MAY 24TH, 2021

XAI FOR ML JET TAGGERS

Garvita Agarwal, **Lauren Hay,** Ia Iashvili, Benjamin Mannix, Christine McLean, Margaret Morris, Salvatore Rappoccio, Ulrich Schubert

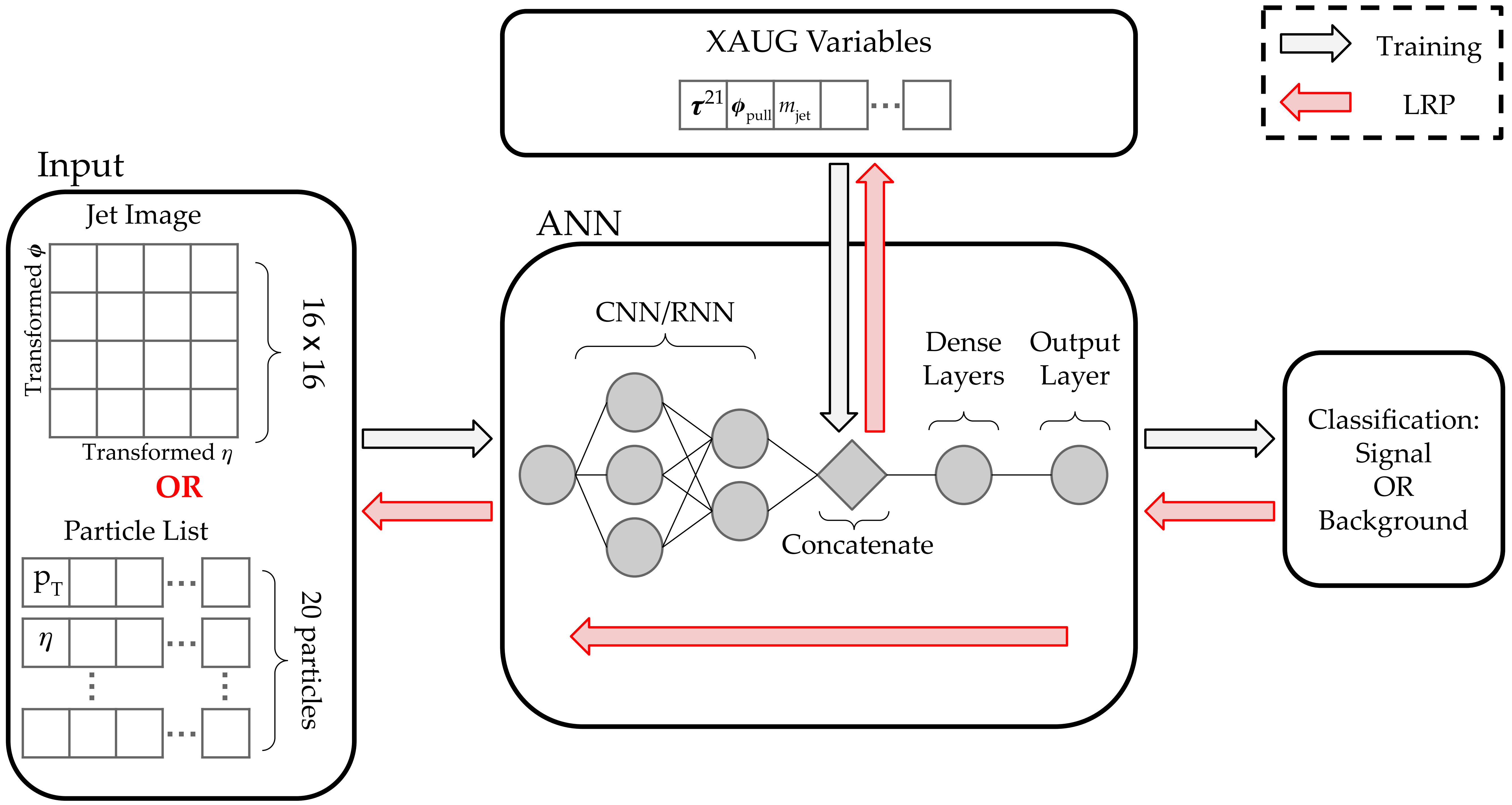
University at Buffalo

https://arxiv.org/abs/2011.13466

INTRODUCTION

XAUG VARIABLES

- Explain ML decisions of a jet classifier using expert augmented (XAUG) variables
- General method: provide XAUG inputs to a jet tagging network, apply LRP to network and compare results to same network without XAUG vars.

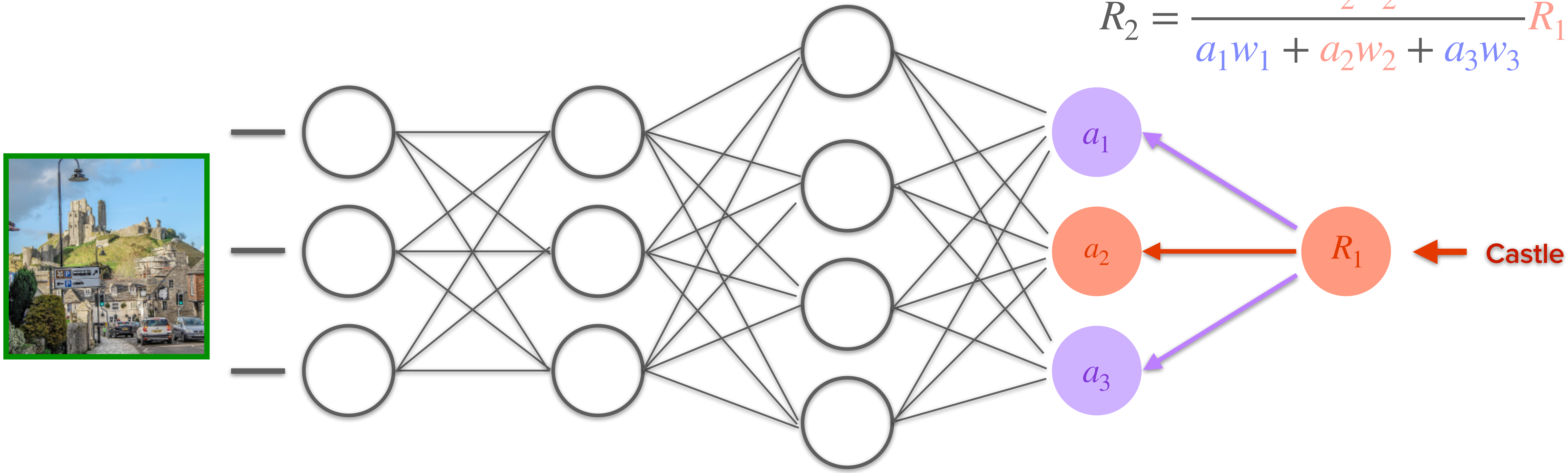


LAYERWISE RELEVANCE PROPAGATION

- **LRP propagates a prediction backwards through the network, assigning a relevance score to each piece of input.**

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

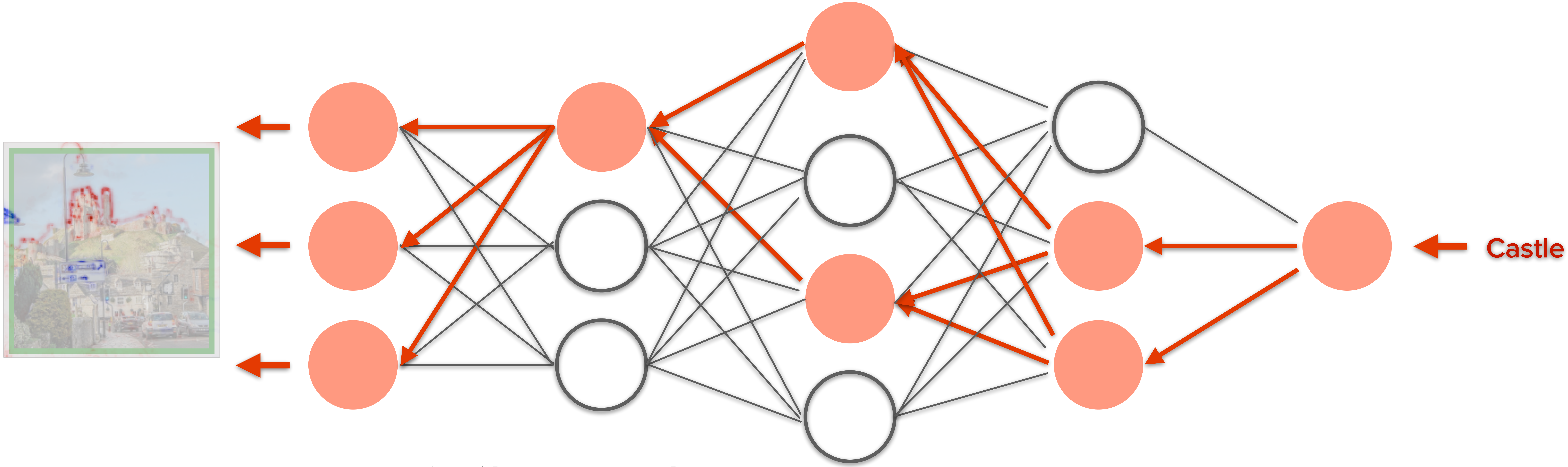
$$R_2 = \frac{a_2 w_2}{a_1 w_1 + a_2 w_2 + a_3 w_3} R_1$$



Layer-Wise Relevance Propagation: An Overvie; G. Montavon , et al. (2019) [https://doi.org/10.1007/978-3-030-28954-6_10]

LRP

- Due to a conservation of relevance, the backwards propagation process does not alter the prediction
- LRP attributes the entirety of the network's decision to the inputs, which can be visualized as a heat map for images

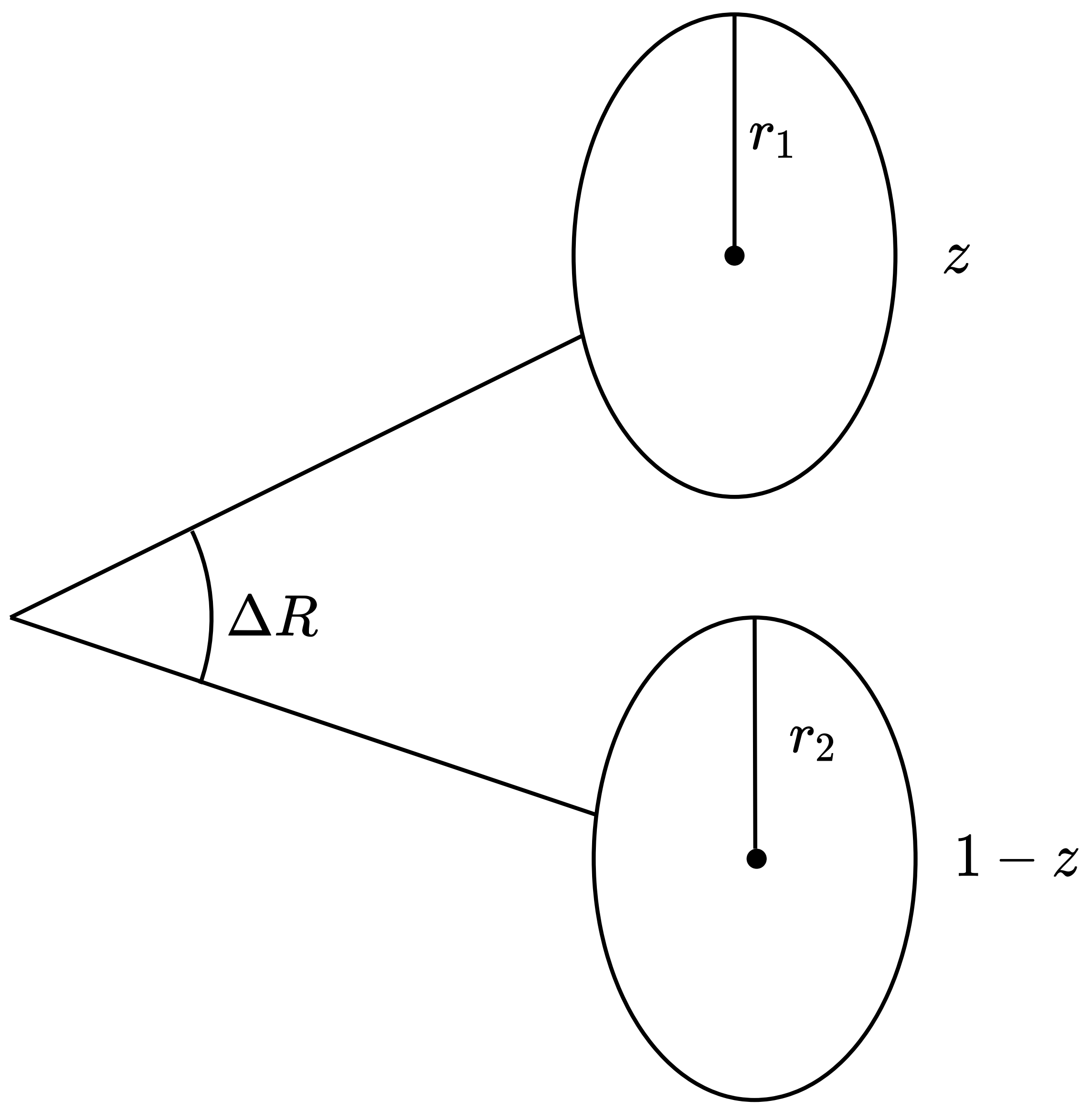


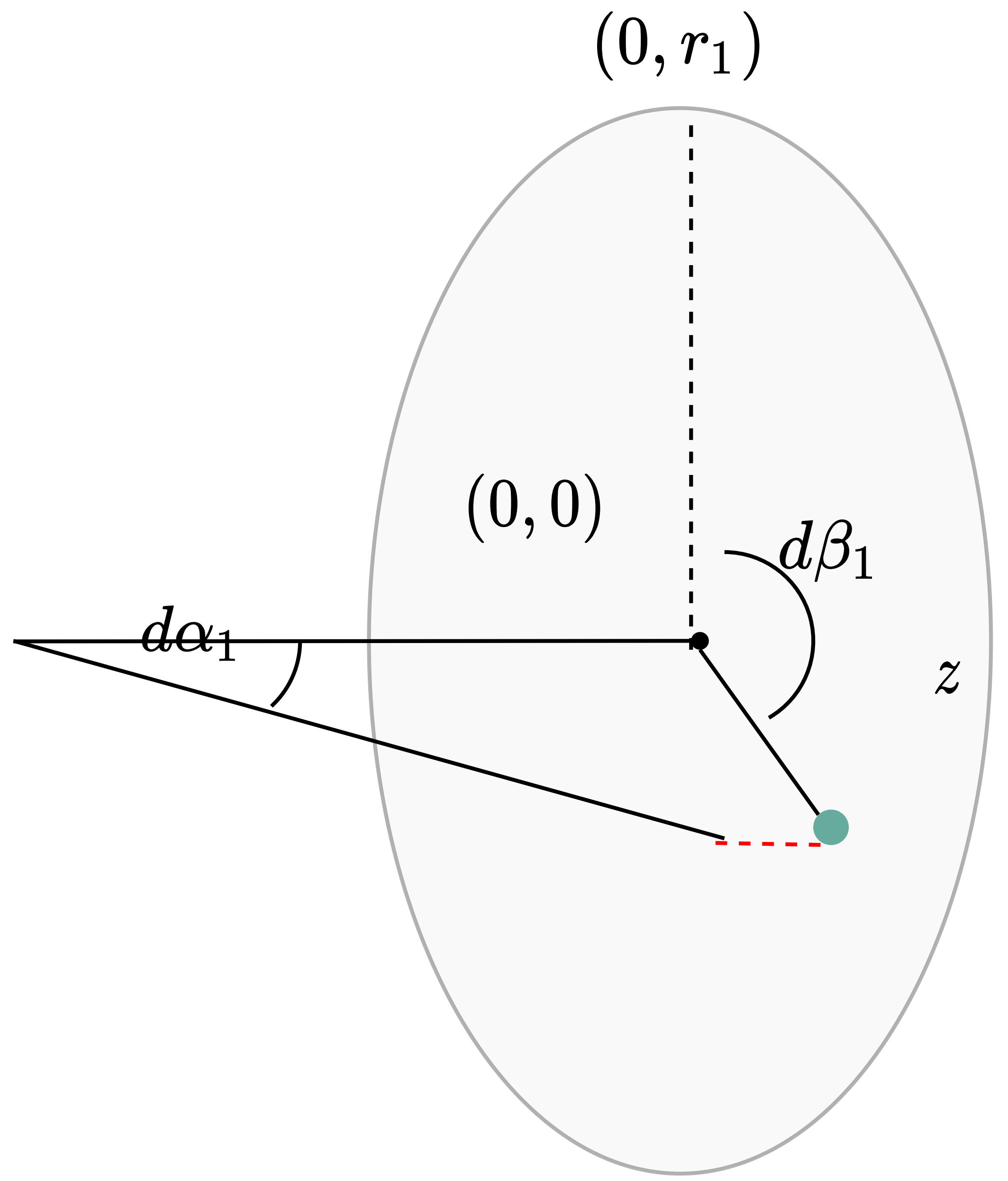
iNNvestigate Neural Networks! M. Alber, et al. (2018) [arXiv:1808.04260]

TOY MODEL

TOY MODEL

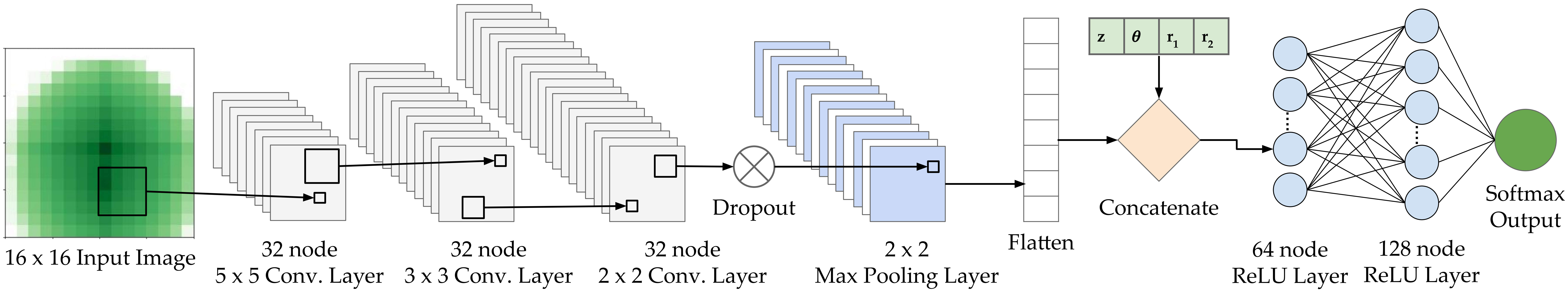
- Toy events simulated to mimic particle level events with 1 jet consisting of 20 particles, divided evenly between 2 subjets
- Goal is to have a small number of variables capture all the information in the event
- The z and θ (ΔR) values are sampled from a normal distribution for "signal-like" images and from exponential distribution for "background-like" images





2DCNN

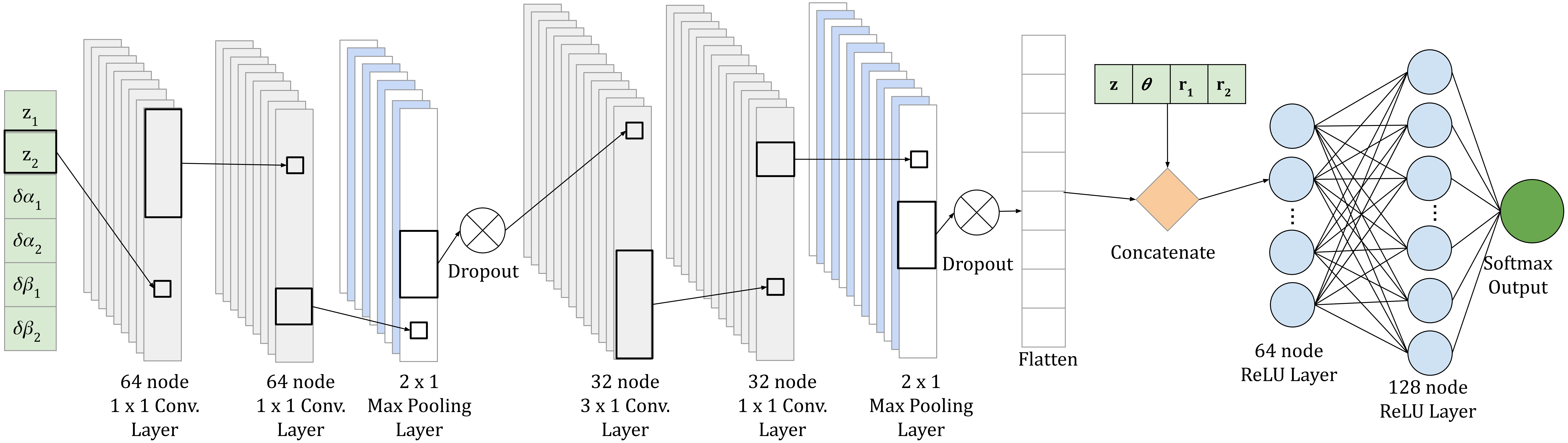
- **Architecture based on ImageTop network**



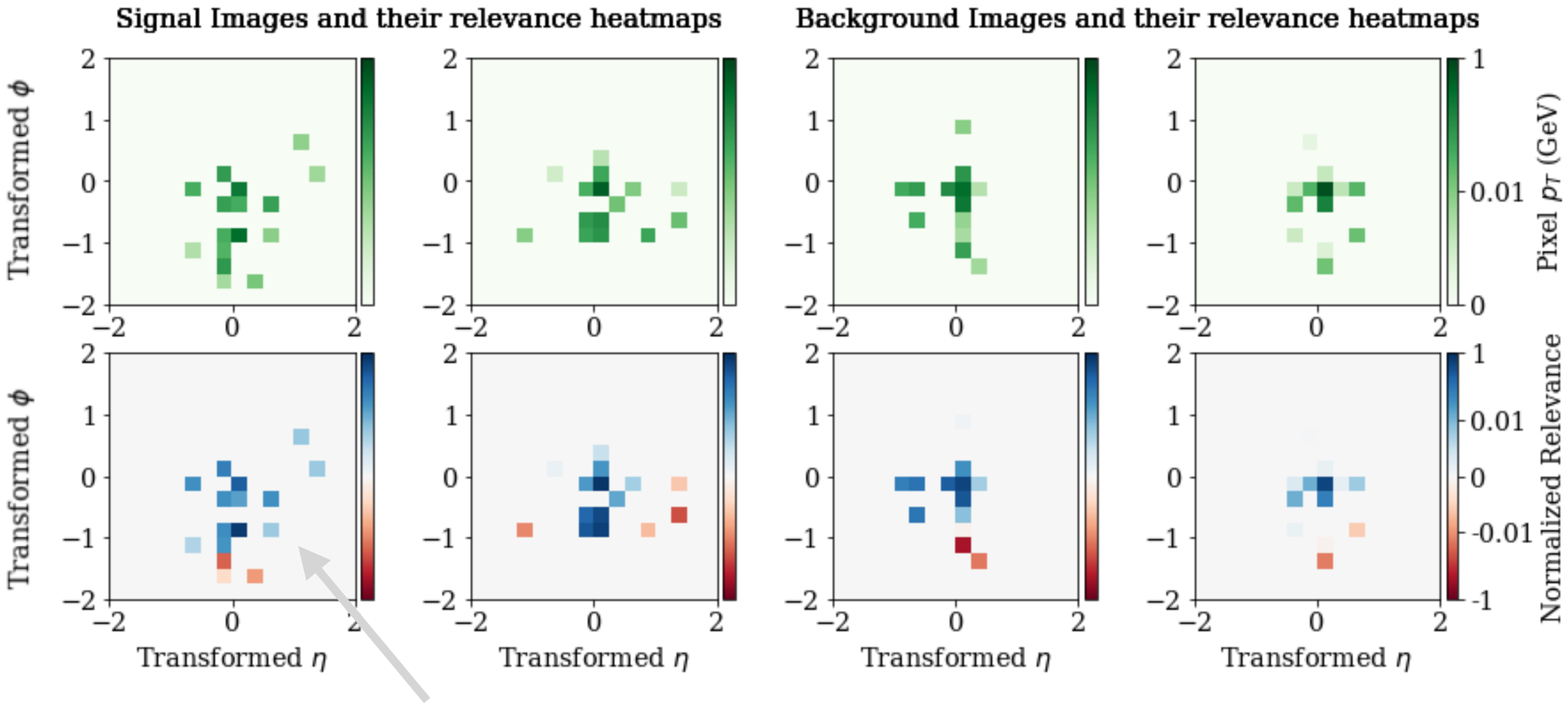
Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques; CMS Collaboration (2020) [arXiv:2004.08262]

1DCNN

- Architecture inspired by DeepAK8 jet tagging algorithm



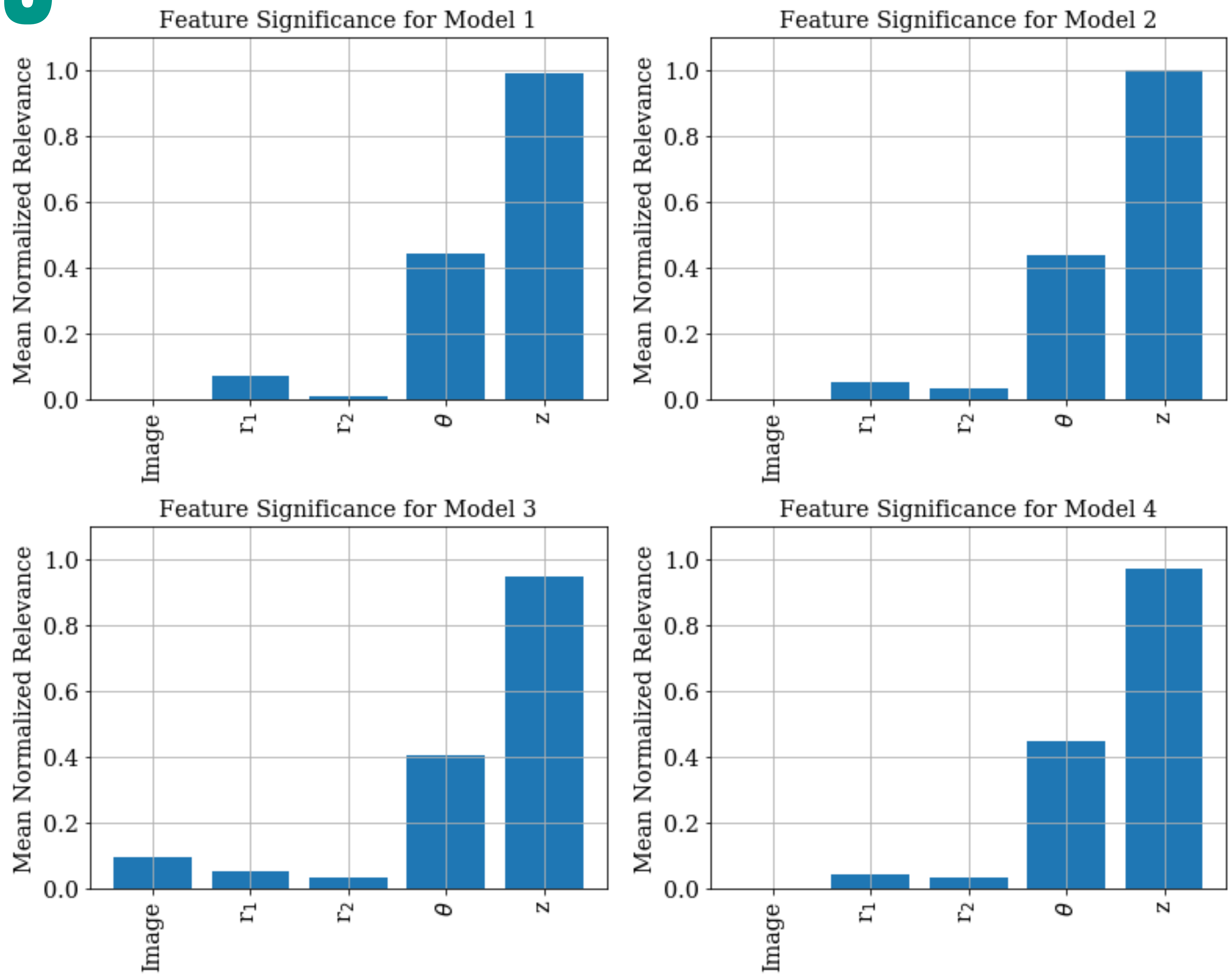
TOY 2DCNN RESULTS



More relevance is given along the ϕ axis in the signal images.

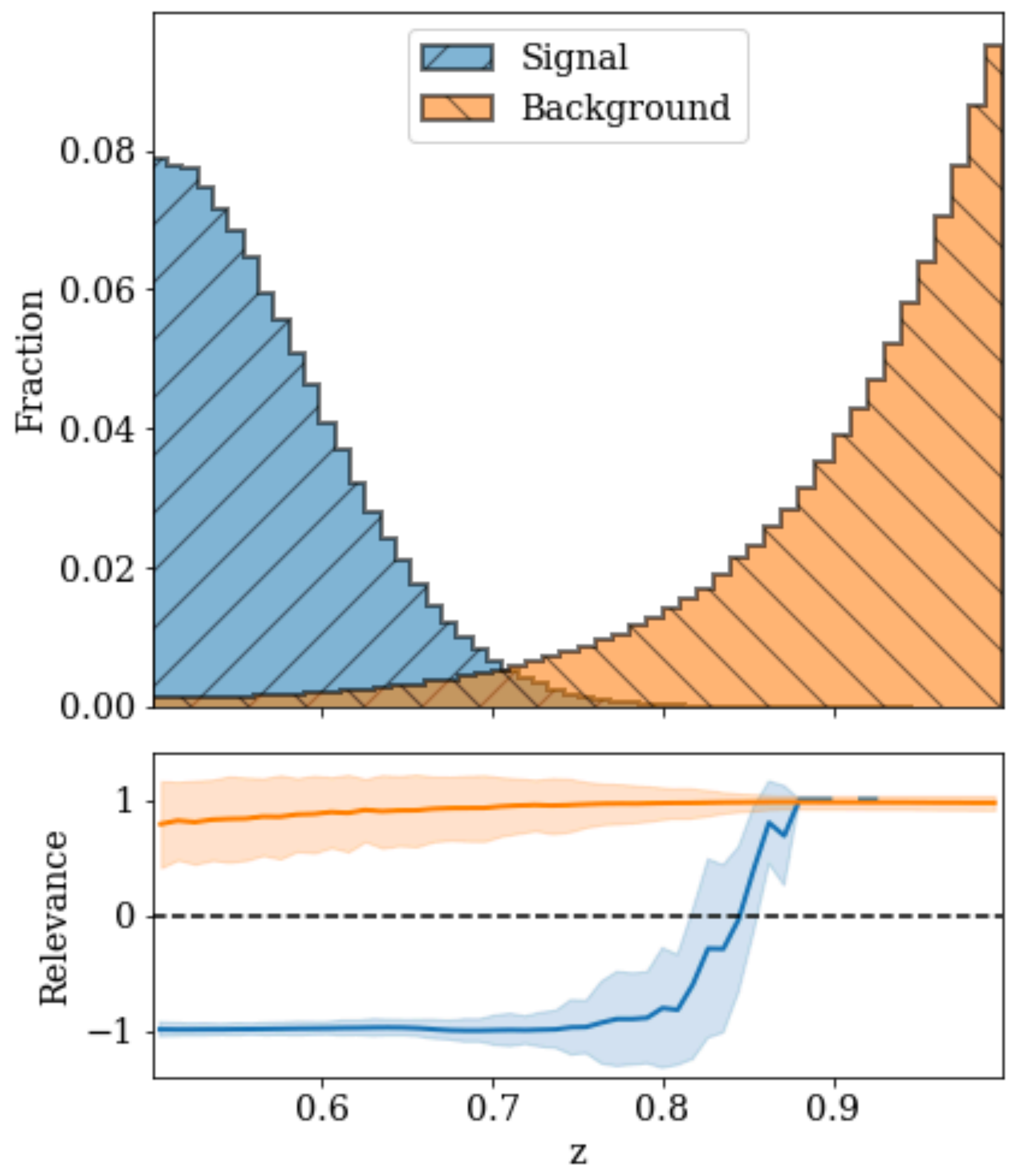
TOY 2DCNN RESULTS

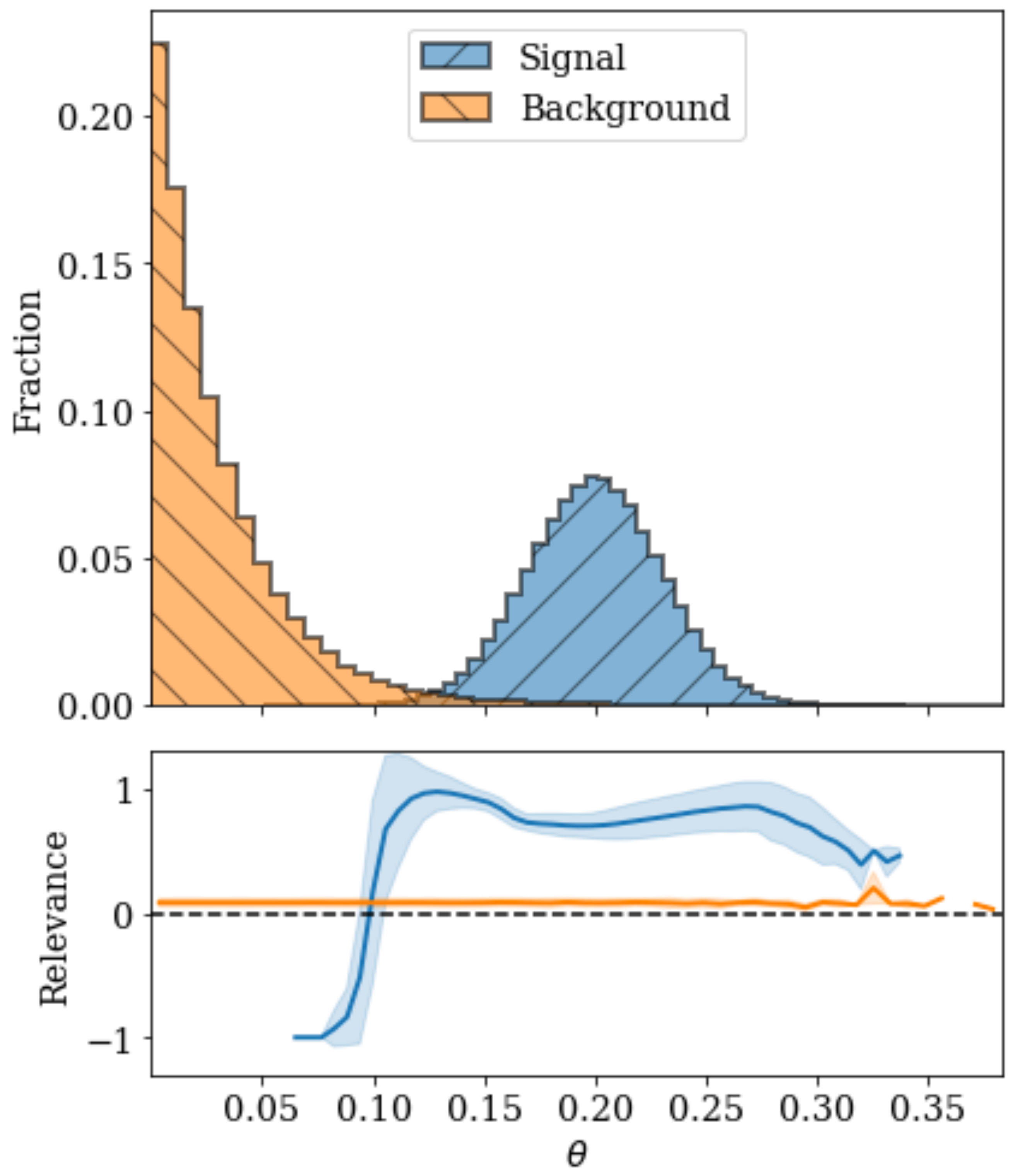
- Mean Normalized Relevance
  - Find feature with max absolute LRP score and divide all scores by this max value
  - For each image, sum absolute value of normalized pixels to get a single number for each image
  - Average absolute relevance scores across all events for each feature



TOY 2DCNN RESULTS

- Profile plots show the relevances vs the corresponding input variables
- For some profiles relevance appears to reflect input distribution, but other don't — networks' decision boundaries live in a higher dimensional space

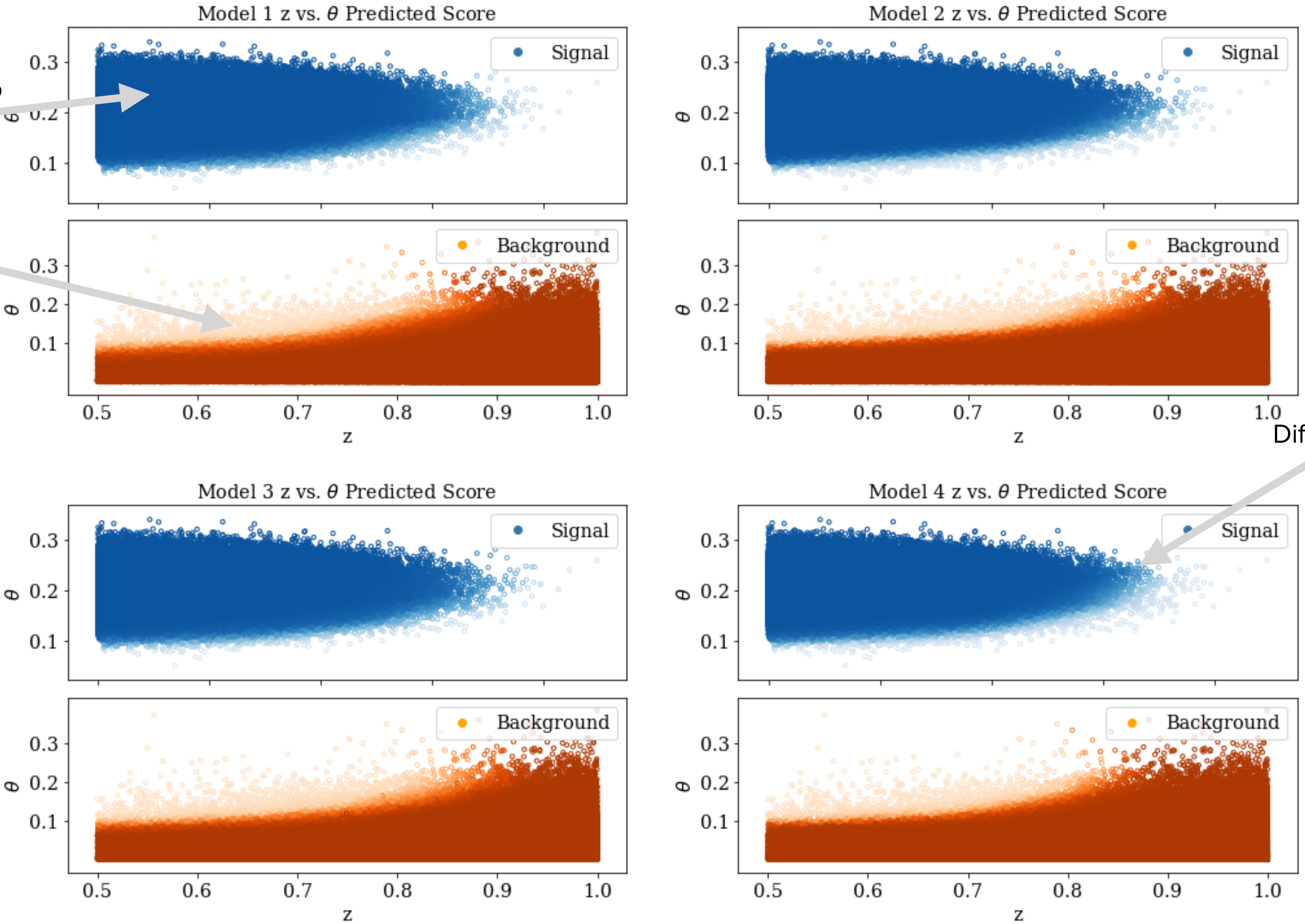




TOY 2DCNN RESULTS

Darker markers corresponds to higher relevance scores.

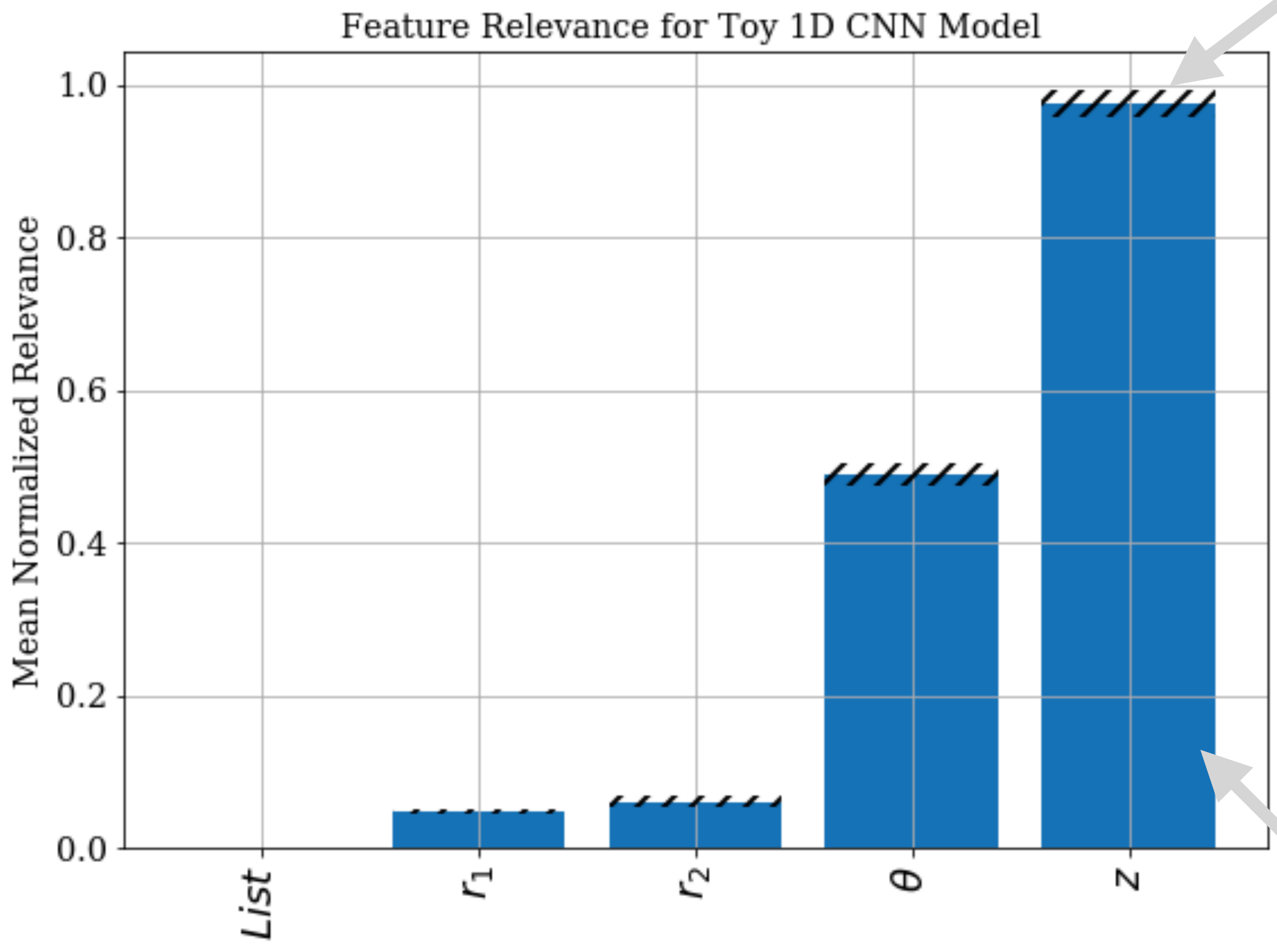
Sharp gradient shows decision boundary for these variables

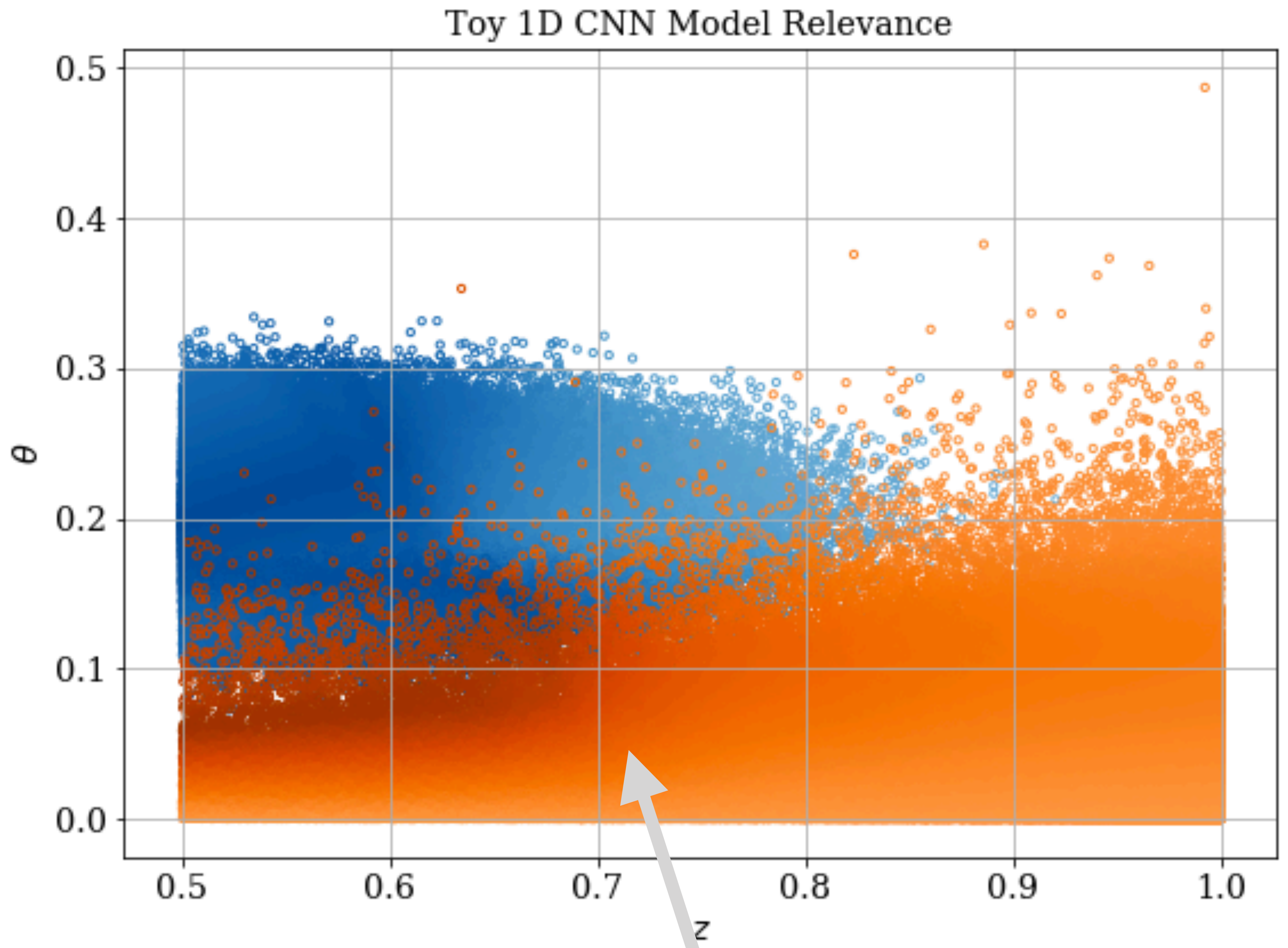


Differences in boundary shape show how trainings vary

TOY 1DCNN RESULTS

Error bars show standard deviation of relevance after multiple trainings.





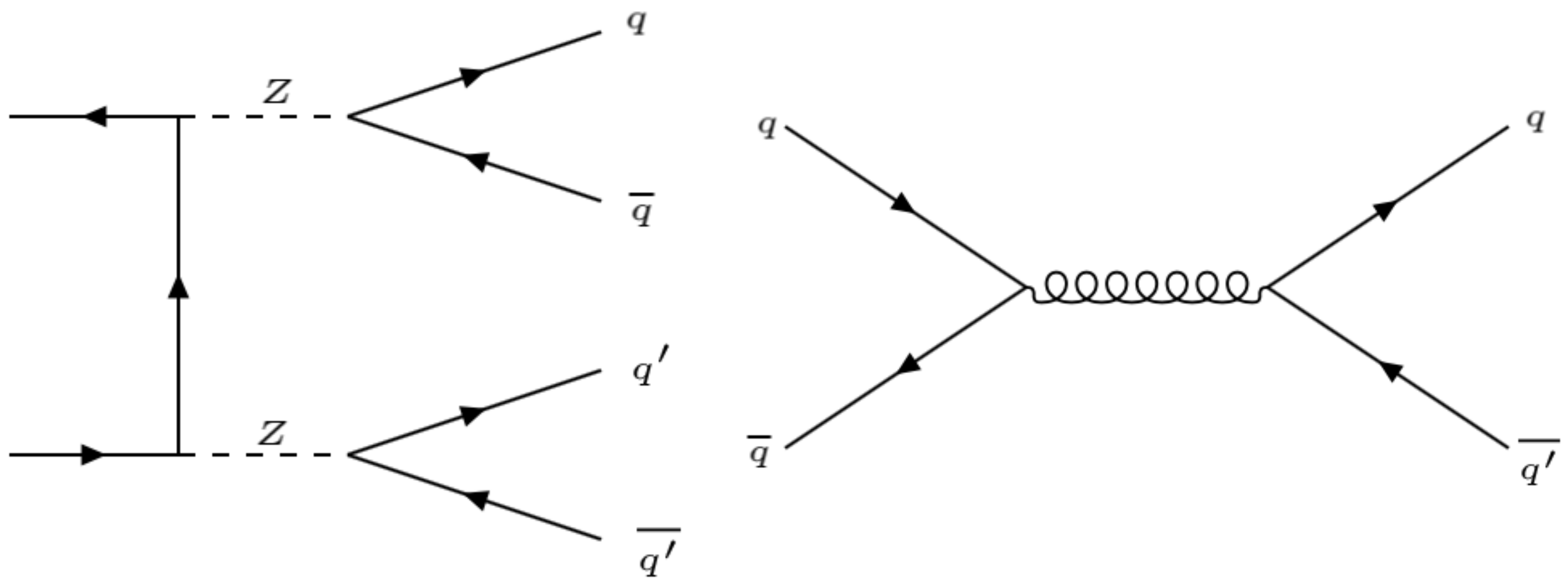
Most relevant features are same as 2DCNN.

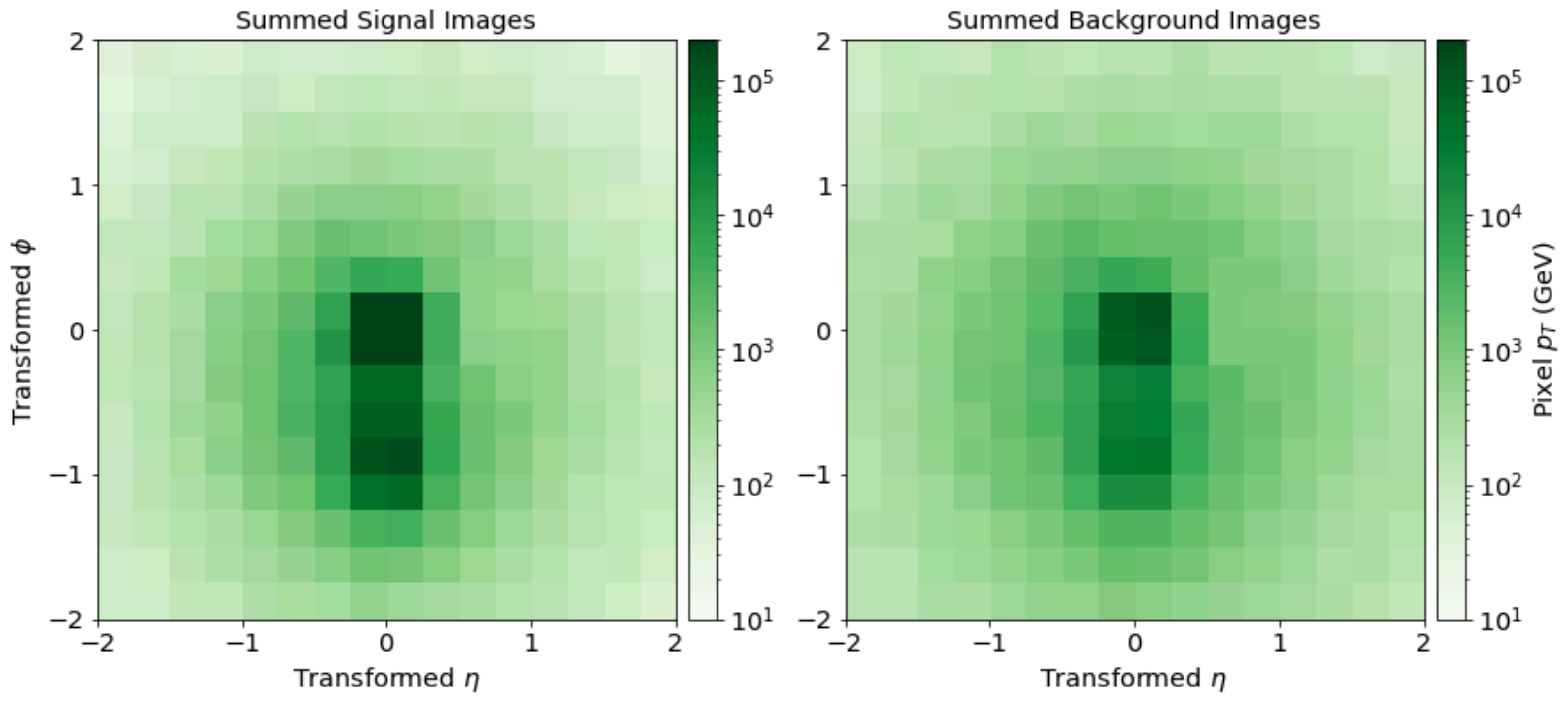
More robust “substructure” within relevance of the top two variables.

PARTICLE SIMULATION

PYTHIA GENERATION

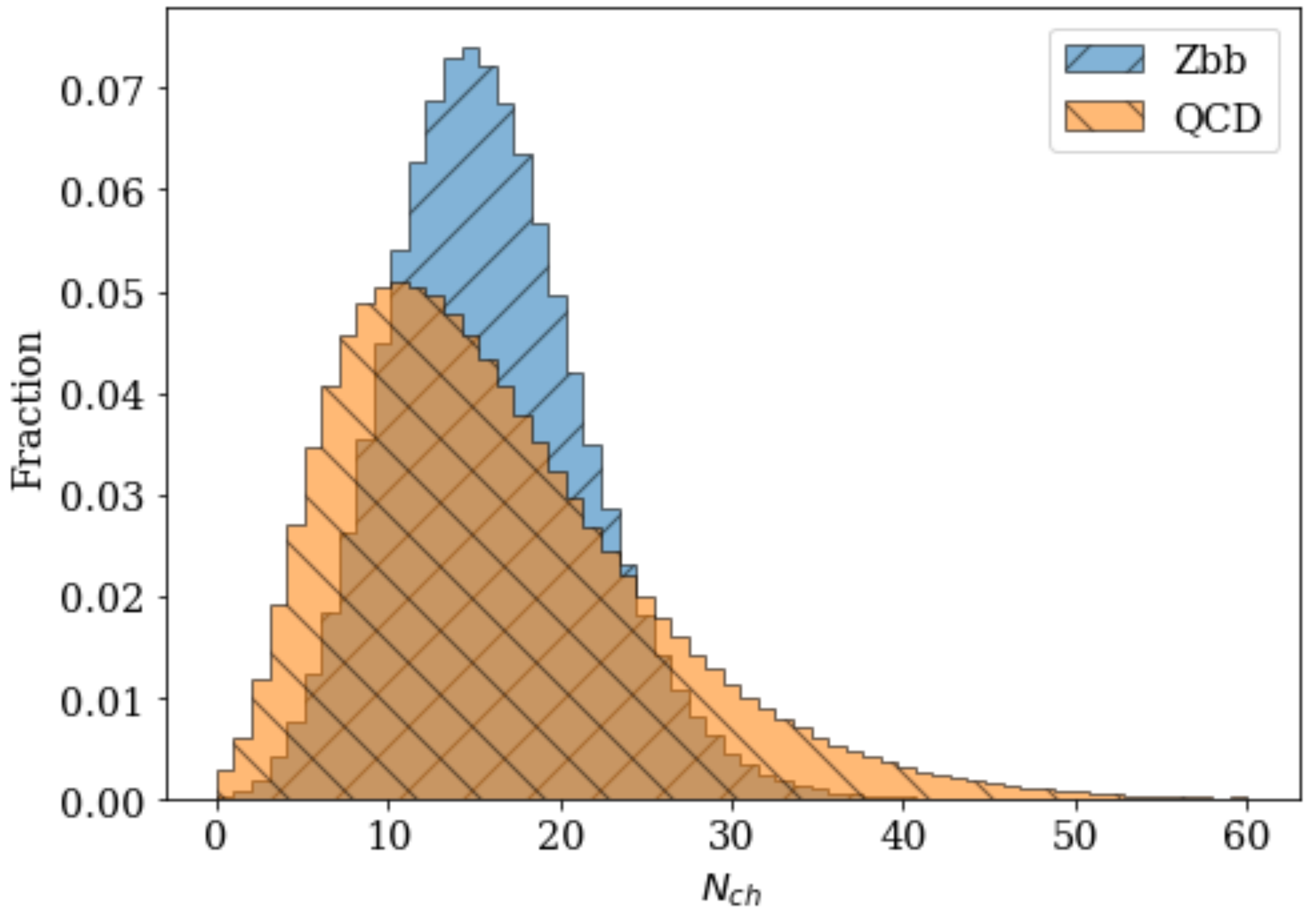
- Simulated with pythia8, SM ZZ and QCD
- AK8 jets from fastjet
- pT > 200 GeV
- mMDT from fastjet-contrib
  - $z = 0.1, \beta = 0$
- Preprocessing for images: rotation and scaling so that lower pT subjet is always at (0,-1), and normalize inputs w.r.t. jet pT, parity flip

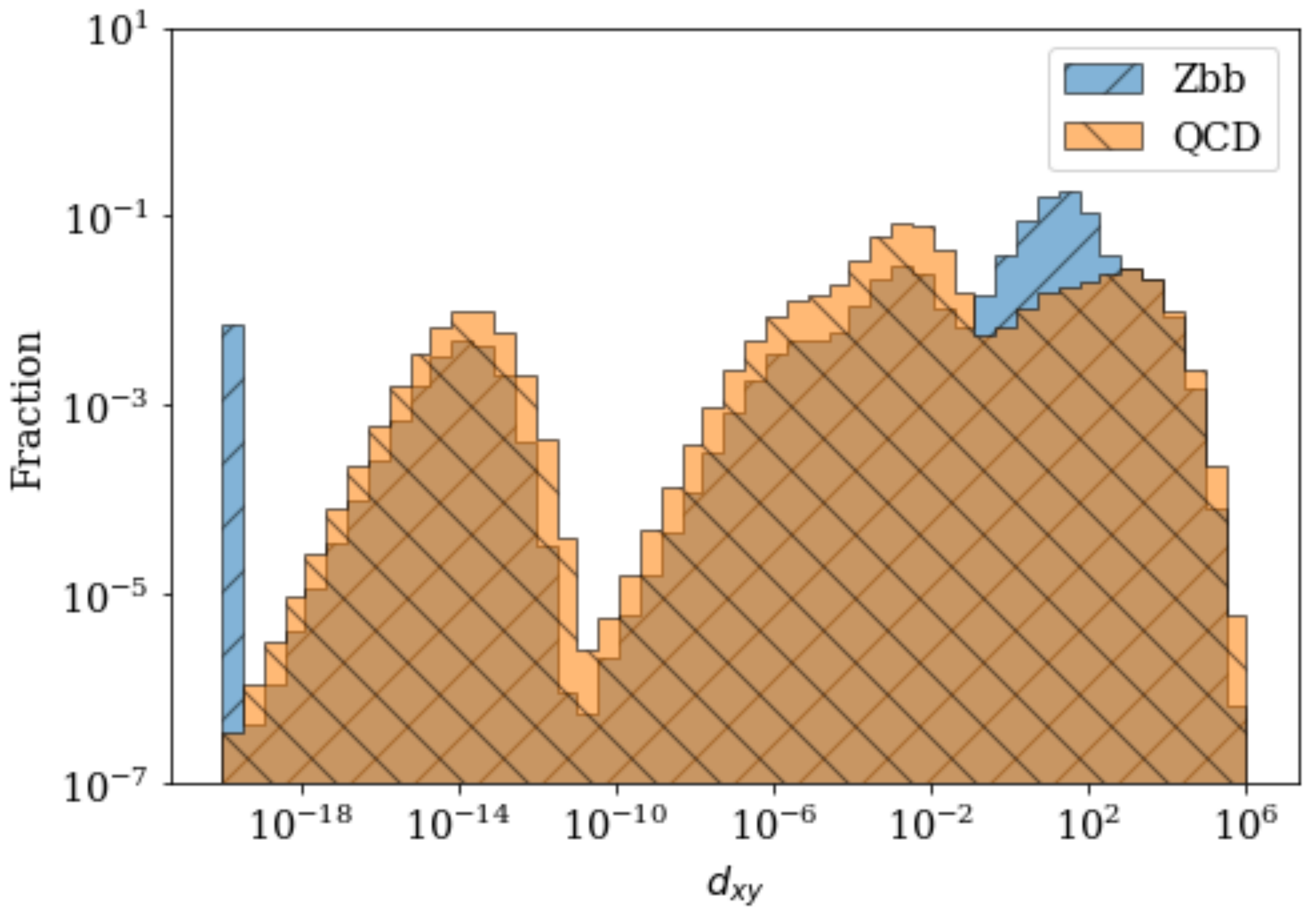


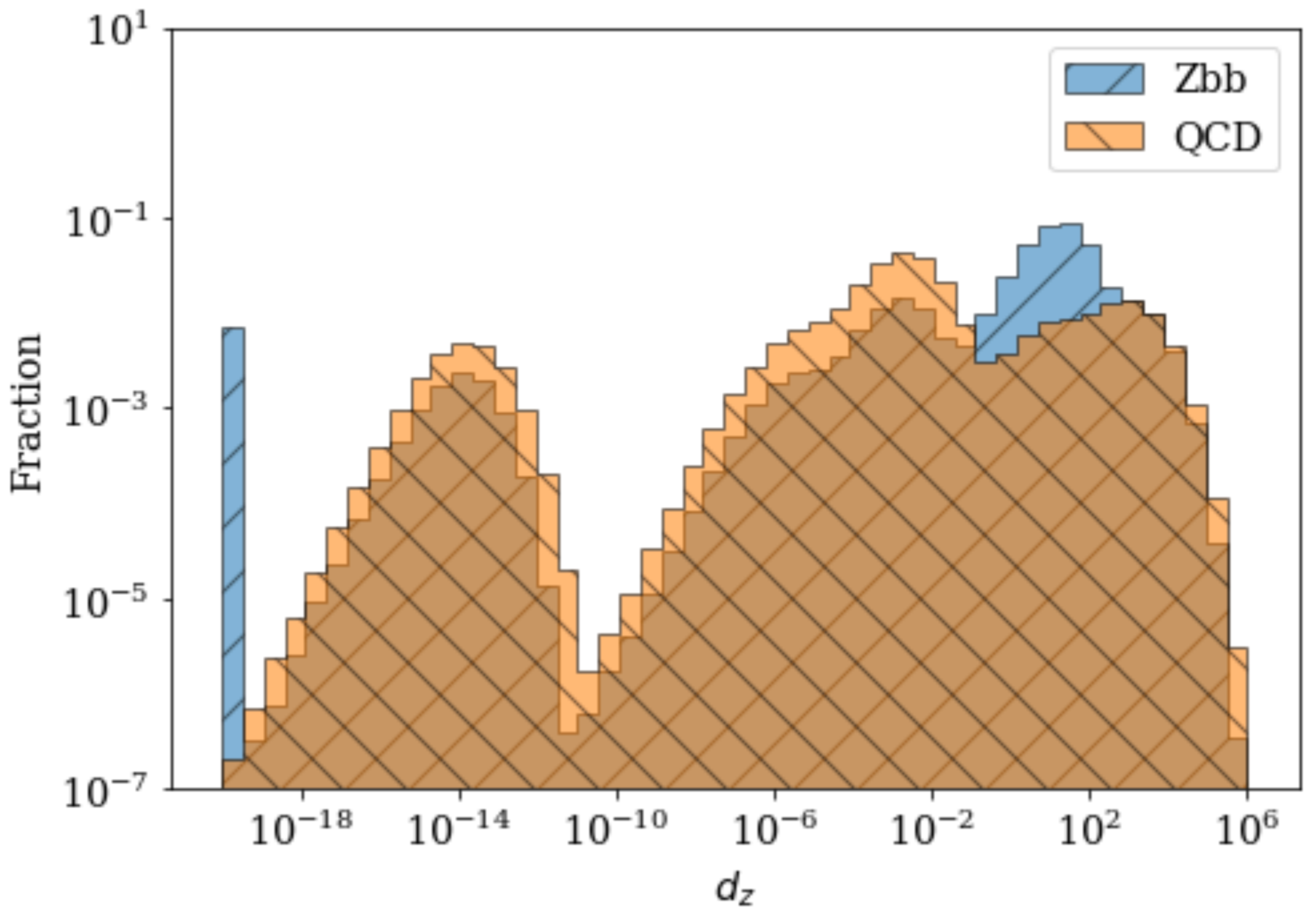


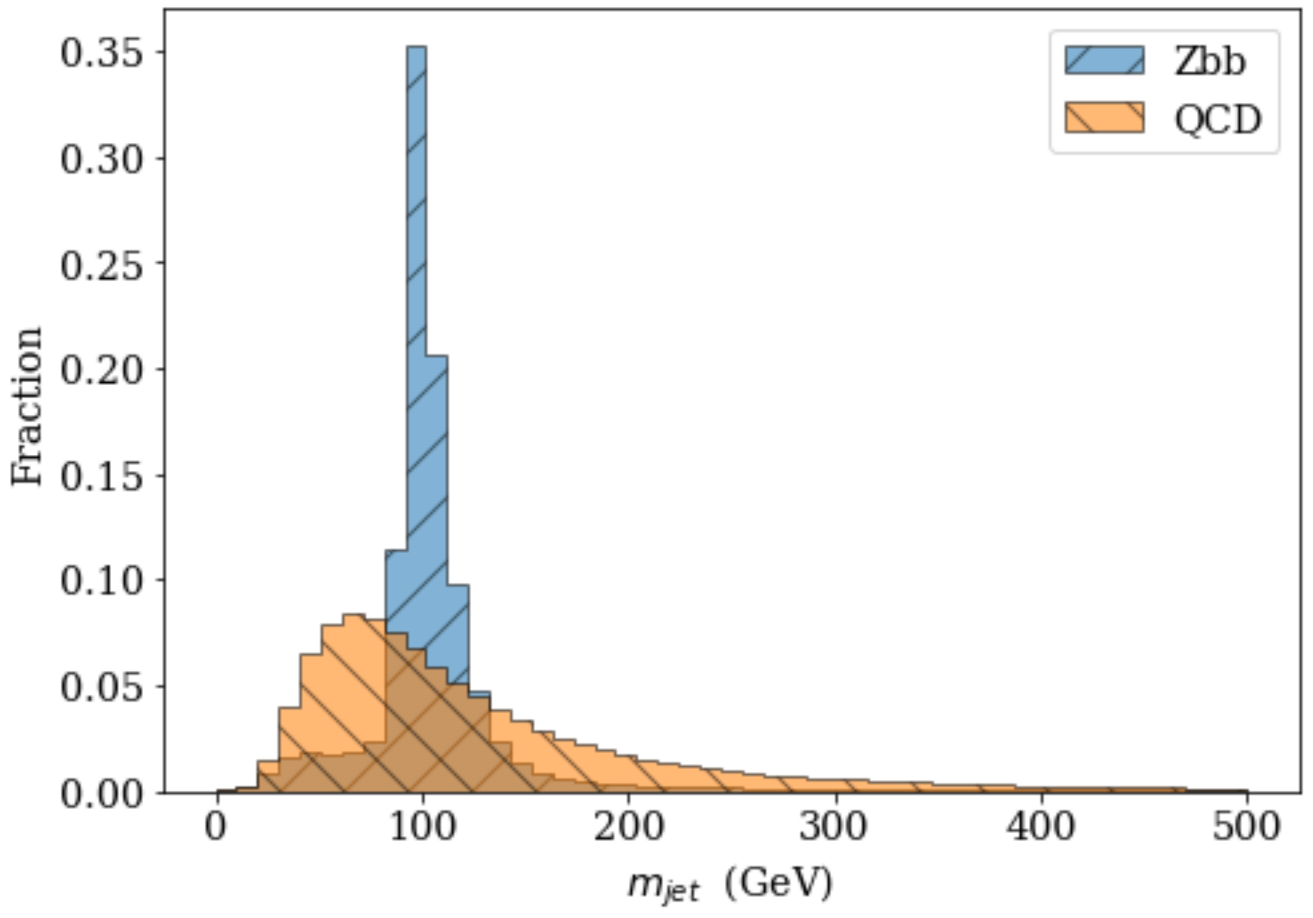
PARTICLE MODEL

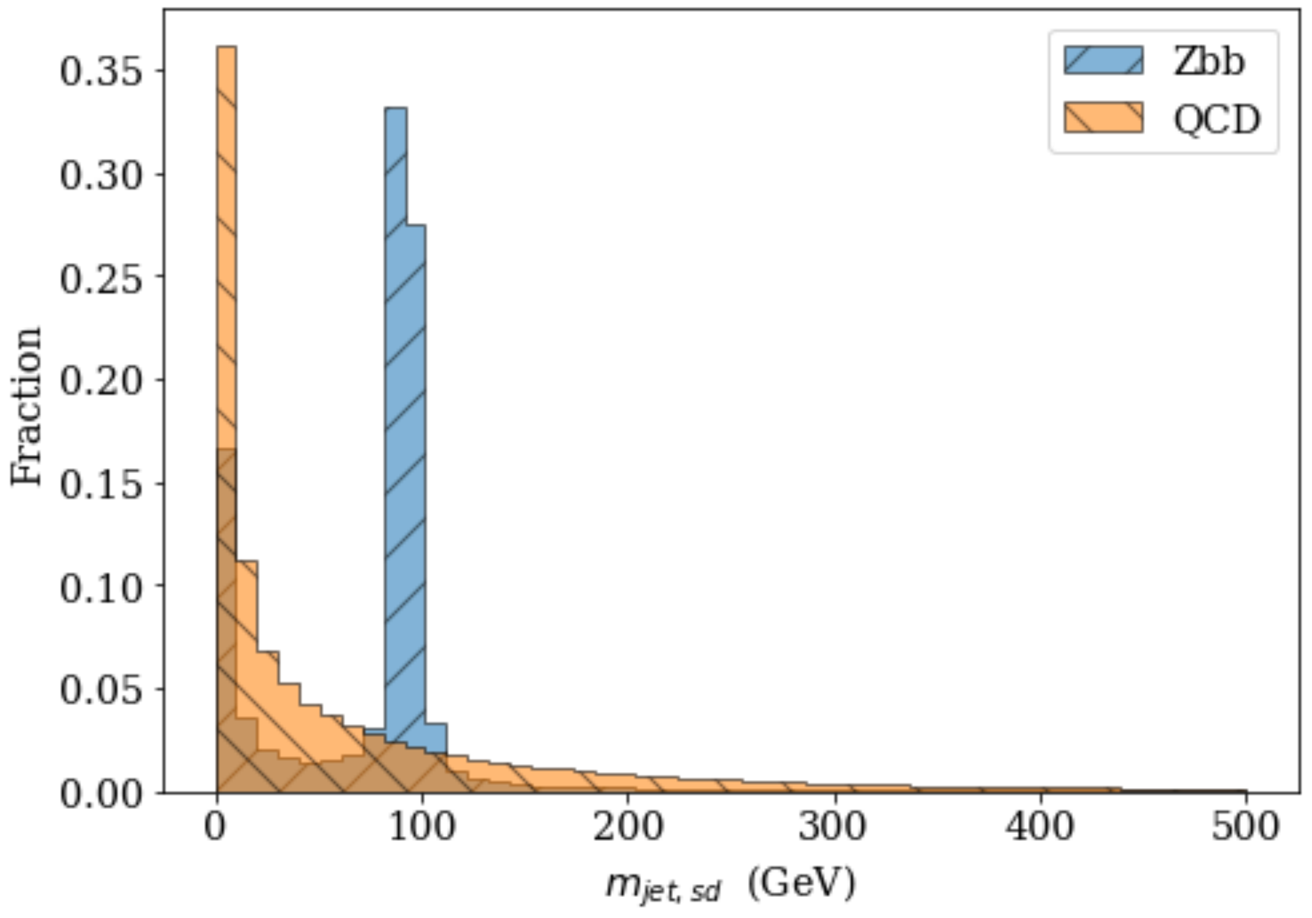
- **Use same network structures as Toy Model, replacing inputs with equivalent counterparts.**

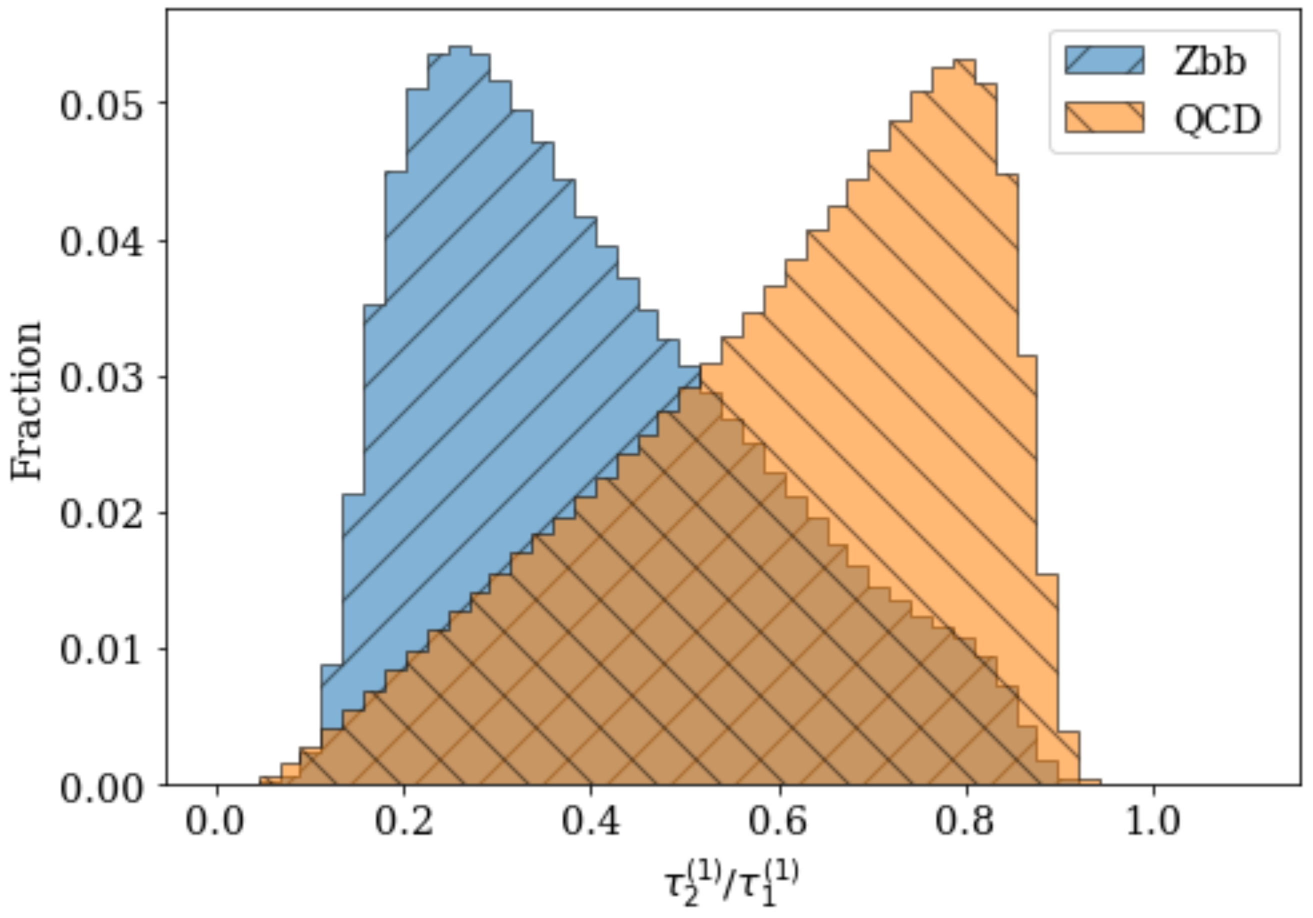




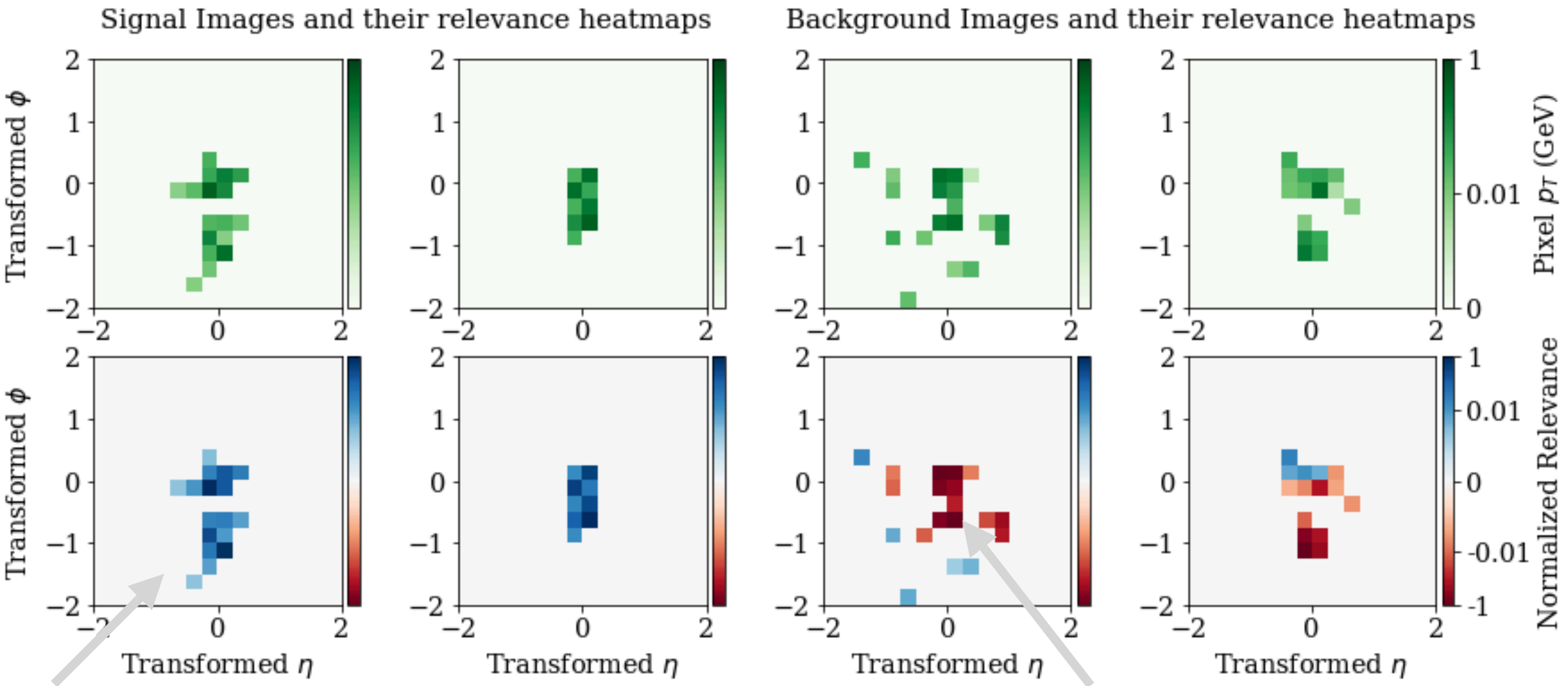








LRP HEATMAPS

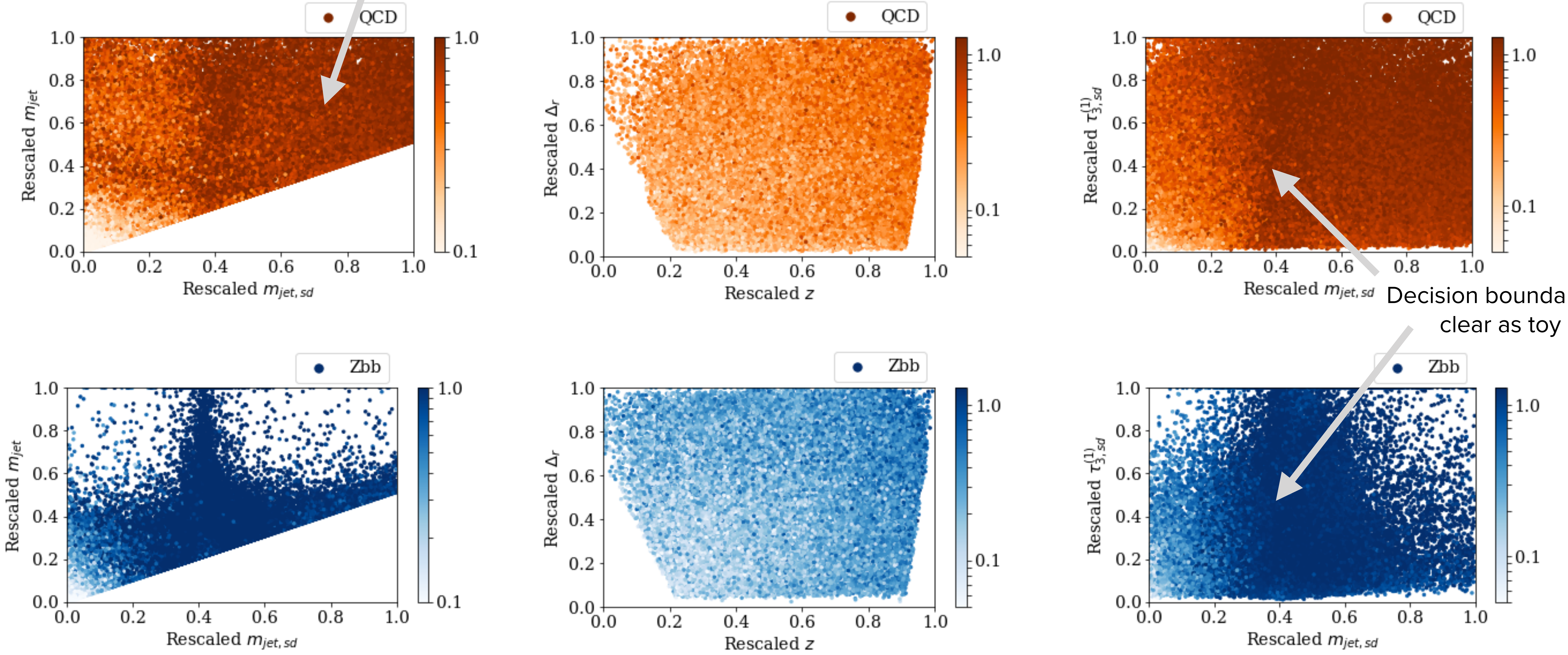


Signal is given mostly positive relevance, primarily along ϕ axis.

Background is given mostly negative relevance, and is more dispersed.

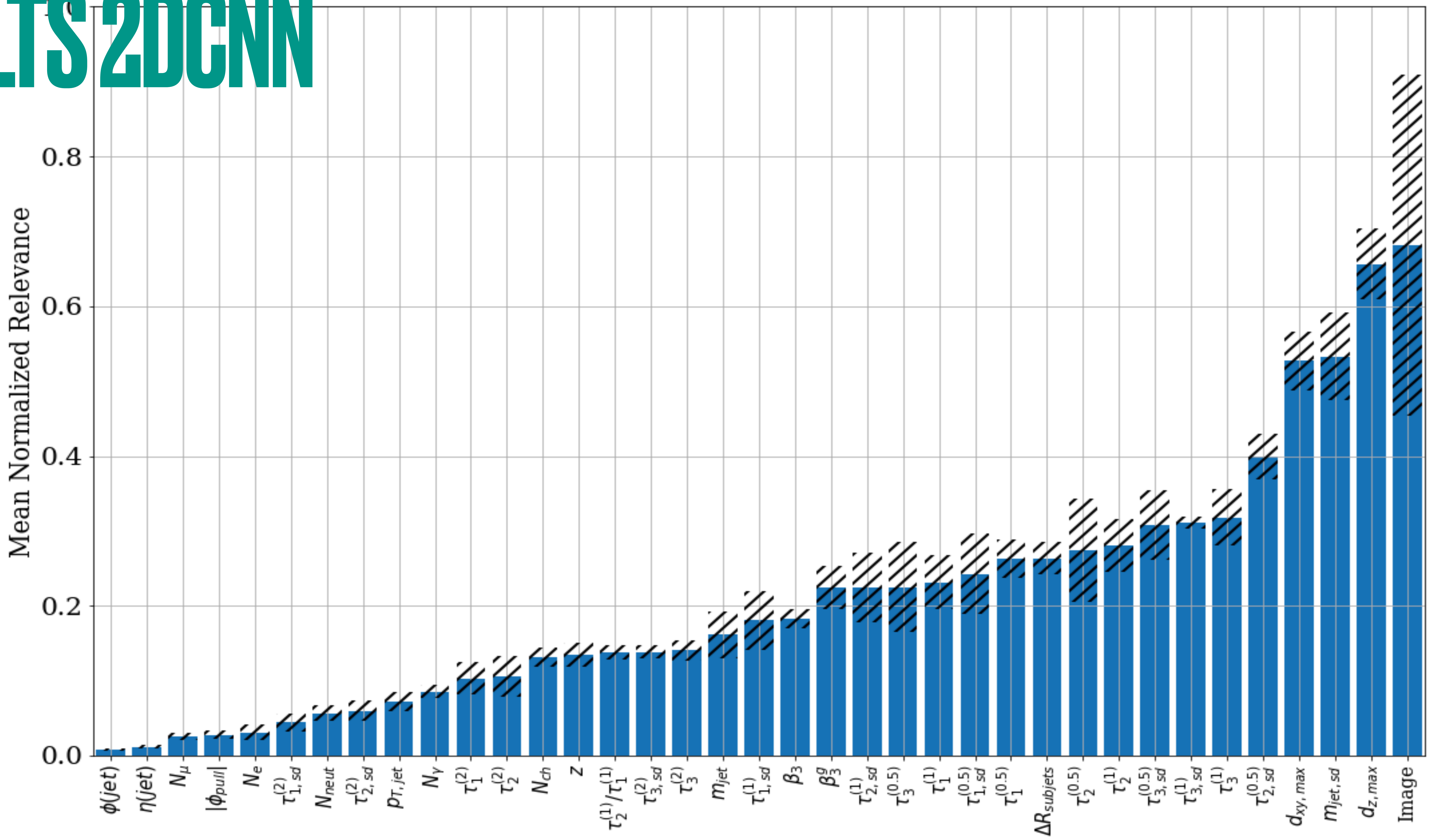
2D SCATTER REPRESENTATIONS

Darker markers correspond to higher abs. relevance scores.

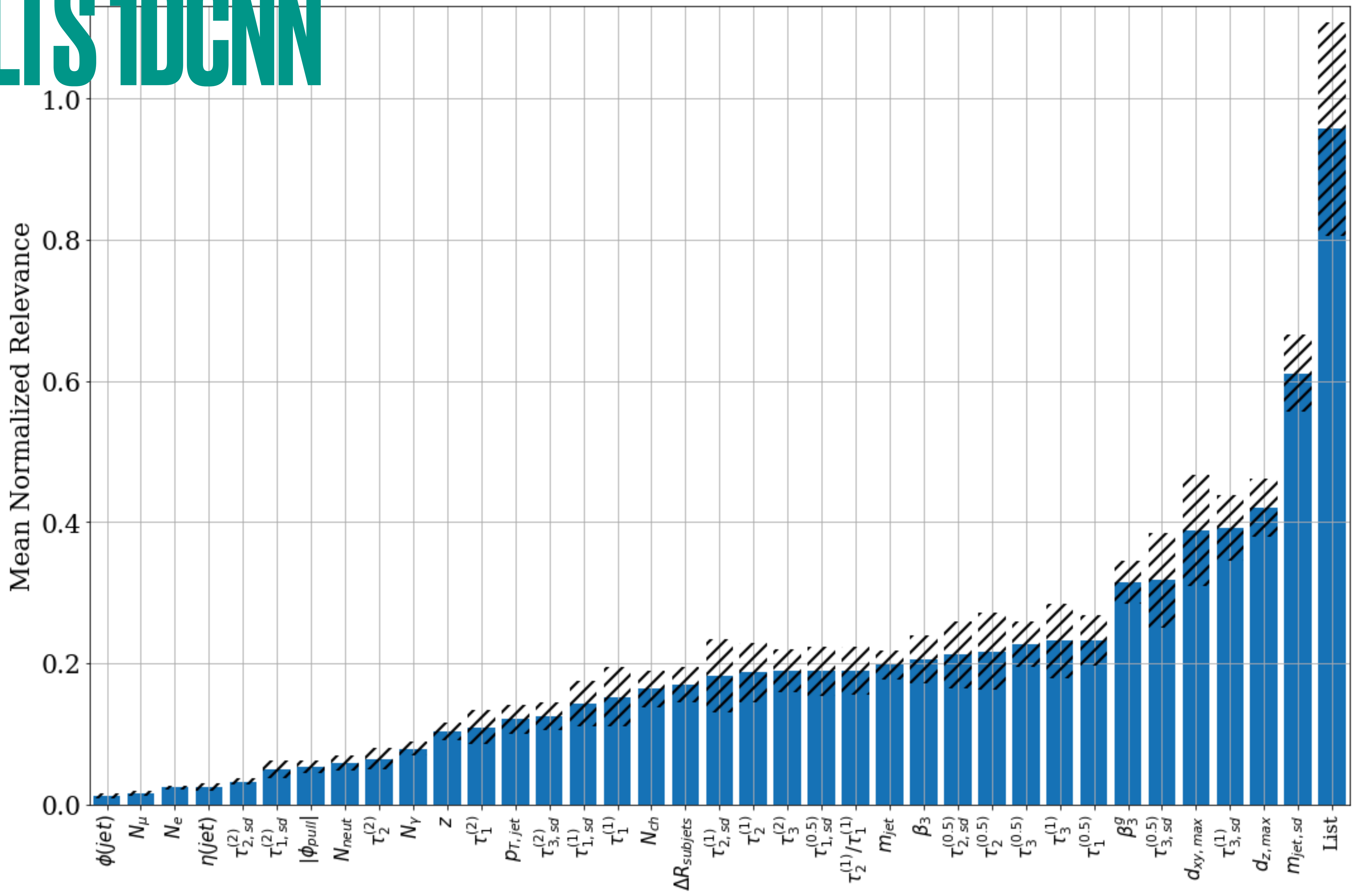


Decision boundaries not as clear as toy case.

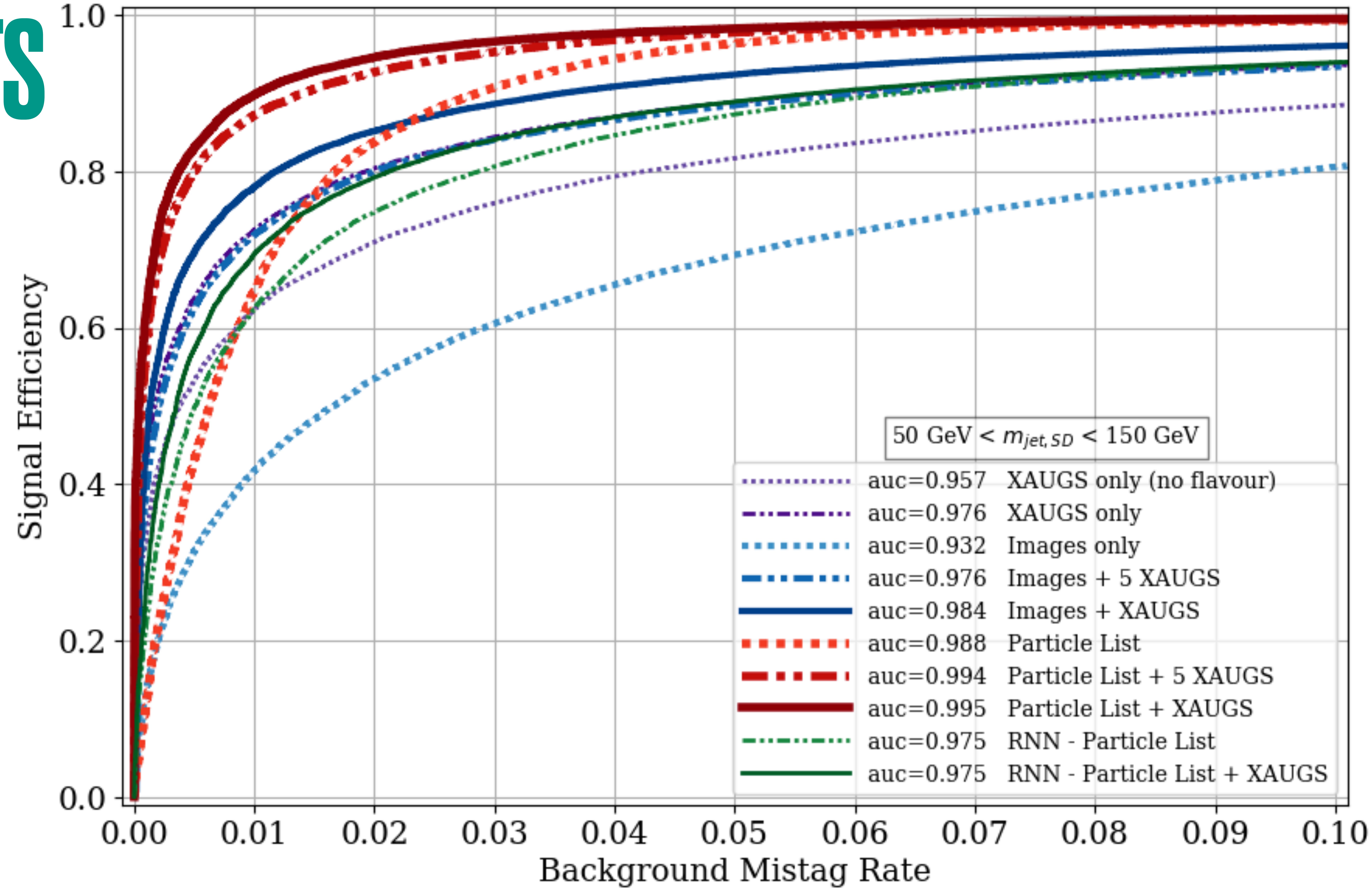
RESULTS 2DCNN



RESULTS 1DCNN



RESULTS



1.0
0.8
0.6
0.4
0.2
0.0
Signal Efficiency
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
Background Mistag Rate
50 GeV < $m_{jet,SD}$ < 150 GeV
auc=0.957 XAUGS only (no flavour)
auc=0.976 XAUGS only
auc=0.932 Images only
auc=0.976 Images + 5 XAUGS
auc=0.984 Images + XAUGS
auc=0.988 Particle List
auc=0.994 Particle List + 5 XAUGS
auc=0.995 Particle List + XAUGS
auc=0.975 RNN - Particle List
auc=0.975 RNN - Particle List + XAUGS

CONCLUSIONS

- Introducing XAUG variables and performing LRP can shed light on network decisions and relevant subspaces in the training
- XAUG variables can be used to boost classification performance
- XAUG variables can capture the information of lower level networks entirely, and a set of XAUG variables can replace long lists of particle-level information while producing comparable network performance
- Use of these techniques together can be used to quantify numerical uncertainty in training of DNNs

BACKUP

INTRODUCTION

- Machine Learning (ML) is commonly used for classification of boosted jets
- Convolutional Neural Networks (CNNs) take greyscale jet images as inputs
- A special case of the CNN is a 1-dimensional CNN which takes list-like inputa
- Decision-making process of the networks is not well understood

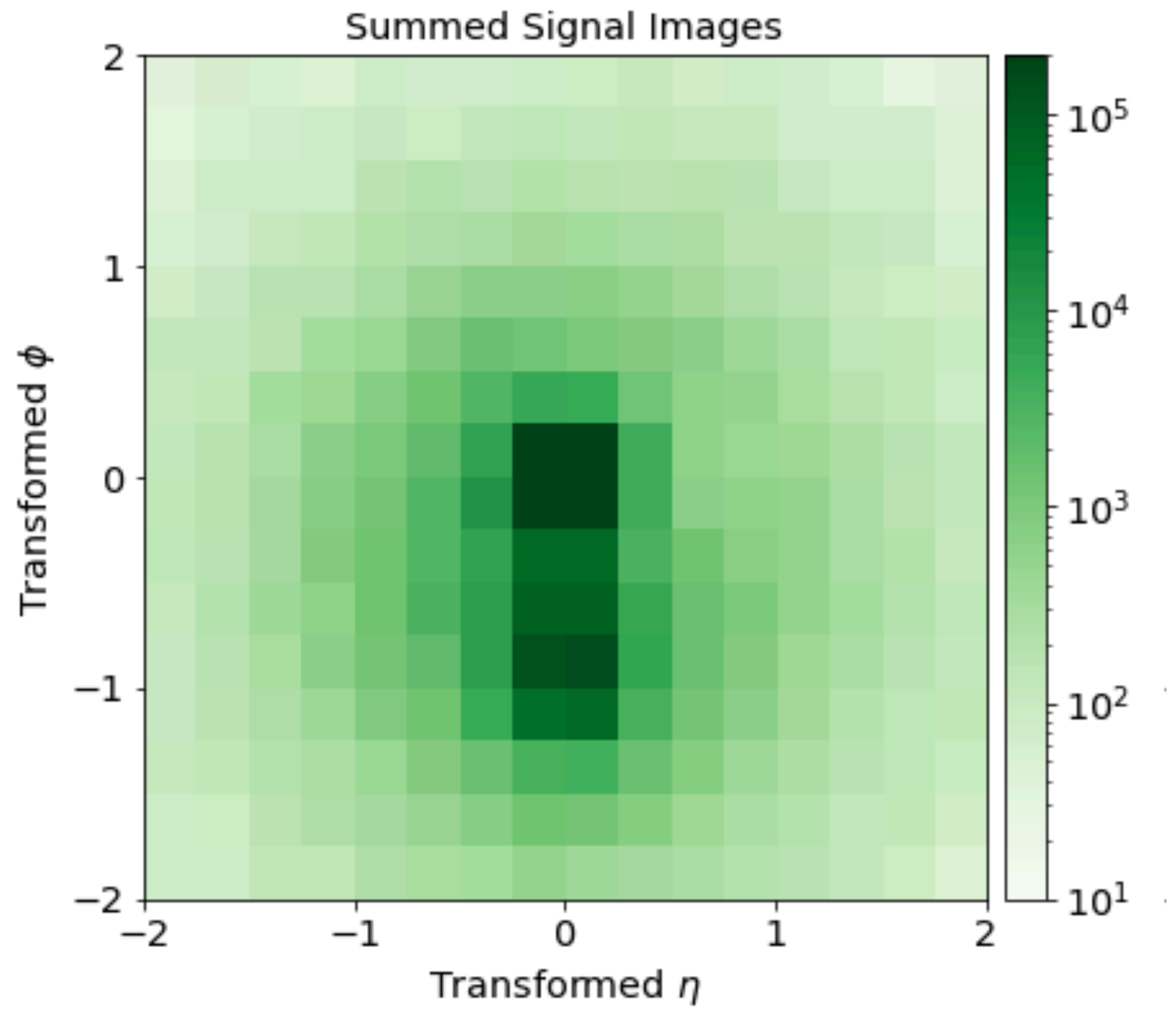


Fig 1: Greyscale jet image

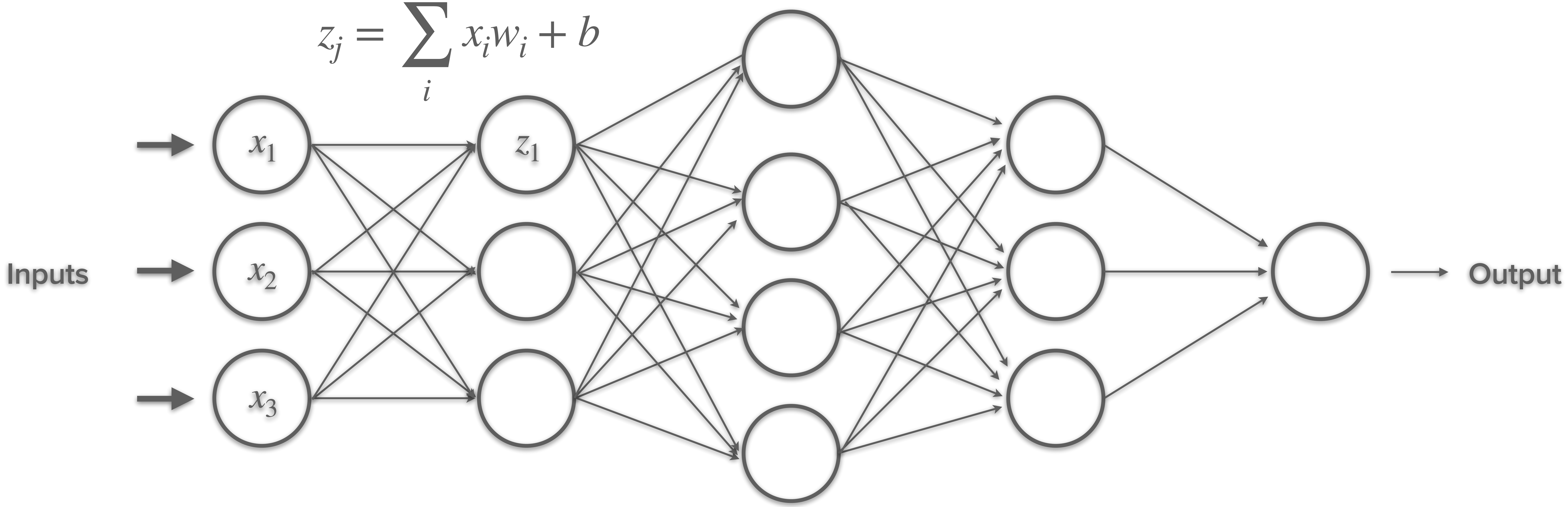
https://arxiv.org/pdf/1709.04464.pdf

MOTIVATION

- Most ML models behave as black boxes
- Augment the inputs to various types of jet classifying NNs with expert variables
- Extract classifying information using Layerwise Relevance Propogation (LRP)
- Understand what subset of information from the inputs and expert variables is relevant to the NN

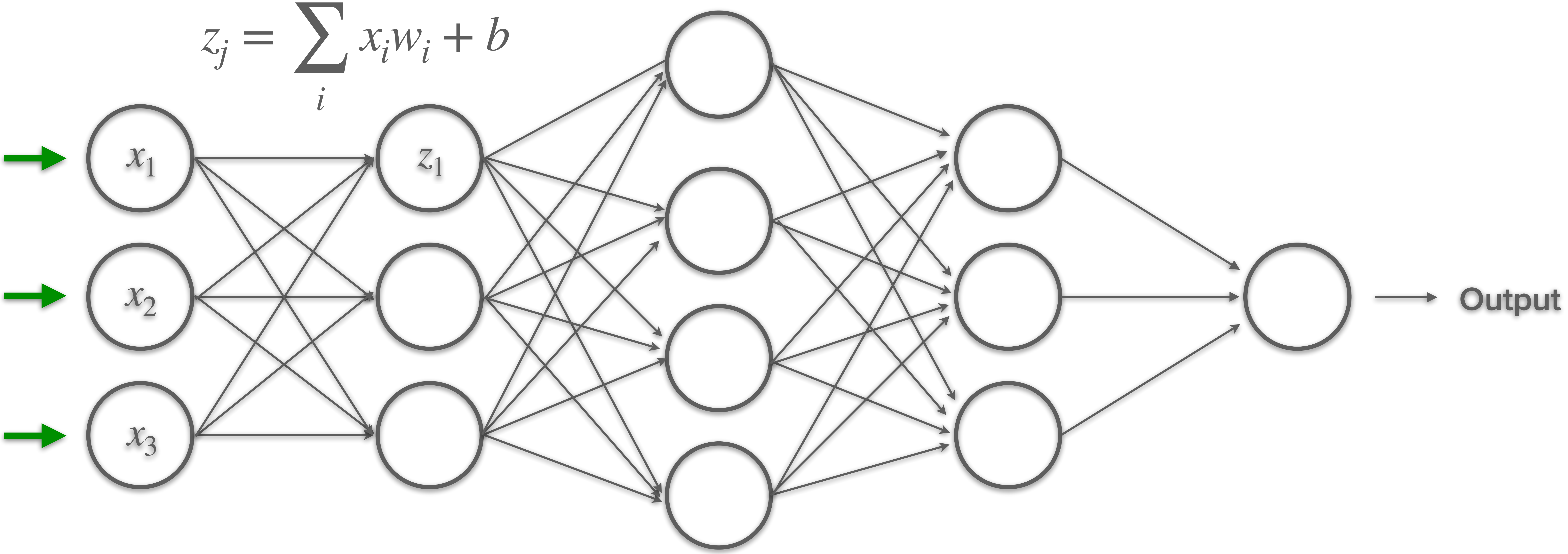
DNN FORWARD PROPAGATION

$$z_j = \sum_i x_i w_i + b$$



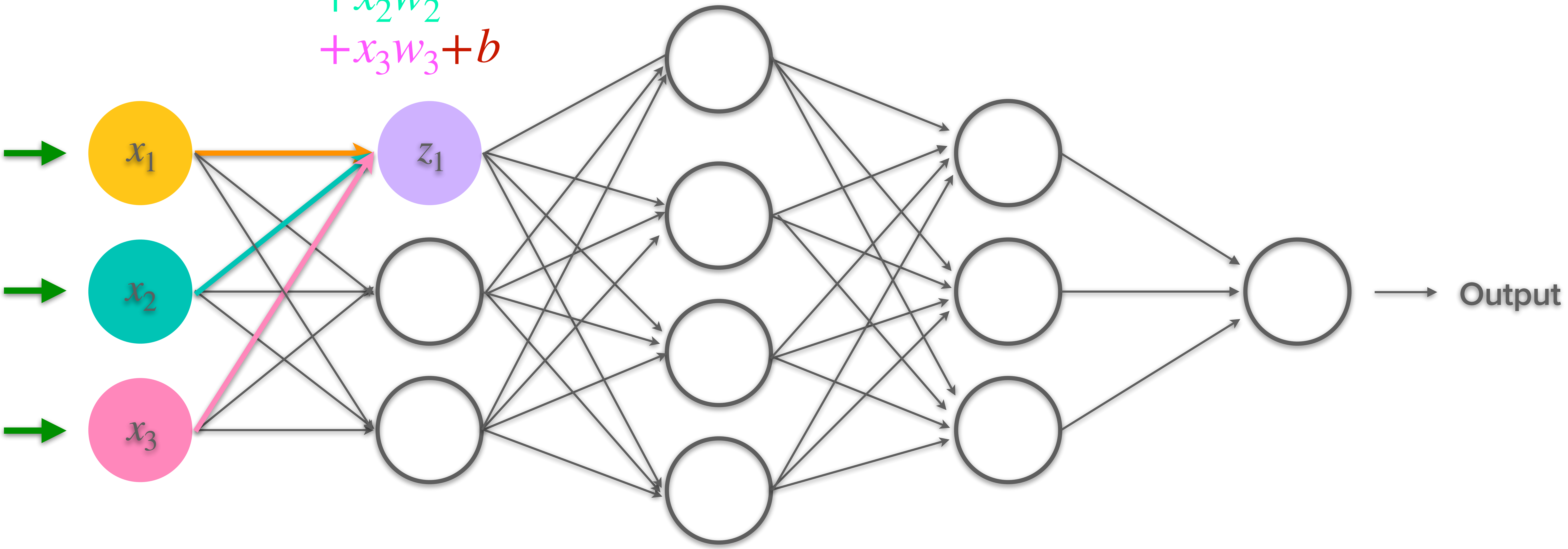
DNN FORWARD PROPAGATION

$$z_j = \sum_i x_i w_i + b$$

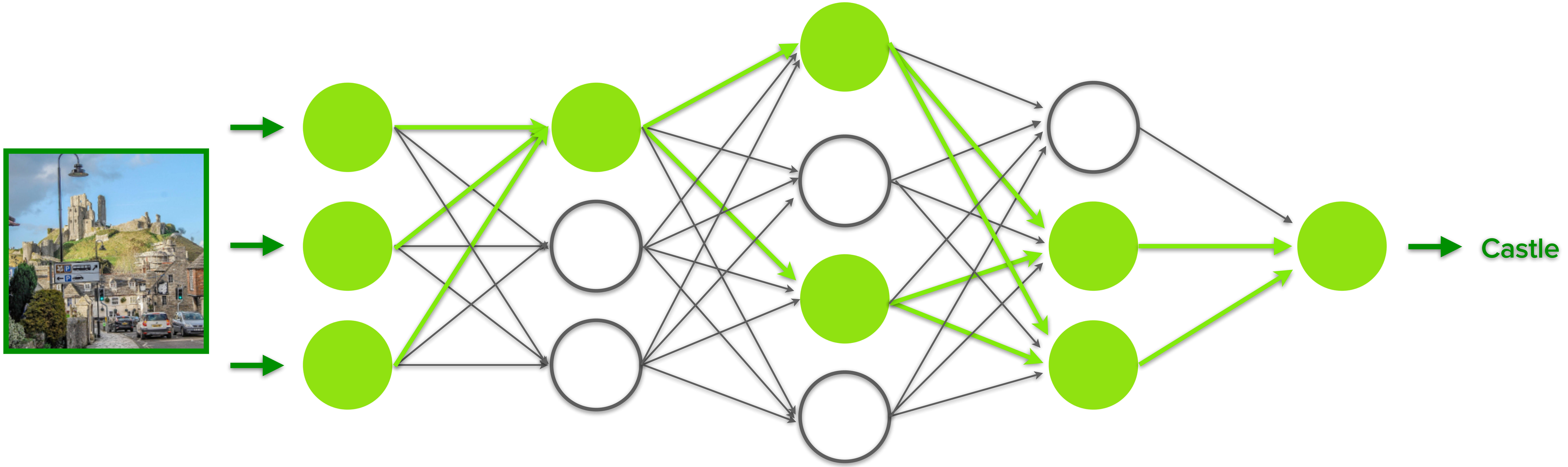


DNN FORWARD PROPAGATION

$$z_1 = x_1 w_1 + x_2 w_2 + x_3 w_3 + b$$

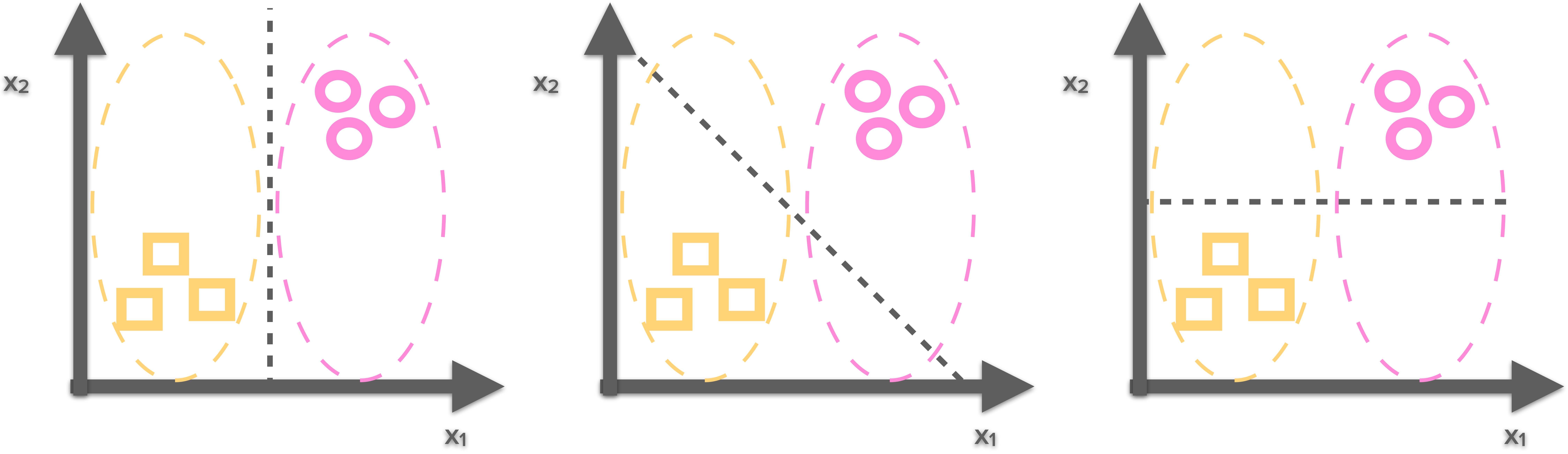


DNN FORWARD PROPAGATION



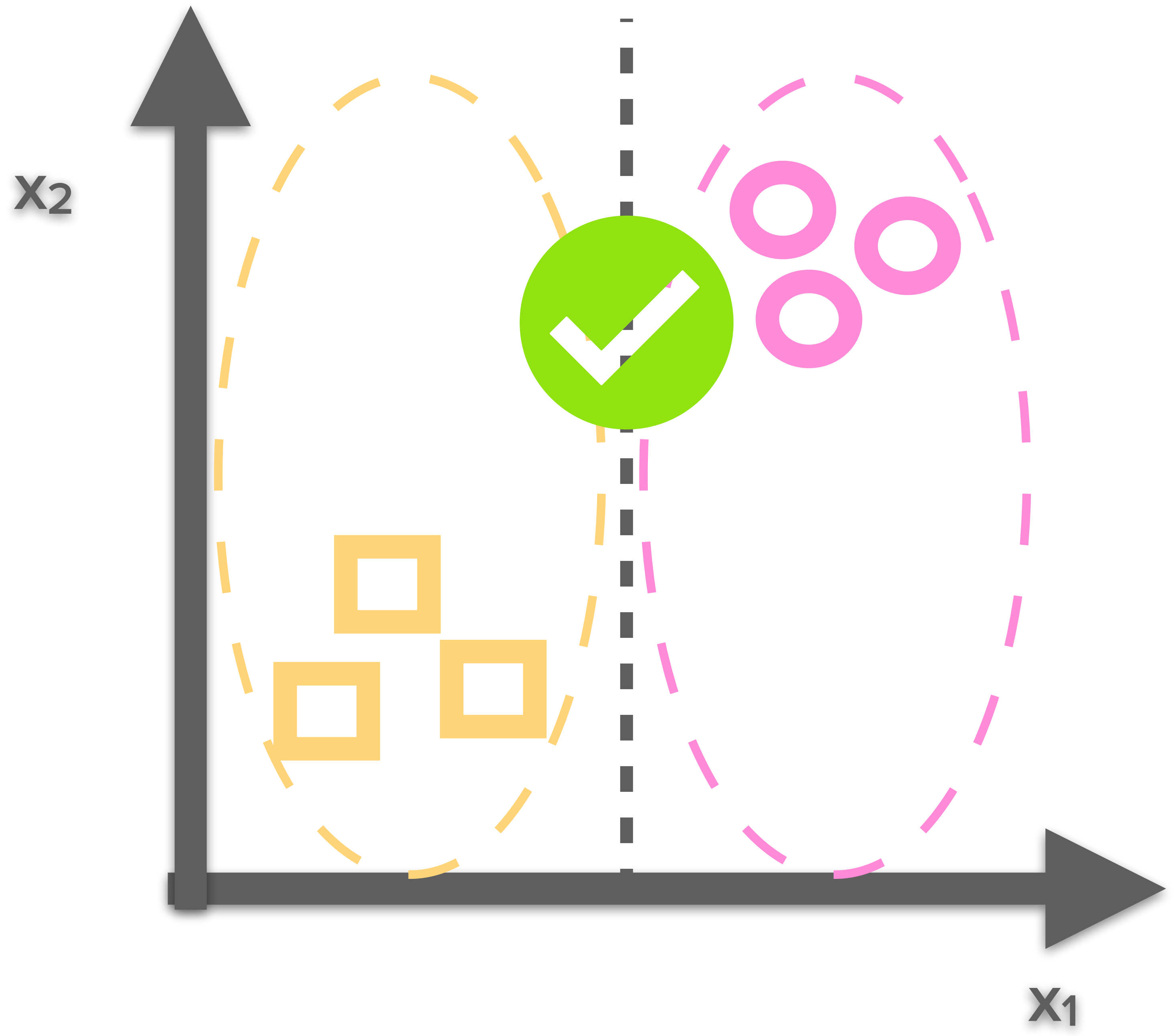
EXPLAINING THE DECISION

- Want to ensure that predictions are supported by meaningful patterns in the data.



EXPLAINING THE DECISION

- LRP is one technique that can be used to tease out if the networks learned patterns are following the intended categorisation.



LRP PROPAGATION RULES

- LRP-z:
  - Redistributes the relevance in proportion to the contributions to the neuron activation.
  - Gradient X Input → Noisy

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k$$

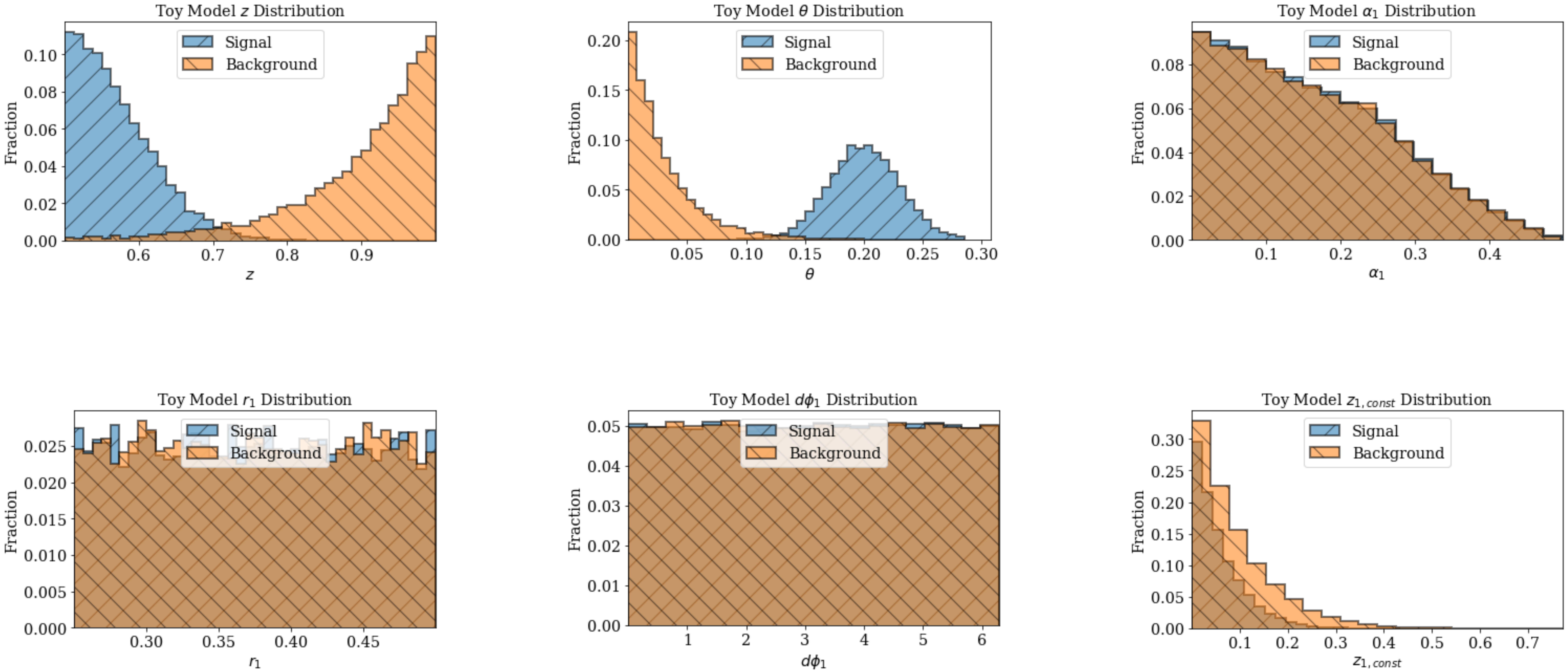
- LRP-ϵ:
  - ϵ absorbs some relevance for weak and/or contradictory contributions.
  - For large ϵ only salient explanation factors survive the absorption → Less Noisy

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k$$

- LRP-$\alpha 1 \beta 0$:
  - Limiting effect on how large positive and negative relevance can grow → Stable Explanations
  - $\alpha(\beta)$ controls by how much positive(negative) contributions are favored.

$$R_j = \sum_k \left(\alpha \frac{(a_j w_{jk})^+}{\sum_{0,j} (a_j w_{jk})^+} - \beta \frac{(a_j w_{jk})^-}{\sum_{0,j} (a_j w_{jk})^-} \right) R_k$$

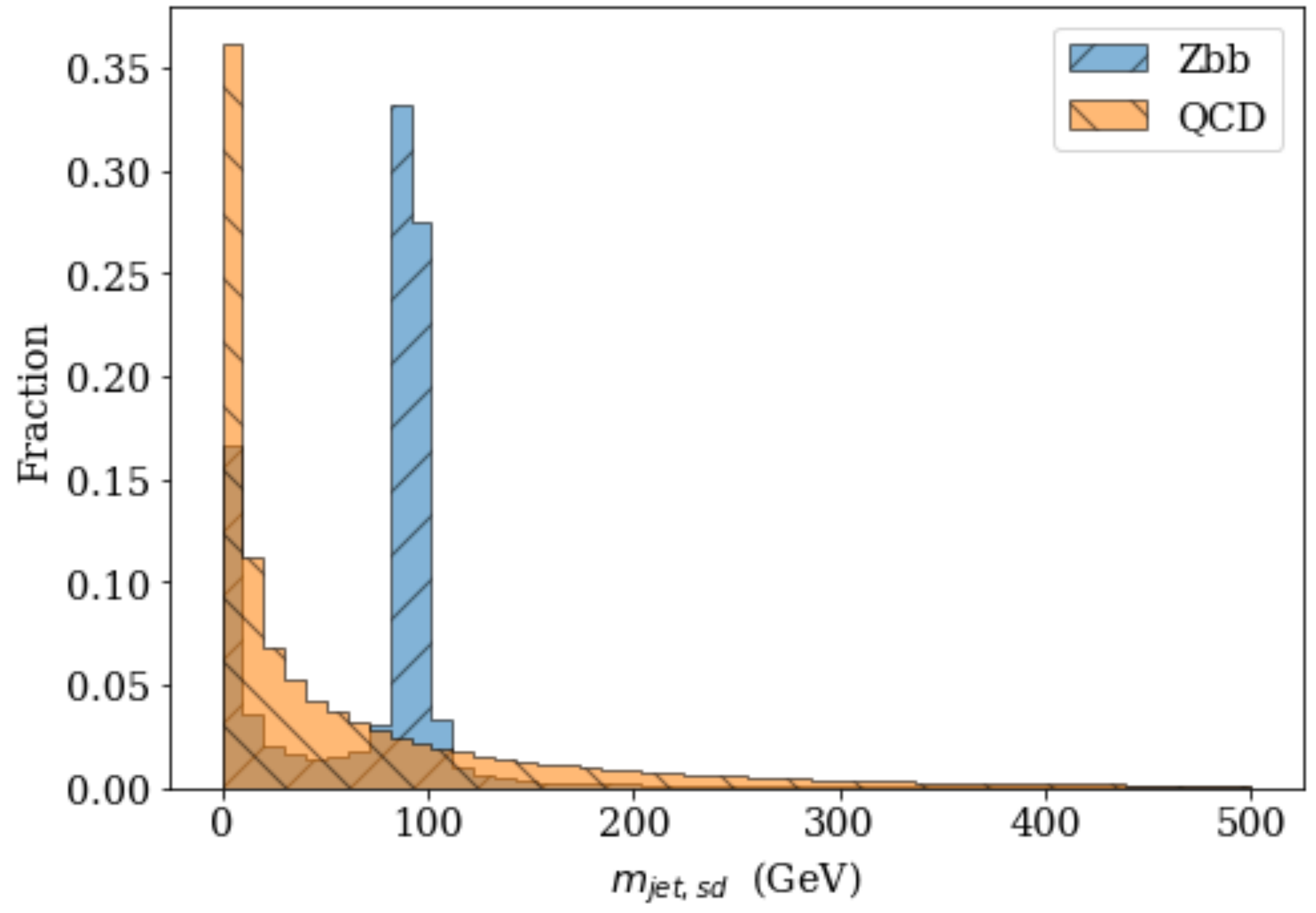
TOY MODEL



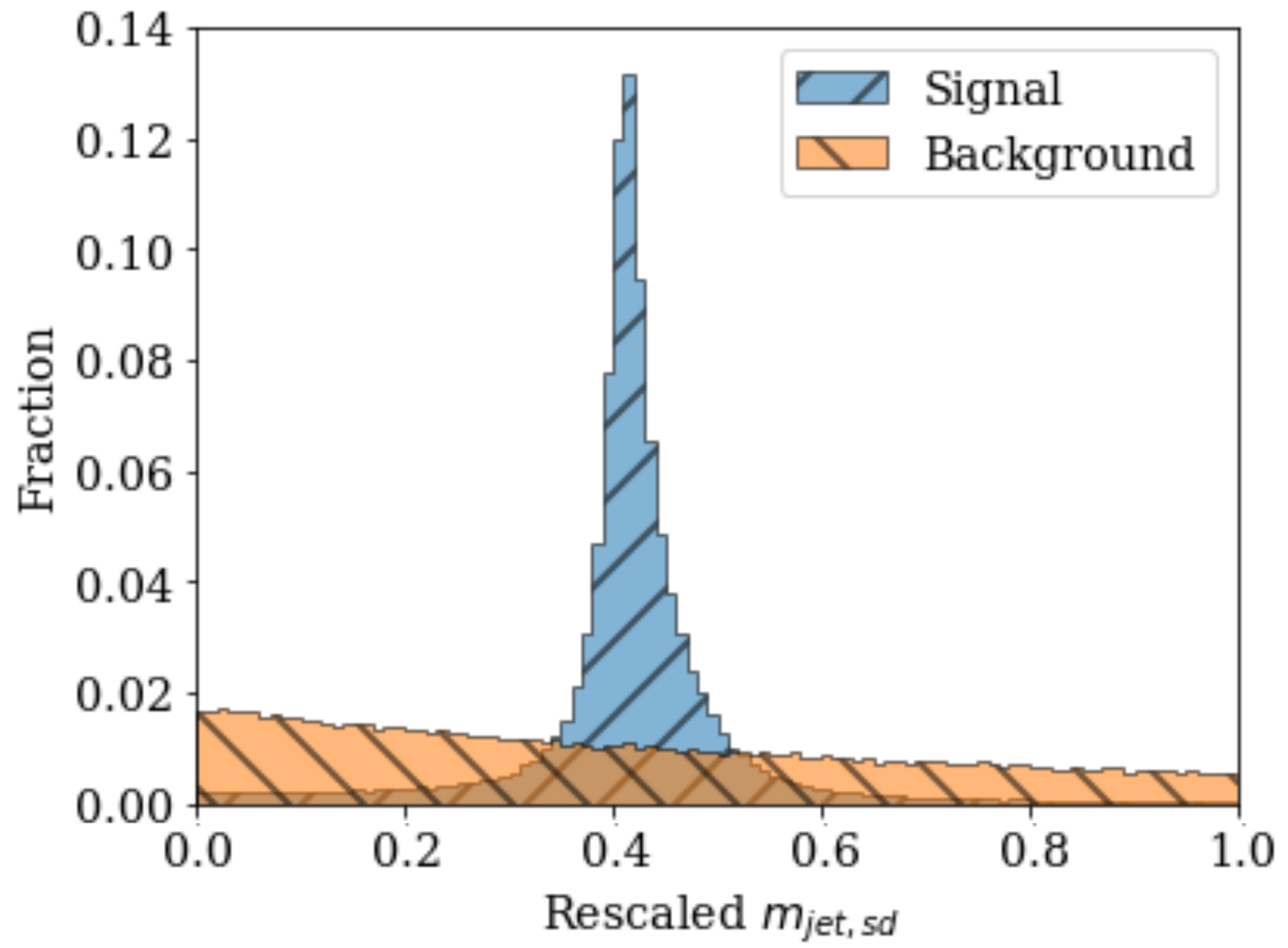
Toy Model z Distribution
Signal
Background
Fraction
z
Toy Model θ Distribution
θ
Toy Model α_1 Distribution
α_1
Toy Model r_1 Distribution
r_1
Toy Model $d\phi_1$ Distribution
$d\phi_1$
Toy Model $z_{1,const}$ Distribution
$z_{1,const}$

PREPROCESSING

1. Cut on softdrop mass: keep jets with m_{SD} 50-150 GeV
2. Numerical rescaling
   1. Rebin outliers to $mean + 3(std)$ and $mean - 3(std)$
   2. Input distributions are then rescaled from 0 to 1: $$\frac{x - x_{min}}{x_{max} - x_{min}}$$

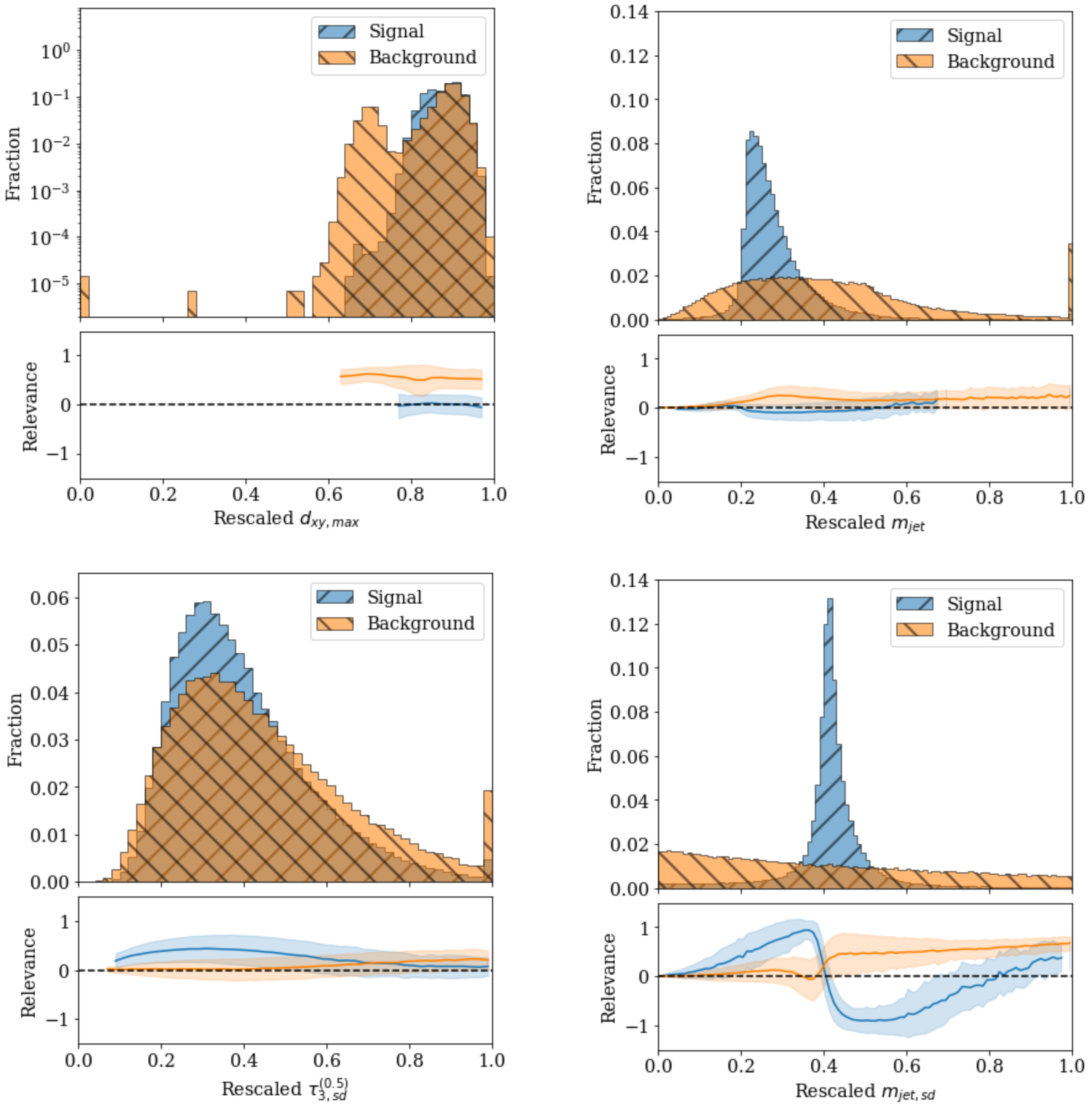


Mass cut + rescaling

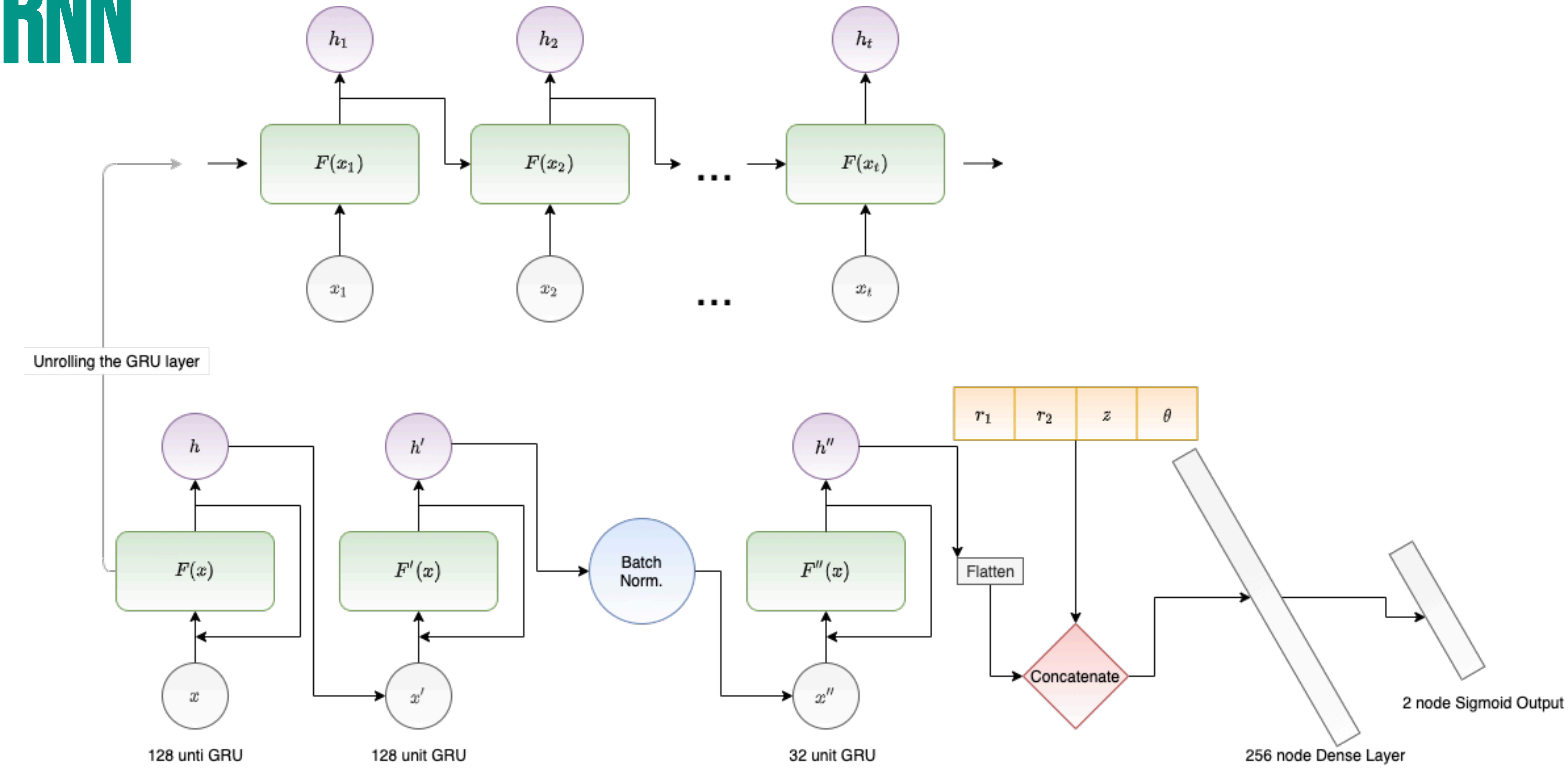


PARTICLE MODEL PROFILE PLOTS

- Profiles do not show a clear decision boundary, prompting the creation higher dimensional plots



RNN



h_1
h_2
h_t
$F(x_1)$
$F(x_2)$
$F(x_t)$
...
x_1
x_2
x_t
...
Unrolling the GRU layer
r_1
r_2
z
θ
h
h'
h''
$F(x)$
$F'(x)$
Batch Norm.
$F''(x)$
Flatten
Concatenate
x
x'
x''
128 unti GRU
128 unit GRU
32 unit GRU
2 node Sigmoid Output
256 node Dense Layer

PARTICLE LIST INPUTS

Variable
$log(p_T)$
$log(p_T/p_{T_{jet}})$
$log(E)$
$\|\eta\|$
$\Delta\phi(jet)$
$\Delta\eta(jet)$
$\Delta R(jet)$
$\Delta R(subjet1)$
$\Delta R(subjet2)$
Charge q
isMuon
isElectron
isPhoton
isChargedHadron
isNeutralHadron
d_{xy}
d_z