Measuring the QCD color structure of the Higgs particle

Myeonghun Park

(Seoultech)

Based on arXiv:2209.03898 (with Ahmed Hammad),
and on-going project (Kayoung Ban and K.C. Kong)

HPNP2023 - The 6th International Workshop on
"**Higgs** as a Probe of New Physics 2023"

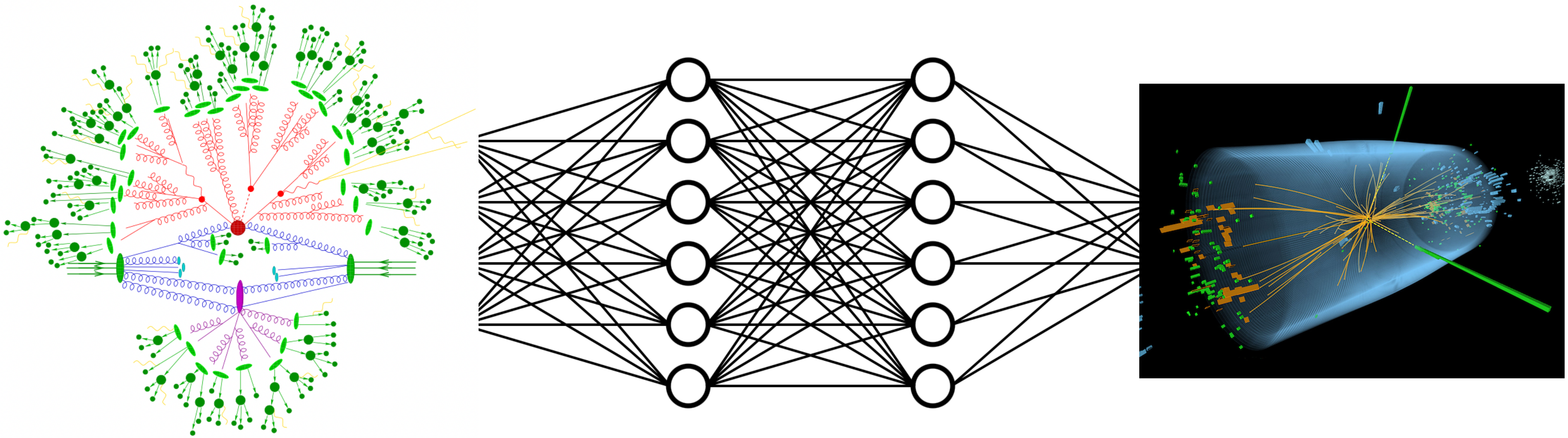
Tagging Higgs

- In this talk I will focusing on tagging $H \to b\bar{b}$.
  - **using a Neural Network**.

- This method is very universal, we can utilize this method to any "**color-singlet**" particle decaying into two QCD particles.

  - We can also "identify" color charge of the Higgs directly.
    (in the region of low backgrounds)

Enhancing signatures over BKG



- With our elaborated **theoretical model**,

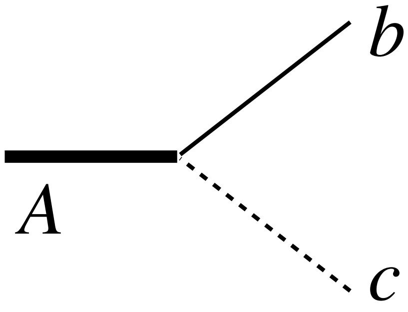
  1) Get **expectations** from **MC simulations**

  2) Get **data** from **experiments** (e.g. the LHC)

  3) Compare our expectation to data with sophisticated computer **algorithms (ML: machine learning)**

Extracting features of a new physics

- **Kinematic variables** to utilize a **different phase-space** structures (signal, v.s. backgrounds)



(invisible)

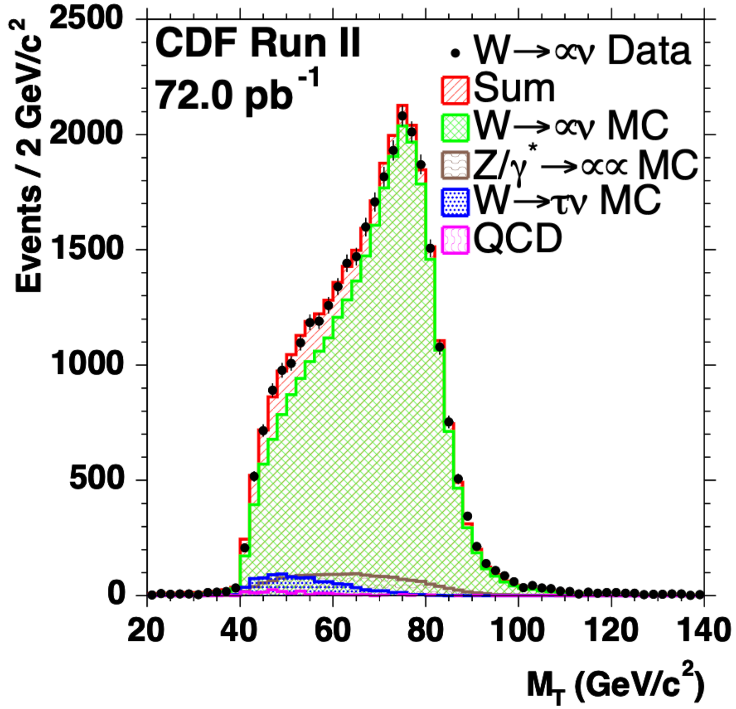
$$\theta = \{m_A\} \longrightarrow X = \{p_b^\mu, p_c^T\}$$

$$\dim(X) = 6 \rightarrow \dim(V) = 1$$

- A human-engineered feature variable, M_T which estimates M_A with an endpoint of its distribution

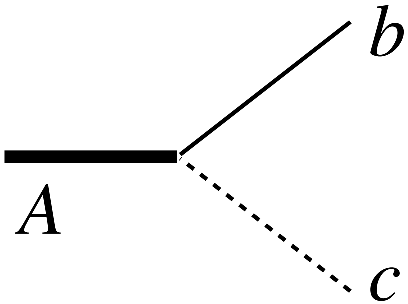
  (highly singular behavior due to its Jacobian peak)

Tao Han's collider pheno lecture



Extracting features of a new physics

- **Kinematic variables** to utilize a **different phase-space** structures (signal, v.s. backgrounds)

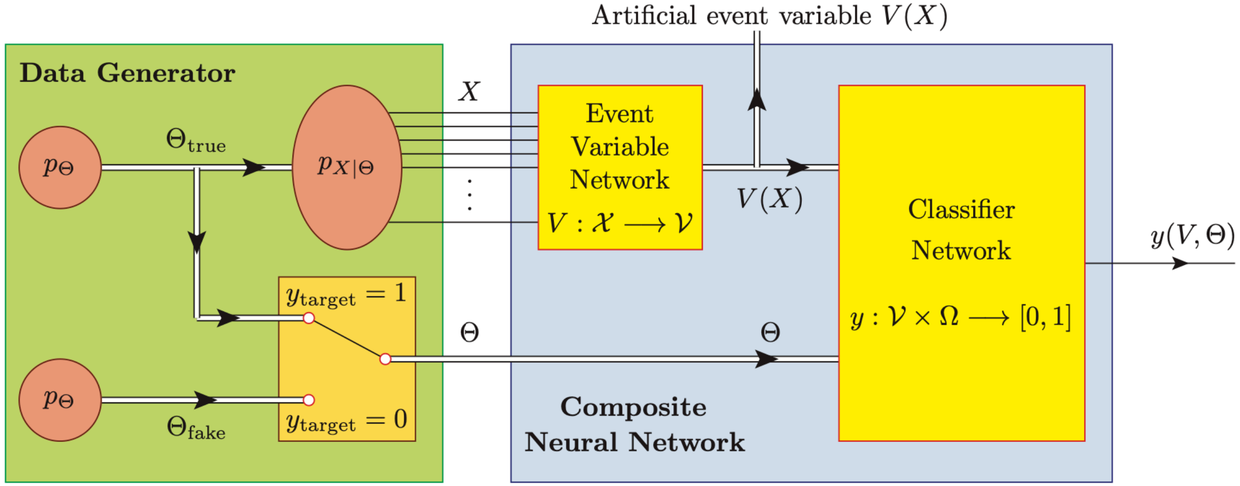


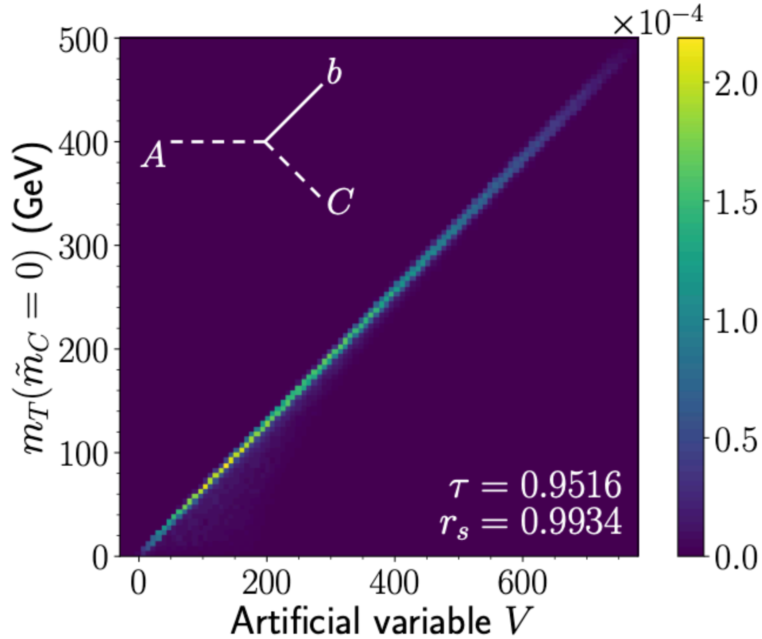
(invisible)

$$\theta = \{m_A\} \longrightarrow X = \{p_b^{\mu}, p_c^{T}\}$$

$$\dim(X) = 6 \to \dim(V) = 1$$

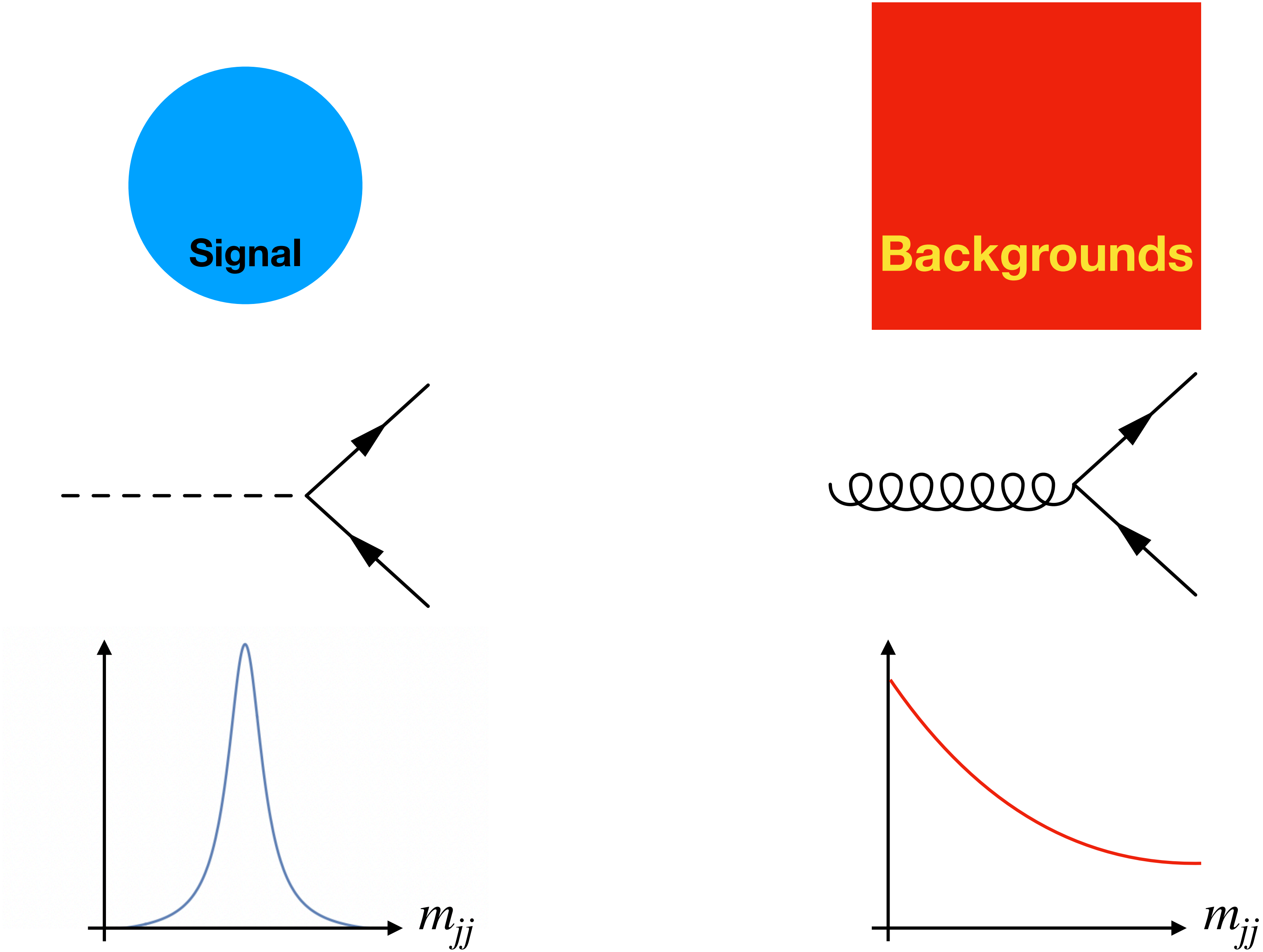
A Neural-Network can design an event-variable
(by enforcing information-bottleneck to NN)



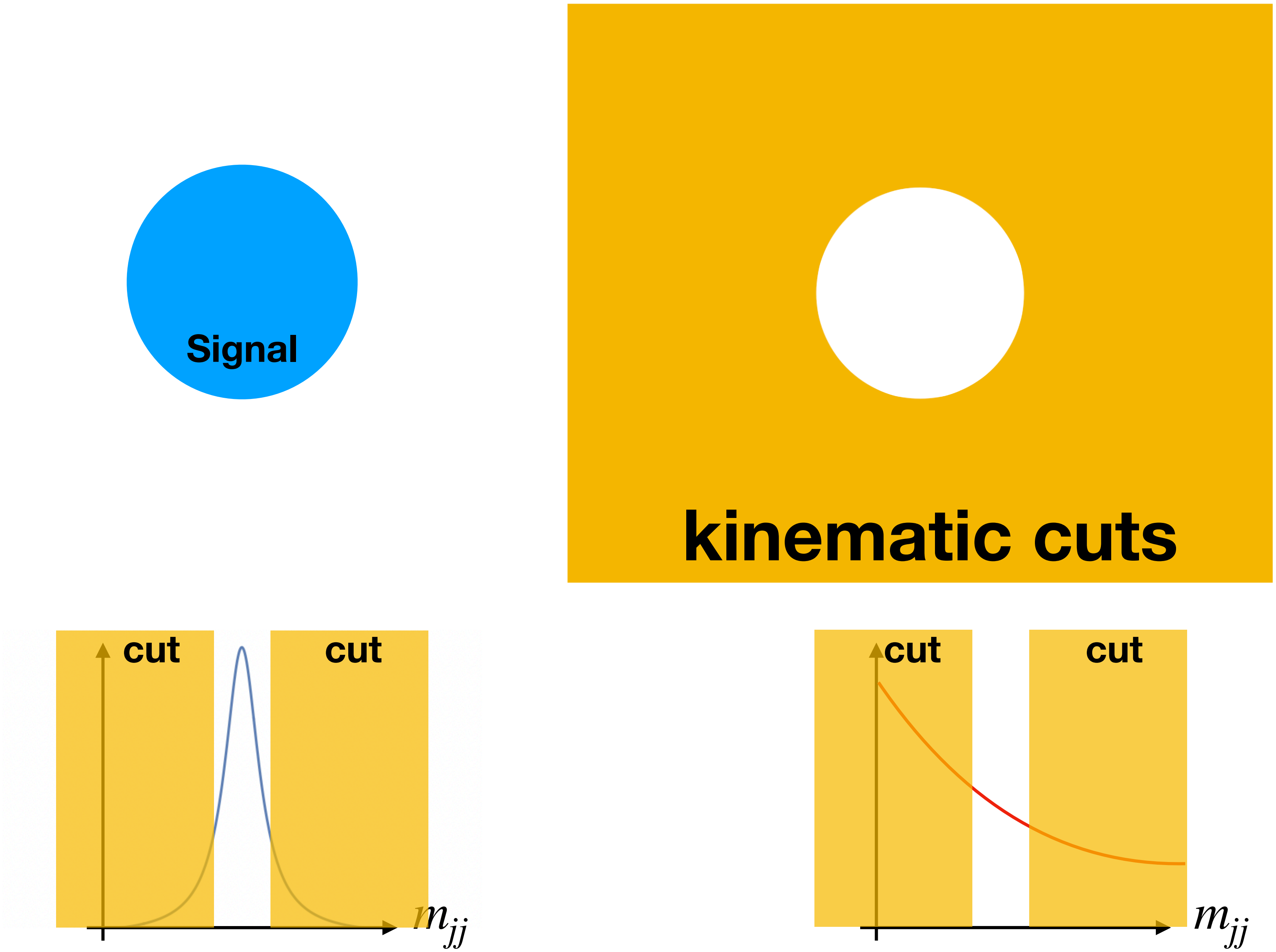


Doojin Kim, KC Kong, Konstantin Matchev, Prasanth, **MP**. (2023)

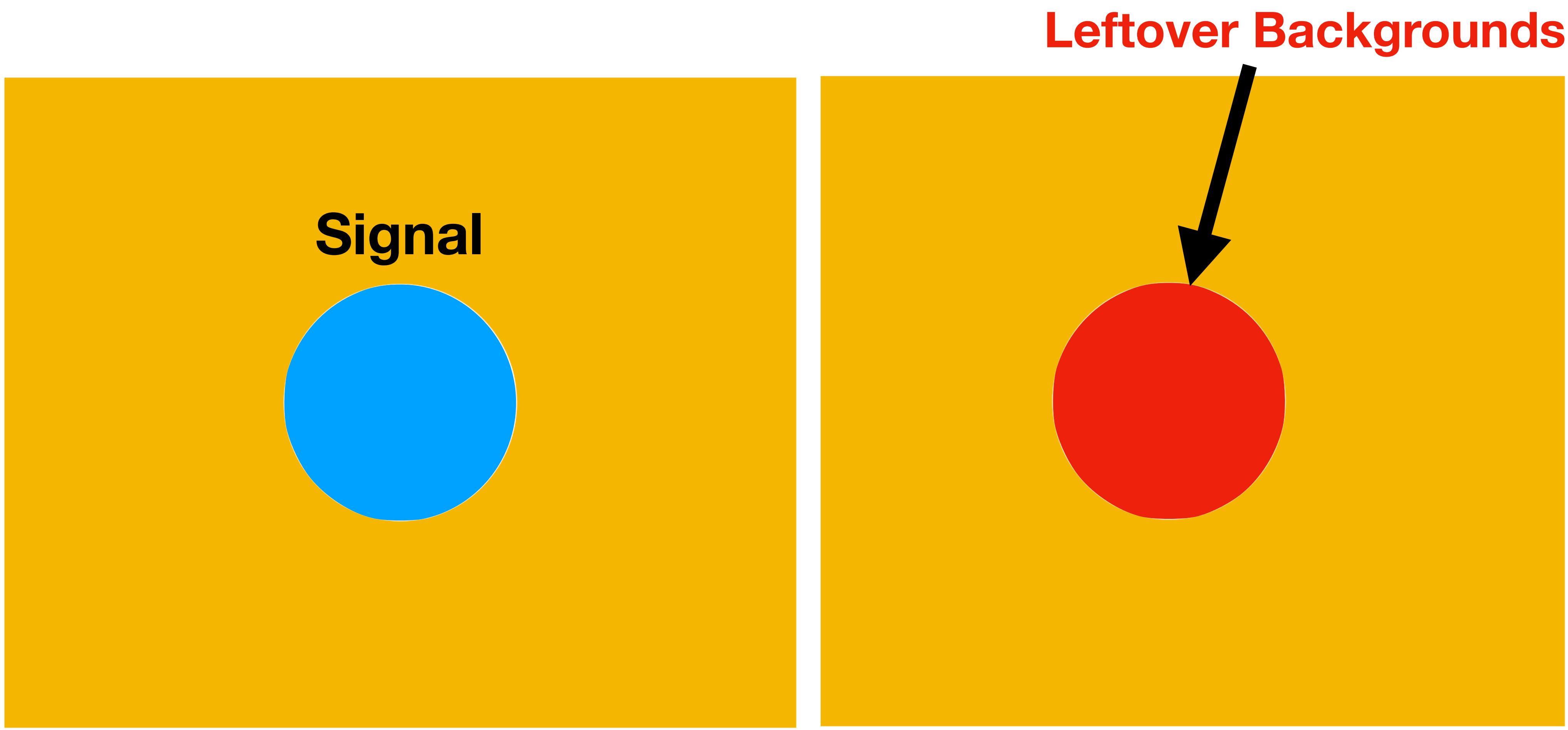
Basic idea of Kinematic cuts



Design Kinematic cuts to reduce BKG while leave signals as many as we can



We are shaping backgrounds into signals...?



Leftover Backgrounds become very similar to signals

(Similar phase-space part of BKG would remain)

The problem is....."BKG is Huge"

Actual problem is that the size of remaining BKG is
huge compared to SIG

Orthogonal information to the Kinematics

- Differences in kinematics are from **"high P_T" region, i.e. reconstructed (reco) level**

  - Telling us about the structure of "Feynman-diagram"
    (Event-topology, Mass spectrum)

- We can further utilize $|\mathscr{M}|^2$ differences (Density bounded by phase-space)
  - e.g.) Decaying angle of the Higgs

- **Differences in QCD radiation patterns** are from **"soft P_T" region**

  - Telling us about the state under a gauge group, $SU(3)_C$

More than Kinematics difference

- In many cases, the **soft QCD radiation patterns** from signals are different from Backgrounds. (e.g. : rapidity gap)

Jason Gallicchio, Matthew D. Schwartz 2010

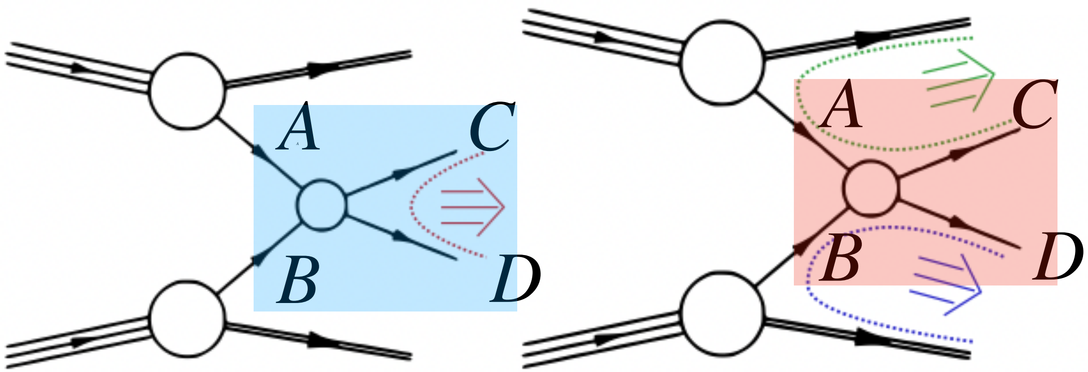
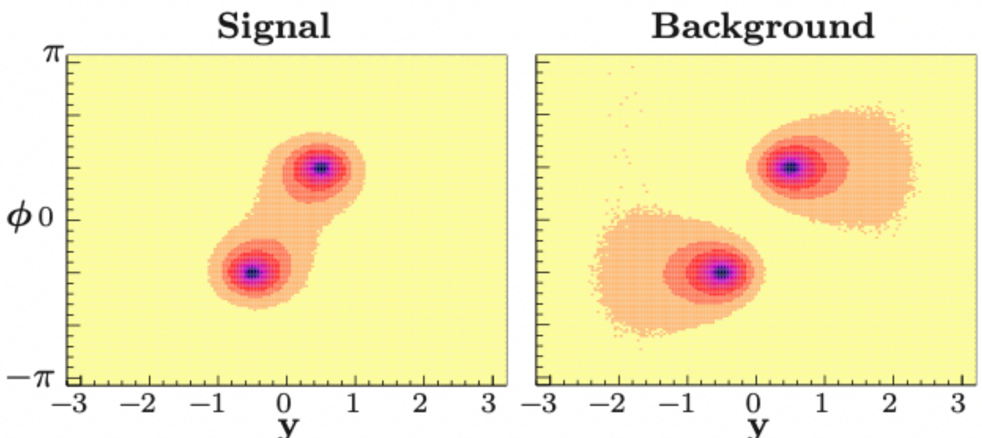


FIG. 1: Possible color connections for signal ($pp \to H \to b\bar{b}$) and for background ($pp \to g \to b\bar{b}$).



$$gg \to h \to b\bar{b}$$

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

V . S .

$$gg \to b\bar{b}$$

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

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

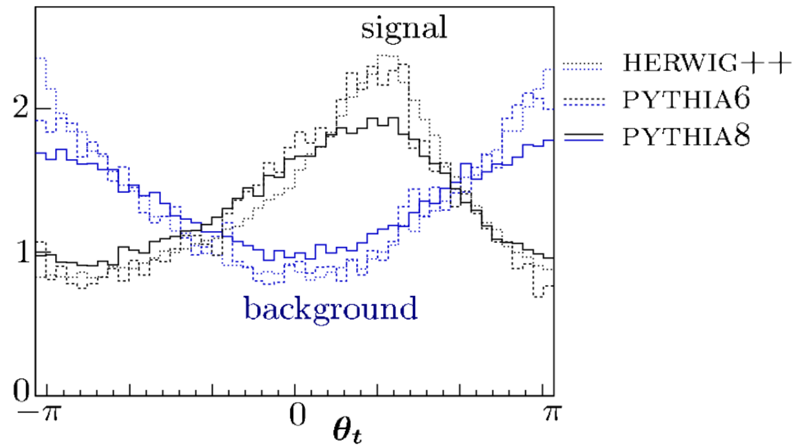
Utilizing QCD information

- One can design a QCD variable, for example a pull-vector

$$\vec{t} \equiv \sum \frac{p_T^i |r_i|}{p_T^{\text{jet}}} \vec{r}_i$$

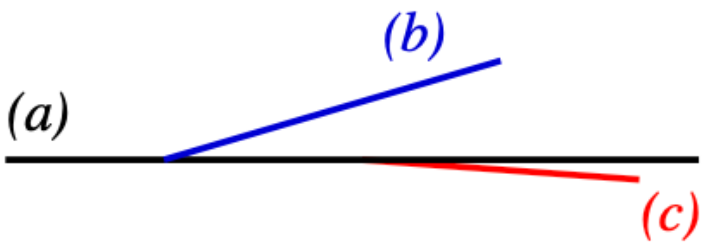
(Jason Gallicchio, Matthew D. Schwartz 2010)

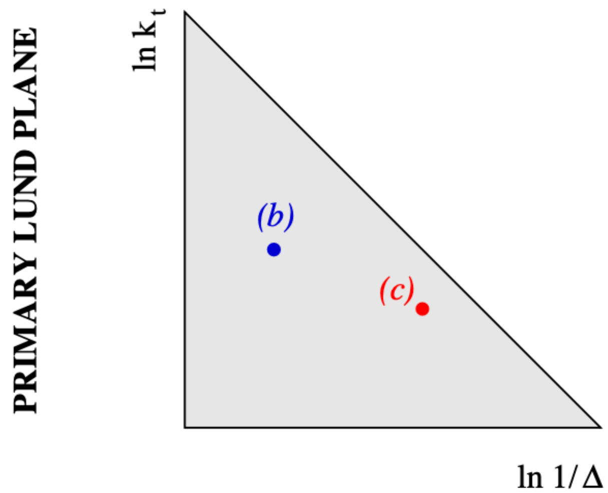
provides an one-dimensional feature



- Or one can get two-dimensional features,

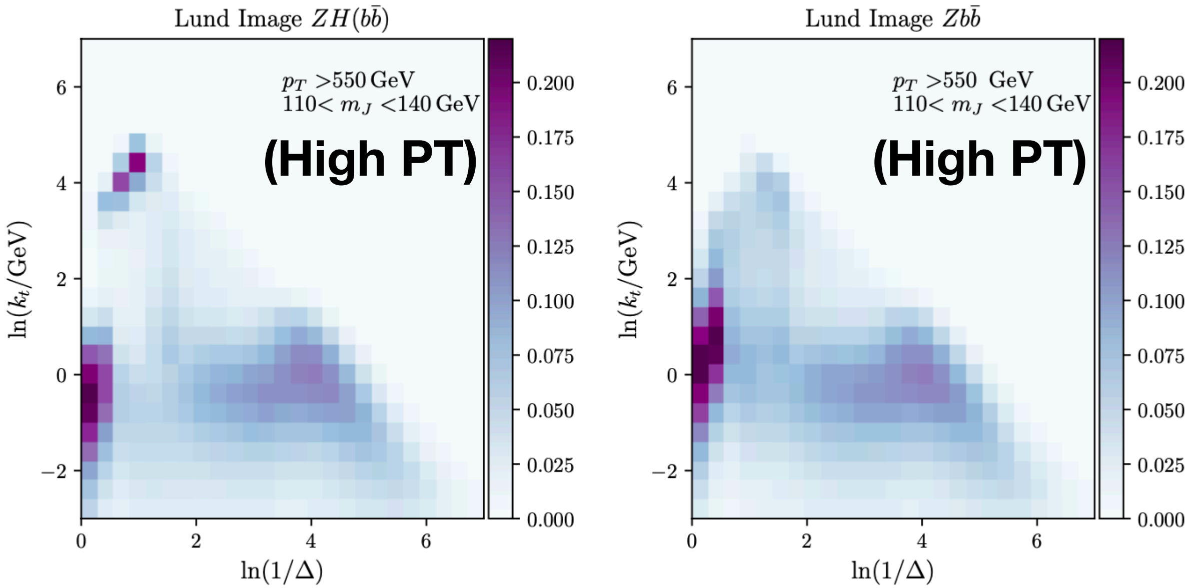
(Frederic A. Dreyer, Gavin P. Salam, Gregory Soyez 2018)



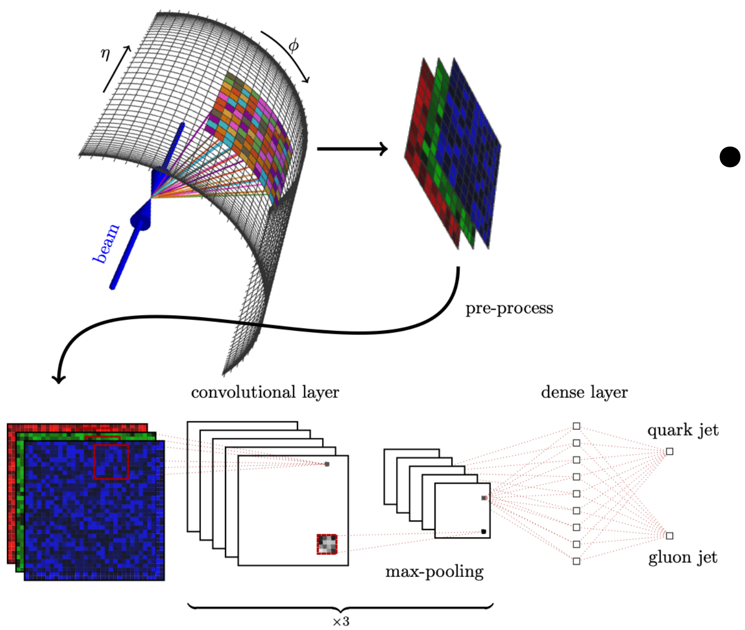


Fully utilizing QCD information?

- One needs to understand differences in **"image"**



(Charanjit K. Khosa, Simone Marzani, 2021)



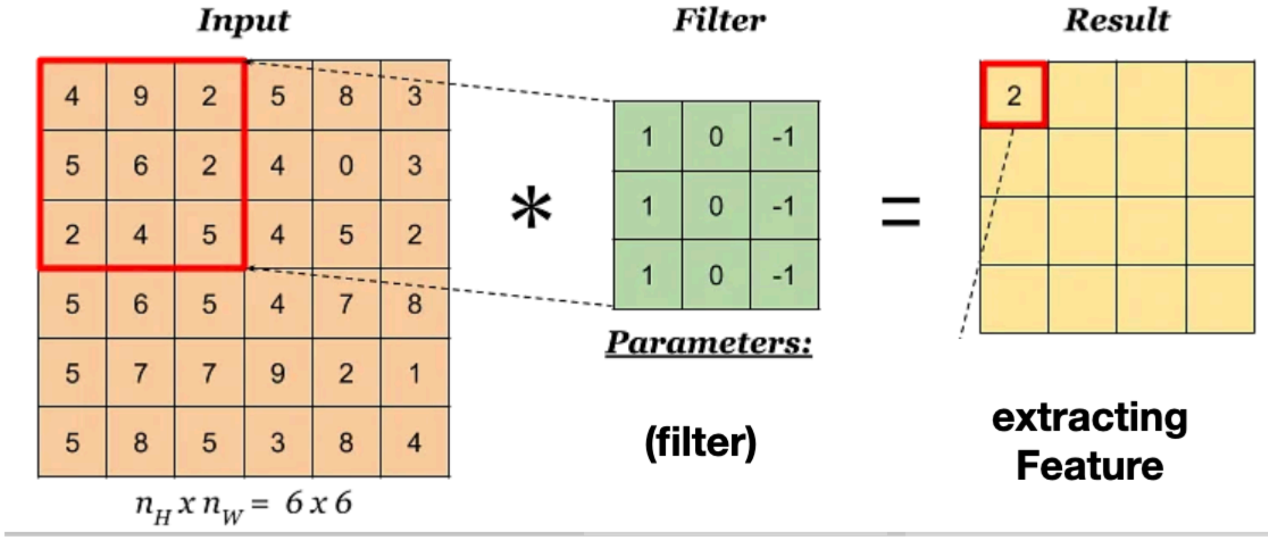
- A **neural network** (designed to understand a **picture**) can tell differences in QCD

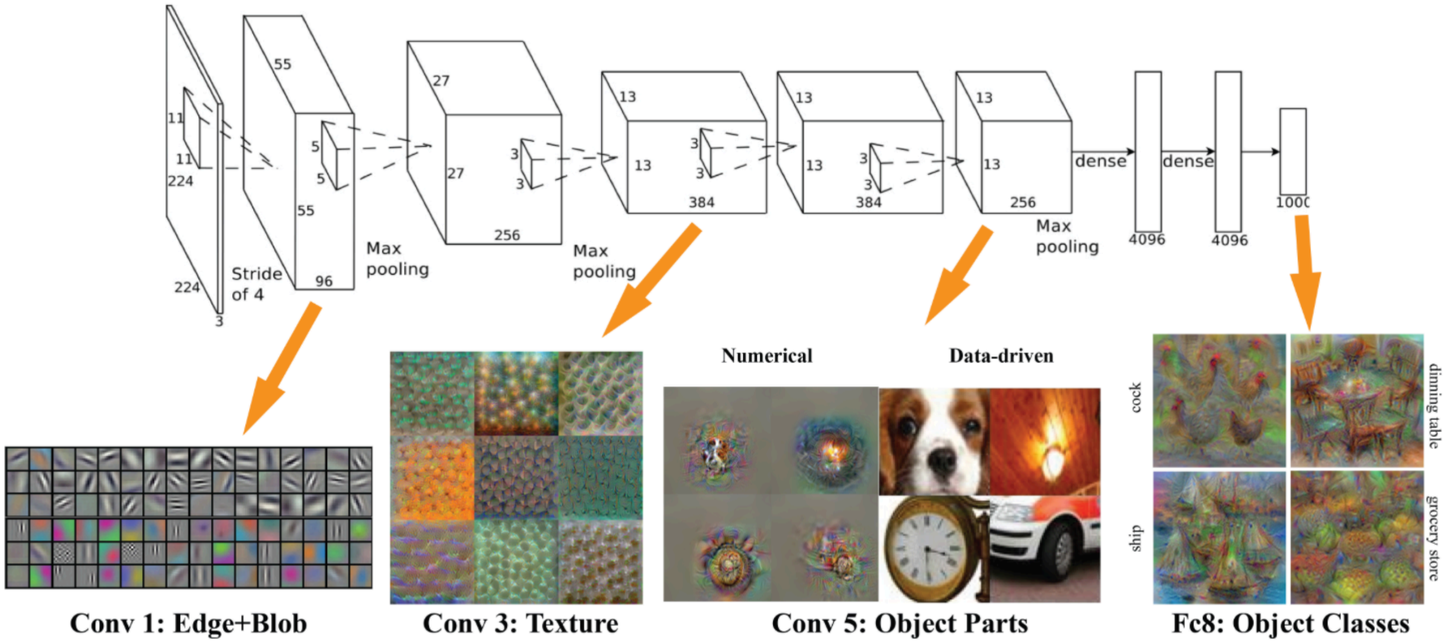
  - Pixels are energy deposits from various sub-detectors (e.g. : tracks, e-cal, h-cal)

M. Schwartz et.al. arXiv:1612.01551

Conventional Image recognition

(applying series of filters)



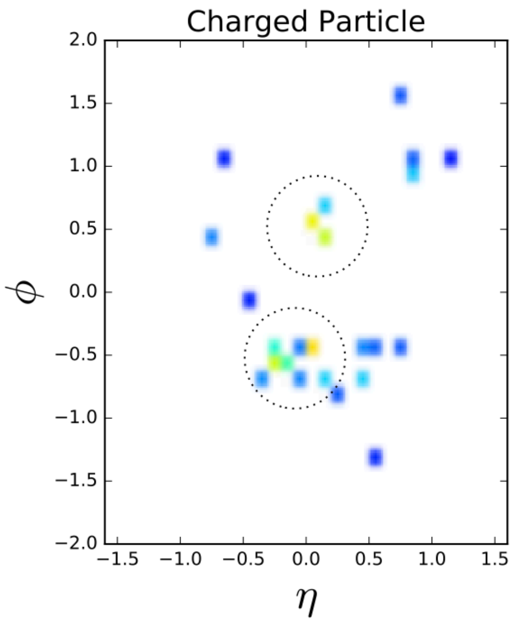


Basic structure of Image recognition using CNN

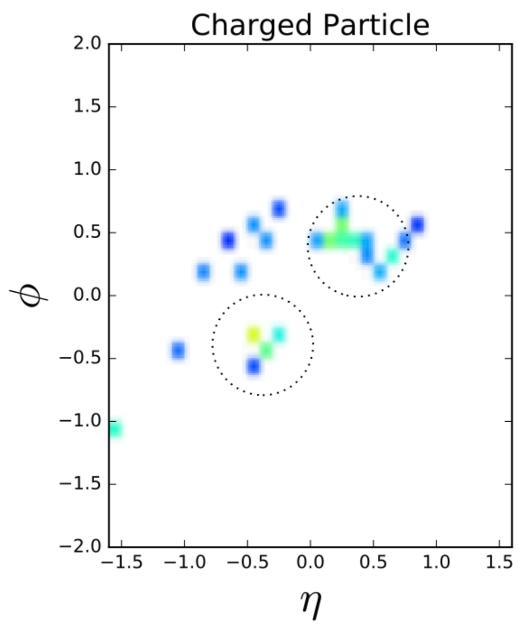
The problem of conventional ML (1)

- The direct use of a neural network **(designed for commercial image)** is **not suitable (= not efficient) to our physics cases.**

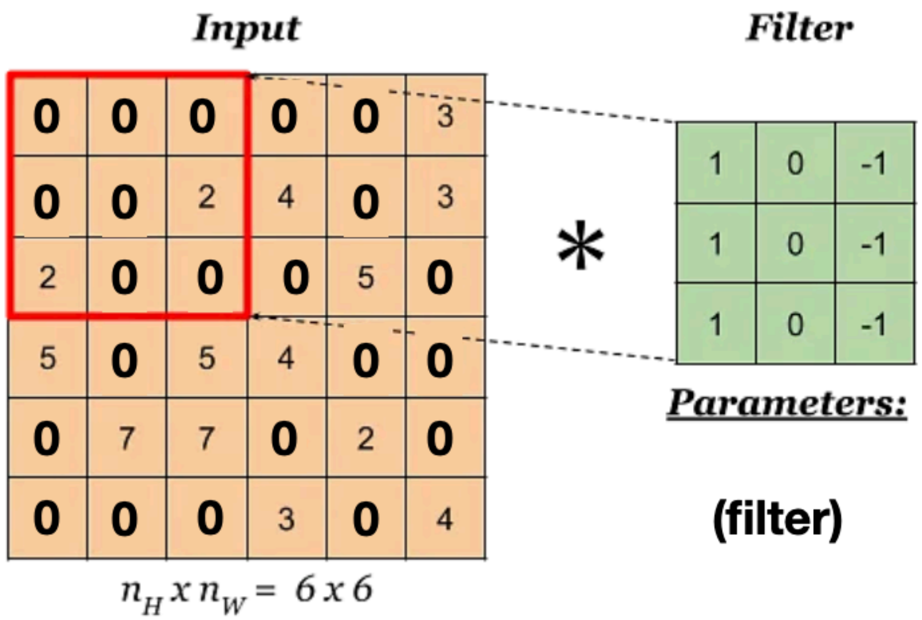
- The "image" from our LHC data is very **sparse**



$h \to b\bar{b}$



$b\bar{b}$
in Backgrounds

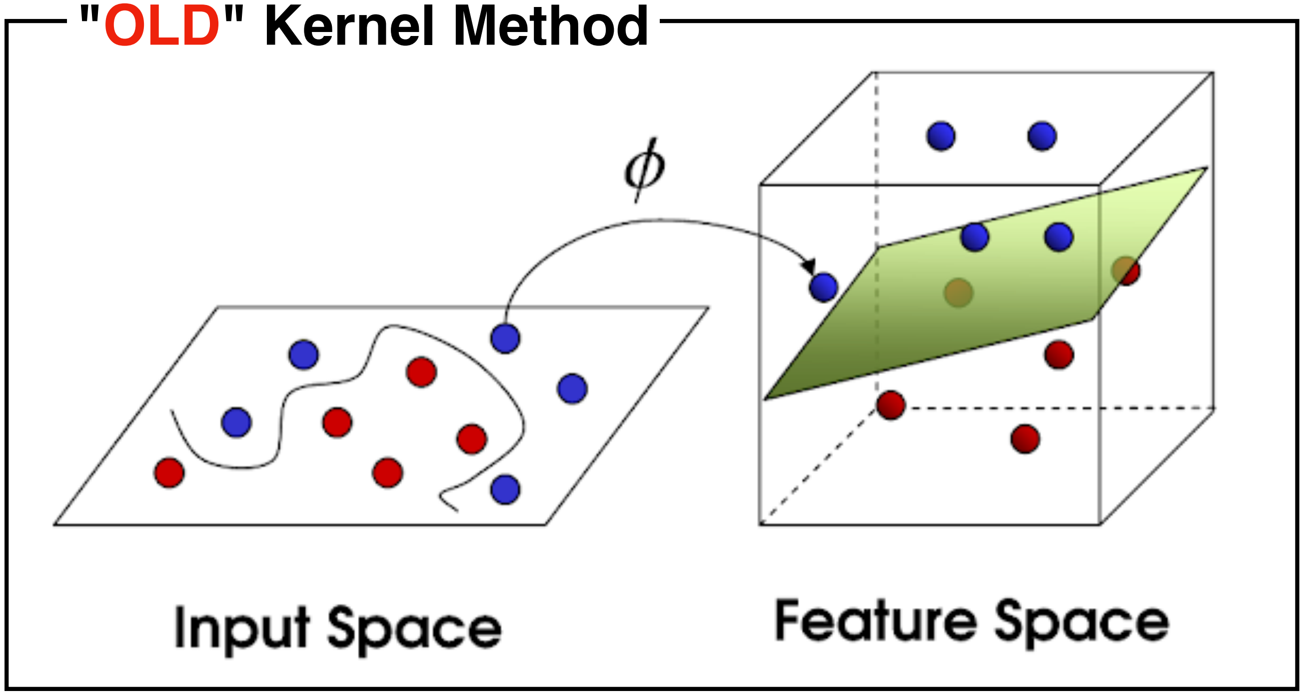


Mostly empty image (value = 0)
Any meaning to apply a "filter"?

One solution: Kernel Method

- We can provide a good kernel to separate data **efficiently,** namely with a few and sparse "image" data by making **"linearly" separable**

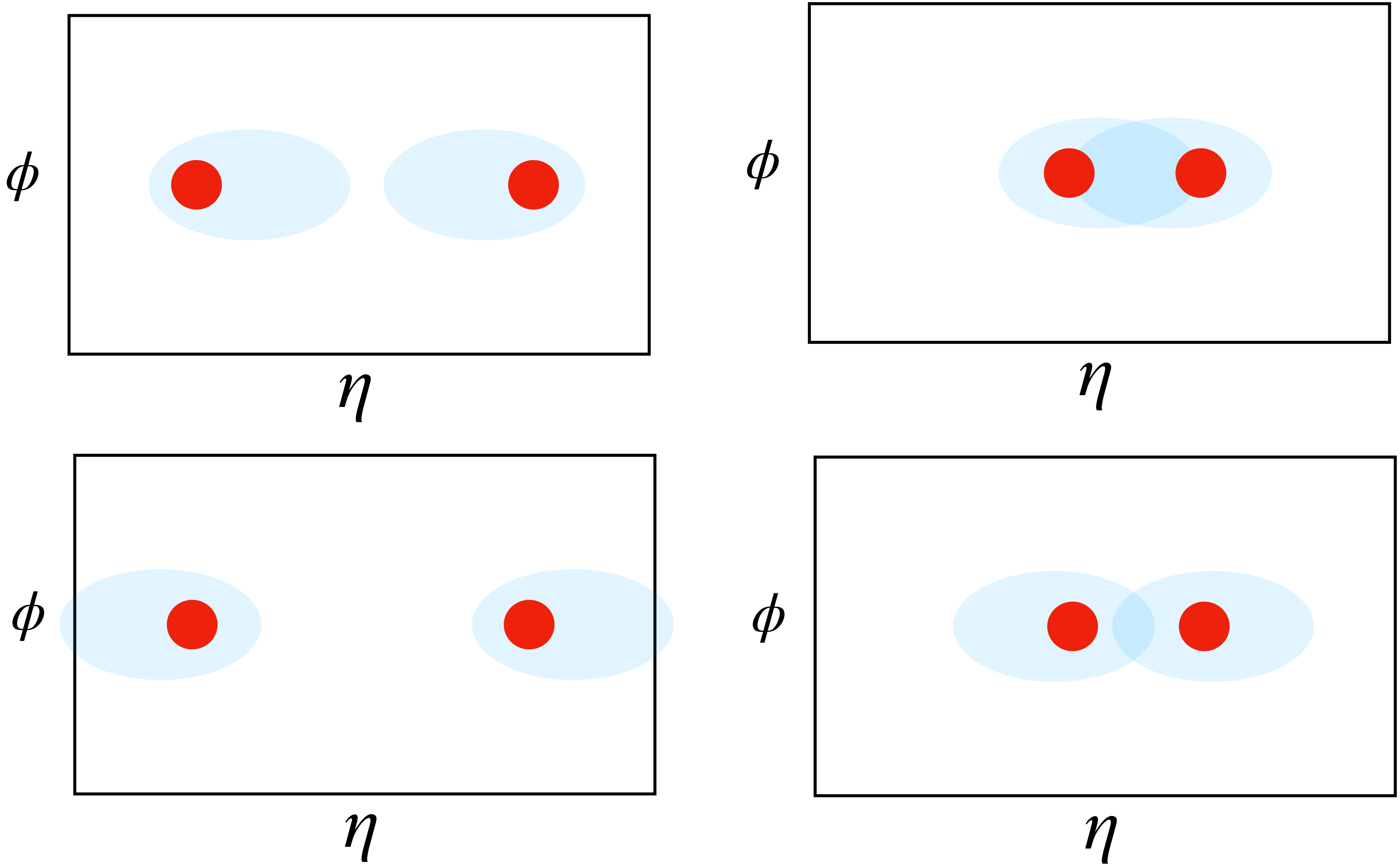
"OLD" Kernel Method



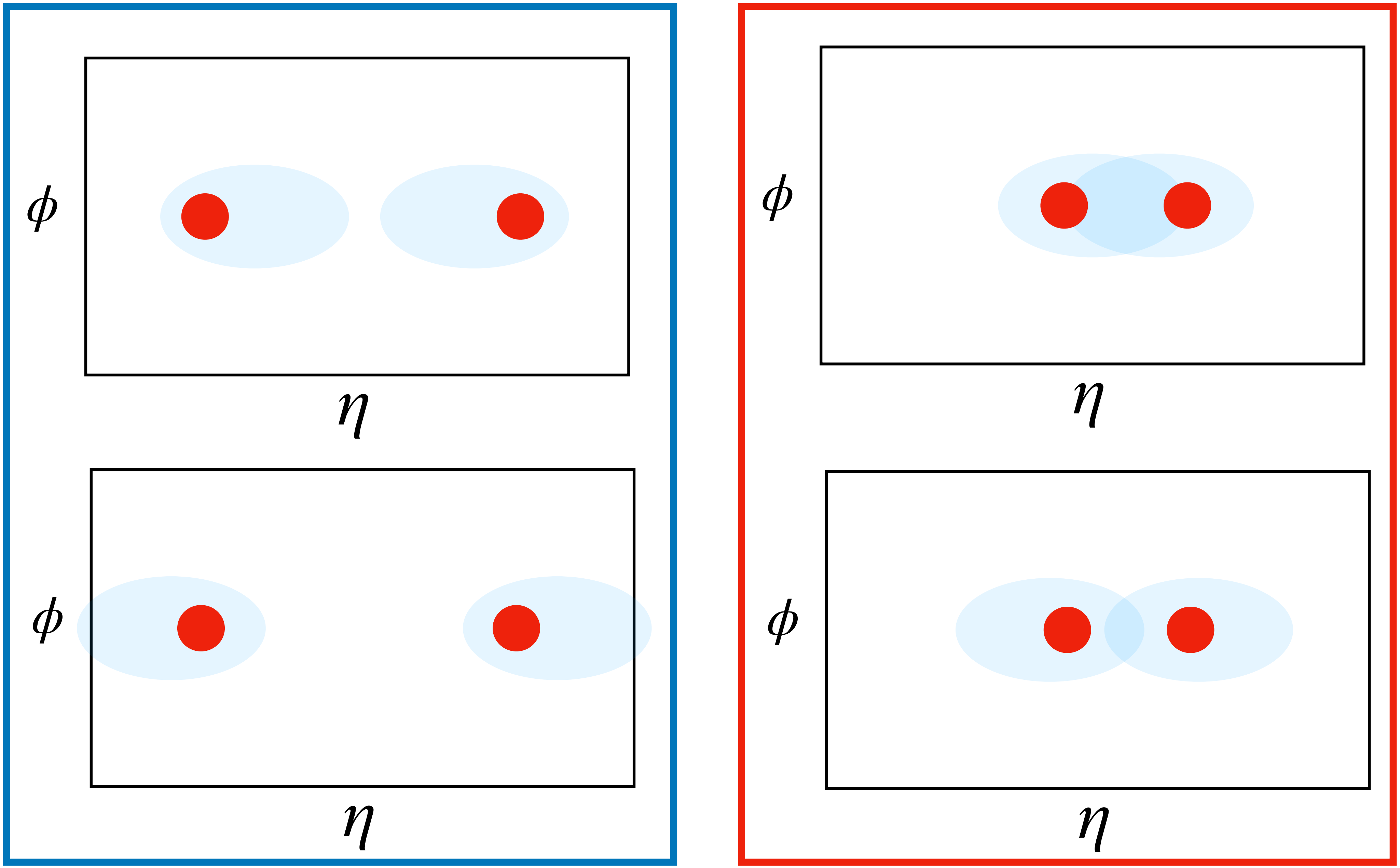
- Designing a kernel requires a domain knowledge (based on our expertise)
  - This means "**old**"
  : **Conventional ML : end - to - end (**Blackbox): No human intervention

The problem in data (2)

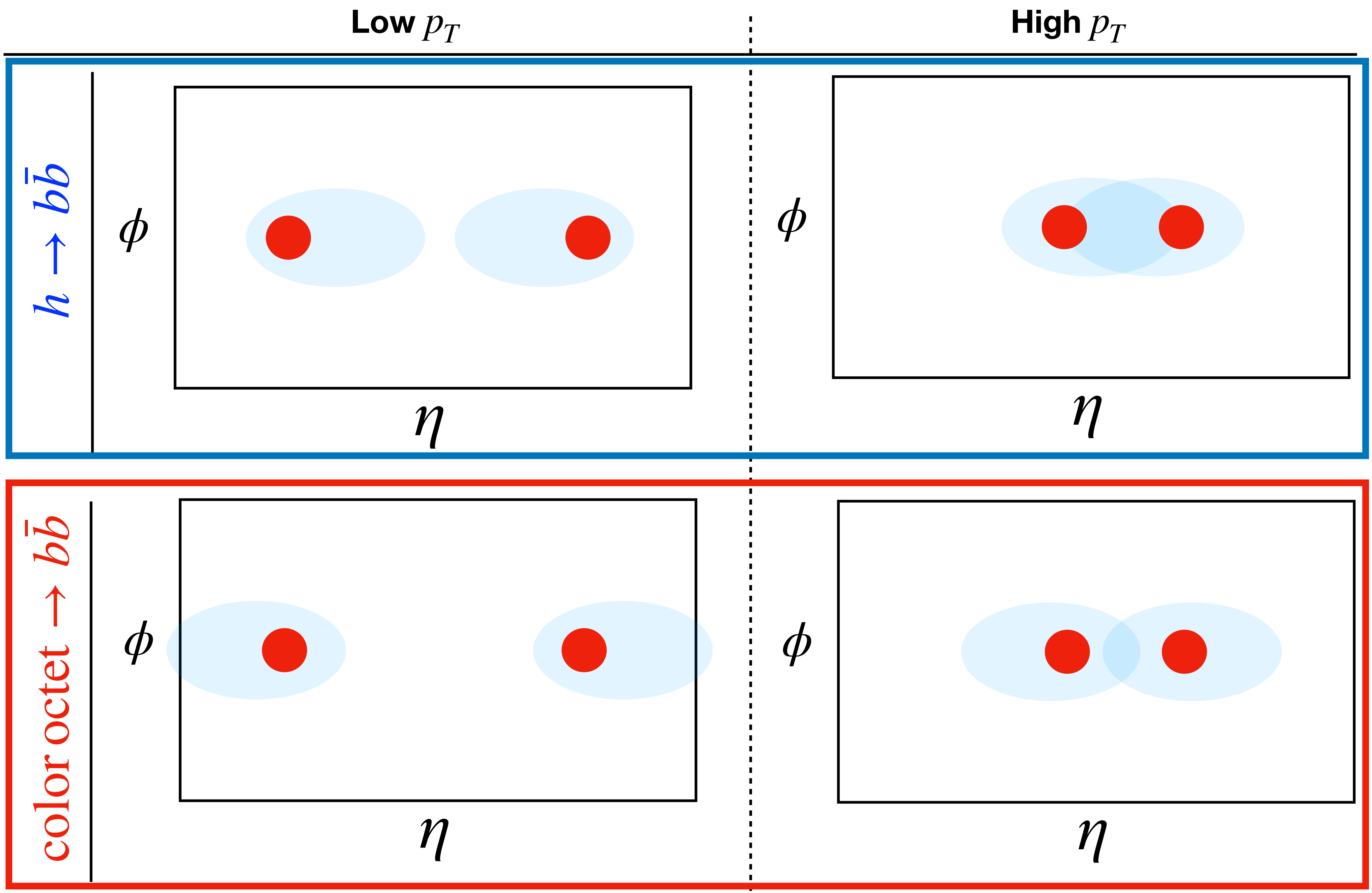
- Let's make two categories : Divide below into two categories

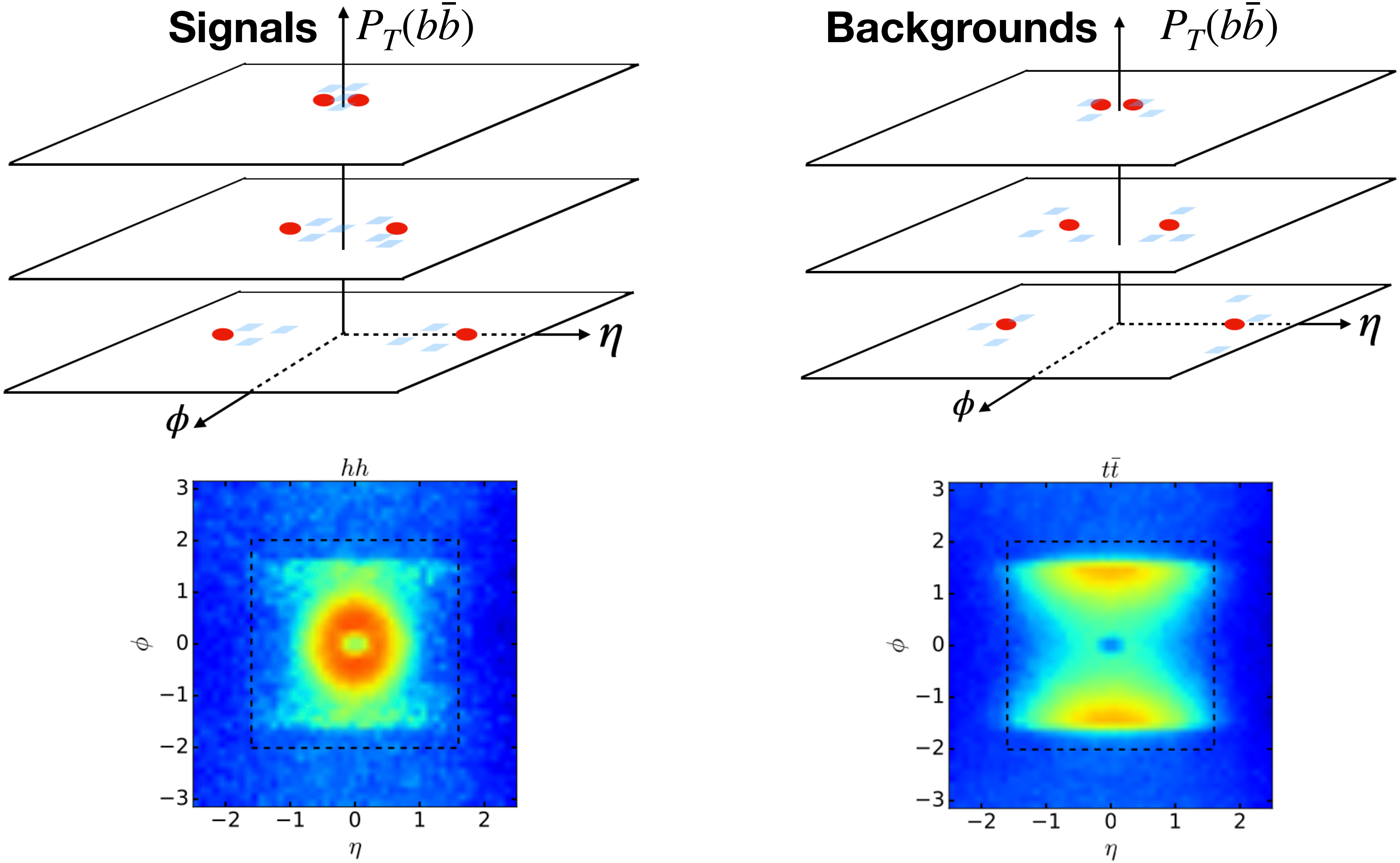


- **A quick trial: Attentions are on hot cores**



- Due to the **softness of radiations**, everyone (even ML) gets focused on **hot cores** ($b/\bar{b}$)

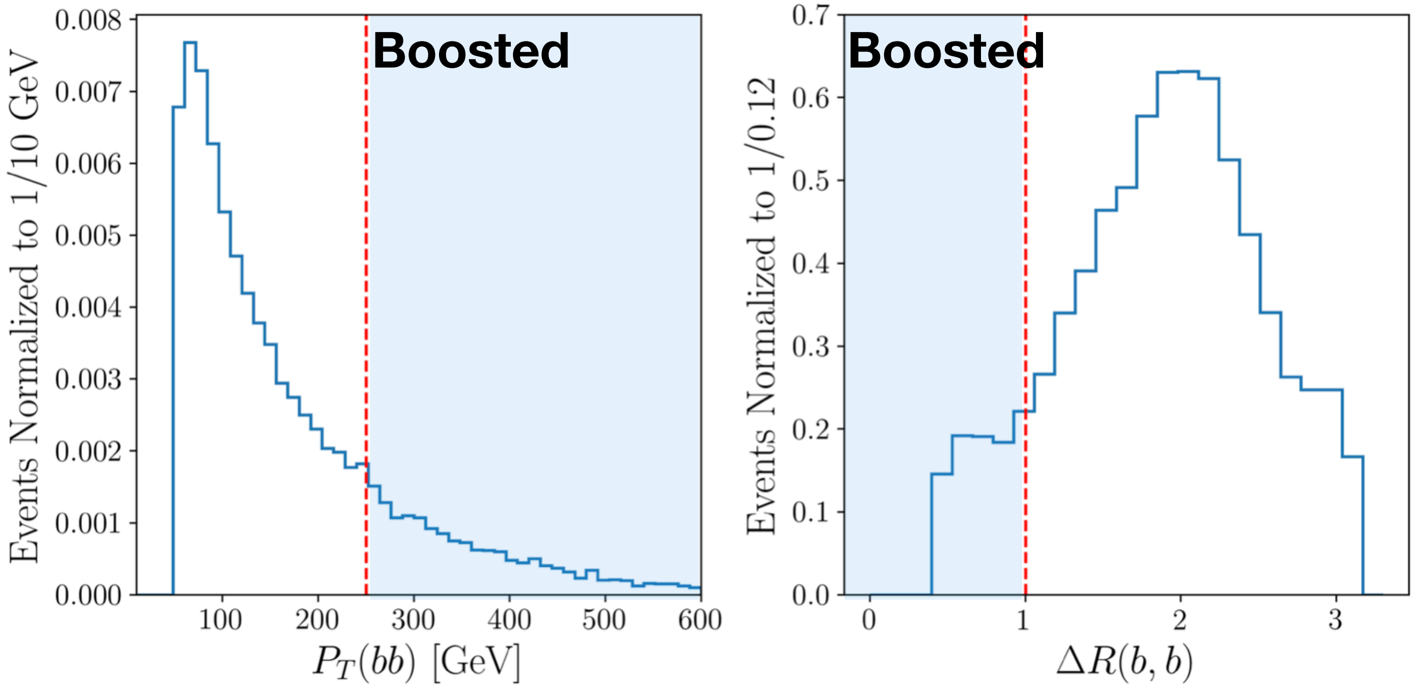


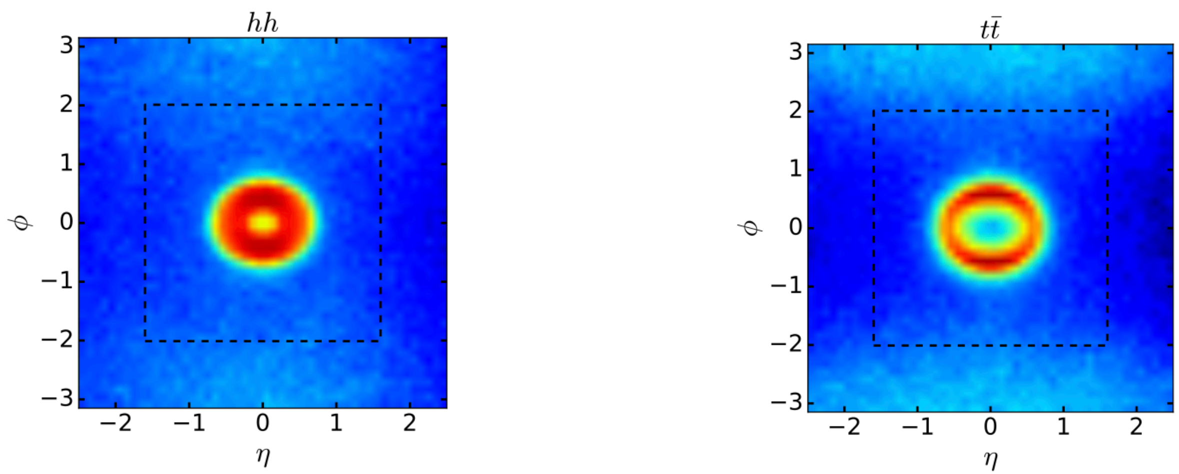


- Conventional Machine Learning can not focus on soft-patterns, **rather on different kinematics.**

- **It requires "BIG" data** to pay attention to **soft patterns.**

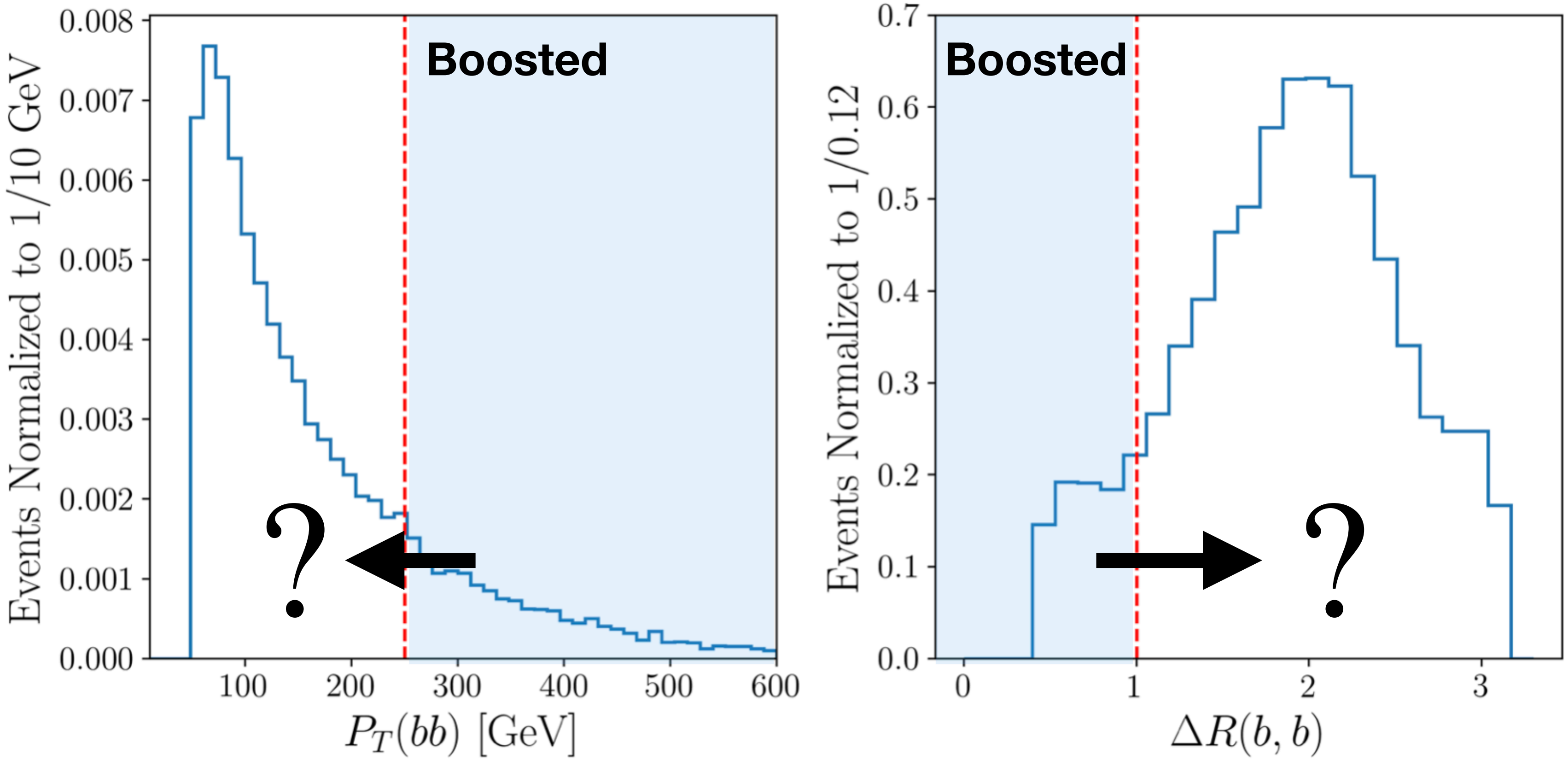
One solution: focusing on small region





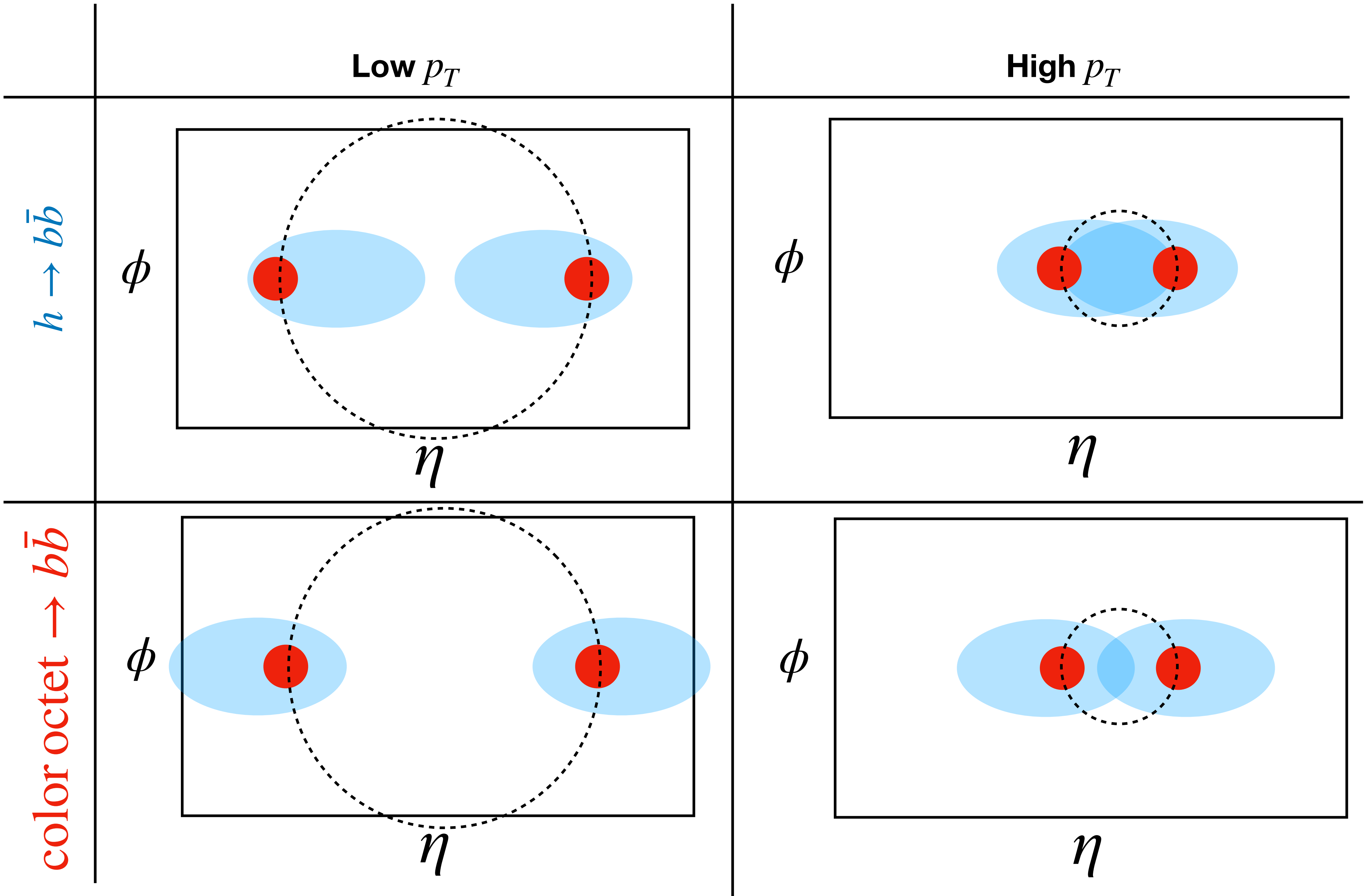
- In a boosted region, the dependency on "kinematics" becomes mild

Want to use "**Full p_T range**" of "Higgs"



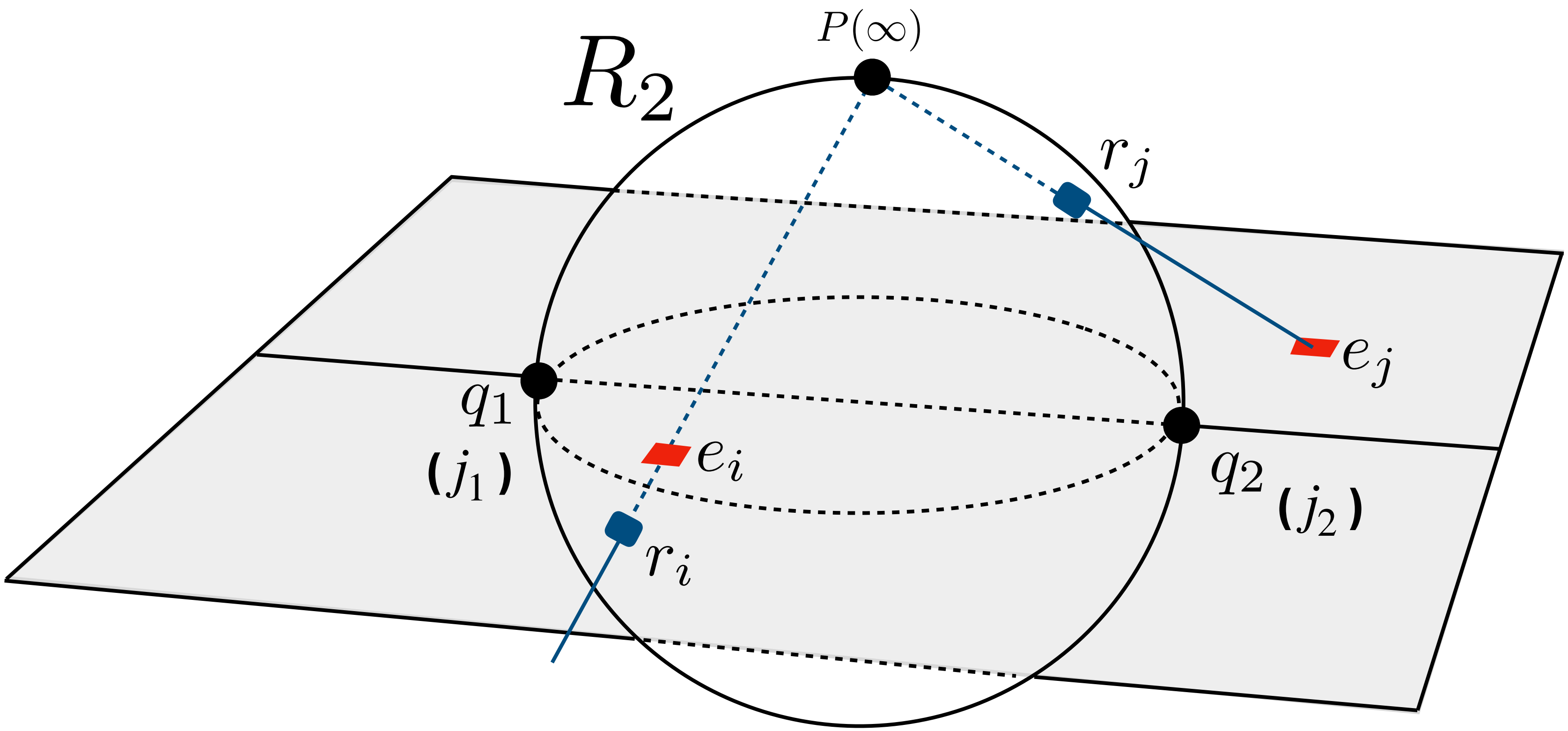
- "Easy" solution demands a **huge price**: the **statistics**. We want **to collect more statistics !**

- A **binary problem**, either "**inside**"or "**outside**" a circle.



Inverse stereographic projection

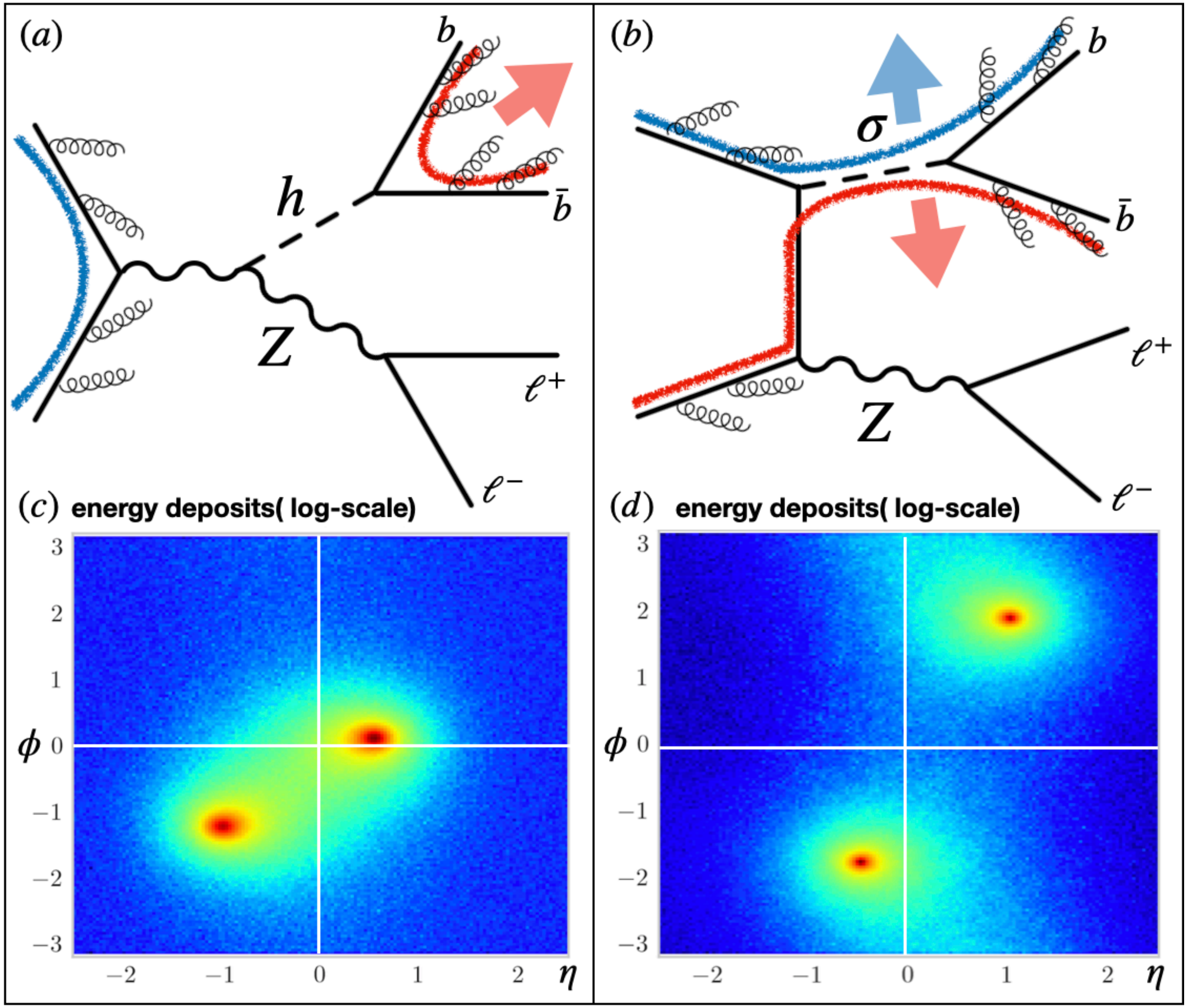
(we call "Riemannian" Kernel)



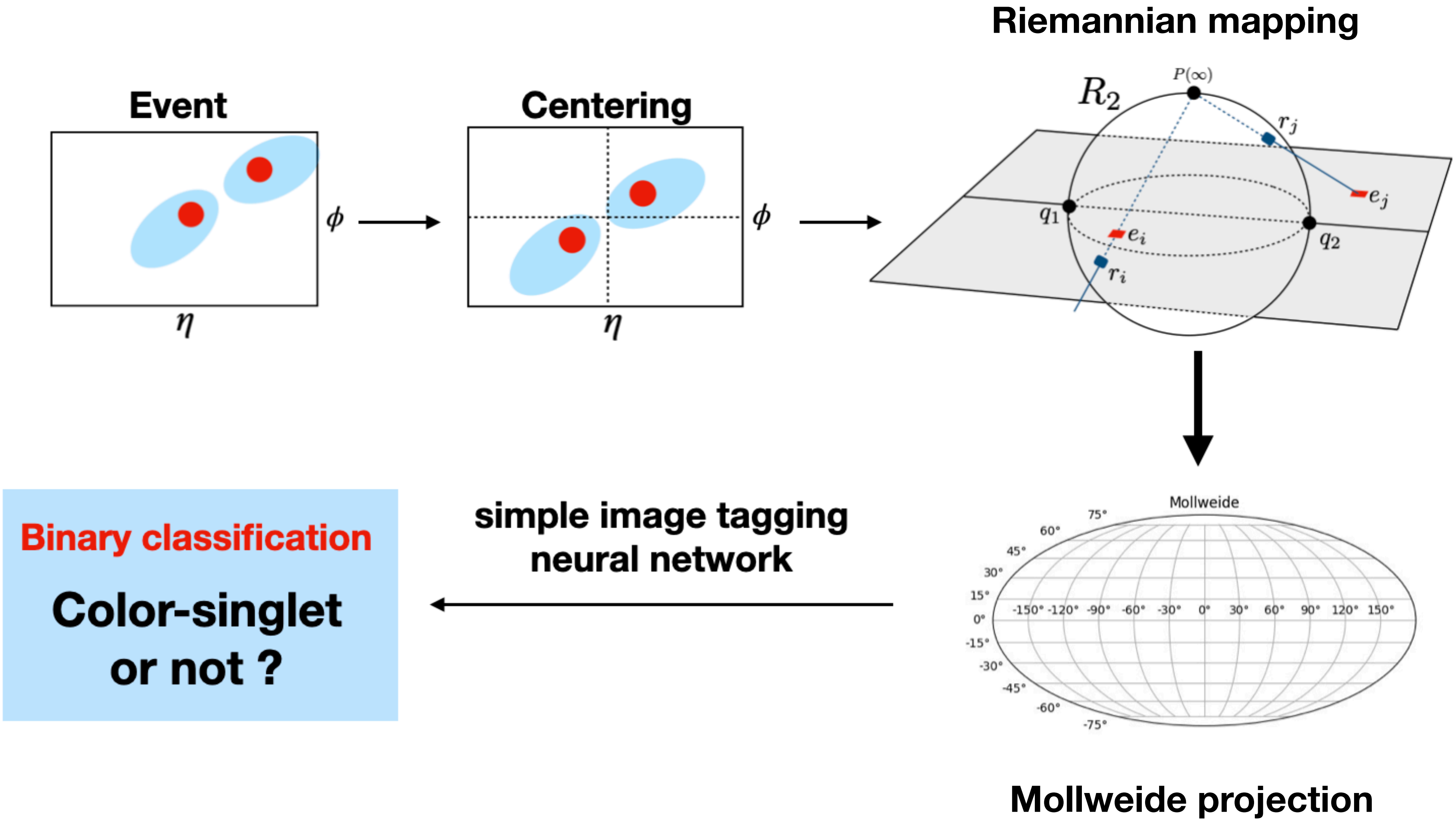
- Soft radiations which are
  inside of a circle → Southern hemisphere (H)
  outside of a circle → North hemisphere (Color octet status)

- Consider **only angular positions**, **totally** **independent from a radius which is proportional to $P_T(jj)$**.

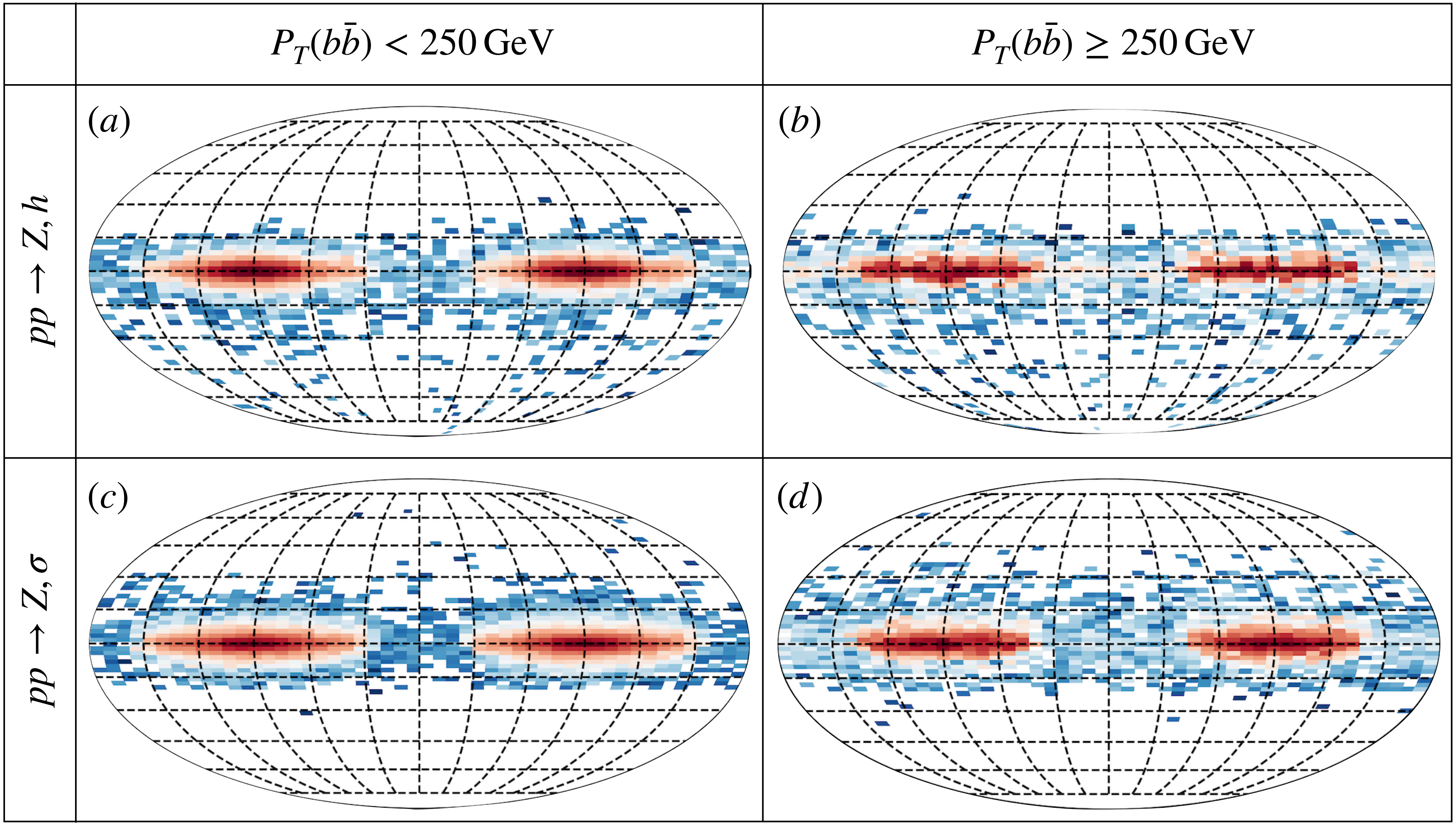
- A toy model of color octet "scalar" particle with $m_\sigma = m_h$ to focus on checking the performance on "QCD". (Also QCD backgrounds, $pp \to (g \to b\bar{b}), Z$ is in this case)



Riemannian preprocessing



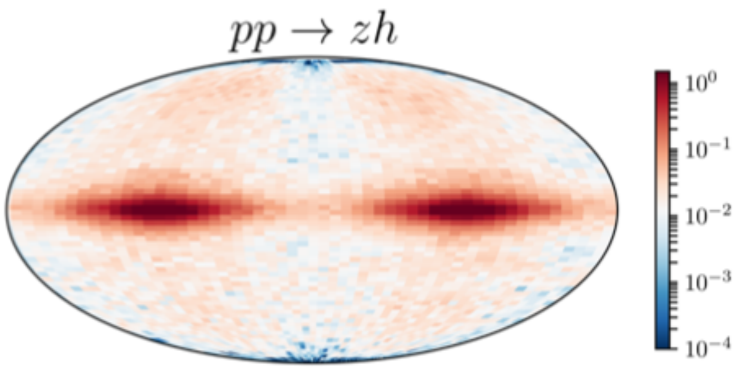
Mollweide projection of Riemannian preprocessing

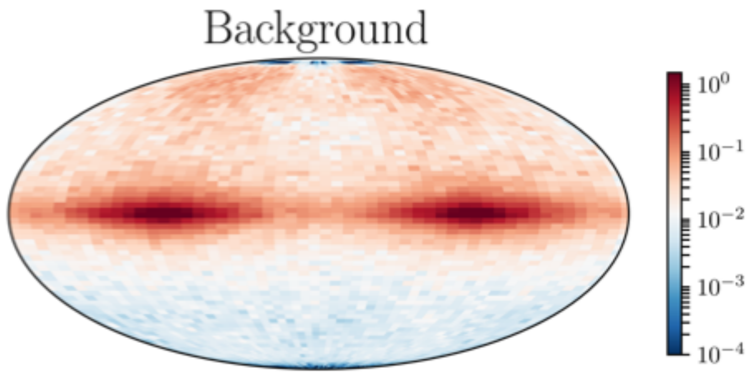


- The distribution of soft patterns does not show a dependency on $P_T(b\bar{b})$

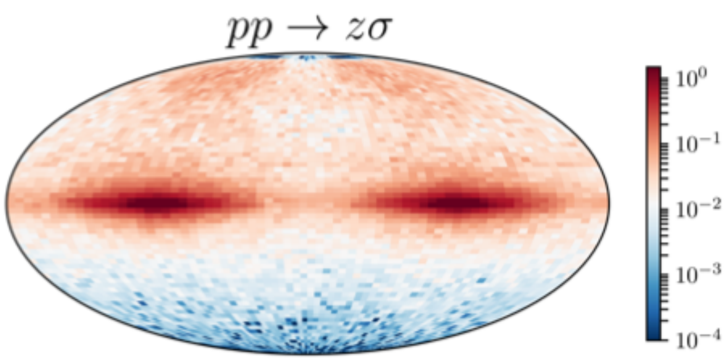
Landscape of Color activity

- Accumulated 5000 events shot



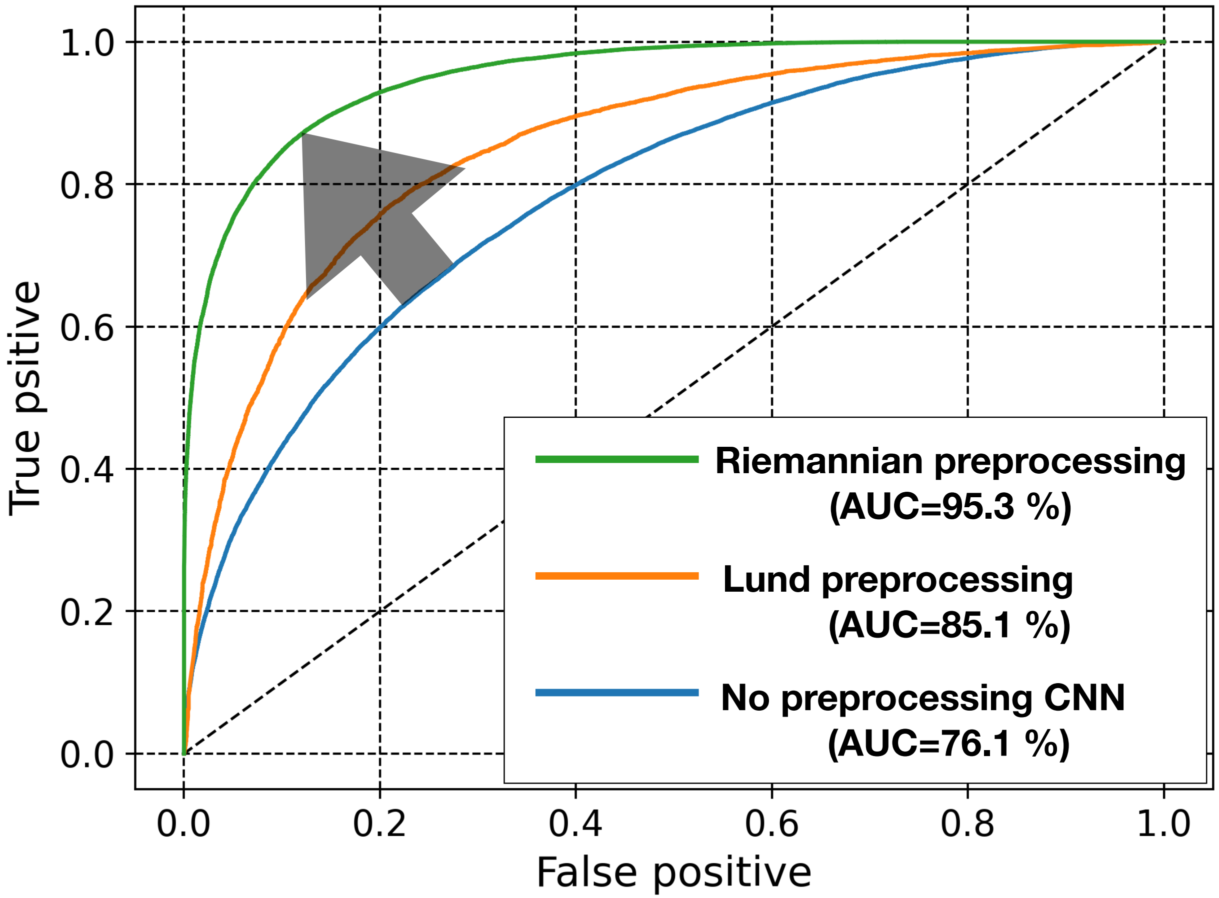


($V + js$, $t\bar{t}$, $t(\bar{t})$, VV, $\cdots$)



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

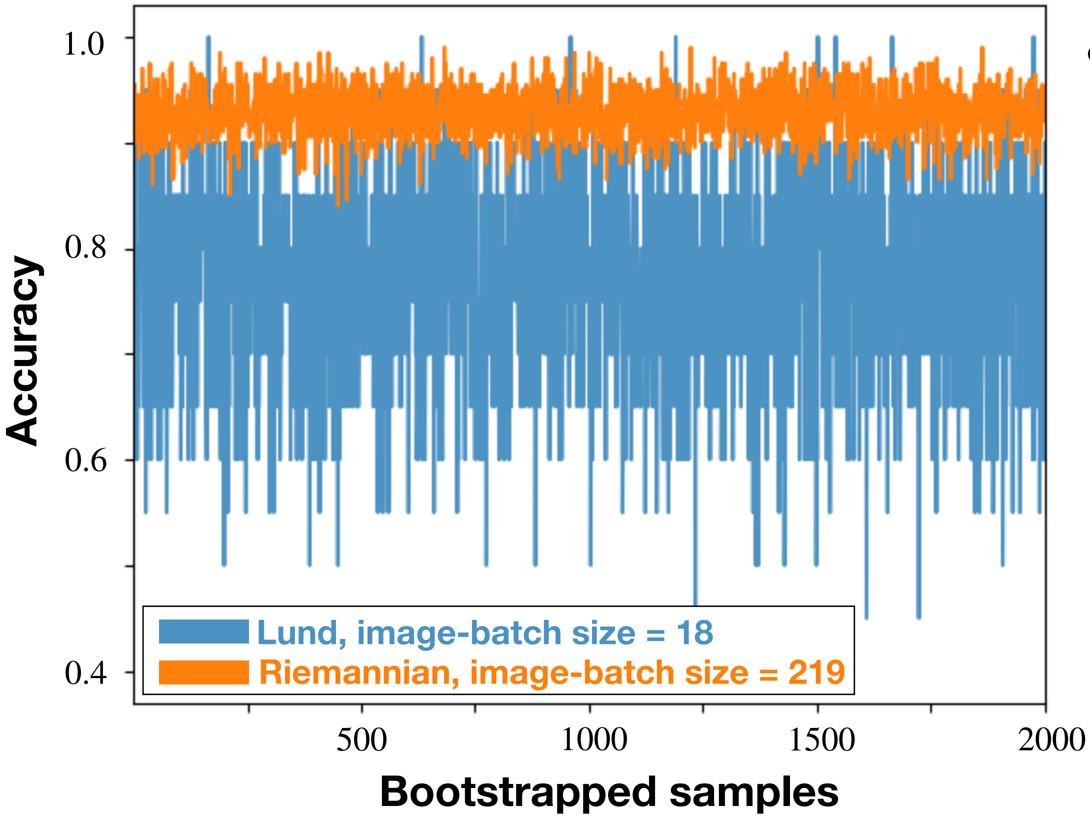
Performance test



- With 100,000 MC data sample each for (1) whole p_T range and for (2) boosted p_T "Riemann" preprocessing has a outperformance.

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

Applying to the LHC test



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

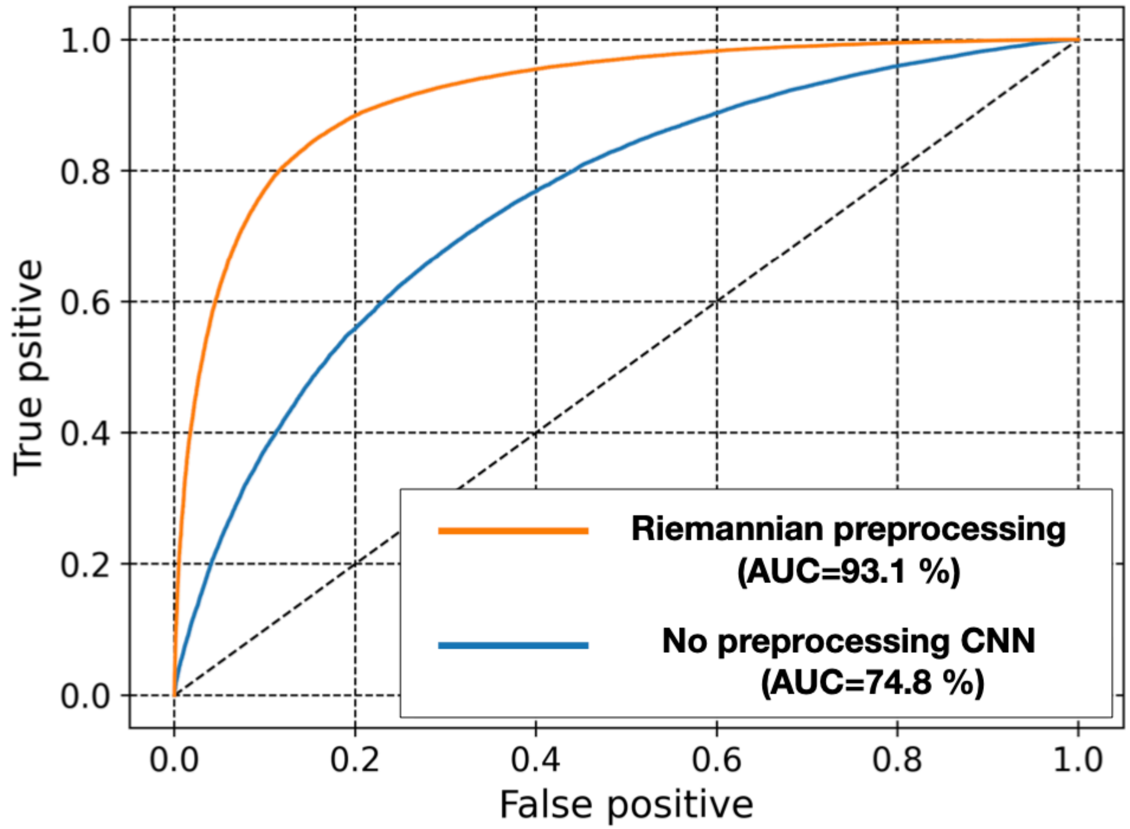
  - **Number of Higgs samples after selection cuts : 219**

  - **Number of Higgs samples in the boosted region ($p_T > 250\mathrm{GeV}$) : 18**

- With well-trained Neural Network, **analysis only with High P_T region** will **suffer from** **"statistical fluctuation"** in the real battle of the LHC.

- Thus, **the method with wide range of $p_T(h)$ would be better**

Reducing QCD backgrounds



- Applying "color singlet ML tagger", we can achieve **"factor 2"** (CNN: 25%) enhancement compared to conventional cut-and-counting based only on kinematic features.

Conclusion

- We are interested in maximizing the discovery chance of the **color singlet particle :**

  - Utilizing QCD information has been known to be helpful to suppress backgrounds.

- In this talk, I present **one simple way of data preprocessing** to use **QCD information over the wide range of kinematics.**
  : Color-singlet tagger.