Physicists Learning from Machines Learning

Smart but Interpretable Neural Networks for Physics at the LHC

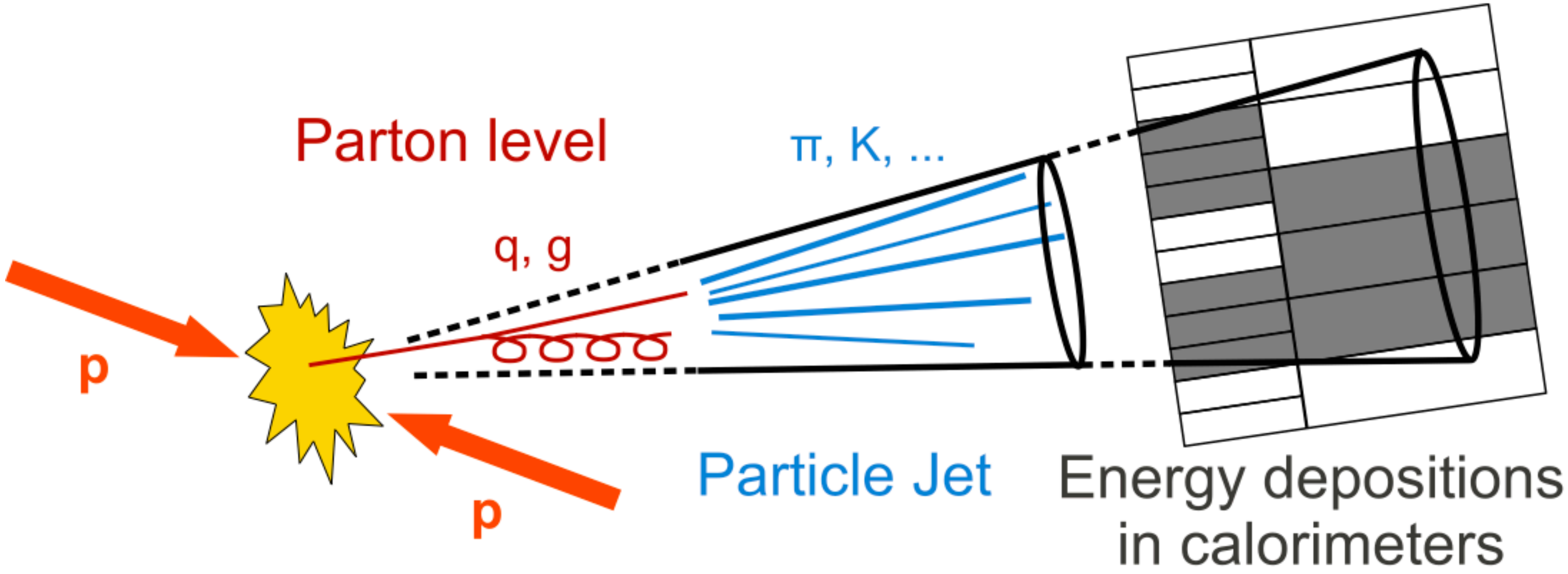
Authors: Taylor Faucett, Daniel Whiteson & Jesse Thaler

Mapping machine-learned physics into a human-readable space
https://journals.aps.org/prd/abstract/10.1103/PhysRevD.103.036020

@ Clermont-Ferrand

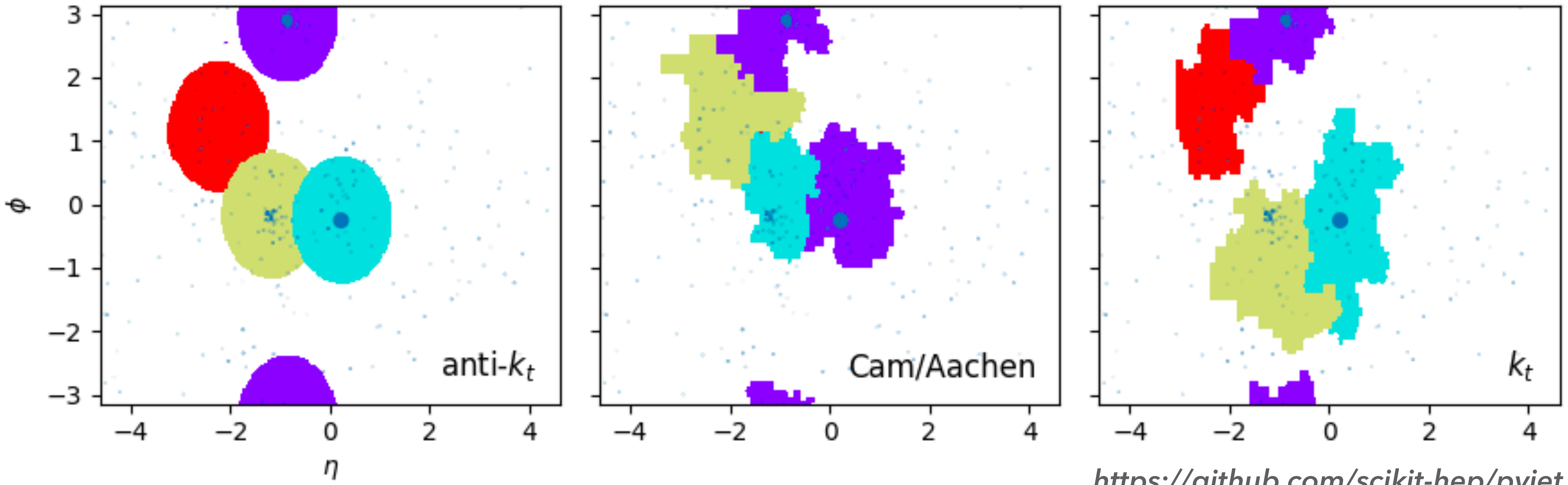
Hadrons & Jets

- Particle collisions create particles with non-zero color charge (i.e. quarks & gluons)
- Free quarks/gluons hadronize to produce hadrons (e.g. mesons and baryons)
- “Jets” are collimated sprays of many hadrons in a cone.
- Identifying jets and different kinds of jets help distinguish high-energy processes.



https://cms.cern/tags/particle-jet

- Jets are isolated in the detector using clustering algorithms
- How do we distinguish different types of jets from one another?



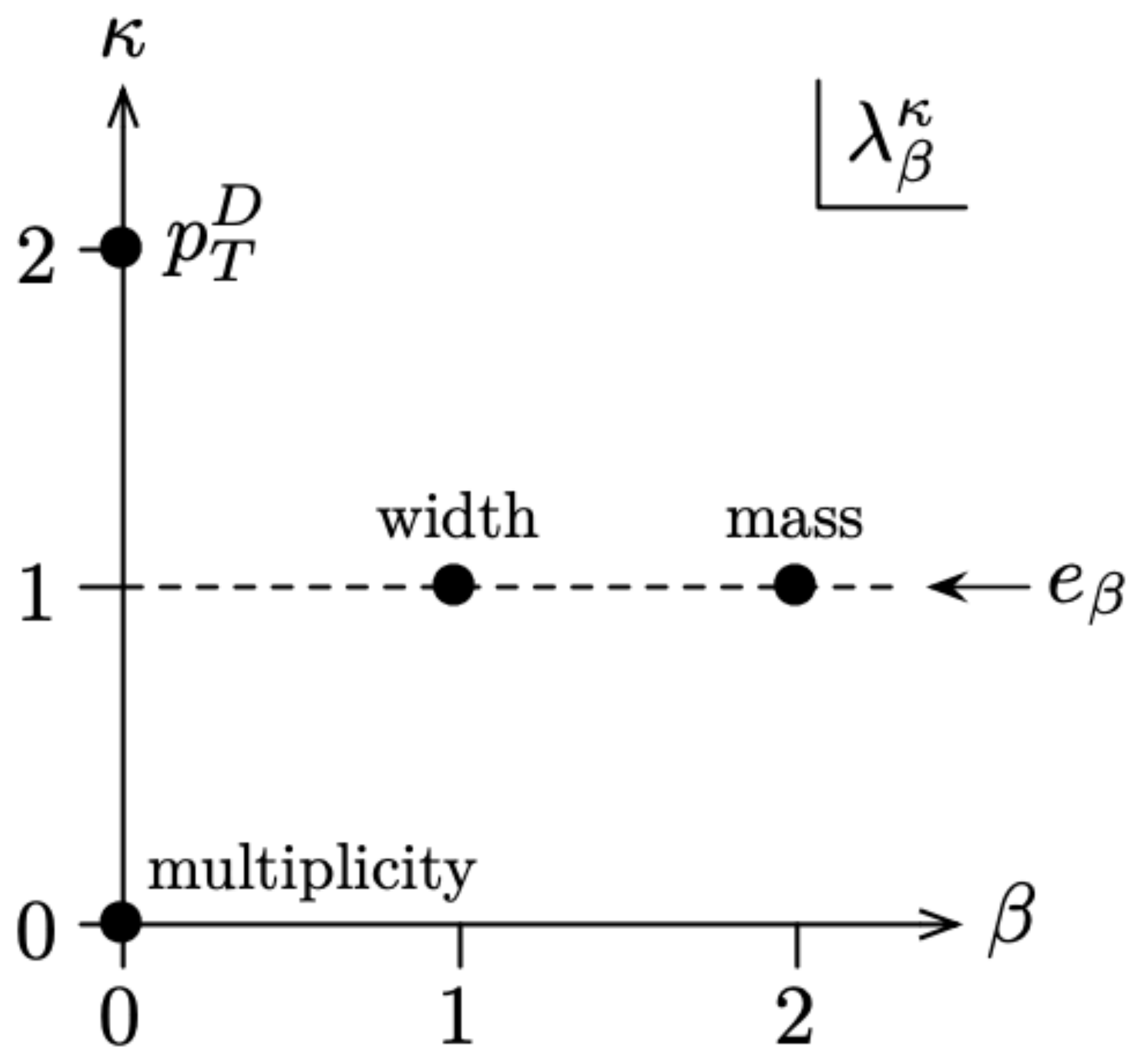
https://github.com/scikit-hep/pyjet

Jet Substructure (JSS)

What variables would be useful for training ML to classify different jets?

$$\lambda_\beta^\kappa = \sum_{i \in jet} z_i^\kappa \theta_i^\beta$$

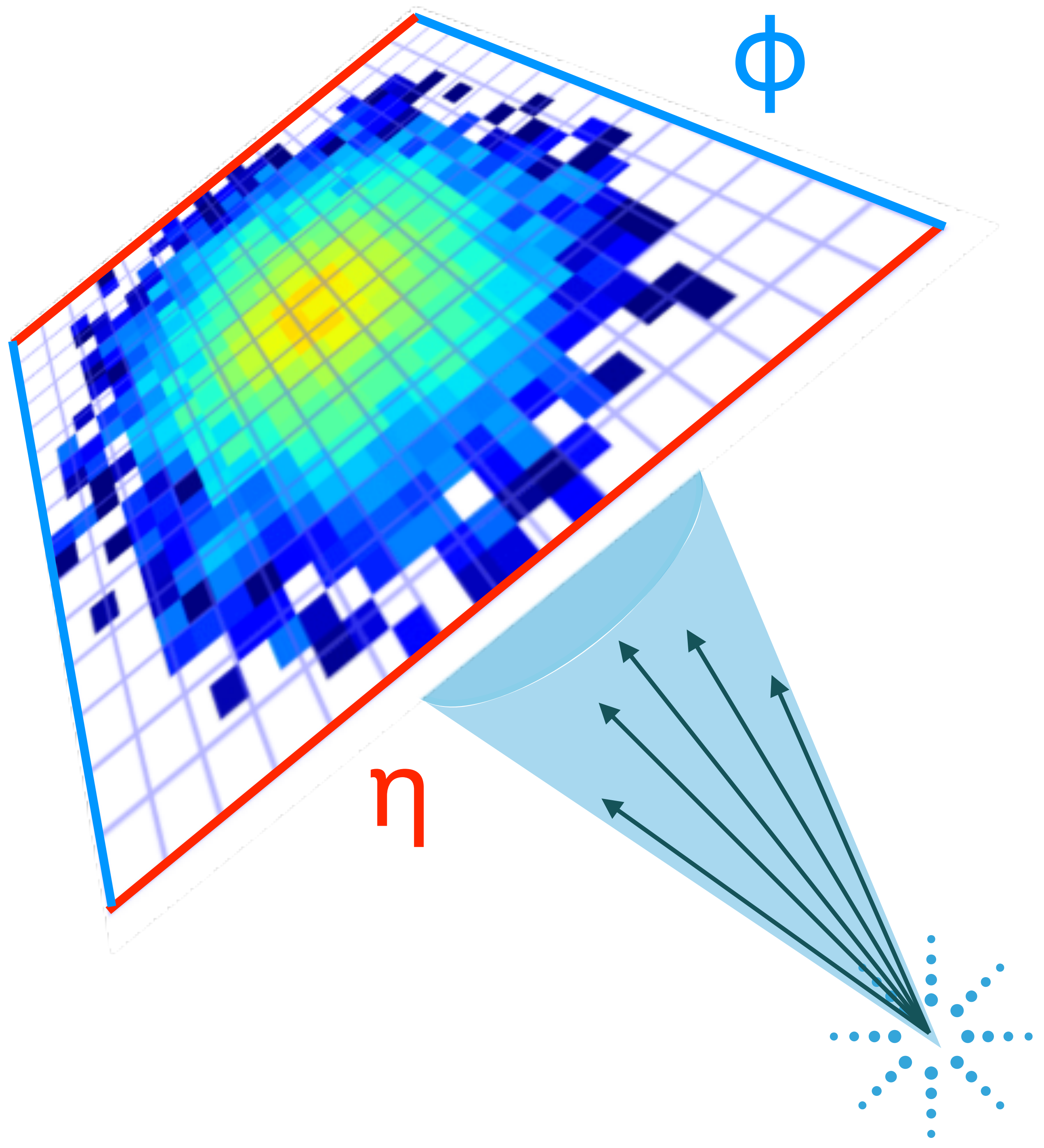
- Quark/Gluon discrimination?
  - Gluons produce more particles.
  - Look at "jet multiplicity" and similar proxies/weightings of momentum/ angular separation
  - From "generalized angularity": multiplicity, LHA, pTD…
- What about Jets with multiple sub-jets?
  - W/Z/h decay to 2 jets
  - We invent "N-subjettiness" to quantify separable jet substructure



$$\tau_N = \frac{1}{d_0} \sum_k p_{T,k} min\left\{\Delta R_{0,k}, \Delta R_{1,k} \ldots \Delta R_{N,k}\right\}$$

What's the alternative?

- Do we need to hand-pick observables in every study?
- What if we let a Deep Neural Network learn to solve the problem?
- Can a CNN to learn to classify directly from the calorimeter data?

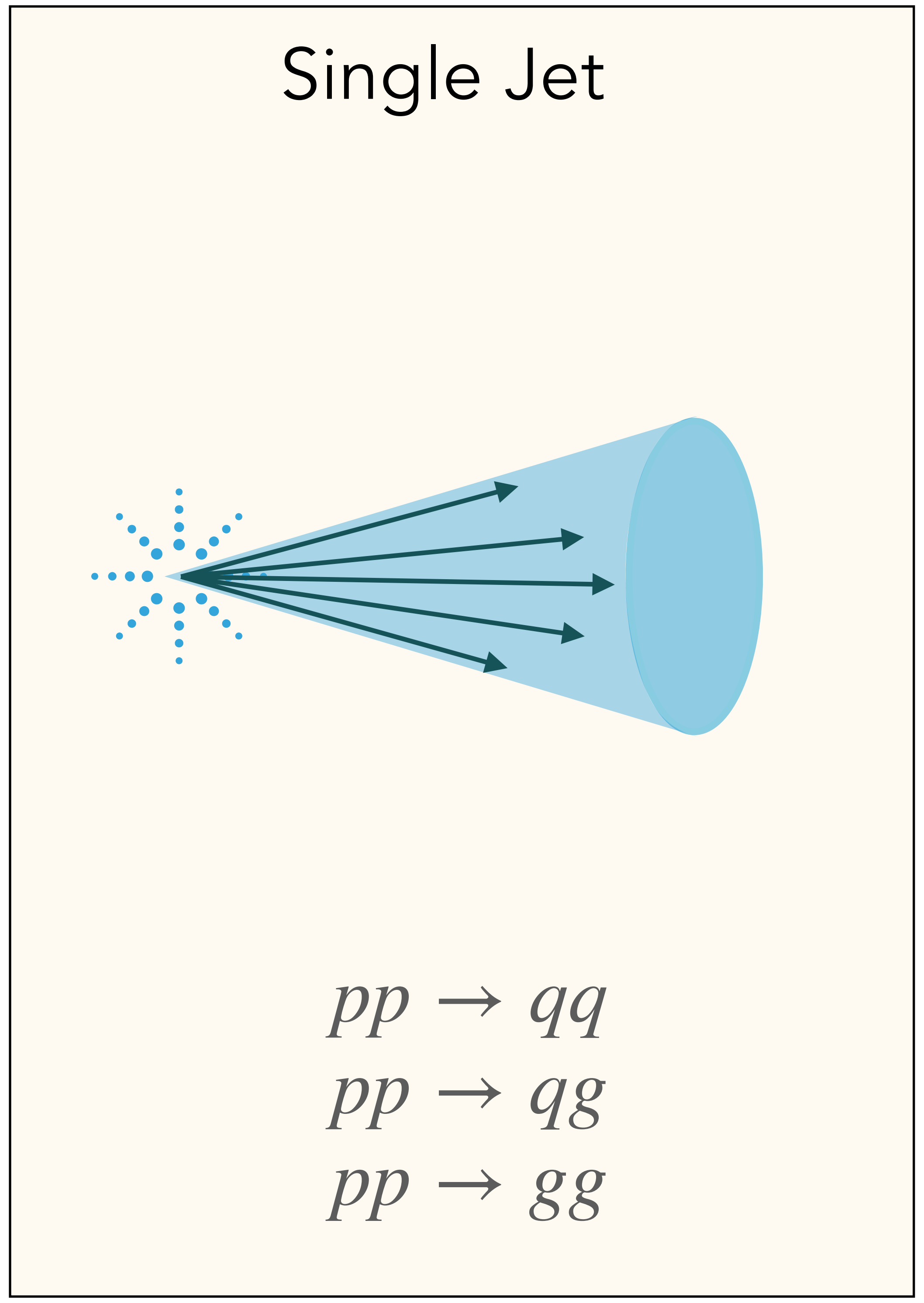


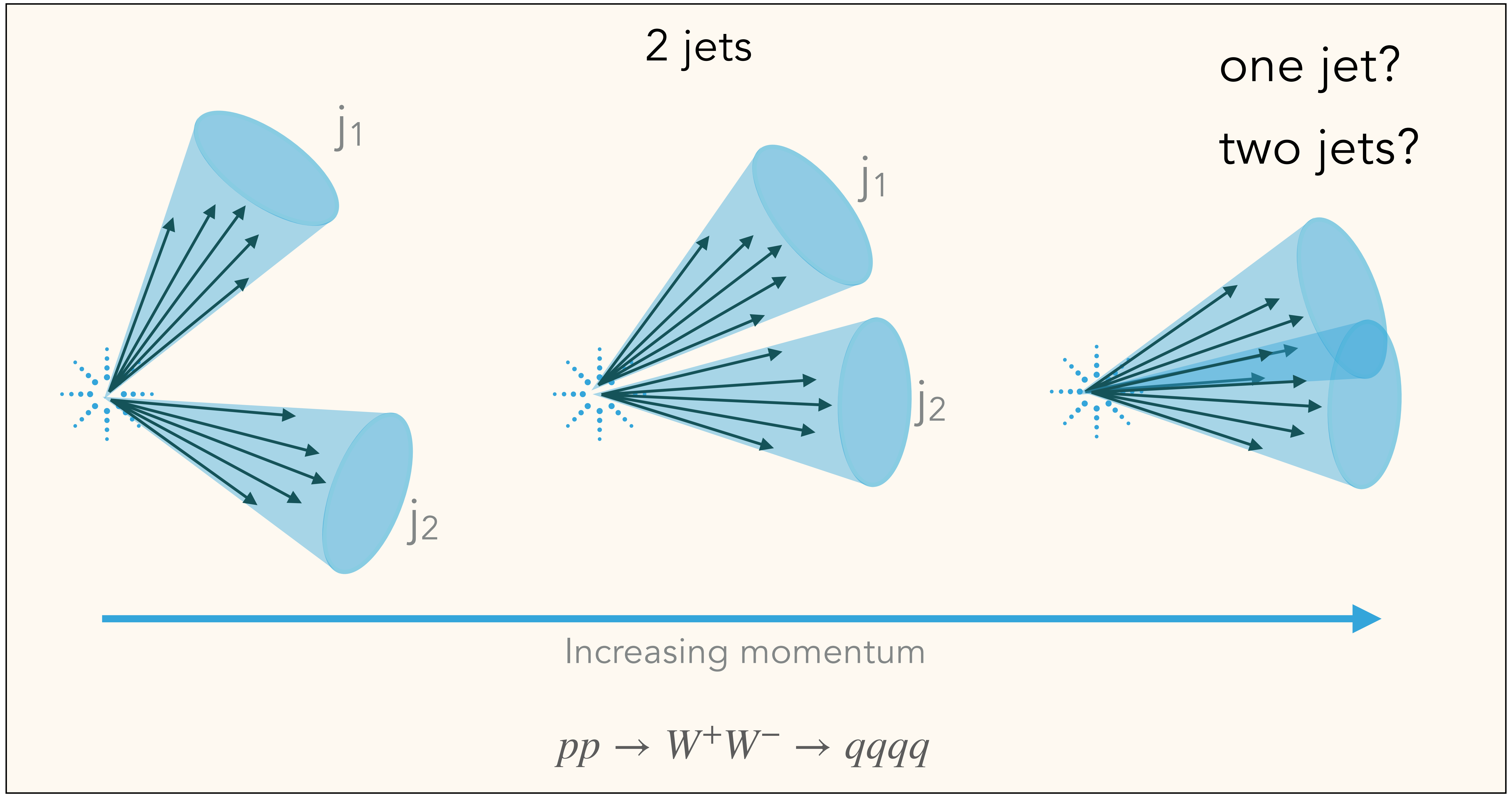
- E_T = Transverse Energy
- Position (η, ϕ)
- $\eta = -\ln(\tan(\theta/2))$

Test Case: Boosted W vs QCD

Boosted W bosons (W → qq') create highly collimated di-jets.

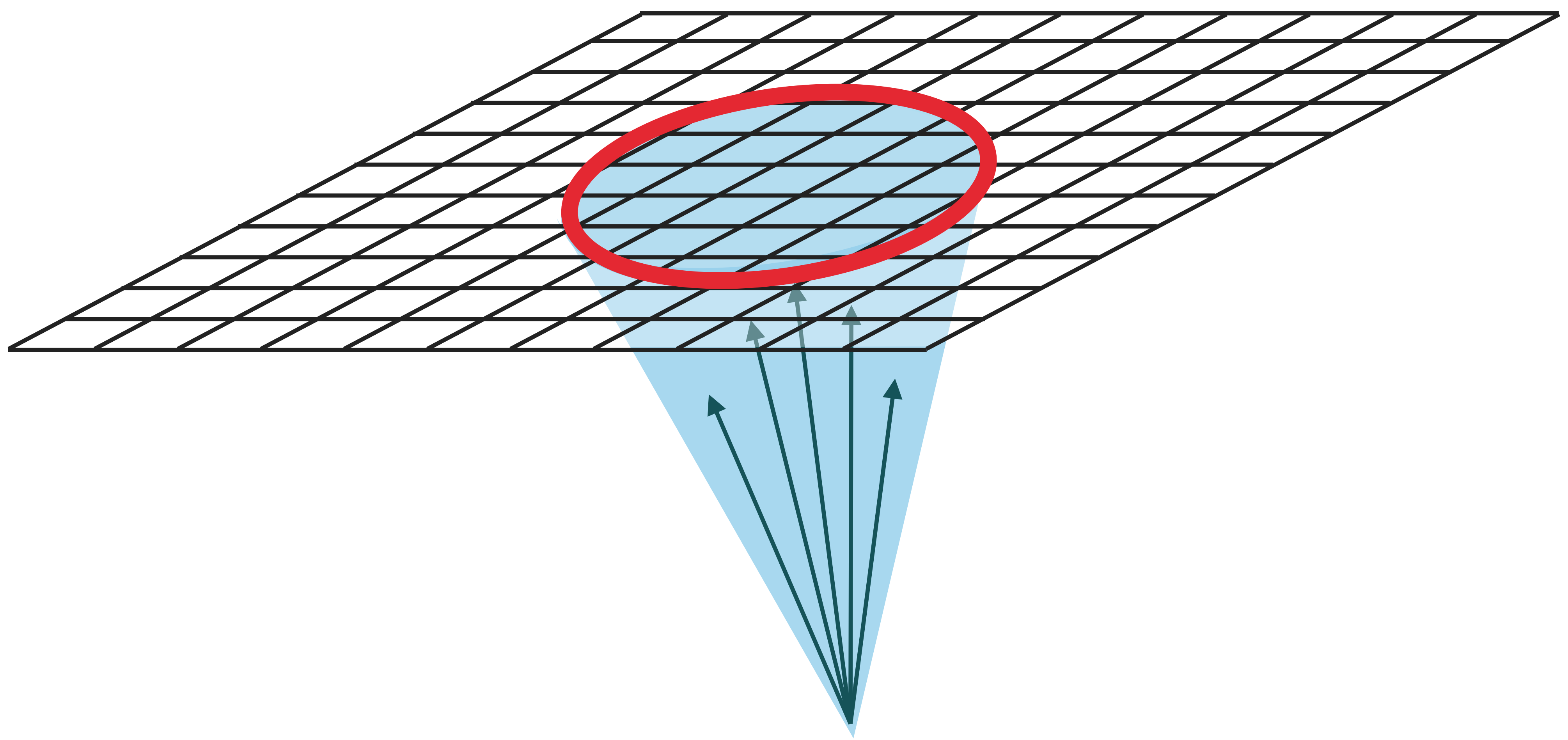
Can a CNN separate boosted W jets from QCD jets?

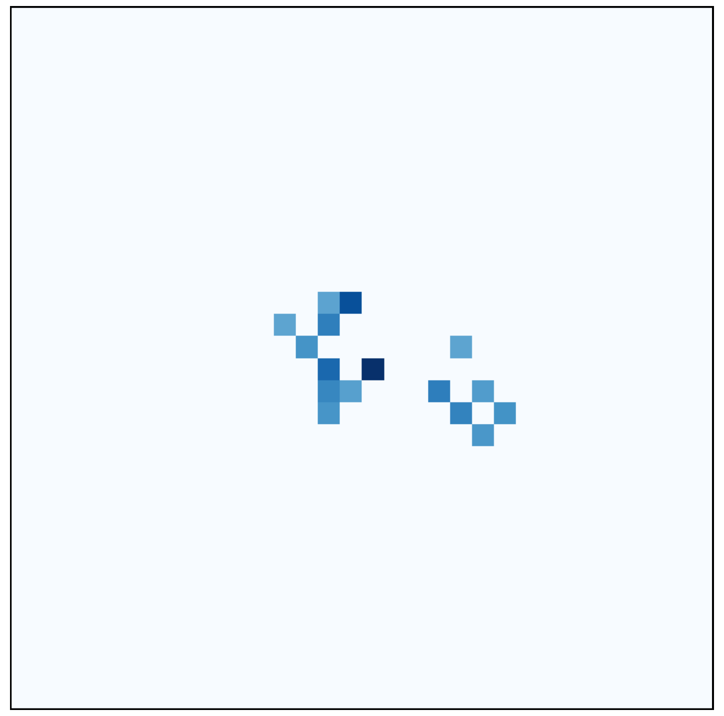




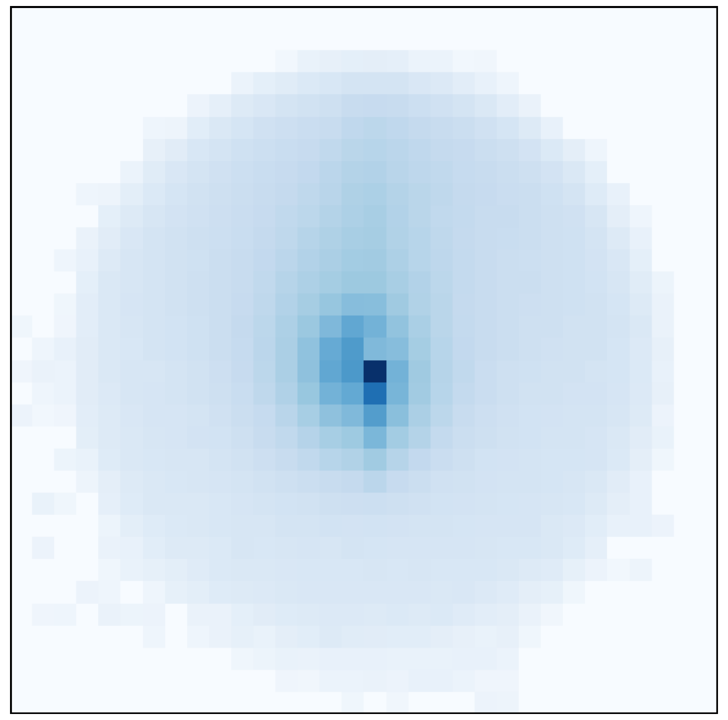
QCD mono-jet vs Boosted W di-jet

QCD Jet (q, g)



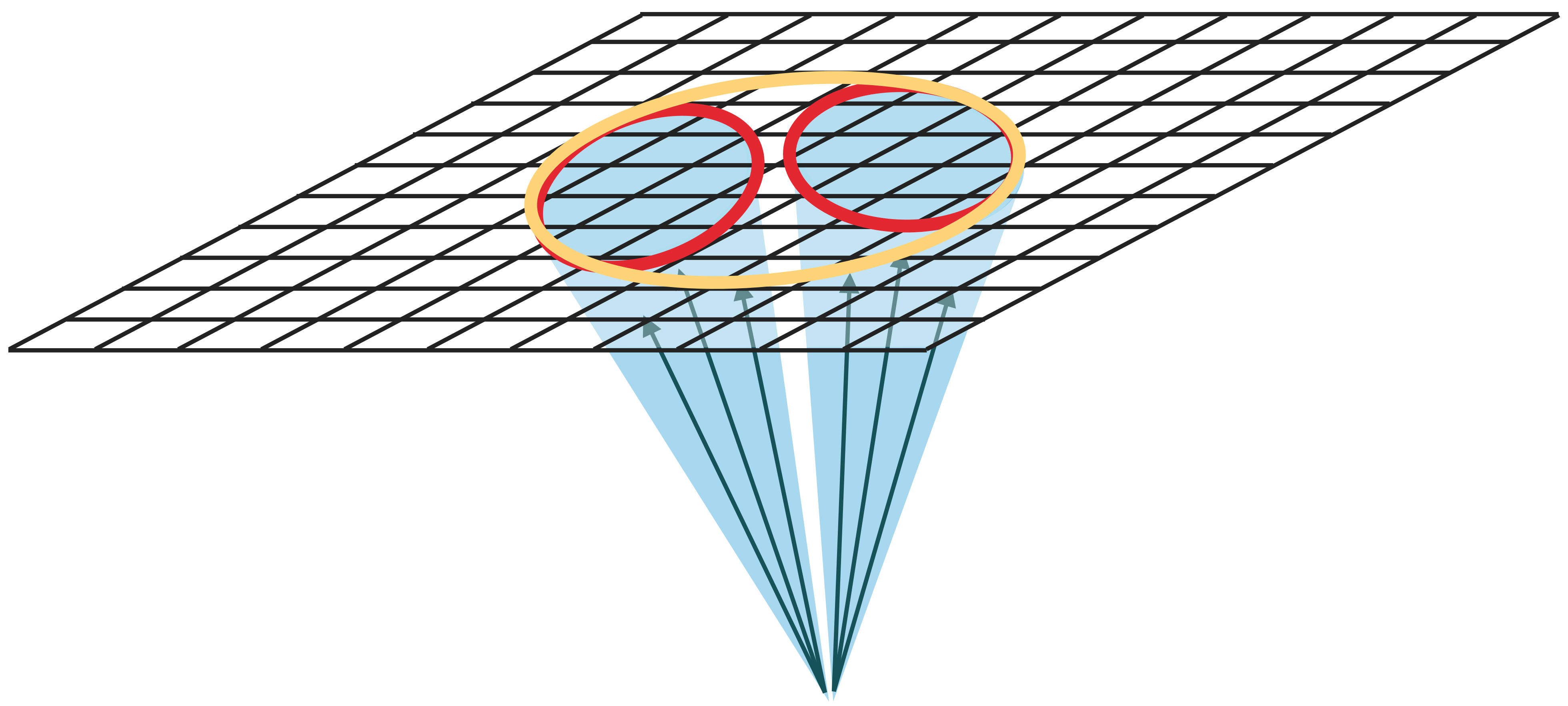


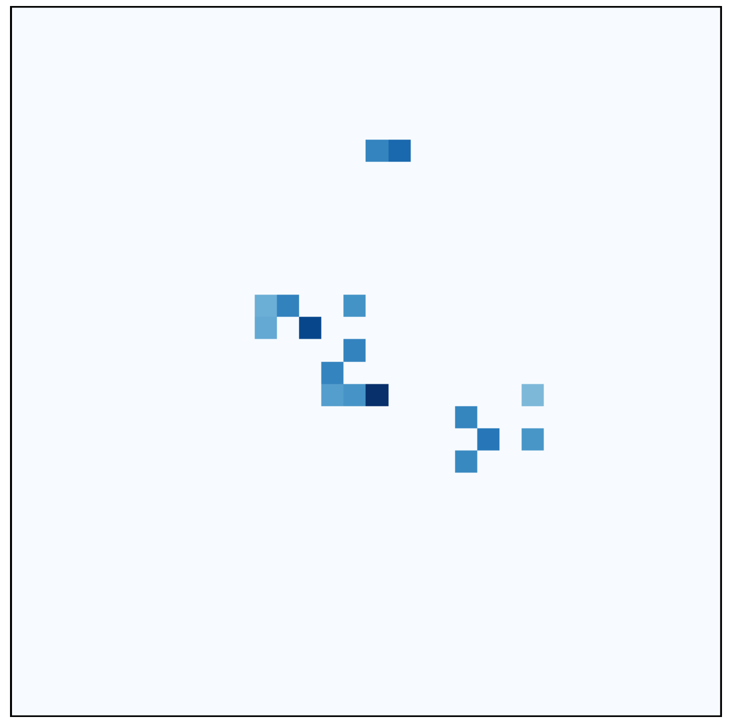
1 Event



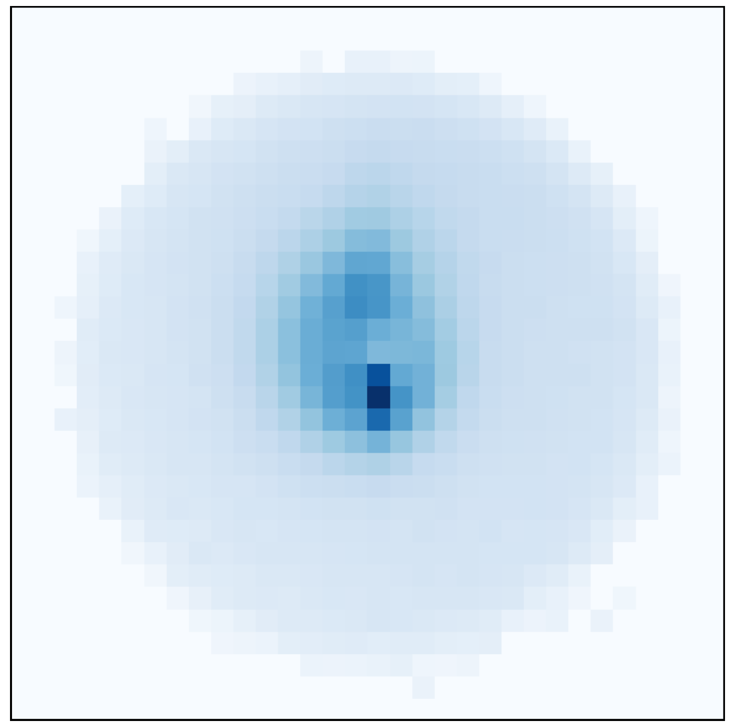
Average of all events

W jet





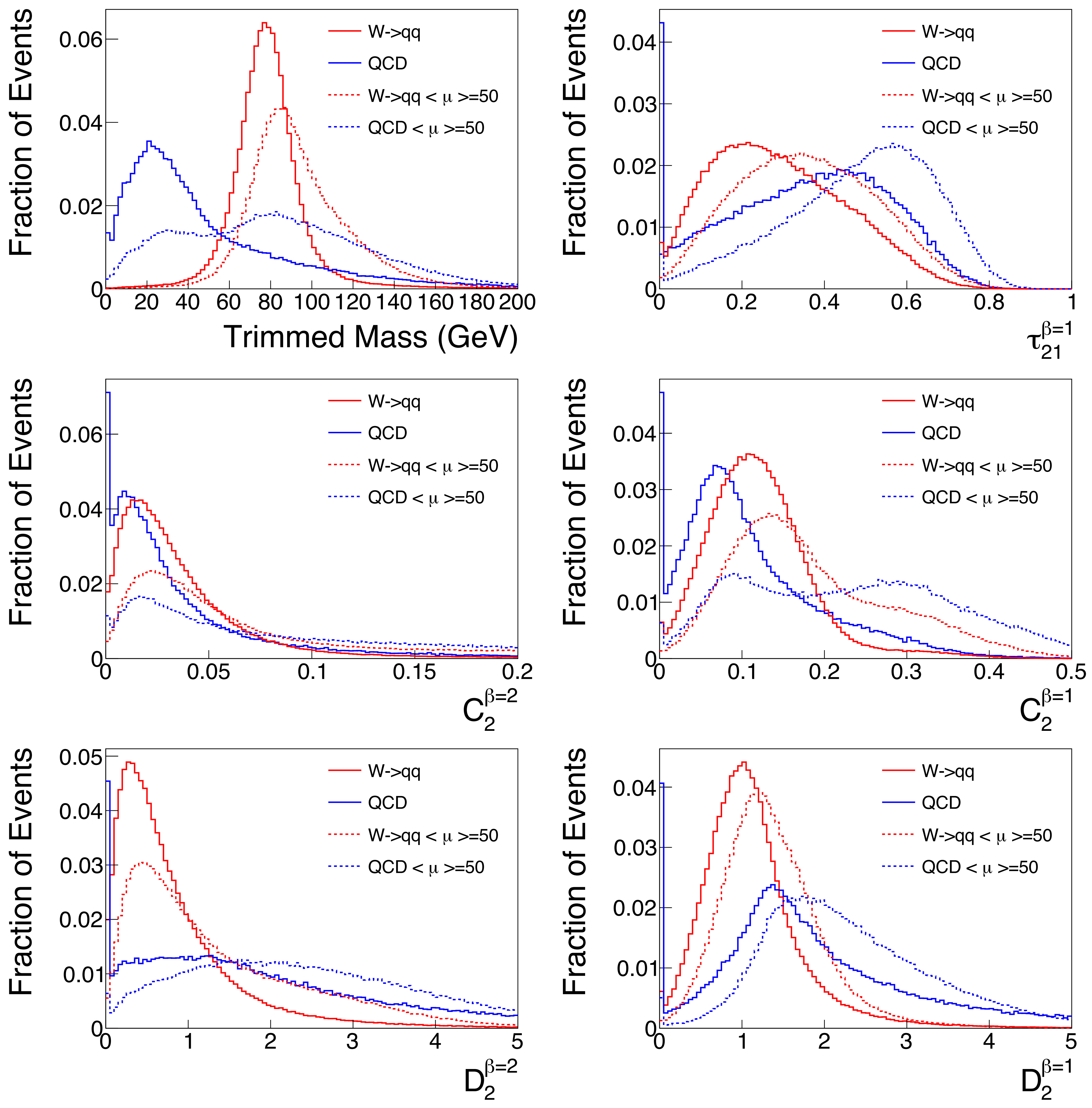
1 Event



Average of all events

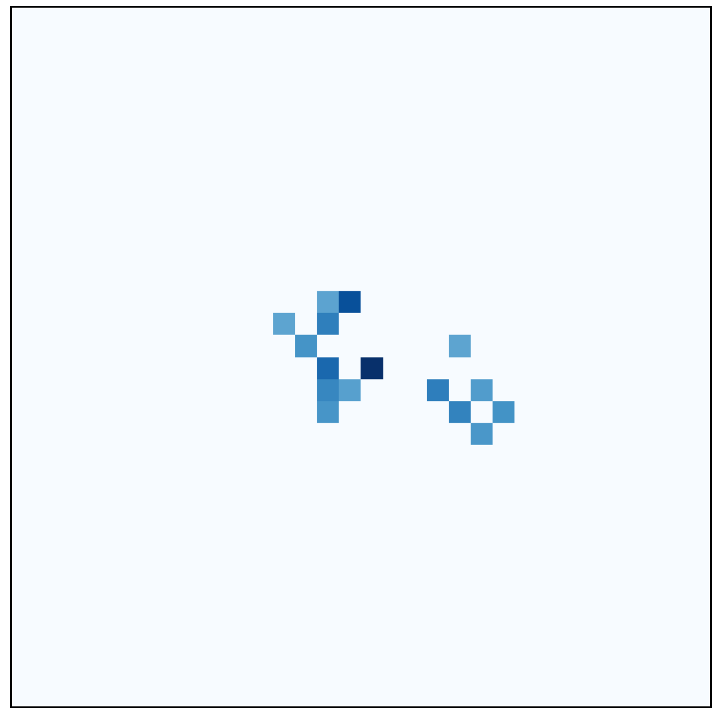
High-Level (HL) vs Low-Level (LL)

6 HL Variables (HL network)

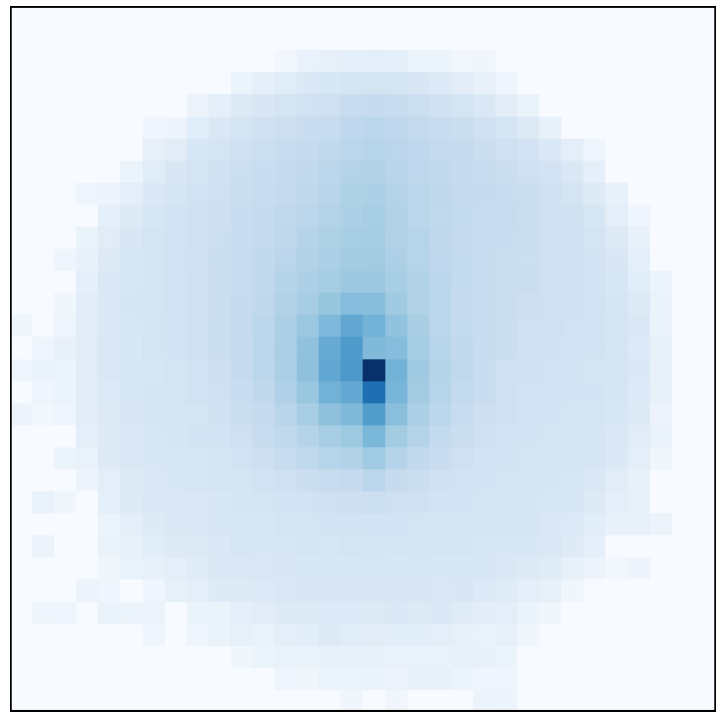


vs.

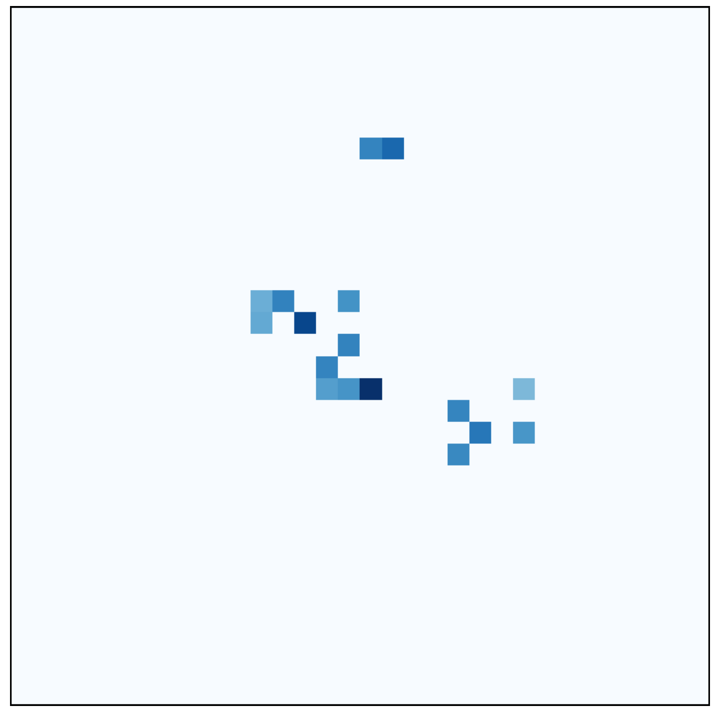
Jet Images (LL network)



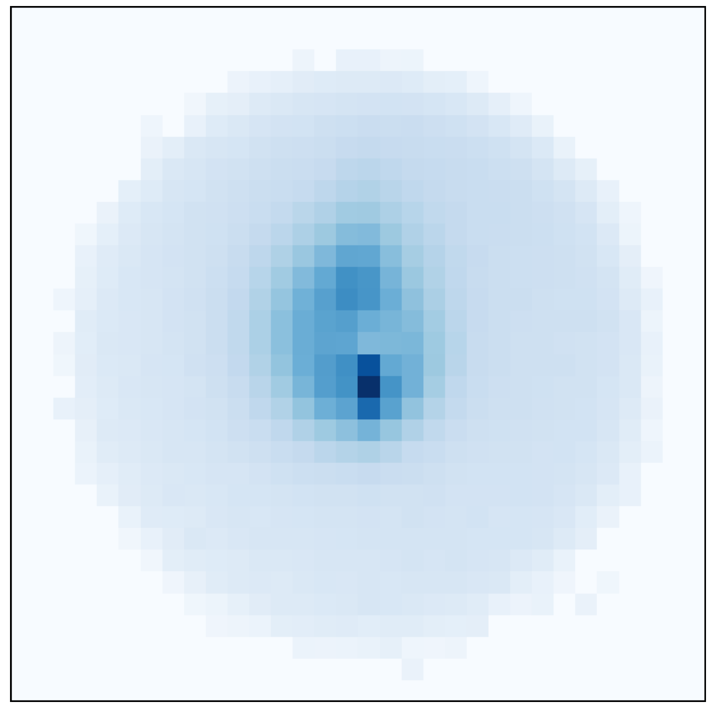
1 Event



Average of all events



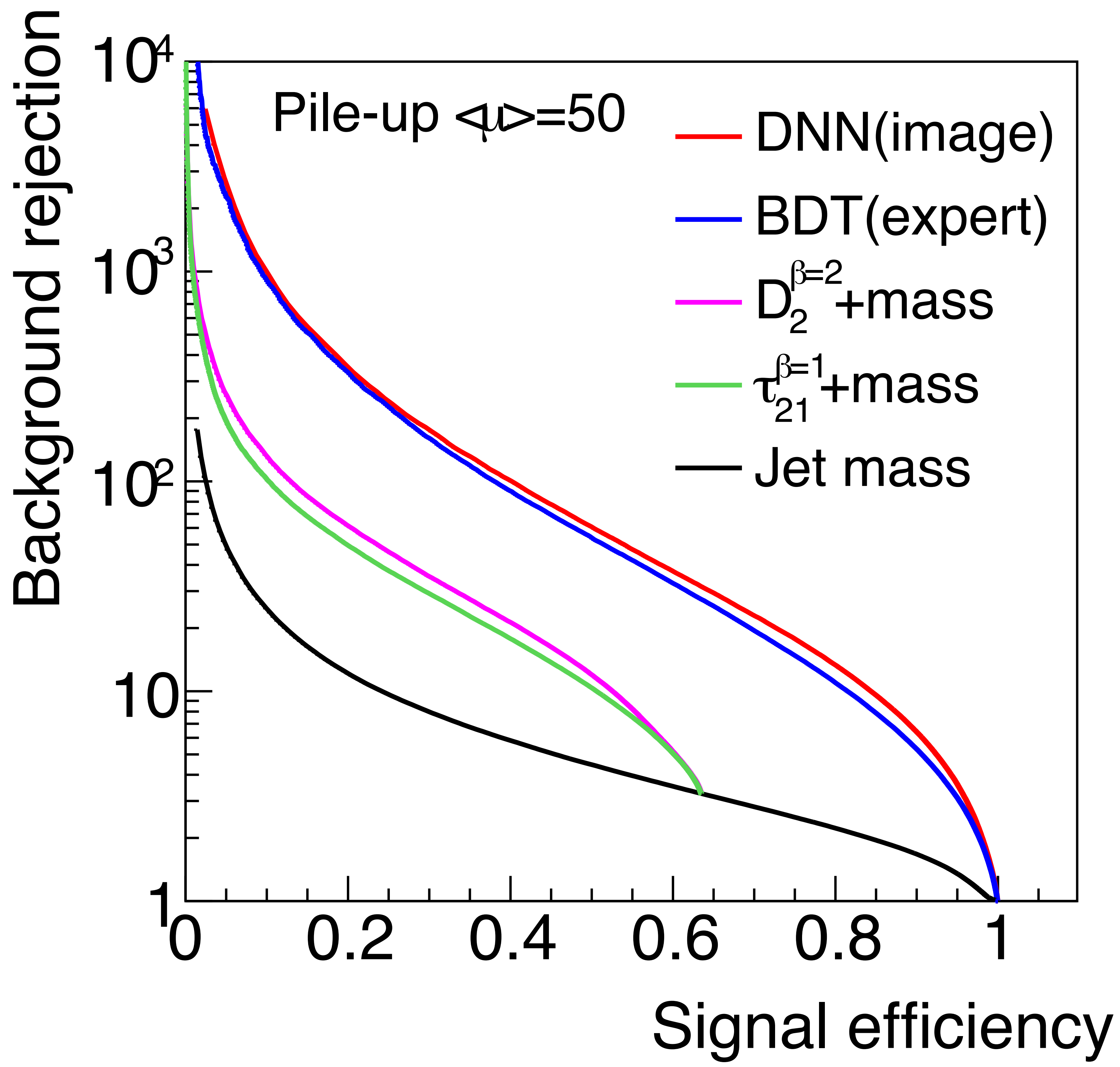
1 Event



Average of all events

Jet Images out-perform JSS

- CNN performs better than Jet Substructure
  - **LL network (red line)**: AUC = 95.30% ± 0.02%
  - **HL network (blue line)**: AUC = 95.00% ± 0.02%
- But wait!
  - Where is that extra information coming from?
  - Why don't Jet Substructure observables contain this information?
- We've used a black box, so how can we investigate?



Baldi, P., Bauer, K., Eng, C., Sadowski, P., & Whiteson, D. (2016, March 30). Jet Substructure Classification in High-Energy Physics with Deep Neural Networks. arXiv.org. http://doi.org/10.1103/PhysRevD.93.094034

Another approach to generalizing JSS

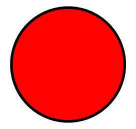
- We weren't the only ones thinking about a generalized approach to understanding JSS!

Jesse Thaler, Patrick Komiske, Eric Metodiev

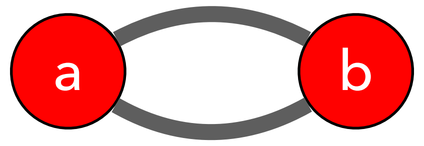
https://arxiv.org/abs/1712.07124

- Energy Flow Polynomials (EFP): A complete linear basis set for jet substructure.

Graph Components

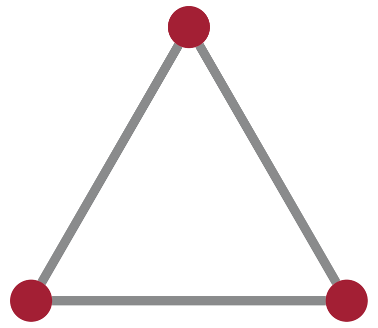
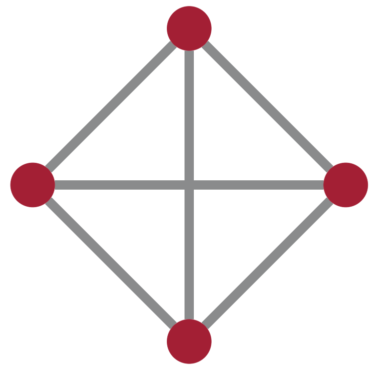
 $= \sum_{a}^{N} z_a$

$= \theta_{ab}$



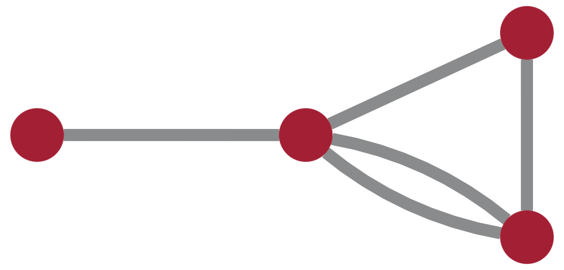
$= (\theta_{ab})^2$

Existing JSS exists in the EFP space

$e_2^{(\beta)} =$  , $e_3^{(\beta)} =$  , $e_4^{(\beta)} =$  ,

Energy Correlation

We can also explore more exotic observables.



$$= \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \sum_{i_3=1}^{M} \sum_{i_4=1}^{M} z_{i_1} z_{i_2} z_{i_3} z_{i_4} \theta_{i_1 i_2} \theta_{i_2 i_3} \theta_{i_2 i_4}^2 \theta_{i_3 i_4}.$$

With variables and parameters (κ,β)

$$z_i^\kappa = \left(\frac{p_{T,i}}{\sum_i p_{T_j}} \right)^\kappa \qquad \theta_{ij}^\beta = \left(\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2 \right)^{\beta/2}$$

Marry the two methods

1. Black-box Learning
    - **Benefits:** Powerful Performance and no need to hand-pick observables
    - **Drawbacks:** Not interpretable. What are we learning about the problem?
2. Energy Flow Polynomials (EFP)
    - **Benefits:**
        - Physics motivated
        - Modeling can be verified
        - Uncertainties can be defined
        - Compact and efficient
    - **Drawbacks:** It's an infinite space. How do we begin to choose observables and their parameters?

Combine them! The DNN has learned how to solve the problem. Let the DNN tell us what EFPs to choose!

Decision Ordering: A New Metric for Decision Similarity

- Consider boosted W vs QCD jet binary classification
- CNN has learned where to draw an ideal decision surface in its feature space.
- We want a HL feature space that makes equally good decisions.
- How do we compare the decision surface for the CNN to the HL features?

Comparing pair orderings

- Take a pair of signal (x) and background (x′) features,
- Predictions for 2 NN (f(x) and g(x)) of the features will increase/decrease relative to input

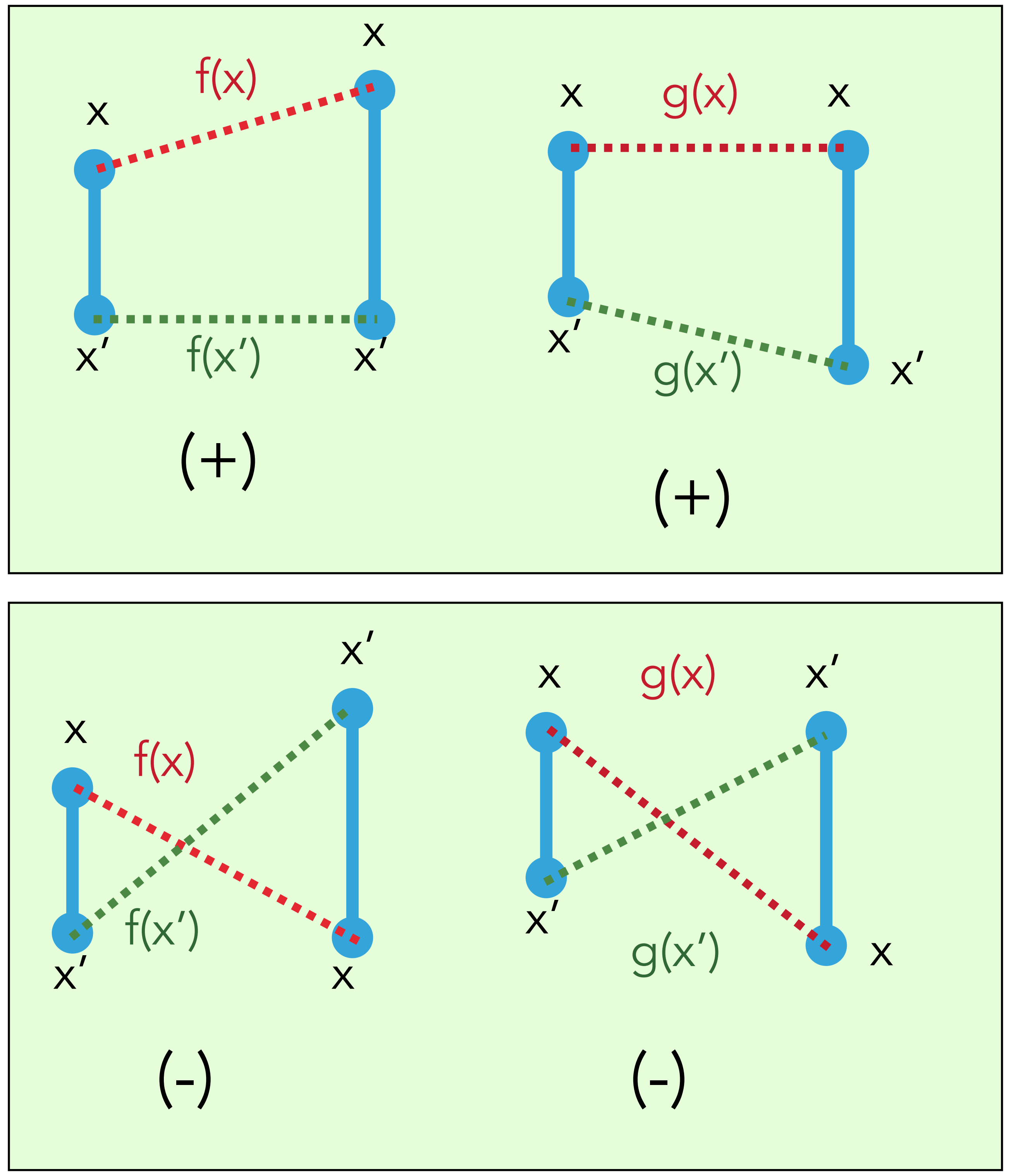
$$\mathrm{DO}(x, x') = \Theta\left[(f(x) - f(x')) \cdot (g(x) - g(x'))\right]$$

Heaviside step function (Θ) sets DO=1 for similar order, DO=0 for dissimilar order.

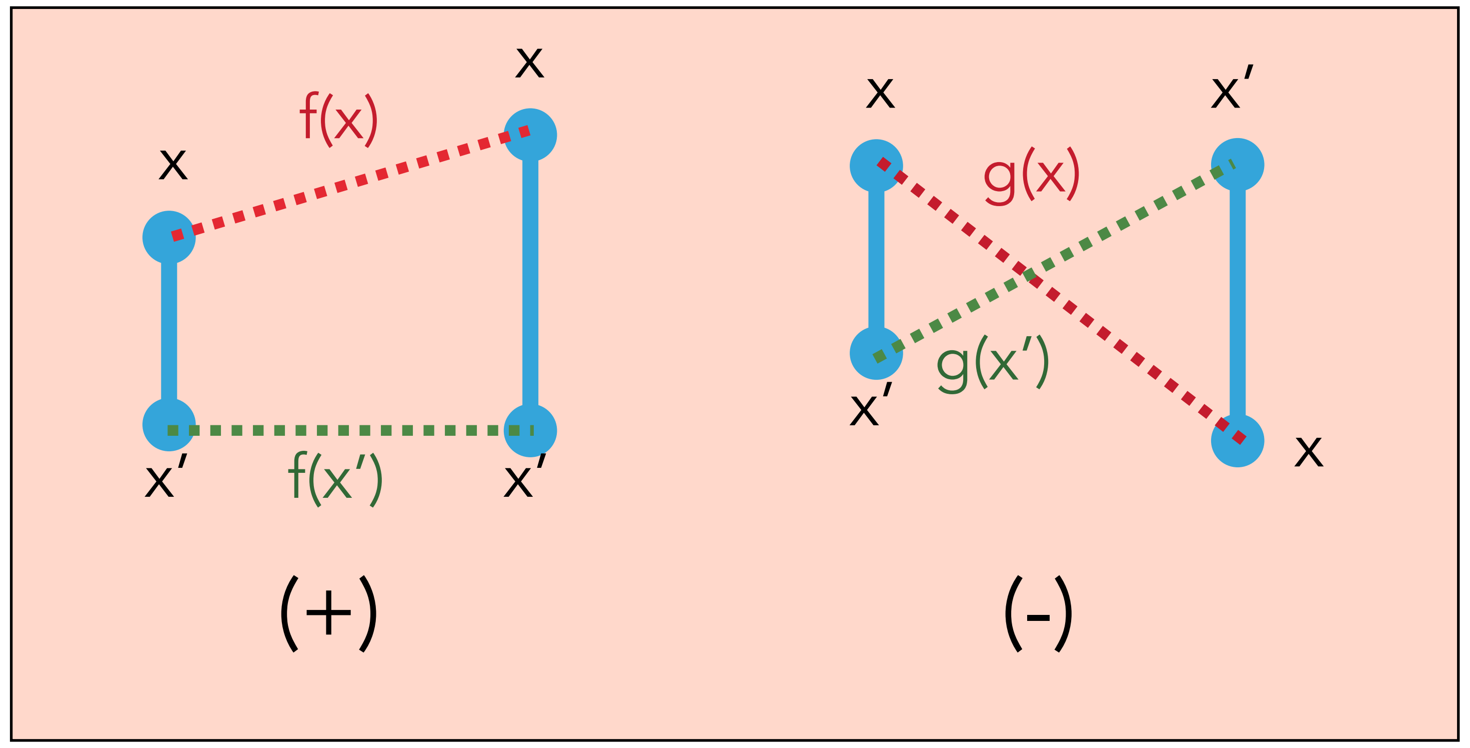
Average many examples = Average Decision Ordering

$$\mathrm{ADO}' = \sum DO(x, x')$$

Similar Orderings



Dissimilar Orderings

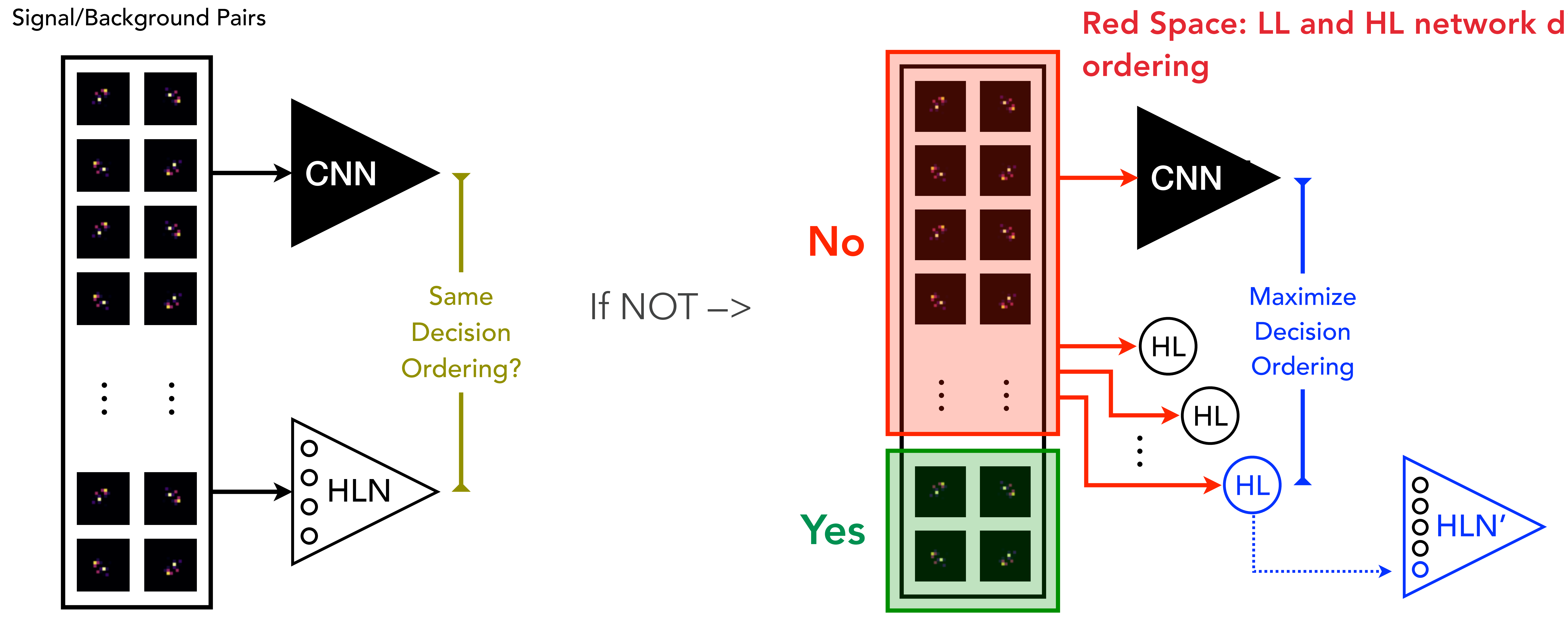


ADO = 1 : *Identical decisions*

ADO = 0.5 : *Random similarity*

Why is ADO useful?

We can compare NN decision making. Where does the HL network and LL network disagree?



Use ADO to choose EFP that makes similar choices to the LL network in the “differently ordered” red space

Finding a new EFP with Guided Search

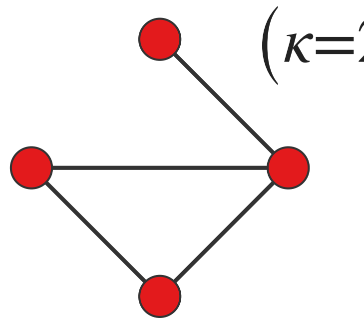
We only need 1 new observable to achieve equal performance with the CNN!

Observable	AUC	ADO[CNN, Obs.]
M_{jet}	0.898 ± 0.004	0.807
$C_2^{\beta=1}$	0.660 ± 0.006	0.584
$C_2^{\beta=2}$	0.604 ± 0.007	0.548
$D_2^{\beta=1}$	0.790 ± 0.005	0.743
$D_2^{\beta=2}$	0.807 ± 0.005	0.762
$\tau_2^{\beta=1}$	0.662 ± 0.006	0.600
6HL	0.9504 ± 0.0002	0.971
CNN	0.9531 ± 0.0002	1.000
$7\text{HL}_{\text{black-box}}$	0.9528 ± 0.0003	0.971

Original HL

Original HL + 1 EFP

Which EFP did we pick?



$$^{(\kappa=2,\ \beta=1/2)} = \sum_{a,b,c,d=1}^{N} z_a^2 z_b^2 z_c^2 z_d^2 \sqrt{\theta_{ab}\theta_{bc}\theta_{ac}\theta_{ad}}$$

Noteworthy details

- EFP is not Infrared-safe (k ≠ 1)
- β=1/2 is probing small-angle behaviour
- Chromatic #3 graph (probing deviations from 2-prong substructure)
- Chromatic Number = Minimum number of prongs to not vanish

Conclusion

Deep networks can identify gaps where low-level data contains unused information

ML Mapping strategies can capture and translate that information into understandable physics

Questions ?