CoDec (**Co**ntrastive **Dec**orrelation)

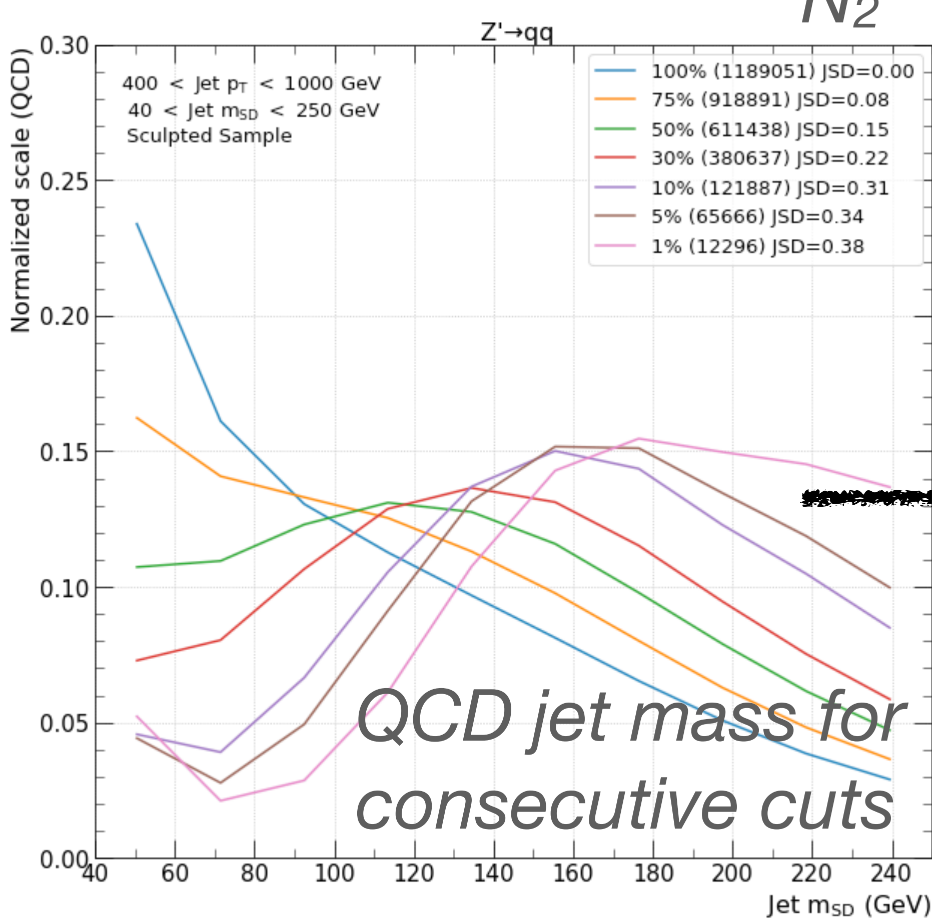
BOOST 2022
August 17, Tagger Session

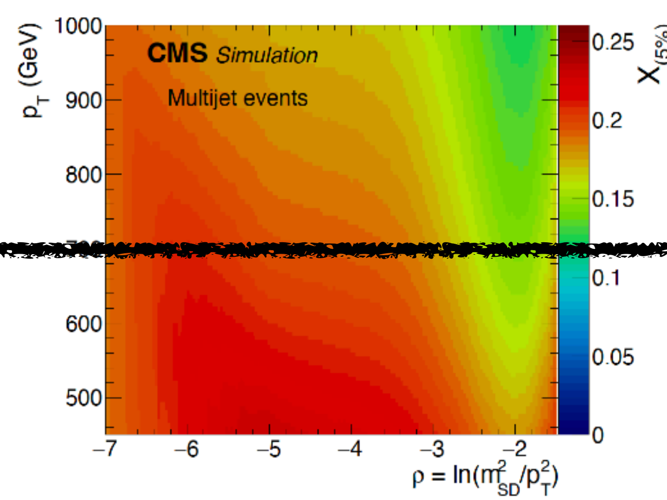
Jeffrey Krupa, Eric Moreno, Phil Harris, Keiran Lewellen, Yihan Liu, Sang Eon Park, Dylan Sheldon Rankin, David Yu

Tagger correlations

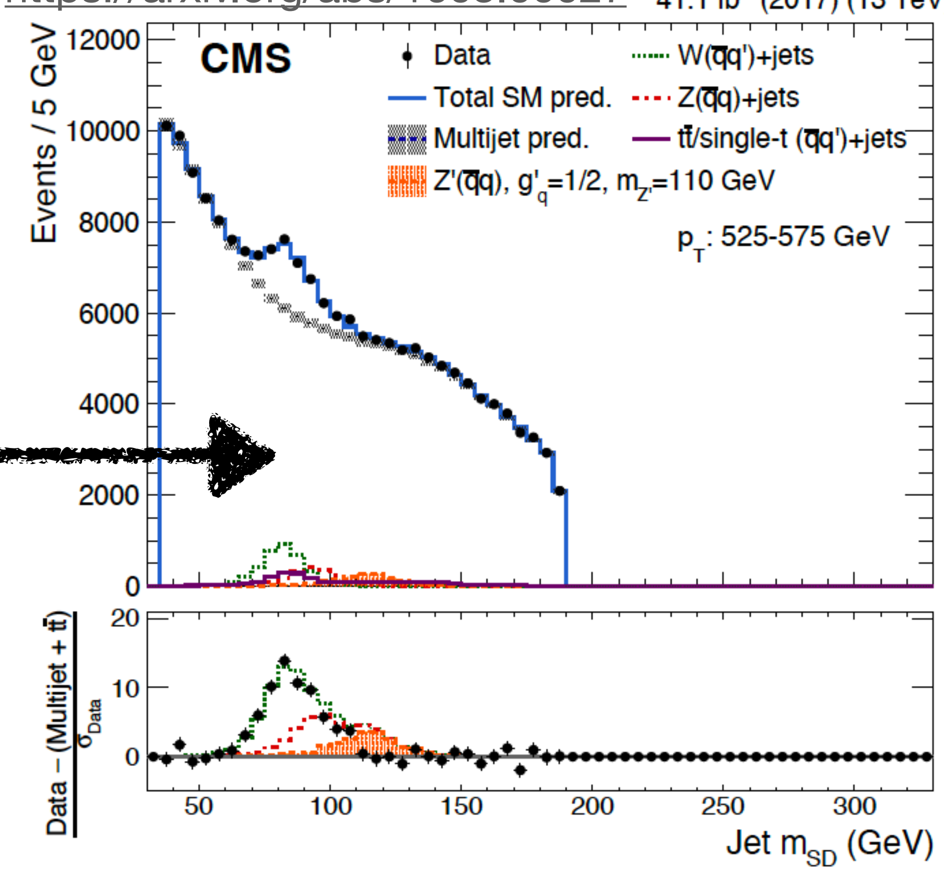
- We use taggers and regressions for low mass boosted dijet searches
- Rely on smoothly falling backgrounds to estimate QCD passing N_2
    - We need to cut tight
    - DDT method: define new variable N_2^{DDT} for which passing and failing regions have the QCD jet mass shape

N_2



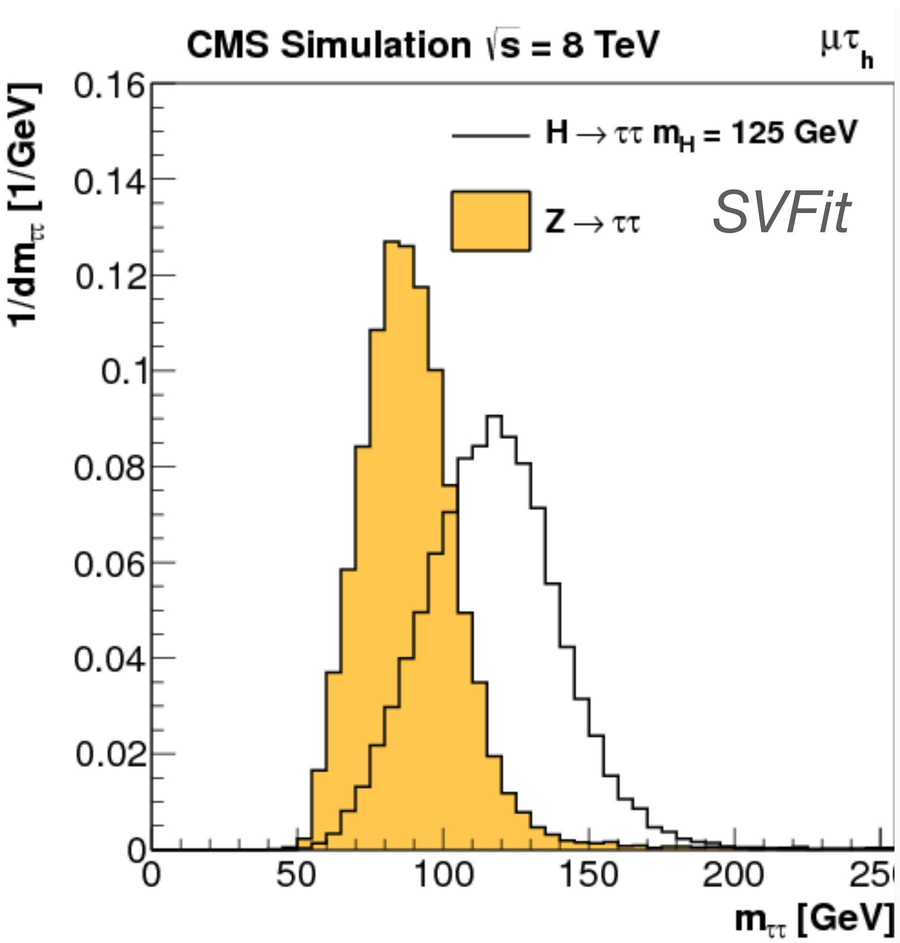


https://arxiv.org/abs/1603.00027

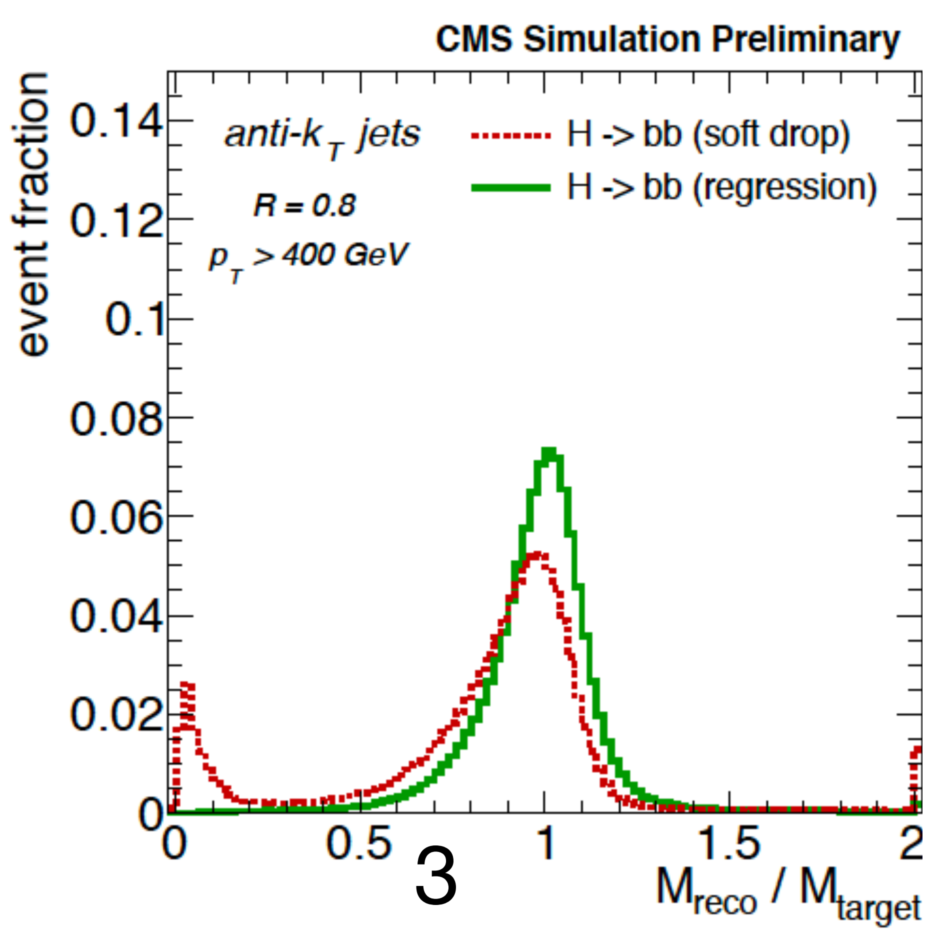


Another application: mass reconstruction

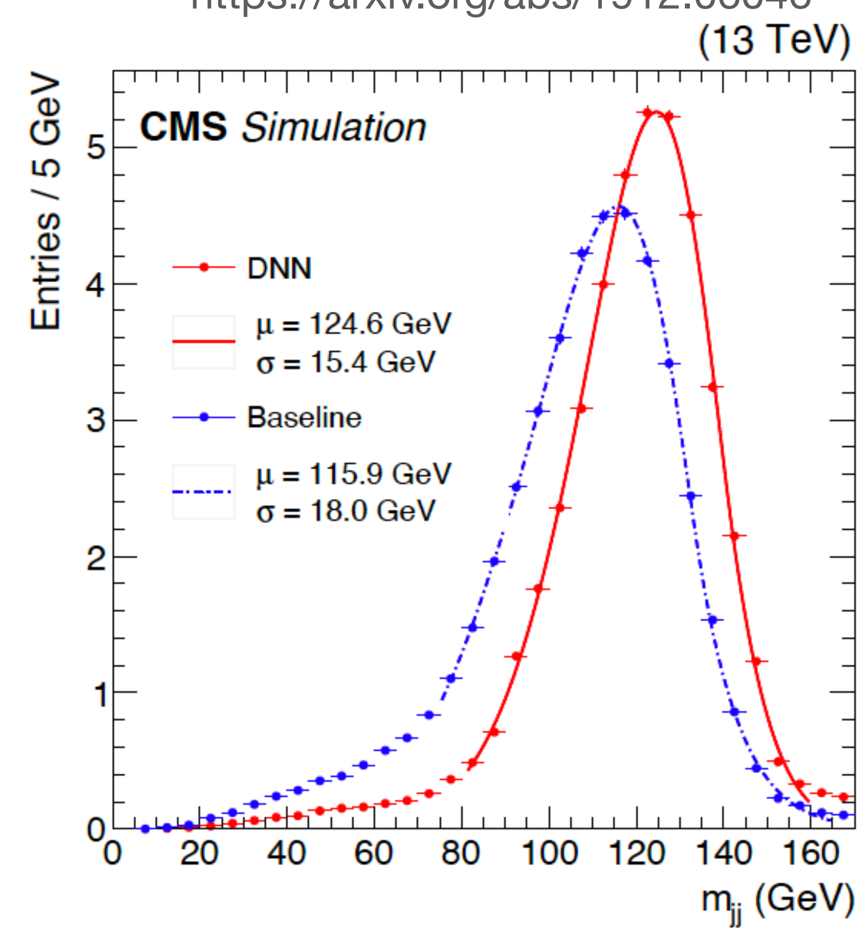
- Sensitivity can be improved by fitting regressed mass
  - Recovers energy from e.g. neutrinos, jet grooming
- Peakless Z' helps us to be **sensitive across large phase space**



https://cds.cern.ch/record/2777006?ln=en

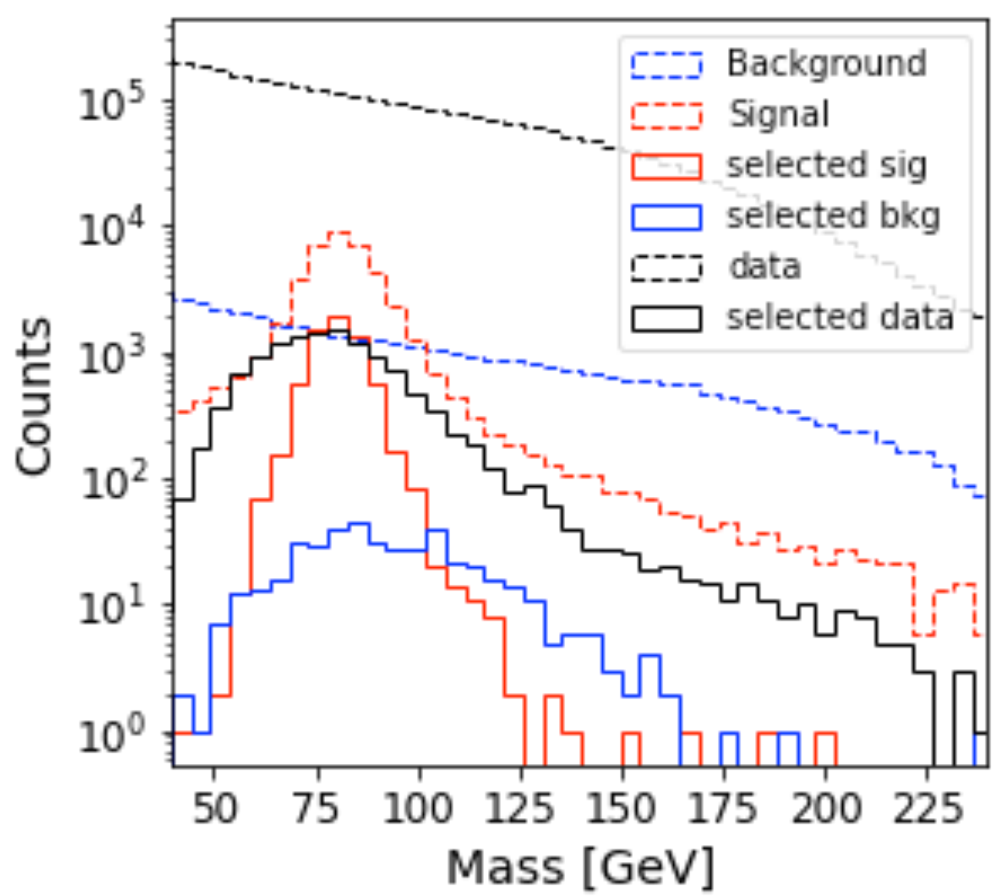


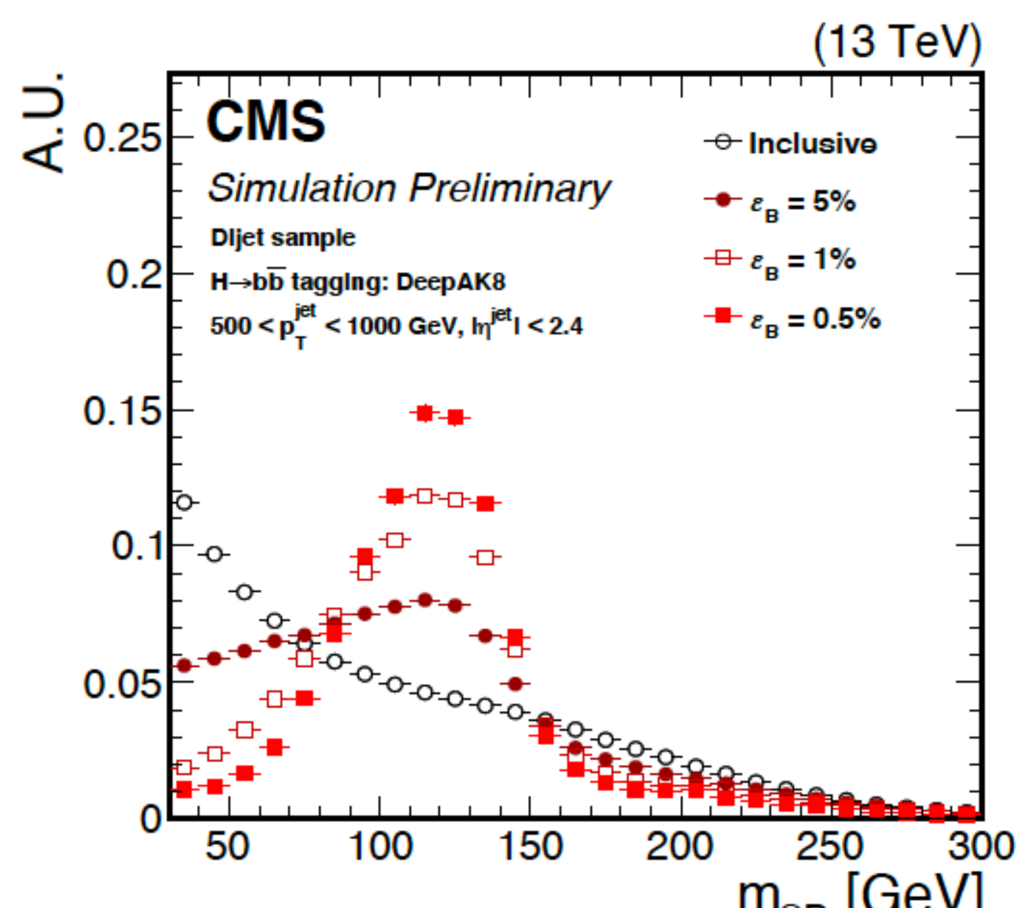
https://arxiv.org/abs/1912.06046



ML

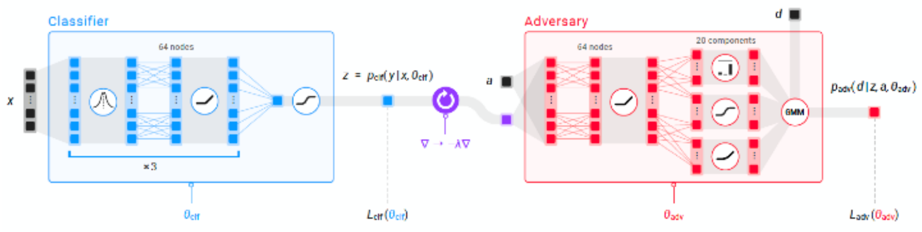
- Correlations grow stronger with ML
  - Mass is quickly learned!
  - Generic problem





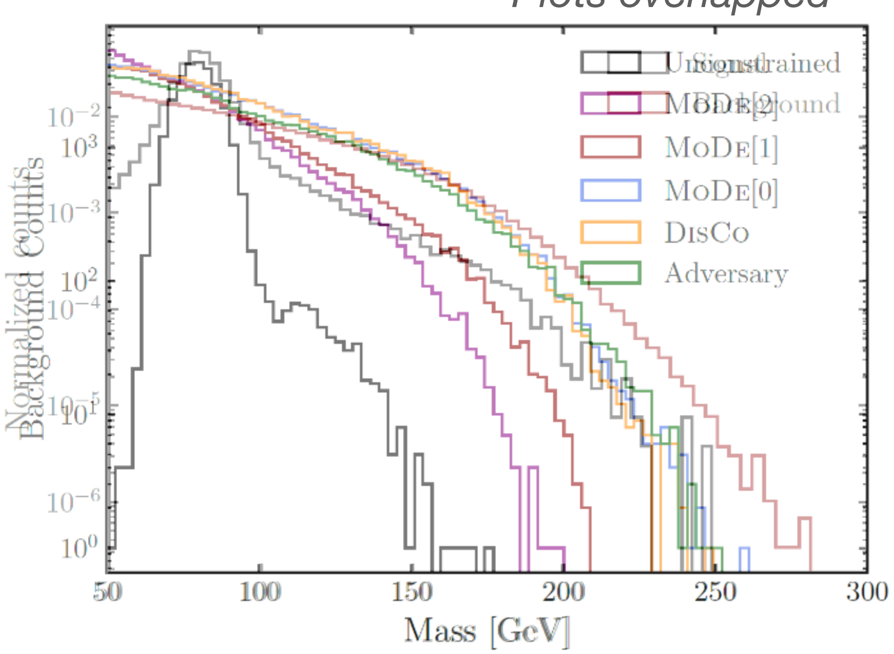
https://cds.cern.ch/record/2707946/files/DP2020_002.pdf
https://github.com/violatingcp/ContraDecorr/blob/main/Jets_v2.ipynb

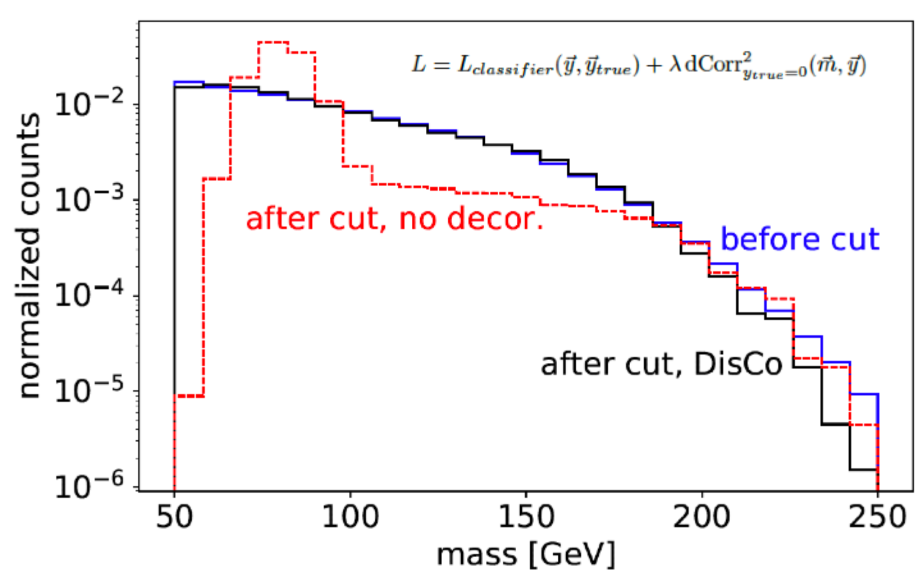
The landscape

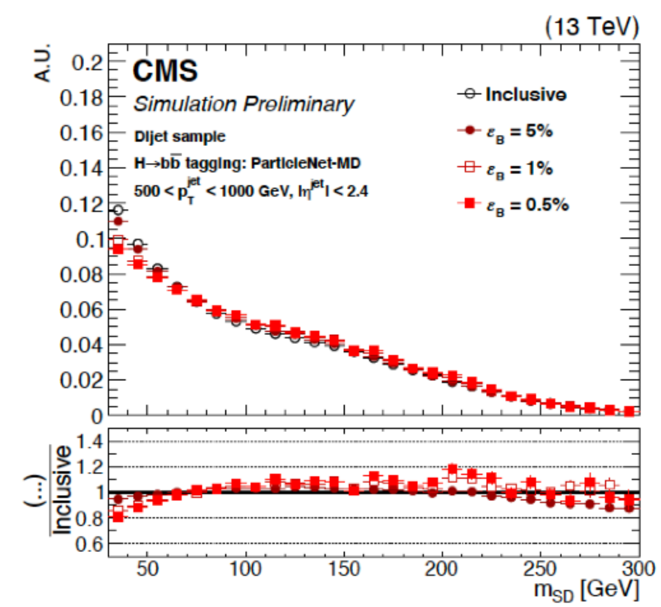


- Typically use DDT method
- Many approaches have been applied to mitigate this, e.g. DisCo, MoDe, KL-divergence, multi-mass-point training sample (CMS particleNetMD), adversaries, …
  - Many emphasize architectures/losses
  - We take a more “old school” approach

Plots overlapped







https://arxiv.org/abs/1603.00027
https://cds.cern.ch/record/2707946/files/DP2020_002.pdf
https://arxiv.org/abs/2010.09745
https://cds.cern.ch/record/2630973/files/ATL-PHYS-PUB-2018-014.pdf

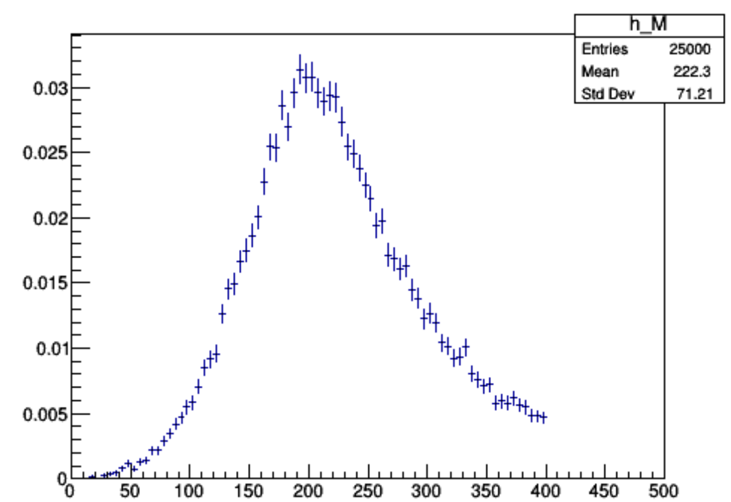
Flat mass sample generation

- Idea: use [Madgraph LO bias weighting] to generate a training sample with a "flat" mass profile, but otherwise identical to desired signal

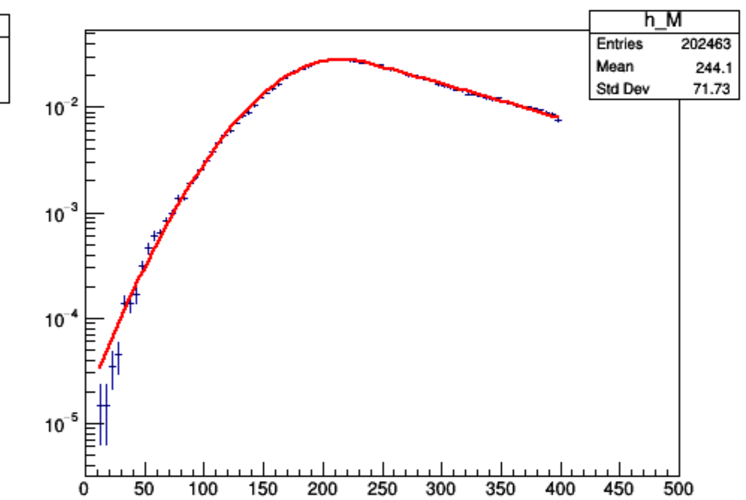
Method

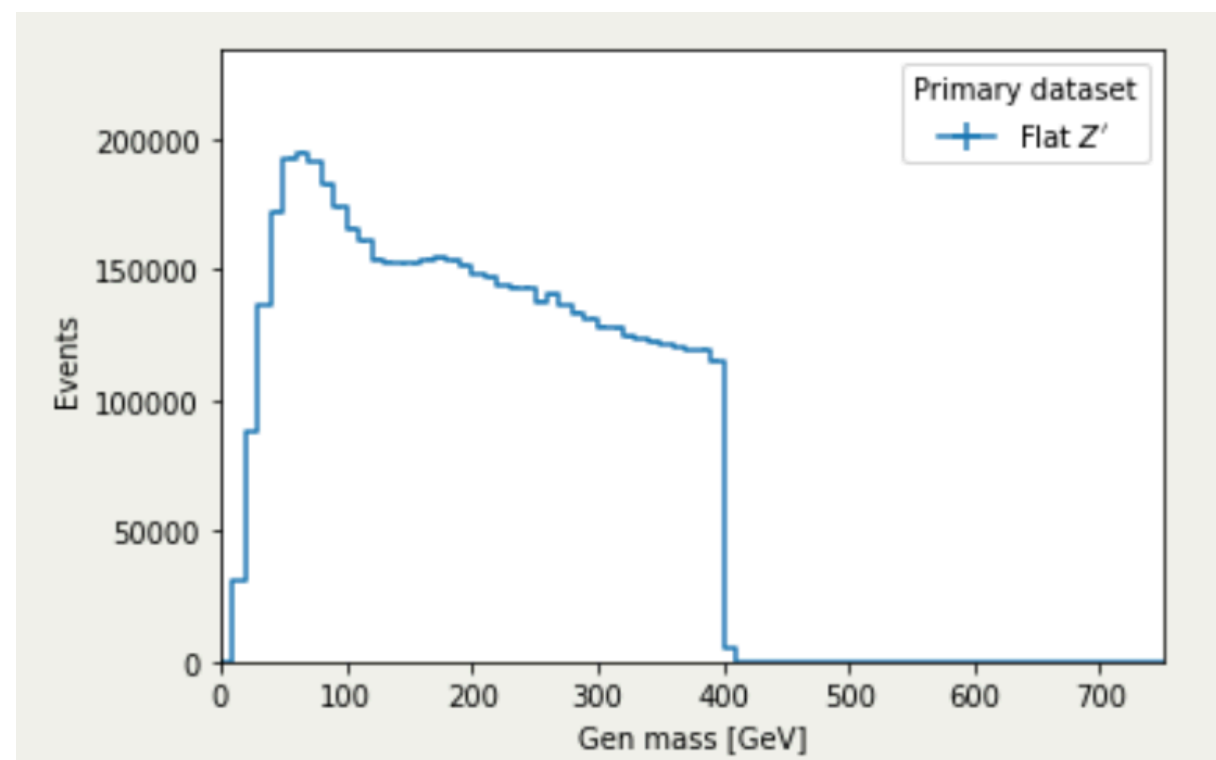
- Start from $Z'(qq)+\gamma$ sample w/large width
  - $m \sim 175$ GeV, $\Gamma \sim 100\,\%$
  - H_{T} cut for boosted events
- **Fit mass shape**: $f(m)$ (e.g. Crystal Ball)
- **Reweight** sample using Madgraph bias weighting: $w(m) = 1/f(m)$
  - Can also reweight p_{T} etc.

Initial sample w/large Γ



Fit mass distribution



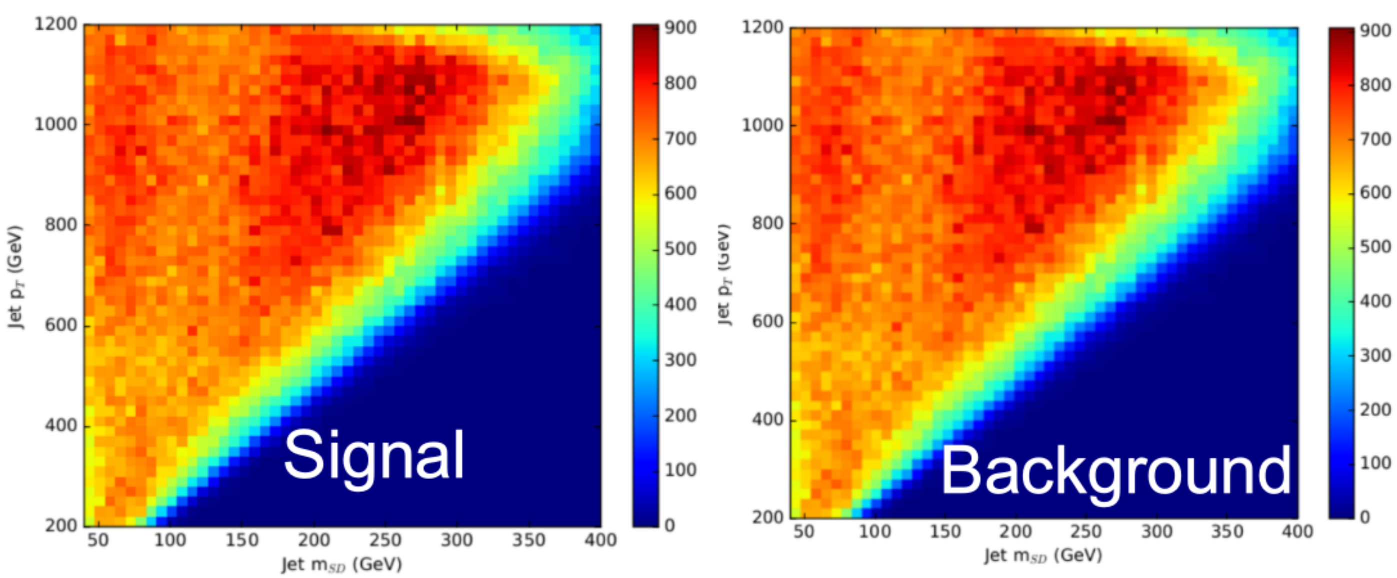


Flat sample

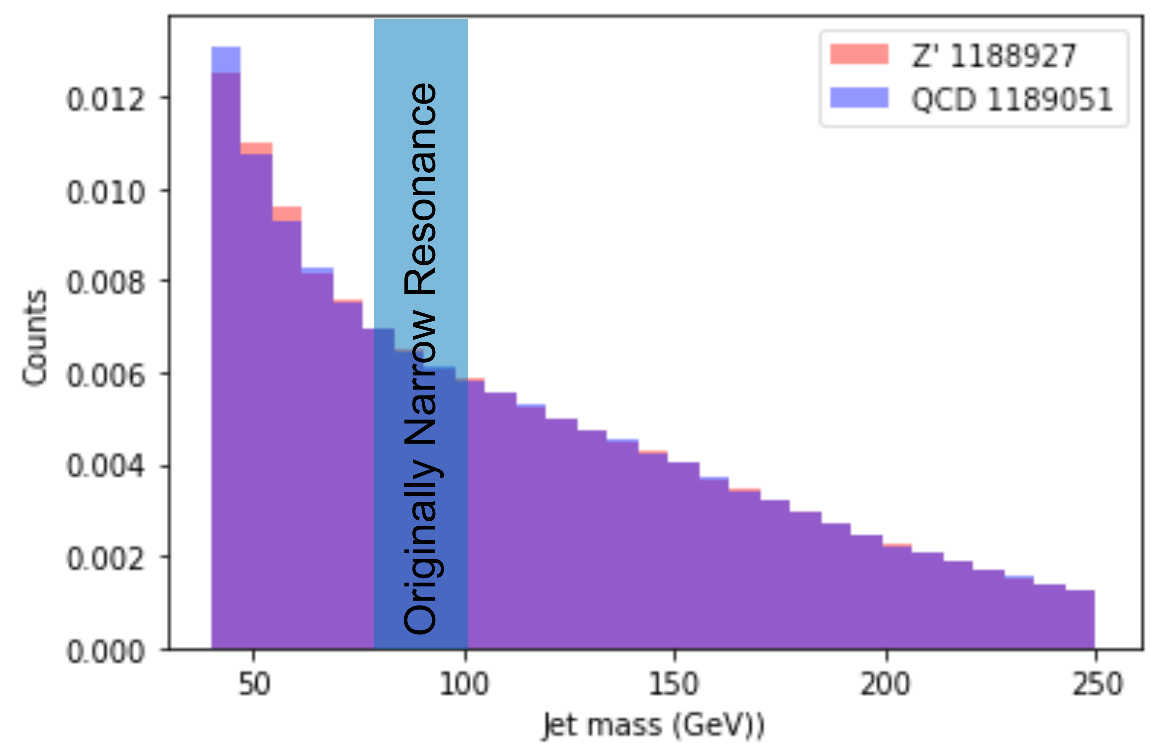
Using flat samples

- We flatten further by sampling events such that signal and QCD match
  - Residual differences are applied as weights
  - We find the 1D approach to be more robust

2D approach: flattening in (mass,pT)

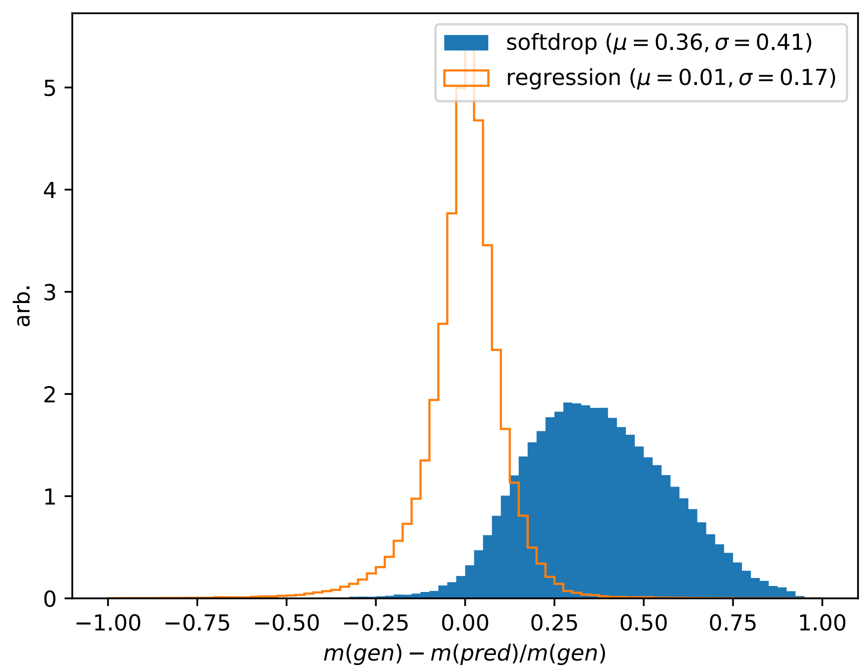


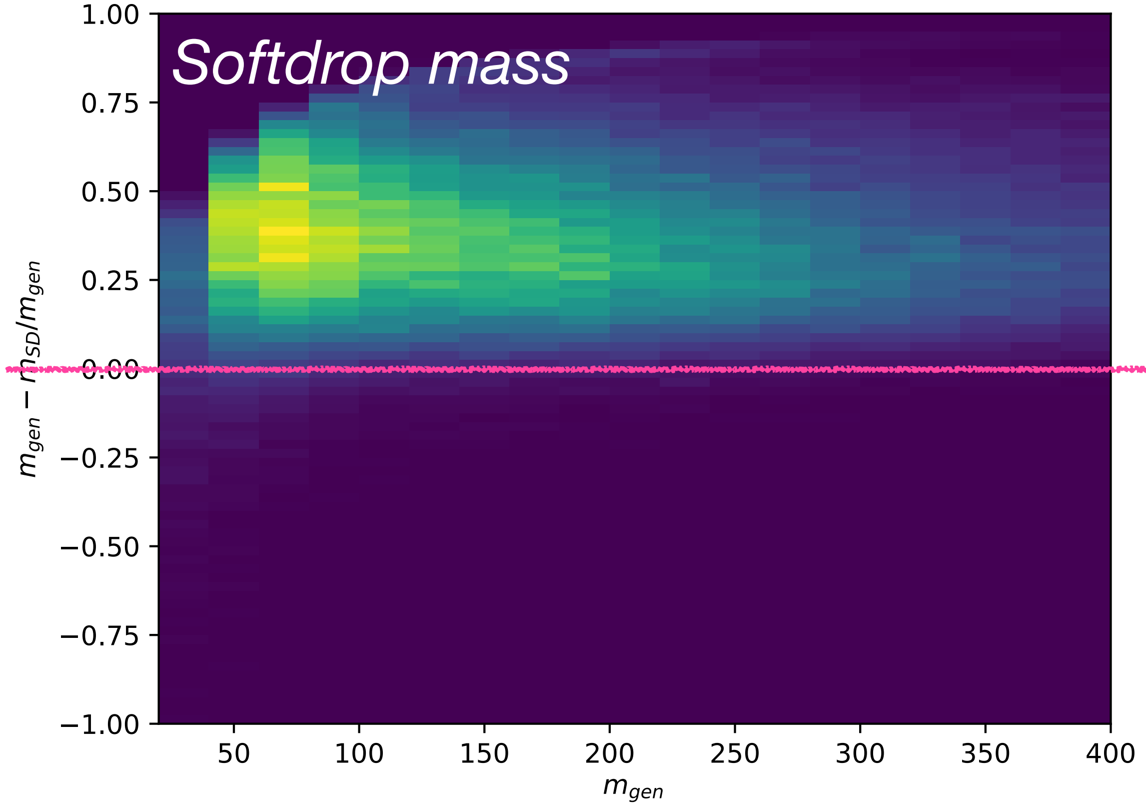
1D approach: flatten in mass only

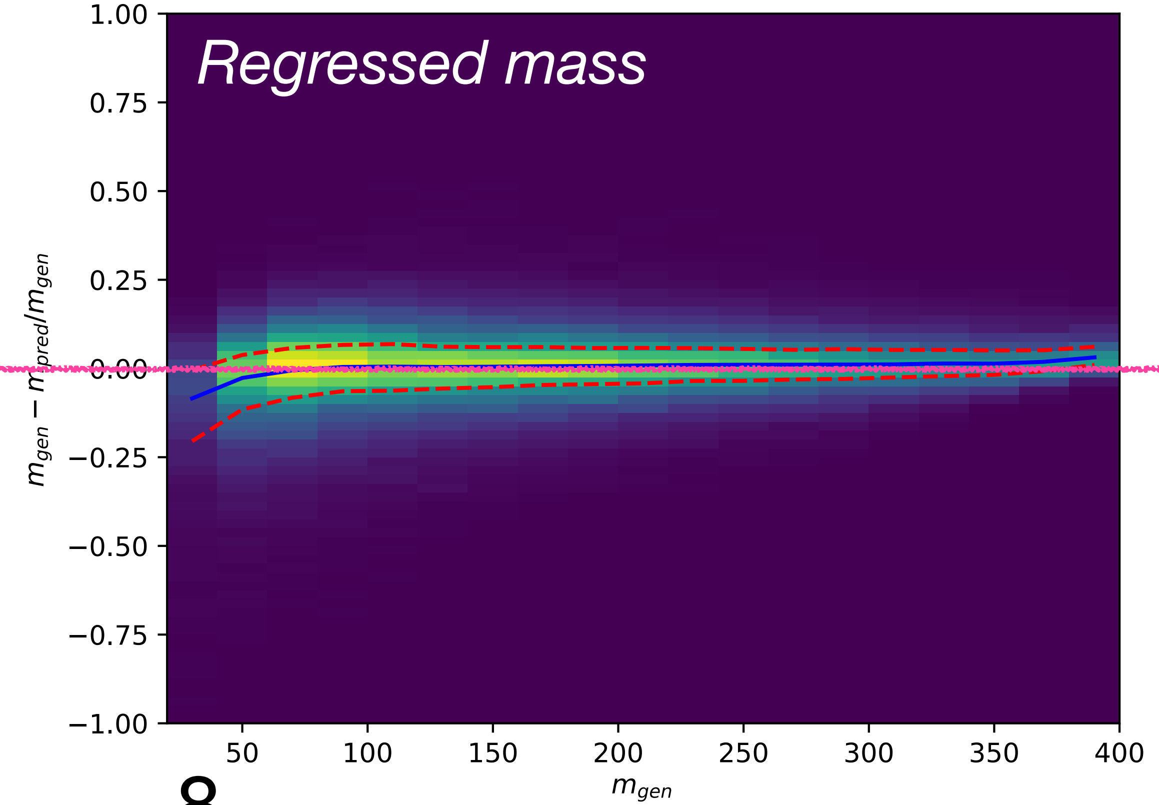


Application 1: regression

- Training network to predict true mass from particle constituents
  - **Recover losses** from grooming/invisible
- Flat samples gives large improvement vs. soft drop
  - Flat Z' sample helps us to be **sensitive across large mass range**



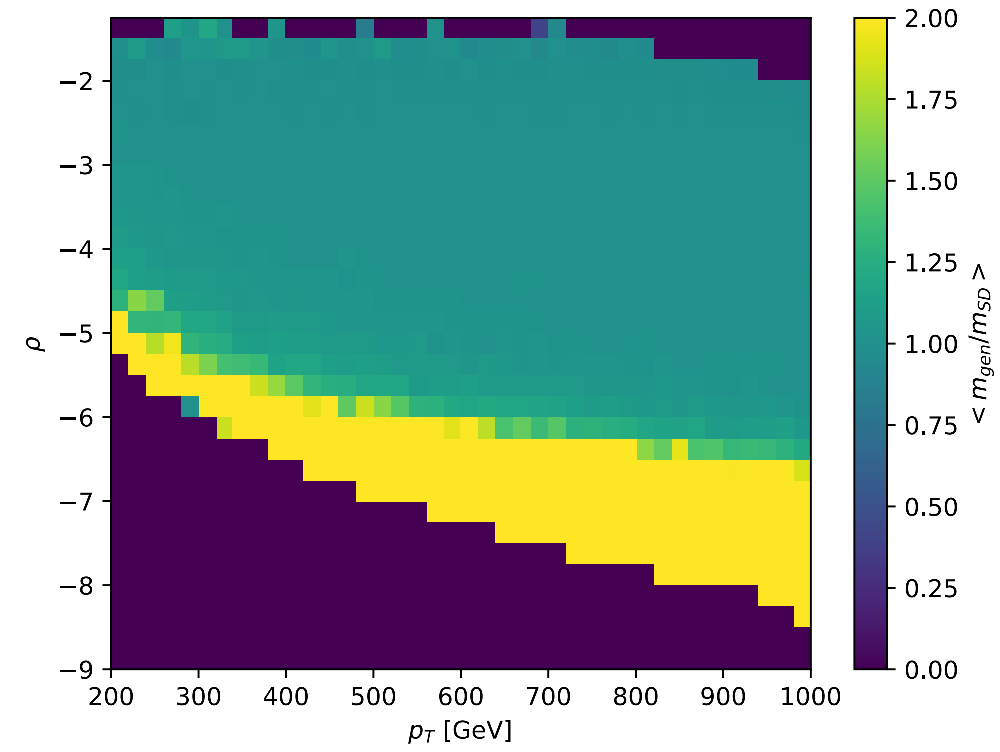


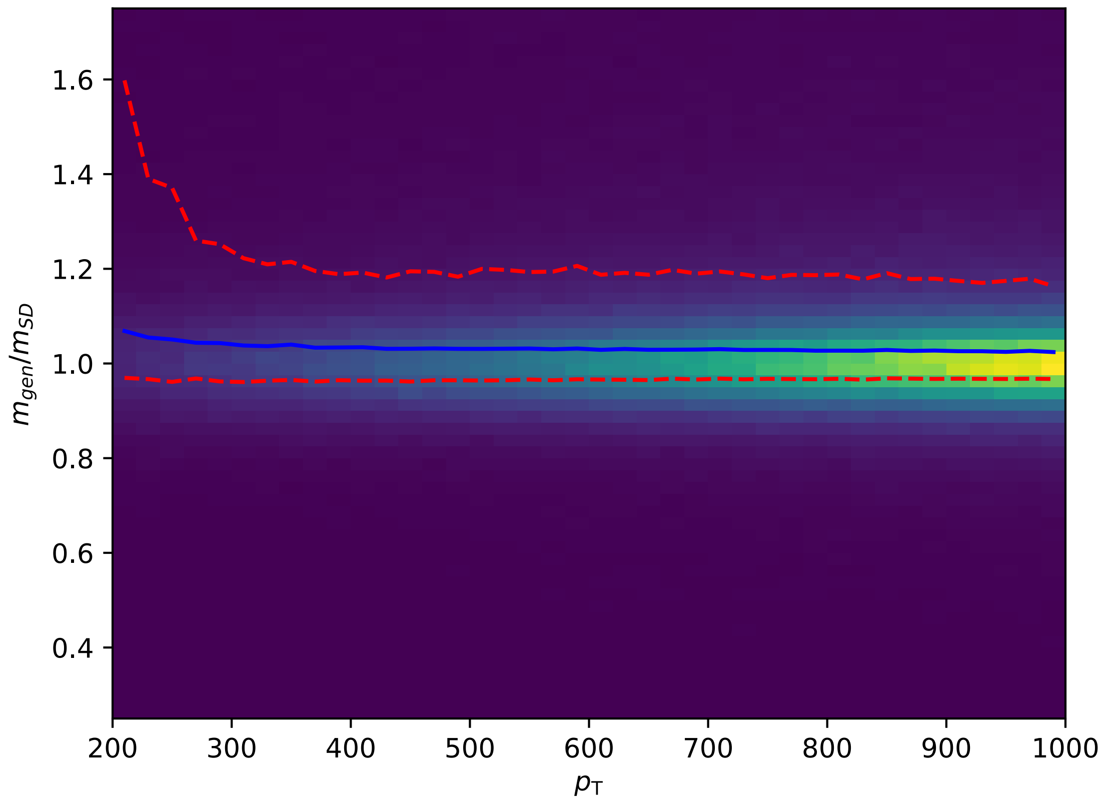


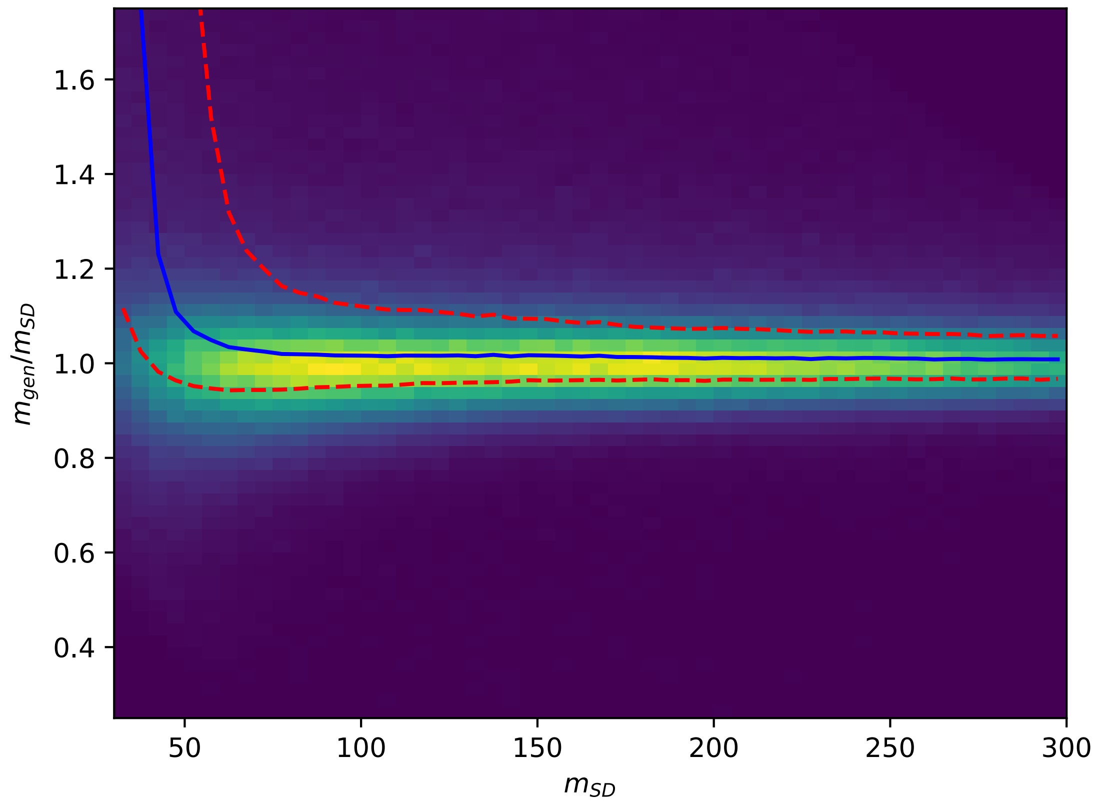
Better

Application 2: calibration

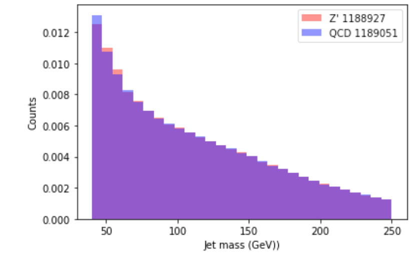
- In analysis we apply truth-level correction to the softdrop jet (mass,pT)
  - Limited statistics from resonance/ bumpy approaches
  - Flat sample has large statistics across entire mass range of interest



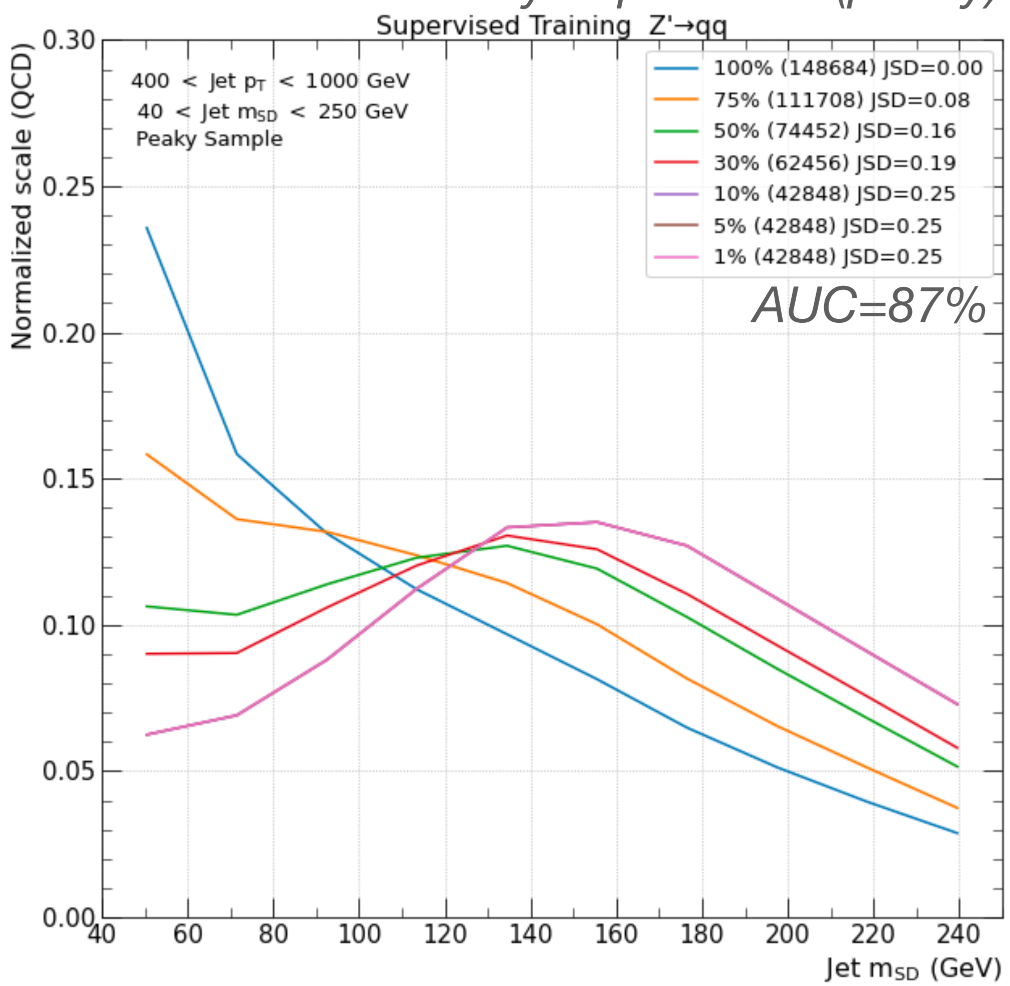




Application 3: tagging

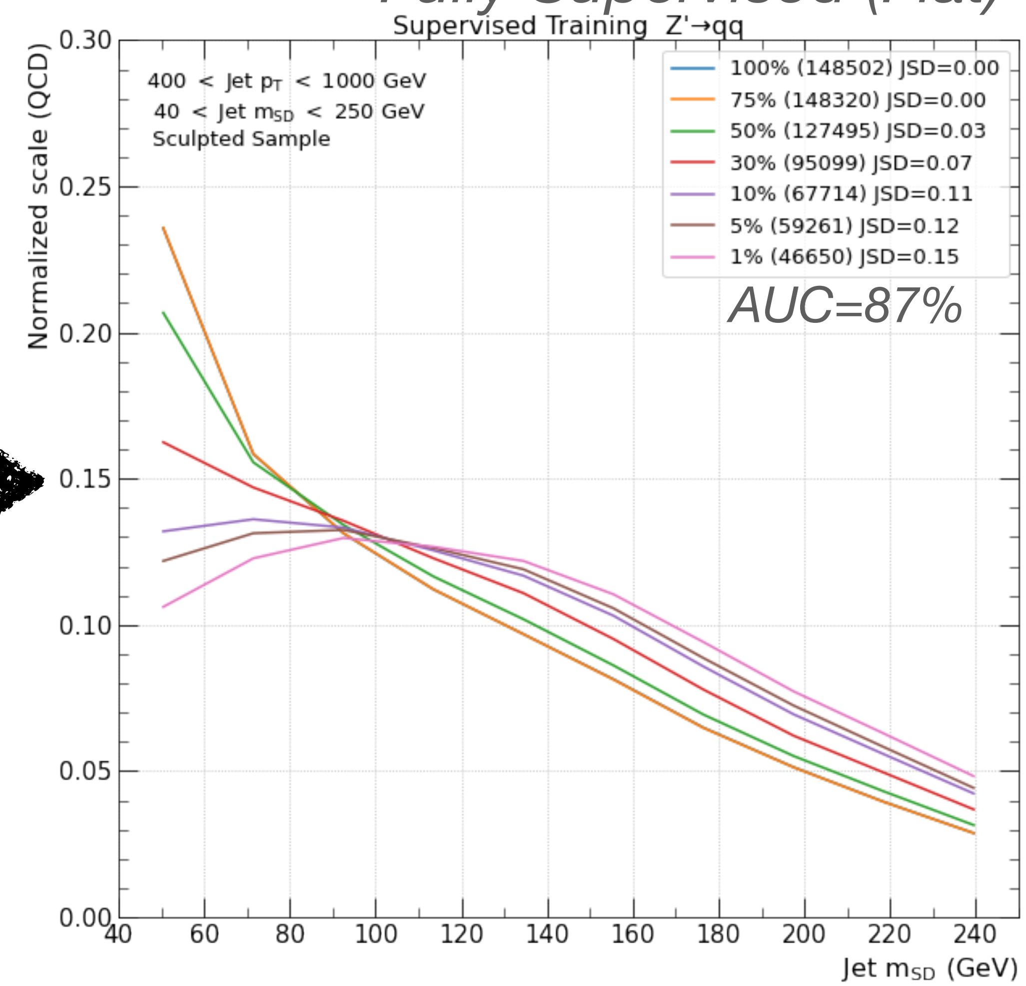


Fully supervised (peaky)



Improvement from flat

Fully Supervised (Flat)

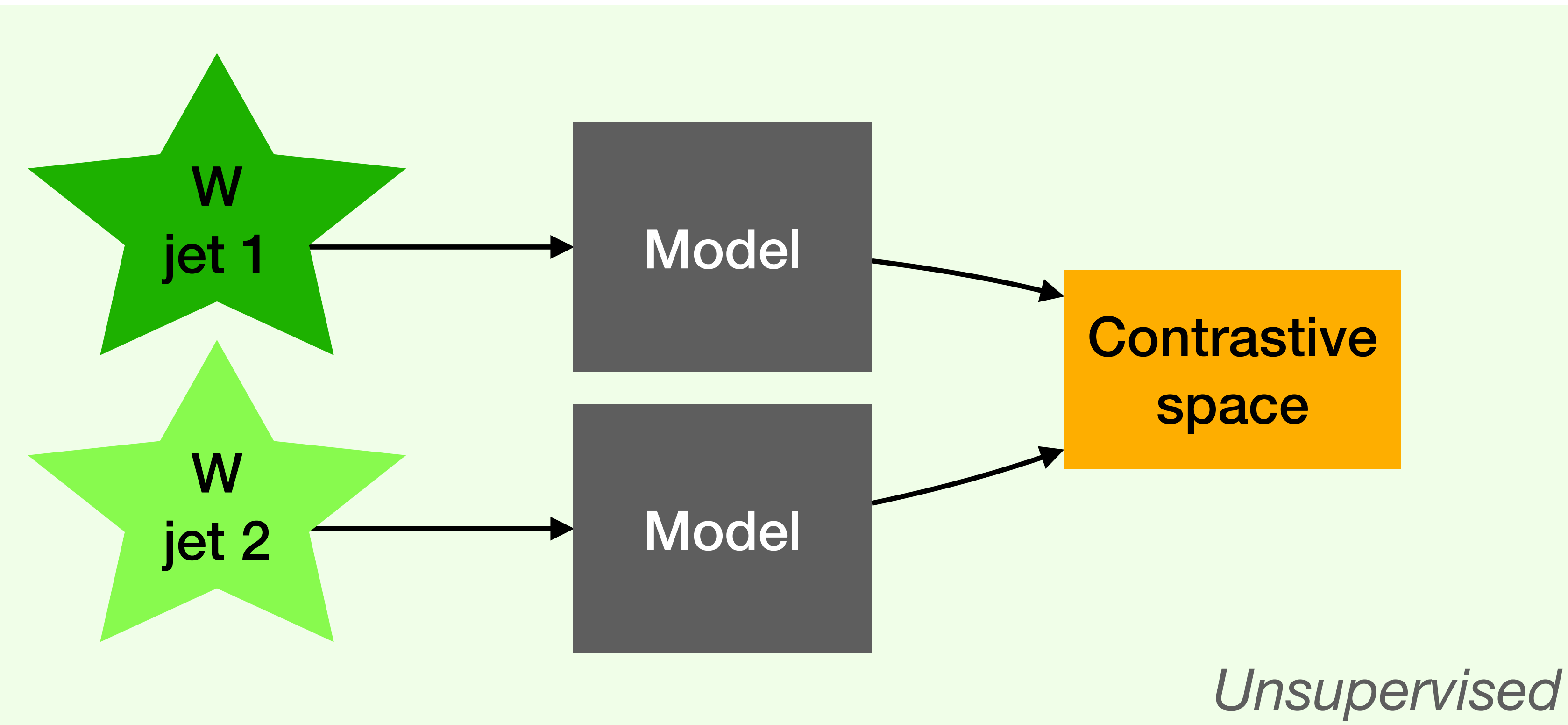


Residual correlations present

- <u>Also seen in N2</u>. Depends on flavour, pT, …

Can we go further with a different space?

The Contrastive Space



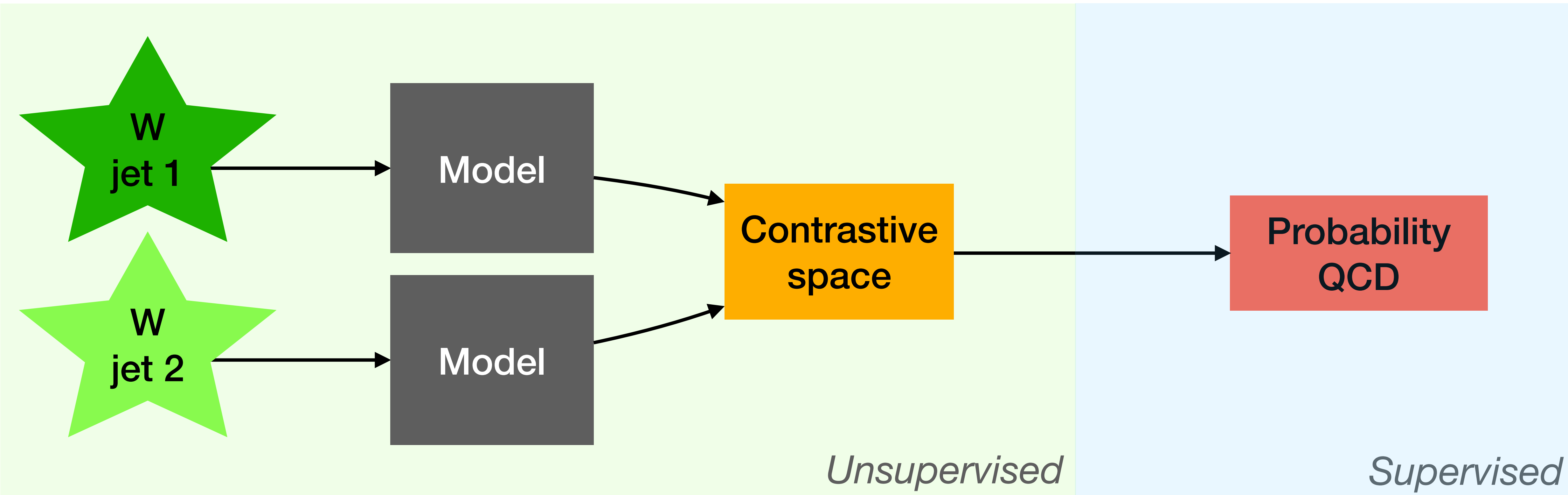
"Attractive" "Repulsive" "Decorrelation"

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]$$

Minimize the difference in representation between two distorted objects of the same origin

**Note: we don't explicitly give mass to the loss!*

The Contrastive Space



https://arxiv.org/abs/2103.03230

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised	25.4	56.4	48.4	80.4
PIRL	-	-	57.2	83.8
SimCLR	48.3	65.6	75.5	87.8
BYOL	53.2	68.8	78.4	89.0
SwAV	53.9	**70.2**	78.5	**89.9**
Barlow Twins (ours)	**55.0**	69.7	**79.2**	89.3

ImageNet tagging

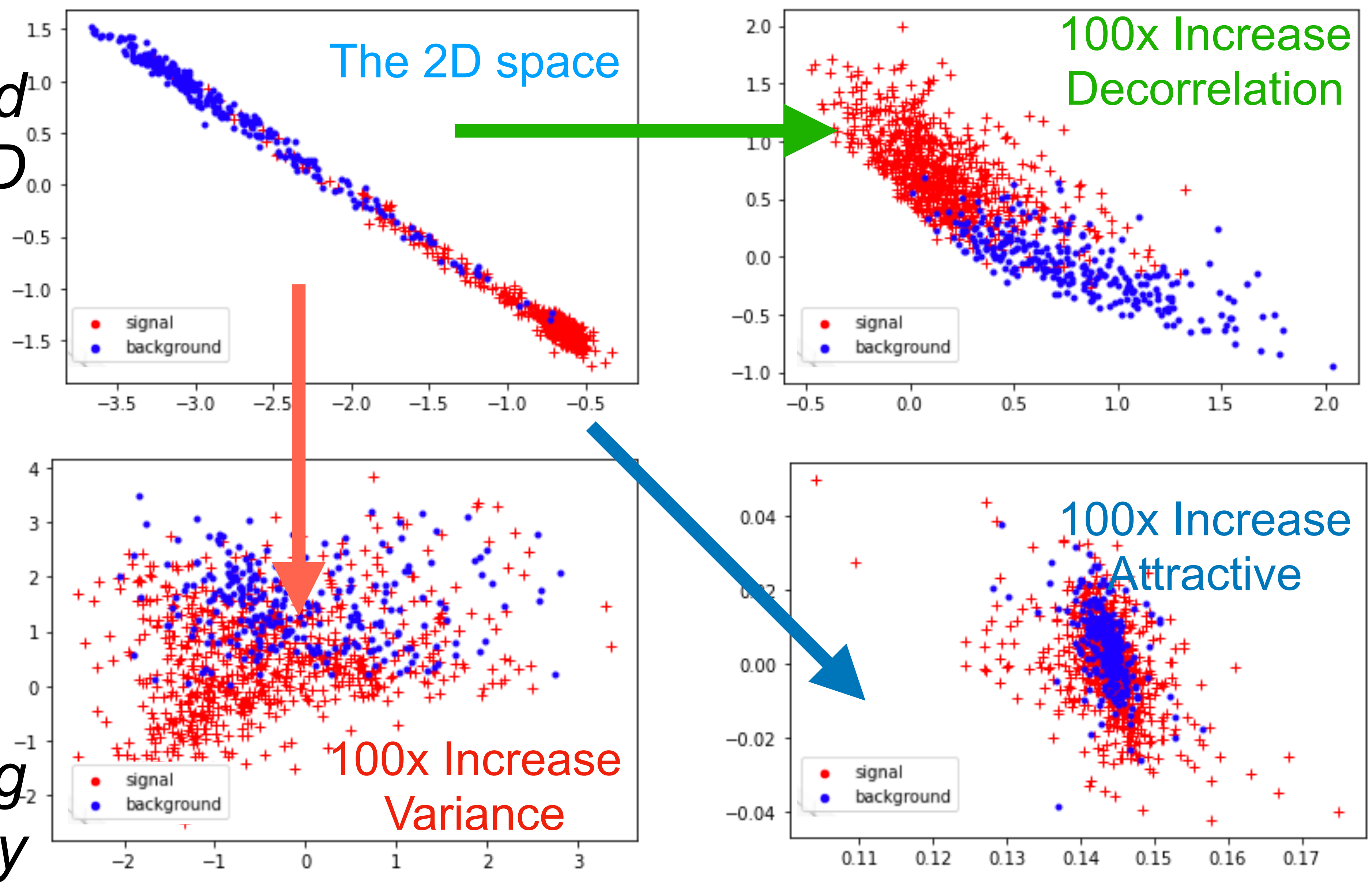
- This idea is the basis of contrastive learning
  - Notion of constructing a “self-supervised” space
- Contrastive learning is currently leading to top ML Perf Algos
- Most well known contrastive method is SimCLR
  - We focus on VICReg and BarlowTwins

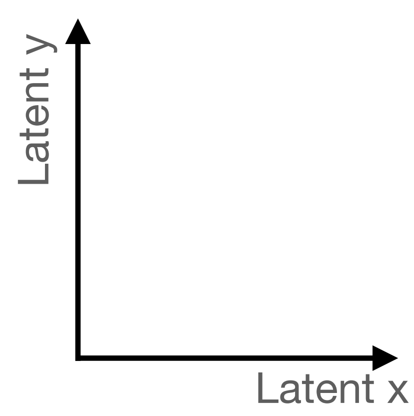
A Toy Dataset

MSE | Variance to 1 | Diagonalize Latent space

$$\ell(Z,Z') = \lambda s(Z,Z') + \mu[v(Z)+v(Z')] + \nu[c(Z)+c(Z')]$$

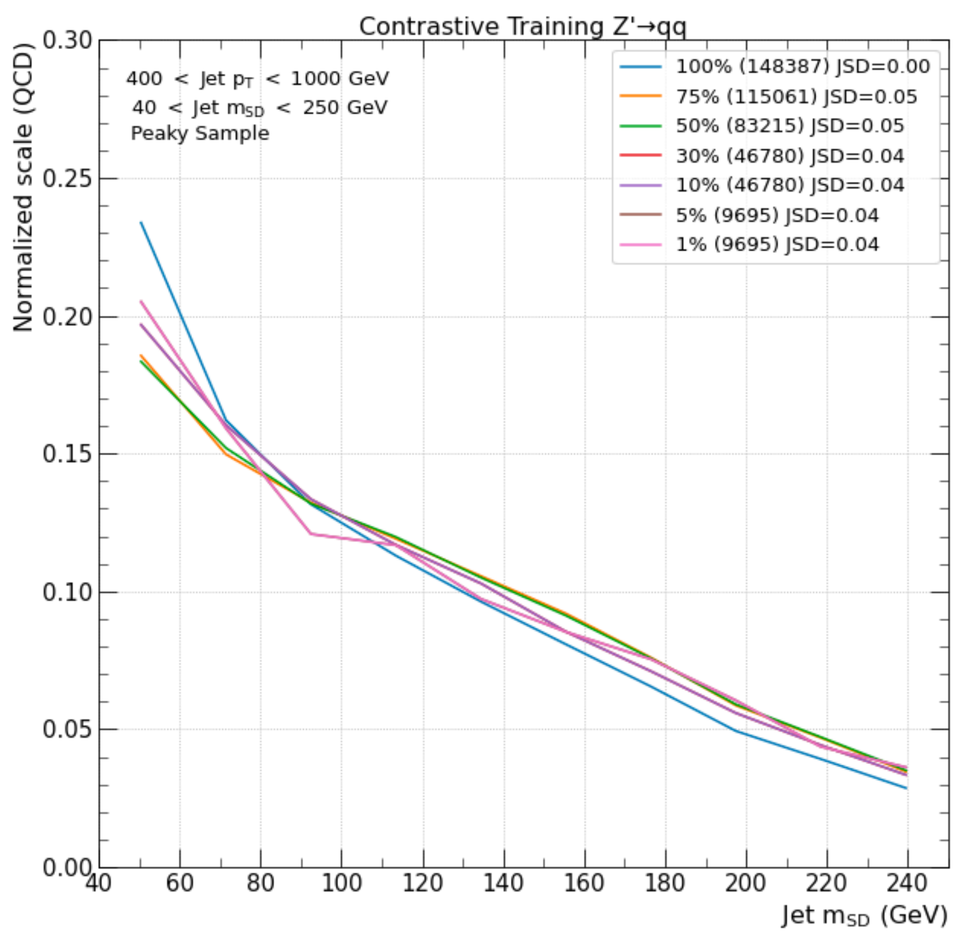
3D space embedded in 2D

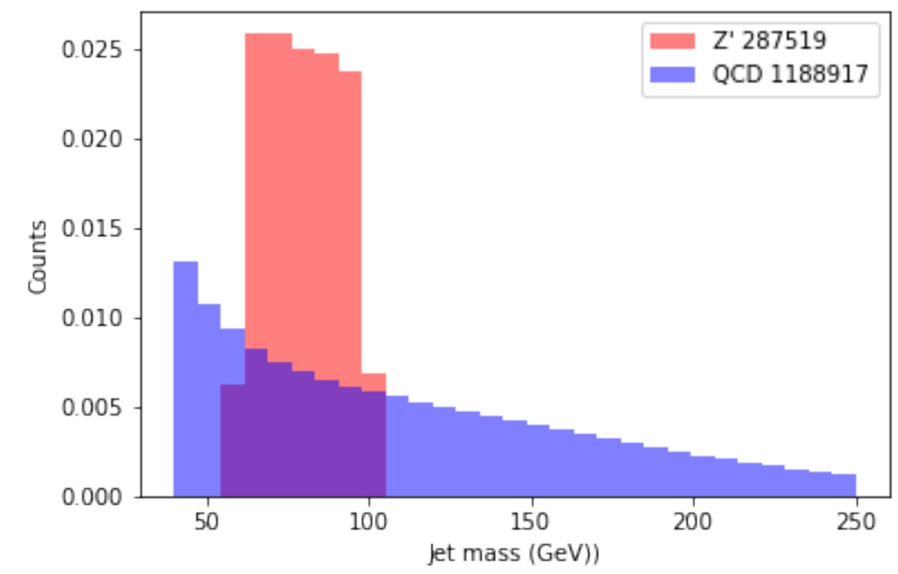




Naively run a training our space magically becomes separated!

Contrastive only

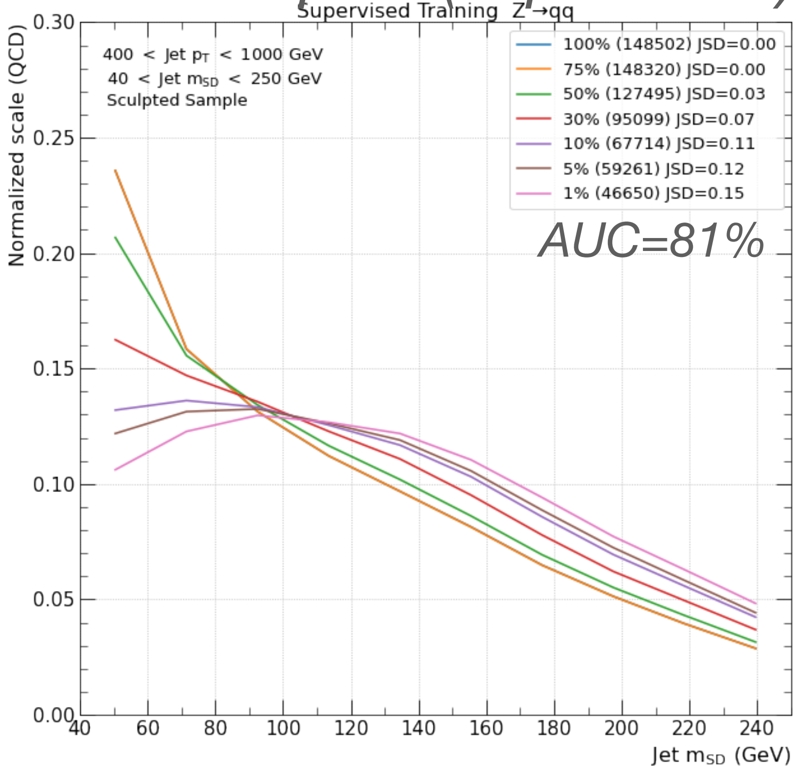


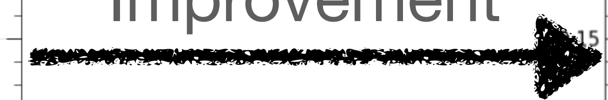


- Contrastive training on peaky sample already decorrelates nicely

Flat+Contrastive

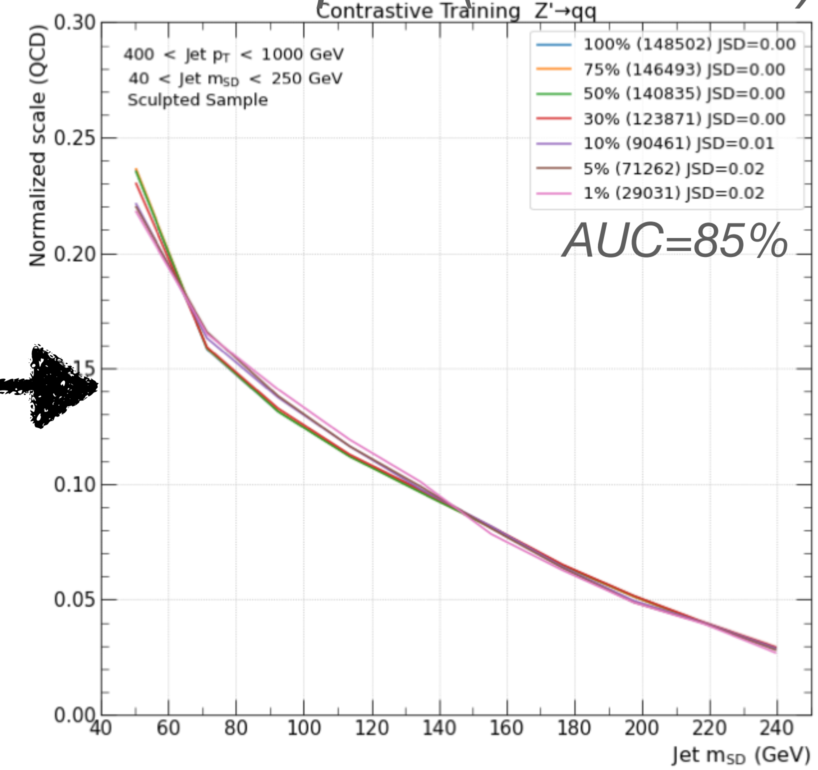
Flat samples (supervised)

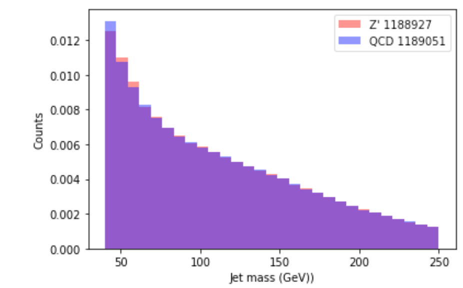


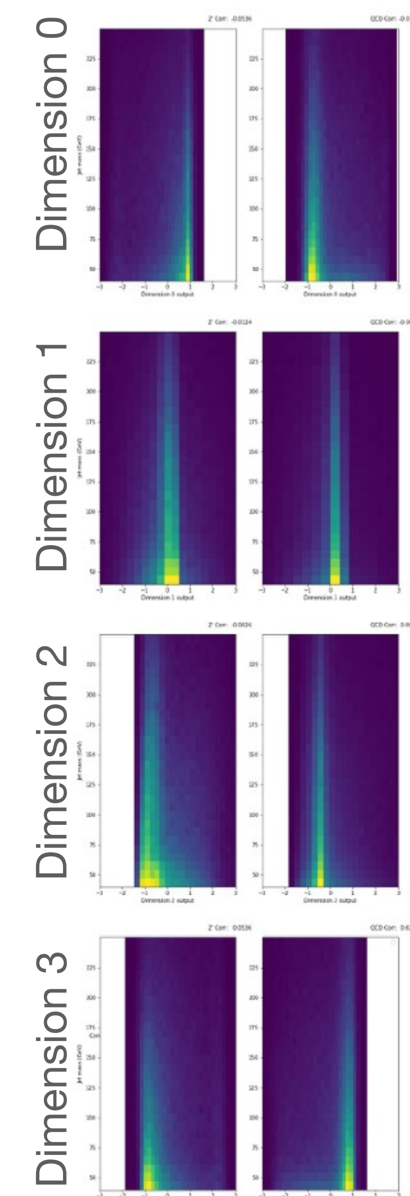


Contrastive Improvement

Flat samples (contrastive)



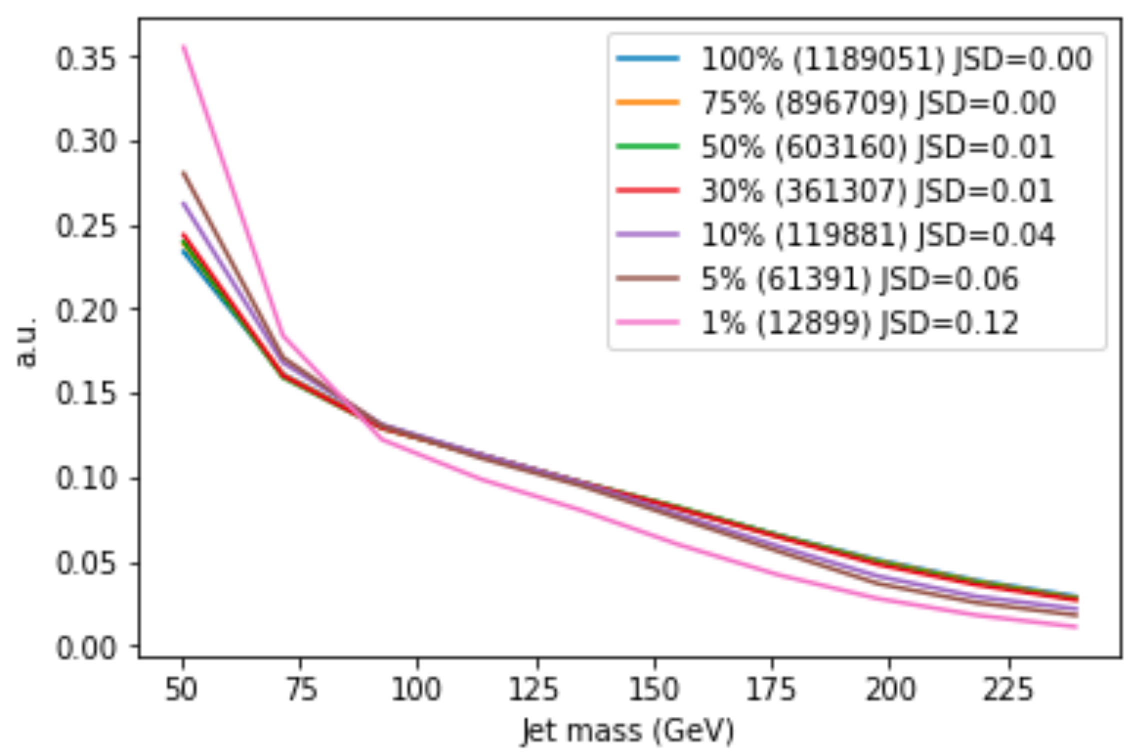




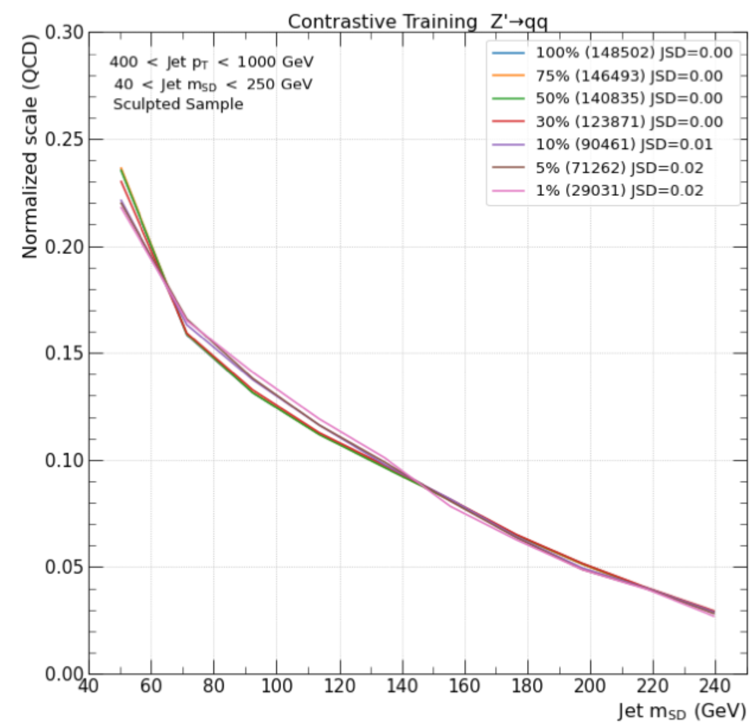
- Best decorrelation scheme is **flat Z' sample + contrastive loss**
  - 4D latent space
  - Tagging similar to MoDe[0]
  - Contrastive alone is insufficient
- With respect to supervised training, contrastive space appears to:
  - **Relies less on on mass** for separation
  - increase performance
  - *Still working to understand the properties of the contrastive space*

MoDe[0] vs contrastive on flat

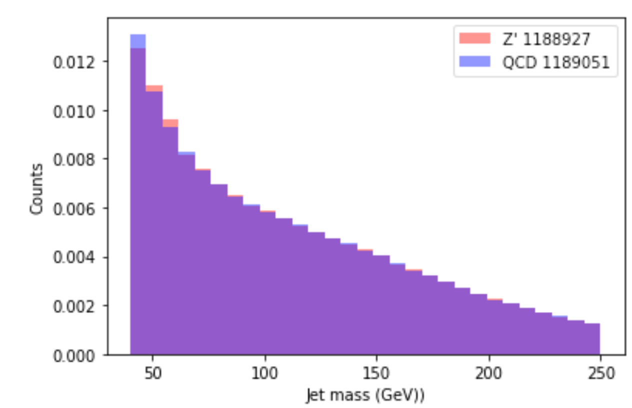
λ mode0 = 10



Contrastive



- As a check, we train flat samples with MoDe[0] loss
  - AUC 86% MoDe[0], 85% contrastive
  - Cutting less than 5% is important

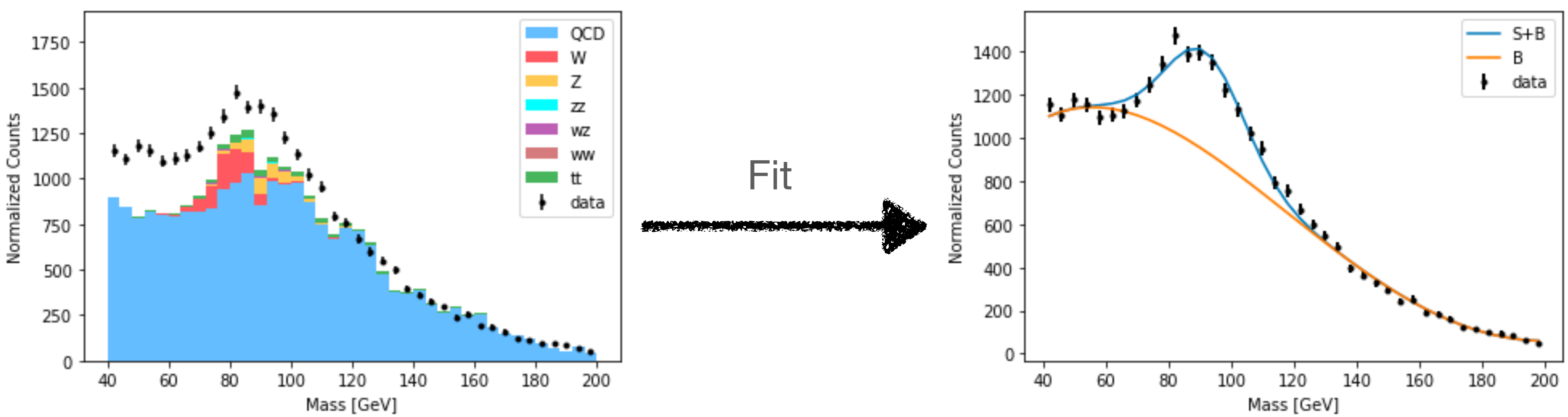


Bonus: Fitting open data

$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]$$

x1 **x1** **x5**

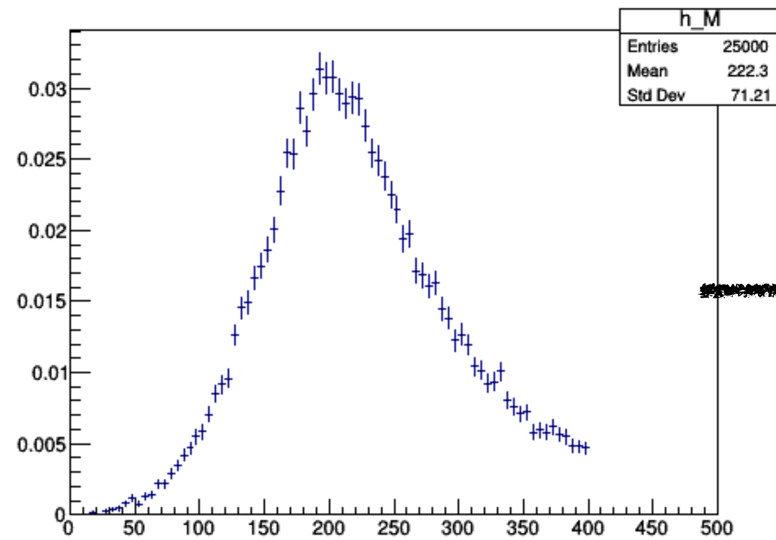
+ No other Terms

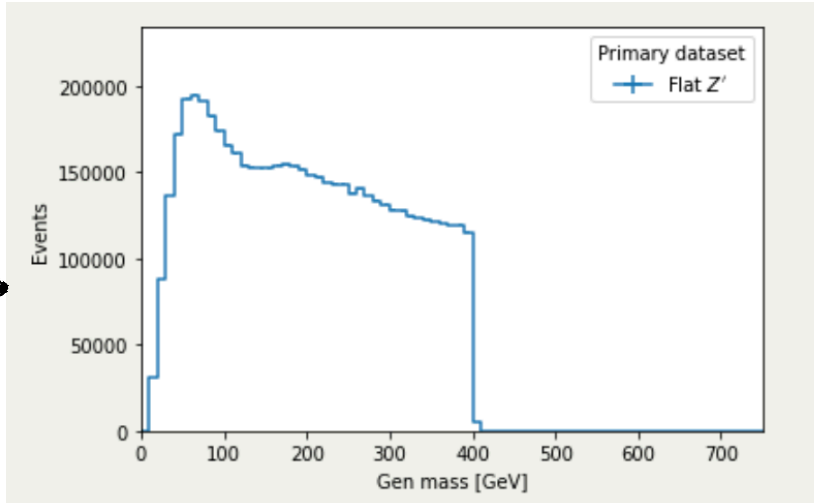


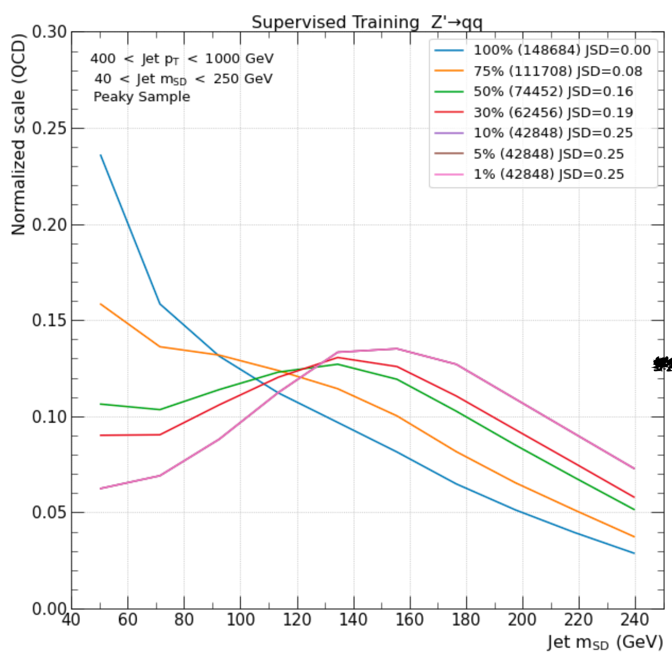
- Some residual correlation
  - Again, mass is not given to the loss

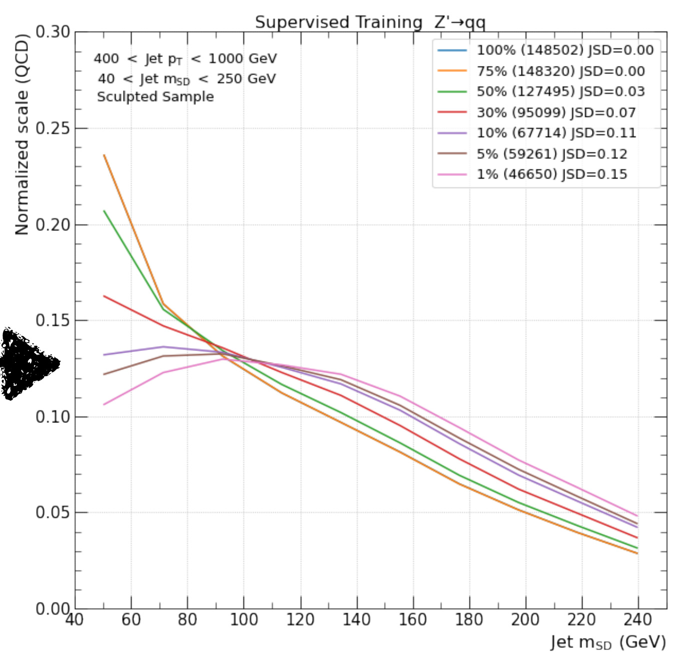
Summary

*We generate **flat samples** through bias weights in madgraph*





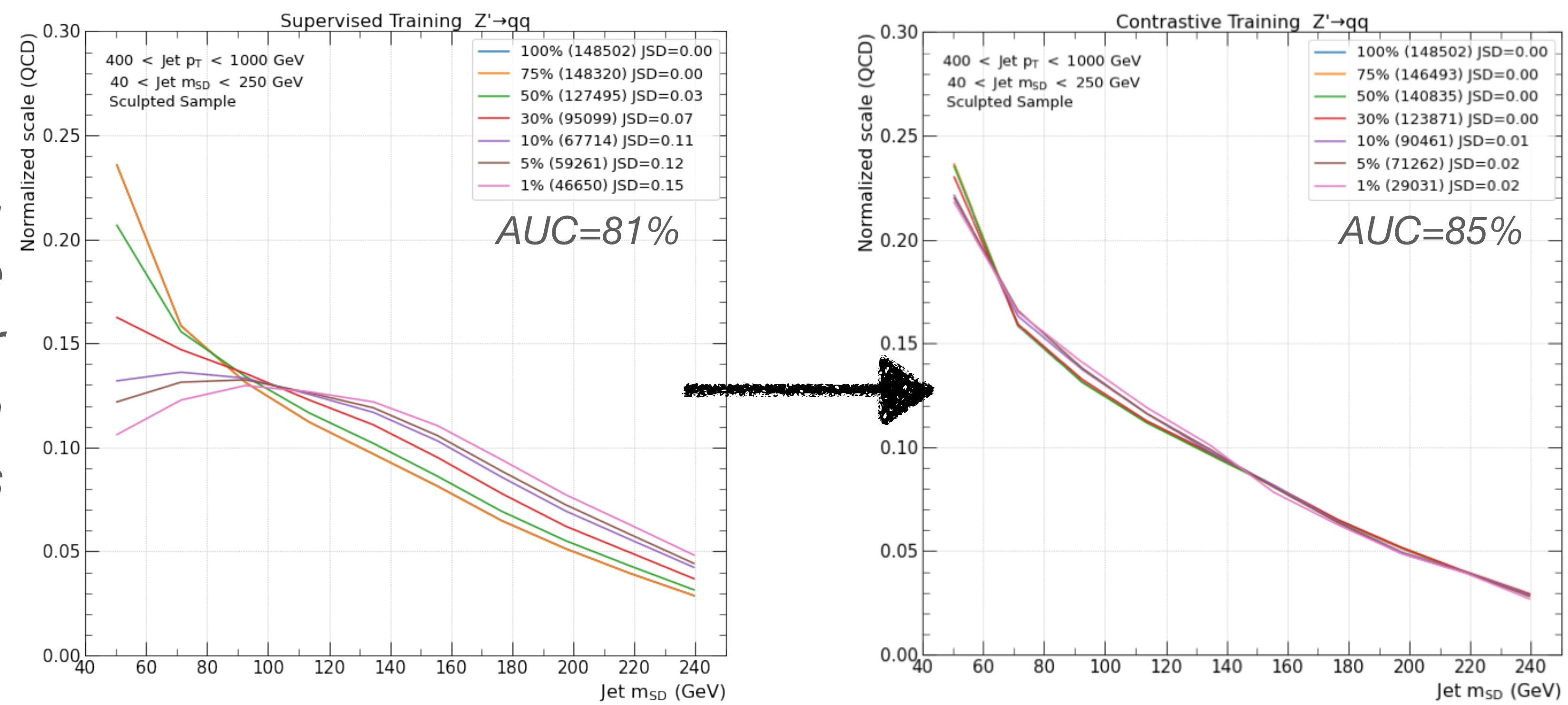


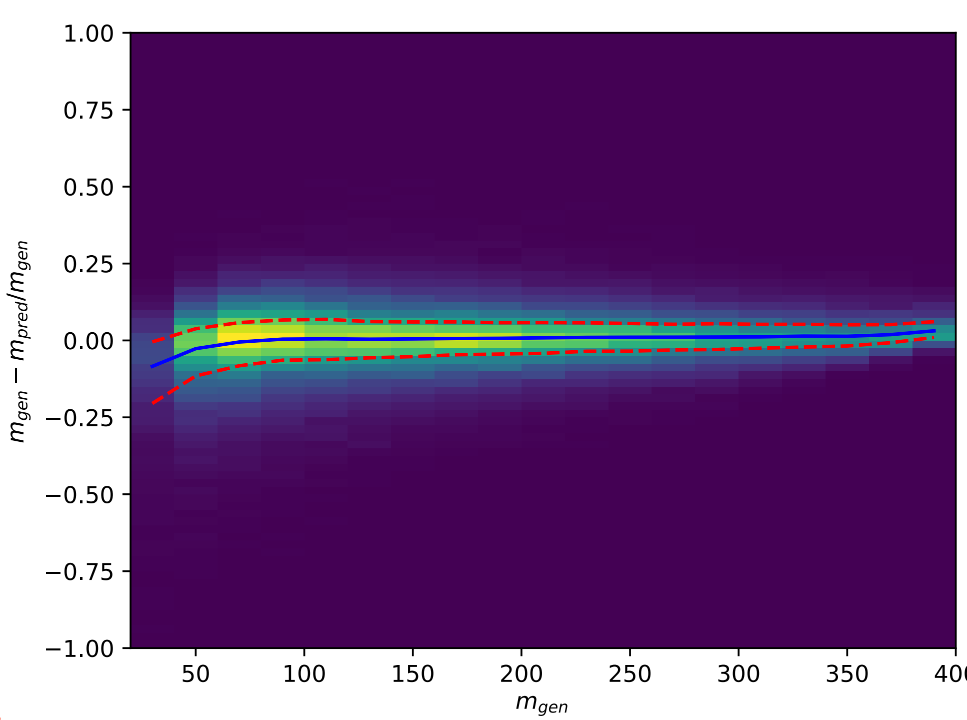


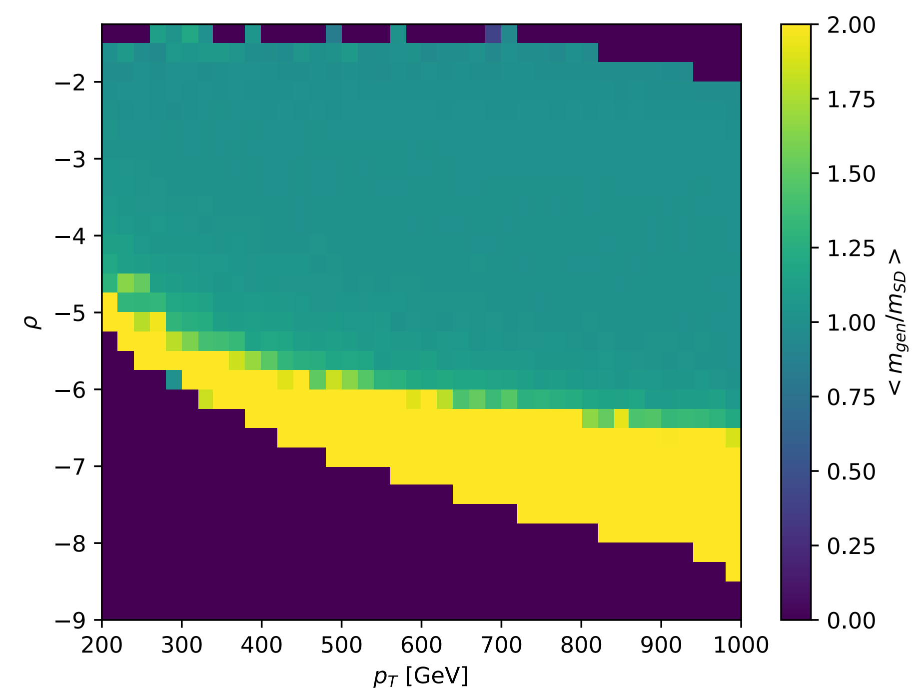
*A naive supervised training on the flat sample is **less correlated with mass***

Summary

We introduce semi-supervised contrastive space which further ***removes correlation with jet mass***





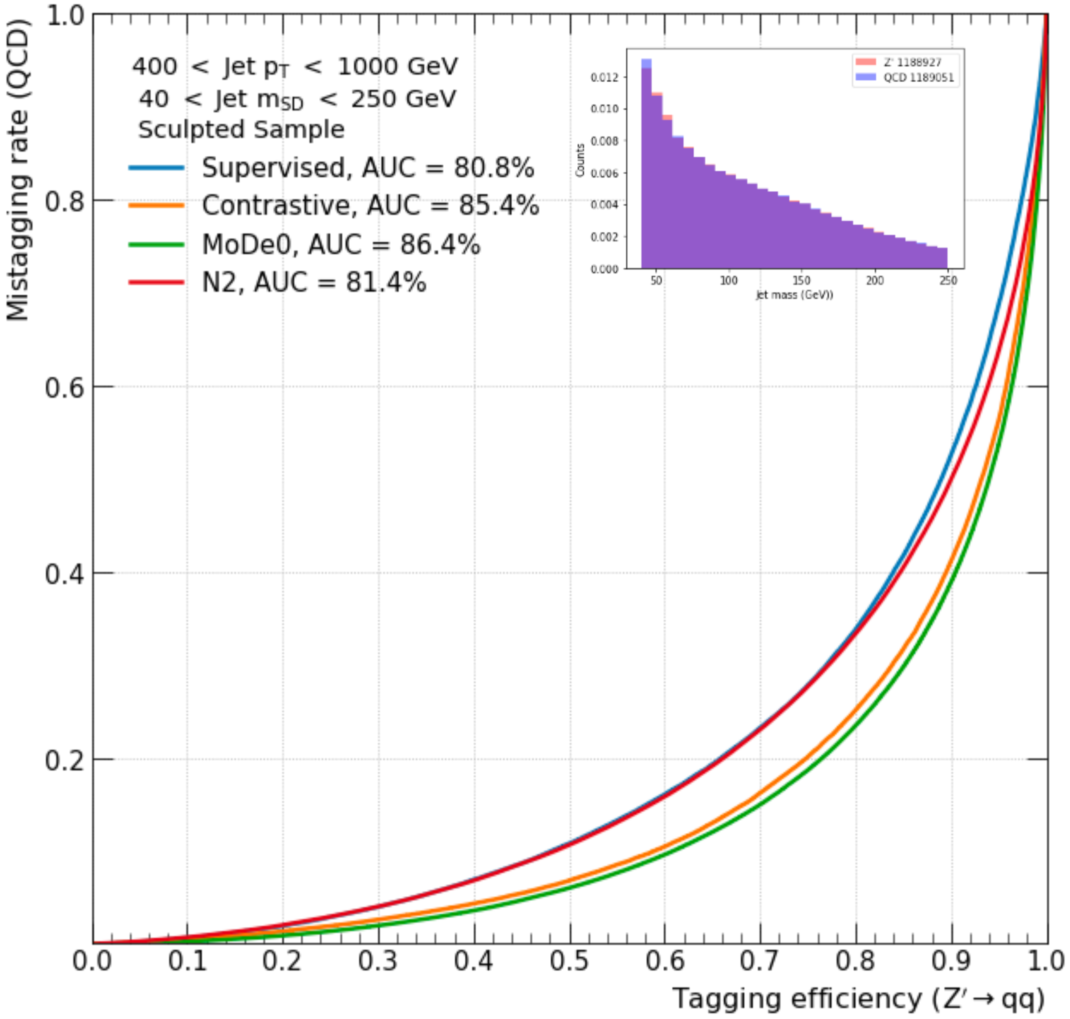


Flat samples also show exciting potential in ***regression*** *and* ***calibration***

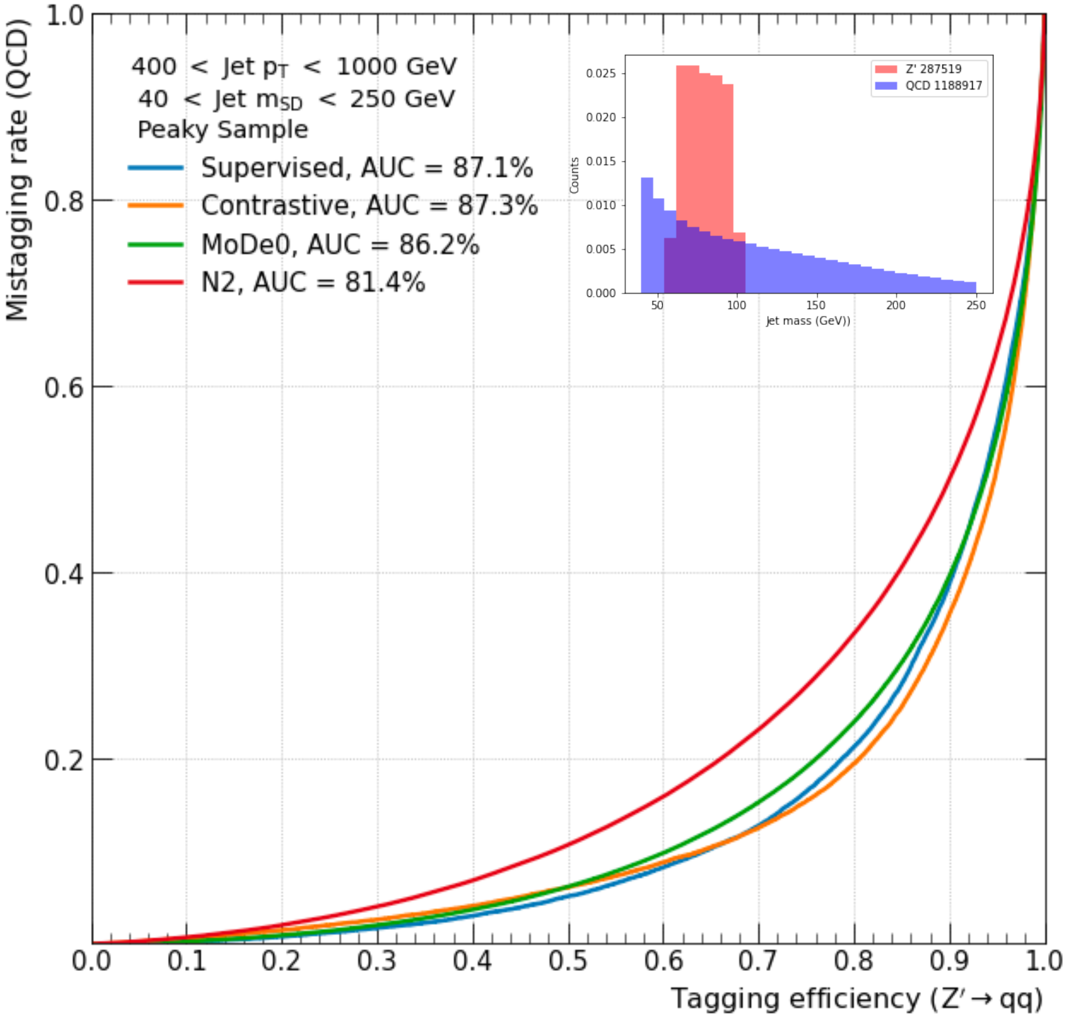
Backup

Tagging ROCs summary

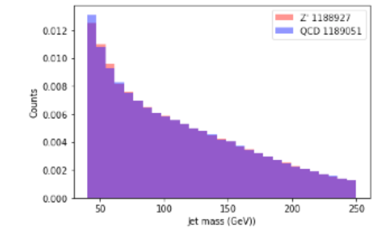
Flat sample

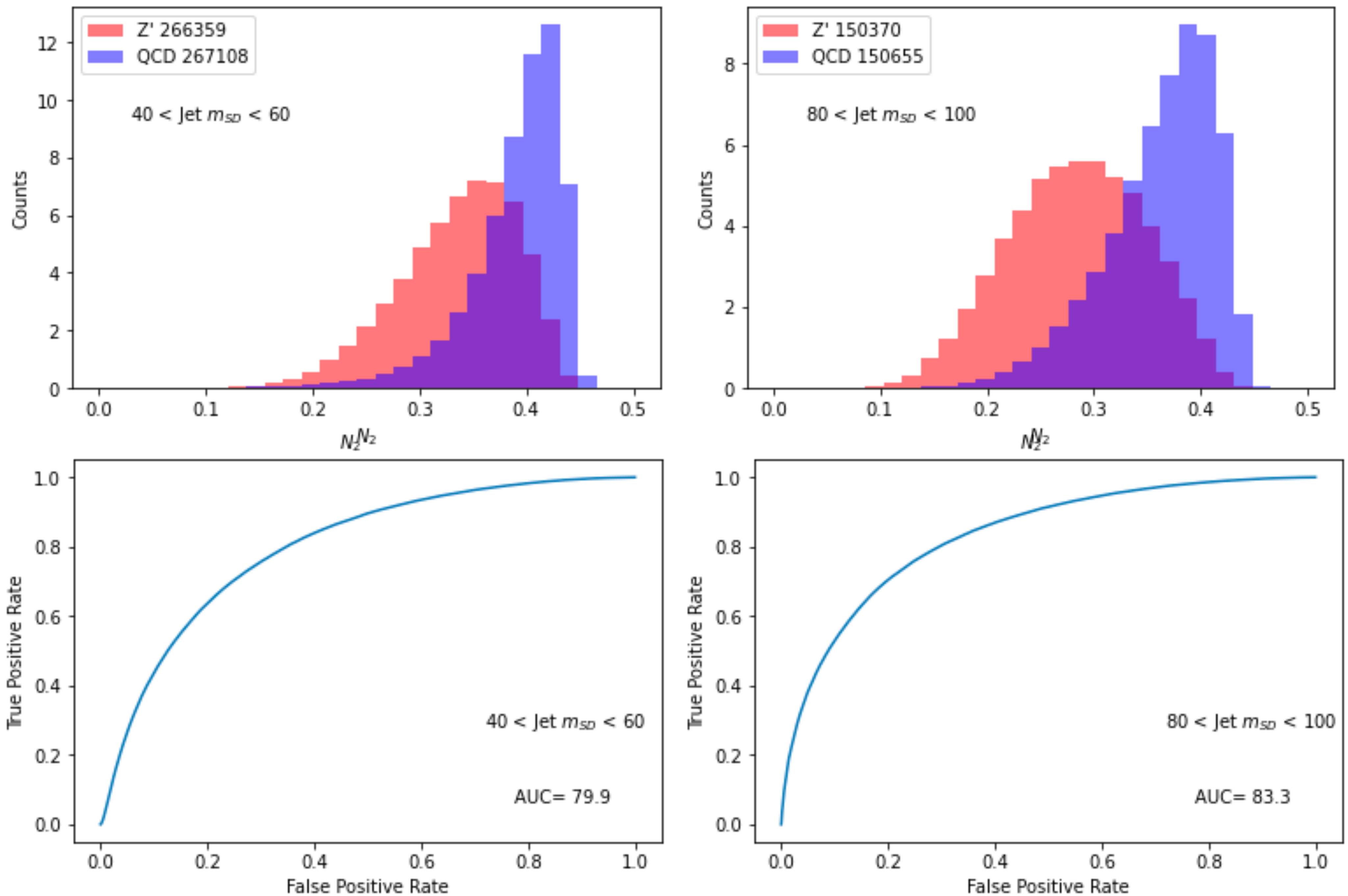


Peaky sample



N2 performance vs mass on flat sample

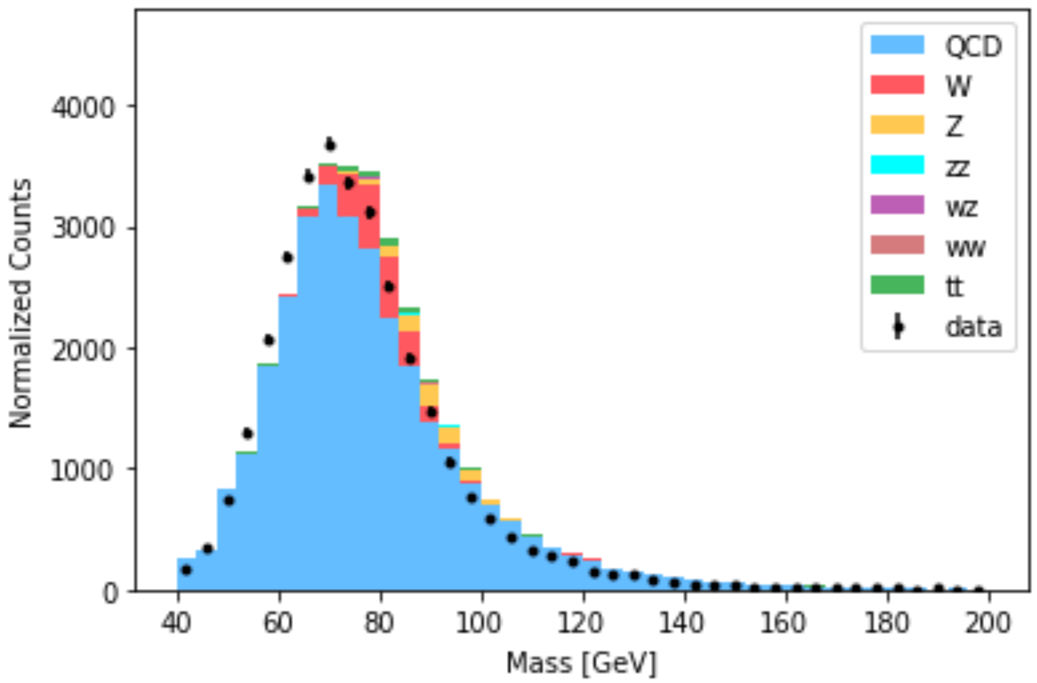


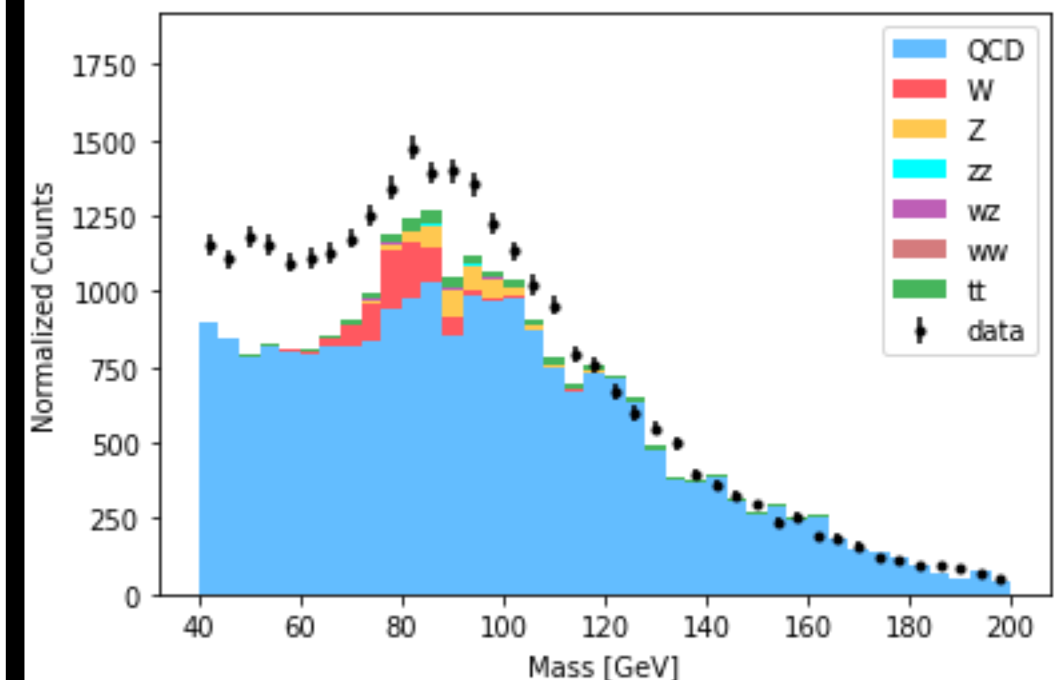


- Lets compare semi-supervised with supervised

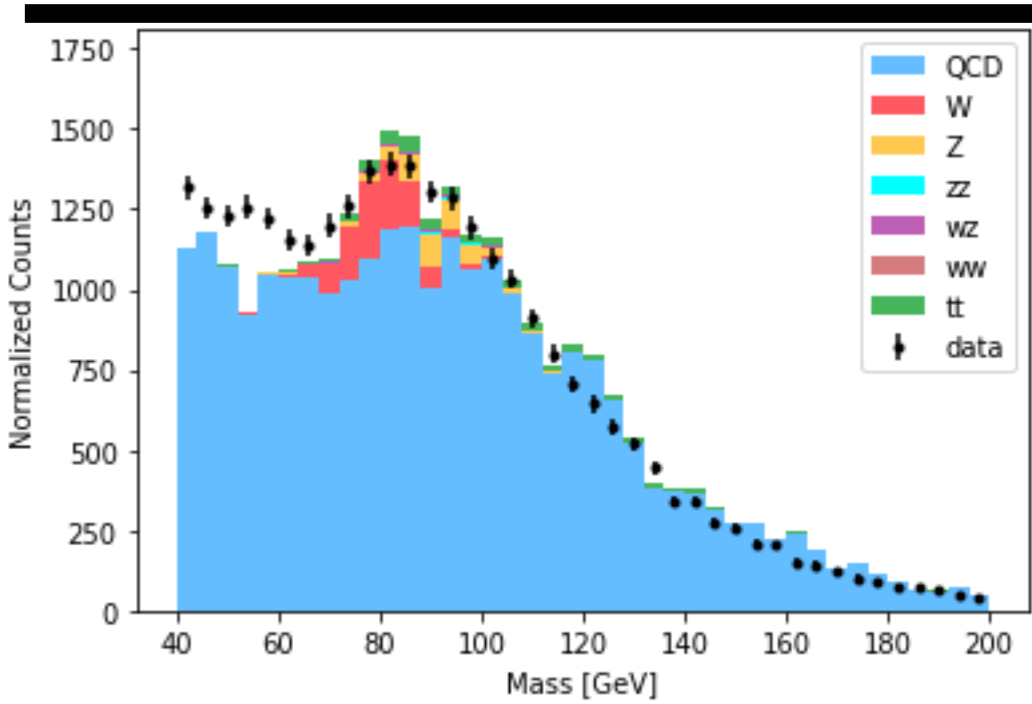
Normal Training | Contrastive Training

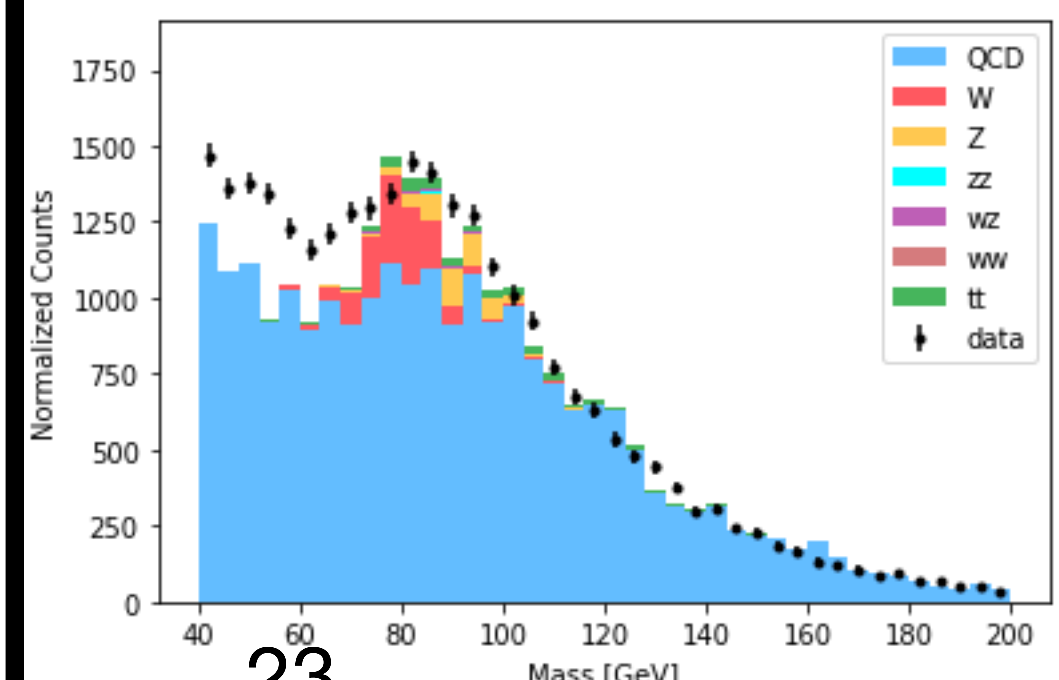
Supervised Loss





Supervised Loss + Explicit Mass Decorrelation





Contrastive Learning