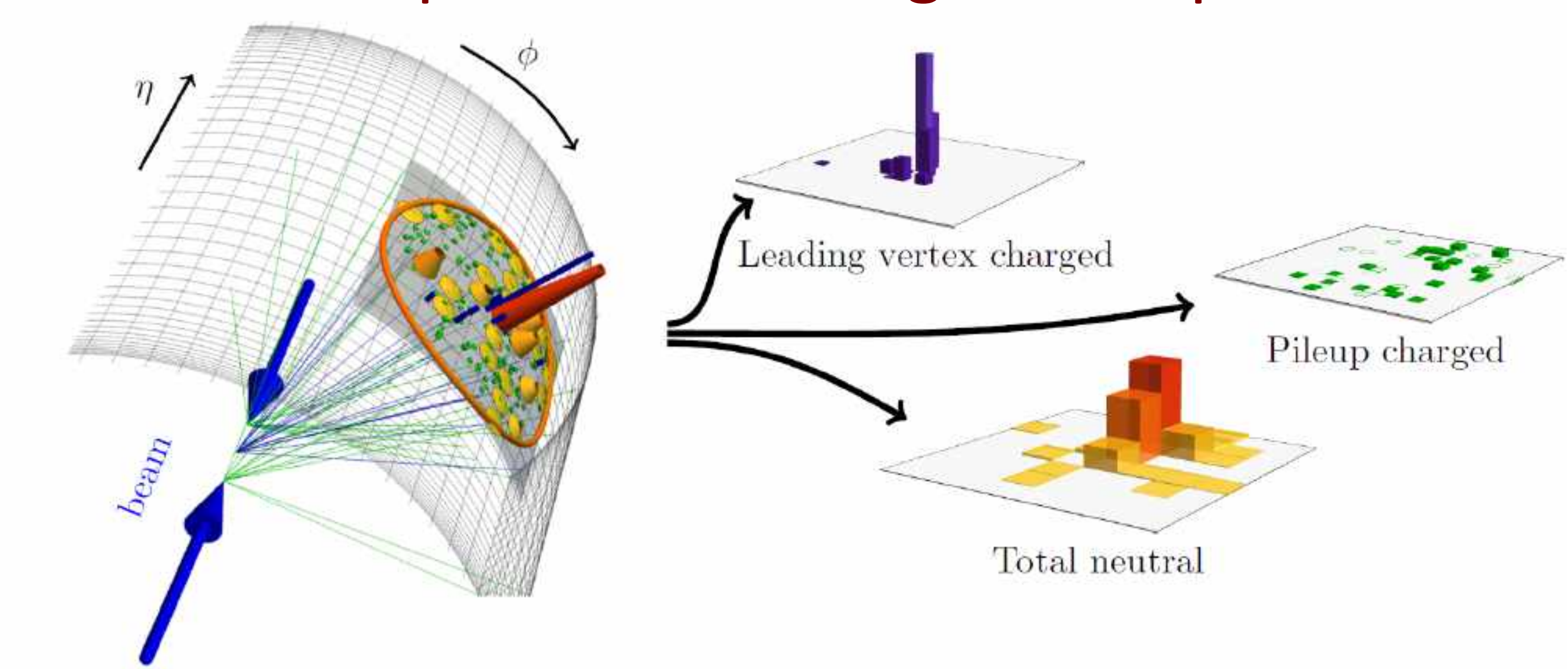
Pileup removal algorithm



- Tracking system can resolve primary collision from secondary "pileup" collisions
- Only charged particles can be seen this way

- Can we use machine learning to remove the pileup radiation?

Pileup removal as regression problem



Can measure

1. Leading vertex charged particles
2. Pileup charged particles
3. Total neutral particles

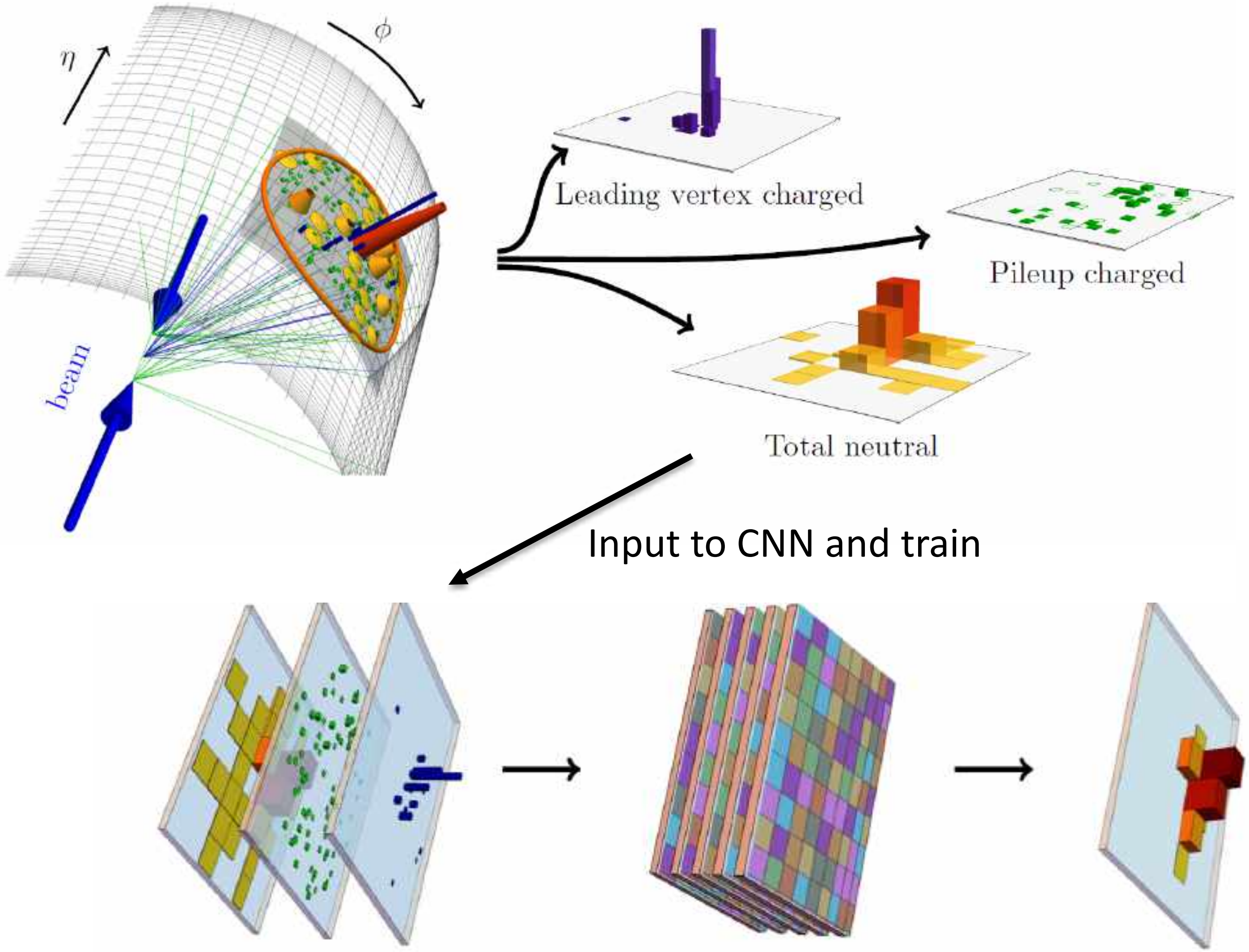
?

Leading vertex neutral particles

Convnets for Pileup Removal

Komiske, Metodiev, Nachman, MDS (arXiv:1707.08600)

- Separate observable energy deposits into 3 images



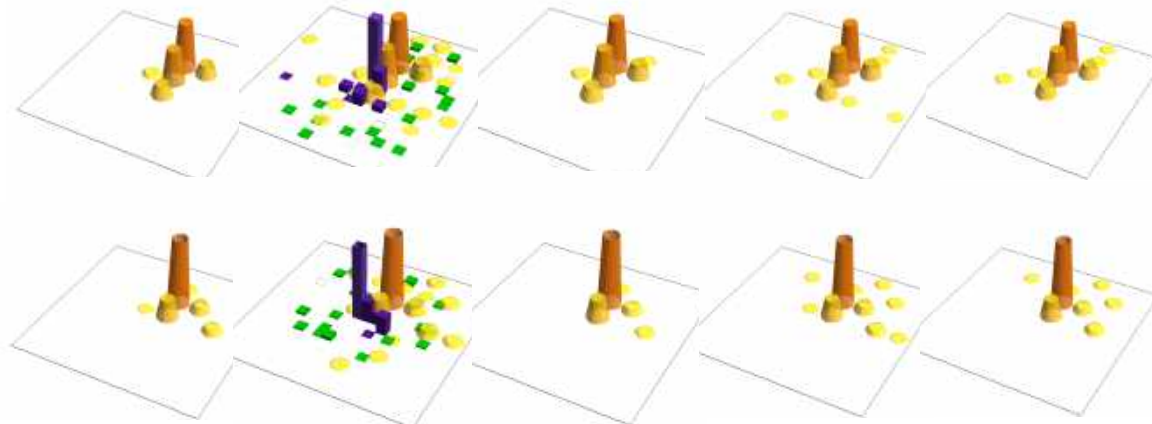
PileUp Mitigation with Machine Learning (PUMML)

Komiske, Metodiev, Nachman, MDS (arXiv:1707.08600)

Traditional approaches

Truth | With 140 piluep events | PUMML | PUPPI | SoftKiller

- Excellent leading vertex (truth) reconstruction

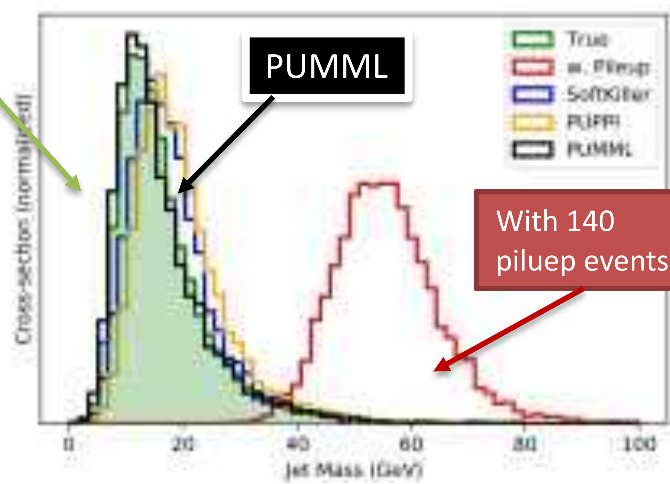


- Excellent observable reconstruction

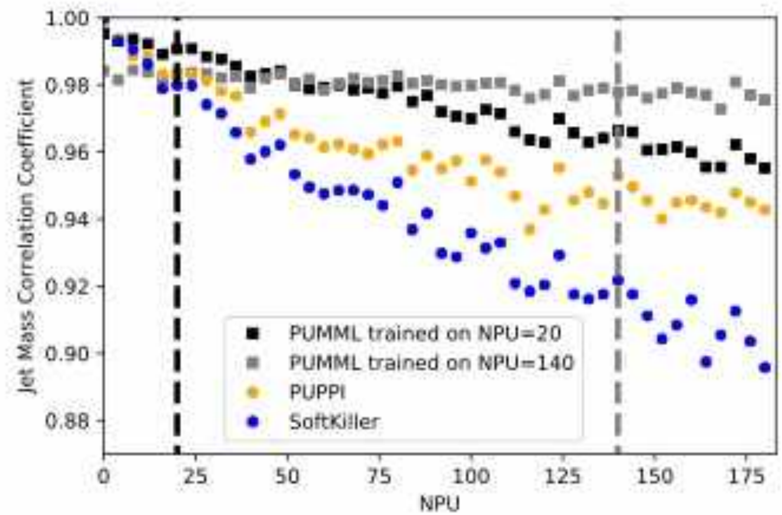
Truth

PUMML

With 140 piluep events



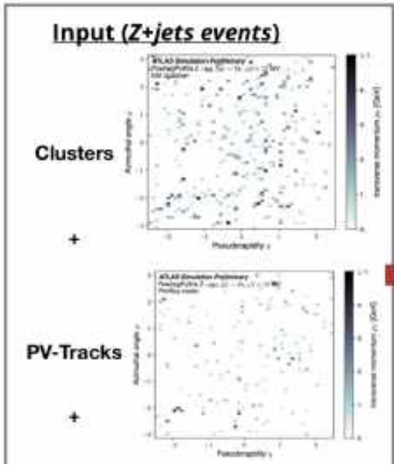
- Excellent stability for variable piluep #

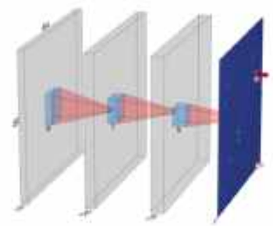


ATLAS uses conv nets to measure MET (in simulation)

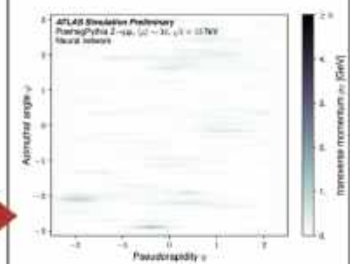
ATL-PHYS-PUB-2019-028

$$E_{\mathrm{T}}^{\mathrm{miss}} = -\left(\sum_{i\,\in\mathrm{muons}} {p_{\mathrm{T}}}^{i} + \sum_{i\,\in\mathrm{electrons}} {p_{\mathrm{T}}}^{i} + \sum_{i\,\in\mathrm{photons}} {p_{\mathrm{T}}}^{i} + \sum_{i\,\in\mathrm{hadronic}\ \tau} {p_{\mathrm{T}}}^{i} + \sum_{i\,\in\mathrm{jets}} {p_{\mathrm{T}}}^{i} + \sum_{i\,\in\mathrm{Soft\ Term}} {p_{\mathrm{T}}}^{i} \right)$$

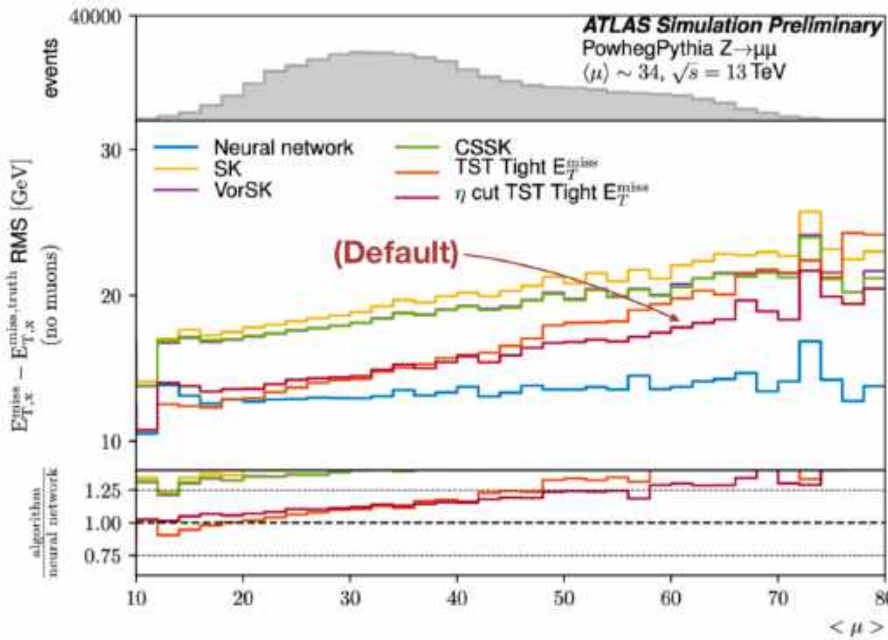




Layer	Kernelsize	Filters	Activation	Parameters
2D convolution	9	15	ReLU	7305
dropout				
2D convolution	7	20	ReLU	14720
dropout				
2D convolution	5	25	ReLU	12525
dropout				
2D convolution	1	1	ReLU	26
Total				34576



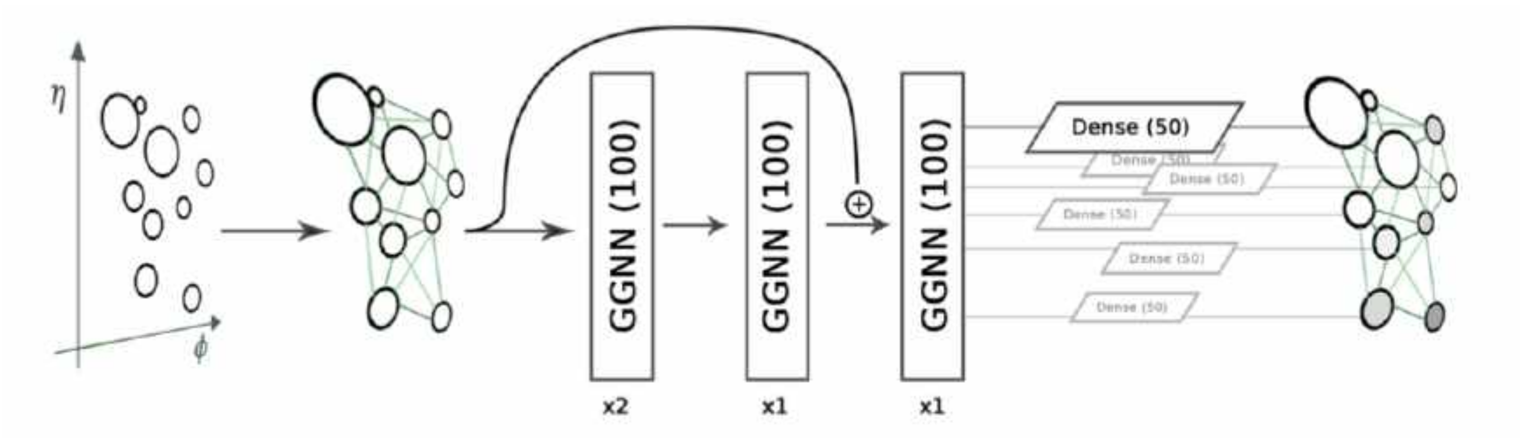
Output: **an event image with MET as-close-as-possible to the truth-level expectation.**

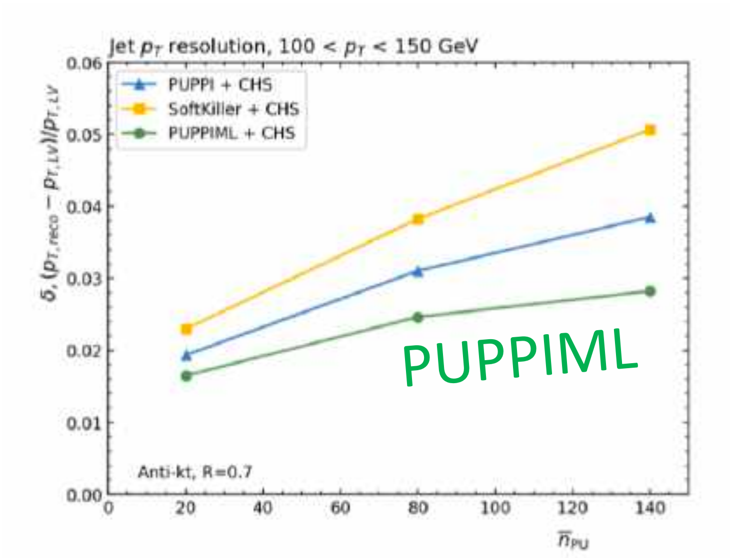


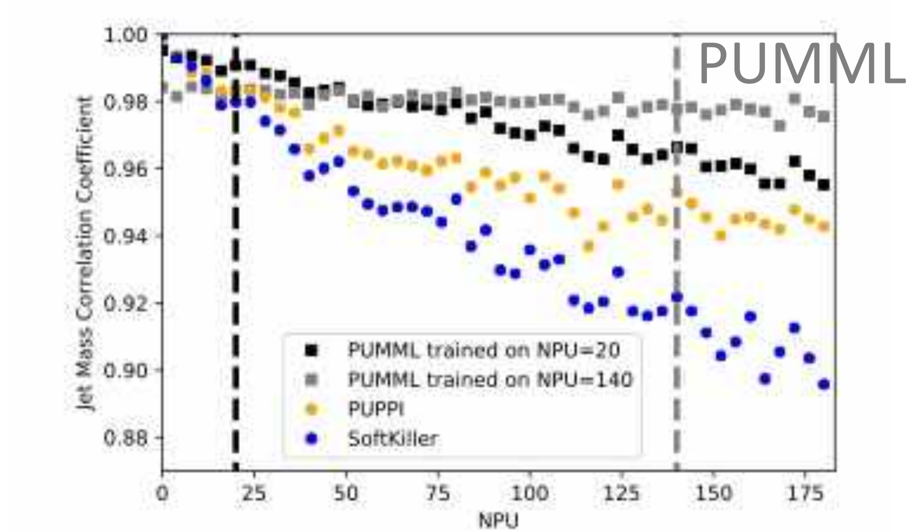
Conv net
image regression

PUPPIML: Graphnet approach

Martinez et al. (arXiv:1810.07988)







- good stability

- comparable to PUMML

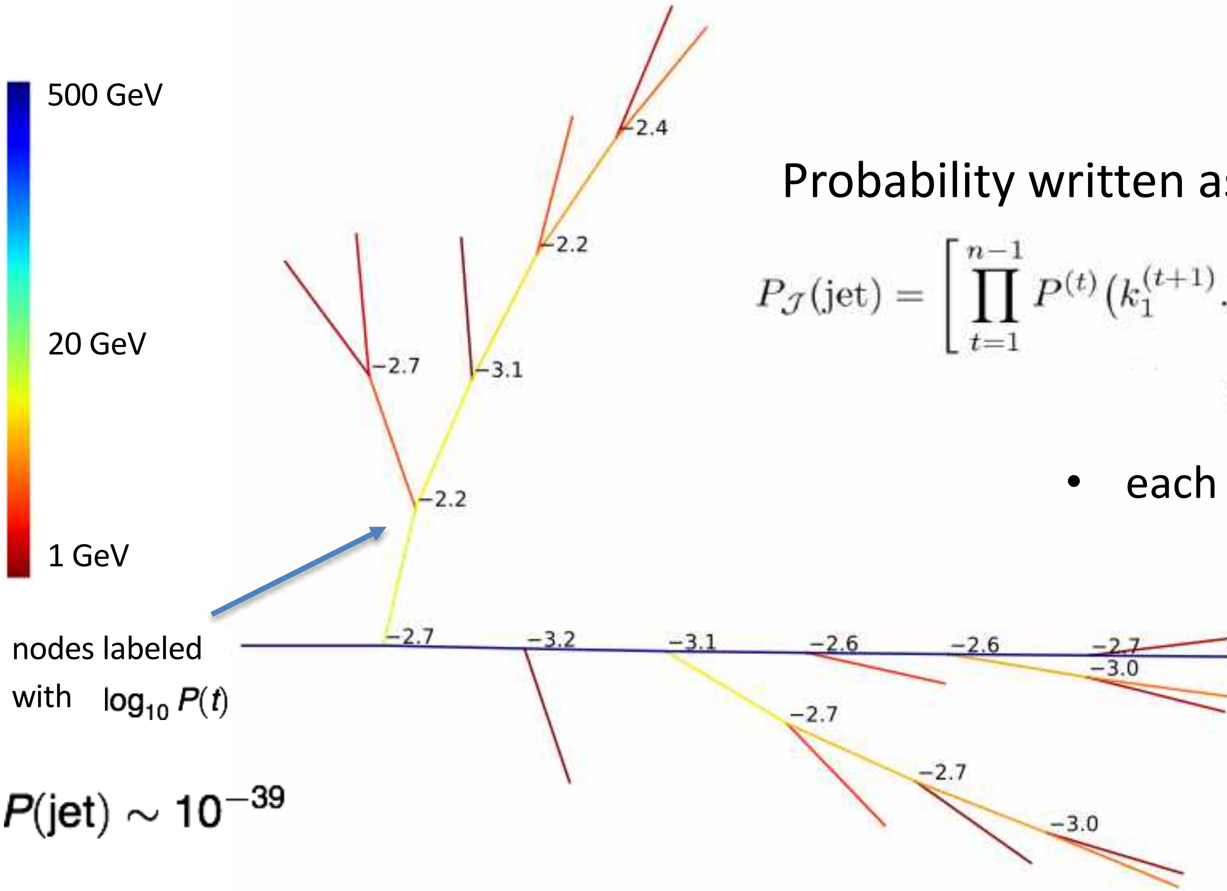
UNSUPERVISED LEARNING

JUNIPR

Andreassen, Feige, Frye, MDS
(arXiv:1804.09720, 1906.10137)

- unsupervised approach: learn probability distribution for each sample
- represent data as clustering tree
- can be used to classify or generate

$$dP_q(x) = \frac{d^n\sigma_q}{dp_1 \cdots dp_n}$$



nodes labeled with $\log_{10} P(t)$

$P(\text{jet}) \sim 10^{-39}$

Probability written as product

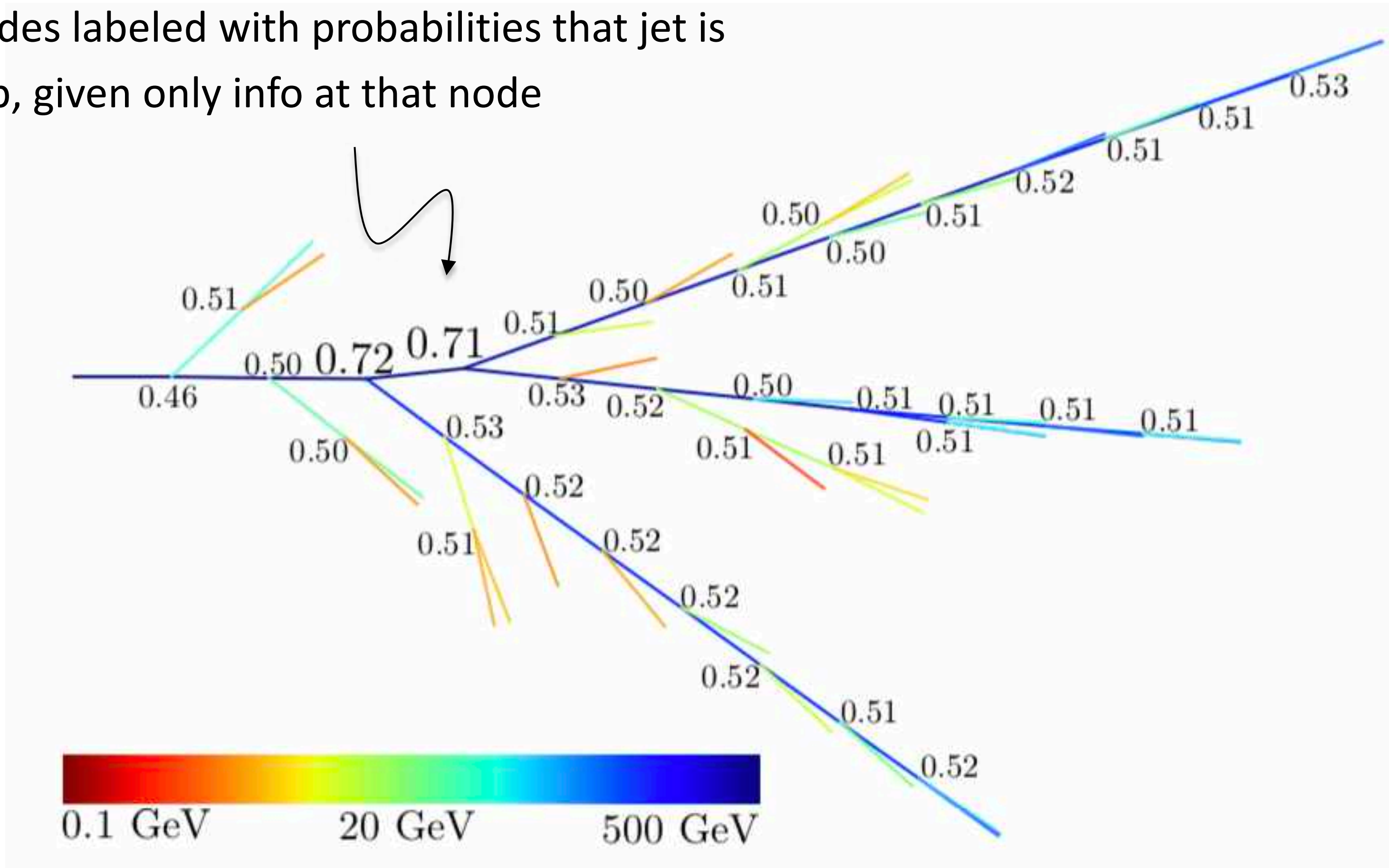
$$P_{\mathcal{J}}(\text{jet}) = \left[\prod_{t=1}^{n-1} P^{(t)}\left(k_1^{(t+1)} \dots k_{t+1}^{(t+1)} \middle| k_1^{(t)} \dots k_t^{(t)}\right)\right] \times P^{(n)}\left(\text{end}\middle| k_1^{(n)} \dots k_n^{(n)}\right)$$

- each term interpretable

JUNIPR

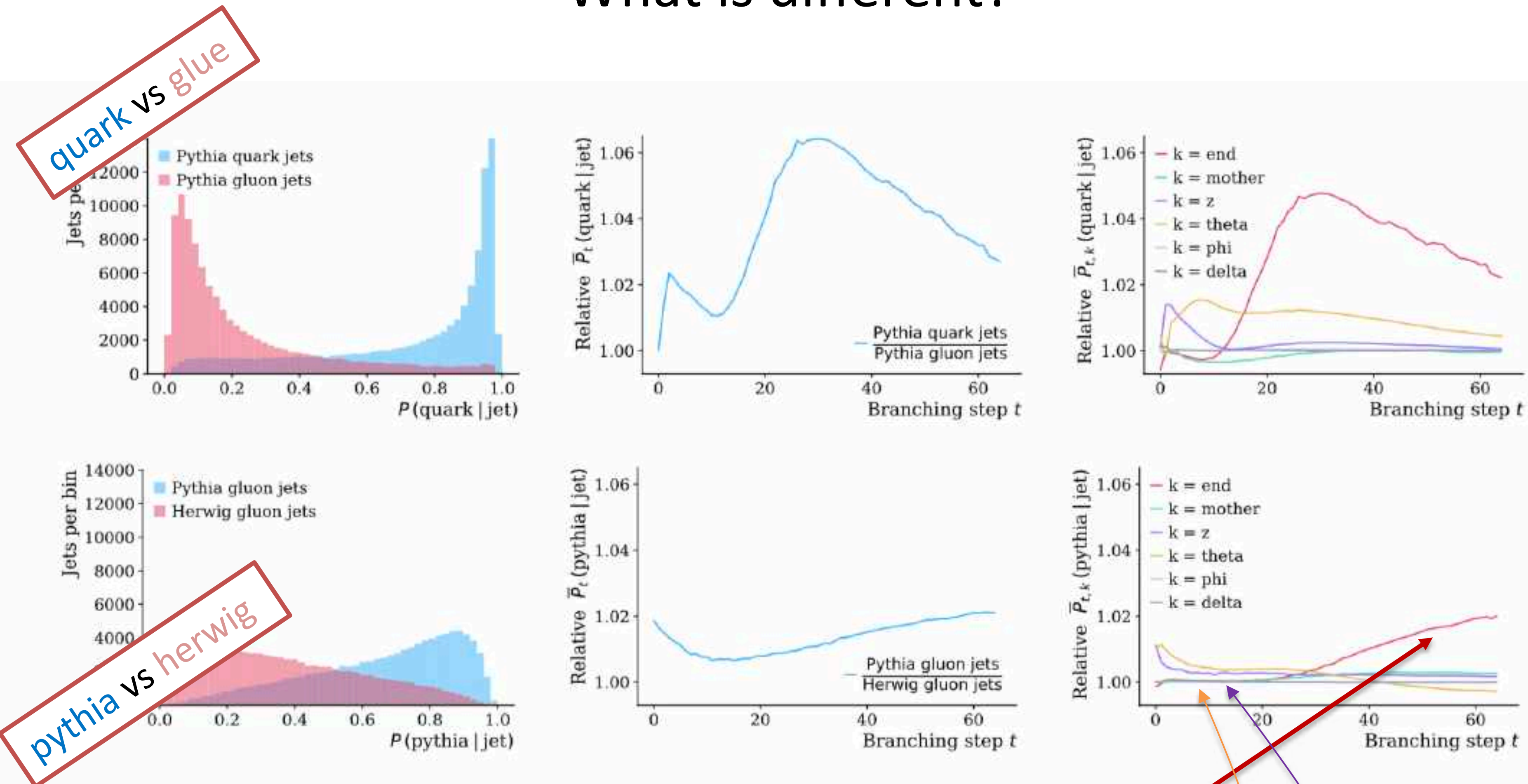
Visualizing discrimination power: boosted top or QCD jet?

- nodes labeled with probabilities that jet is top, given only info at that node



JUNIPR

What is different?

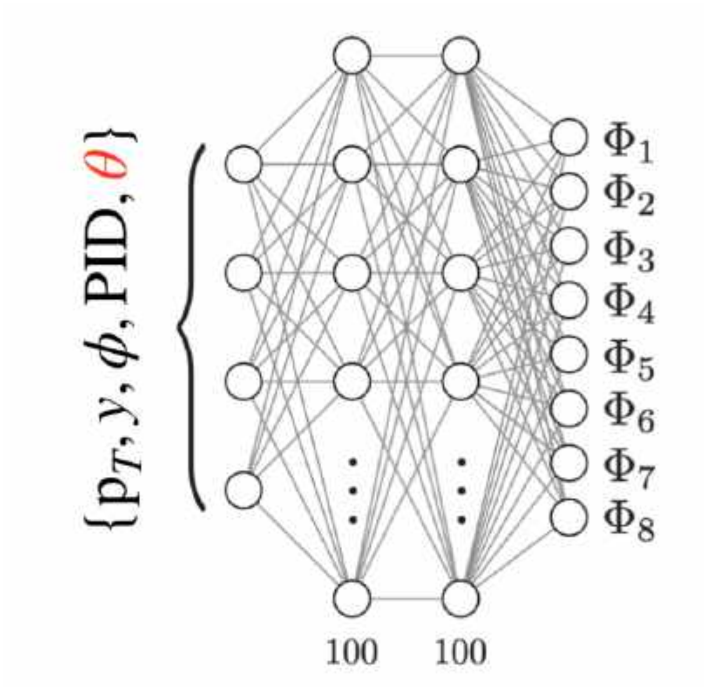


- Length of shower important
- angular distribution early on
- energy sharing different early on

Andreassen and Nachman
(arXiv:1907.08209)

DCTR: Use relative weights for tuning

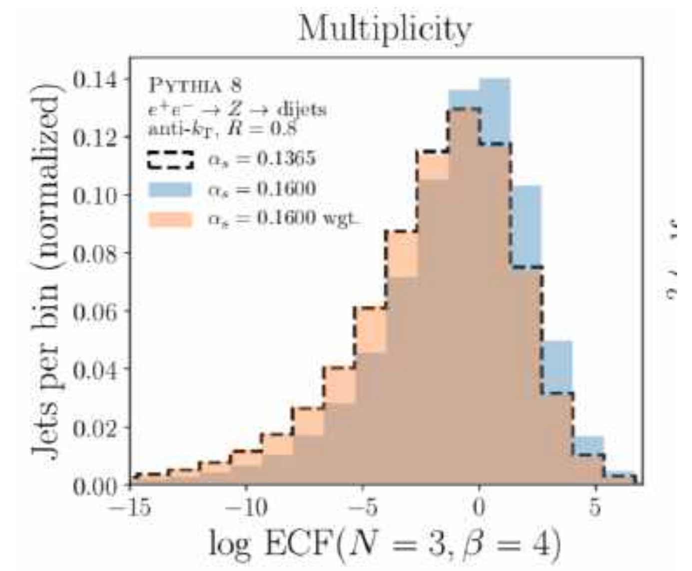
- includes simulation parameters θ in truth data



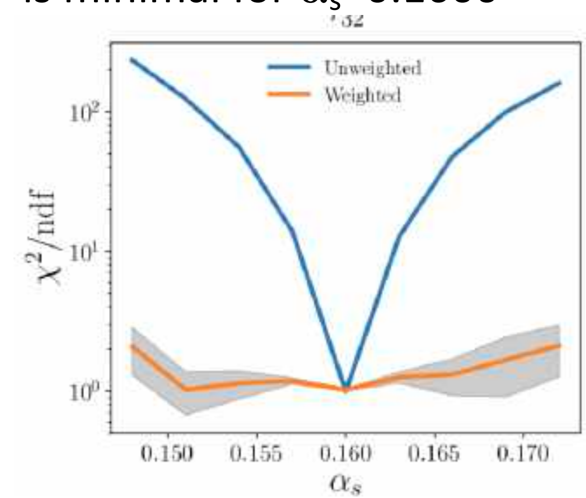
- learns relative weight $P(x, \theta)/P(x,\theta_0)$

- Could be a very efficient way to tune simulations
- or to reweight simulations to data

Reweights α_s=0.1365 distribution back to α_s=0.1600



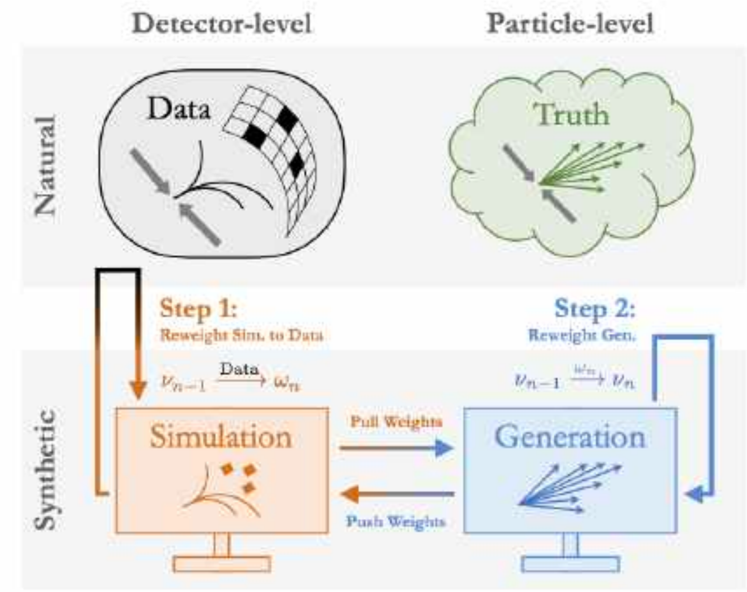
χ^2 is minimal for α_s=0.1600

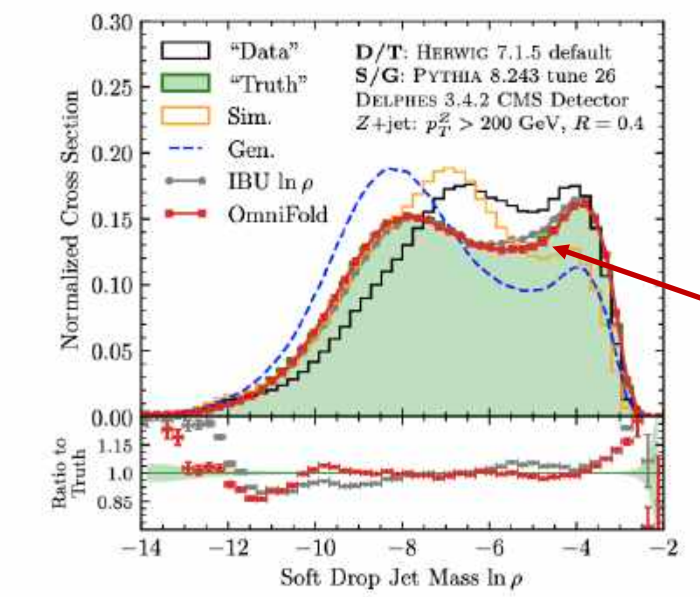


OmniFold: Unfold events

Andreassen et al. (arXiv:1911.09107)

- Uses ML to learn mapping from generator to detector
- Can then unfold any distribution to truth level
- Previous unfolding techniques are observable-by-observable



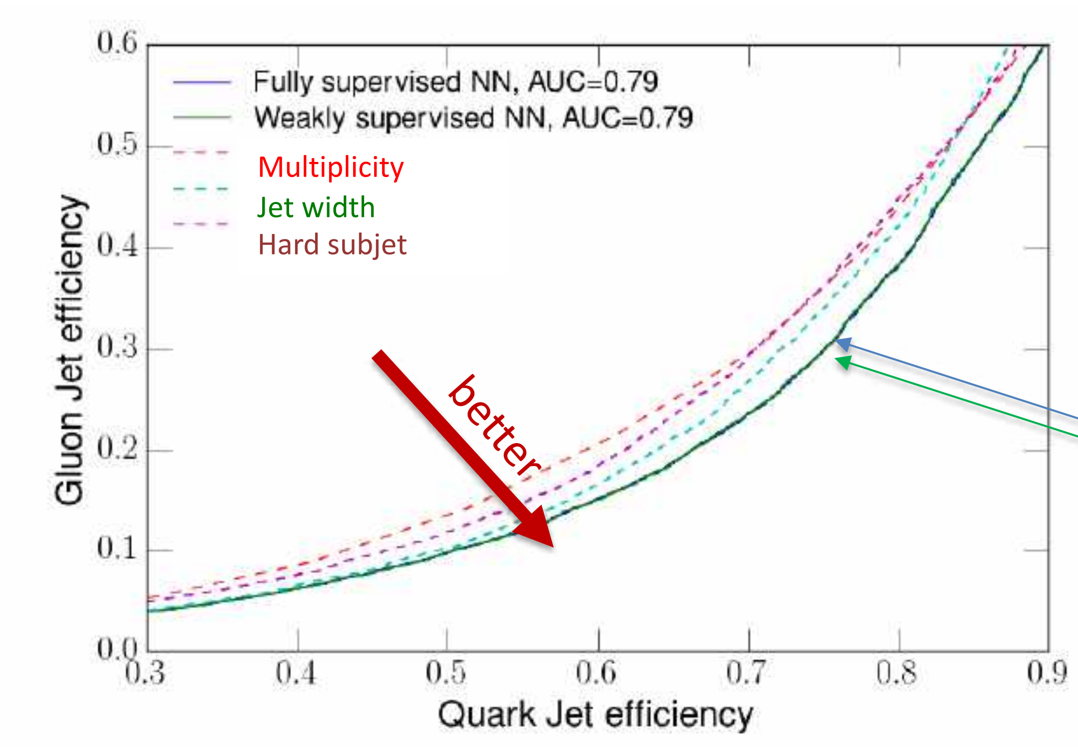


- Examples given used
  - Herwig as "truth" + delphes -> "data"
  - pythia as "sim" + delphes -> "gen"
  - Truth and omnifold agree
- Trying out on actual data is work in progess

Dery at al. (arXiv:1702.00414)

Weak supervision

Supervised learning: pure samples of quark and gluons used
Weak learning: mixed samples of quarks or gluons used
Unsupervised learning: no labels at all, just find patterns

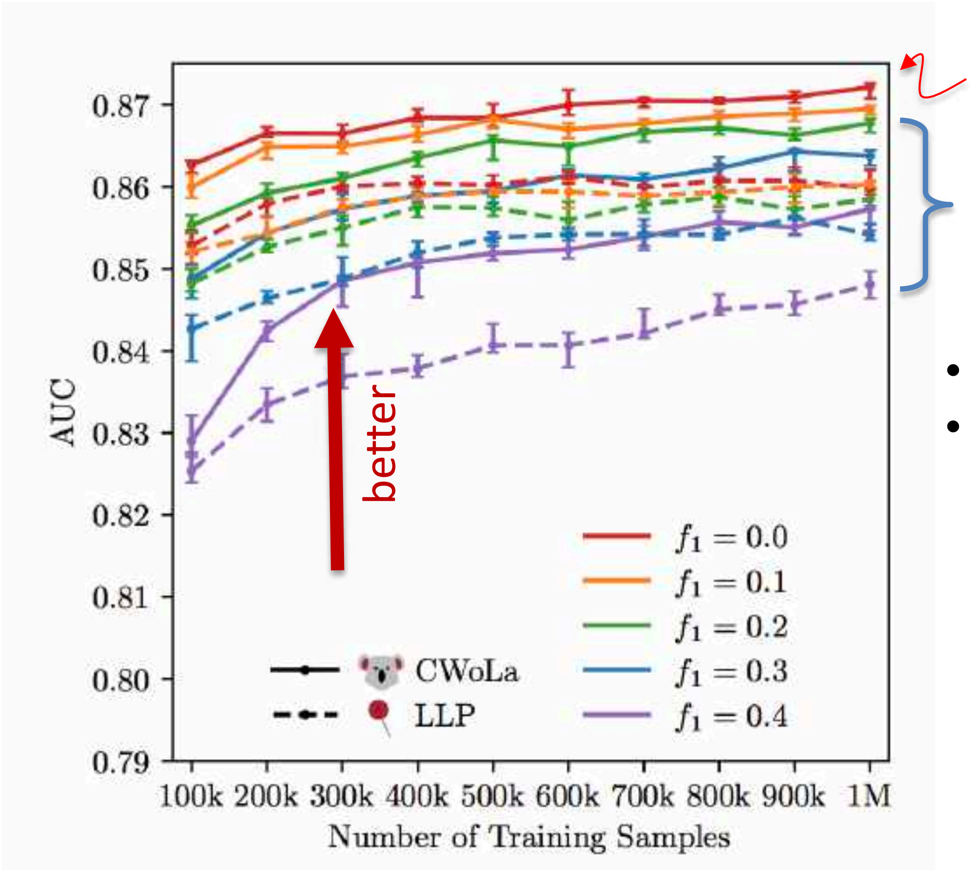


- **Weak learning** works with three Q/G discriminants

Weakly supervised works as well as fully supervised

Jet images + weak supervision

MDS, Komiske, Metodiev, Nachman (arXiv:1801.10158)



- Works as well as with full supervision
- Labels not needed even for complex inputs

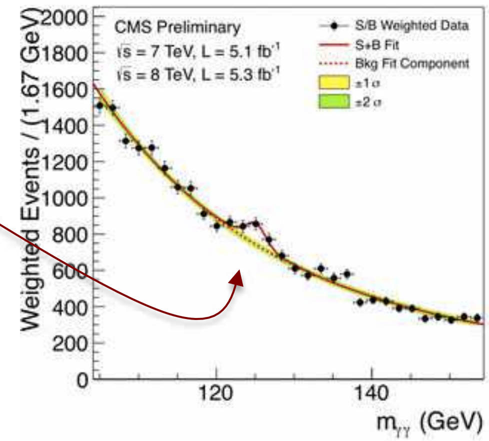
Weak supervision is a **breakthrough** for particle physics:
Can learn complex discrimination directly from data

DECORRELATION

Adversarial networks

Louppe, Kagan, Cranmer (arXiv:1611.01046)
Shimmin et al. (arXiv:1703.03507)

Want to find bump



- Train network on signal and background
- Can sculpt background to look like signal

No longer trust bump

Discovery significance reduced

Adversarial network

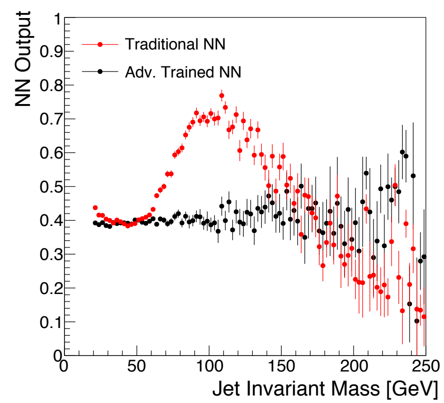
Network 1

Network 2

Tries to separate signal from background

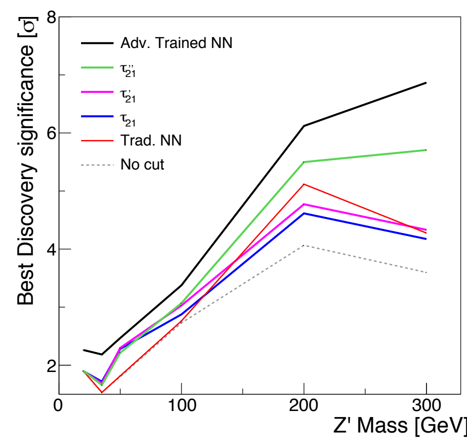
Tries to guess mass from network 1 output

Penalty in loss function



Decorrlates NN output from mass

Discovery Significance Improved!

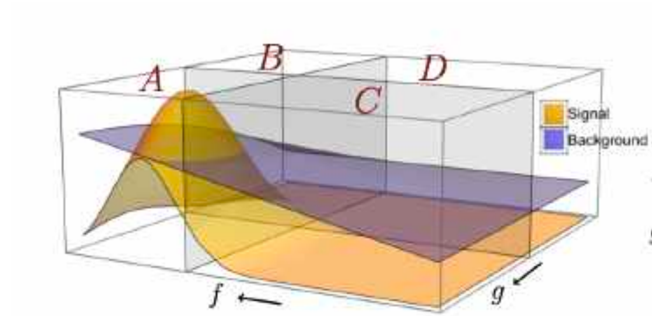


ABCDisCo: ML the ABCD method

Kaciecka, Nachman, MDS, Shih (to appear..)

ABCD method:

- Standard experimental sideband technique
- Estimate background in region A via $N_A = \frac{N_B N_C}{N_D}$
- Requires two features f and g to be uncorrelated
  - E.g. f = mass and g = rapidity

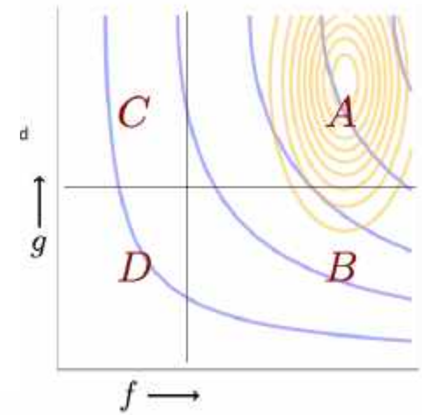


- Single DisCo

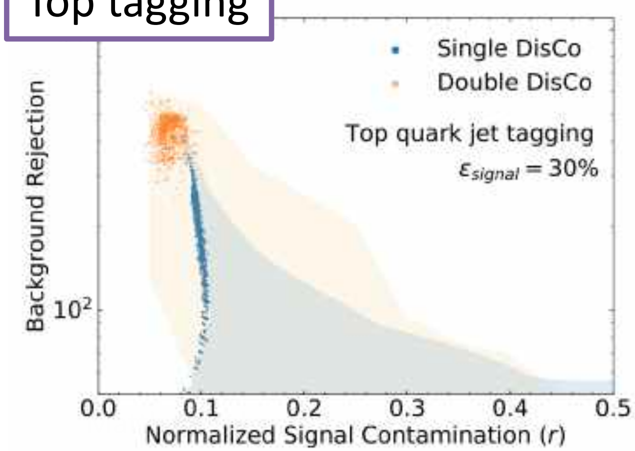
f is fixed (e.g. mass) g is learned

- Double DisCo

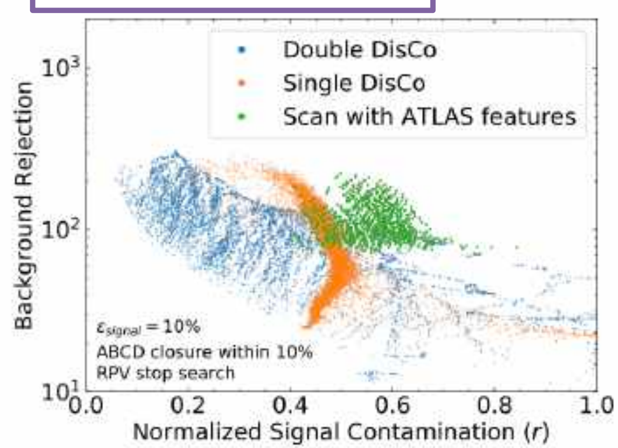
f and g are learned



Top tagging



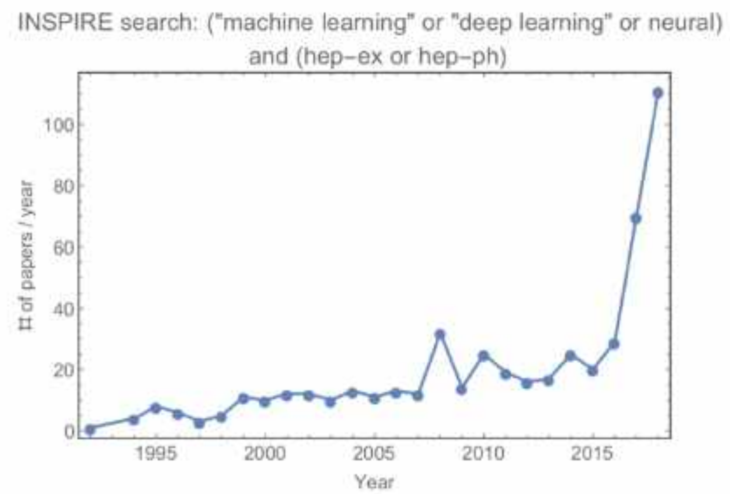
RPV squark search



Works great!

Conclusions

- Modern machine learning is growing rapidly
- "Traditional" collider physics is dead



- Much progress on standard nails
  - top-tagging
  - quark/gluon
  - anomaly detection

- State-of-the-art methodology
  - New observables (ML top tagging…)
  - New data representations (JUNIPR, point clouds …)
  - Improving experimental analyses (Omnifold, ABCDisCo..)
- See https://iml-wg.github.io/HEPML-LivingReview for a comprehensive list of papers

- Past: apply hammers to nails

- Future: learn some new physics