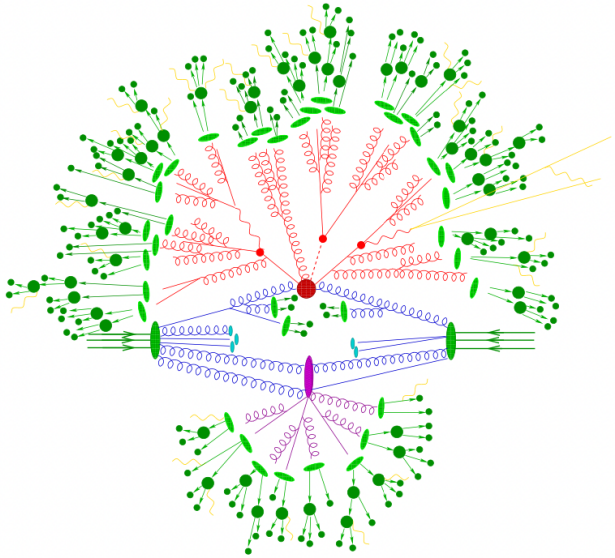
Riemannian Data preprocessing in ML to focus on QCD color structure

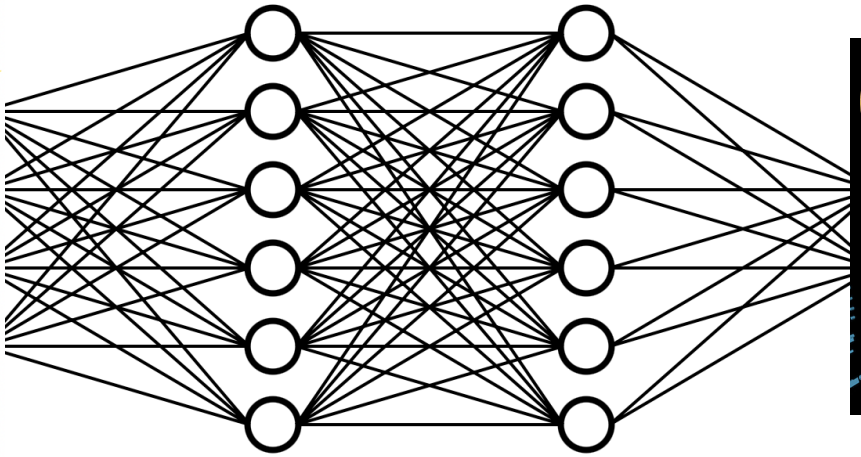
Myeonghun Park

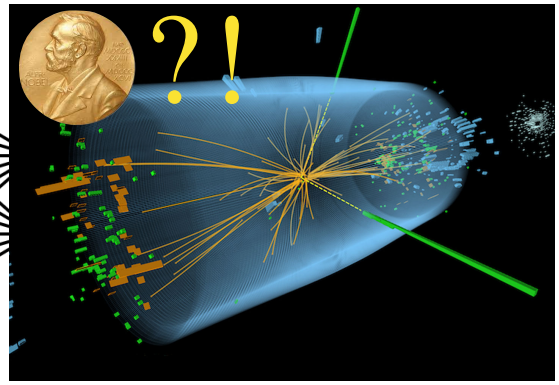
With A. Hammad, arXiv:2209.03898

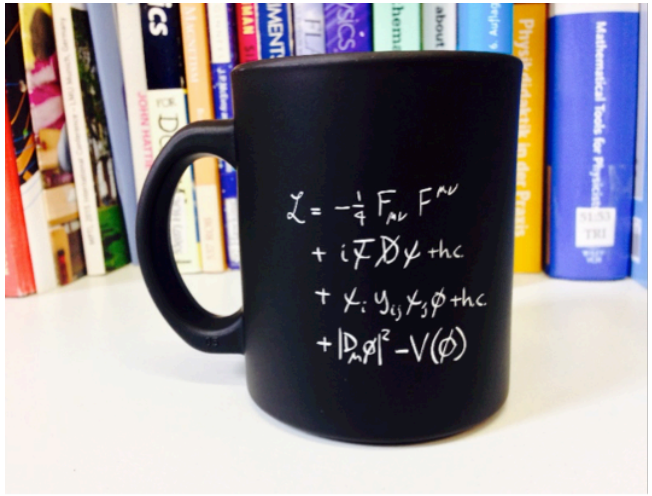
2022 Workshop on Physics of Dark Cosmos: dark matter, dark energy, and all

Theory, Data, Machine Learning









- With our elaborated **theoretical model**, (of course, we need lots of coffees)

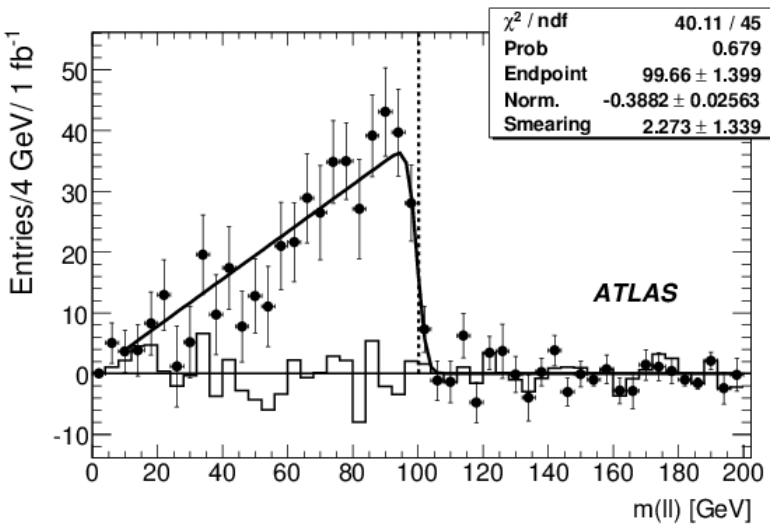
  1) Get expectations from **simulations**

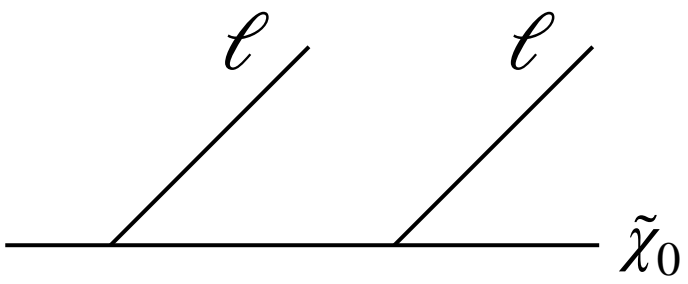
  2) Get **data** from **experiments** (e.g. the LHC)

  3) **compare** our expectation to data with sophisticated computer **algorithms**.

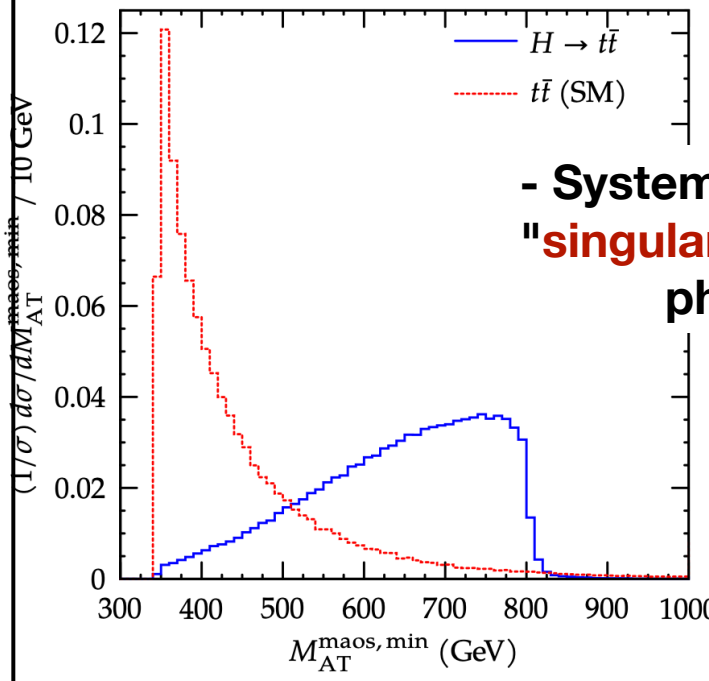
Extracting features of a new physics

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

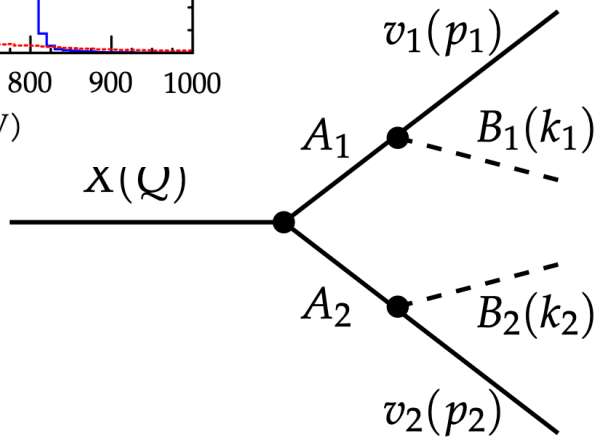




Chan Beom Park 2020

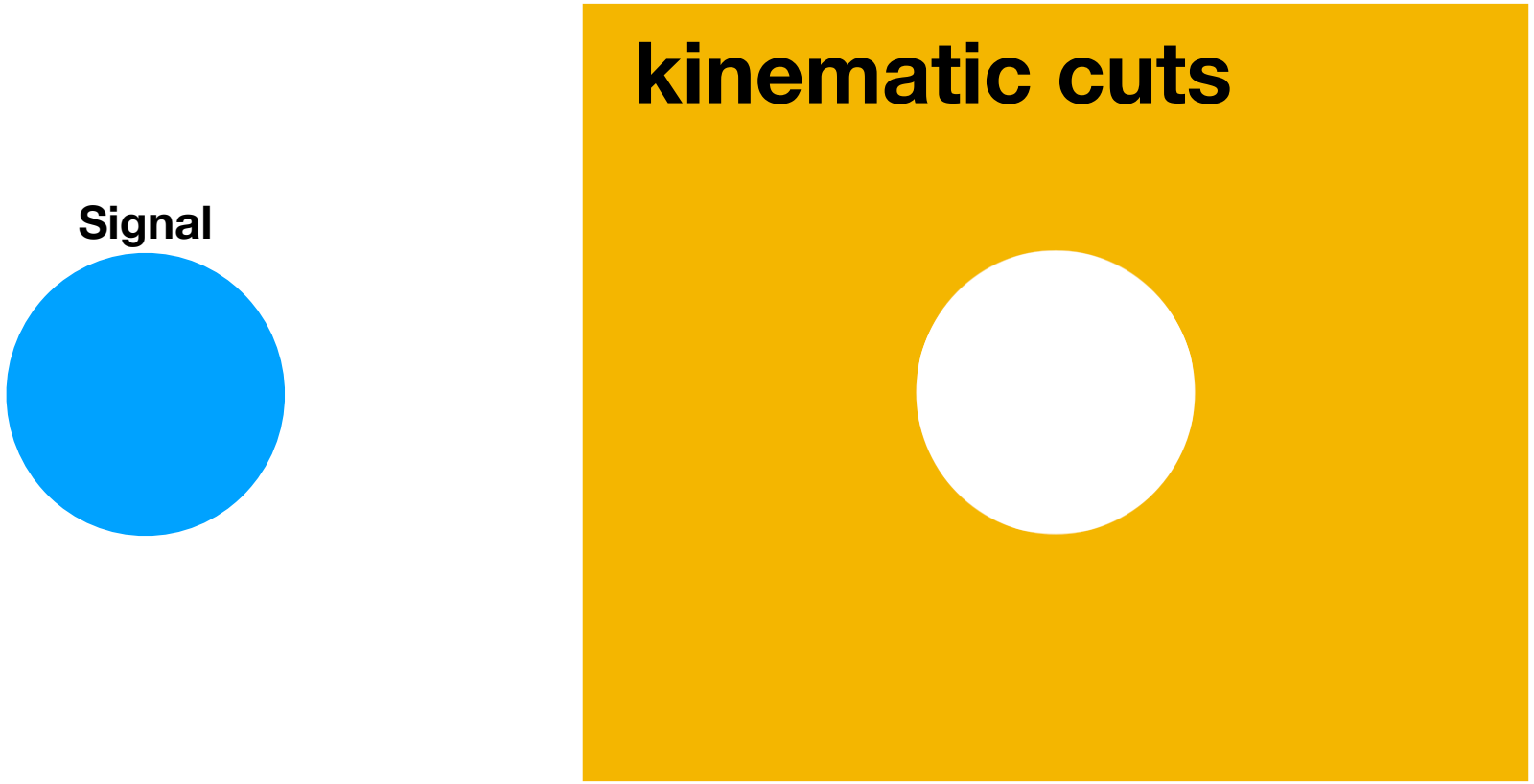


- Systematic method to find "singular structure" for new physics search

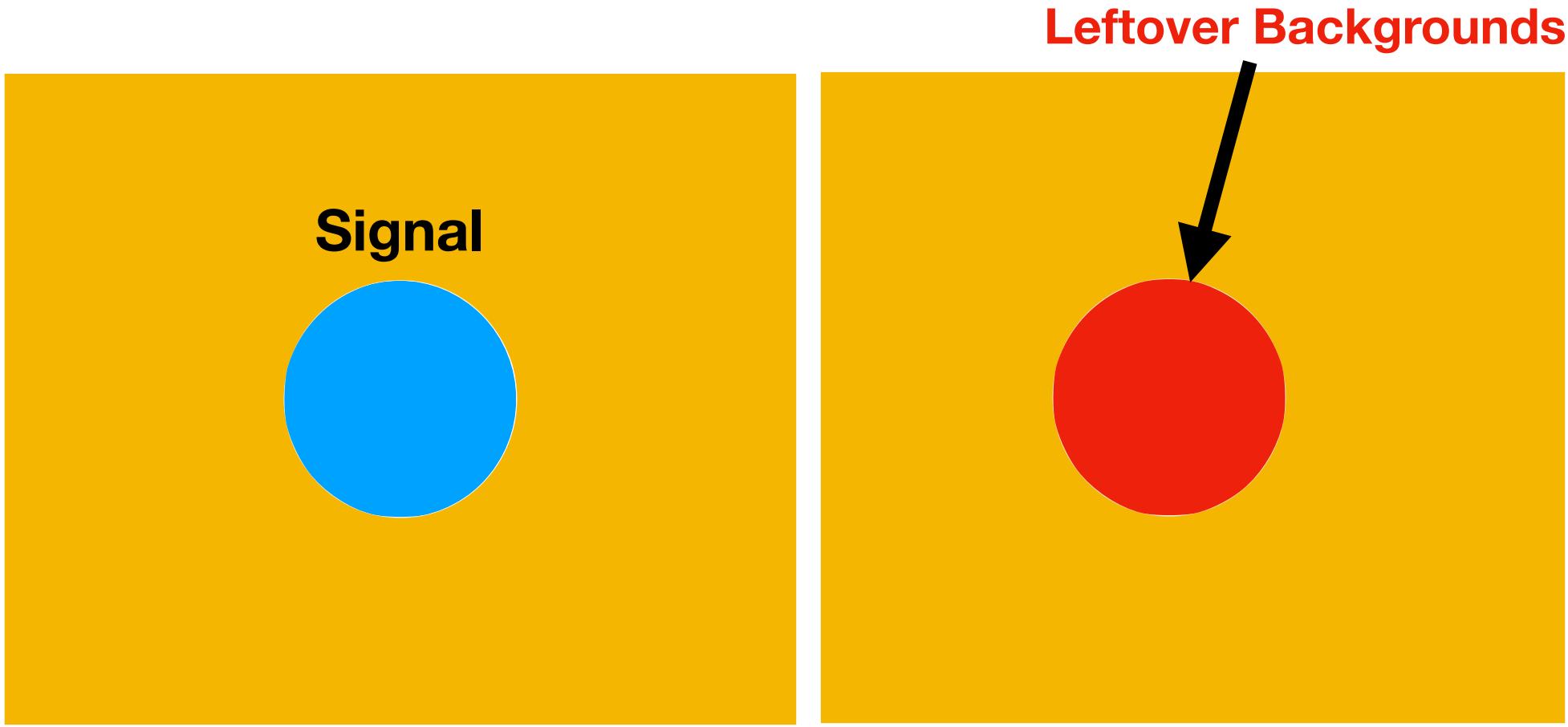


Basic idea of Kinematic cuts

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



Shaping backgrounds into signals ?



Leftover Backgrounds become very similar to signals

More than phase-space difference

- In many cases, the **soft QCD radiation patterns** from signals are different from Backgrounds.

Jason Gallicchio, Matthew D. Schwartz PRL 2010

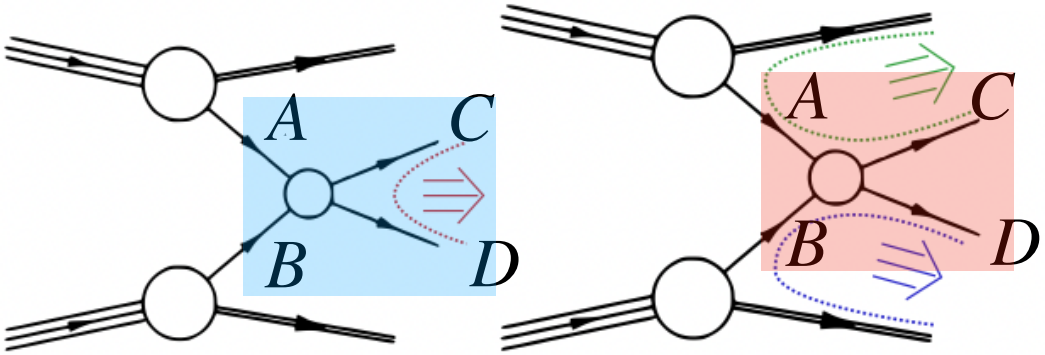
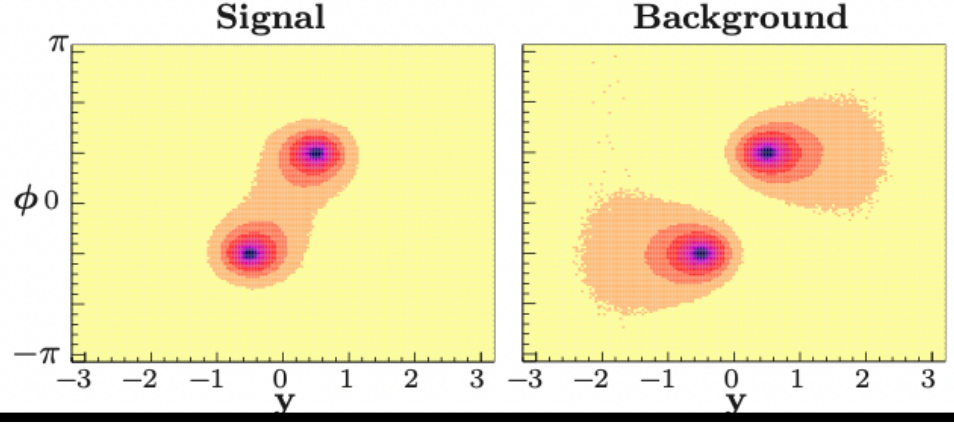


FIG. 1: Possible color connections for signal ($pp \to H \to b\bar{b}$) and for background ($pp \to g \to b\bar{b}$).



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

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

V . S .

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

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

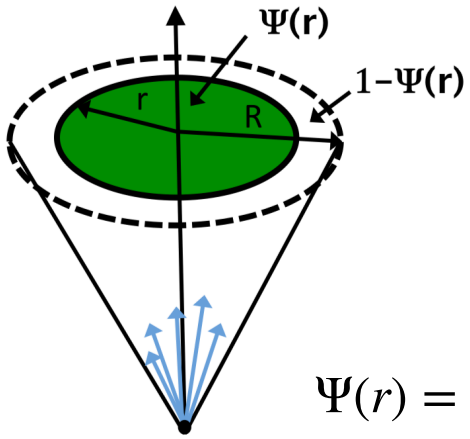
$$\mathrm{Tr}[T^A T^D]\mathrm{Tr}[T^B T^C]$$

Orthogonal information

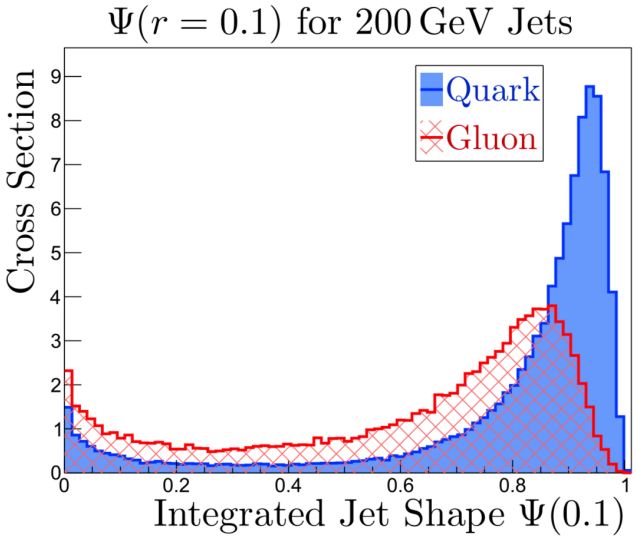
- Difference in kinematics is from "high P_T" region.
  - phase space
  - We have very sophisticated cuts (singular variables)

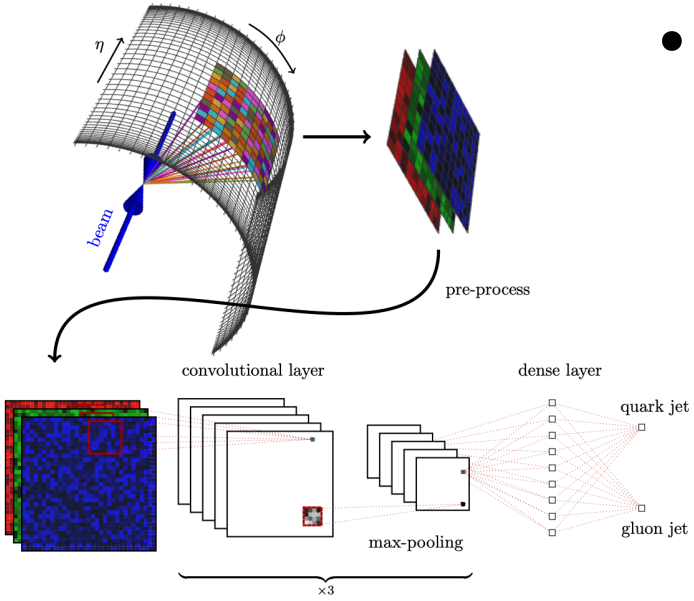
- Difference in QCD radiation patterns is from "soft P_T" region
  - status under a gauge group, $SU(3)_C$
  - We have good computer algorithms (Deep Learning with image)

Deep learning for QCD images: q vs g



$$\Psi(r) = \frac{\sum_i p_T(0 < r_i < r)}{p_T}$$



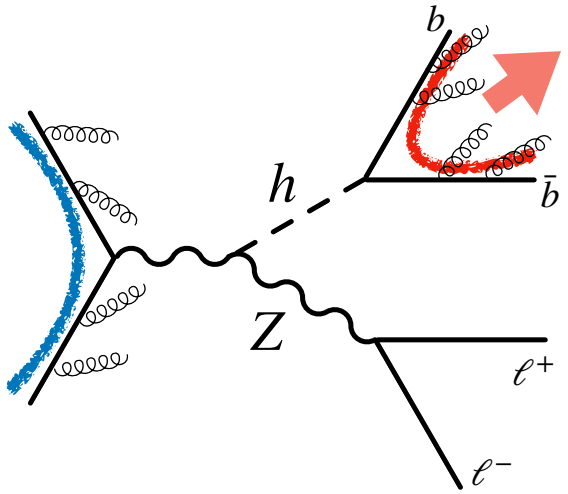


- CNN from industry **works very well**

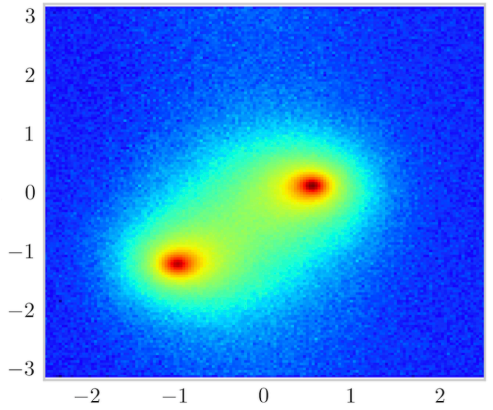
  in differentiating quark-jet vs gluon-jet
  - Pixels are energy deposits
    from various sub-detectors
    (e.g. : tracks, e-cal, h-cal)
  - **Energy deposits are well localized within ΔR_J**

M. Schwartz et.al. arXiv:1612.01551

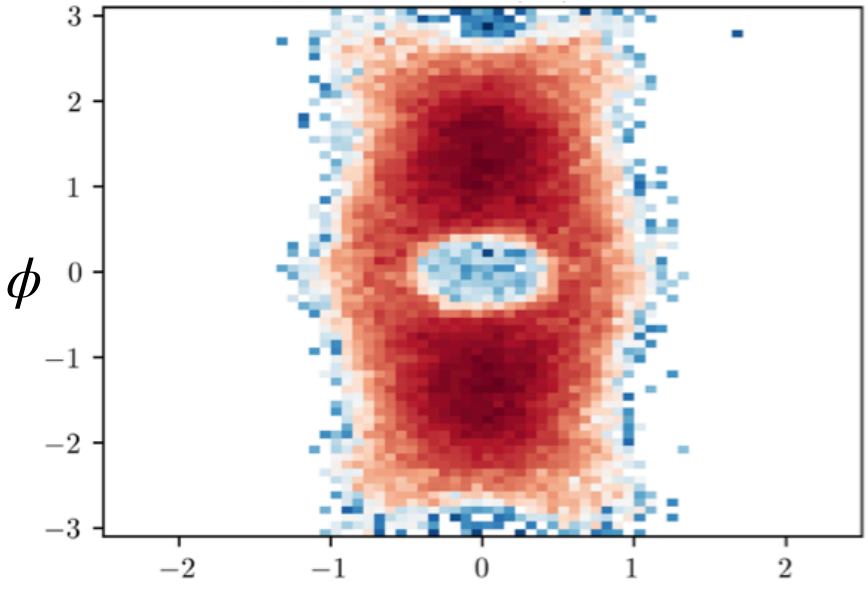
Can we use CNN in our case?



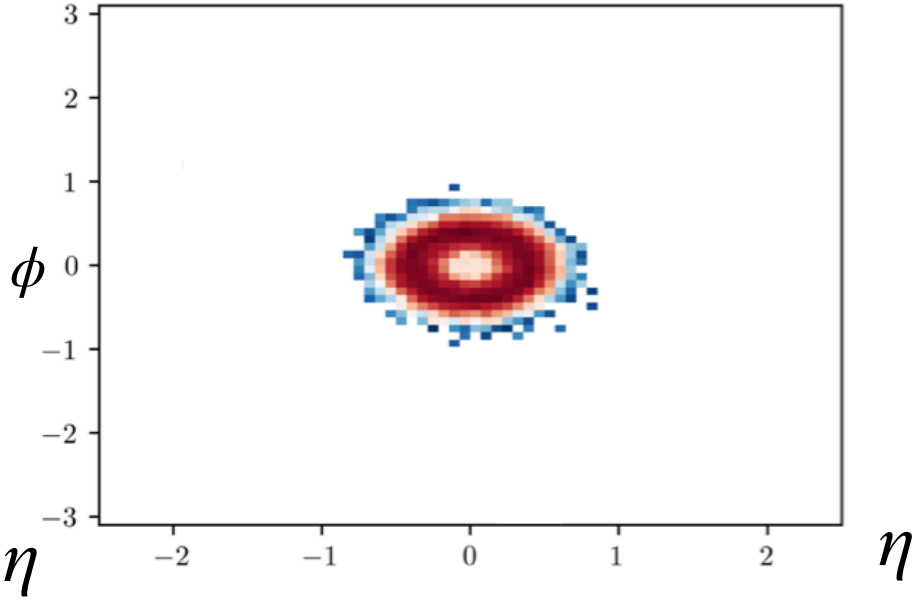
For fixed P_T



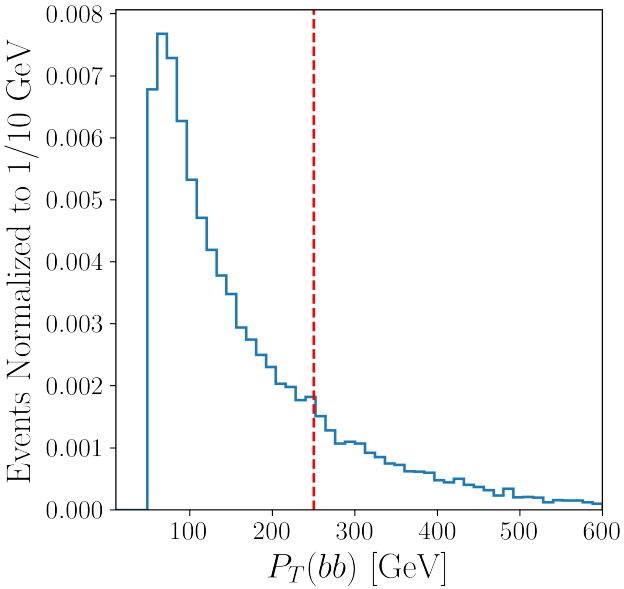
$p_T < 250\text{GeV}$

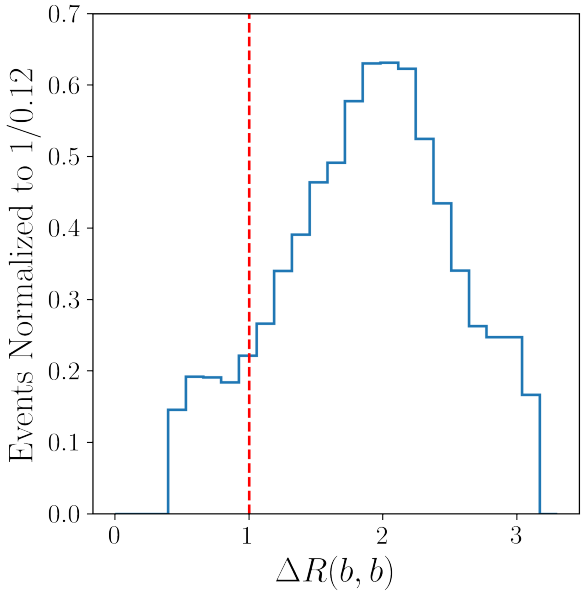


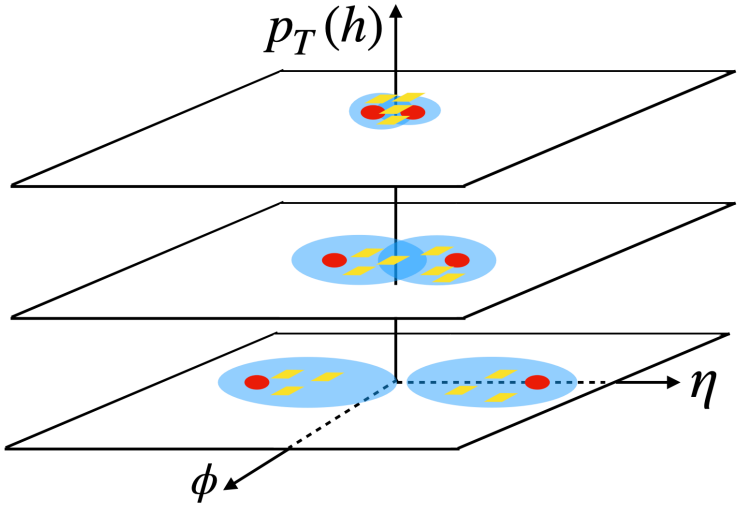
$p_T > 250\text{GeV}$



Can we use CNN in our case?



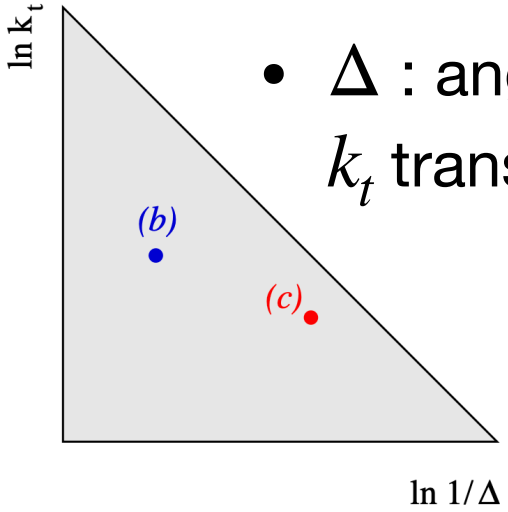




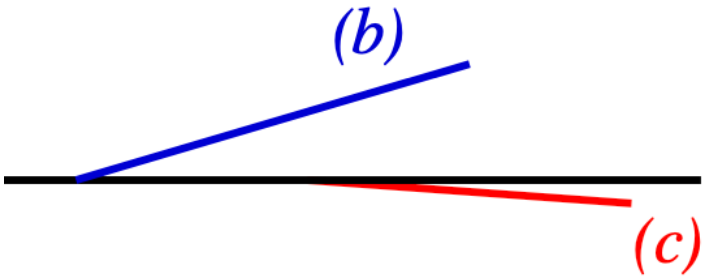
- Distance ΔR between b and $\bar{b}$ becomes smaller with large $P_T(h)$
- **We may focus on "localized jet"** (high P_T case: Boosted analysis)
- Or we works hard to find a **QCD pattern observable invariant under** $P_T(h)$

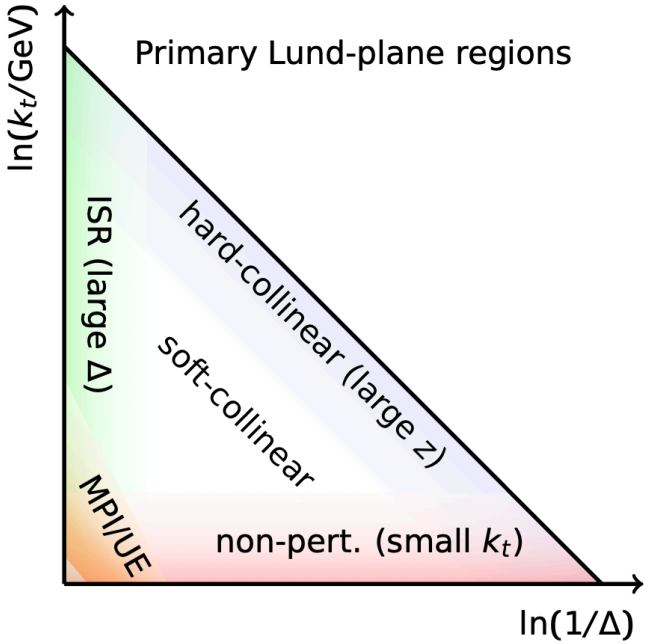
Study for Boosted region

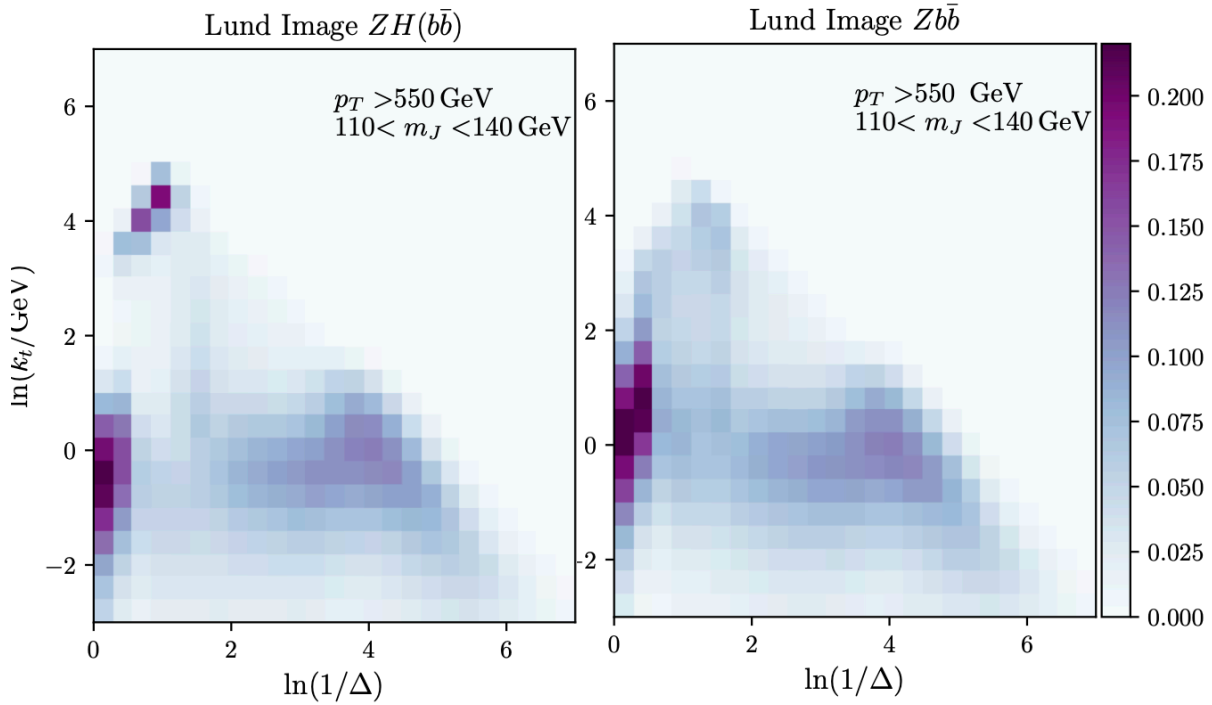
Gavin Salam et.al, (2018),
Charanjit K. Khosa & Simone Marzani (2021)



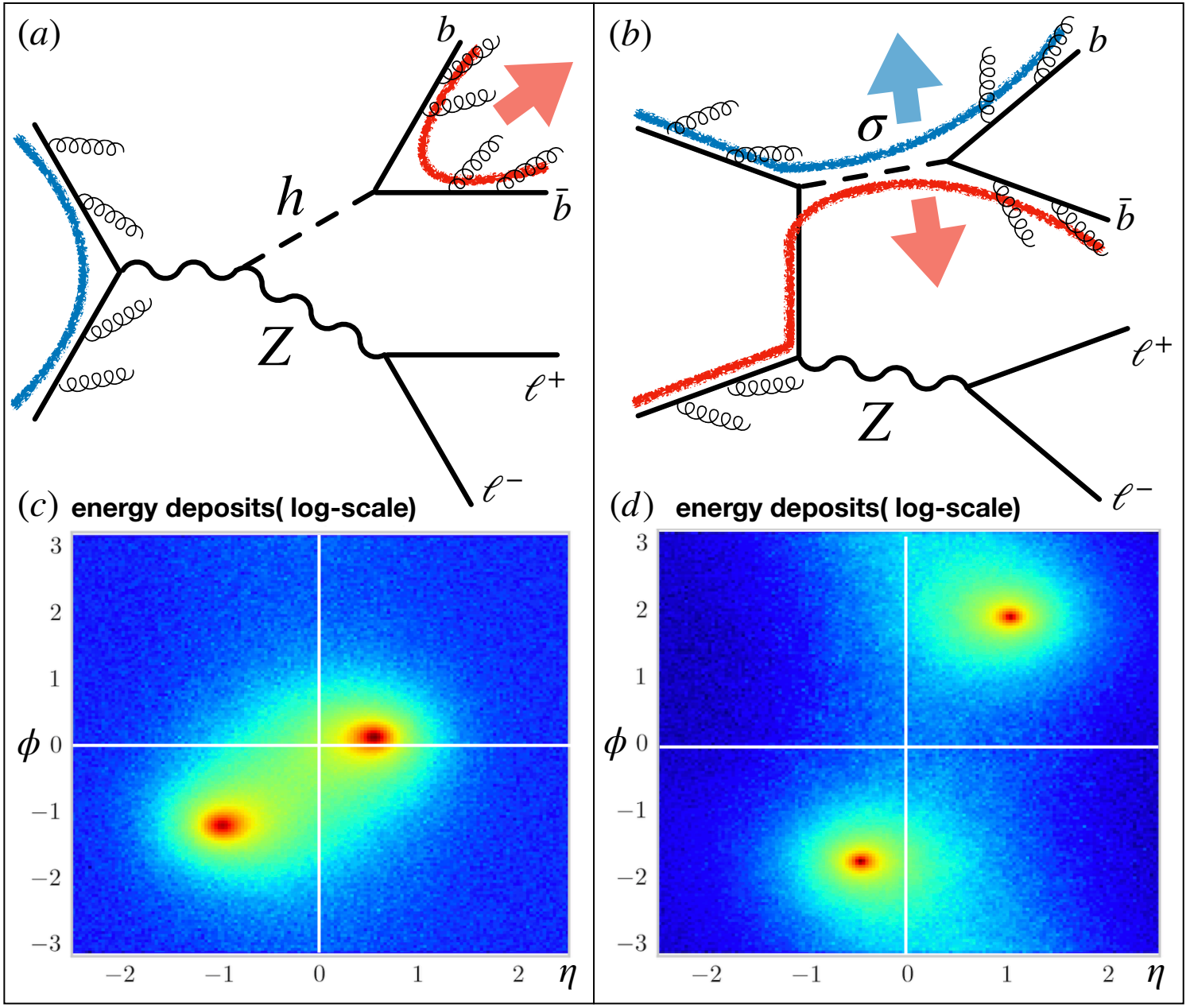
- Δ : angular distance,
  k_t transverse momentum

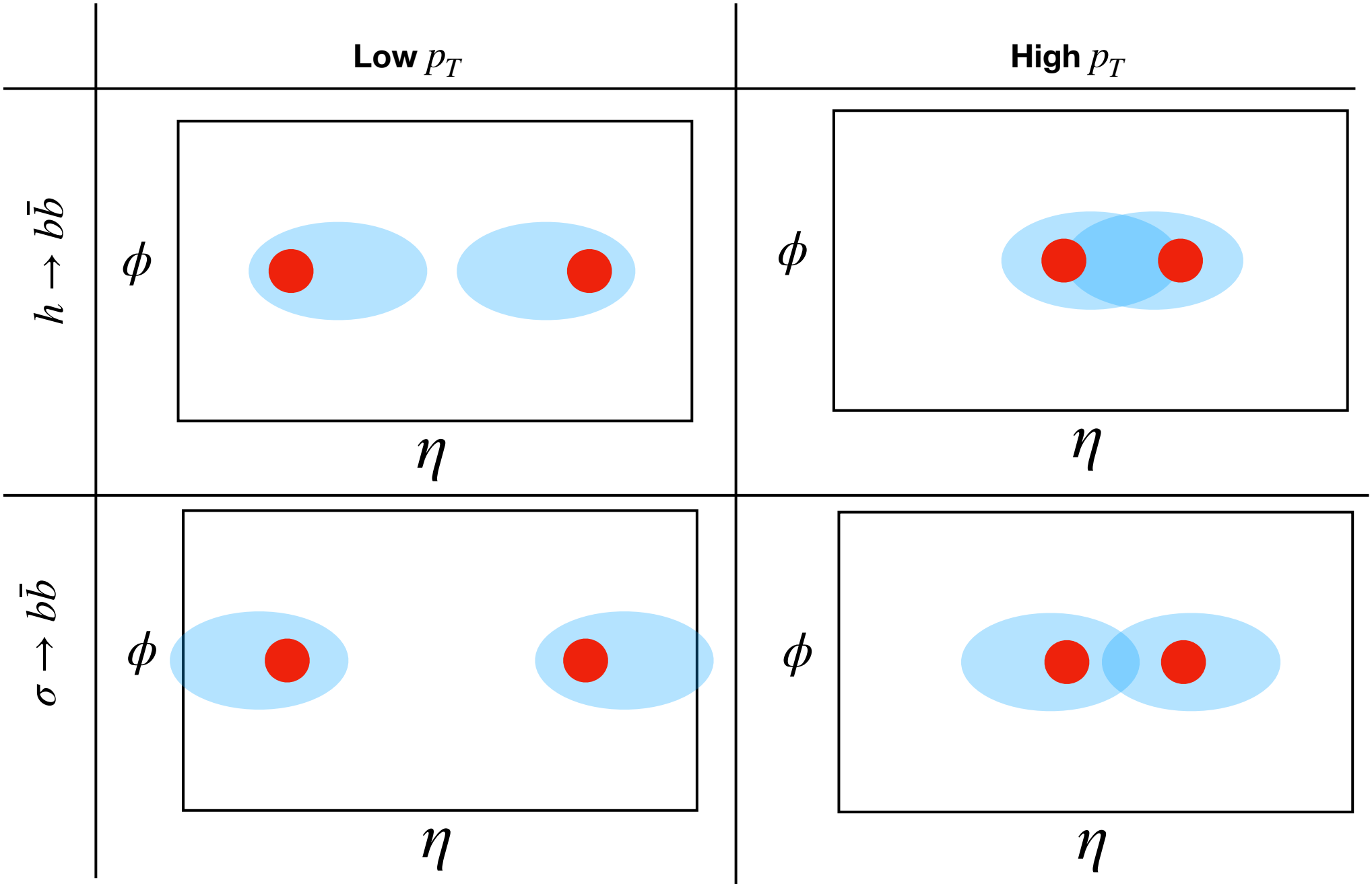




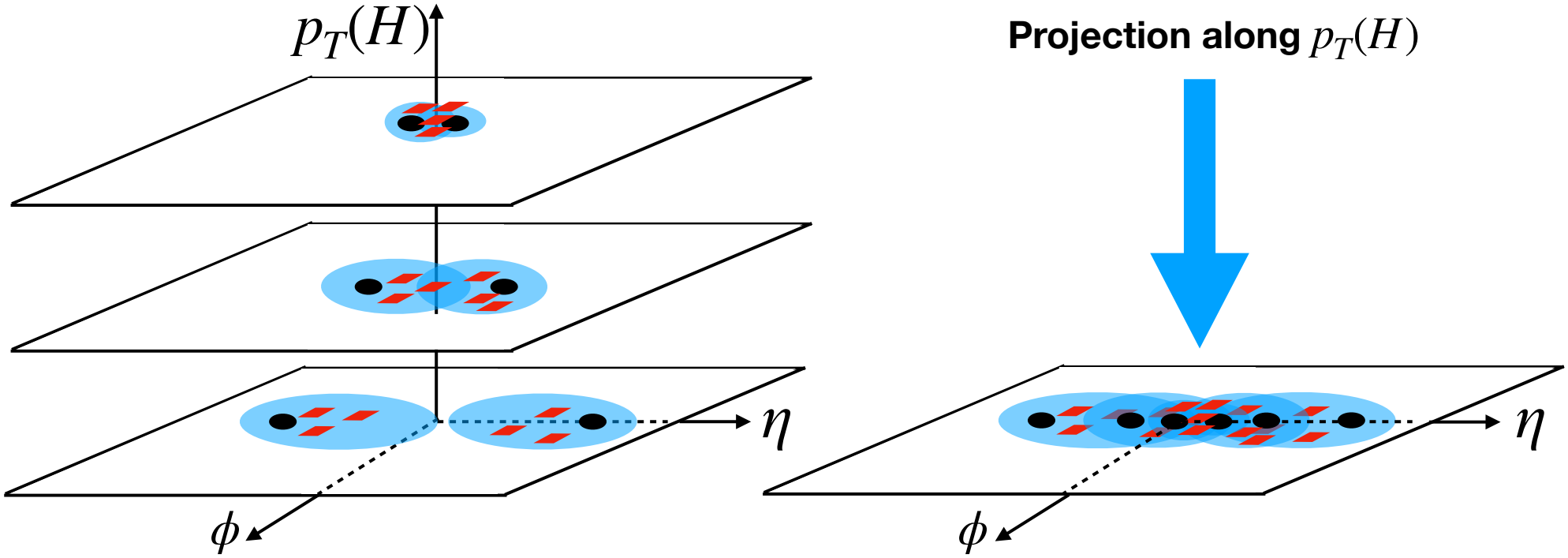


Toy example


(a)
b
h
$\bar{b}$
Z
ℓ^+
ℓ^-
(b)
b
σ
$\bar{b}$
ℓ^+
Z
ℓ^-
(c) energy deposits(log-scale)
(d) energy deposits(log-scale)
3
2
1
ϕ 0
−1
−2
−3
−2
−1
0
1
2
η



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



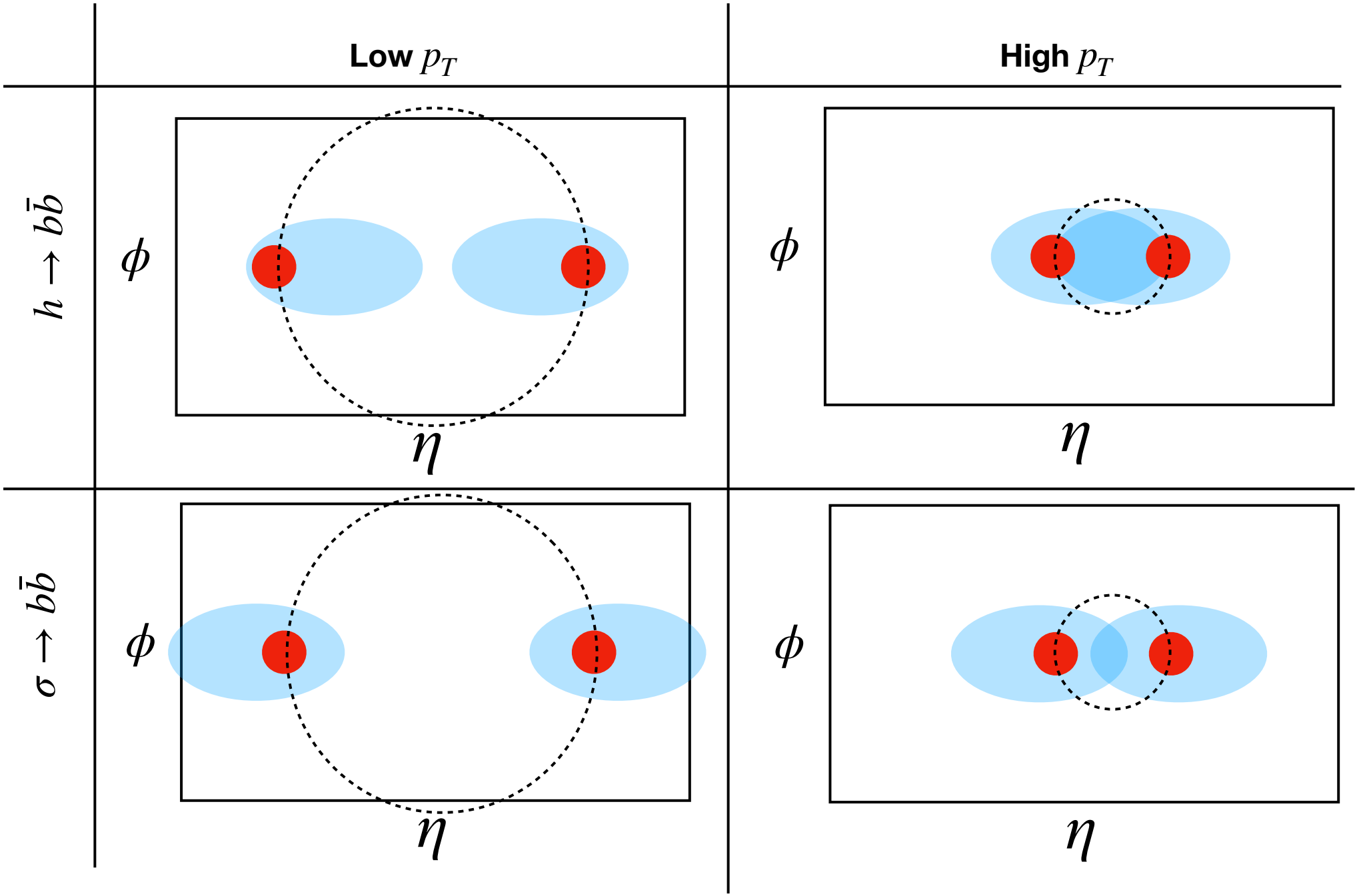
- If one tries to study various $p_T(H)$ ranges,

  even ML will not give a good performance.

How can we use

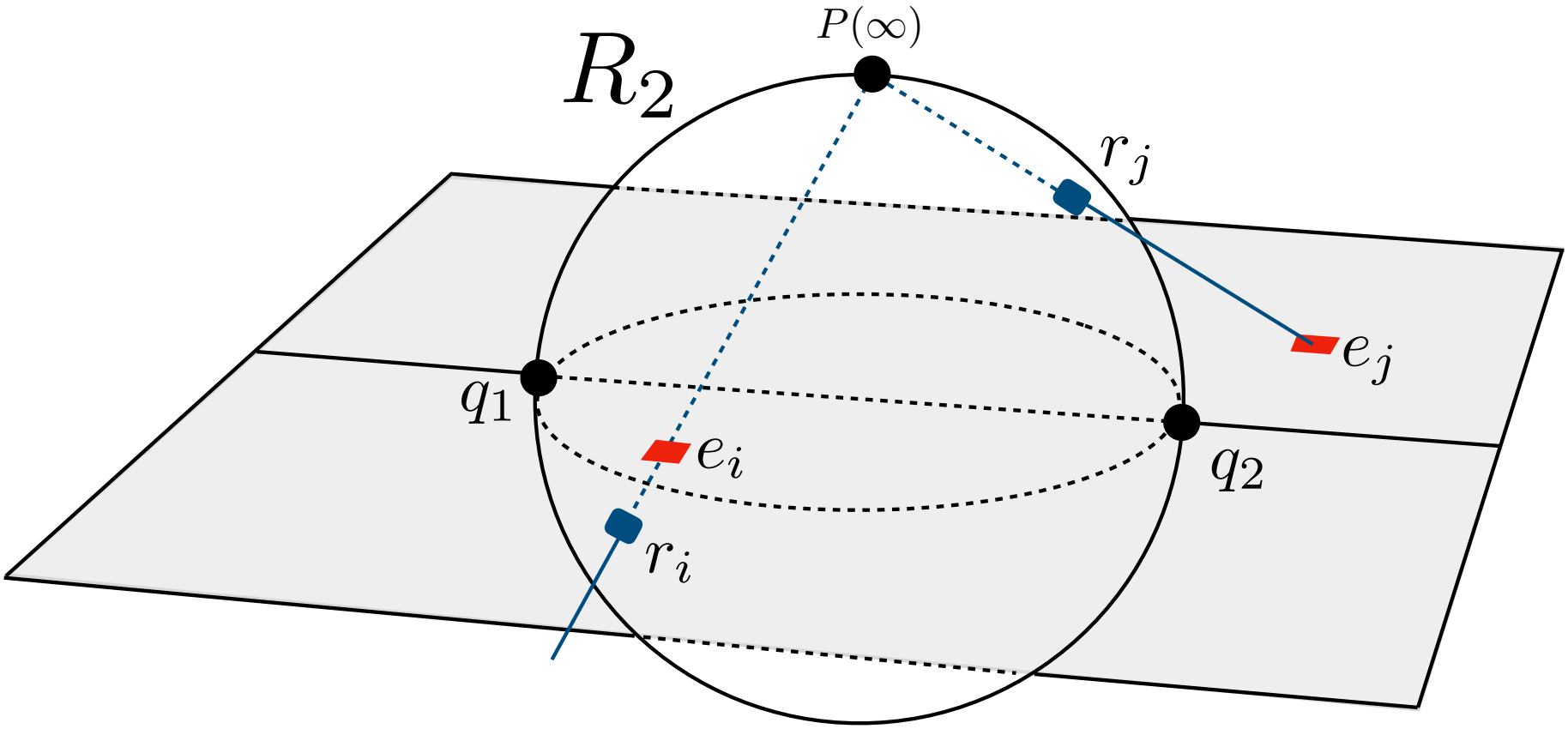
"**full p_T range**" of "Higgs"?

- for the actual LHC test

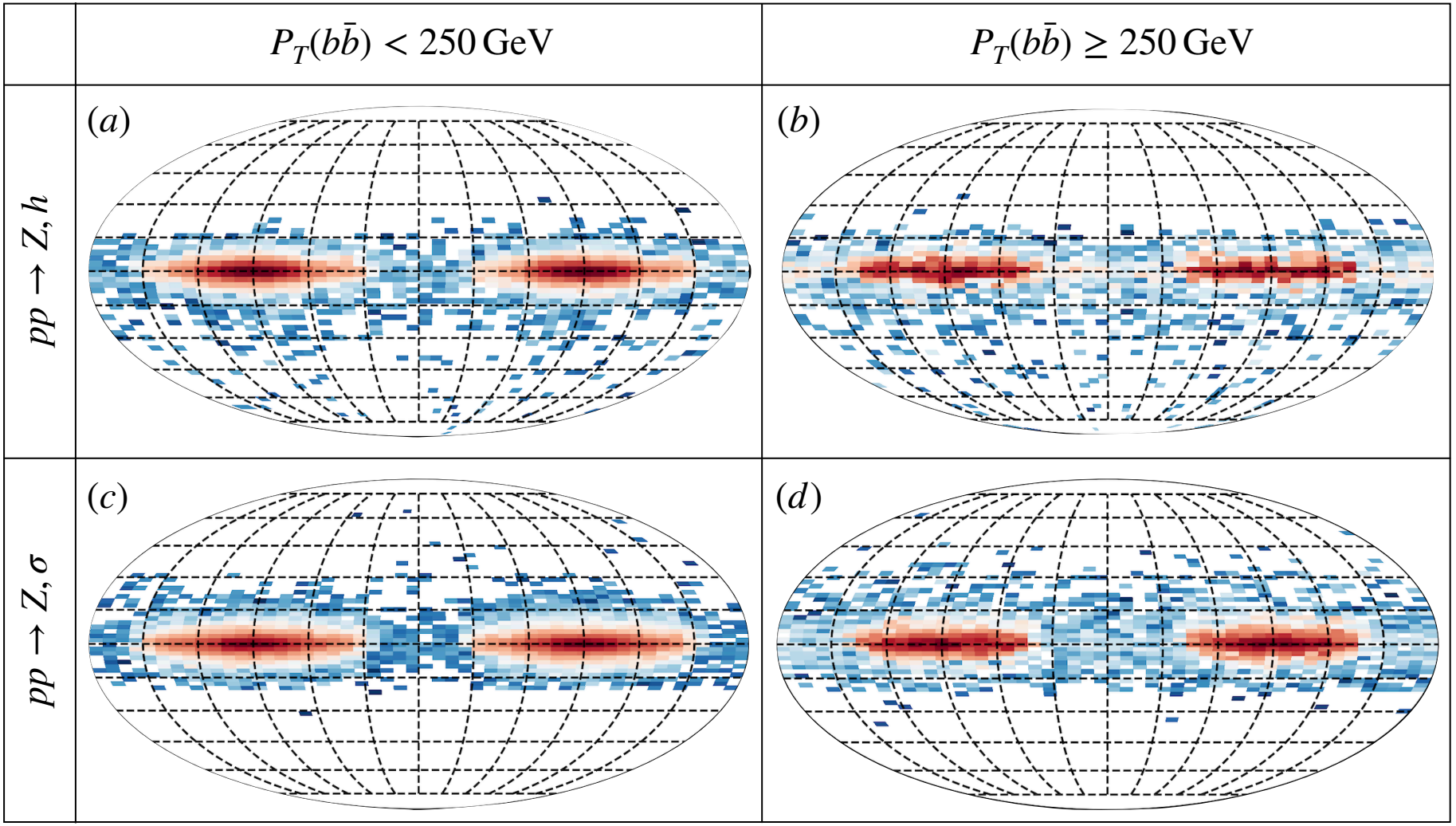


- A **binary problem**, either "**inside**"or "**outside**" a circle.

Inverse stereographic projection



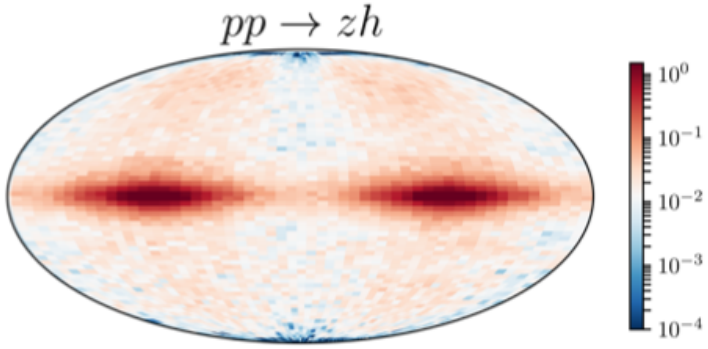
- Soft radiations which are
  inside of a circle → Southern hemisphere (H)
  outside of a circle → North hemisphere (σ)

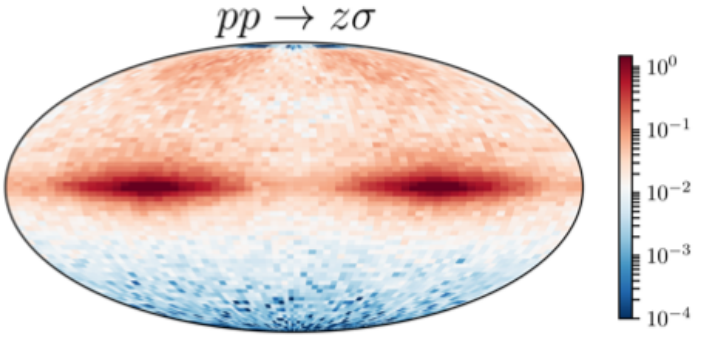


- Our image is invariant under $P_T(b\bar{b})$!

Landscape of Color activity

- Accumulated 5000 events shot

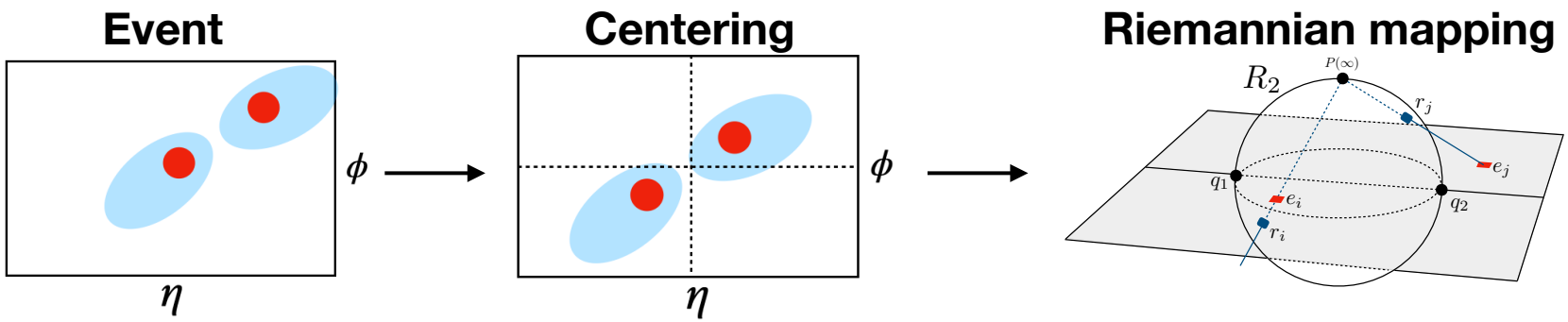


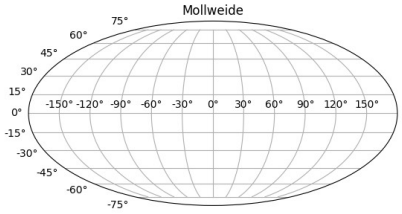


- Corruptions in North hemisphere are from ISR / MPI QCD activities.

- Color-connected case (right) has populated on Northern hemisphere
  Color-disconnected has population in a southern hemisphere.

Riemannian preprocessing with CNN



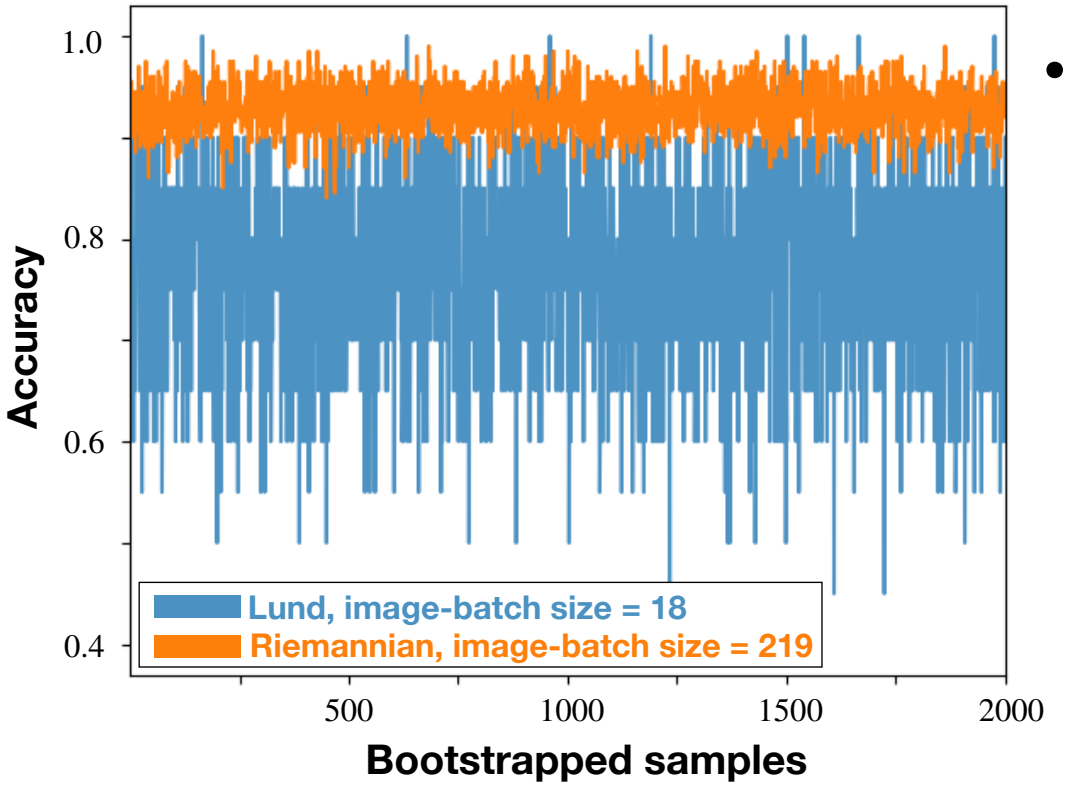


Mollweide projection

simple CNN

Binary classification

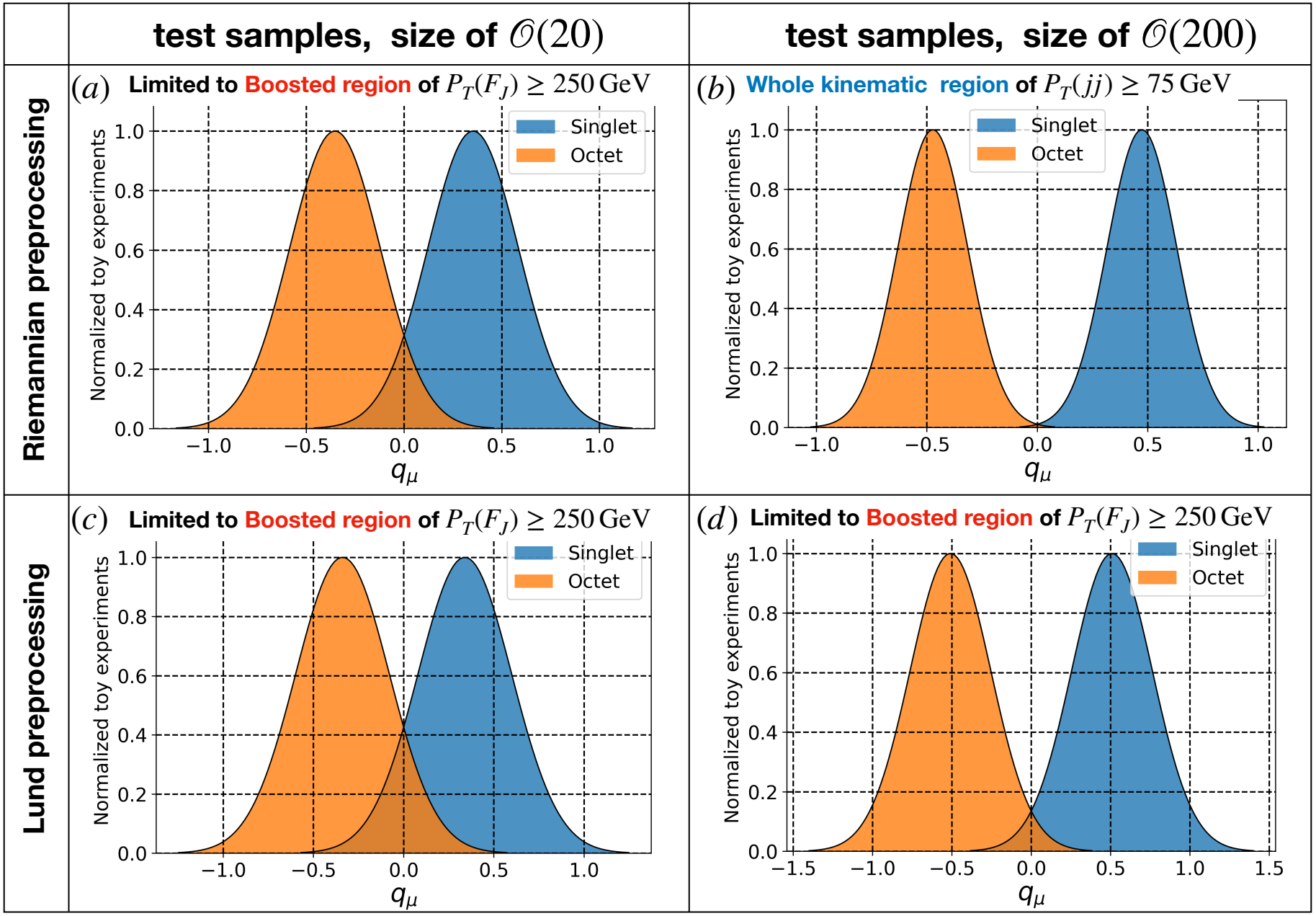
Applying NN to the LHC



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

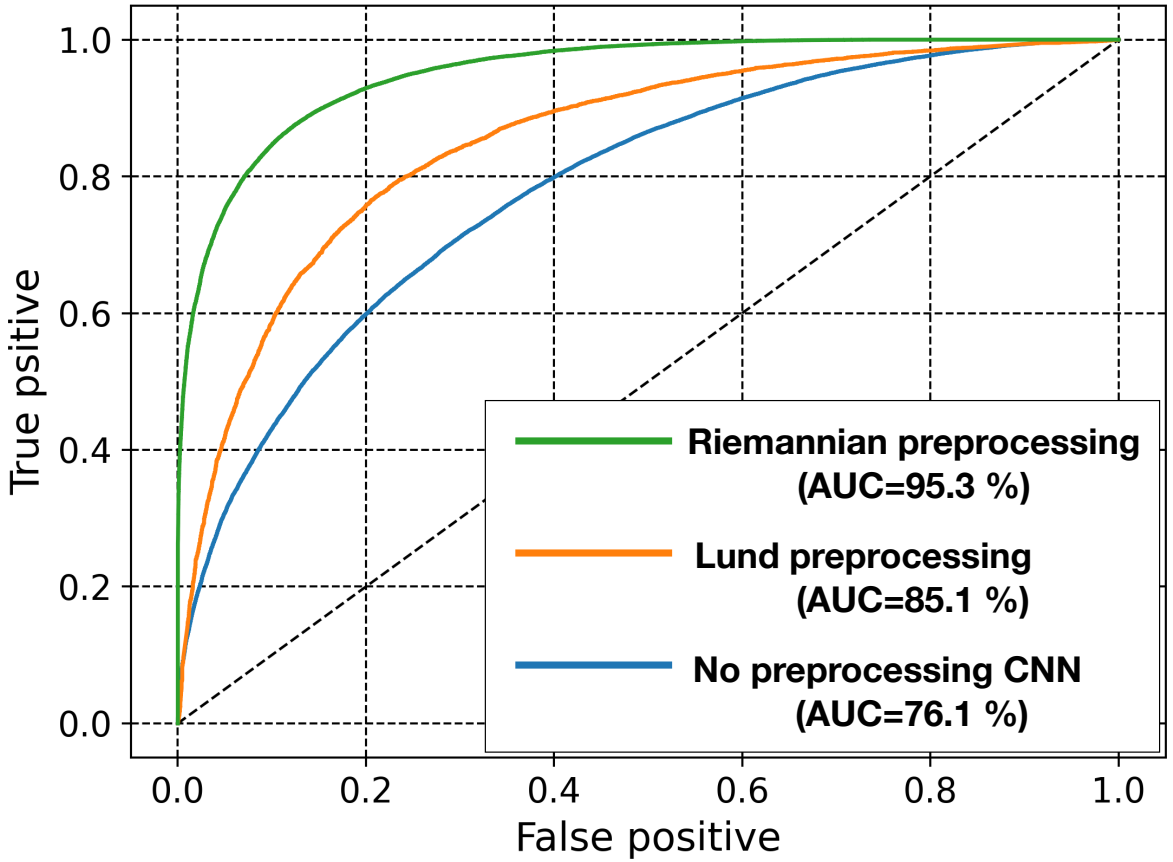
  Number of Higgs samples after selection cuts : 219
  Number of Higgs samples in the boosted region ($p_T > 250\text{GeV}$) : 18

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



- Our simple mapping is better even with the same statistics.

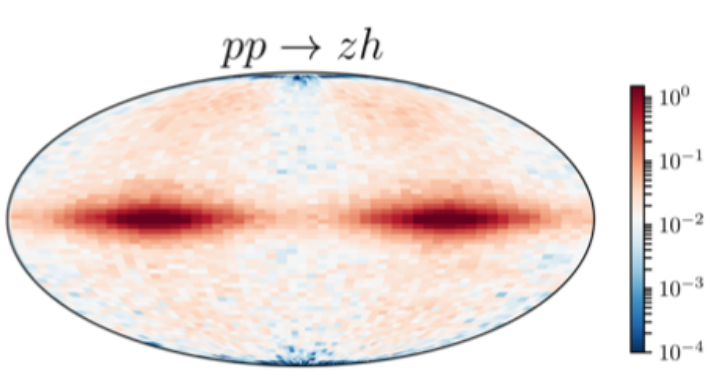
Performance test

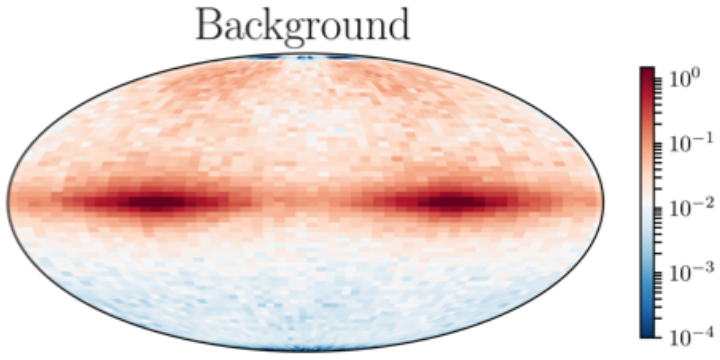


- With 100,000 MC data sample each for (1) whole p_T range and for (2) boosted p_T (60% training, 20% validation, 20% test),
  Riemannian preprocessing has a outperformance.

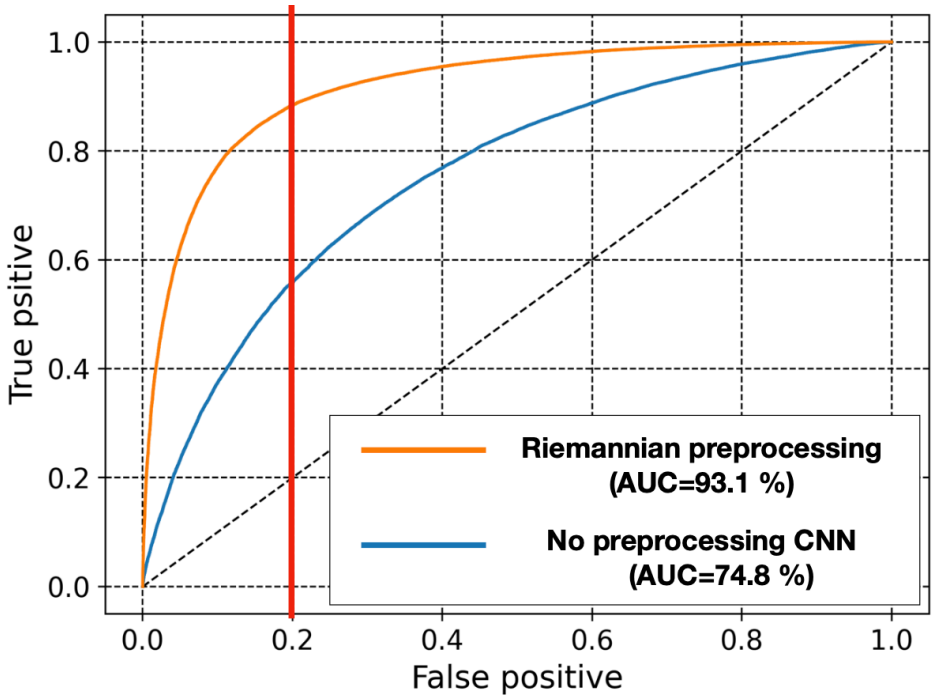
- Lund preprocessing ("double-logarithmic plane") is from [arXiv:2105.03989] for a boosted Higgs (Data preprocessing with selected QCD features)

Signal / BKG separation





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



- After various kinematic cuts (ATLAS), we test

  - Ordinary CNN : 25% better
  - Riemannian preprocessed CNN : **Factor 2 better!**

Conclusion

- I presented a **simple mapping** to make QCD information independent on a phase-space.

- **Data Preprocessing** is still required
  - theoretical point of view: Better understanding
  - experimental point of view: for the actual statistics@LHC

- Thus, in applying various Artificial Neural Network (ANN) techniques in collider physics, our **domain knowledge** plays a key role.