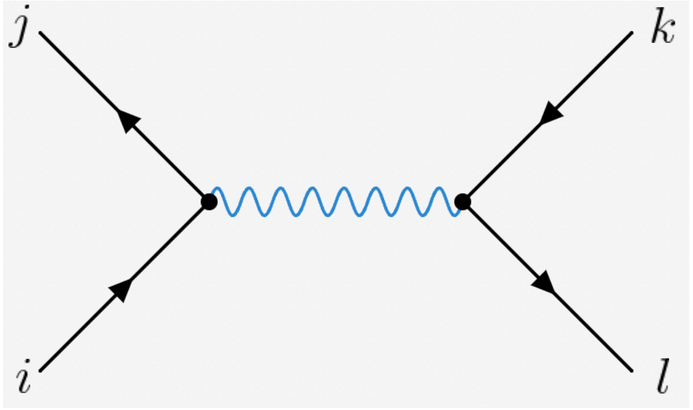
# Riemannian Data preprocessing in Machine Learning to focus on QCD color structure

**Ahmed Hammad**

Seoul national university of science and technology

Based on: arXiv:2209.03898
In collaboration with Myeonghun Park

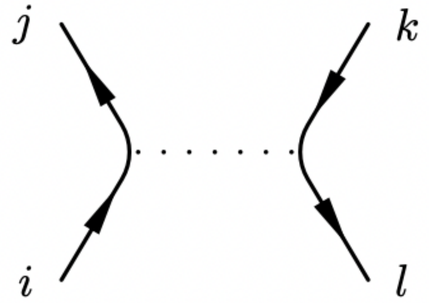
The QCD color flow for particle interactions identified by the contraction of the color indices of the process



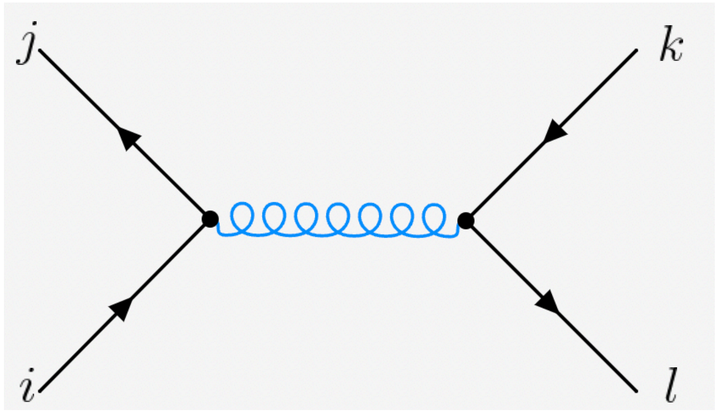

$$Tr[T^i T^j]Tr[T^k T^l]$$

Color lines

arXiv:1206.3700v2



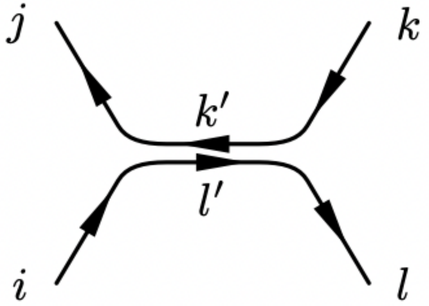

$$\frac{-1}{\sqrt{2}}\delta_i^j \qquad \frac{-1}{\sqrt{2}}\delta_k^l$$




$$Tr[T^i T^k]Tr[T^j T^l]$$

$$Tr[T^i T^l]Tr[T^k T^j]$$

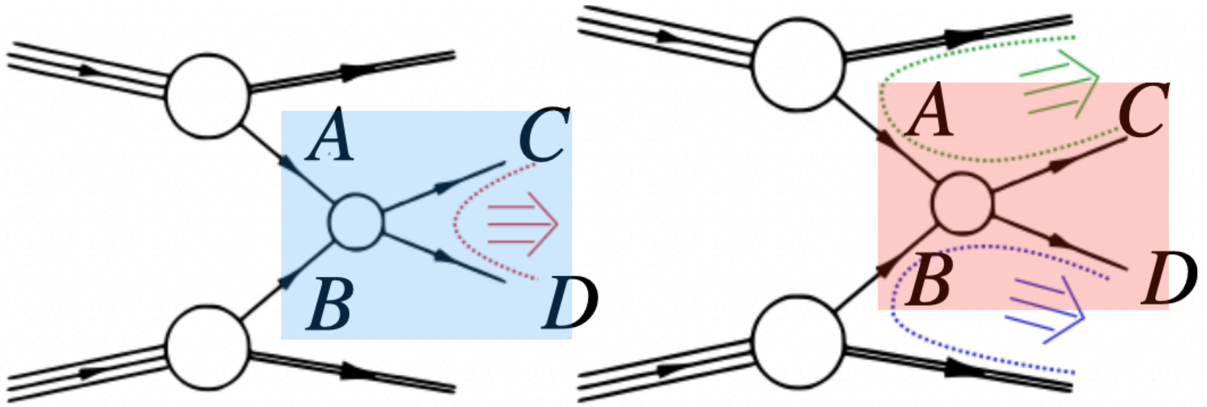
Color lines




$$\frac{1}{\sqrt{2}}\delta_i^{l'}\delta_{k'}^j \qquad \frac{1}{\sqrt{2}}\delta_{l'}^l\delta_k^{k'}$$

# Color flow is physical and can be used to identify the color charge for different mediators

Arxiv:1001.5027v3

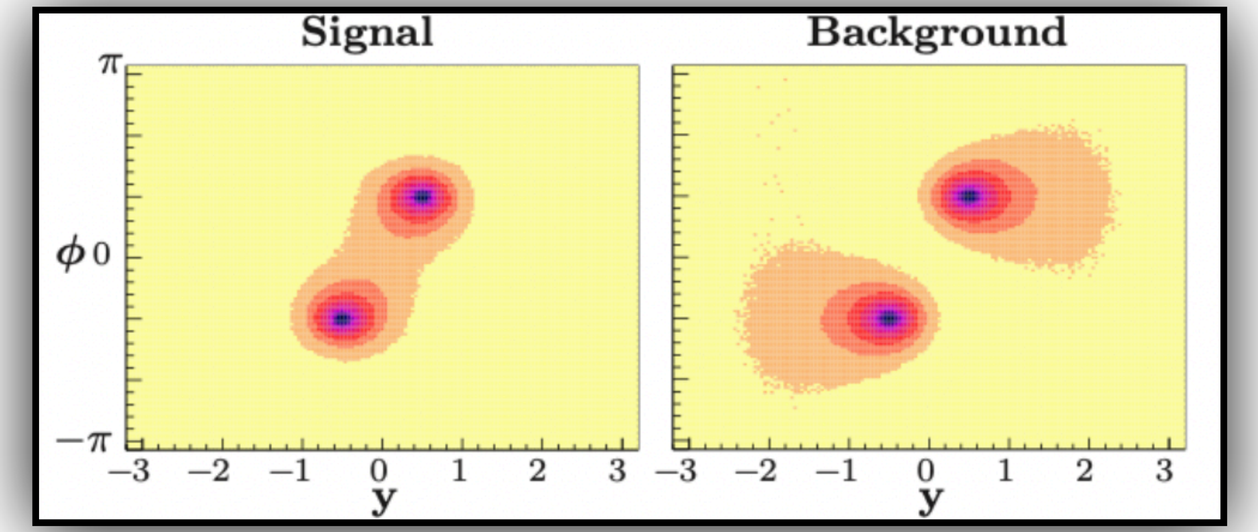



$$\mathrm{Tr}[T^A T^B]\mathrm{Tr}[T^C T^D]$$

$$\mathrm{Tr}[T^A T^C]\mathrm{Tr}[T^B T^D]$$

$$\mathrm{Tr}[T^A T^D]\mathrm{Tr}[T^B T^C]$$

$pp \to H \to \bar{b}b$




$pp \to g \to \bar{b}b$

# How to capture & analyze the color flow at the LHC ?

Consider the LHC detectors as live camera that picture event by event in the eta-phi plane.
Constructed pictures are two dimensions array and the pixel intensity is weighted by the energy deposit of each particle in the corresponding part of the detector

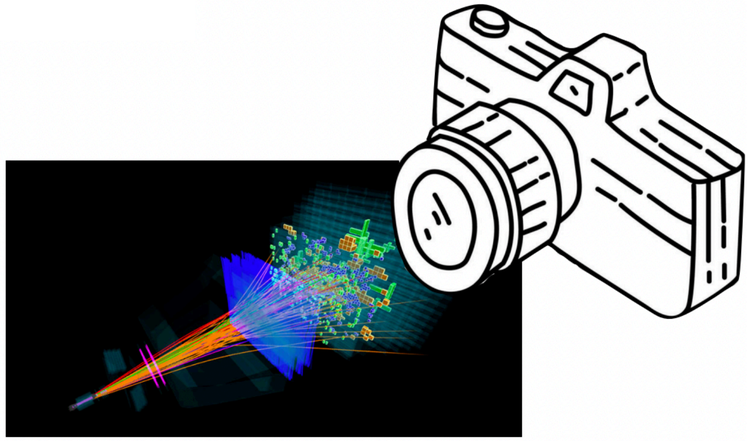
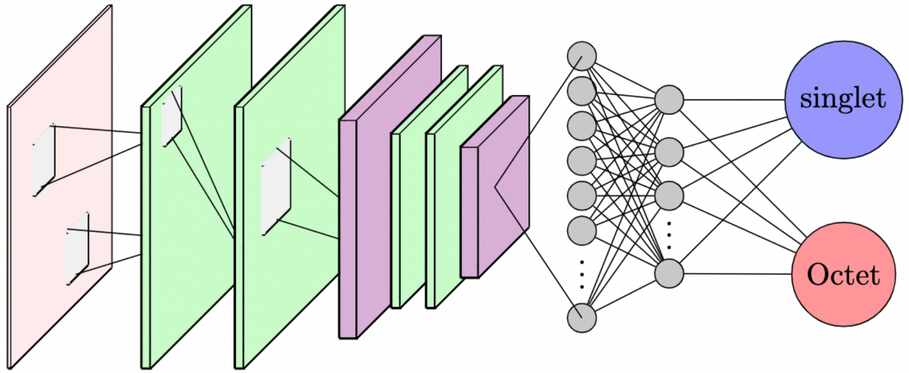
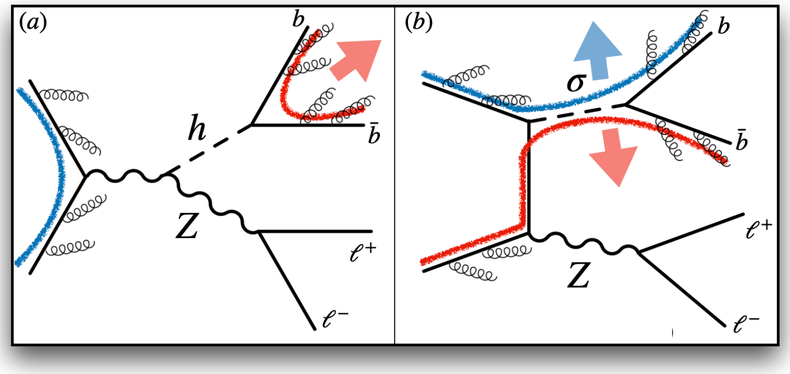


Image preprocessing




Analyze the model output

# Color Singlet Vs Color Octet



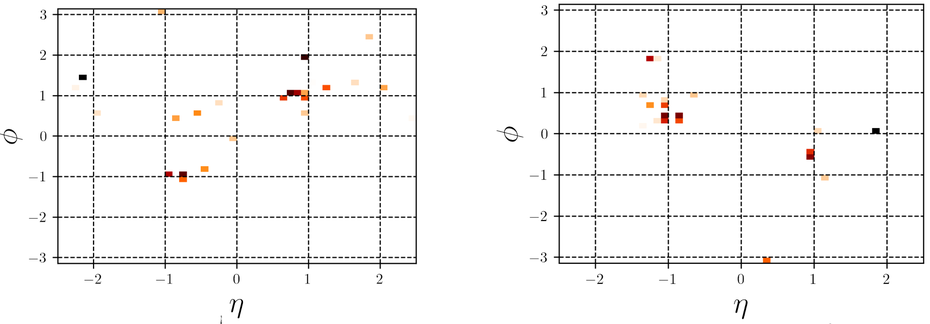
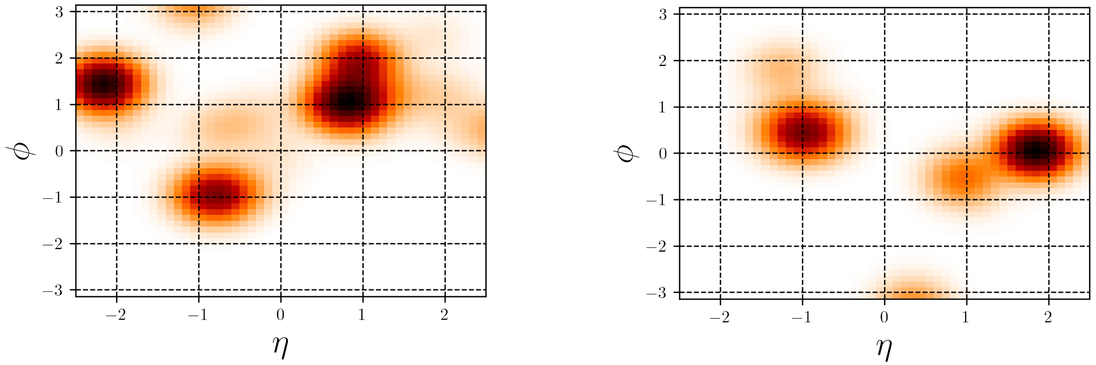
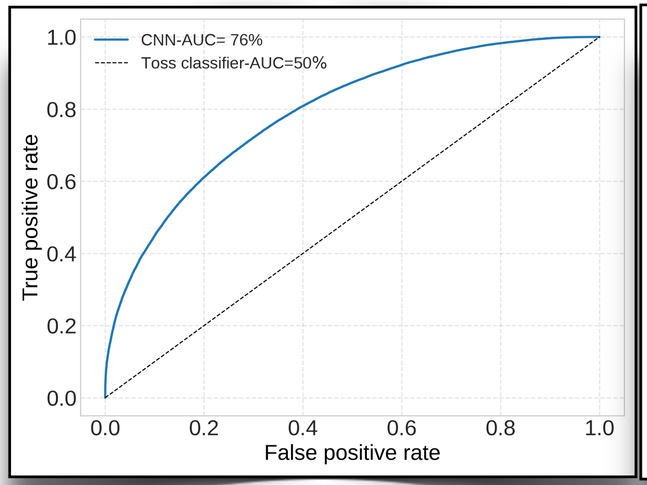
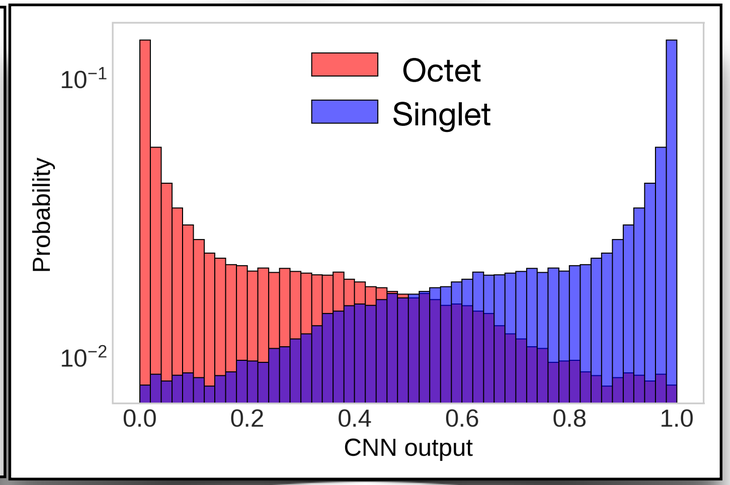
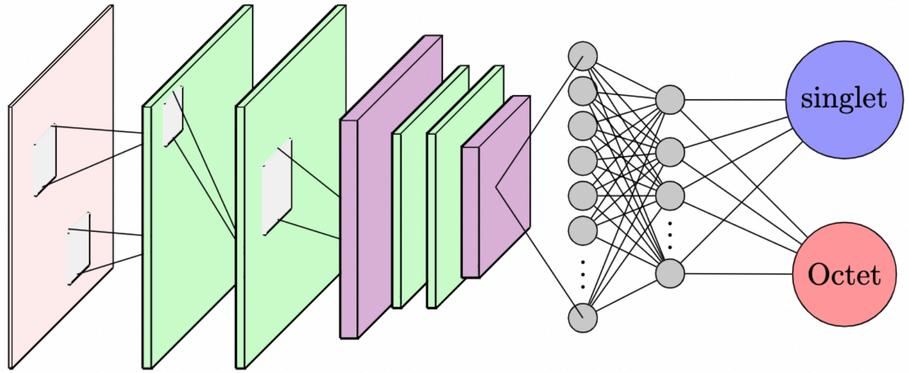




Image preprocessing



# Why the performance is low ??!!!!

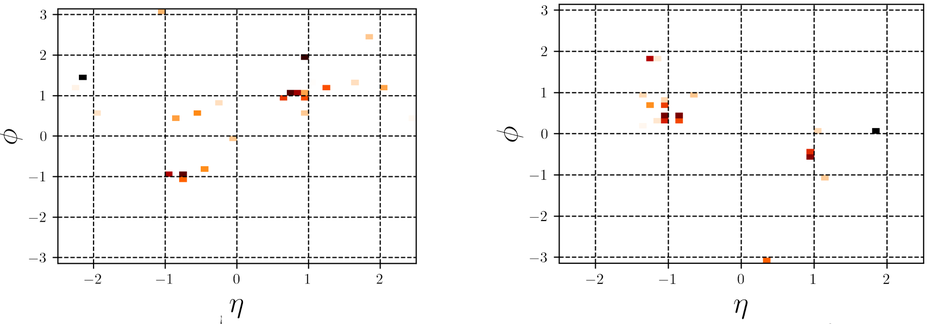
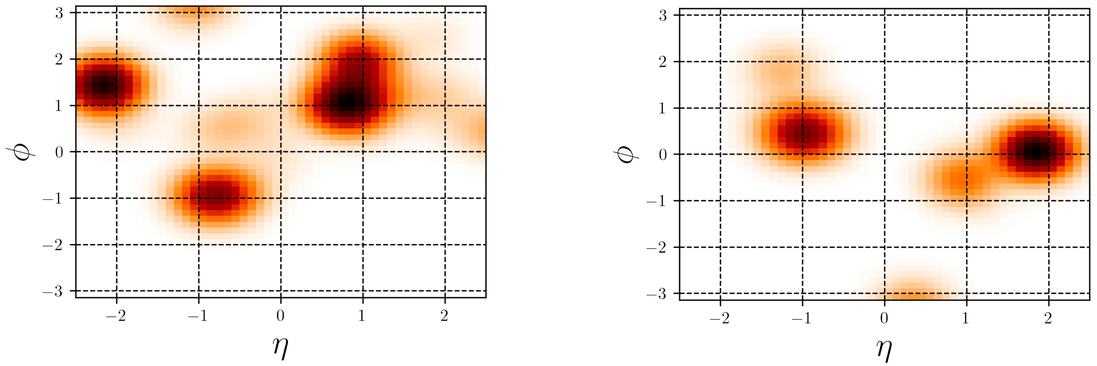
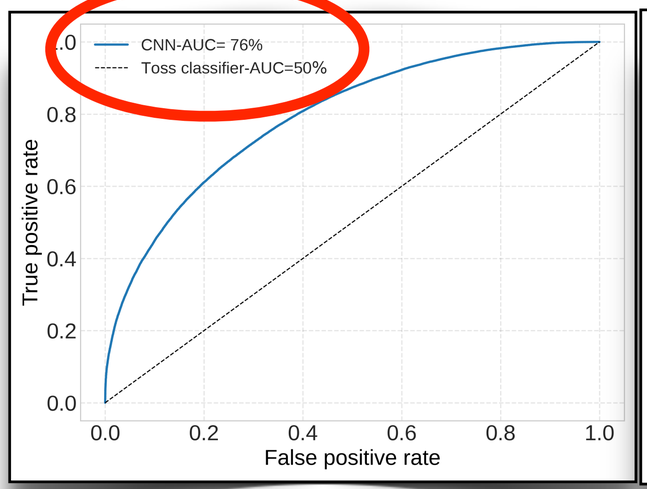
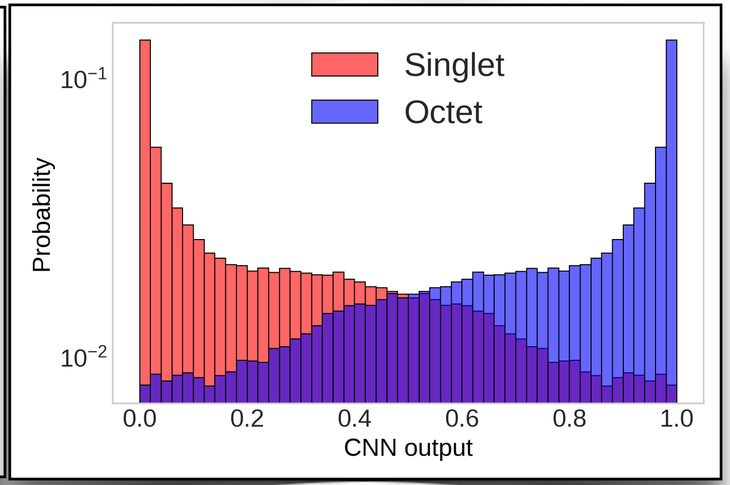
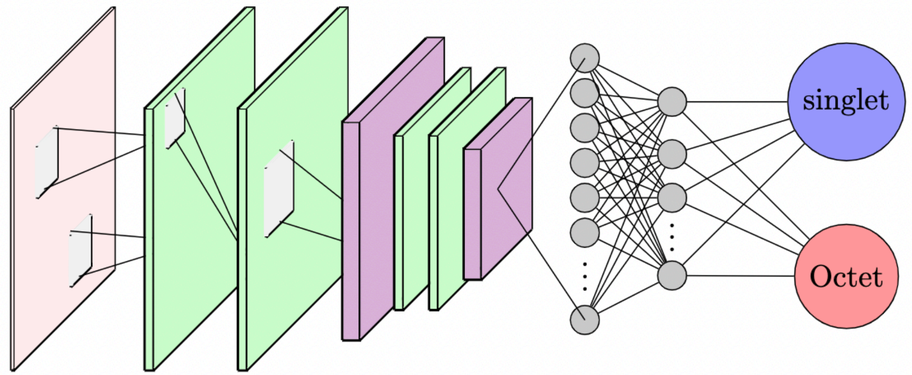



Image preprocessing

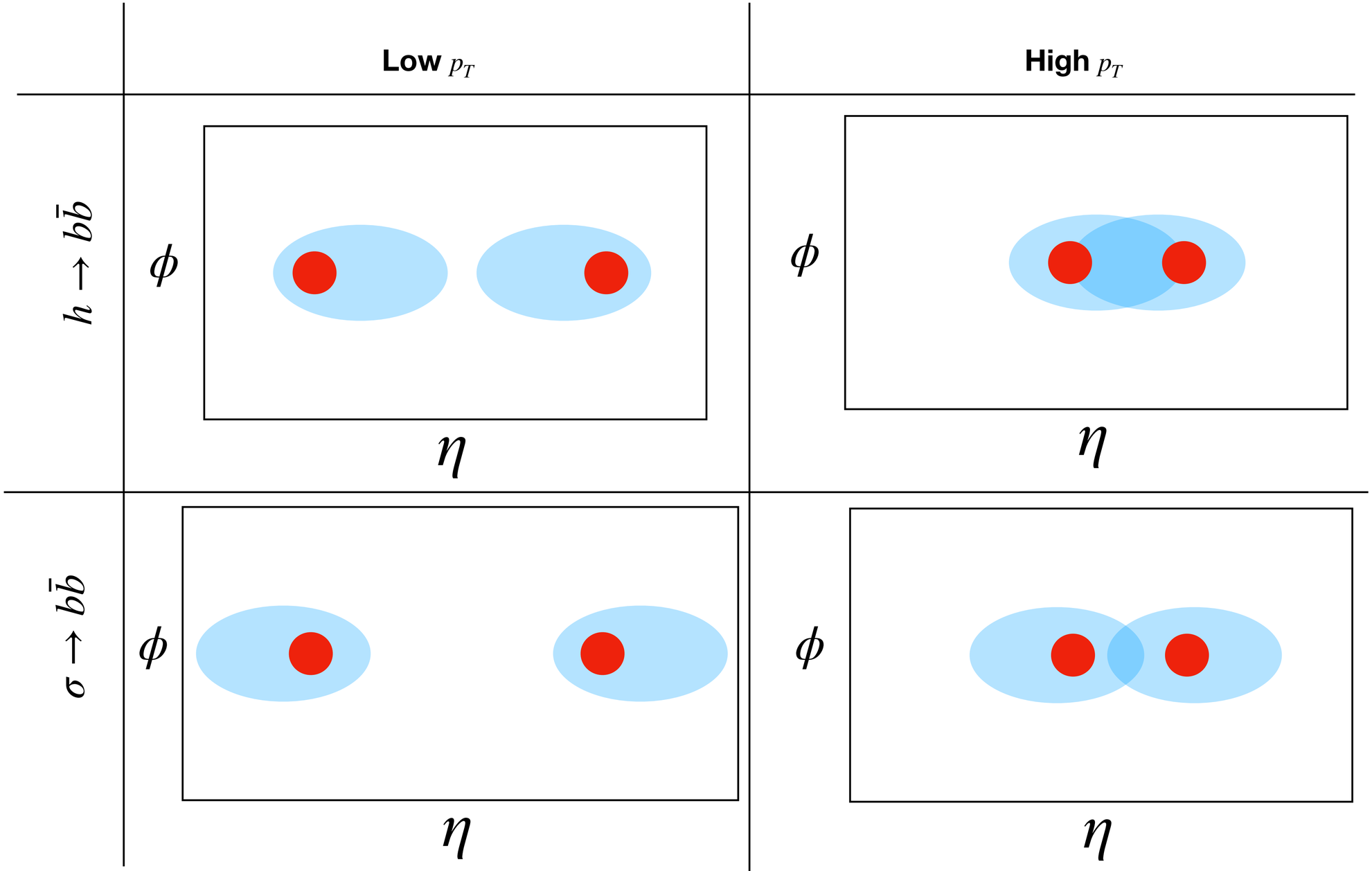


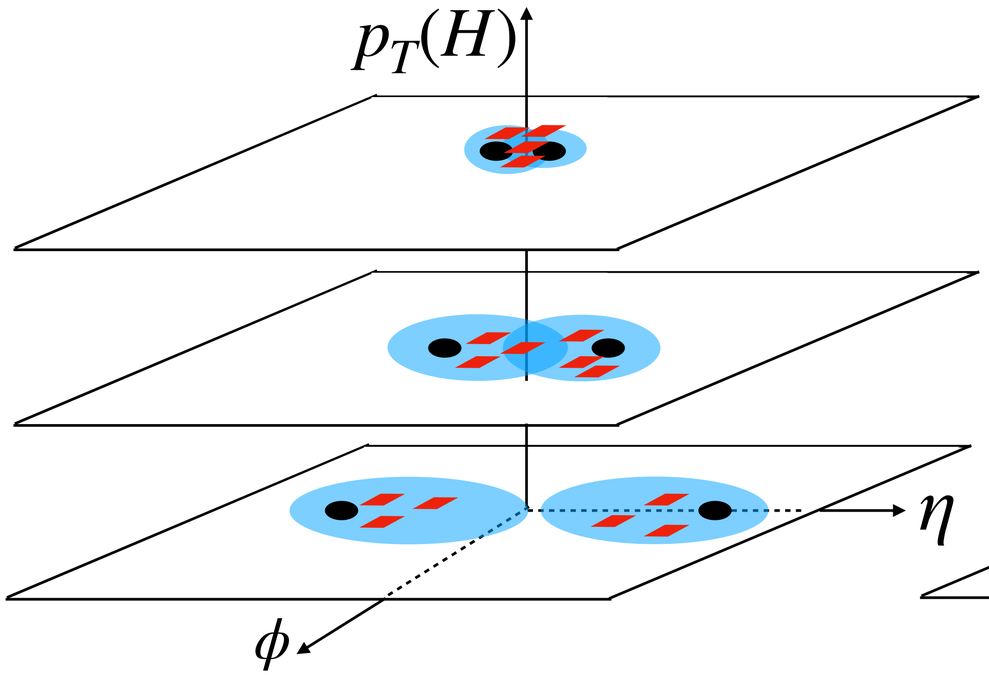



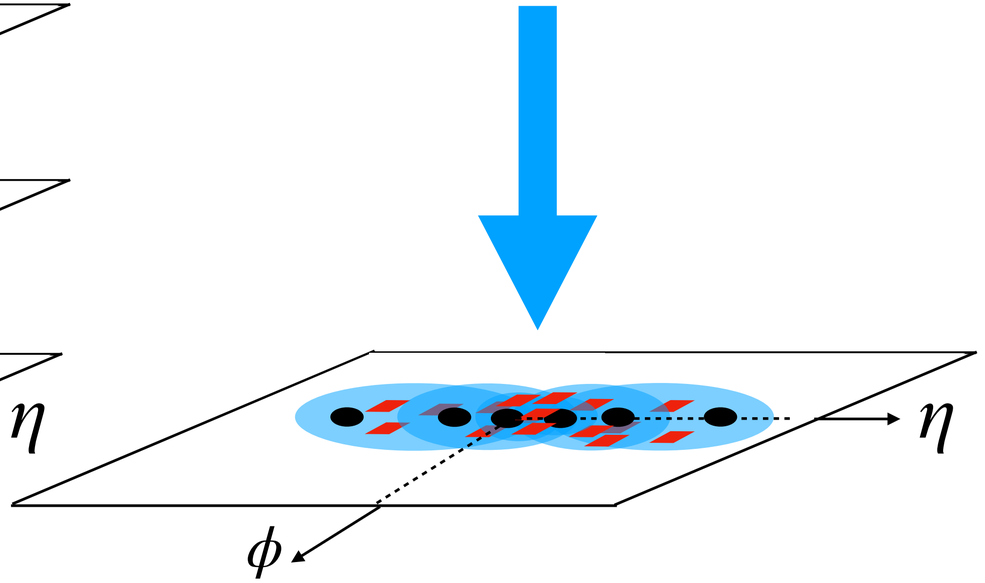



Distribution of the pixels intensities depend on the transverse momentum



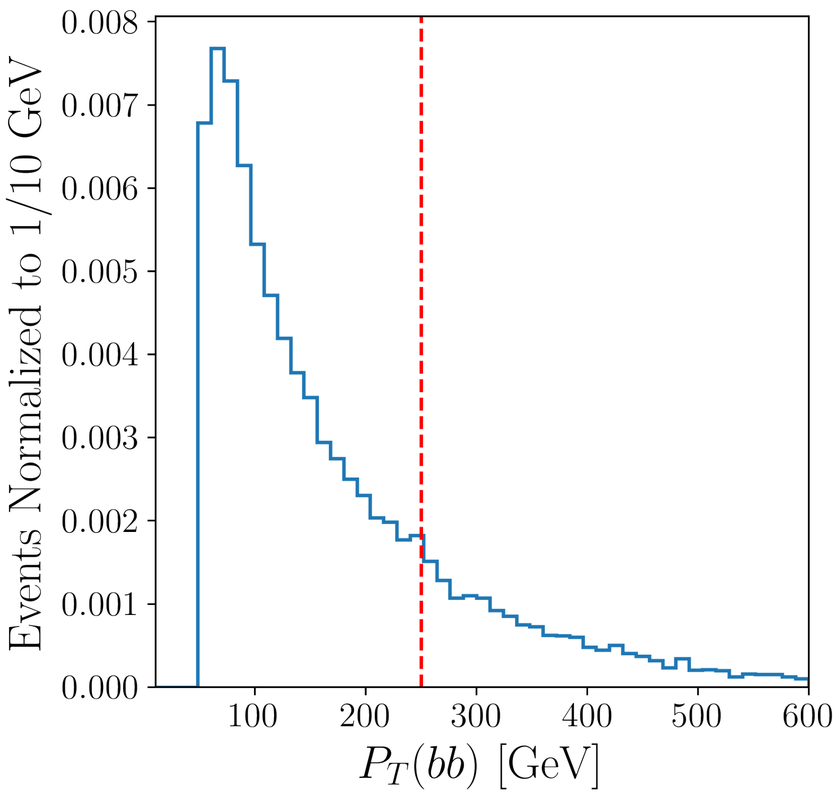


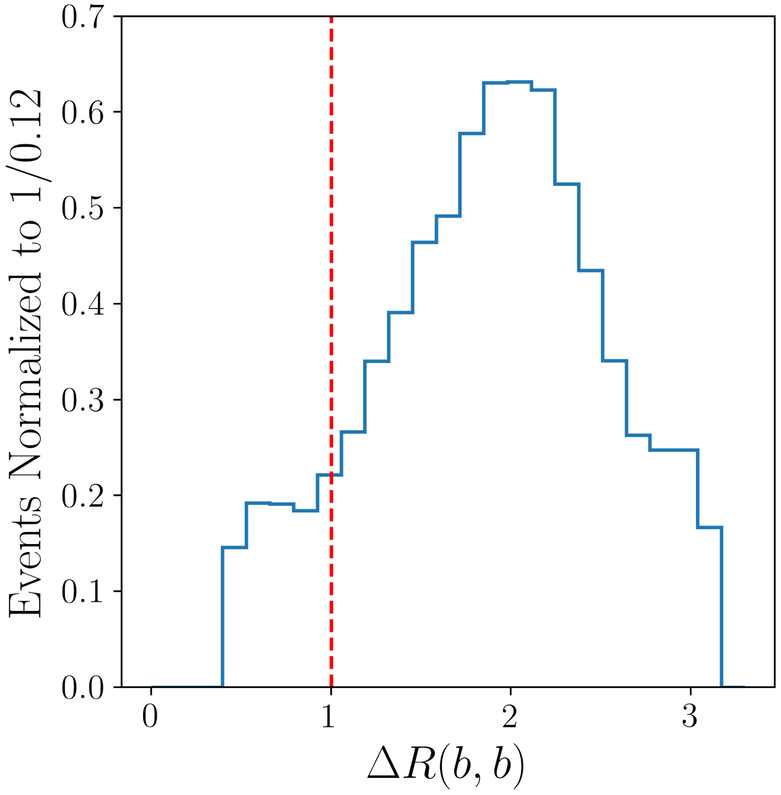

**Projection along** $p_T(H)$

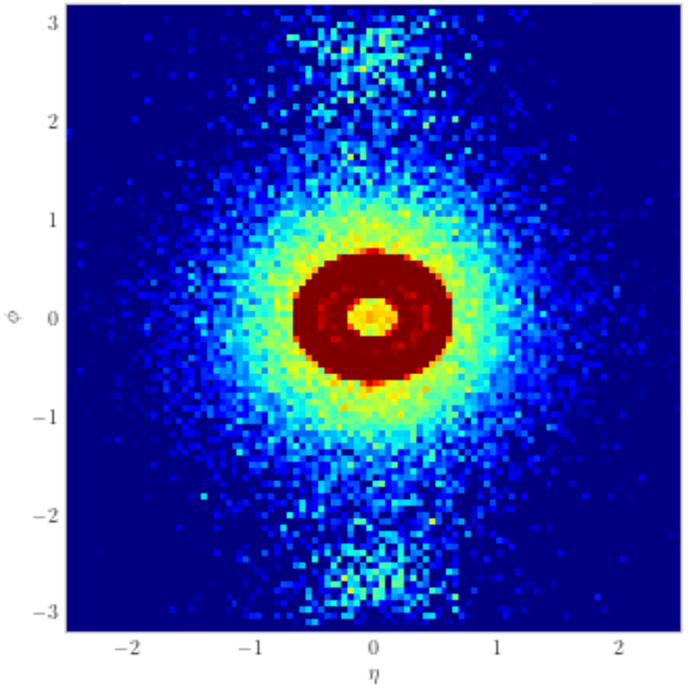



The CNN tries to learn very complicated pattern

# Solution 1 : Focus on the boosted region only



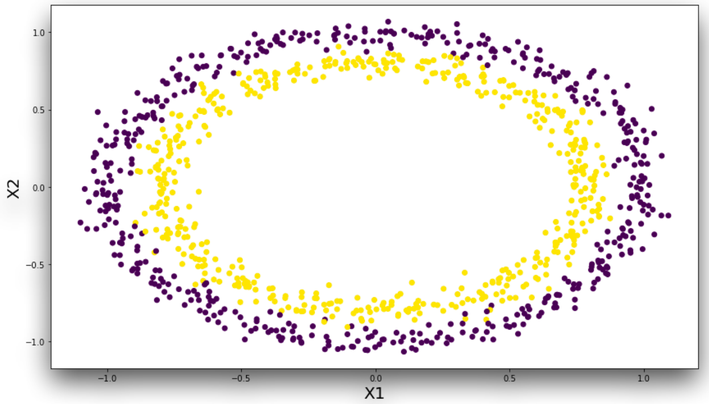



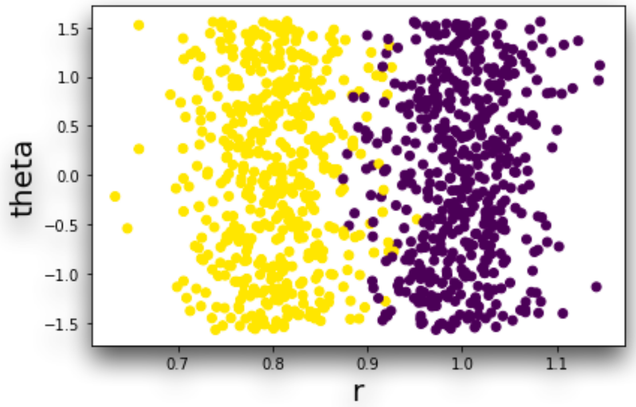




We loose statistics !!!

# Solution 2 : Kernel mapping

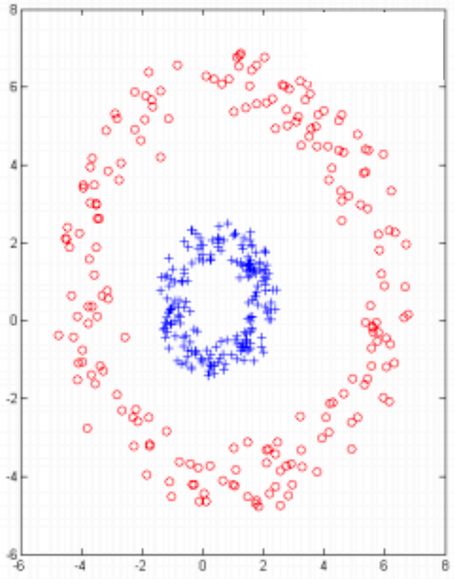
By mapping non-linear separable data from low dimensional space to other coordinates
by using specific kernel, one can find a hyper-plane that can easily separate between the data

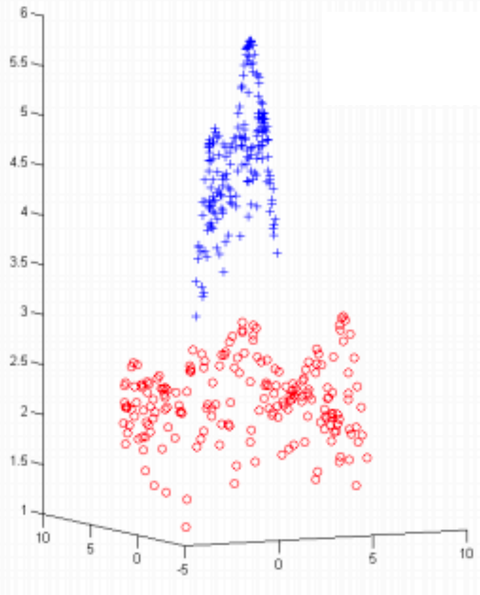



Cartesian coordinates -> polar coordinates

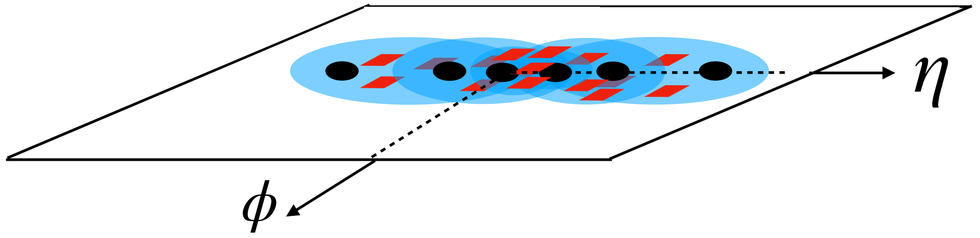
$$\theta = \tan^{-1}(\frac{x_1}{x_2}) \quad r = \sqrt{(x_1^2 + x_2^2)}$$



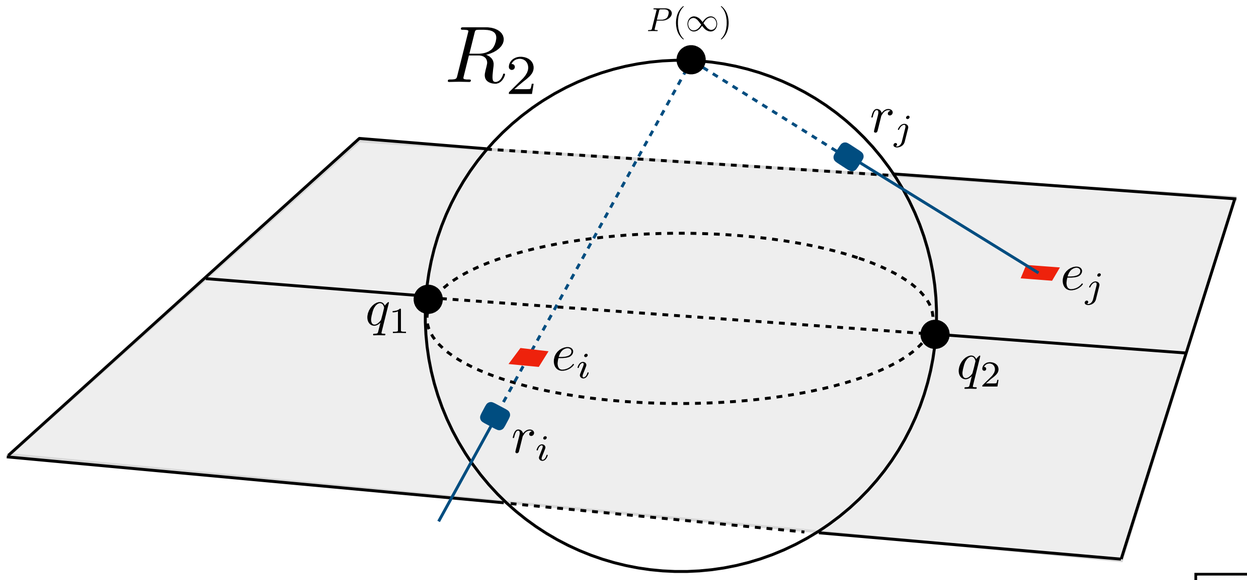


2d -> 3d  using Gaussian RBF kernel

Which kernel shall we use in our case ?



# Inverse stereographic projection

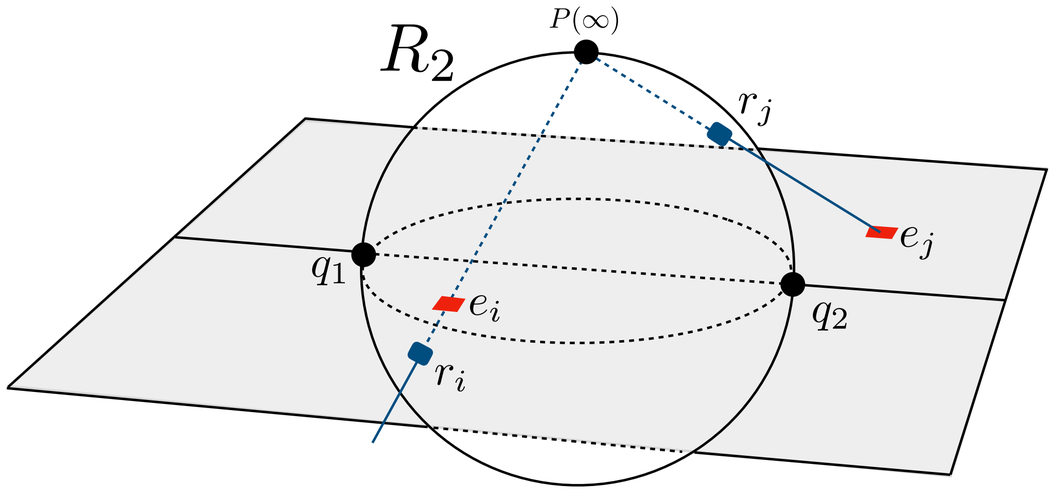


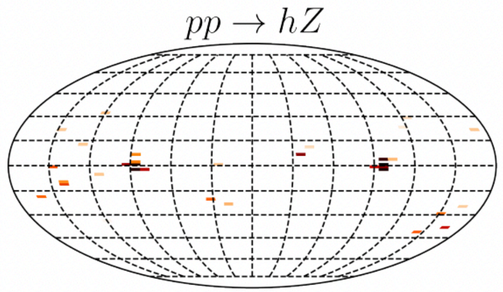

Stereographic projection, Six Books Of Optics
By François d'Aguilon

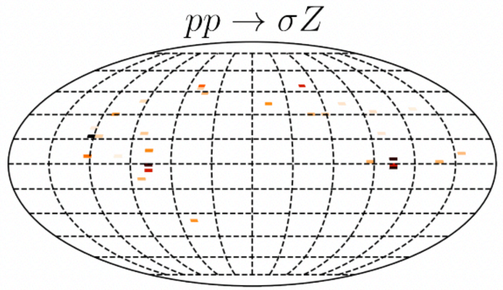
$$\cos\theta_i^R = \left(\frac{\eta_i^2 + \phi_i^2 - r_R^2}{\eta_i^2 + \phi_i^2 + r_R^2}\right), \quad \phi_i^R = \arctan\left(\frac{\phi_i}{\eta_i}\right)$$

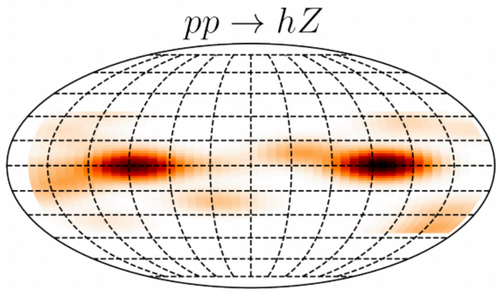
- Soft radiations which are
  inside of a circle → Southern hemisphere
  outside of a circle → North hemisphere

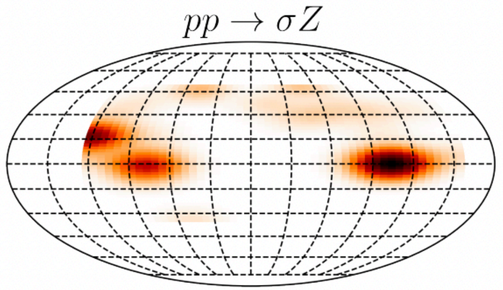
# Inverse stereographic projection



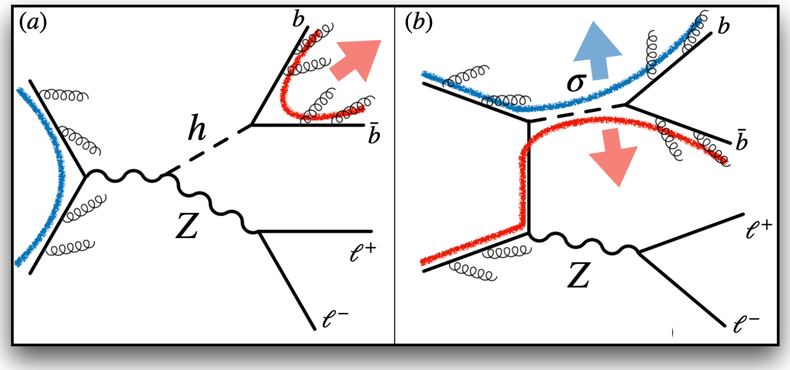



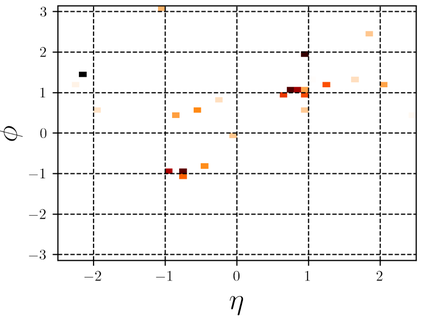



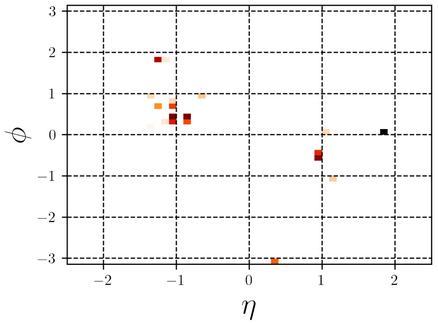




Position of the hot cores is fixed to $\pm\frac{\pi}{2}$



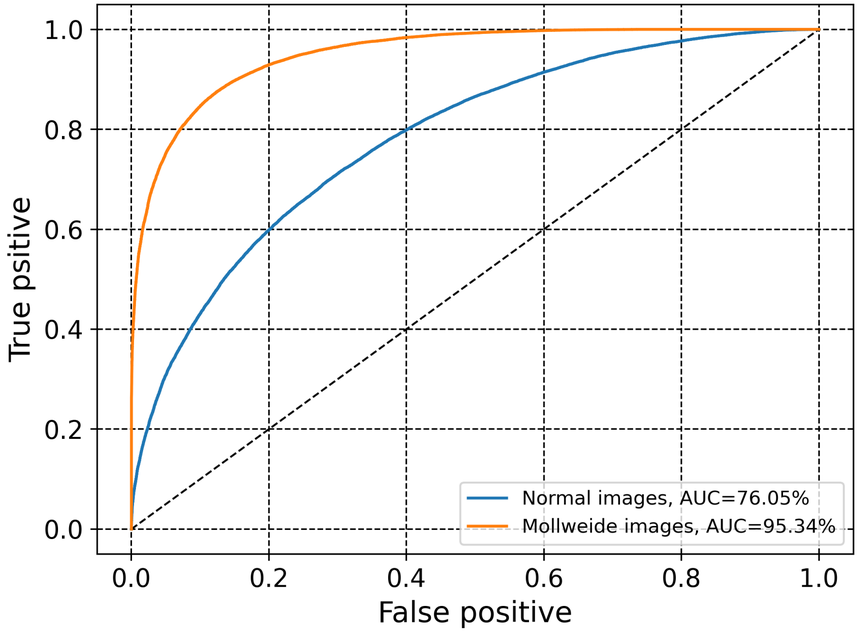
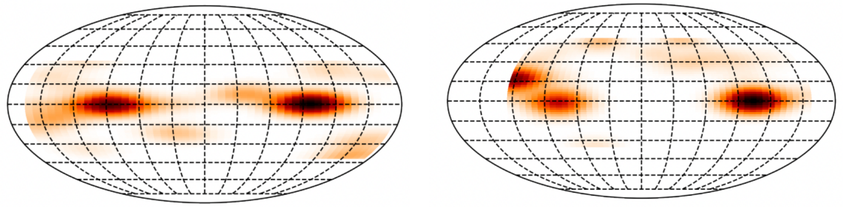
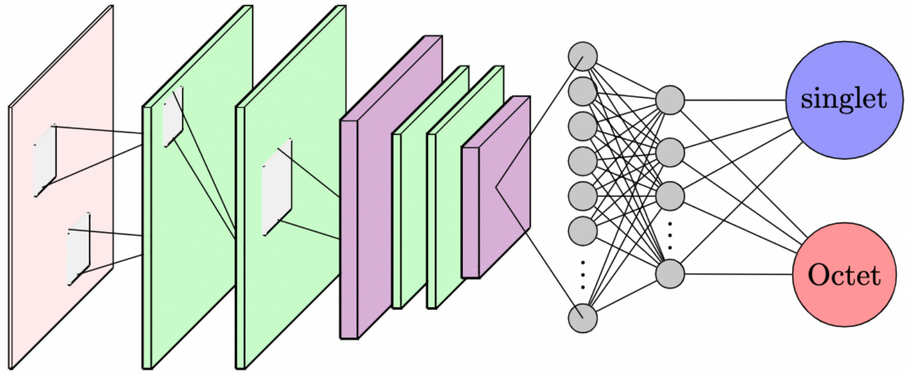
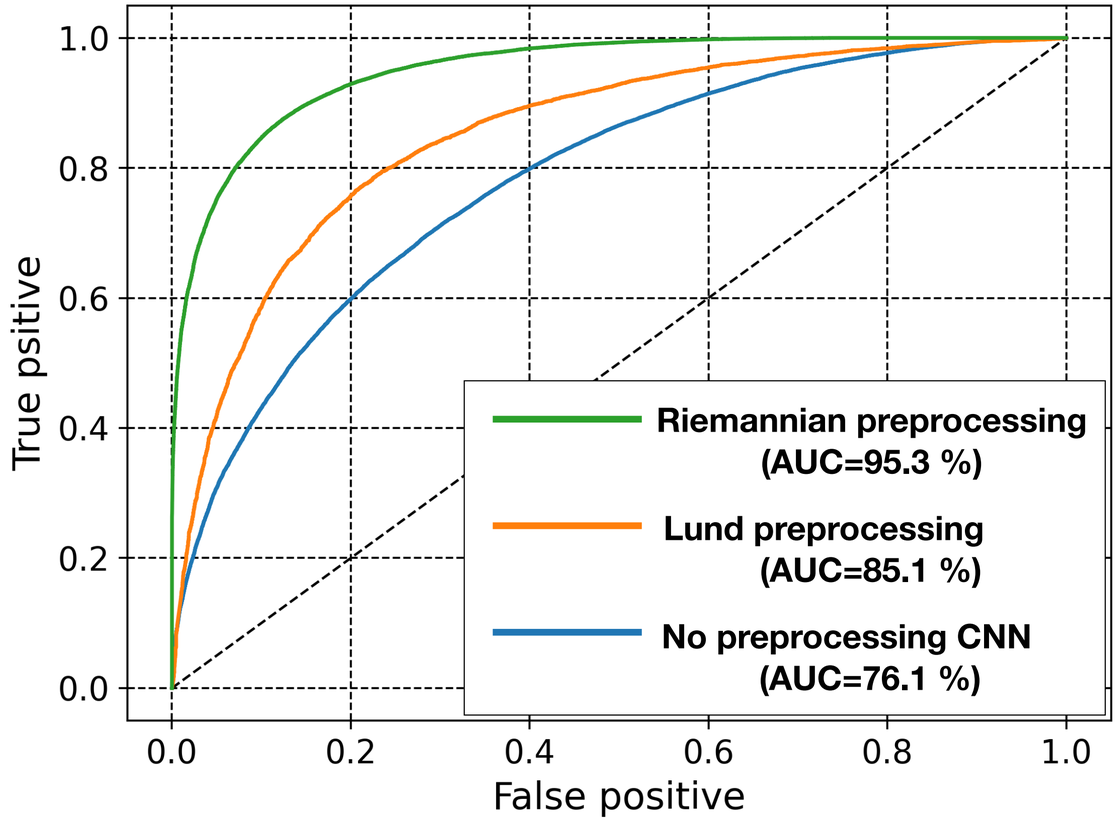




Image preprocessing
+
Riemannian mapping





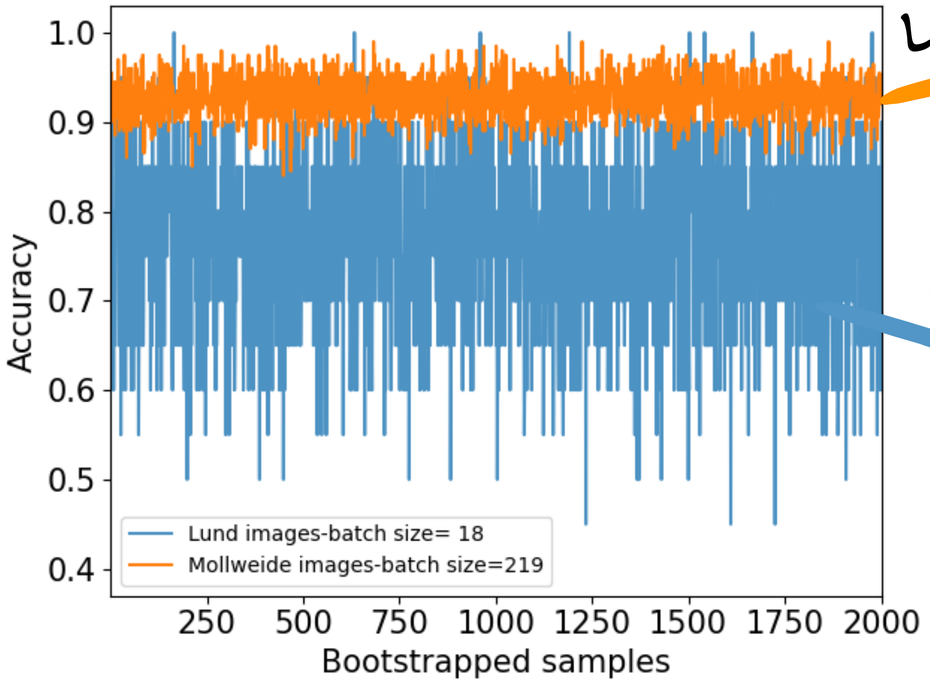
# Performance test



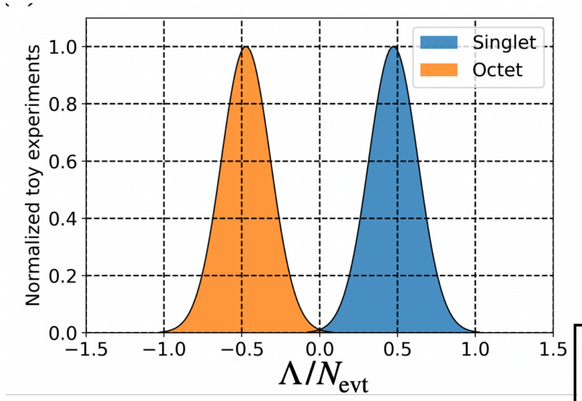

- With 100,000 MC data sample each for (1) whole $p_T$ range and for (2) boosted $p_T$ (60% training, 20% validation, 20% test),
  Riemannian preprocessing has an outperformance.

- Lund preprocessing ("double-logarithmic plane") is from [arXiv:2105.03989] for a boosted Higgs (Data preprocessing with selected QCD features)

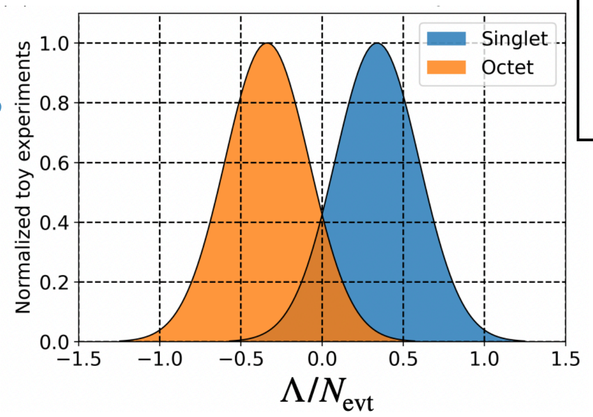
# Effects of low statistics




Large statistics




Low statistics




- Based on the ATLAS work (Measurement of $WH/ZH$ in $H \to b\bar{b}$, 13TeV with $139\text{fb}^{-1}$: arXiv:2007.02873)

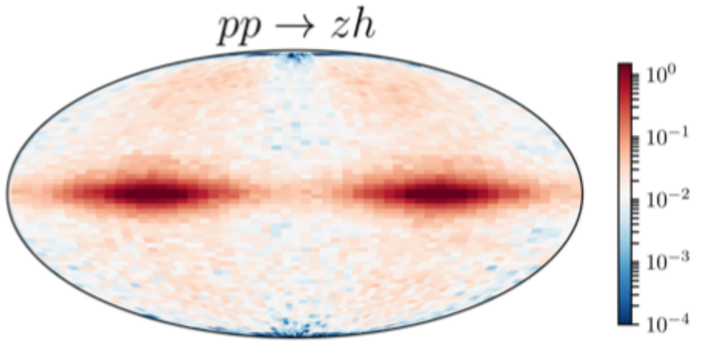
  Number of Higgs samples after selection cuts : 219
  Number of Higgs samples in the boosted region ($p_T > 250\text{GeV}$) : 18

$$\Lambda = \ln \frac{\mathcal{L}(\mathbb{H}_1)}{\mathcal{L}(\mathbb{H}_0)} = \ln \frac{\Pi_i P_{\mathbb{H}_1}(\vec{X}_i)}{\Pi_i P_{\mathbb{H}_0}(\vec{X}_i)} = \sum_{i=1}^{N_{\text{evt}}} \ln \frac{P_{\mathbb{H}_1}(\vec{X}_i)}{P_{\mathbb{H}_0}(\vec{X}_i)}$$

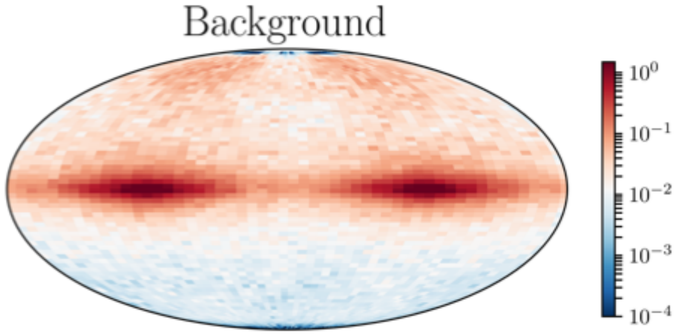
- With well-trained Neural Network, we may suffer from **"statistical fluctuation"** in the real battle of the LHC.

# Background rejection

- Accumulated 5000 events shot



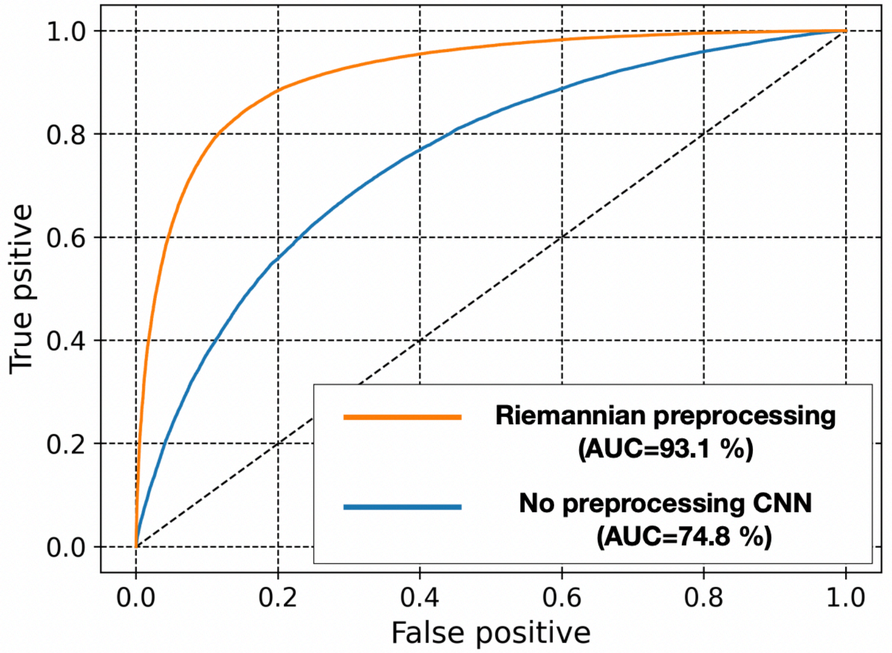

Corruptions in North hemisphere are from ISR / MPI QCD activities.




Combined:

Di-boson, Single top, Z+jets, tuba

- arXiv:2007.02873



# Conclusion

- Color flow analysis with full momentum range leads to low performance
- Cut the phase space to the boosted region can enhance the performance but leads to small statistics
- Riemannian mapping is a simple mapping by using inverse stereographic projection
- Riemannian mapping as data preprocessing can localize the hot cores and enhance the analysis performance on the full momentum range
- Riemannian mapping is a generic process and can be used for background rejection analysis